Symbolic Principal Components for Interval-Valued Data

L. Billard*, J. Le-Rademacher**

*Department of Statistics, University of Georgia USA
lynne@stat.uga.edu
**Division of Biostatistics, Medical College of Wisconsin USA
jlerade@mcw.edu

Abstract. The centers method (Cazes *et al.*, 1997, Chouakria, 1998) was the first principal component analysis for interval-valued data (where it is implicitly assumed that values within an interval are uniformly distributed across that interval). Many other methods have since been proposed. All fail in various ways to capture fully all the information contained in the data. Here, we set these in context against a new method which calculates the covariance matrix exactly. This new method also includes a new visualization of the projection of the observations onto the principal component space.

1 Introduction

There have been a number of methods proposed in the literature for obtaining principal components for interval-valued data. More recently, Le-Rademacher and Billard (2012) has developed a so-called symbolic covariance principal component analysis for such data, based on the exact calculation of the covariance matrix which matrix is fundamental to any principal component methodology. A brief description of the standard principal component methodology is provided in Section 2.1. After describing the calculation of this exact covariance matrix in Section 2.2, we review in Section 2.3 the various methods that currently exist against the backdrop of that exact covariance matrix. Then, in Section 3, we illustrate the new symbolic covariance principal component analysis method on the familiar oils data (Ichino, 1988). A new visualization of the resulting projections of the observations onto the principal component space based on polytopes is also shown. One consequence of the polytope projections is that the separation of the observations is better than when the traditional maximal covering area rectangles are used. Finally, in Section 4, we provide a symbolic-valued output of the principal components which better explains the output more accurately than that obtained using the maximal covering area rectangles.

2 Background

2.1 Standard Principal Component Analysis

Let $\mathbf{Y} = (Y_1, \ldots, Y_p)$ be random variables taking values in $\mathcal{R}^p$. The basic idea behind a principal component analysis is to transform the p-dimensional observations into a set of s-

dimensional functions, with $s \leq p$. Specifically, these functions, called principal components PC_ν, are

$$PC_\nu = e_{\nu_1} Y_1 + \cdots + e_{\nu_p} Y_p, \;\; \nu = 1, \ldots, p, \tag{1}$$

where λ_ν and $\boldsymbol{e}_\nu = (\boldsymbol{e}_{\nu_1}, \ldots, \boldsymbol{e}_{\nu_p})$ are the ν^{th} eigenvalue and ν^{th} eigenvector, respectively, of the covariance matrix $\boldsymbol{\Sigma}$, with $\sum_j e_{\nu_j}^2 = 1$, and where Var$(PC_\nu) = \lambda_\nu$, and Cov$(PC_\nu, PC_{\nu'})$ $= 0, \;\; \nu \neq \nu'$. Typically, rather than using the covariances directly, the data can be normalized, so that the covariance matrix $\boldsymbol{\Sigma}$ becomes the correlation matrix $\boldsymbol{\Sigma}$ with elements calculated from $\Sigma_{jk} =$ Cov$(Y_j, Y_k)/$[Cov(Y_j, Y_j)Cov$(Y_k, Y_k)]^{1/2}$, $j, k = 1, \ldots, p$.

There are two key steps in this method. The first is the calculation of the $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ whose elements are Cov(Y_j, Y_k) (or equivalently, the correlation matrix with elements Σ_{jk}), $j, k = 1, \ldots, p$. The second is the projection of the observation $\mathbf{Y}$ through (1) onto the s-dimensional principal component space $\mathcal{R}^s$. Typically, $s = 3$ or 4, for most applications. For complete details of this methodology, see any of the numerous texts on the subject, e.g., Anderson (1984), Jolliffe (1986), and Johnson and Wichern (2002).

2.2 Symbolic Covariance Function

Suppose we have a set of random variables $\mathbf{Y} = (Y_1, \ldots, Y_p)$ with realizations $\mathbf{Y}_u$, $u = 1, \ldots, m$, where each Y_{ju} takes values in the interval $[a_{ju}, \; b_{ju}]$, $j = 1, \ldots, p, \; u = 1, \ldots, m$.

