Metodološki zvezki, Vol. 11, No. 1, 2014, 1-20

Symbolic Covariance Matrix for Interval-valued Variables and its Application to Principal Component Analysis: a Case Study

Katarina Košmelj[1], Jennifer Le-Rademacher[2] and Lynne Billard[3]

Abstract

In the last two decades, principal component analysis (PCA) was extended to interval-valued data; several adaptations of the classical approach are known from the literature. Our approach is based on the symbolic covariance matrix Cov for the interval-valued variables proposed by Billard (2008). Its crucial advantage, when compared to other approaches, is that it fully utilizes all the information in the data. The symbolic covariance matrix can be decomposed into a within part $CovW$ and a between part $CovB$. We propose a further insight into the PCA results: the proportion of variance explained due to the within information and the proportion of variance explained due to the between information can be calculated. Additionally, we suggest PCA on $CovB$ and $CovW$ to be done to obtain deeper insights into the data under study.

In the case study presented, the information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is about 45%. It turns out that, for these data, the uniformity assumption over intervals does not hold and so analysis of the data represented by histogram-valued variables is suggested.

1 Introduction

1.1 Principal component analysis for classical data

Principal component analysis (PCA) was first described by Pearson (1901) as an analogue of the principal axes theorem in mechanics; it was later independently developed and named by Harold Hotelling in the 1930s. It is a very popular exploratory tool in classical multivariate data (see e.g., Chatfield and Collins, 1980; Johnson and Wichern, 2002). Its major objective is to reduce the dimension of the variable space: the original p random variables $\mathbf{X} = (X_1, X_2, ..., X_p)$ are transformed into s random variables $\mathbf{Y} = (Y_1, Y_2, ..., Y_s)$, called Principal Components, where $s \ll p$, and the $\mathbf{Y}$ variables are uncorrelated. This transformation is defined in such a way that the first principal component (PC_1) accounts for as much of the variability, i.e., variance, in the data as

[1] Biotechnical Faculty, University of Ljubljana, Slovenia; katarina.kosmelj@bf.uni-lj.si
[2] Medical College of Wisconsin, Milwaukee, USA; jlerade@mcw.edu
[3] University of Georgia, Athens, USA; lynne@stat.uga.edu

possible, and each succeeding component in turn has the highest variance possible, under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components.

The solution of the problem described above is given by the eigenvalues and eigenvectors of the covariance matrix of $X_1, X_2, ..., X_p$. Principal components are linear combinations of the original variables, defined by the eigenvectors of this covariance matrix. From the basic linear algebra it follows: there are p eigenvalues ordered: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$; eigenvalues of the covariance matrix are the variances of the principal components. The eigenvalues add up to the sum of the diagonal elements, i.e., to the trace of the covariance matrix. This means that the sum of the variances of the principal components is equal to the sum of the variances of the original variables. The i-th principal component accounts for $\lambda_i / \sum_{j=1}^{p} \lambda_j$ of the total variance in the original data. When the decision on the reduced dimension s is taken, we calculate the proportion of variance accounted for by the first s principal components, $\sum_{j=1}^{s} \lambda_j / \sum_{j=1}^{p} \lambda_j$.

As the covariance on standardized variables equals the correlation, therefore, in this case, eigenvalues and eigenvectors of the correlation matrix are used. It is recommended to perform PCA on standardized variables when the original variables are measured on scales with different ranges.

1.2 Principal component analysis for symbolic data

In the second part of the 20^{th} century, the need to analyze massive datasets emerged. Symbolic data analysis started as a response to that demand; see Bock and Diday (2000), Billard and Diday (2003, 2006), among others. Symbolic analytical methods are often generalizations of their classical approach counterparts. A symbolic method should give the same results as its classical counterpart when applied to classical data (Billard, 2011, Le-Rademacher and Billard, 2012).

In the last two decades, PCA was adapted for symbolic data, first in the context of interval-valued data. A number of approaches were proposed. Le-Rademacher and Billard (2012) give a short overview of its historical development; let us review them briefly. Cazes et al. (1997) proposed the first adaptations of PCA known as the *centers method* and the *vertices method*, see also Douzal-Chouakria et al. (2011); Zuccolotto (2007) applied the vertices method to a dataset on job satisfaction; Lauro and Palumbo (2000) introduced a Boolean matrix to account for the interdependency of the vertices, Palumbo and Lauro (2003) and Lauro et al. (2008) proposed the *midpoint-radii* method treating interval midpoints and interval midranges as two separate variables; Gioia and Lauro (2006) proposed a PCA version based on an interval algebra approach.

Le-Rademacher and Billard (2012) describe these approaches in detail and discuss their characteristics in the context of symbolic data analysis: namely, these approaches fail in different ways to utilize the entire information included in the interval-valued data.

These deficiencies can be avoided when the symbolic covariance matrix Cov is used. Its calculation in the interval setting was first presented in Billard (2008). The crucial advantage of this symbolic covariance matrix is that it fully utilizes all the information in the data; also it is shown that the symbolic covariance matrix can be decomposed into a within part $CovW$ and a between part $CovB$. Two papers on this topic (Le-Rademacher and Billard, 2012 and Billard and Le-Rademacher, 2012) also provide a new approach to constructing the observations in PC space allowing for a better visualization of the

results. Le-Rademacher and Billard (2013a) propose an approach to construct histogram values from the principal components of interval-valued observations. Le-Rademacher (2008) and and Le-Rademacher and Billard (2013b) extend these ideas to histogram-valued observations. In a different direction, Giordani and Kiers (2006) consider fuzzy data, which is a different domain from symbolic data and so is outside the purview of the present work. Likewise, a different domain is the PCA of time series data of Irpino (2006).

1.3 Objective of this study

We want to compare PCA results obtained on different data types. To enable the comparison of the results, the data were aggregated from the same dataset. For each observation and each variable, we aggregated the data in two different ways:

- the mean value;
- the $[min, max]$ interval which is based on the minimal and maximal value under observation.

The main objective of this study is to find out what is the information gain when analyzing the $[min, max]$ interval instead of the mean value.

In the next section, some well known characteristics of interval-valued data are summarized. Covariance in the interval setting will be illustrated and compared to the covariance in the classical setting. For PCA on interval-valued variables, a simple measure of the information gain will be introduced and additional PCA analyses will be suggested. These approaches allow for a deeper insight into the dataset under study.

The third section presents a simple case study. It consists of seven meteorological stations in Slovenia, they are described by seven variables, the data are from the 40 year-period 1971-2010. The results of different PCA analyses will be presented and compared. To facilitate the comparison of the results, the dataset is very small, however, the stations are chosen according to subject-matter knowledge. The last section gives some conclusions and suggestions for further work.

