Adv Data Anal Classif (2017) 11:327–351
DOI 10.1007/s11634-016-0255-9

REGULAR ARTICLE

Principal component analysis for histogram-valued data

J. Le-Rademacher[1] · L. Billard[2]

Received: 14 August 2013 / Revised: 30 April 2016 / Accepted: 11 May 2016 /
Published online: 26 May 2016

Abstract This paper introduces a principal component methodology for analysing histogram-valued data under the symbolic data domain. Currently, no comparable method exists for this type of data. The proposed method uses a symbolic covariance matrix to determine the principal component space. The resulting observations on principal component space are presented as polytopes for visualization. Numerical representation of the resulting polytopes via histogram-valued output is also presented. The necessary algorithms are included. The technique is illustrated on a weather data set.

Keywords Principal components · Histogram observations · Polytopes

Mathematics Subject Classification 62H25 · 60-08

1 Introduction

In an age of "big data", we are faced with challenges to develop statistical methodology that can analyse the data accurately. One approach is to aggregate the data in some meaningful way even if only to reduce the size of the data set. There are a myriad of

Electronic supplementary material The online version of this article (doi:10.1007/s11634-016-0255-9) contains supplementary material, which is available to authorized users.

✉ L. Billard
lynne@stat.uga.edu

J. Le-Rademacher
Le-Rademacher.Jennifer@mayo.edu

[1] Mayo Clinic, Rochester, MN 55905, USA

[2] University of Georgia, Athens, GA 30602, USA

ways to effect this aggregation. Clearly, it is preferable to carry out any such aggregation according to scientific questions driving the analysis. The resulting aggregated data will perforce be in the form of so-called symbolic data, such as lists, intervals, histograms, and the like. While the aggregation of data sets (small or large) will produce symbolic data, symbolic data can also arise naturally (e.g., species data, many medical entities such as blood pressure, etc.). A detailed description of symbolic data along with numerous examples can be found in Bock and Diday (2000) and Billard and Diday (2003, 2006); Billard (2011) provides a non-technical illustrated introduction to symbolic data. It is important to note the distinction between symbolic data and fuzzy data. In symbolic data, an observation contains multiple values either by natural occurrence or by data aggregation. The values within a symbolic observation follow a probability distribution. By contrast, the values in a fuzzy set are caused by the imprecision or the fuzziness of the definition of the event. The values in a fuzzy set have associated grades of membership. See Zadeh (1965, 1968) and Shapiro (2009) for further information about fuzzy sets. The focus of this paper is on symbolic data and not on fuzzy data.

Often-times, in the absence of knowing what else to do, analysts have taken the average of the aggregated values, or some other seemingly suitably selected "representative" value as a classical surrogate, for a given category. However, it is known that this approach, while giving answers, give answers that are not necessarily correct, since some of the variations present in the data are ignored. For example, Billard (2008) has shown that for a data set of interval or histogram-valued observations, the total variation equals the sum of the within observation variation and the between observation variation. The between observation variation is a measure of the variation obtained when using the average as a classical surrogate to represent the aggregated values of a category (though this can change slightly depending on varying underlying assumptions, but the sense remains). The within observation variation is a measure of the internal variation of each observation. It is this within observation variation that is ignored when using classical surrogates only. To illustrate this further, suppose we have two samples each of size $n = 1$; the first sample contains the observation $X_{(1)} = [9, 11]$ and the second is $X_{(2)} = [0, 20]$. Both have the same average value (=10). Clearly, using the average as the basis for the analysis will give the same answer for both samples, yet also clearly, the two samples have differing values and so should produce differing answers. That is, symbolic data require symbolic methods which use all the information contained in the data.

Our focus is on principal component analysis for histogram-valued observations. The basic principles of principal component analyses are unchanged from those for classically-valued observations. See any of the many texts, e.g., Jolliffe (2004) and Johnson and Wichern (2002) for an applied approach, and Anderson (1984) for a theoretical approach. Recently, Le-Rademacher and Billard (2012) developed methodology for interval data; this included obtaining representations of the projections of the observed hyperrectangles in $\mathcal{R}^p$ as polytopes in principal component space. These contrast sharply from the point projections in principal component space obtained for classical point data. Le-Rademacher and Billard (2012) also provided an illustrative comparison of their method with previous attempts to analyse interval data, including, e.g., Cazes et al. (1997), Chouakria (1998), Lauro and Palumbo (2000), Irpino et al.

(2003), Palumbo and Lauro (2003), Lauro et al. (2008), and Douzal-Chouakria et al. (2011). A brief non-illustrative description of these differences is given in Billard and Le-Rademacher (2013). This comparison showed that these earlier efforts (while advancing the science) had failed in different ways to meet the difficult challenge of capturing all the variations inherent to the data. To extend the projected polytopes beyond visualization, Le-Rademacher and Billard (2013) introduced an algorithm to translate the resulting polytopes into histogram-valued output that can be used as numerical input in further statistical analyses.

To date, no comparable methodology exists for handling histogram-valued data. Note that, the histogram-valued data here are different from the modal categorical data of Cazes (2002), Ichino (2011), and Makosso-Kallyth and Diday (2012). In this work, we extend the interval PCA methodology of Le-Rademacher and Billard (2012, 2013) to obtain principal components for histogram-valued observations. This method uses all the variations contained in the data set by using the symbolic covariance matrix; see Sect. 2. Also, in Sect. 2, a visualization of the resulting principal components and computation of the histogram-valued output are suggested. Since hyperrectangles for histogram data can be viewed as consisting of sets of weighted sub-hyperrectangles, expansion of the polytope methods developed for interval data is proposed. However, the generalization of the methods from interval data to histogram data is not trivial. Challenges encountered in the proposed expansion are addressed in Sect. 2. The algorithms for constructing the polytopes and the histogram-valued output are summarized in Sect. 3 (detailed algorithms are given in the Appendix). In Sect. 4, the new methodology is illustrated on a real data set describing monthly temperatures at weather stations in China. Conclusions are given in Sect. 5.

2 Basics and methodology

2.1 Histogram-valued data: basic statistics

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a p-dimensional random variable taking values in $\mathcal{R}^p$. When a realization is a histogram-valued (or, simply, histogram) observation, it takes the form, for each X_j, $j = 1, \ldots, p$,

$$X_{ij} = \{[a_{ij1}, b_{ij1}), p_{ij1}; \ldots; [a_{ijs_{ij}}, b_{ijs_{ij}}], p_{ijs_{ij}}\}, \quad i = 1, \ldots, n, \tag{1}$$

with $\sum_{k_j=1}^{s_{ij}} p_{ijk_j} = 1$. The disjoint histogram subintervals $[a_{ijk_j}, b_{ijk_j})$ in Eq. (1), with $a_{ijk_j} \le b_{ijk_j}$, can be open or closed at either end, and occur with probability or relative frequency p_{ijk_j}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, and $k_j = 1, \ldots, s_{ij}$. Typically, the number of subintervals s_{ij} varies across different observations i and variables j.

The sample mean and sample variance for each X_j, $j = 1, \ldots, p$, were obtained by Billard (2008), respectively, as

$$\bar{X}_j = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k_j=1}^{s_{ij}} (a_{ijk_j} + b_{ijk_j}) p_{ijk_j} \tag{2}$$

and

$$S_j^2 = \frac{1}{3n}\sum_{i=1}^{n}\sum_{k_j=1}^{s_{ij}}(a_{ijk_j}^2 + a_{ijk_j}b_{ijk_j} + b_{ijk_j}^2)p_{ijk_j} - \frac{1}{4n^2}\left[\sum_{i=1}^{n}\sum_{k_j=1}^{s_{ij}}(a_{ijk_j} + b_{ijk_j})p_{ijk_j}\right]^2. \tag{3}$$

The sample symbolic covariance (or, simply, sample covariance) between two variables X_{j_1} and X_{j_2} was obtained in Billard (2008) as

$$\begin{aligned} Cov(X_{j_1}, X_{j_2}) = \frac{1}{6n}\sum_{i=1}^{n}\sum_{k_1=1}^{s_{ij_1}}\sum_{k_2=1}^{s_{ij_2}} \{&[2(a_{ij_1k_1} - \bar{X}_{j_1})(a_{ij_2k_2} - \bar{X}_{j_2}) \\ &+ (a_{ij_1k_1} - \bar{X}_{j_1})(b_{ij_2k_2} - \bar{X}_{j_2}) \\ &+ (b_{ij_1k_1} - \bar{X}_{j_1})(a_{ij_2k_2} - \bar{X}_{j_2}) \\ &+ 2(b_{ij_1k_1} - \bar{X}_{j_1})(b_{ij_2k_2} - \bar{X}_{j_2})]p_{ij_1k_1j_2k_2}\} \end{aligned} \tag{4}$$

where $p_{ij_1k_1j_2k_2}$ is the relative frequency of the rectangle formed by subinterval $[a_{ij_1k_1}, b_{ij_1k_1})$ of X_{j_1} and subinterval $[a_{ij_2k_2}, b_{ij_2k_2})$ of X_{j_2}. For simplicity of notation in Eq. (4), $k_1 = k_{j_1}$ and $k_2 = k_{j_2}$. When $j_1 = j_2 = j$, then $Cov(X_{j_1}, X_{j_2}) = S_j^2$. When $s_{ij} = 1$ and hence $p_{ij1} = 1$ for all $i = 1, \ldots, n$, $j = 1, \ldots, p$, the histogram value of Eq. (1) reduces to interval data as a special case. In this case, the sample statistics $\bar{X}$, S_j^2 and $Cov(X_{j_1}, X_{j_2})$ of Eqs. (2)–(4) reduce to their respective formula for interval-valued observations (obtained by Bertrand and Goupil (2000), Billard (2008); likewise, when the data are classically valued, where now $s_{ij} = 1$, and $b_{ij1} = a_{ij1}$ for all i, j, these statistics reduce to their well known classical counterparts, as a second special case. An underlying assumption in these formulae is that within a given subinterval, values for X_{ij} are uniformly distributed across those subintervals (even though the random variable X_j itself is not necessarily uniformly distributed; indeed, the $\mathbf{X}$ is often assumed to follow a multivariate normal distribution, at least asymptotically).