Bertrand and Goupil (2000) derived the sample variance of Y_j, S_j^2, as

$$S_j^2 = \frac{1}{3m} \sum_{u=1}^{m} [a_{ju}^2 + a_{ju} b_{ju} + b_{ju}^2] - \frac{1}{4m^2} [\sum_{u=1}^{m} (a_{ju} + b_{ju})]^2. \tag{2}$$

An implicit assumption of this derivation is that values within an interval are uniformly distributed across that interval. A theoretical justification underlying this result is provided in Le-Rademacher and Billard (2011). Except where so stated, in the sequel, this assumption will be assumed.

Later, Billard (2008) showed (2) could be re-written, in terms of sum of squares (SS), as

$$\text{Total SS} = \text{Between SS} + \text{Within SS} \tag{3}$$

where Total SS $= mS_j^2$ from (2) and

$$\text{Between SS} = \sum_{u=1}^{m} [(a_{ju} + b_{ju})/2 - \bar{Y}_j]^2, \tag{4}$$

$$\text{Within SS} = \sum_{u=1}^{m} [(a_{ju} - \bar{Y}_{ju})^2 + (a_{ju} - \bar{Y}_{ju})(b_{ju} - \bar{Y}_{ju}) + (b_{ju} - \bar{Y}_{ju})^2]/3 \tag{5}$$

with

$$\bar{Y}_j = \sum_{u=1}^{m} (a_{ju} + b_{ju})/(2m), \;\; j = 1, \ldots, p, \tag{6}$$

$$\bar{Y}_{ju} = (a_{ju} + b_{ju})/2, \; j = 1, \ldots, p, \; u = 1, \ldots, m. \tag{7}$$

Note that the Within SS can also be written as

$$\text{Within SS} = \sum_{u=1}^{m}(b_{ju} - a_{ju})^2/12. \tag{8}$$

Likewise, when considering two variables, Y_j and Y_k, the sum of products (SP) can be shown to satisfy

$$\text{Total SP} = \text{Between SP} + \text{Within SP} \tag{9}$$

where, for $j, k = 1, \ldots, p$,

$$\text{Between SP} = \sum_{u=1}^{m}[(a_{ju} + b_{ju})/2 - \bar{Y}_j][(a_{ku} + b_{ku})/2 - \bar{Y}_k], \tag{10}$$

$$\text{Within SP} = \sum_{u=1}^{m}(b_{ju} - a_{ju})(b_{ku} - a_{ku})/12. \tag{11}$$

Hence, the sample covariance $\text{Cov}(Y_j, Y_k) = S_{jk}^2 = \text{Total SP}/m$ is given by

$$\begin{aligned}\text{Cov}(Y_j, Y_k) = \frac{1}{6m}\sum_{u=1}^{m}[&2(a_{ju} - \bar{Y}_j)(a_{ku} - \bar{Y}_k) + (a_{ju} - \bar{Y}_j)(b_{ku} - \bar{Y}_k) \\ &+ (b_{ju} - \bar{Y}_j)(a_{ku} - \bar{Y}_k) + 2(b_{ju} - \bar{Y}_j)(b_{ku} - \bar{Y}_k)].\end{aligned} \tag{12}$$

When $j = k$, the sample covariance function in (12) simplifies to the sample variance function in (2). Also, in the special case that the data are classical points in $\mathcal{R}^p$, we can write the observation $Y = a$ as $Y = [a,\ a]$. It is easily verified that the formulas for S_j^2 and S_{jk}^2 in (2) and (12), respectively, reduce to their classical counterparts.

Clearly, the sample covariance functions calculated from (12) give the exact covariance values for interval-valued data. As such, when applying the standard theory of Section 2.1, exact eigenvalues and exact eigenvectors emerge. Hence, exact principal components from (1) are obtained for each of the hypercubes governing each particular observation $\mathcal{H}_u$, $u = 1, \ldots, m$, in $\mathcal{R}^p$. This gives so-called symbolic covariance principal components. See Le-Rademacher and Billard (2012).