2 Interval-valued variables

Let us first note that an interval-valued random variable is just a standard random variable but its values are intervals. Let $\mathbf{X} = (X_1, X_2, ..., X_p)$ be a p-dimensional random variable taking values in R^p. Let X_j be an interval-valued random variable, its data exist for a random sample of size n and is in the form $X_{ij} = [a_{ij}, b_{ij}]$, $a_{ij} \leq b_{ij}$, $i = 1, ..., n$. In the case $a_{ij} = b_{ij}$, for any $i = 1, 2, ..., n$, X_{ij} has a classical value. Each observation described by a p-dimensional interval-valued variable can be visualized as a hypercube in R^p.

2.1 Mean and variance

The mean and the variance for an interval-valued variable are based on the assumption that the distribution of the values within each interval is uniform. They were first defined by Bertrand and Goupil (2000). The sample variance of X_j is:

$$S_j^2 = \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2 + a_{ij}b_{ij} + b_{ij}^2) - \overline{X}_j^2, \tag{2.1}$$

where the sample mean $\overline{X}_j$ is the average of the interval midpoints

$$\overline{X}_j = \frac{1}{2n}\sum_{i=1}^{n}(a_{ij} + b_{ij}). \tag{2.2}$$

Billard (2008) showed that (2.1) can be rewritten as

$$S_j^2 = \frac{1}{3n}\sum_{i=1}^{n}[(a_{ij} - \overline{X}_j)^2 + (a_{ij} - \overline{X}_j)(b_{ij} - \overline{X}_j) + (b_{ij} - \overline{X}_j)^2], \tag{2.3}$$

and proved that the Total Sum of Squares SST can be decomposed into a within part SSW and a between part SSB :

$$nS_j^2 \ = SST_j \ = SSW_j \ + SSB_j. \tag{2.4}$$

The Within Sum of Squares SSW measures the internal variation and can be expressed as follows:

$$\begin{aligned} SSW_j = \ & \frac{1}{3}\sum_{i=1}^{n}[(a_{ij} - \frac{a_{ij}+b_{ij}}{2})^2 + (a_{ij} - \frac{a_{ij}+b_{ij}}{2})(b_{ij} - \frac{a_{ij}+b_{ij}}{2}) + (b_{ij} - \frac{a_{ij}+b_{ij}}{2})^2] \\ = \ & \sum_{i=1}^{n}\frac{(b_{ij}-a_{ij})^2}{12}. \end{aligned} \tag{2.5}$$

Thus, as expected, SSW is based on an implicit assumption that the distribution of values within each observed interval is uniform, $X_{ij} \tilde{} \ U(a_{ij}, b_{ij})$, $i = 1, 2, ..., n$. Other distributions are also relevant; e.g., Billard (2008) presents the formulae for SSW and SST when observations within each interval follow a triangular distribution.

The Between Sum of Squares SSB describes the between variation, i.e., the variation of the interval midpoints:

$$SSB_j = \sum_{i=1}^{n}(\frac{a_{ij}+b_{ij}}{2} - \overline{X}_j)^2, \tag{2.6}$$

and is independent of the distribution within the intervals.

2.2 Covariance

Let X_{j_1} and X_{j_2} be two interval-valued random variables with pairwise observations: $X_{j_1} = [a_{ij_1}, b_{ij_1}]$ and $X_{j_2} = [a_{ij_2}, b_{ij_2}]$ on a random sample of size n. The following holds: $a_{ij} \leq b_{ij}$, for $j = j_1, j_2$, and $i = 1, 2, ..., n$. Total Sum of Products SPT is decomposed into two components, the Sum of Products Within, SPW, and the Sum of Products Between, SPB; it is connected to the covariance Cov:

$$nCov_{j_1j_2} = SPT_{j_1j_2} = SPW_{j_1j_2} + SPB_{j_1j_2}. \tag{2.7}$$

The Sum of Products Within SPW and Sum of Products Between SPB are related to $CovW$ and $CovB$, respectively, which are expressed as follows:

$$CovW_{j_1j_2} = \frac{SSW_{j_1j_2}}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{(b_{ij_1} - a_{ij_1})(b_{ij_2} - a_{ij_2})}{12}, \tag{2.8}$$

$$CovB_{j_1j_2} = \frac{SSB_{j_1j_2}}{n} = \frac{1}{n}\sum_{i=1}^{n}(\frac{a_{ij_1} + b_{ij_1}}{2} - \overline{X}_{j_1})(\frac{a_{ij_2} + b_{ij_2}}{2} - \overline{X}_{j_2}). \tag{2.9}$$

It may be interesting to notice that the entries of the $CovW$ matrix are always positive, their magnitudes depend on the ranges, $R_{ij} = b_{ij} - a_{ij}$, $j = j_1, j_2$; the greater the ranges of the two variables the greater is the entry of $CovW$. It should be pointed out that $CovW$ is not a true covariance matrix on the ranges; the terms for the true covariance matrix on the ranges would be $(R_{ij_1} - \overline{R}_{j_1})(R_{ij_2} - \overline{R}_{j_2})$. However, the $CovW$ matrix incorporates information on the size of the rectangles.

The entries of $CovB$ are classical covariances (divided by n not by $n - 1$) on the interval midpoints. When, instead of the intervals $[a, b]$, PCA is performed on the interval midpoints: $[(a + b)/2, (a + b)/2]$, $CovW$ is zero and $Cov = CovB$; in this case, the symbolic PCA results are the same as for a classical PCA on the interval midpoints.

Billard (2008) showed that the covariance between two interval-valued variables X_{j_1} and X_{j_2} can be calculated directly, using the following expression:

$$\begin{aligned} Cov_{j_1j_2} = \frac{1}{6n}\sum_{i=1}^{n}[2\,(a_{ij_1} - \overline{X}_{j_1})(a_{ij_2} - \overline{X}_{j_2}) + (a_{ij_1} - \overline{X}_{j_1})(b_{ij_2} - \overline{X}_{j_2}) \\ + (b_{ij_1} - \overline{X}_{j_1})(a_{ij_2} - \overline{X}_{j_2}) + 2\,(b_{ij_1} - \overline{X}_{j_1})(b_{ij_2} - \overline{X}_{j_2})] \end{aligned} \tag{2.10}$$

Two special cases are easily checked: a) covariance of two identical variables equals its variance; b) covariance of two classical variables equals the well known classical covariance.

