EXTRACTING INFORMATION FROM INTERVAL DATA USING SYMBOLIC PRINCIPAL COMPONENT ANALYSIS

M.R. Oliveira[1], M. Vilela[2], A. Pacheco[3], R. Valadas[4], P. Salvador[5]
[1,2,3]*CEMAT and* [1,2,3]*DM,* [4]*Instituto Superior Técnico, Universidade de Lisboa*
[5]*Universidade de Aveiro*
[4,5]*Instituto de Telecomunicações*
[1,2,3,4]*Lisbon and* [5]*Aveiro, PORTUGAL*
e-mail: [1]rosario.oliveira@tecnico.ulisboa.pt,
[2]margarida.azeitona@tecnico.ulisboa.pt,
[3]apacheco@math.tecnico.ulisboa.pt,
[4]rui.valadas@tecnico.ulisboa.pt, [5]salvador@av.it.pt

Abstract

We address the definition of symbolic variance and covariance for random interval-valued variables, and present four known symbolic principal component estimation methods using a common insightful framework. In addition, we provide a simple explicit formula for the scores of the symbolic principal components, equivalent to the representation by Maximum Covering Area Rectangle. Furthermore, the analysis of a real dataset leads to a meaningful characterization of Internet traffic applications.

1 Introduction

The low cost of information storage combined with recent advances in search and retrieval technologies has made huge amounts of data available, the so-called *big data* explosion. New statistical analysis techniques are now required to deal with the volume and complexity of this data. One promising technique is Symbolic Data Analysis (SDA), introduced in the late 1980s by Edwin Diday.

In conventional data analysis, the variables that characterize an object can only take single values. SDA introduces symbolic random variables which can take values over complex data structures like lists, intervals, histograms or even distributions. Symbolic data may exist on their own right or may result from the aggregation of a base dataset according to the researchers interest.

For example, suppose that our goal is to characterize the ages of university teachers. The variable that records the teachers' age will have as many observations as teachers, and these can differ among universities. Let us assume that a given university has 1000 teachers, and the values $\omega_1, \ldots, \omega_{1000}$ are the teachers' ages. SDA calls these values *micro-data*. In conventional statistical analysis, the universities would have to be characterized by single-valued variables, e.g. the mean teachers' age. SDA can deal with more complex data structures, called *macro-data*. For example, the teachers' age can be aggregated into one interval or various intervals. Our main interest in this paper is on interval-valued data, where macro-data corresponds to the interval between minimum and maximum of micro-data values: $[a, b] = [\min\{\omega_1, \ldots, \omega_{1000}\}, \max\{\omega_1, \ldots, \omega_{1000}\}]$.

The paper is organized as follows. Section 2 presents basic descriptive statistics, including symbolic variances and covariances, for interval-valued data. Section 3 introduces Symbolic Principal Component Analysis (SPCA) for interval-valued data. Section 4 uses SPCA on the analysis of Internet data produced by six different Internet applications. Finally, some conclusions are drawn in Section 5.

2 Basic descriptive statistics

There have been several proposals for definitions of symbolic versions of sample mean, variance, covariance, and correlation, according to various types of symbolic data and including interval-valued data [1].

We assume that the collected interval-valued data are realizations of random vectors. As such, we consider a random interval-valued vector $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, where $X_j = [A_j, B_j]$, with A_j and B_j being random variables verifying $P(A_j \leq B_j) = 1$, denotes the j-th random interval-valued variable of $\boldsymbol{X}$. Even though this is the common representation of random interval-valued variables, we follow the approach of [2, 3, 6] and write the intervals X_j in terms of their centers, $C_j = (A_j + B_j)/2$, and their ranges, $R_j = B_j - A_j$. This choice leads to a clear interpretation of an interval in terms of its "location" on the real line along with its length; moreover it enables for the unification of several results in the literature (cf. [2, 3, 6] and references therein). Likewise, the random vector $\boldsymbol{X}$ is equivalently represented by the random vector of centers, $\boldsymbol{C} = (C_1, \ldots, C_p)^t$, and the random vector of ranges, $\boldsymbol{R} = (R_1, \ldots, R_p)^t$.