2.3 Literature Review

The first method introduced for interval-valued data was the "centers" method, by Cazes *et al.* (1997) and Chouakria (1998). Here, the interval-values $[a_{ju},\ b_{ju}]$ were replaced by the interval centers or midpoints $X_{ju}^c = (a_{ju} + b_{ju})/2$, for each $j = 1, \ldots, p$ and $u = 1, \ldots, m$. The covariance matrix was calculated from these X^c values and hence the principal components from (1). By referring to (3) and (9), it is clear that the covariances are based on the Between variations only and that the Within variations are not used. That is, there is a loss of information.

Cazes *et al.* (1997) and Chouakria (1998) also introduced a "vertices" method. In this case, the interval-values were replaced by two classical values, viz., the two interval end-points, $Y_{ju}^{(1)} = a_{ju}$ and $Y_{ju}^{(2)} = b_{ju}$. The elements of the covariance matrix now corresponded

to the (Between SS/SP + Error SS/SP) where their Between SS/SP are as given in (4) and (10), respectively. However, their Error SS/SP $\neq$ Within SS/SP. Thus, while more of the variation in the data is included compared with the centers method, some information is still lost. Douzal-Chouakria *et al.* (2011) developed some refinements to the vertices method and compared the results with several classical surrogates; but they did not address the loss of information issue.

Later, a symbolic-object method, developed by Lauro and Palumbo (2000), uses the vertices $(Y_{ju}^{(1)}, Y_{ju}^{(2)})$ of the vertices method to calculate their covariance matrix based on the centers Y_{ju}^{c}. As for each of the vertices and centers methods, this symbolic object method also does not use all the variation information contained in the data. Lauro and Palumbo (2000) also developed a range-tranformation approach, and then combined this with their symbolic-object method to give a "mixed" strategy.

Palumbo and Lauro (2003) converts the intervals $[a_{ju},\ b_{ju}]$ into two classical values, the midpoint Y_{ju}^{c} of the centers method and the range $Y_{ju}^{r} = (b_{ju} - a_{ju})$ (or equivalently, $Y_{ju}^{r}/2$), $j = 1, \ldots, p,\ u = 1, \ldots, m$, to give a midpoints-range method. It is easy to show that Range SS/SP $\neq$ Within SS/SP. Two covariance matrices are calculated, one based on the midpoints and one on the ranges.

A number of other approaches has been introduced, e.g., Gioia and Lauro (2006) and Lauro *et al.* (2008) have tried interval arithmetic ideas. However, this approach only works when the intervals are short.

Unfortunately, all these methods in the literature fail in some way to use all the variations inherent in the data; therefore, there is a loss of information. There are also some further considerations.

First, there are implicit independence assumptions between the endpoints $Y_{ju}^{(1)}$ and $Y_{ju}^{(2)}$ (in the vertices and related methods), and between the midpoints Y_{ju}^{c} and the ranges Y_{ju}^{r} (in the range and related methods). Clearly, these independence assumptions are not sustainable.

Furthermore, suppose the data fit the special case that all intervals have the same midpoints; e.g., $[9, 11], [1, 19], [2, 18], \ldots$. In this case, the eigenvalues of the corresponding midpoint covariance matrix are zero. Hence, any of the methods which use the centers Y_{ju}^{c} will not work, i.e., the centers, symbolic-object, mixed, and range-transformation methods fail. Likewise, when the data have intervals with common ranges, e.g., $[0, 10], [20, 30], [120, 130], \ldots$, the eigenvalues of the range covariance matrix are zero. Hence, methods which use the range values will not work, e.g., the center-range and range-transformation methods fail. Notice however that intervals with common midpoints, or intervals with common ranges, are still differing observations with inherent variations; so any viable method has to be able to accommodate these special cases. The vertices method and the symbolic-covariance method work on these special cases, as well as for classical data.