Figure 1 gives some insight into the calculation of the covariance in the classical and interval setting. Covariance in the classical setting is based on the position of the points, in the interval setting it is based on the rectangles: the location of the midpoints determines the between part, the size of the rectangles determines the within part, which

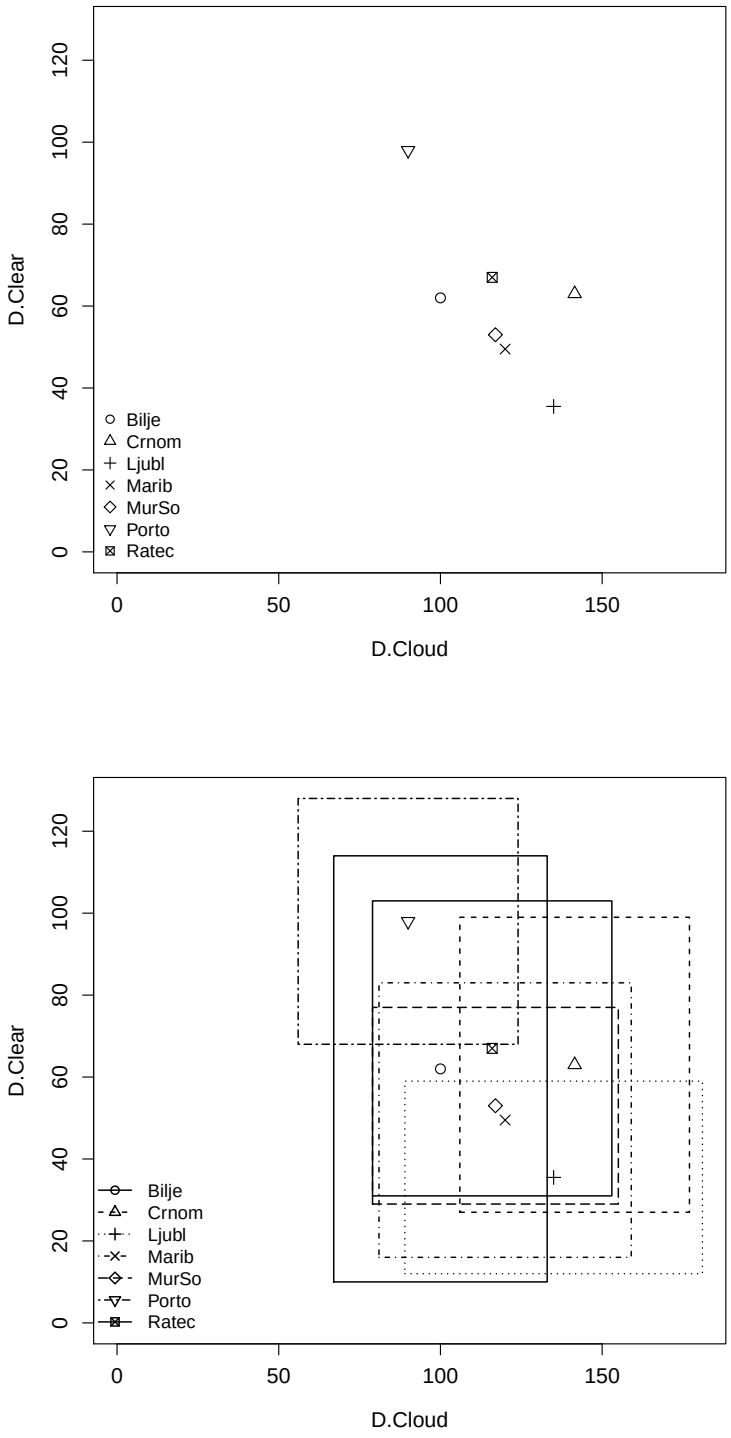

Figure 1: Calculation of the covariance in the classical setting (upper part) is based on the position of the points; in the interval setting (lower part) it is based on the rectangles: position of the midpoints and size of the rectangles determine its value.

is always positive. Covariance is calculated on the number of clear days (D.Clear) and the number of cloudy days (D.Cloud) for seven meteorological stations (for details, see next section). Figure 1 (upper part) illustrates the classical covariance, the position of the points suggests that the covariance is negative, the same would be expected from the subject-matter knowledge; the obtained value is $Cov(D.Cloud, D.Clear) = -212.0$. However, the covariance in the interval setting is positive, $Cov(D.Cloud, D.Clear) = +202.2$. This is due to a large within interval component $CovW(D.Cloud, D.Clear) = 411.7$, the between component is the same as classical covariance on the midpoints, $CovB(D.Cloud, D.Clear) = -212.0$; see Figure 1 (lower part).

2.3 Principal component analysis in the context of interval-valued data

A crucial advantage of the symbolic covariance matrix Cov is that it fully utilizes all the information in the data. It can be decomposed into a within part $CovW$ and a between part $CovB$. This decomposition allows for a deeper insight into the PCA results from the traces of these matrices. Since the trace of a matrix is a linear operator, the following holds:

$$tr(Cov) = tr(CovW) + tr(CovB). \tag{2.11}$$

Hence, we can assess the proportion of variance explained due to the within information and the proportion of variance explained due to the between information. The information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is due to the within information.

Additional PCA analysis can be done on $CovB$, these results are equivalent to the classical PCA results on the interval midpoints. A PCA analysis can also be performed on $CovW$; the interpretation of these results may enlighten some of the aspects of the within information.

3 A case study

We consider yearly data from the period 1971-2010 in Slovenia, data were collected by Slovenian Environment Agency (http://meteo.arso.gov.si/met/sl/archive/), and are shown in the Appendix. The following variables are taken into account: number of cold days (D.Cold), number of warm days (D.Warm), number of days with storms (D.Storm), number of days with precipitations (D.Prec), number of days with snow cover (D.SnCov), number of clear days (D.Clear), and number of cloudy days (D.Cloud). According to meteorological definitions, for a cold day the minimal daily air temperature is below $0\ ^0$C, for a warm day the maximal daily temperature is above $25\ ^0$C; a clear day has under 20% of cloudiness, a cloudy day has over 80%. Hence, D.Cold and D.Warm are based on the same variable, i.e., air temperature, the same holds for D.Clear and D.Cloud which are based on cloudiness.