Note that, although X_j is symbolic-valued (a histogram in this case), the mean and the (co)variance of X_j are classically valued. Using conditional expectation and an internal parameters approach, Le-Rademacher and Billard (2011) showed that the variance of a symbolic-valued random variable is a sum of the mean and the variance of classically valued internal parameters. Similarly, Xu (2010) showed that the covariance of interval-valued random variables is a sum of the mean and the covariance of classically valued internal parameters. These (co)variances are similar to the (co)variances of classically-valued random variables with mixture distribution. Hence, the properties of the covariance matrix for classical data carry through to symbolic data, including the fact that the values of the correlation are between (-1) and $(+1)$. Furthermore, the sample mean $\bar{X}_j$ of Eq. (2) and the sample variance of Eq. (3) are maximum likelihood estimators (MLE) of the mean and the variance of histogram-valued random variable X_j Le-Rademacher and Billard (2011). The covariance $Cov(X_{j_1}, X_{j_2})$ of Eq. (4) is a histogram data generalization of the MLE of the covariance of interval data from Xu (2010) and Billard et al. (2011). Since the symbolic sample covariance matrix, with elements shown in Eq. (4), accounts for the total variation of histogram data, we propose constructing the principal components using this covariance matrix as described in the following sections.

2.2 Principal components

When the dimension p is large and some of the variables are correlated, principal component analysis is often used to reduce the dimension and the collinearity in the data by creating uncorrelated linear combinations of the p variables, called principal components. In particular, the νth principal component, PC_ν, is the linear transformation of an observation $\mathbf{X}$ satisfying

$$PC_\nu = e_{\nu_1} X_1 + \cdots + e_{\nu_p} X_p, \quad \nu = 1, \ldots, p, \tag{5}$$

where λ_ν and $\mathbf{e}_\nu = (e_{\nu_1}, \ldots, e_{\nu_p})$ are the νth eigenvalue and νth eigenvector, respectively, of the covariance matrix Σ with $\lambda_1 \geq \cdots \geq \lambda_p$, $\sum_j e_{\nu_j}^2 = 1$, and where $Var(PC_\nu) = \lambda_\nu$, and $Cov(PC_\nu, PC_{\nu'}) = 0$, $\nu \neq \nu'$. That is, the eigenvalue λ_ν represents the amount of variation explained by PC_ν. In practice, the first two or three principal components account for most of the total variation in the data. Hence, subsequent analyses using the first few principal components account for most of the data variability.

Instead of using the covariances directly, the data can be normalized; in this case, the covariance matrix Σ is replaced by the correlation matrix Σ with elements

$$\Sigma_{j_1 j_2} = Cov(X_{j_1}, X_{j_2})/[Cov(X_{j_1}, X_{j_1})Cov(X_{j_2}, X_{j_2})]^{1/2}, \quad j_1, j_2 = 1, \ldots, p. \tag{6}$$

Anderson (1963, 1984), Mardia (1979), Johnson and Wichern (2002), and Jolliffe (2004) provide comprehensive treatments of classical PCA.

Once the covariance matrix, or equivalently the correlation matrix, is obtained, the derivation of the principal components through Eq. (5) is analogous to that for classical data and for interval data. The difference across these data formats is in the relevant formula to calculate the covariance/correlation matrix Σ. For histogram-valued data, the elements of Σ are given in Eq. (4). Since the covariance matrix Σ is classically valued, the eigenvalues and the eigenvectors of Σ have the same properties and the same interpretation as those of classical data. Specifically, the eigenvectors are orthogonal; hence the principal components are orthogonal linear transformations of the original variables. Moreover, the eigenvector $\mathbf{e}_\nu$ represents the direction with the νth largest variation in the observed data. See Jolliffe (2004) for a discussion of mathematical and statistical properties of PCs. Note that, in this work, we propose using the symbolic sample covariance matrix to calculate the principal components. Therefore, the principal component axes align in directions of maximum total variation for histogram data.

Analogous to the classical case, the principal components of histogram-valued variables are linear combinations of those variables. Whereas for classical data, these linear combinations are linear transformations of single data points, for histogram data, they are transformations of all the data points in the histograms. In other words, since the histogram observations are hyperrectangles which are convex sets, the linear combinations of histogram observations are linear transformations of these convex sets. Section 2.3 describes the construction of the transformed convex sets (polytopes) for visualization.

2.3 Visualization

When the data are classical points in $\mathcal{R}^p$, the projections are points onto the principal component space; see any text on multivariate statistics, e.g., Jolliffe (2004). When the data are intervals so forming hyperrectangles in $\mathcal{R}^p$, the projections onto principal component space are polytopes; see Le-Rademacher and Billard (2012). They further showed that these polytopes are convex sets with boundaries defined by the transformed vertices of the hyperrectangles. The polytopes can be reconstructed in the principal component space by connecting the transformed vertices of the observed hyperrectangles. The vertices are used as the first step in the reconstruction of the observations because they define the boundaries of the hyperrectangles. After the vertices are transformed, the entire polytope representing a transformed interval-valued observation can be reconstructed in the principal component space. This allows reconstruction of the polytopes without having to transform each individual data point within the observations, which can be infinitely many. The vertices are part of the hyperrectangle representing an observation. They are not by themselves the entire observation. This same idea is extended to reconstructing histogram-valued observations in the principal component space as described in the following.

To construct polytopes for histogram data, we first recognize that each histogram is in effect a weighted set of sub-hyperrectangles in $\mathcal{R}^p$. Within each sub-hyperrectangle, the density is assumed to be uniform, but that these densities vary across the sub-hyperrectangles. For example, when the number of subintervals $s_{ij} = 2$ and $p = 2$, we have the hyperrectangle as shown in Fig. 1a with the differing weights of the sub-rectangles reflected by their differing colors. This is contrasted with the single uniformly weighted rectangle for interval data in Fig. 1b. In general, there will be r_i sub-hyperrectangles for the histogram $\mathbf{X}_i$ where

$$r_i = \prod_{j=1}^{p} s_{ij}. \tag{7}$$

Since a histogram-valued observation can be represented by a hyperrectangle in the sample space $\mathcal{R}^p$, an observation $\mathbf{X}_i$ can be expressed in terms of the vertices of its sub-hyperrectangles. Denote the set of subinterval endpoints by $\mathbf{Q}_i = (Q_{i1}, \ldots, Q_{ip})$ with $Q_{ij} = \{a_{ij1}, b_{ij1}, \ldots, a_{ijs_{ij}}, b_{ijs_{ij}}\}$, $j = 1, \ldots, p$. Let $\mathbf{W}_i^v$ be the $(2^p r_i \times p)$ matrix whose rows consist of the vertex coordinates of all possible permutations of the elements of $\mathbf{Q}_i$. From Eq. (7), it follows that $\mathbf{W}_i^v$ has $2^p r_i = 2^p(\prod_{j=1}^{p} s_{ij})$ rows.

However, usually the upper endpoint of a subinterval equals the lower endpoint of the next subinterval. Therefore, without loss of generality, for each X_{ij}, assume $b_{ij(k_j-1)} = a_{ijk_j}$ for $k_j = 2, \ldots, s_{ij}$ and write $b_{ijs_{ij}} = a_{ij(s_{ij}+1)}$. Then, the set of endpoints $\mathbf{Q}_i$ can be replaced by $\mathbf{E}_i = (E_{i1}, \ldots, E_{ip})$ with $E_{ij} = \{a_{ij1}, \ldots, a_{ij(s_{ij}+1)}\}$,

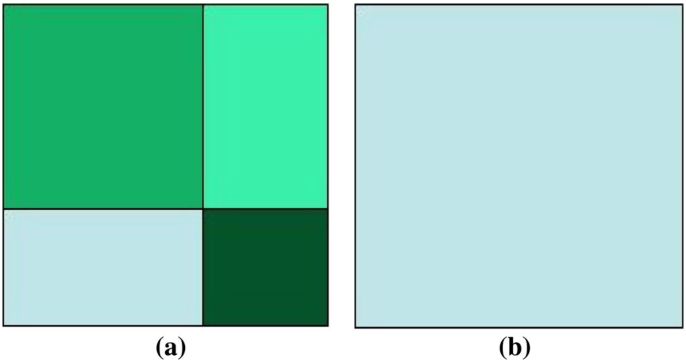

Fig. 1 **a** Histogram sub-rectangles. **b** Interval rectangle

$j = 1, \dots, p$, $i = 1, \dots, n$. Let $\mathbf{X}_i^v$ be the matrix whose rows include all possible permutations of the elements of $\mathbf{E}_i$. Then, $\mathbf{X}_i^v$ has $N_i = \prod_{j=1}^{p} (s_{ij} + 1)$ rows. The quantity $\prod_{j=1}^{p} (s_{ij} + 1)$ is less than or equal to $2^p (\prod_{j=1}^{p} s_{ij})$. The magnitude of the difference between these two quantities depends on the number of subintervals for each histogram X_{ij} and the number of variables, p. When $s_{ij} = 1$ for all $j = 1, \dots, p$, as for interval data,

$$\frac{\prod_{j=1}^{p} (s_{ij} + 1)}{2^p (\prod_{j=1}^{p} s_{ij})} = \frac{\prod_{j=1}^{p} 2}{2^p (\prod_{j=1}^{p} 1)} = \frac{2^p}{2^p} = 1.$$

