Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data

by

Jennifer G. Le-Rademacher

(Under the direction of Lynne Billard)

Abstract

Unlike a classical random variable which takes a single value, a symbolic random variable takes multiple values. The values within a symbolic random variable form an internal distribution that does not exist in a classical random variable. Statistical methodologies developed for classical data can not be readily applied to symbolic data. Therefore, new methodologies must be developed to take into account the internal structure of symbolic variables. In this dissertation, we propose three methods of symbolic data analysis. The first proposed method extends classical principal component analysis (PCA) to an analysis of interval-valued observations, using a so-called symbolic variance-covariance structure. Using the symbolic covariance structure ensures that the principal components explain the total variance of interval-valued data. Furthermore, two representations of the principal components resulting from the proposed PCA method are introduced. The first representation shows interval-valued observations as polytopes in a principal components space. The polytopes constructed in this method represent the true structure of interval-valued observations

in a principal components space. The second representation gives histogram-valued principal components constructed from a 2-dimensional projection of the polytopes resulting from the first representation. Algorithms to construct the polytopes and to compute the histograms representing the principal components are given in this dissertation along with two examples to illustrate the method. The second method extends the PCA method proposed for interval-valued data to a PCA method for histogram-valued data. This method treats histogram-valued observations as a generalization of interval-valued observations. The two representations proposed for interval-valued observations are then extended to represent histogram-valued observations. An algorithm for the extension along with an example to illustrate this method are included. The third method proposes a construction of likelihood functions for symbolic data. The proposed likelihood function is then used to derive maximum likelihood estimators for the mean and the variance of three common types of symbolic data: interval-valued data, histogram-valued data, and triangular-distribution-valued data.

INDEX WORDS: Symbolic data analysis, Interval-valued data, Histogram-valued data, Principal component analysis, Likelihood functions, Linear transformation, Convex hull, Polytopes

Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data

by

Jennifer G. Le-Rademacher

B.S., Georgia State University, 1997

B.A., Georgia State University, 1997

M.S., Georgia State University, 2004

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2008

Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data

by

Jennifer G. Le-Rademacher

Major Professor: Lynne Billard

Committee: Jeongyoun Ahn
Cheolwoo Park
T.N. Sriram
Xiangrong Yin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

Dedication

This dissertation is dedicated to my parents who gave me the precious gift of life and have shown me how to live an honest and productive life. They have sacrified their dreams so their children can achieve theirs. I would not be who and where I am today without them.

Acknowledgments

I would first like to thank Dr. Lynne Billard for her guidance and for her intellectual as well as financial support during my research. It is a humbling experience and an honor to have a mentor of her caliber. Her enthusiasm and continuous quest for new ideas inspire me to keep an open mind and to keep exploring new concepts in my future career as a researcher. I admire her energy and tireless work ethics.

Next, I would like to thank Dr. Sriram, Dr. Yin, Dr. Park, and Dr. Ahn for serving on my dissertation committee. I appreciate their comments and especially the time they have to spend reading my dissertation. I would like to thank all my friends in the department for the laughs during stressful times. I would also like to thank the professors and staff of this department for their help throughout my years in the program.

Lastly, I would like to thank my family for their love and encouragement. This journey would have not been accomplished without the endless help from my parents, my sister, Vivian, and my brother-in-law, Chris; without the unconditional support from my husband, Tom; and without the inspiration from my niece, Emma, and my children, Ella and Brooks.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The focus of this dissertation is to propose an approach to principal component analysis (PCA) for symbolic data, specifically for interval-valued data and histogram-valued data and to introduce the likelihood functions for symbolic data. Theory for principal component analysis and the likelihood functions are well developed in the classical setting. However, in the symbolic data setting, only a few adaptations of PCA exist in the literature for interval-valued observations while none exists for the histogram-valued case. Furthermore, no theoretical framework has been established for symbolic data to this point, not even the likelihood functions which are fundamental foundation for many statistical methods. Unlike a classical observation which takes a single value, a symbolic observation takes multiple values. As a consequence, symbolic data have an internal structure which does not exist in classical data. Traditional methods of analysis of classical data do not account for this structure. Therefore, new analytical methods need to be developed to account for this special characteristic of symbolic data.