For illustrative simplicity, only seven meteorological stations are chosen for this case study. They are: Bilje (Bilje), Črnomelj (Crnom), Ljubljana (Ljubl), Maribor (Marib),

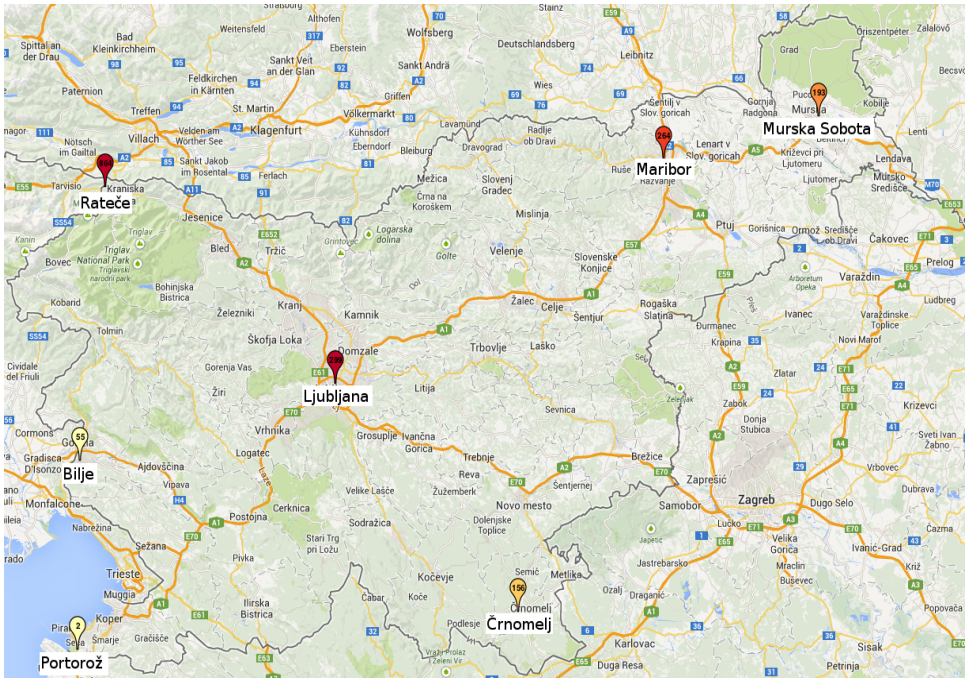

Figure 2: Geographical position of seven meteorological stations under study: Bilje (Bilje), Črnomelj (Crnom), Ljubljana (Ljubl), Maribor (Marib), Murska Sobota (MurSo), Portorož-airport (Porto), and Rateče (Ratec); elevation (in meters) is pinned to each station.

Murska Sobota (MurSo), Portorož-airport (Porto), and Rateče (Ratec). Their location is shown in Figure 2. Portorož-airport is situated at sea level (elevation 2 m), Rateče is in the Alps (elevation 864 m), the other stations have elevation from 55 m to 299 m. The dataset is slightly incomplete: data for Portorož-airport started in 1975, for Bilje, Črnomelj, Maribor and Murska Sobota data for some years are inconsistently missing.

As already stated, we want to compare PCA results obtained on different data types which were aggregated from the same dataset. For each station and each variable, we aggregated the data in two different ways: the mean value and the $[min, max]$ interval which is based on the minimal and maximal values in the period under observation.

3.1 PCA on the Means

In Table 1, the classical covariance matrix calculated on the means is presented; the sum of variances (3891.8) is given below the matrix. Dominant variances are as follows: $Var(D.SnCov) = 1449.4$, $Var(D.Cold) = 1284.6$; dominant covariances are: $Cov(D.SnCov, D.Cold) = 1232.5$ (positive), $Cov(D.SnCov, D.Warm) = -679.3$ and $Cov(D.Warm, D.Cold) = -558.2$ (negative).

In Table 2, the PCA results are given. The first two principal components explain about 92% of total variance, the first three around 97%. The loads for the first three principal components are also presented; we shall interpret the first two principal components only. For the first principal component (PC_1) D.Cold and D.SnCov are dominant, for the second principal component (PC_2) D.Clear and D.Cloud show up. We can deduce that PC_1 is positively correlated with low air temperature and PC_2 with the surplus of cloudy

Table 1: Covariance matrix calculated on the means. The sum of the variances (in the table in bold) is given below the matrix.

	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**1284.6**	-558.2	-182.0	308.1	1232.5	-262.3	204.8
D.Warm	-558.2	**356.7**	45.1	-96.1	-679.3	70.1	-28.3
D.Storm	-182.0	45.1	**47.3**	-29.3	-109.4	51.4	-29.5
D.Prec	308.1	-96.1	-29.3	**212.9**	337.8	-148.6	195.5
D.SnCov	1232.5	-679.3	-109.4	337.8	**1449.4**	-216.3	177.6
D.Clear	-262.3	70.1	51.4	-148.6	-216.3	**296.8**	-212.0
D.Cloud	204.8	-28.3	-29.5	195.5	177.6	-212.0	**244.1**

Sum of variances = **3891.8**

Table 2: PCA on the means, results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

	PC_1	PC_2	PC_3
Cum.% of var. exp.	79.2	92.2	96.7
D.Cold	**0.625**	0.014	**0.661**
D.Warm	-0.307	0.274	0.192
D.Storm	-0.071	-0.057	-0.391
D.Prec	0.172	0.385	-0.322
D.SnCov	**0.670**	-0.218	-0.473
D.Clear	-0.138	**-0.598**	-0.090
D.Cloud	0.113	**0.607**	-0.193

over clear days.

Figure 3 presents the seven stations in the space of PC_1 by PC_2. There is a positive trend with low air temperature along PC_1: Portorož-airport reveals few days with low air temperature and snow cover, Rateče the opposite. This is consistent with the fact that Portorož-airport is located near the Adriatic sea, Rateče is located in the Alps. There is a positive trend in the surplus of cloudy over clear days along PC_2; here, Portorož-airport has the lowest surplus (it has more clear than cloudy days), Ljubljana and Črnomelj have the highest (here, there are more cloudy than clear days).

3.2 Symbolic PCA on interval-valued variables

3.2.1 Symbolic covariance matrix and its decomposition

The symbolic covariance matrix Cov for the intervals is given in Table 3; also shown is the decomposition into $CovB$ and $CovW$. The term $CovB$ is identical to the classical covariance matrix on the interval midpoints. Values of $CovW$ reflect the internal variability and are all positive. Consequently, the terms in Cov are always larger than the corresponding terms in $CovB$; thus, there are fewer negative terms in Cov than in $CovB$.

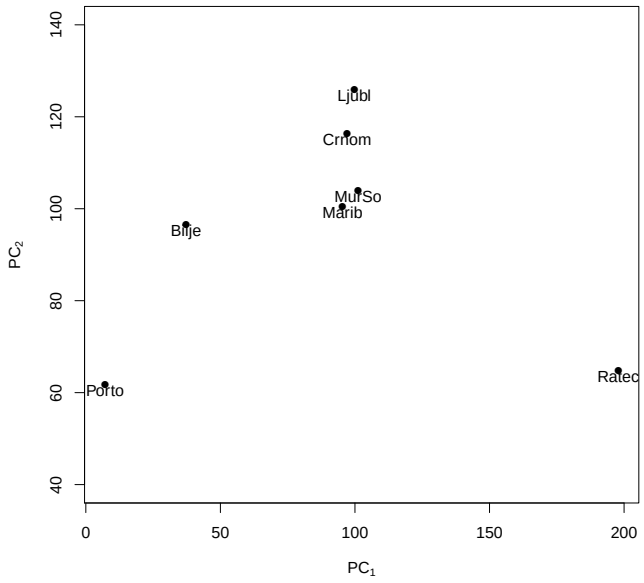

Figure 3: PCA on the means: presentation of seven stations in two-dimensional space of PC_1 by PC_2; 92.2% of total variance is explained. PC_1 reflects positive impact of low air temperature, PC_2 reflects positive impact of surplus of cloudy over clear days.