When s_{ij} is large for most X_{ij}, it is easily shown that

$$\frac{\prod_{j=1}^{p} (s_{ij} + 1)}{2^p (\prod_{j=1}^{p} s_{ij})} \to \frac{1}{2^p}.$$

Hence, when p is large, it is more efficient to use the matrix of vertices $\mathbf{X}_i^v$ which is given by

$$\mathbf{X}_i^v = \begin{bmatrix} a_{i11} & a_{i21} & \dots & a_{ip1} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i11} & a_{i21} & \dots & a_{ip(s_{ip}+1)} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1(s_{i1}+1)} & a_{i2(s_{i2}+1)} & \dots & a_{ip1} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1(s_{i1}+1)} & a_{i2(s_{i2}+1)} & \dots & a_{ip(s_{ip}+1)} \end{bmatrix}. \tag{8}$$

Figure 2a shows these vertices when $s_{ij} = 2$ and $p = 2$ (see Fig. 1a) using the endpoints of $\mathbf{E}_i$; and Fig. 2b gives the vertices for the interval data with $s_{ij} = 1$ and $p = 2$ (see Fig. 1b) using the endpoints of $\mathbf{Q}_i$. The matrix of endpoints for Fig. 2a has $N_i = \prod_{j=1}^{2} (2 + 1) = 9$ rows and is given by, from Eq. (8),

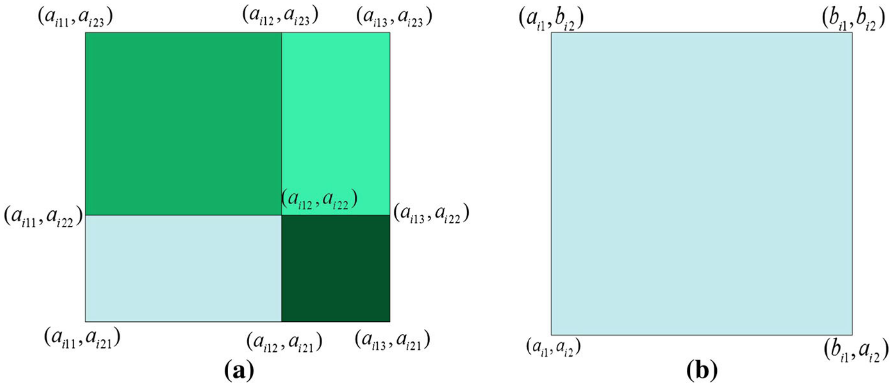

Fig. 2 **a** Vertices **E** for Histogram sub-rectangles. **b** Vertices **Q** for Interval rectangle

$$\mathbf{X}_i^v = \begin{bmatrix} a_{i11} & a_{i11} & a_{i11} & a_{i12} & a_{i12} & a_{i12} & a_{i13} & a_{i13} & a_{i13} \\ a_{i21} & a_{i22} & a_{i23} & a_{i21} & a_{i22} & a_{i23} & a_{i21} & a_{i22} & a_{i23} \end{bmatrix}^T \tag{9}$$

where T stands for the transpose function. Construction of $\mathbf{X}_i^v$, and hence of $\mathbf{X}_i^h$, is given in “Constructing the matrix of vertices” in the Appendix.

Equation (5) can now be applied to the set of vertices in $\mathbf{X}_i^v$ transforming them into points in the principal component space. Let $\mathbf{Y}_i^v$ be the matrix of transformed vertices of observation i. Then, from Eq. (5), we have

$$\mathbf{Y}_i^v = \mathbf{X}_i^v \mathbf{e}. \tag{10}$$

Le-Rademacher and Billard (2012) proved that a hyperrectangle representing an interval-valued observation when transformed into a principal components space becomes a polytope. Similarly, each sub-hyperrectangle belonging to a histogram-valued observation i becomes a polytope in principal component space. The vertices of the polytope are the transformed vertices, through Eq. (10), of its corresponding sub-hyperrectangle in the sample space. To construct the polytopes for a histogram-valued observation i, we treat the r_i sub-hyperrectangles that belong to histogram-valued observation i as a set of r_i interval-valued observations but with weights, i.e., with densities d_i^h (say), $h = 1, \ldots, r_i$. Let the matrix of vertices for the hth sub-hyperrectangle be $\mathbf{X}_i^h$. These sub-matrices are extracted from $\mathbf{X}_i^v$. Consider the example of Fig. 2a and its associated matrix of vertices in Eq. (9). For $h = 1$ (the lower left-hand sub-rectangle of Fig. 2a), the matrix of vertices is

$$\mathbf{X}_i^{h=1} = \begin{bmatrix} a_{i11} & a_{i11} & a_{i12} & a_{i12} \\ a_{i21} & a_{i22} & a_{i21} & a_{i22} \end{bmatrix}^T$$

corresponding to the 1st, 2nd, 4th, 5th rows of $\mathbf{X}_i^v$ in Eq. (9); for $h = 2$ (the upper left-hand sub-rectangle of Fig. 2a), we have

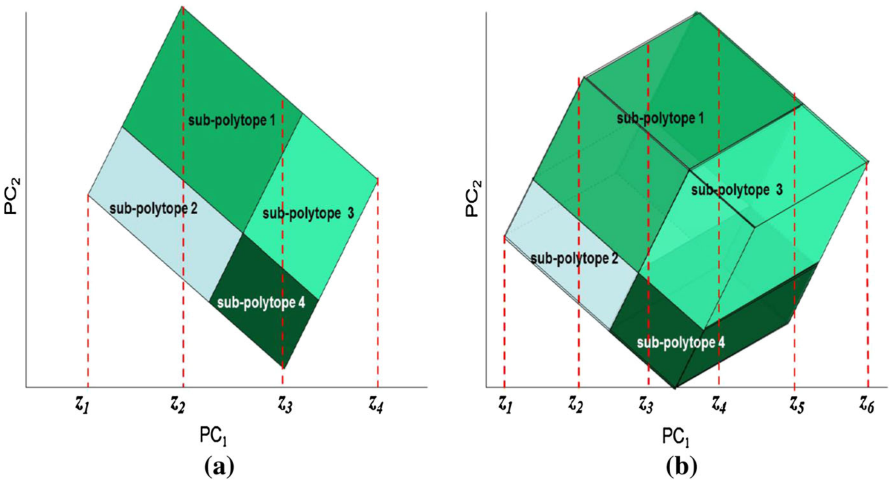

Fig. 3 **a** Subintervals of PC Histogram for $p = 2$. **b** Subintervals of PC Histogram for $p = 3$

$$\mathbf{X}_i^{h=2} = \begin{bmatrix} a_{i11} & a_{i11} & a_{i12} & a_{i12} \\ a_{i22} & a_{i23} & a_{i22} & a_{i23} \end{bmatrix}^T$$

corresponding to the 2nd, 3rd, 5th, 6th rows of $\mathbf{X}_i^v$ in Eq. (9); and so on. Analogous to Eq. (10), the matrix of vertices of the sub-polytope representing sub-hyperrectangle $h, h = 1, \ldots, r_i$, is

$$\mathbf{Y}_i^h = \mathbf{X}_i^h \mathbf{e}. \tag{11}$$

For computational purposes, let $\mathbf{Y}_i$ be a three-dimensional array composed of the collection of $\mathbf{Y}_i^h$ of Eq. (11), i.e., $\mathbf{Y}_i = (\mathbf{Y}_i^1, \ldots, \mathbf{Y}_i^{r_i})$. Construction of $\mathbf{Y}_i$ is given by the algorithm in the Appendix. The sub-polytope representing sub-hyperrectangle h is then constructed by connecting corresponding vertices of $\mathbf{Y}_i^h$ as described in Le-Rademacher and Billard (2012). Figure 3a shows the polytope projection of a two-dimensional hyperrectangle with four sub-rectangles onto a principal component plane. Figure 3b shows the projection of a three-dimensional hyperrectangle.

Although densities of the sub-hyperrectangles, equivalently densities of the sub-polytopes, in observation i vary, illustrating this variability presents some challenges. The first challenge is due to the fact that many of the sub-polytopes in an observation are hidden behind other sub-polytopes. Therefore, only densities of the sub-polytopes on the boundary can be visualized. The second challenge includes computational complexity. If densities are specified by different colors, then all polytopes must be filled with color associated with their density. Figure 3b illustrates that this complexity exists even with a simple three-dimensional observation. Writing a program to automate this process is a time-consuming and a challenging project. This can be a potential future project.

With these challenges in mind, we propose separately constructing polytopes (without densities) for visualization and constructing a matrix of densities associated with the sub-polytopes as an alternative way to understand the variability in their densities.

The resulting densities are further taken into account when computing the histogram-valued output. Computation of the sub-polytope densities and the output histograms are described in the following section.