Symbolic data was first introduced by Diday (1987). Whereas a classical variable takes a single value, a symbolic variable may take a finite or an infinite set of values. A random variable that takes a finite set of values is called a multi-valued variable. The values in a finite set can be either quantitative or qualitative. Another type of symbolic variable may take an infinite set of numerical values ranging from a low to a high value. This type of variable is called an interval-valued variable. A more complex symbolic variable may have weights, probabilities, or even a distribution associated with the values it takes. This class of symbolic variable is called a modal-valued variable. An example of a modal-valued variable

is a histogram-valued variable. Classical data can be thought of as a special case of symbolic data where the internal distribution puts the probability 1 on a single value.

Symbolic data can occur naturally or can be created by aggregating a very large dataset. Billard and Diday (2006) gives many examples of natural and aggregated symbolic data. Some data inherently take multiple values. Two examples of naturally occuring symbolic data include daily temperatures measured in a city and colors of bird species. Daily temperatures are reported at the lowest to the highest point in the day. The resulting observation is then recorded as an interval. In the other example, some species of birds have more than one color so the value for each observation is a finite list of colors which is a multi-valued variable. Some data become symbolic after some processing such as data resulting from a database query. One example of aggregated symbolic data is of claims records retained by an insurance company. Such a dataset may contain thousands if not millions of individual observations, i.e., claims. It is difficult to extract knowledge from a dataset this large. If the insurer's main interest is to understand expenses paid to the policy holders based on age and gender, then the data can be aggregated into gender by age groups. The resulting dataset will then contain variables with multiple values, hence, symbolic.

Symbolic data of the second type has become more common in recent years due to advances in technology which enable storage of extremely large datasets. Larger datasets provide more information about the subjects of interest, however they also present challenges in understanding all the information available. Performing even simple exploratory procedures on these datasets requires a lot of computing power. As a result, much research effort in recent years has been steered toward finding more efficient methods of analysis to accommodate these enormous datasets. One of the methods to make an extremely large dataset manageable is to aggregate the observations into groups of interest. Traditionally, when data are aggregated, only a single value is used for each variable. This value is typically the mean or the median of the group. Using one value to represent all values belonging to a group naturally leads to some loss of information. The information lost during this process

may produce misleading results when the aggregated dataset is analyzed. With the introduction of symbolic data, much of this information is retained by including all values taken by observations in the group.

Principal component analysis (PCA) is a popular dimension reduction method in classical data analysis. Some adaptations of PCA to interval-valued data exist in the literature. Cazes et al. (1997) proposes two methods known as the centers and the vertices methods. The centers method computes the principal components using the centers of the original intervals. Although the order of computation using this method is low, it ignores the internal variation within each observation. The vertices method computes the principal components using all vertices, treating each of them as an independent observation. An advantage of the vertices method over the centers method is its partial accounting for the internal variation by including all vertices belonging to the observations. However, it still does not account for the total variance of interval-valued observations. Another drawback of the vertices method is its treating all vertices as independent observations. Lauro and Palumbo (2000) attempts to account for the dependency among vertices of the same observation by prosposing a multistage method called the symbolic object PCA (SO-PCA). This method includes two separate procedures. The first is a modification of the vertices method and the second involves transformation of the interval ranges. The SO-PCA method still does not account for the total variation structure of interval-valued data. Palumbo and Lauro (2003) proposes yet another method known as the midpoint-radii method and Gioia and Palumbo (2006) introduces an inteval matrix approach. Both the midpoint-radii and the interval matrix methods are based on interval algebra. Interval algebra only works for very narrow intervals which poses a limitation to these methods. To avoid the drawbacks associated with all current methods, in the first part of this dissertation, we propose a new PCA method for interval-valued observations using a so-called symbolic variance-covariance structure, referred to as the symbolic covariance method from hereon. Our proposed method accounts for the total covariance structure

of interval-valued data as well as the dependency among vertices of the same observation. This method can also be applied to data of all ranges.