Let $(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n)^t$ and $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)^t$ denote the vectors of centers and ranges obtained from a random sample of size n from $\boldsymbol{X}$, where $\boldsymbol{C}_i = (C_{i1}, \ldots, C_{ip})^t$ and $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{ip})^t$ characterizes the i-th entity or object of the sample. In this setting, a natural proposal for sample symbolic mean of the interval-valued variable X_j is to use the traditional sample mean of the centers, $\overline{X}_j = \overline{C}_j$ with $\overline{C}_j = \sum_{i=1}^n C_{ij}/n$.

As concerns the sample symbolic variance of the interval-valued variable X_j, we express the proposals available in the literature as the sum of two components, the first accounting for the variability of the associated centers and the second for the size of the associated ranges, in the form

$$S_{jj}^{(\alpha)} = \sum_{i=1}^{n} \frac{\left(C_{ij} - \overline{C}_j\right)^2}{n} + \alpha \sum_{i=1}^{n} \frac{R_{ij}^2}{n}, \tag{1}$$

with the nonnegative weight α accounting for the relevance given to the ranges. In particular, we address three cases, with respective values 0, 1/4, 1/12 for the weight α. The first case ($\alpha = 0$) ignores the values of the ranges, simply turning the symbolic variance into the variance of the centers. Concerning the second case ($\alpha = 1/4$), we note that as $R_{ij}/2$ represents the radius of the interval associated with i-th entity, measured on the j-th random interval-valued variable, $\sum_{i=1}^n R_{ij}^2/(4n)$ may be interpreted as the sample second order moment of the radius of the j-th random interval-valued variable. The third case ($\alpha = 1/12$) corresponds to choosing the weight derived in [2] assuming that micro-data are uniformly distributed on the random intervals.

In the same manner, we consider proposals for the sample symbolic covariance between two interval-valued variables X_j and X_l that express it as the sum of two components, the first accounting for the sample covariance of the associated centers and the second for the size of the associated ranges, in the form

$$S_{jl}^{(\beta)} = \sum_{i=1}^{n} \frac{(C_{ij} - \overline{C}_j)(C_{il} - \overline{C}_l)}{n} + \beta \sum_{i=1}^{n} \frac{R_{ij} R_{il}}{n}, \tag{2}$$

with the nonnegative weight β accounting for the relevance given to the ranges associated to the interval-valued variables X_j and X_l.

In sequence, we may use (1)-(2) to construct a sample symbolic covariance matrix $\boldsymbol{S}^{(\alpha,\beta)}$ having on the diagonal the sample symbolic variances $S_{jj}^{(\alpha)}$, given in (1), and outside the diagonal the sample symbolic covariances $S_{jl}^{(\beta)}$, $j \neq l$, given in (2), leading to

$$\boldsymbol{S}^{(\alpha,\beta)} = \boldsymbol{S}_{CC} + (\alpha - \beta)\, \mathrm{Diag}\left(\frac{\boldsymbol{\mathcal{R}}^t \boldsymbol{\mathcal{R}}}{n}\right) + \beta\, \frac{\boldsymbol{\mathcal{R}}^t \boldsymbol{\mathcal{R}}}{n}, \tag{3}$$

with $\boldsymbol{S}_{CC}$ denoting the sample covariance matrix of the centers and $\boldsymbol{\mathcal{R}} = [R_{ij}]$ the $(n \times p)$ matrix of observed ranges. Particular cases of sample symbolic covariance matrices, $\boldsymbol{S}^{(\alpha,\beta)}$, with $\alpha \in \{0, 1/4, 1/12\}$ and $\beta = \alpha$ or $\beta = 0$, have been introduced in the literature ([2, 6] and references therein). Details about the links between these sample symbolic covariance matrices and SPCA for interval-valued data are discussed in the next section.

3 Symbolic Principal Component Analysis

Principal component analysis (PCA) is one of the most popular statistical methods to analyse real data. There have been several proposals to extend this methodology to the symbolic data analysis framework, in particular to interval-valued data. The majority of the available methods rely on a strategy called symbolic-conventional-symbolic, meaning that: (i) input data is symbolic (interval-valued, in here), (ii) the data is converted into conventional, to which the conventional PCA method is applied, and (iii) at the end, the PCA results are turned into symbolic, usually by a method called Maximum Covering Area Rectangle (MCAR), see [3, 6] and references therein for details.