2.4 Numerical output

Let $\mathbf{d}_i$ be an r_i-vector of densities. Then, the hth element of $\mathbf{d}_i$ is the density of the hth sub-hyperrectangle of observation i. For $j = 1, \ldots, p$ and $k_j = 1, \ldots, s_{ij}$,

$$d_i^h = \prod_{j=1}^{p} p_{ijk_j} \tag{12}$$

where

$$h = \sum_{j=1}^{p-1} \left[(k_j - 1) \prod_{l=j+1}^{p-1} s_{il} \right] + k_p. \tag{13}$$

That is, e.g., the first sub-hyperrectangle of observation i is formed by the first subinterval of histograms X_{ij} for all $j = 1, \ldots, p$. Then, $k_j = 1$ for all j. Thus, from Eq. (13),

$$h = \sum_{j=1}^{p-1} \left[(k_j - 1) \prod_{l=j+1}^{p-1} s_{il} \right] + k_p = 0 + 1 = 1.$$

Hence, the first element of $\mathbf{d}_i$ is, from Eq. (12),

$$d_i^1 = \prod_{j=1}^{p} p_{ijk_j} = \prod_{j=1}^{p} p_{ij1}.$$

Similarly, the second sub-hyperrectangle of observation i is formed by the first subinterval of histograms X_{ij} for all $j = 1, \ldots, p - 1$ and the second subinterval of X_{ip}. Then, $k_j = 1$ for $j = 1, \ldots, p - 1$ and $k_p = 2$. Therefore,

$$h = \sum_{j=1}^{p-1} \left[(k_j - 1) \prod_{l=j+1}^{p-1} s_{il} \right] + k_p = 0 + 2 = 2.$$

Hence, the second element of $\mathbf{d}_i$ is

$$d_i^2 = \prod_{j=1}^{p} p_{ijk_j} = \left(\prod_{j=1}^{p-1} p_{ij1} \right) p_{ip2}.$$

Other elements of $\mathbf{d}_i$ can be found the same way. The value of $\mathbf{d}_i$ is the density of the sub-hyperrectangle whose vertices are expressed in matrix $\mathbf{X}_i^h$. Equivalently, $\mathbf{d}_i$ is the density of the sub-polytope whose vertices are expressed in matrix $\mathbf{Y}_i^h$ of Eq. (11).

Prior to incorporating the sub-polytope densities into the relative frequencies of the histogram output, construct the principal component (PC) histogram corresponding to each sub-polytope of observation i as detailed in Le-Rademacher and Billard (2013). Let

$$Z_{i\nu}^{h} = \{[z_1, z_2), p_1; \ldots ; [z_{s-1}, z_s], p_s\} \tag{14}$$

denote the resulting PC_ν histogram representing sub-polytope h of observation i where s is the number of subintervals in $Z_{i\nu}^{h}$ and p_k is the relative frequency of subinterval $[z_k, z_{k+1})$ ignoring the sub-polytope density. Then, the relative frequency of $[z_k, z_{k+1})$ is $d_i^h p_k$ after adjusting for the density. Now, the adjusted PC_ν histogram for the sub-polytope h is $Z_{i\nu}^{h} = \{[z_1, z_2), d_i^h p_1; \ldots ; [z_{s-1}, z_s], d_i^h p_s\}$.

Next, we propose combining the r_i adjusted histograms $Z_{i\nu}^{h}$ for all $h = 1, \ldots, r_i$ into one histogram representing PC_ν of observation i. For convenience, we propose creating a combined histogram of equal subinterval widths. Due to the linear transformation, each subinterval along the PC_ν axis may contain parts of several sub-polytopes. For example, in Fig. 3a the first subinterval $[z_1, z_2)$ includes part of polytopes 1 and 2; whereas subinterval $[z_2, z_3)$ includes part of all four sub-polytopes. The subinterval relative frequencies for the combined histogram must correctly account for the proportion of various sub-polytopes that contribute to the subinterval. Figure 3b shows the process is more complex with higher dimensional data. Steps to compute the combined histograms are described in "Constructing the PC Histograms" in the Appendix.

3 Algorithm

The algorithm needed to carry out the proposed analysis can be divided into four main components. For each observation, do the following:

A.1 Construct the matrix of vertices $\mathbf{X}_i^v$ of the observed sub-hyperrectangles.

A.2 Construct the polytopes by:
- computing the transformed vertices $\mathbf{Y}_i^v (= \mathbf{X}_i^v \mathbf{e})$ and then connect the appropriate vertices to build the polytopes and
- computing the density vector $\mathbf{d}_i$.

A.3 Plot two- or three-dimensional projection of the polytopes from A.2 onto a principal component space.

A.4 Compute principal component histograms by:
- computing the principal component histogram for each sub-polytope from A.2 and then
- combining the sub-polytope histograms into one histogram using the densities computed from A.2.

Details of the algorithm are given in the Appendix. An R-program and an example data are given as supplemental materials.

Although the structure of histogram observations can be complex, especially with a large number of variables, we carefully define the matrix of vertices in Eq. (8) to reduce the size of the data matrix to increase the efficiency of the algorithm. Computation of the polytopes and the principal histograms is straight forward. The algorithm involves permutation of the vertices (part A.1) and direct matrix computations (parts A.2−A.4).

Table 1 Covariance matrix for China weather variables

Variable	X_1 Nov–Mar	X_2 Apr & Oct	X_3 May & Sep	X_4 Jun–Aug	X_5 Elevation
X_1	206.24	127.52	89.73	51.94	−15.88
X_2	127.52	82.75	59.68	37.07	−40.18
X_3	89.73	59.68	45.10	29.60	−50.68
X_4	51.94	37.07	29.60	22.17	−59.71
X_5	−15.88	−40.18	−50.69	−59.71	390.77

Table 2 Eigenvalues λ_ν and eigenvectors $\mathbf{e}_\nu = (e_{\nu 1}, \ldots, e_{\nu p})$

ν:	1	2	3	4	5
λ_ν	3.877	1.078	0.028	0.010	0.008
Proportion of variance	0.775	0.215	0.006	0.002	0.002
$\mathbf{e}_\nu^T$	$\mathbf{e}_1^T$	$\mathbf{e}_2^T$	$\mathbf{e}_3^T$	$\mathbf{e}_4^T$	$\mathbf{e}_5^T$
X_1	0.465	0.378	0.430	0.617	0.276
X_2	0.493	0.220	0.205	−0.310	−0.756
X_3	0.505	0.057	−0.071	−0.628	0.585
X_4	0.486	−0.248	−0.751	0.356	−0.105
X_5	−0.223	0.863	−0.452	−0.047	0.003

Table 3 Correlation between PCs and original variables

Variable	X_1	X_2	X_3	X_4	X_5
PC_1	0.915	0.971	0.995	0.957	−0.439
PC_2	0.392	0.228	0.060	−0.258	0.896

With current computing resources, these procedures are routinely done and do not take much time. As a reference, it took 30 s to produce the results shown in Tables 1, 2, 3, 4 and Fig. 4 of Sect. 4 using a personal computer with a 3.00 GHz processor and 8 GB of RAM.

The efficiency of this algorithm follows the development of Le-Rademacher and Billard (2012, 2013). First, the polytopes are constructed by connecting the vertices of the polytopes for all observations in the same sequence; see Le-Rademacher and Billard (2012) for a detailed exposition. This allows us to simply recreate the true shape of the observations in the PC space without using a search procedure, e.g., the parallel edge connected shapes of Irpino et al. (2003), which is often much more computationally intensive. Secondly, the output histograms are computed using a two-dimensional projection of the polytopes onto the $PC_1 \times PC_\nu$ plane to ensure the resulting histograms reflect the largest source of variation in the data while keeping the computation manageable. Furthermore, the symmetry of the sub-polytopes means it is only necessary to compute the subinterval endpoints and the subinterval frequencies of the first half of each sub-histogram. The endpoints and the frequencies of the

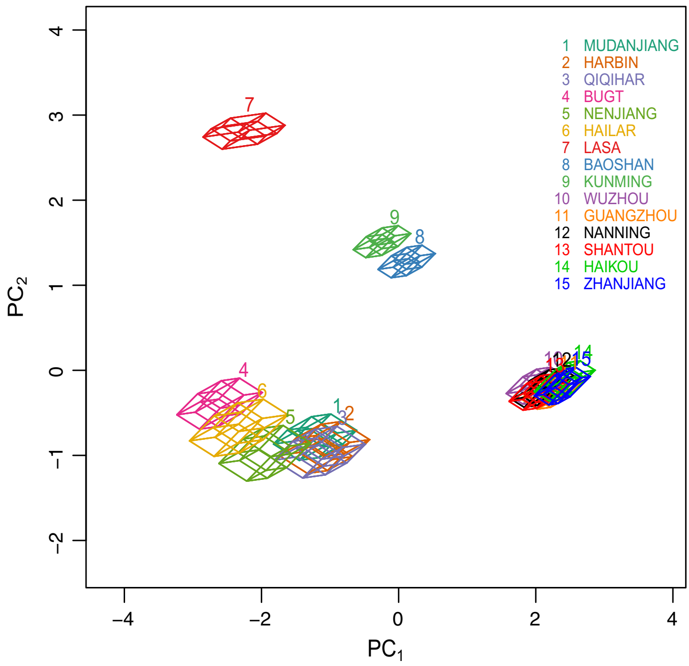

Fig. 4 Polytope projections onto $PC_1 \times PC_2$ space

second half of each sub-histogram can be directly copied from the first half; see Le-Rademacher and Billard (2013) for details. This further improves the efficiency of our algorithm.

4 Illustration

The foregoing theory is illustrated on a data set consisting of temperatures at a subset of 15 weather stations in China available from the website <http://dss.ucar.edu/datasets/ds578.1>. At each location, the average monthly temperatures from 1951 to 2000 were aggregated into histograms with five subintervals. Months represented the variables; however, since the temperatures for the months November–March were essentially the same and so not useful as a distinguishing feature, these five months were combined into one variable "Nov–Mar" (X_1). Similarly, the temperatures for April and October are combined into "Apr & Oct" (X_2), May and September into "May & Sep" (X_3), and June through August into "Jun–Aug" (X_4). Elevation for each station was also added as a 5th variable (X_5), with the single classical value expressed as an interval $z = [z, z]$. Thus, we have $p = 5$ variables, $n = 15$ histogram-valued observations with $s_{ij} = 5$ subintervals, for each $i = 1, \ldots, n$, $j = 1, \ldots, 4$, and $s_{ij} = 1$ when $j = 5$. For illustrative purposes, we make the not-unreasonable assumption that all variables and observations are equally weighted.