Despite the existence of multiple PCA approaches to interval-valued data, none has been proposed for histogram-valued data. In the second part of this dissertation, we propose a method to compute the principal components for histogram-valued observations based on the fact that a histogram is a generalization of an interval. We generalize our proposed PCA method for interval-valued data to a PCA method for histogram-valued data. The method we propose for histogram-valued data inherits all benefits from the symbolic covariance PCA method for interval-valued data. The principal components computed from this method account for the total variance of histogram-valued observations. This method can also be applied to histograms of all sizes.

Likelihood functions play an imperative role in a statistical framework. They are tools for solving problems from estimation to regression. Likelihood functions are well studied in the classical environment. However, no extension to symbolic data exists at this time. Further development of some symbolic methodologies can not proceed without these functions. As the third and final part of this dissertation, we propose an approach to finding the likelihood functions of symbolic data. We also derive the maximum likelihood estimators for some common types of symbolic data based on the proposed likelihood functions in this part.

This dissertation consists of six chapters. Following the introduction, we give a review of current literature on symbolic data, principal component analysis, and PCA methods for interval-valued data in Chapter 2. Chapter 3 describes our proposed method for interval-valued PCA. Besides details of the theoretical framework, Chapter 3 also includes an algorithm to compute the principal components for interval-valued observations as well as application of the symbolic covariance PCA to two real datasets. Chapter 4 describes an extension of the symbolic covariance PCA method of Chapter 3 to histogram-valued data. Chapter 4 provides the framework for the proposed extension, an algorithm, as well as an application of the proposed method to a histogram-valued dataset. In Chapter 5, we propose a method

to construct the likelihood functions for symbolic data. We also derive the maximum likelihood estimators for the mean and the variance of some common types of symbolic variables. Finally, a summary of contributions to symbolic data analysis resulting from this dissertation along with a brief discussion of future work stemming from this work are presented in Chapter 6.

1.1 References

Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-Valued Data. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, G. Cucumel, P. Bertrand and F. de Carvalho). Springer-Verlag, Berlin, 3-12.

Billard, L., Chouakia-Douzal, A., and Diday, E. (2007). Symbolic Principal Component Analysis for Interval-Valued Observations. *Journal of the American Statistical Association*, pending acceptance.

Billard, L. and Diday E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, **98**, 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, New York.

Bock, H.-H. and Diday, E. (eds.) (2000). *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.

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*, **XLV** (3). 5-24.

Chouakria, A., Cazes, P., and Diday, E. (2000). Symbolic Principal Component Analysis. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 200-212.

Diday, E. (1987). Introduction à l'Approache Symbolique en Analyse des Données. *Premières Journées Symbolic-Numérique*, CEREMADE, Université Paris, 21-56.

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

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

Jolliffe, I.T. (2004). *Principal Component Analysis*, 2nd edition. Springer, New York.

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

Lauro, C. and Palumbo, F. (2003). Some Results and New Perspectives in Principal Component Analysis for Interval Data. *CLADAG '03 Book of Short Papers* 237-244.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, New York.

Moore, R. (1966). *Interval Analysis*. Prentice-Hall, New Jersey.

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

Chapter 2

Literature Review

To establish a foundation for Chapters 3, 4, and 5, we give a review of current literature on material relating to our work in this dissertation. Section 2.1 introduces notation and some basic terms necessary to discuss symbolic data. This section also provides descriptive statistics for three different types of symbolic data. A brief summary of principal component analysis methodology is given in Section 2.2. Five methods of PCA have been proposed for interval-valued data up to this point. They are summarized in Section 2.3.

2.1 Symbolic Data

Before starting the literature review, it is necessary to define common notation used in the rest of this work. Let $\boldsymbol{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(p)})$ denote a p-variate random variable where $X_{(j)}$ is the j^{th} variable for $j = 1, 2, \ldots, p$. Let $\boldsymbol{X}_i$ denote the i^{th} observation of a data matrix $\boldsymbol{X}$ where $i = 1, 2, \ldots, n$. Note the distinction between the subscripts. The subscript enclosed in parentheses as in $X_{(j)}$ signifies the variable index whereas the subscript without parentheses as in $\boldsymbol{X}_i$ represents the observation index. Using this notation, a data matrix $\boldsymbol{X}$ can be expressed as a vector of variables or a vector of observations, respectively,

$$\boldsymbol{X} = \left[\boldsymbol{X}_{(1)}\boldsymbol{X}_{(2)}\ldots\boldsymbol{X}_{(p)}\right] = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}.$$