We study four SPCA methods: CPCA, VPCA, CIPCA, and SymCovPCA. CPCA and VPCA corresponds to the first SPCA methods proposed in the literature and the last two are among the most recent alternatives. All these four methods rely on the symbolic-conventional-symbolic strategy, which can be specified as follows: (i) compute the associated $(p \times p)$ sample symbolic covariance matrix $\boldsymbol{S}^{(\alpha,\beta)}$ (see Table 1 and [3]); (ii) obtain the spectral decomposition of $\boldsymbol{S}^{(\alpha,\beta)}$, as in the conventional PCA, and (iii) transform the conventional scores into symbolic scores, e.g. using MCAR.

Note that $\boldsymbol{S}^{(\frac{1}{4},0)}$ and $\boldsymbol{S}^{(\frac{1}{12},0)}$ (see Table 1) are covariance matrices that use a definition of symbolic variance of an interval-valued variable that does not coincide with the definition of symbolic covariance between the same interval-valued variable and itself.

Table 1: Sample symbolic covariance matrices $\boldsymbol{S}^{(\alpha,\beta)}$, defined by the combination of several proposals for symbolic variances and covariances along with the corresponding SPCA method.

(α,β)	$\boldsymbol{S}^{(\alpha,\beta)}$	SPCA Method
(0,0)	$\boldsymbol{S}_{CC}$	CPCA
$(\frac{1}{4},\frac{1}{4})$	$\boldsymbol{S}_{CC}+\frac{1}{4}\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}$	—
$(\frac{1}{12},\frac{1}{12})$	$\boldsymbol{S}_{CC}+\frac{1}{12}\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}$	SymCovPCA
$(\frac{1}{4},0)$	$\boldsymbol{S}_{CC}+\frac{1}{4}\text{Diag}\left(\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}\right)$	VPCA
$(\frac{1}{12},0)$	$\boldsymbol{S}_{CC}+\frac{1}{12}\text{Diag}\left(\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}\right)$	CIPCA

This violates a basic rule in the conventional framework, namely that the variance of a variable equals the covariance of the variable with itself. In spite of this fact, the CIPCA's authors, who proposed $\boldsymbol{S}^{(\frac{1}{12},0)}$ [3], argue that this is an advantage of their method.

Similarly to the conventional PCA, it may be interesting to define the SPCA based on standardized interval-valued variables, and to do so we introduce the sample correlation matrix as: $\boldsymbol{P}^{(\alpha,\beta)}=\boldsymbol{U}_{(\alpha)}^{-1}\boldsymbol{S}^{(\alpha,\beta)}\boldsymbol{U}_{(\alpha)}^{-1}$, where $\boldsymbol{U}_{(\alpha)}=\text{Diag}\left(S_{11}^{(\alpha)},\ldots,S_{pp}^{(\alpha)}\right)^{1/2}$, for $\boldsymbol{S}^{(\alpha,\beta)}=[S_{jl}^{(\alpha,\beta)}]$, where $S_{jj}^{(\alpha,\beta)}=S_{jj}^{(\alpha)}$ and $S_{jl}^{(\alpha,\beta)}=S_{jl}^{(\beta)}$, for $j\neq l$. Equivalently, $\boldsymbol{S}^{(\alpha,\beta)}=\boldsymbol{U}_{(\alpha)}\boldsymbol{P}^{(\alpha,\beta)}\boldsymbol{U}_{(\alpha)}$. Thus, SPCA methods based on standardized interval-valued variables just have to use $\boldsymbol{P}^{(\alpha,\beta)}$ instead of $\boldsymbol{S}^{(\alpha,\beta)}$.