The sample covariance matrix can then be calculated from Eq. (4); see Table 1.

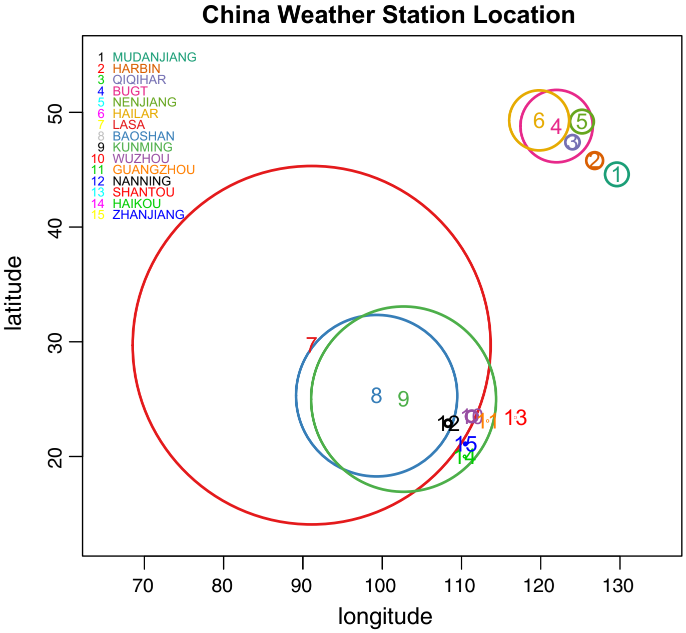

Fig. 5 Latitude × longitude location of weather stations

Then, from Eq. (6), we can obtain the correlation matrix Σ as

$$\Sigma = \begin{bmatrix} 1.00 & 0.98 & 0.93 & 0.77 & -0.06 \\ 0.98 & 1.00 & 0.98 & 0.87 & -0.22 \\ 0.93 & 0.98 & 1.00 & 0.94 & -0.38 \\ 0.77 & 0.87 & 0.94 & 1.00 & -0.64 \\ -0.06 & -0.22 & -0.38 & -0.64 & 1.00 \end{bmatrix}.$$

The eigenvalues of Σ are as shown in Table 2, along with the associated eigenvectors. Hence, the principal components can be calculated through Eq. (5). Table 2 also shows that together the first two principal components explain 99 % of the total variance in the data with PC_1 accounting for 77.5 % and PC_2 an additional 21.5 %. The correlation between these two principal components and the original variables (Table 3) shows a strong correlation between PC_1 and monthly temperatures whereas PC_2 is strongly correlated with station elevation.

The algorithm of Sect. 3 can now be applied to produce the polytopes in the principal component space corresponding to each observation. These are displayed in Fig. 4. The latitude × longitude locations of these stations are shown in Fig. 5, where the differing size of the circle reflects the differing station elevations. From Fig. 4, it is clear that observations $i = 1, \ldots, 6$ (i.e., stations 1 = Mudanjiang, 2 = Harbin, 3 = Qiqihar, 4 = Bugt, 5 = Nenjiang, and 6 = Hailar, respectively) form one cluster; observation $i = 7$ (i.e., station 7 = Lasa) is a one-station cluster; observations $i = 8, 9$ (i.e., 8 = Baoshan, and 9 = Kunming, respectively) form a third cluster, and observations

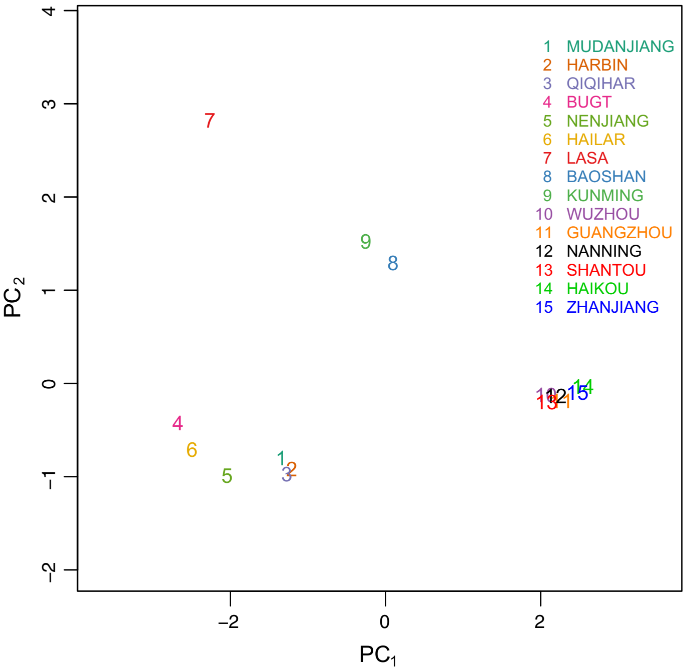

Fig. 6 Projections onto $PC_1 \times PC_2$ space of classical surrogates

$i = 10, \dots, 15$ (i.e., 10 = Wuzhou, 11 = Guangzhou, 12 = Nanning, 13 = Shantou, 14 = Haikou, and 15 = Zhanjiang, respectively) comprise the fourth cluster.

When the locations of Fig. 5 are considered, it is evident that the first cluster consists of stations all located in the north-east and so can be reasonably expected to have somewhat similar weather (though the elevations, especially for Bugt and Hailar, are not all comparable; but this difference is reflected in the polytopes for these two stations being in the left side of the cluster of polytopes for this location). Likewise, the stations in the fourth cluster are all located in the south-east part of the country (see Fig. 5), all with the same low elevations. This similarity is reflected in the comparable polytopes in the $PC_1 \times PC_2$ space in Fig. 4.

The stations Baoshan and Kunming (#8–9), though located in the south, are further west and at higher elevations; hence, the separate cluster matches their actual geography. Here, the elevations as well as the temperature histograms are very similar, with the result that the polytopes are almost identical. Lasa is situated even further west and is much higher in elevation than any of the other 14 stations. Therefore, its temperature patterns would be quite distinct. Thus, the polytope in this case is quite different and quite isolated from those of other stations; see top-left of Fig. 4.

If instead of keeping the variations in the temperatures across years (as in the histogram observations), overall averages were used as classical surrogates, then the resulting projections of these surrogates gave the point projections on $PC_1 \times PC_2$ space of Fig. 6. The clusters are again clear, and match those obtained from the histogram values, thus validating the methodology proposed herein.

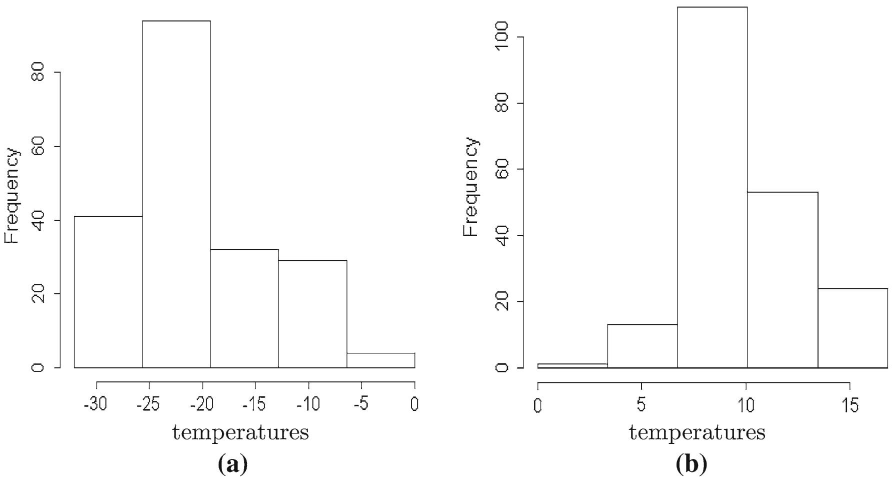

Fig. 7 **a** March histogram for Hailar (#6). **b** March histogram for Kunming (#9)

Finally, relative sizes of polytopes provide an additional interpretation. For example, consider the stations Hailar (#6) and Kunming (#9). The polytope for Hailar is larger in size than is that for Kunming on $PC_1 \times PC_2$ space. This difference is explained by the fact that the range of temperatures and the internal variations of the histograms for Hailar are larger than those for Kunming. (This difference in variations is evident in Fig. 7 which shows the relevant histograms for these two stations for the March temperatures; likewise for other months). In contrast, the classical principal component values for these two stations are but points on $PC_1 \times PC_2$, reflecting that classical observations themselves have no internal variation. Note that although Kunming is at a higher elevation than is Hailar, the elevation variable is classically valued and so has no internal variation. That is, the elevation variable value here is not a factor in the size of the polytope (only in its position on $PC_1 \times PC_2$ space); it is the different internal values of the histograms that are reflected in the different sizes of the polytopes. This additional interpretation from an histogram-valued principal component analysis is not possible from classically valued analyses.