Additionally, the random variable X_{ij} represents the j^{th} variable of the i^{th} observation and the lower case x_{ij} is the realized value of X_{ij} in the classical case and ξ_{ij} denotes a realized

value of a symbolic random variable X_{ij}. The notation ξ_{ij} is used in the symbolic case to emphasize that ξ_{ij} takes multiple values unlike the single value x_{ij} in classical data. Some examples follow. For classical data, suppose $X_{(1)}$ is the birth-weight of a newborn in pounds; then the 5^{th} observed value may take a value $x_{51} = 8.2$ and suppose $X_{(3)}$ is the newborn's eyes color then the observed value for the same baby may be $x_{53} = blue$. For symbolic data, suppose $X_{(1)}$ is the daily temperature; then the recorded temperatures for the 1^{st} observation may take values $\xi_{11} = [44, 62]$ and suppose $X_{(2)}$ is the color of species of birds; then the observed value for the 1^{st} species may take values $\xi_{12} = \{blue, green\}$.

In the following subsections we give formal definitions necessary to discuss distribution functions and descriptive statistics for symbolic data. For a comprehensive treatment of the topic, refer to Bertrand and Goupil (2000), Bock and Diday (2000), and Billard and Diday (2006).

Let $\mathcal{X}_{(j)}$ be the domain of $X_{(j)}$ and $\mathcal{X} = \mathcal{X}_{(1)} \times \mathcal{X}_{(2)} \times \ldots \times \mathcal{X}_{(p)} = \times_{j=1}^{p} \mathcal{X}_{(j)}$ be the domain of $\boldsymbol{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(p)})$. Then,

Definition 2.1.1. Every point $\boldsymbol{x} = (x_{(1)}, x_{(2)}, \ldots, x_{(p)}) \in \mathcal{X}$ is called a *description vector*.

Definition 2.1.2. Every $D \subseteq \mathcal{X}$ such that $D = D_{(1)} \times D_{(2)} \times \ldots \times D_{(p)}$ where $D_{(j)} \subseteq \mathcal{X}_{(j)}$ is called a *description set*.

When $D = D_{(1)} \times D_{(2)} \times \ldots \times D_{(p)}$ is the Cartesian product, D is called a *Cartesian description set*.

When $D_{(j)}$ is a singleton, i.e., $D_{(j)} = \{x_{(j)}\}$ for all $j = 1, 2, \ldots, p$, a description vector $\boldsymbol{d} = (D_{(1)}, D_{(2)}, \ldots, D_{(p)}) = (\{x_{(1)}\}, \{x_{(2)}\}, \ldots, \{x_{(p)}\})$ is called an *individual description vector*.

Definition 2.1.3. Let $A \subseteq D$ and $B \subseteq D$ be two description sets and $\boldsymbol{x} \in \mathcal{X}$. Define a *logical dependency rule* v as

$$v : [\boldsymbol{x} \in A] \Rightarrow [\boldsymbol{x} \in B] .$$

Equivalently, v is a mapping from $\mathcal{X}$ onto $\{0,1\}$ such that

$$v(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{x} \in (A \cap B) \text{ or } \boldsymbol{x} \notin A, \\ 0, & \text{otherwise.} \end{cases}$$

The set of all logical dependency rules v defined on $\mathcal{X}$ is denoted by $V_{\mathcal{X}}$.

Definition 2.1.4. The *virtual description* of the description vector $\boldsymbol{d}$, *vir($\boldsymbol{d}$)*, is the set of all individual description vectors $\boldsymbol{x}$ that satisfy all the rules v in $\mathcal{X}$. That is,

$$vir(\boldsymbol{d}) = \{\boldsymbol{x} \in D | v(\boldsymbol{x}) = 1, \forall v \in V_{\mathcal{X}}\}.$$

Now that some preliminary terms have been established, the observed frequencies, the empirical distribution functions, and some descriptive statistics can be defined. In Subsection 2.1.1 we give the definitions of these statistics for multi-valued variable. For detailed derivations and examples, refer to Bertrand and Goupil (2000) and Billard and Diday (2006). However, for interval-valued variable more details of these statistics will be given in Subsection 2.1.2. In Subsection 2.1.3, we give the empirical distribution function and some descriptive statistics for histogram-valued variabel without derivation. Derivation of the distribution function and descriptive statistics for this type of variable can be generalized from the interval-valued case; therefore, they are not presented in this dissertation. Again, to learn more, see Bertrand and Goupil (2000) and Billard and Diday (2006).