The sum of symbolic variances is 5754.2, the between component explains 3170 (55.1%), and the remaining 2563.2 (44.9%) is due to the within component. In this case, we can conclude that the gain in information, when we analyze the intervals instead of the interval midpoints, is large, it is nearly 45%. Let us find out the corresponding impact on the PCA results.

3.2.2 PCA on symbolic covariance matrix

Table 4 shows the PCA results based on the symbolic covariance matrix. The first two principal components explain 86.4% of variance, the first three 95.1%. For PC_1, the loads for D.Cold and D.SnCov are dominant, for PC_2 the dominant loads are D.Warm and D.Clear (positive), for PC_3 D.Clear (negative). Hence, the PC_1 is positively correlated with low air temperature, as in the PCA on the means; however, other results are different: PC_2 is positively correlated with D.Warm and D. Clear, PC_3 is negatively correlated with D.Clear.

Visualisation of these PCA results in two-dimensional space is based on the approach presented in Le-Rademacher and Billard (2012). For each station, a 7-dimensional polytope is obtained. Figure 4 (upper plot) presents the projection of these polytopes onto the PC_1 by PC_2 plane. Considerable overlapping is presented. The plot shows that the variability in PC_1 (D.Cold and D.SnCov) is dominant, for Rateče it is the greatest; however, variability in PC_2 (D.Warm and D.Clear) is comparable for all stations. Only two pairs of stations do not overlap: Portorož-airport and Rateče, Bilje and Rateče. The polytopes for two extreme stations, Portorož-airport and Rateče, are presented on Figure 4 (lower plot).

Table 3: Covariance matrix Cov for interval-valued variables, variances are in bold; it is decomposed into $CovB$ and $CovW$ (below). The respective sum of variances is presented below the corresponding matrix.

Cov	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**1363.7**	-74.1	30.8	537.4	1313.6	105.0	630.2
D.Warm	-74.1	**720.3**	224.4	262.1	-53.4	399.3	407.2
D.Storm	30.8	224.4	**153.9**	146.1	127.6	265.9	191.4
D.Prec	537.4	262.1	146.1	**443.2**	615.6	163.6	557.6
D.SnCov	1313.6	-53.4	127.6	615.6	**1595.7**	187.5	750.6
D.Clear	105.0	399.3	265.9	163.6	187.5	**724.6**	202.2
D.Cloud	630.2	407.2	191.4	557.6	750.6	202.2	**752.9**

Sum of variances = **5754.2**

$CovB$	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**1056.0**	-435.8	-144.9	258.0	941.1	-233.8	251.5
D.Warm	-435.8	**286.8**	19.2	-67.0	-482.5	8.8	-39.2
D.Storm	-144.9	19.2	**46.3**	-18.4	-68.8	60.1	-24.8
D.Prec	258.0	-67.0	-18.4	**183.1**	282.0	-148.7	209.8
D.SnCov	941.1	-482.5	-68.8	282.0	**997.7**	-196.6	270.7
D.Clear	-233.8	8.8	60.1	-148.7	-196.6	**322.3**	-209.5
D.Cloud	251.5	-39.2	-24.8	209.8	270.7	-209.5	**278.9**

Sum of between variances = **3171.0**

$CovW$	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**307.7**	361.7	175.8	279.4	372.5	338.8	378.7
D.Warm	361.7	**433.5**	205.2	329.1	429.2	390.5	446.5
D.Storm	175.8	205.2	**107.6**	164.5	196.3	205.9	216.1
D.Prec	279.4	329.1	164.5	**260.2**	333.6	312.2	347.8
D.SnCov	372.5	429.2	196.3	333.6	**598.0**	384.1	479.9
D.Clear	338.8	390.5	205.9	312.2	384.1	**402.2**	411.7
D.Cloud	378.7	446.5	216.1	347.8	479.9	411.7	**474.1**

Sum of within variances = **2583.2**

Table 4: PCA on the intervals, results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

	PC_1	PC_2	PC_3
Cum.% of var. exp.	62.0	86.4	95.1
D.Cold	**0.569**	-0.277	-0.102
D.Warm	0.081	**0.663**	0.279
D.Storm	0.079	0.261	-0.119
D.Prec	0.309	0.172	0.271
D.SnCov	**0.636**	-0.220	-0.180
D.Clear	0.127	**0.512**	**-0.767**
D.Cloud	0.384	0.275	0.452

From these plots, it is observed that the internal variability for Rateče is greater than it is for Portorož-airport.

3.2.3 PCA on $CovB$ and $CovW$

We proceed with PCA on $CovB$, this is identical to the classical PCA on the interval midpoints, the results are in Table 5 (left) and are plotted in Figure 5 (upper plot); they are consistent with the PCA results on the means.

Since $CovW$ depicts the within interval information, PCA on $CovW$ allows an insight into the variability within the interval variables, see Table 5 (right) and Figure 5 (lower plot). In this case, the PC_1 explains 93.2%, the PC_2 explains additional 5.4%. The loads for PC_1 for all variables have similar magnitude, while for PC_2 the dominant load is D.SnCov; accordingly, PC_1 is positively related to all the variables, PC_2 is positively related to D.SnCov. The scores are calculated using the midpoints. The stations are located along the diagonal, from Portorož-airport at the lower end to Rateče at the upper end, revealing the increase of interval variability from the lower to the upper end. This result is consistent with the fact that Portorož-airport has tighter intervals, Rateče has larger intervals.

3.2.4 Programs used

Algorithms for deriving the PCA results on the symbolic covariance matrix along with the corresponding polytops are available at Le-Rademacher and Billard (2012, Supplementary material - online version). Their R script (R Core Team, 2013) was upgraded with PCA on $CovW$ and $CovB$ and adapted for our case-study.

3.3 Other PCA approaches for interval-valued variables

Other PCA approaches on interval data are described in the literature. As stated before, Le-Rademacher and Billard (2012) give a detailed insight into these methods. We shall limit ourselves to only some of them.