We further computed the PC histograms for each weather station. The histograms for the first principal component with five subintervals of equal width are shown in Table 4. Note that the purpose of the PC histograms is to provide numerical output for the resulting principal components. Figure 8 provides a graphical depiction of the PC_1 histograms to illustrate the internal distribution of the output histograms. The polytopes of Fig. 4 are a better visualization tool. In addition to the clustering pattern seen in Fig. 4, the resulting PC histograms indicate similarity in the distribution of the PC_1 histograms within each cluster. The distribution of PC_1 values for observations $i = 1, \ldots, 6$ have larger spread and are slightly skewed to the right with more than 90 % of the values spread between the second and third subintervals. In contrast, the distribution for observations $i = 10, \ldots, 15$ are more narrow and somewhat more symmetric with 50–60 % of values located near the center of the histograms. Distributions for the observations, $i = 8, 9$ are again similar. When these PC histograms are input into

Table 4 Histogram output for principal component 1

Obs.	**PC**$_1$ Histogram
1	[−1.83, −1.58), 0.038; [−1.58, −1.34), 0.514; [−1.34, −1.10), 0.414; [−1.10, −0.86), 0.034; [−0.86, −0.61], 0.0002
2	[−1.75, −1.48), 0.018; [−1.48, −1.22), 0.471; [−1.22, −0.95), 0.481; [−0.95, −0.69), 0.030; [−0.69, −0.42], 0.0001
3	[−1.81, −1.55), 0.024; [−1.55, −1.28), 0.487; [−1.28, −1.02), 0.458; [−1.02, −0.75), 0.032; [−0.75, −0.49], 0.0001
4	[−3.23, −2.99), 0.007; [−2.99, −2.74), 0.291; [−2.74, −2.49), 0.612; [−2.49, −2.24), 0.089; [−2.24, −1.99], 0.0004
5	[−2.62, −2.35), 0.014; [−2.35, −2.08), 0.363; [−2.08, −1.82), 0.555; [−1.82, −1.55), 0.067; [−1.55, −1.28], 0.0004
6	[−3.05, −2.76), 0.038; [−2.76, −2.48), 0.516; [−2.48, −2.19), 0.419; [−2.19, −1.90), 0.027; [−1.90, −1.61], 0.0001
7	[−2.85, −2.61), 0.005; [−2.61, −2.38), 0.194; [−2.38, −2.14), 0.656; [−2.14, −1.90), 0.144; [−1.90, −1.66], 0.0012
8	[−0.30, −0.13), 0.003; [−0.13, 0.03), 0.274; [0.03, 0.20), 0.638; [0.20, 0.37), 0.084; [0.37, 0.53], 0.0004
9	[−0.66, −0.49), 0.004; [−0.49, −0.33), 0.230; [−0.33, −0.16), 0.623; [−0.16, 0.01), 0.142; [0.01, 0.18], 0.0017
10	[1.57, 1.75), 0.002; [1.75, 1.93), 0.130; [1.93, 2.11), 0.641; [2.11, 2.29), 0.224; [2.29, 2.46], 0.0033
11	[1.79, 1.97), 0.004; [1.97, 2.14), 0.177; [2.14, 2.32), 0.644; [2.32, 2.50), 0.172; [2.50, 2.68], 0.0027
12	[1.65, 1.83), 0.001; [1.83, 2.01), 0.058; [2.01, 2.20), 0.610; [2.20, 2.38), 0.325; [2.38, 2.57], 0.0051
13	[1.61, 1.78), 0.004; [1.78, 1.96), 0.177; [1.96, 2.13), 0.644; [2.13, 2.31), 0.172; [2.31, 2.48], 0.0030
14	[1.98, 2.15), 0.001; [2.15, 2.33), 0.034; [2.33, 2.51), 0.463; [2.51, 2.69), 0.479; [2.69, 2.86], 0.0229
15	[1.98, 2.14), 0.003; [2.14, 2.31), 0.071; [2.31, 2.47), 0.570; [2.47, 2.63), 0.345; [2.63, 2.79], 0.0118

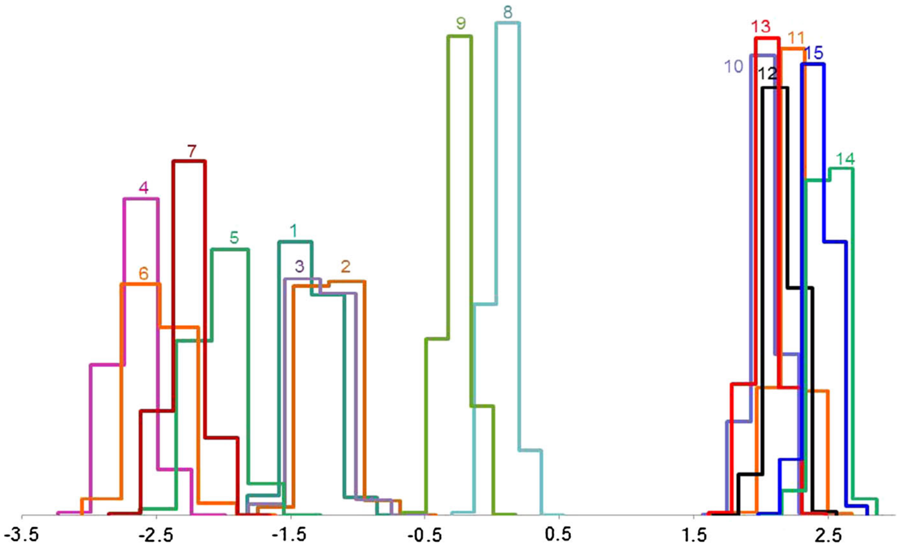

Fig. 8 Graphical depiction of PC_1 histogram output of Table 4

models for further analysis, they reflect the varying distributions within observations along the PC axes.

5 Conclusions

In this paper, we propose a principal component analysis methodology for histogram-valued data by expanding the symbolic covariance method that Le-Rademacher and Billard (2012, 2013) proposed for interval-valued data. Currently, there is no comparable PCA methodology for histogram-valued observations under the symbolic data domain.

The principal components in this method are computed from the symbolic covariance matrix to account for the total variance of histogram data. The polytopes of Le-Rademacher and Billard (2012) are adapted to represent the observations in the principal component space by treating the observed histograms as collections of weighted sub-hyperrectangles. We further propose computing a density matrix associated with the sub-hyperrectangles to reflect the varying densities of the sub-hyperrectangles within each observation. The density matrix is also used when computing the numerical output to ensure the relative frequencies of the output histograms, adapted from Le-Rademacher and Billard (2013), correctly accounting for the sub-hyperrectangle densities.

By basing our method on the polytopes of Le-Rademacher and Billard (2012) and the histogram output of Le-Rademacher and Billard (2013), our method inherits the computational efficiency of their methods by using routine matrix computation and avoiding search algorithms. However, generalizing the polytopes and histogram output to histogram-valued observations is not a trivial process as described in Sect. 2 and as illustrated by the detailed algorithm of the Appendix. By carefully defining the matrix of vertices, our method requires little additional power and retains computational efficiency.

Supplemental materials

R program: The R program to compute the principal components, the polytopes, and the output histograms as proposed in the article.
Example data: The temperature data used for illustration in the article.

Appendix: Algorithm

The algorithm to construct the polytope representation of the observations on principal component space has essentially two parts. The first part ("Constructing the matrix of vertices" in the Appendix) constructs the matrices of vertices needed to build the actual polytopes. Then ("Constructing the polytopes" in the Appendix) the construction of the polytopes per se is described. Extensions to two- and three-dimensional polytope plots are given in "Constructing Two and Three Dimensional Plots" in the Appendix. The algorithm to compute the histograms from the resulting polytopes is given in "Constructing the PC histograms" in the Appendix. The indexing notation used in these algorithms is similar to that of the R language. Therefore, the position for an element of a vector, a matrix or an array is specified in a pair of square brackets, [ ]. The index for an element of a vector is enclosed in the brackets. An element of a matrix is specified by a pair of numbers separated by a comma. The first number specifies the row and the second number specifies the column. The position of an array is specified by three numbers separated by commas corresponding to row, column, and matrix, respectively. Also, we use the lower case to represent an observed data matrix [e.g., $\mathbf{x}_i^v$ to distinguish it from the random data matrix $\mathbf{X}_i^v$ of Eq. (8)].

Constructing the matrix of vertices

First, assume that the observed data vector $\mathbf{x}_i$ has been separated into a vector of subinterval endpoints and a vector of the relative frequencies. That is, let $\mathbf{x}_{ep}$ be the vector of subinterval endpoints and $\mathbf{x}_{rf}$ be the vector of subinterval relative frequencies. Then, $\mathbf{x}_{ep}$ has $\sum_{j=1}^{p} (s_{ij} + 1)$ elements and has the form

$$\mathbf{x}_{ep} = \left[a_{i11} \; \ldots \; a_{i1(s_{i1}+1)} \; \ldots \; a_{ip1} \; \ldots \; a_{ip(s_{ip}+1)} \right]$$

where a_{ijk}, for $k = 1, \ldots, s_{ij} + 1$ and $j = 1, \ldots, p$, are elements of the set E_{ij}. The vector $\mathbf{x}_{rf}$ has $\sum_{j=1}^{p} s_{ij}$ elements and has the form

$$\mathbf{x}_{rf} = \left[p_{i11} \; \ldots \; p_{i1s_{i1}} \; \ldots \; p_{ip1} \; \ldots \; p_{ips_{ip}} \right]$$

where p_{ijk} is the relative frequency of the kth subinterval of the observed histogram x_{ij}. Before creating the matrix of vertices for observation i, a p-vector whose elements are the number of subintervals for X_{ij} is also needed. Let **ns** denote the vector of number of subintervals of X_{ij}. Then,

$$\mathbf{ns} = \left[s_{i1} \, \ldots \, s_{ip} \right].$$

With the information in $\mathbf{x}_{ep}$, $\mathbf{x}_{rf}$, and **ns**, we can proceed with constructing the matrix of vertices $\mathbf{x}_i^v$ using the following five steps: *Step* 1 : Create a $(p+1)$-vector **nr** whose $(j+1)^{\text{th}}$ element, for $j = 1, \ldots, p$, is the number of times that points a_{ijk}, for $k = 1, \ldots, s_{ij} + 1$, must be repeated in Step 5 below. The first element of **nr** is the number of rows of the matrix of observed vertices, $\mathbf{x}_i^v$.