2.1.1 Multi-valued data

Given a random sample $\boldsymbol{X}_i$ for $i = 1, 2, \ldots, n$, let the j^{th} variable, $X_{(j)}$, be a multi-valued random variable and $\xi_{(ij)}$ be a realization of X_{ij}. Let W be a value in $X_{(j)}$. Then, the observed frequency of W taking value $\xi \in \xi_{ij}$ is given by

$$O_W(\xi) = \sum_{i=1}^{n} \frac{|\{\boldsymbol{x} \in vir(\boldsymbol{d}_i) | x_{(j)} = \xi\}|}{|vir(\boldsymbol{d}_i)|}$$

where $|A|$ is the number of elements belonging to set A. Hence, the empirical distribution function of W is

$$F_W(\xi) = \frac{1}{n'} \sum_{\xi_k \leq \xi} O_W(\xi_k) \tag{2.1}$$

where $n' = n - n_0$ and n_0 is the number of i for which $|vir(\boldsymbol{d}_i)| = 0$. Moreover, if W is quantitative, the symbolic sample mean and the symbolic sample variance can be derived using the empirical distribution function in Equation (2.1). Let $\bar{W}$ and S^2 be the sample mean and the sample variance of W. Then, they are given by, respectively,

$$\bar{W} = \frac{1}{n'} \sum_{\xi_k \in \mathcal{X}_{(j)}} \xi_k O_W(\xi_k)$$

and

$$S^2 = \frac{1}{n'} \sum_{\xi_k \in \mathcal{X}_{(j)}} (\xi_k - \bar{W})^2 O_W(\xi_k).$$

2.1.2 Interval-valued data

Since Chapter 3 of this dissertation focuses on principal component analysis for interval-valued data, more detailed information is given for the statistics of this type of variables. We will show some derivations of the empirical distribution function, the mean, and the variance-covariance for interval-valued variable in this subsection. Again, let $\boldsymbol{X}_i, i = 1, \ldots, n$, be a random sample. Let the j^{th} variable, $X_{(j)}$, be an interval-valued variable. Then, a realization ξ_{ij} of X_{ij} takes an interval of values $[a_{ij}, b_{ij}]$ where $a_{ij} \leq b_{ij}$. Let W be a point in $X_{(j)}$. Assume W is uniformly distributed over the interval $X_{ij} = [a_{ij}, b_{ij}]$ for all individual description vectors $\boldsymbol{x} \in vir(\boldsymbol{d}_i)$. Then, for each ξ,

$$P\{W \leq \xi | \boldsymbol{x} \in vir(\boldsymbol{d}_i)\} = \begin{cases} 0, & \xi < a_{ij}, \\ \frac{\xi - a_{ij}}{b_{ij} - a_{ij}}, & a_{ij} \leq \xi < b_{ij}, \\ 1, & b_{ij} \leq \xi. \end{cases}$$

Furthermore, assume each object is equally likely to be observed with probability $1/n$, then the empirical distribution function of W is,

$$\begin{aligned} F_W(\xi) &= \frac{1}{n} \sum_{i=1}^{n} P\{W \leq \xi | \boldsymbol{x} \in vir(\boldsymbol{d}_i)\} \\ &= \frac{1}{n} \left\{ \sum_{i:\xi \in \xi_{ij}} (\frac{\xi - a_{ij}}{b_{ij} - a_{ij}}) + |(i|\xi \geq b_{ij})| \right\}. \end{aligned}$$

Differentiating $F_W(\xi)$ with respect to ξ produces the empirical density of W,

$$f_W(\xi) = \frac{1}{n} \sum_{i:\xi \in \xi_{ij}} (\frac{1}{b_{ij} - a_{ij}}). \tag{2.2}$$