The most common way to transform conventional objects into symbolic ones for methods following the symbolic-conventional-symbolic strategy is the MCAR representation. Following the same line of work as before, in [3] we deduced an explicit formulation of the MCAR representation in terms of centers and ranges. Furthermore, the sample scores of the i-th object on the j-th symbolic principal component (SPC), according with MCAR, are:

$$\widehat{\text{SPC}}_{ij}=\left[\hat{\boldsymbol{\gamma}}_j^t(\boldsymbol{C}_i-\hat{\boldsymbol{\mu}}_C)-\frac{1}{2}|\hat{\boldsymbol{\gamma}}_j|^t\boldsymbol{R}_i,\ \hat{\boldsymbol{\gamma}}_j^t(\boldsymbol{C}_i-\hat{\boldsymbol{\mu}}_C)+\frac{1}{2}|\hat{\boldsymbol{\gamma}}_j|^t\boldsymbol{R}_i\right], \tag{4}$$

where $\hat{\boldsymbol{\gamma}}_j$ is the j-th eigenvector of $\boldsymbol{S}^{(\alpha,\beta)}$, the sample symbolic covariance matrix under consideration, $|\hat{\boldsymbol{\gamma}}_j|=(|\hat{\gamma}_{1j}|,\ldots,|\hat{\gamma}_{pj}|)^t$, and $\hat{\boldsymbol{\mu}}_C$ is the vector of center sample means.

As a direct consequence of (4), the centers of the scores, $\hat{\boldsymbol{\gamma}}_j^t(\boldsymbol{C}_i-\hat{\boldsymbol{\mu}}_C)$, are a linear combination of the centers of the original interval-valued variables, whose weights are given by the eigenvectors of the corresponding symbolic covariance matrix. Additionally, the scores ranges, $|\hat{\boldsymbol{\gamma}}_j|^t\boldsymbol{R}_i$, are also a linear combination of the original ranges, whose weights have the same magnitude as the centers but are all positive. This formulation makes clear that MCAR' score ranges are never negative.

4 Analysis of Internet Data

In this section we illustrate the use of SPCA through a dataset of Internet traffic, typically observed in backbone networks, and measured during July 2014. Specifically, the dataset contains traffic produced by six different Internet applications, namely Web browsing (produced by HTTP), file sharing (produced by Torrent), streaming, video (YouTube), port scans (produced by NMAP), and snapshots. The first four applications correspond to regular traffic and the last two to Internet attacks. The analysis usually aims at detecting the various Internet applications within a traffic aggregate and/or the separation between regular and illicit traffic.

The dataset comprises 917 traffic objects, corresponding to packet flows of specific applications, which we call *datastreams*. For each datastream, we registered five different traffic characteristics observed in 0.1 seconds intervals, during 5 minutes. The traffic characteristics registered were the following: number of upstream packets (PUp), number of downstream packets (PDw), number of upstream bytes (BUp), number of downstream bytes (BDw), and number of active TCP sessions (Ses). Thus, each object is characterized by a total of 3000 observations per traffic characteristic, which constitutes our micro-data.

The conventional approach to analyse this data is based on summary statistics of each traffic characteristic. In particular, [4, 5] used 8 summary statistics (minimum, 1^{st} quartile, median, mean, 3^{rd} quartile, maximum, standard deviation, and median absolute deviation) for the above five traffic characteristics, giving a total of 40 variables to describe the datastreams. This approach usually requires a pre-processing step to remove irrelevant and redundant variables; Pascoal [5] used a robust feature selection method based on mutual information for that purpose.

This dataset is naturally symbolic, since each traffic characteristic is multi-valued. SDA takes into consideration the complex structure of these data, and may lead to clearer interpretation and new insights. In our case, we will use interval-valued variables for each traffic characteristic (our macro-data), instead of the 8 summary statistics listed above.

Given the nature of the data and the existence of potential atypical observations among the micro-data, we decided to trim 1% of the lower and 1% of the higher values. This was only done for the regular applications given that illicit ones have few datastreams and small variability and would be completely eliminated from the dataset, even for such small trimming percentiles. Apart from that, and following the recommendations in [4, 5], data was smoothed using a logarithm transformation ($\ln(x+1)$, to overcome the existence of zeros). SPCs were estimated using the four methods under study. The conventional analysis of the eigenvalues of the various sample symbolic covariance matrices (not shown here, see [3] for details) suggests to retain two principal components, which explain between 80.3% (CIPCA) and 95.6% (SymCovPCA) of the total sample variance associated with $\boldsymbol{S}^{(\alpha,\beta)}$.