1. For $j = 1, \ldots, p$, set $\mathbf{nr}[p - j + 1] = \prod_{l=p-j+1}^{p} (s_{il} + 1)$.
2. Set $\mathbf{nr}[p + 1] = 1$.

Step 2 : Create a $(p+1)$-vector $\mathbf{nr}_p$ whose $(j+1)^{th}$ element, for $j = 1, \ldots, p$, is the number of sub-hyperrectangles present in observation i when all variables up to j are excluded.

1. For $j = 1, \ldots, p$, set $\mathbf{nr}_p[p - j + 1] = \prod_{l=p-j+1}^{p} s_{il}$.
2. Set $\mathbf{nr}_p[p + 1] = 1$.

Step 3 : Create a p-vector **sp** whose jth element is the position of the element of $\mathbf{x}_{ep}$ which is the first subinterval endpoint for variable j.

1. Set $\mathbf{sp}[1] = 1$.
2. For $j = 1, \ldots, p - 1$, set $\mathbf{sp}[j + 1] = \sum_{l=1}^{j} (s_{il} + j + 1)$.

Step 4 : Initialize the matrix of observed vertices $\mathbf{x}_i^v$ by letting $\mathbf{x}_i^v$ be an $(N_i \times p)$ matrix of zeros where $N_i = \prod_{j=1}^{p} (s_{ij} + 1)$.
Step 5 : Update the elements of $\mathbf{x}_i^v$ by

```
1. For $j = 1, \ldots, p$, do
   (a) Let $nj = \mathbf{ns}[j]$.
   (b) Let $rj = \mathbf{nr}[j + 1]$.
   (c) Let $sj = \mathbf{sp}[j]$.
   (d) For $l = 0, \ldots, nj$,
         – For $k = 1, \ldots, rj$,
         – set $\mathbf{x}_i^v[l(rj) + k, j] = \mathbf{x}_{ep}[sj + l]$.
2. For $j = 2, \ldots, p$, do
   (a) Let $tj = \frac{\mathbf{nr}[1]}{\mathbf{nr}[j]} - 1$.
   (b) Let $rj = \mathbf{nr}[j]$.
   (c) For $l = 1, \ldots, tj$,
         – For $k = 1, \ldots, rj$,
         – set $\mathbf{x}_i^v[l(rj) + k, j] = \mathbf{x}_i^v[k, j]$.
```

End of Step 5. At the end of Step 5, we obtain the matrix $\mathbf{x}_i^v$ whose rows are the coordinates of the vertices of observation i.

Constructing the polytopes

The following algorithm includes seven steps:
Step 1 : First, compute the matrix of transformed vertices, $\mathbf{y}_i^v (= \mathbf{x}_i^v \mathbf{e})$, for the polytope representing observation i in a principal components space.
Step 2 : Next, create a three-dimensional array $\mathbf{y}_i$ to store the transformed vertices that belong to a sub-polytope together. The array $\mathbf{y}_i$ is a result of combining $r_i = \prod_{j=1}^{p} s_{ij}$ matrices $\mathbf{y}_i^h$ where $h = 1, \ldots, r_i$. Each matrix $\mathbf{y}_i^h$ of dimension $(2^p \times p)$ contains coordinates of all vertices that belong to sub-polytope h.

```
1. Initialize array $\mathbf{y}_i$ by letting $\mathbf{y}_i$ be an array of zeros with dimension $(2^p \times p \times r_i)$.
2. Update the elements of $\mathbf{y}_i$ by running the following nested loop,
   (a) Set $kr_0 = 0$ and $ni_0 = 0$.
   (b) For $j = 1, \ldots, p-1$,
         – For $l_j = 0, \ldots, s_{ij}$,
             i. Let $kr_j = kr_{j-1} + (\mathbf{nr}[j+1])l_j$.
            ii. Let $ni_j = ni_{j-1} + (\mathbf{nr}_p[j+1])l_j$.
           iii. For $k = 1, \ldots, \mathbf{ns}[p]$,
             A. Set $kr = kr_{p-1} + k$.
             B. Set $ni = ni_{p-1} + k$.
             C. Set $\mathbf{y}_i[1, , ni] = \mathbf{y}_i^v[kr, ]$
             D. For $o = 1, \ldots, p$, do
                  For $r = 1, \ldots, 2^{(o-1)}$,
                  set $\mathbf{y}_i[2^{(o-1)} + r, , ni] = \mathbf{y}_i^v[kr[r] + \mathbf{nr}[p-o+2], ]$ and
                  set $kr = (kr, kr[r] + \mathbf{nr}[p-o+2])$.
```

Step 3 : Next, reconstruct polytopes corresponding to sub-hyperrectangles of observation i by following the next two sub-steps.
Step 3 − A. Construct the matrix of connected vertices $\mathbf{C}$ associated with $\mathbf{y}_i^v$ as follows:

```
1. Initialize $\mathbf{C}$ as a $2^p \times p$ matrix of zeros.
2. Update $\mathbf{C}$ by doing the following step for $j = 1, \ldots, p$,
   – For $j_1 = 0, \ldots, 2^{(j-1)} - 1$, do
       – For $j_2 = ((2^{(p-j+1)})j_1 + 1), \ldots, ((2^{(p-j+1)})j_1 + 2^{(p-j)})$, set $\mathbf{C}[j_2, j] = j_2 + 2^{(p-j)}$.
       – For $j_2 = ((2^{(p-j+1)})j_1 + 2^{(p-j)} + 1), \ldots, ((2^{(p-j+1)})j_1 + 2^{(p-j+1)})$,
       – set $\mathbf{C}[j_2, j] = j_2 - 2^{(p-j)}$.
```

Step 3 − B A p-dimensional plot of the polytopes is constructed in the principal component space by the following two steps:

```
1. Make a scatter plot of $\mathbf{y}_i^v$.
2. Construct the vertices of each sub-polytope as follows, for each $h = 1, \ldots, r_i$,
     For $v_1 = 1, \ldots, 2^p$, do
     for $j_1 = 2, \ldots, p+1$,
     set $v_2 = \mathbf{C}[v_1, j_1]$, and
     connect the points $\mathbf{y}_i[v_1, h]$ and $\mathbf{y}_i[v_2, h]$ with a line.
```

End of Step 3. We now have a plot of the r_i polytopes representing of observation $\mathbf{x}_i$, $i = 1, \ldots, n$, in PC space. This step is an adaptation of Steps 3–4 for obtaining the polytope for interval-valued data; see Le-Rademacher (2008) and (Le-Rademacher and Billard (2012), Supplemental Material)

At the end of Step 3, polytopes representing observation i in a principal component space are plotted. While these polytopes are now constructed, we recall that the densities of a histogram observation vary across the hyperrectangles. To create the vector of densities for these polytopes, follow the next 4 steps.

Step 4 : Create a p-vector $\mathbf{sp}_p$ whose jth element is the position of the element of $\mathbf{x}_{rf}$ which is the first subinterval relative frequency for variable j.

```
1. Set $\mathbf{sp}_p[1] = 1$.
2. For $j = 1, \ldots, p-1$,
       set $\mathbf{sp}_p[j+1] = \sum_{l=1}^{j} s_{il} + 1$.
```

Step 5 : Let $\mathbf{x}_p^v$ be an $(r_i \times p)$ matrix of relative frequencies. The row h of $\mathbf{x}_p^v$ contains the relative frequencies of subintervals making up sub-hyperrectangle h. Initialize $\mathbf{x}_p^v$ by setting all elements of $\mathbf{x}_p^v$ to zeros.

Step 6 : Update the elements of $\mathbf{x}_p^v$ by

```
1. For $j = 1, \ldots, p$, do
     – Let $nj = \mathbf{ns}[j]$.
     – Let $rj = \mathbf{nr}_p[j+1]$.
     – Let $sj = \mathbf{sp}_p[j]$.
     – For $l = 0, \ldots, nj-1$,
         – For $k = 1, \ldots, rj$,
         – set $\mathbf{x}_p^v[l(rj)+k, j] = \mathbf{x}_{rf}[sj+l]$.
2. For $j = 2, \ldots, p$, do
     – Let $tj = \frac{\mathbf{nr}_p[1]}{\mathbf{nr}_p[j]} - 1$.
     – Let $rj = \mathbf{nr}_p[j]$.
     – For $l = 1, \ldots, tj$,
         – For $k = 1, \ldots, rj$,
             set $\mathbf{x}_p^v[l(rj)+k, j] = \mathbf{x}_p^v[k, j]$.
```

Step 7 : Let $\mathbf{d}_i$ be an r_i-vector whose elements are densities of the sub-hyperrectangles belonging to observation i. The density for each sub-hyperrectangle is the product of relative frequencies of the p subintervals making up that sub-hyperrectangle. That is, for $h = 1, \ldots, r_i$, $\mathbf{d}_i[h] = \prod_{j=1}^{p} \mathbf{x}_p^v[h, j]$.

At the end of Step 7, we obtain a vector of densities $\mathbf{d}_i$ whose hth element is the density of sub-hyperrectangle h of observation i.