3 Illustration - Oils Data

3.1 Symbolic Principal Components

The symbolic covariance method is illustrated on the oils data of Ichino (1988). There are $p = 4$ random variables, Y_1 = specific gravity, Y_2 = freezing point, Y_3 = iodine value, and Y_4 =

Variable	Eigenvector ν_k			
Y_j	ν_1	ν_2	ν_3	ν_4
Y_1	0.5395	0.4173	-0.0256	0.7308
Y_2	-0.5280	-0.3203	0.5192	0.5908
Y_3	0.5105	-0.0902	0.8018	-0.2973
Y_4	-0.4117	0.8457	0.2948	-0.1686

TAB. 1 – *Eigenvectors for oils data*

saponification. Measurements were taken on each of $m = 8$ oils, specifically, linseed, perilla, cotton, sesame, camellia, olive, beef, and hog.

The symbolic correlation matrix $\mathbf{\Sigma}$, calculated from (2) and (12), for these data is

$$\mathbf{\Sigma} = \begin{bmatrix} 1.0000 & -0.9126 & 0.7714 & -0.4338 \\ -0.9126 & 1.0000 & -0.6431 & 0.5126 \\ 0.7714 & -0.6431 & 1.0000 & -0.5874 \\ -0.4338 & 0.5126 & -0.5874 & 1.0000 \end{bmatrix}. \tag{13}$$

The eigenvalues of $\mathbf{\Sigma}$ are

$$\lambda_1 = 0.7385, \;\; \lambda_2 = 0.1636, \;\; \lambda_3 = 0.0857, \;\; \lambda_4 = 0.0121;$$

and the eigenvectors corresponding to the ν^{th} principal component are as shown in Table 1. Hence, by application of (1), the principal component space for each observation can be obtained.

3.2 Visualization

In a standard analysis, each observation $\mathbf{Y}_u$ in $\mathcal{R}^p$ produces a corresponding principal component $PC_{\nu u}$ from (1). If the observations are classical points, then the projection of the point observation $\mathbf{Y}_u$ is a point in the principal component space. However, when the observation is a hypercube in $\mathcal{R}^p$, then the projection of that observation is a polytope on the principal component space. This is seen in Fig.1, where the shaded area refers to the projection of one such observation H onto the $PC_1 \times PC_2$ space. Previous methods, including those reviewed in Section 2.3, have adopted the maximal covering area rectangle (MCAR) approach. This rectangle is obtained according to $PC_{\nu u} = [PC^a_{\nu u}, PC^b_{\nu u}]$ with, for each $u = 1, \ldots, m$,

$$PC^a_{\nu u} = \min_{\mathbf{Y}_u \in \mathcal{H}_u} (PC_{\nu u}), \;\; PC^b_{\nu u} = \max_{\mathbf{Y}_u \in \mathcal{H}_u} (PC_{\nu u})$$

where $PC_{\nu u}$ is calculated from (1) for each $\mathbf{Y}_u \in \mathcal{H}_u$. Thus, the MCAR is the rectangle covering the hypercube of Fig. 1.

Notice however that the MCAR principal component for an observation includes regions (the unshaded parts of Fig. 1) that are projections of points not in the observed hypercube $\mathbf{Y}$. A first step in removing these unshaded parts was given by Irpino *et al.* (2003) in their parallel edges connected shape (PECS) projection. Later, Le-Rademacher and Billard (2012) provide

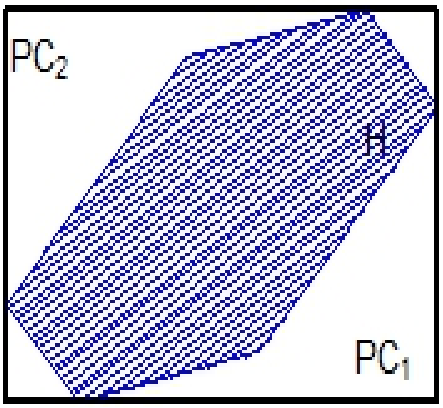

FIG. 1 – *Maximum covering area rectangle (MCAR).*

an algorithm based on polytope theory that allows for a projection into the principal component space that matches the shaded region (of Fig.1). A comparison of these three projections is shown in Fig. 2 (from Le-Rademacher and Billard, 2012, Figure 2). A consequence of the MCAR representation is that observations may appear to overlap when they do not. The polytope representation minimizes any overlapping to real/valid overlappping, and so produces less cluttered plots.