The results obtained with CPCA and SymCovPCA are similar, and so are the results obtained with VPCA and CIPCA. Moreover, these similarities are easily explained by the expressions of Table 1. For these reasons, only the estimates associated

Table 2: Eigenvectors of the sample symbolic covariance matrices for each estimation method, called loadings.

	SymCovPCA		CIPCA	
	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
$\ln(\text{PDw}+1)$	-0.264	-0.171	-0.125	-0.059
$\ln(\text{BDw}+1)$	**-0.730**	-0.043	**-0.932**	0.337
$\ln(\text{PUp}+1)$	-0.255	-0.168	-0.113	-0.070
$\ln(\text{BUp}+1)$	**-0.571**	0.075	-0.318	**-0.937**
$\ln(\text{Ses}+1)$	-0.079	**0.967**	-0.029	-0.027

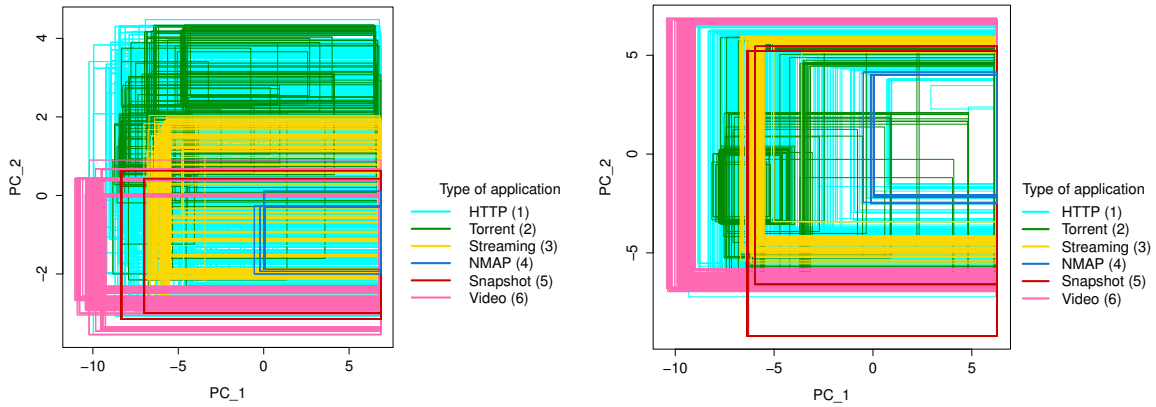
with the most recent methods (SymCovPCA and CIPCA) are shown in this paper.

Table 2 shows the loadings of the first and second SPC, obtained with SymCovPCA and CIPCA. In the case of SymCovPCA, the number of upstream and downstream bytes (BUp, BDw) have the highest loading (on absolute value) in the definition of the first SPC. Thus, the center and range of the first SPC can be interpreted as a weighted sum of the number of upstream and downstream bytes. The number of bytes is sometimes referred to as the traffic volume. For the center, the negative coefficients indicate that datastreams with high (low) number of bytes in both directions have low (high) center values on the first SPC. For the range, the coefficients are taken in absolute value, so datastreams with high (low) number of bytes in both directions have high (low) range values on the first SPC. Recall that the range expresses the inner variability of micro-data. As for the second SPC, the loading associated with number of sessions stands out. Thus, datastreams characterized by an high (low) number of sessions have high (low) center and range values on the second SPC.