Constructing two and three dimensional plots

Usually, visualization of the projections of observations onto the principal component space is limited to two dimensions, $PC_{\nu_1} \times PC_{\nu_2}$. This is achieved by replacing the substeps 1 and 2 in Step 3-B of the polytope algorithm of Sect. 1, by the following three substeps:

1. Let $\mathbf{y}_i^{(2)}$ be the $r_i 2^p \times 2$ matrix whose first and second columns are, respectively, columns ν_1 and ν_2 of $\mathbf{y}_i^v$.

2. Make a scatter plot of $\mathbf{y}_i^{(2)}$.
3. Connect corresponding points of $\mathbf{y}_i^{(2)}$ by using substep 2 of Step 3-B of Sect. 1, with $\mathbf{y}_i^{v}$ replaced by $\mathbf{y}_i^{(2)}$; now $p = 2$.

To construct a three-dimensional plot of $PC_{\nu_1} \times PC_{\nu_2} \times PC_{\nu_3}$, follow the same three steps as here for constructing two-dimensional plots except that $\mathbf{y}_i^{(2)}$ is replaced by $\mathbf{y}_i^{(3)}$ where now, in substep 1, $\mathbf{y}_i^{(3)}$ is an $(r_i 2^p \times 3)$ matrix with columns ν_1, ν_2, and ν_3 of $\mathbf{y}_i^{v}$. In substep 3, $p = 3$.

Constructing the PC histograms

The following algorithm constructs a histogram representing the ν^{th} principal component for observation i by first computing the PC histograms corresponding to the sub-polytopes of observation i, then combine the r_i histograms into one PC_ν histogram representing observation i.

Step 1 : Follow the algorithm of Le-Rademacher and Billard (2013) to create the $(r_i \times 3s)$ matrix $\mathbf{z_{i\nu}}$ whose hth row contains the subinterval endpoints and the relative frequencies for sub-polytope h as specified in Eq. (14). Here, elements $3k, k = 1\ldots, s$, of $\mathbf{z}_{i\nu}[h,]$ are the unadjusted relative frequencies of the histogram representing sub-polytope h.

Step 2 : Update the relative frequencies from Step 1 by setting $\mathbf{z}_{i\nu}[h, 3k] = \mathbf{d}_i[h]\mathbf{z}_{i\nu}[h, 3k]$.

Step 3 : This next step combines the s histograms in $\mathbf{z}_{i\nu}$ into one histogram with subintervals of equal width.

1. Let lo and hi be the lowest and the highest endpoints of the r_i histograms of observation i. Then, $lo = min(\mathbf{z}_{i\nu}[, 1])$ and $hi = max(\mathbf{z}_{i\nu}[, 3s - 1])$.
2. Let sn denote the desired number of subintervals for the combined histogram. Then, the widths of the subintervals are $sw = (hi - lo)/sn$.
3. Let **hm** be an $(sn \times 3)$ transition matrix whose columns 1 and 2 contain the subinterval endpoints and column 3 contains the relative frequencies of the combined PC_ν histogram for observation i. Initialize **hm** by setting its elements to zero.
4. Update **hm** as follows, For $t = 1, \ldots, ns$, do
    (a) Set the endpoints of subinterval t by letting $\mathbf{hm}[t, 1] = lo + (sw)(t - 1)$ and $\mathbf{hm}[t, 2] = lo + (sw)t$.
    (b) Let **fr** be an $(r_i \times s)$ matrix whose (h, q) element corresponds to the proportion of subinterval q of sub-polytope h that falls within the subinterval t. Initialize **fr** by setting its elements to zero.
    - For $h = 1, \ldots, r_i$, do
    - For $q = 1, \ldots, s$, do
      Case a : If $(\mathbf{z}_{i\nu}[h, 3q - 2] \geq \mathbf{hm}[s, 1])$ and $(\mathbf{z}_{i\nu}[h, 3q - 1] \leq \mathbf{hm}[s, 2])$, set $\mathbf{fr}[h, q] = \mathbf{z}_{i\nu}[h, 3q]$.
      Case b : If $(\mathbf{z}_{i\nu}[h, 3q-2] \geq \mathbf{hm}[s, 1])$ and $(\mathbf{z}_{i\nu}[h, 3q-2] < \mathbf{hm}[s, 2])$ and $\mathbf{z}_{i\nu}[h, 3q - 1] > \mathbf{hm}[s, 2]$, set $\mathbf{fr}[h, q] = \frac{(\mathbf{z}_{i\nu}[h,3q])(\mathbf{hm}[s,2]-\mathbf{z}_{i\nu}[h,3q-2])}{\mathbf{z}_{i\nu}[h,3q-1]-\mathbf{z}_{i\nu}[h,3q-2]}$.

Case c : If ($\mathbf{z}_{i\nu}[h, 3q-2] < \mathbf{hm}[s, 1]$) and ($\mathbf{z}_{i\nu}[h, 3q-1] > \mathbf{hm}[s, 1]$) and ($\mathbf{z}_{i\nu}[h, 3q-1] \leq \mathbf{hm}[s, 2]$), set $\mathbf{fr}[h, q] = \frac{(\mathbf{z}_{i\nu}[h,3q])(\mathbf{z}_{i\nu}[h,3q-1]-\mathbf{hm}[s,1])}{\mathbf{z}_{i\nu}[h,3q-1]-\mathbf{z}_{i\nu}[h,3q-2]}$.
Case d : If ($\mathbf{z}_{i\nu}[h, 3q-2] < hm[s, 1]$) and ($\mathbf{z}_{i\nu}[h, 3q-1] > \mathbf{hm}[s, 2]$), set $\mathbf{fr}[h, q] = \frac{(\mathbf{z}_{i\nu}[h,3q])(\mathbf{hm}[s,2]-\mathbf{hm}[s,1])}{\mathbf{z}_{i\nu}[h,3q-1]-\mathbf{z}_{i\nu}[h,3q-2]}$.

(c) Let $\mathbf{hm}[t, 3] = \sum_{h=1}^{r_i} \sum_{q=1}^{s} \mathbf{fr}[h, q]$.

5. Let $sh = \sum_{t=1}^{ns} \mathbf{hm}[t, 3]$.
6. Update $\mathbf{hm}[t, 3] = \mathbf{hm}[t, 3]/sh$.

At the end of this step, we have the subinterval endpoints and the relative frequencies for the combined histogram. Let $\mathbf{pc}_\nu$ be the $(n \times ns)$ matrix whose ith row contains the PC_ν histogram for observation i. Then, for $t = 1, \ldots, ns$, do

1. Let $\mathbf{pc}_\nu[i, 3t-2] = \mathbf{hm}[t, 1]$.
2. Let $\mathbf{pc}_\nu[i, 3t-1] = \mathbf{hm}[t, 2]$.
3. Let $\mathbf{pc}_\nu[i, 3t] = \mathbf{hm}[t, 3]$.

This step concludes the histogram algorithm. Repeat these steps for all observations.

References

Anderson TW (1963) Asymptotic theory for principal components analysis. Ann Math Stat 34:122–148

Anderson TW (1984) An introduction to multivariate statistical analysis, 2nd edn. John Wiley, New York

Bertrand P and Goupil F (2000) Descriptive statistics for symbolic data. In: Bock H-H, Diday E (eds) Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer, Berlin, pp 103–124

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

Billard L (2011) Brief overview of symbolic data and analytic issues. Stat Anal Data Min 4:149–156

Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98:470–487

Billard L, Diday E (2006) Symbolic data analysis: conceptual statistics and data mining. John Wiley, Chichester

Billard L, Guo JH, Xu W (2011) Maximum Likelihood Estimators for Bivariate Interval-Valued Data. Technical Report, University of Georgia, Athens, GA, under revision

Billard L, Le-Rademacher J (2013) Symbolic principal components for interval-valued data. Revue des Nouvelles Technologies de l'Information 25:31–40

Bock HH, Diday E (2000) Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer, Berlin

Cazes P (2002) Analyse Factorielle d'un Tableau de Lois de Probabilité. Rev Stat Appl 50:5–24

Cazes P, Chouakria A, Diday E, Schecktman Y (1997) Extension de l'analyse en composantes principales à des données de type intervalle. Rev Stat Appl 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, Billard L, Diday E (2011) Principal component analysis for interval-valued observations. Stat Anal Data Min 4:229–246

Ichino M (2011) The quantile method for symbolic principal component analysis. Stat Anal Data Min 4:184–198

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

Johnson RA, Wichern DW (2002) Applied multivariate statistical analysis, 5th edn. Prentice Hall, New Jersey

Jolliffe IT (2004) Principal component analysis, 2nd edn. Springer, New York

Lauro NC, Palumbo F (2000) Principal component analysis of interval data: a symbolic data analysis approach. Comput Stat 15:73–87

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

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

Le-Rademacher J, Billard L (2011) Likelihood functions and some maximum likelihood estimators for symbolic data. J Stat Plan Inference 141:1593–1602

Le-Rademacher J, Billard L (2012) Symbolic-covariance principal component analysis and visualization for interval-valued data. J Comput Graph Stat 21:413–432

Le-Rademacher J, Billard L (2013) Principal component histograms from interval-valued observations. Comput Stat 28:2117–2138

Makosso-Kallyth S, Diday E (2012) Adaptation of interval PCA to symbolic histogram variables. Adv Data Anal Classif 6:147–159

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, New York

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

Shapiro AF (2009) Fuzzy random variables. Insur Math Econ 44:307–314

Xu W (2010) Symbolic Data Analysis: Interval-Valued Data Regression. PhD thesis, University of Georgia

Zadeh LA (1965) Fuzzy Sets. Inf Control 8:338–353

Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 23:421–427