The projections of the interval-valued data, or hypercubes, for the oils data onto the $PC_1 \times PC_2$ space, obtained by the polytope approach, are shown in Fig. 3. From this Fig. 3, it is clear that the linseed and perilla oils form one group, beef and hog oils form another group, and cotton, sesame, camellia and olive oils form a third group. Visualization of the variables will be covered elsewhere.

3.3 Symbolic Representation of Output Principal Components

Let us return to the projection of an observation onto the principal component space as illustrated by the shaded region of Fig.1. Consider the 1^{st} principal component, PC_1, axis. As specific values for PC_1 change along this axis, the portion of the shaded region at those values also changes. That is, the distribution of the output principal component, here PC_1, is not uniform (as is implicitly presumed when using the MCARs which also includes the unshaded regions in the output principal component). Tab. 2 gives the output histogram for PC_1 for the oils data when PC_1 is divided into 7 histogram sub-intervals. The details of this calculation can be found in Le-Rademacher (2008).

4 Conclusion

For more complete details of this methodology as it applies to interval-valued data, including more extensive illustrative comparisons with previous methods, see Le-Rademacher and

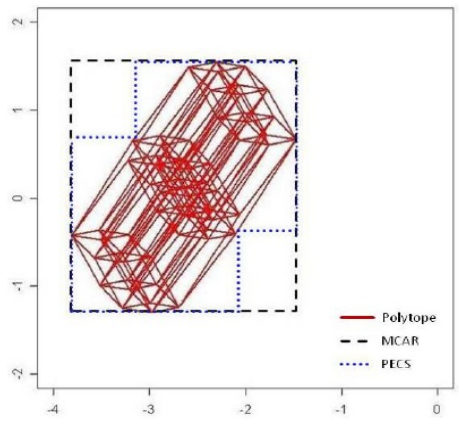

FIG. 2 – *Comparison of MCAR, PECS and polytope projection.*

Billard (2012). Extensions to histogram-valued data are developed in Le-Rademacher (2008).

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd. ed. New York: John Wiley and Sons.

Billard, L. (2008). Sample covariance functions for complex quantitative data. In: *Proceedings World Congress, International Association of Statistical Computing* (eds. M. Mizuta and J. Nakano) Japanese Society of Computational Statistics, Japan, p. 157-163.

Bertrand, P. and F. Goupil (2000). Descriptive statistics for symbolic data. In: *Analysis of Symbolic Data: Exploratory methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Berlin: Springer, 103-124.

Cazes, P., A. Chouakria, E. Diday and Y. Schechtman (1997). Extension de l'analyse en composantes principales à des données de type intervalle. *Revue Statistique Appliqueé* 45, 5-24.

Chouakria, A (1998). *Extension des Méthodes d'Analyse Factorielle à des Données de Type Intervalle*. Thése de doctorat., Université Paris Dauphine, Paris.

Douzal-Chouakria, A., L. Billard and E. Diday (2011). Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining* 4, 229-246.

Ichino, M (1988). General metrics for mixed features - The Cartesian space theory for pattern recognition. In: *Proceedings of the 1988 Conference on Systems, Man, and Cybernetics*. Oxford: Pergamon, 494-497.

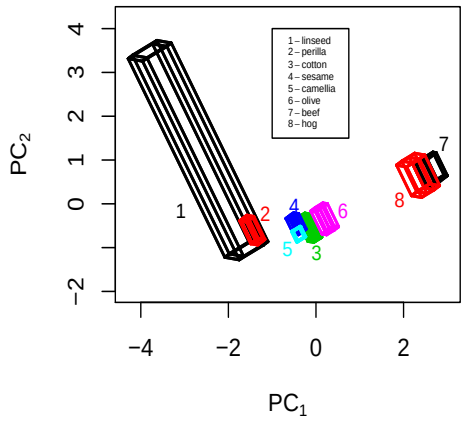

FIG. 3 – *Principal component projections for oils data, based on polytopes.*

Irpino, A., C. Lauro and R. Verde (2003). Visualizing symbolic data by closed shapes. In: *Between Data Science and Applied Data Analysis* (eds. M. Schader, W. Gaul and M. Vichi). Berlin: Springer, 244-251.