The SymCovPCA scores are shown in Figure 1(a). Each datastream is represented by a rectangle, defined by the centers and ranges of the first two SPC. It can be said that the various Internet applications are, in general, well identified, since the datastreams show similar patterns for the same application. Most datastreams have a small minimum traffic volume (number of bytes), with the corresponding rectangles leaning to the right side. HTTP shows no distinctive characteristic, since the datastreams spread over all score ranges. This can be explained by the heterogeneity of user behaviours and accessed Web pages, typical of Web browsing. Torrent is concentrated on the upper part of the graph, due to its high number of sessions. The high number of sessions and large variability of the traffic volume is mostly explained by the variation on the number of available peers during traffic sharing sessions. The graph also suggests the existence of several Torrent groups, but this pattern will become clearer with the CIPCA method. The behaviour of video related with the second SPC contrasts with that of Torrent: it is concentrated in the lower part of the graph, due to its low number of sessions. Moreover, video is the application with the highest traffic volume. We may say that video datastreams are characterized by a low number of high volume sessions, and Torrent by a high number of high volume sessions. Streaming has a behaviour similar to video, but with higher number of sessions and lower traffic volume. NMAP is the application with smallest volume and variability, and has also

a relatively low number of sessions. Finally, the behaviour of snapshot is in-between video and streaming, both in terms of volume and number of sessions. Snapshot has two clear groups, that differ on the peak traffic volume, and correspond to full desktop and partial desktop uploads, respectively.

Table 2 shows that the loadings obtained with CIPCA are much higher (in absolute value) for BDw (first component) and BUp (second component). Thus, the first SPC can be interpreted as the number of bytes down (BDw) and the second one as the number of bytes up (BUp). The CIPCA scores are shown in Figure 1(b). Snapshot has the highest upstream peak traffic volume, and is now better separated from video and streaming. NMAP is again the application with smaller rectangles. However, it is now better separated from HTTP, since most HTTP datastreams have higher traffic volume range simultaneously in the upstream and downstream directions. Video and streaming are also well separated, since video datastreams have consistently higher traffic volume ranges simultaneously in both directions. Regarding Torrent, it is now possible to distinguish among three groups: the group centers occur at approximately the same upstream traffic volume; one group has small traffic range in both directions (small rectangles) and high downstream volume, another has high traffic ranges in the downstream direction but small in the upstream direction, and a third one has small downstream volumes but high upstream traffic ranges. These groups emerge from differences on the relative location of peers and the quality/stability of links. The first group corresponds to closer peers from which it is possible to download at higher speeds, the third to farther peers for which the links are less stable and unable to download at high speeds, and the third group is a mixture of the two previous ones.



(a) SymCovPCA. (b) CIPCA.

Figure 1: Symbolic scores, estimated by MCAR method.

5 Conclusion

Starting from the definition of symbolic variance and covariance for random interval-valued variables, we have used a common insightful framework to present four symbolic

principal component estimation methods that rely on a symbolic-conventional-symbolic strategy: CPCA, VPCA, CIPCA, and SymCovPCA.

The analysis of a symbolic dataset containing Internet traffic lead to a clear interpretation of the underlying Internet applications (Web browsing, file sharing, streaming, video, port scans, and snapshots). The analysis highlighted the difficulties in separating illicit traffic from regular one, suggesting the need to develop outlier detection methods for symbolic data.

.

Acknowledgements

This work has been supported by Fundação para a Ciência e Tecnologia (FCT), Portugal, through the projects UID/Multi/04621/2013 and PTDC/EEI-TEL/5708/2014.

References

[1] Billard L., Diday E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons, Chichester.

[2] Oliveira M.R., Vilela M., Pacheco A., Valadas R., Salvador P., (20XX). Population Symbolic Covariance Matrices for Interval Data. *In preparation*.

[3] Vilela M., Oliveira M.R., Pacheco A., Valadas R., Salvador P., (20XX). Population Symbolic Principal Component Analysis for Interval Data. *In preparation*.

[4] Pascoal C., Oliveira M.R., Valadas R., Filzmoser P., Salvador P., Pacheco A. (2012). Robust feature selection and robust PCA for Internet traffic anomaly detection. *In INFOCOM, 2012 Proceedings IEEE, Orlando, USA*, pp. 1755-1763.

[5] Pascoal C. (2014). Contributions to Variable Selection and Robust Anomaly Detection in Telecommunications. *PhD Thesis*, Instituto Superior Técnico, Universidade de Lisboa, Portugal.

[6] Vilela M. (2015). Classical and Robust Symbolic Principal Component Analysis for Interval Data. *Master Thesis*, Instituto Superior Técnico, Universidade de Lisboa, Portugal.