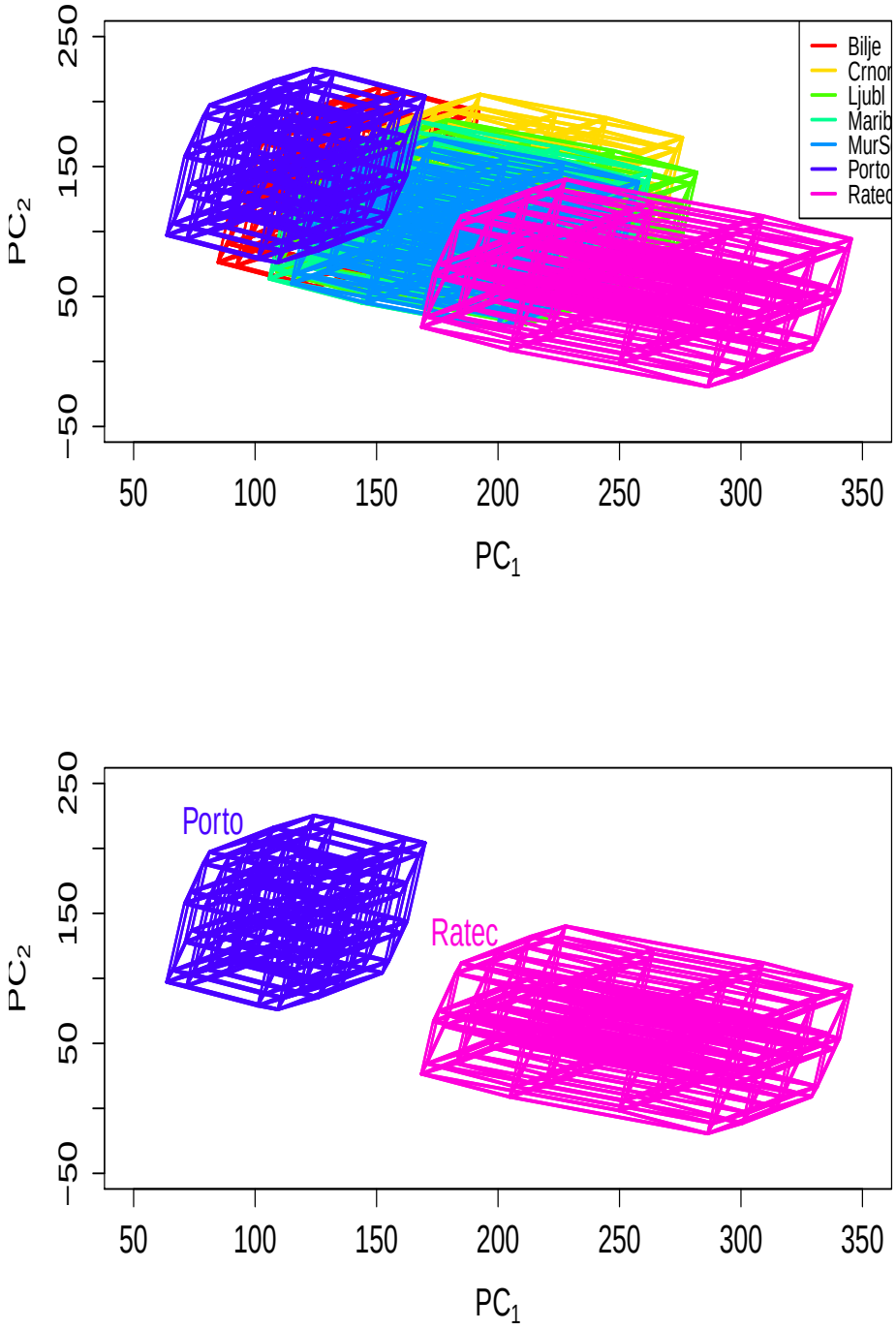

Figure 4: Projection of 7-dimensional polytopes onto 2-dimensional space of PC_1 by PC_2 , upper plot: for all 7 stations; lower plot: for Portorož-airport and Rateče. PC_1 explains 62.0% of variance, it reflects the positive impact of D.Cold and D.SnCov; PC_2 explains 24.4% of variance, it reflects the positive impact of D.Warm and D.Clear.

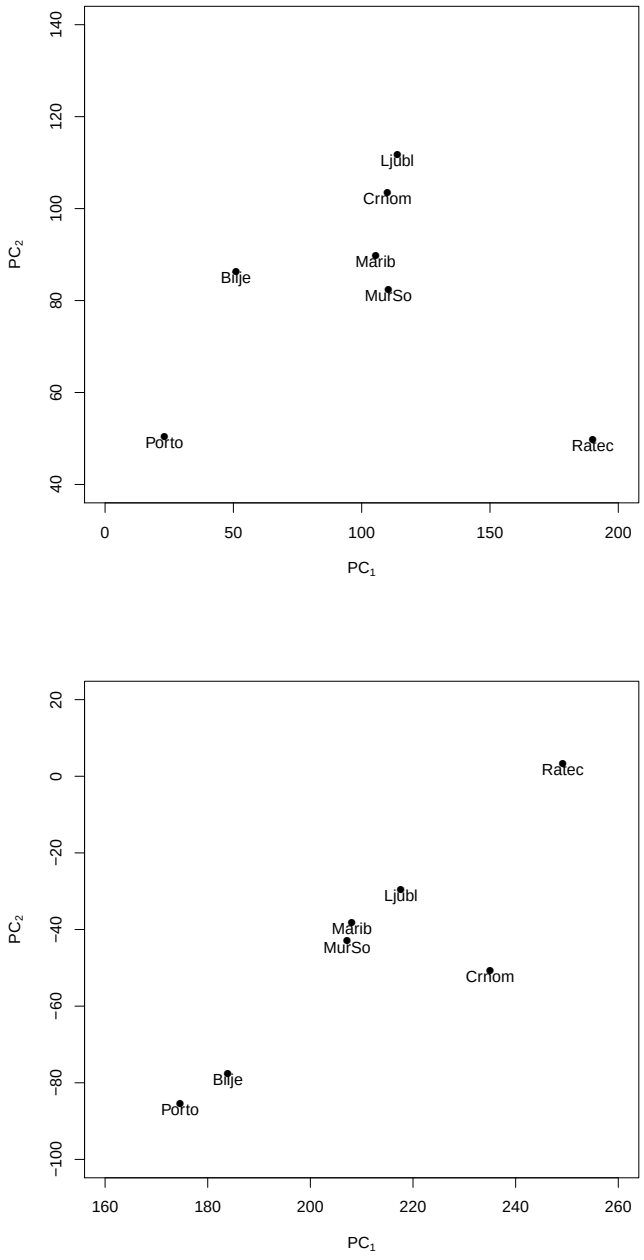

Figure 5: Upper plot: PCA on CovB (this is identical to classical PCA on midpoints); PC_1 reflects the positive impact of D.Cold and D. Sncov; PC_2 reflects the surplus of D.Cloud over D.Clear. Lower plot: PCA on CovW: PC_1 reflects the positive impact of all variables; PC_2 the positive impact of D.SnCov.