Johnson, R. A. and D. W. Wichern (2002). *Applied Multivariate Statistical Analysis* (5th ed.). New York: Prentice Hall.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

Lauro, N. C. and F. Palumbo (2000). Principal component analysis of interval data: A symbolic data analysis approach. *Computational Statistics* 15, 73-87.

Lauro, N. C., R. Verde and A. Irpino (2008). Principal component analysis of symbolic data described by intervals. In: *Symbolic Data Analysis and the SODAS Software* (eds. E. Diday and M. Noirhomme-Fraiture). Chichester: John Wiley and Sons, 279-311.

Le-Rademacher, J. and L. Billard (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference* 141, 1593-1602.

Le-Rademacher, J. and L. Billard (2012). Symbolic-covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics*, in press.

Le-Rademacher, J (2008). *Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data*. Doctoral Dissertation, University of Georgia.

Palumbo, F. and N. C. Lauro (2003). A PCA for interval-valued data based on midpoints and radii. In: *New Developments in Psychometrics* (eds. H. Yanai, A. Okada, K. Shigemasu, Y. Kano and J. Meulman). Tokyo: Springer, 641-648.

Oil	Histogram Output for PC_1: $\{[a_k,\ b_k], p_k; k = 1 \ldots, 7\}$			
linseed	{[-4.290, -4.190], 0.003;	[-4.190,-3.634], 0.105;	[-3.634, -3.304], 0.126;	[-3.304, -2.082], 0.534;
	[-2.082,-1.752], 0.126;	[-1.752, -1.196], 0.105;	[-1.196, -1.096], 0.003}	
perilla	{[-1.744 -1.603], 0.127;	[-1.603, -1.574], 0.058;	[-1.574, -1.489], 0.214;	[-1.489, -1.419], 0.203;
	[-1.419, -1.334], 0.214;	[-1.334, -1.305], 0.058;	[-1.305, -1.164], 0.127}	
cotton	{[-0.441, -0.401], 0.011;	[-0.401, -0.255], 0.210;	[-0.255, -0.187], 0.180;	[-0.187, -0.119], 0.199;
	[-0.119,-0.051], 0.180;	[-0.051, 0.096], 0.210;	[0.096,0.136], 0.011}	
sesame	{[-0.688, -0.567], 0.145;	[-0.567, -0.518], 0.142;	[-0.518, -0.509], 0.031;	[-0.509, -0.401], 0.364;
	[-0.401, -0.392], 0.031;	[-0.392, -0.343], 0.142;	[-0.343, -0.222], 0.145}	
camellia	{[-0.559, -0.539], 0.008;	[-0.539, -0.446], 0.251;	[-0.446, -0.427], 0.090;	[-0.427, -0.364], 0.300;
	[-0.364, -0.344], 0.090;	[-0.344, -0.251], 0.251;	[-0.251, -0.231], 0.008}	
olive	{[-0.119, -0.019], 0.053;	[-0.019, 0.136], 0.280;	[0.136, 0.157], 0.055;	[0.157, 0.242], 0.225;
	[0.242, 0.263], 0.055;	[0.263, 0.418], 0.280;	[0.418, 0.518], 0.053}	
beef	{[2.235, 2.435], 0.151;	[2.435, 2.489], 0.092;	[2.489, 2.567], 0.153;	[2.567, 2.670], 0.210;
	[2.670, 2.747], 0.153;	[2.747, 2.801], 0.092;	[2.801, 3.002], 0.151}	
hog	{[1.840, 1.960], 0.033;	[1.960, 2.179], 0.225;	[2.179, 2.253], 0.114;	[2.253, 2.412], 0.255;
	[2.412, 2.486], 0.114;	[2.486, 2.705], 0.225;	[2.705, 2.825], 0.033}	

TAB. 2 – *Output histogram for PC_1 for oils data*