Table 5: PCA on the CovB (left), PCA on CovW (right); results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

	PC_1	PC_2	PC_3	PC_1	PC_2	PC_3
Cum.% of var. exp.	75.9	91.5	96.8	93.2	98.6	99.7
D.Cold	**0.642**	-0.118	**0.508**	0.355	-0.126	0.118
D.Warm	-0.286	0.376	0.175	0.416	-0.189	**0.603**
D.Thund	-0.068	-0.050	-0.423	0.203	-0.221	-0.194
D.Prec	0.193	0.349	-0.348	0.324	-0.172	0.049
D.SnCov	**0.630**	-0.161	-0.363	0.454	**0.847**	-0.207
D.Clear	-0.168	**-0.627**	-0.380	0.390	-0.391	**-0.700**
D.Cloud	0.198	**0.549**	-0.368	0.443	-0.018	0.221

3.3.1 Centers method

The centers method transforms the interval-valued matrix into a classical matrix of the interval midpoints. The results of this method are given as a part of the PCA approach on symbolic covariance matrix: see $CovB$ in Table 3, PCA results in Table 5 (left) and Figure 5 (upper plot). As already stated, in this approach the internal interval variance is completely ignored.

3.3.2 Vertices method

In this approach, the vertices of the hyper-rectangles (instead of the interval midpoints) are considered as the data-input. In our case, seven variables were taken into account; therefore, there are $2^7 = 128$ vertices. Thus, the dimension of the input matrix is $n = 128$, $p = 7$; classical PCA is performed on this matrix.

Here, we do not present the covariance matrix. The sum of variances equals 10932.8, which is approximately twice the value in the symbolic context (5754.2). Table 6 presents the PCA results for the first three principal components. In this case, PC_1 explains *only* 34.8% of the total variance; the first two principal components 51.3% and the first three 64.7%. For PC_1, D.Cold and D.SnCov are dominant; PC_2 is positively correlated with D.Cloud and negatively with D.Clear (as in the PCA on the means or midpoints); for PC_3 D.Warm and D.SnCov show up, surprisingly, both loads are positive.

We can summarize, that this approach is simple, it always works, but it fails to use all the variation in the data. The results reflect that the data matrix is artificially inflated; the vertices are treated as independent observations, this assumption is not sustainable. Our results are consistent with Douzal-Chouakria et al. (2011), who showed that the variance of the vertices in fact includes some but not all of the internal variation.

3.3.3 The midpoint-radii method

The *midpoint-radii* approach treats a single interval-valued variable as two variables: midpoints and midranges. A PCA can be performed on either of them. This is similar to the

Table 6: PCA on the vertices; results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

	PC_1	PC_2	PC_3
Cum.% of var. exp.	34.8	51.3	64.7
D.Cold	**0.525**	0.037	-0.460
D.Warm	-0.276	0.284	**0.660**
D.Thund	-0.044	-0.024	0.041
D.Prec	0.151	0.181	-0.012
D.SnCov	**0.750**	-0.256	**0.579**
D.Clear	-0.149	**-0.514**	0.111
D.Cloud	0.196	**0.745**	0.050

PCA on $CovB$ and $CovW$; the only difference is, that $CovW$ is an uncentered covariance matrix on the ranges.

To analyze the midpoint and the range data simultaneously, Palumbo and Lauro (2003) propose to superimpose the PCs of the midrange onto the PCs on the midpoint and then rotate the midrange PC axes to maximize the connection between the midpoints and the midranges. It turns out the choice of rotation operator is subjective; the midpoints and the midranges are treated as independent (see Lauro et al., 2008). Le-Rademacher and Billard (2012) showed that the midpoint-radii approach is deficient and not working properly. Due to these facts, we believe that this approach should be replaced by the PCA on the symbolic covariance matrix; see the results given in Table 3 above, Table 4 and Figure 4.

4 Conclusions

A crucial advantage of the symbolic covariance matrix Cov is that it fully utilizes all the information in the data. It can be decomposed into a within part $CovW$ and a between part $CovB$. In the interpretation of the Cov term, we should recognise that: it is the sum of the classical covariance on the interval midpoints and a measure of variability (i.e., the size) of the intervals. Therefore, the sign of $CovB$ may be negative and the sign of Cov positive. Figure 1 illustrates such a case.

However, this decomposition allows for a deeper insight into the interval-valued dataset: from the traces of these matrices, the proportion of variance explained due to the within information and the proportion of variance explained due to the between information can be calculated. The information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is due to the within information.

We can summarize the PCA approach on Cov as follows: the interpretation of the PC should be the "symbolic context"; visualization of the results using the projection of the polytopes is suitable for lower dimensions of p and n, for higher dimensions the plot can be unreadable. We suggest that separate PCA's on both the $CovB$ and the $CovW$ should be done additionally to allow for a deeper understanding of the between and within information. The analysis of PCA results on $CovB$ is straightforward, as in the classical context on the interval midpoints. However, the PCA results on $CovW$ are interpretable

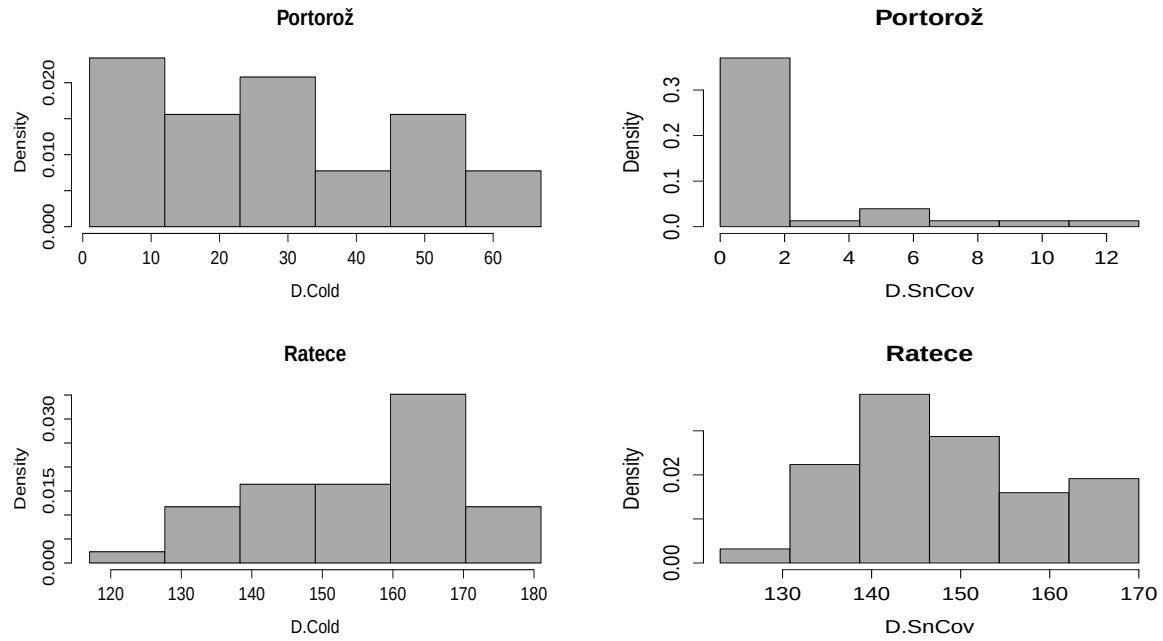

Figure 6: Histograms for D.Cold and D.SnCov for Portorož-airport and Rateče revealing different types of distribution.

in the context of the size of the rectangles.

In the case study presented, the information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is about 45%. For the PCA results on Cov, it may be difficult to grasp the meaning of the PC_2; however, the PCA results obtained on $CovB$ and $CovW$ are consistent with the subject-matter knowledge.

There is an important assumption hidden in this analysis: the distribution of the values along each $[min, max]$ interval should be uniform. This is often not the case, in particular when data for meteorological variables over a longer period are under study; for illustration, see some histograms of the raw data used herein in Figure 6. It is obvious that the uniformity assumption does not hold. Therefore, it may be interesting to analyze the histogram-valued variables and compare the results with the results obtained on the interval-valued variables.

References

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

[2] Billard L. and Diday E. (2003): From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of American Statistical Association*, **98**, 470-487.

[3] Billard L. and Diday E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley Series in Computational Statistics.

[4] Billard L. (2008): Sample Covariance Functions for Complex Quantitative Data. In Mizuta M. and Nakano J. (Ed): *Proceedings of the International Association of Statistical Computing Conference 2008*, 157-163. Yokohama.

[5] Billard L. (2011): Brief overview of symbolic data and analytical issues. *Statistical Analysis and Data Mining*, **4**, 149-156.

[6] Billard L. and Le-Rademacher J. (2012): Principal component analysis for interval data. *WIREs Comput Stat 2012*, 4:535-540. doi: 10.1002/wics.1231.

[7] Bock, H.-H. and Diday, E. (eds.) (2000): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information form Complex Data*. Berlin: Springer-Verlag.

[8] Cazes, P., Chouakria, A., Diday, E. and Schektman, Y. (1997): Extension de l'Analyse en Composantes Principales à des Données de Type Intervalle. *Revue de Statistique Appliquée*, **45(3)**, 5-24.

[9] Chattfield C. and Collins A. J. (1980): *An Introduction to Multivariate Analysis*. Chapman and Hall.

[10] Douzal-Chouakria, A., Billard, L. and Diday, E. (2011): Principal Component Analysis for Interval-valued Observations. *Statistical Analysis and Data Mining*, **4**, 229-246.

[11] Gioia, F. and Lauro, C. (2006): Principal Component Analysis on Interval Data. *Computational Statistics*, **21**, 343-363.

[12] Giordani, P. and Kiers, H. A. L. (2006). A comparison of three methods for Principal Component Analysis of fuzzy interval data. *Computational Statistics and Data Analysis*, special issue *The Fuzzy Approach to Statistical Analysis*, **51(1)**, 379-397.

[13] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, **24**, 417-441, and 498-520.

[14] Hotelling, H. (1936). Relations between two sets of variates. Biometrika, **27**, 321-377.

[15] Irpino, A. (2006). Spaghetti PCA analysis: An extension of principal component analysis to time dependent interval data. Pattern Recognition Letters, **27**, 504-513.

[16] Johnson, R. A. and Wichern D.W. (2002): *Applied Multivariate Statistical Analysis* (5^{th} edition). Prentice Hall.

[17] Lauro, N. C. and Palumbo, F. (2000). Principal Component Analysis of Interval Data: A Symbolic Data Analysis Approach. *Computational Statistics*,**15**, 73-87.

[18] Lauro, N. C., Verde, R. and Irpino, A. (2008): Principal Component Analysis of Symbolic Data Described by Intervals. In E. Diday, and M. Noirhomme-Fraiture (Ed.): *Symbolic Data Analysis and the SODAS Software*, 279-311. Chichester: Wiley.

[19] 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.

[20] Le-Rademacher J. and Billard L. (2012): Symbolic-covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics*, **21(2)**, 413-432. Epub 14 Jun 2012.

[21] Le-Rademacher J. and Billard L. (2013): Principal component histograms from interval-valued observations. *Computational Statistics 2013*, **28**, 2117-2138. doi 10.1007/s00180-013-0399-4.

[22] Le-Rademacher, J. and Billard, L. (2013). Principal component analysis for histogram-valued data. Technical report.

[23] Palumbo, F. and Lauro, N. C. (2003): A PCA for Interval-Valued Data Based on Midpoints and Radii. In H. Yanai, A. Okada, K. Shigemasu, Y. Kanu, and J. Meulman (Ed): *New Developments in Psychometrics*, 641-648. Tokyo: Psychometric Society and Springer-Verlag.

[24] Pearson, K. (1901): On lines and planes of closest fit to a system of points in space. *Philosophical Magazine*, **2**, 557-72.

[25] R Core Team (2013): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[26] Zuccolotto, P. (2007): Principal components of sample estimates: an approach through symbolic data analysis. Statistical Methods and Applications, **16 (2)**, 173-192.

A Appendix

The data used are yearly data from the period 1971-2010 for seven stations in Slovenia. The following variables are taken into account: number of cold days (D.Cold), number of warm days (D.Warm), number of days with storms (D.Storm), number of days with precipitations (D.Prec), number of days with snow cover (D.SnCov), number of clear days (D.Clear), and number of cloudy days (D.Cloud). For each station, min and max values are given for each variable under study.

	D.Cold		D.Warm		D.Storm		D.Prec		D.SnCov		D.Clear		D.Cloud	
Station	min	max	min	max	min	max	min	max	min	max	min	max	min	max
Bilje	38	96	67	125	9	63	97	160	0	12	10	114	67	133
Crnom	65	120	48	118	23	59	128	185	6	88	27	99	106	177
Ljubl	52	112	38	109	30	63	119	186	2	110	12	59	89	181
Marib	56	123	37	110	23	52	110	162	3	92	16	83	81	159
MurSo	77	131	33	109	18	47	107	154	0	85	29	77	79	155
Porto	1	67	34	125	37	71	88	143	0	13	68	128	56	124
Ratec	117	181	6	67	22	52	123	170	43	171	31	103	79	153