Bertrand and Goupil (2000) further defines the symbolic sample mean and symbolic sample variance of W as,

$$\bar{W} = \frac{1}{2n} \sum_{i=1}^{n} (a_{ij} + b_{ij}) \tag{2.3}$$

and

$$S^2 = \frac{1}{3n} \sum_{i=1}^{n} (a_{ij}^2 + a_{ij} b_{ij} + b_{ij}^2) - \frac{1}{4n^2} \left[\sum_{i=1}^{n} (a_{ij} + b_{ij}) \right]^2. \tag{2.4}$$

These two statistics are due to the following as shown in Bertrand and Goupil (2000) and Billard and Diday (2006). First, the symbolic sample mean based on the empirical density function in Equation (2.2) is,

$$\begin{aligned} \bar{W} &= \int_{-\infty}^{\infty} \xi f(\xi) \partial \xi \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{b_{ij} - a_{ij}} \int_{-\infty}^{\infty} \xi \partial \xi \right] \\ &= \frac{1}{2n} \sum_{i=1}^{n} \left[(\frac{1}{b_{ij} - a_{ij}}) (\xi^2 |_{a_{ij}}^{b_{ij}}) \right] \\ &= \frac{1}{2n} \sum_{i=1}^{n} (b_{ij} + a_{ij}). \end{aligned}$$

Then, the symbolic sample variance can be written as,

$$\begin{aligned} S^2 &= \int_{-\infty}^{\infty} (\xi - \bar{W})^2 f(\xi) \partial \xi \\ &= \int_{-\infty}^{\infty} \xi^2 f(\xi) \partial \xi - \bar{W}^2. \end{aligned}$$

Now,

$$\begin{aligned}\int_{-\infty}^{\infty}\xi^2 f(\xi)\partial\xi &= \frac{1}{n}\sum_{i=1}^{n}\left[(\frac{1}{b_{ij}-a_{ij}})\int_{-\infty}^{\infty}\xi^2\partial\xi\right]\\ &= \frac{1}{3n}\sum_{i=1}^{n}\left[(\frac{1}{b_{ij}-a_{ij}})(\xi^3|_{a_{ij}}^{b_{ij}})\right]\\ &= \frac{1}{3n}\sum_{i=1}^{n}\left[(\frac{1}{b_{ij}-a_{ij}})({b_{ij}}^3-{a_{ij}}^3)\right]\\ &= \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2+a_{ij}b_{ij}+b_{ij}^2).\end{aligned}$$

Therefore,

$$S^2 = \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2+a_{ij}b_{ij}+b_{ij}^2)-\frac{1}{4n^2}\left[\sum_{i=1}^{n}(a_{ij}+b_{ij})\right]^2.$$

Now, extend Equation (2.4) to the bivariate case. Let $S_{jj'}$ be the covariance for $W_{(j)}$ and $W_{(j')}$ where $W_{(j)}$ is a point in $X_{(j)}$ and $W_{(j')}$ is a point in $X_{(j')}$ for $j, j' = 1, 2, \ldots, p$. The empirical symbolic covariance for $W_{(j)}$ and $W_{(j')}$ analogous to S^2 is given by

$$S_{jj'} = \frac{1}{3n}\sum_{i=1}^{n}G_{ij}G_{ij'}[Q_{ij}Q_{ij'}]^{1/2} \tag{2.5}$$

where,

$$\begin{aligned}Q_{ij} &= (a_{ij}-\bar{W}_{(j)})^2+(a_{ij}-\bar{W}_{(j)})(b_{ij}-\bar{W}_{(j)})+(b_{ij}-\bar{W}_{(j)})^2,\\ G_{ij} &= \begin{cases}-1, & \bar{W}_{ij}\le\bar{W}_{(j)},\\ 1, & \bar{W}_{ij}>\bar{W}_{(j)},\end{cases}\end{aligned}$$

and $\bar{W}_{ij} = (a_{ij}+b_{ij})/2$; see Billard (2007).

To verify that $S_{jj} = S^2$ when $j = j'$, let us look at

$$\begin{aligned}S_{jj} &= \frac{1}{3n}\sum_{i=1}^{n}G_{ij}G_{ij}[Q_{ij}Q_{ij}]^{1/2}\\ &= \frac{1}{3n}\sum_{i=1}^{n}Q_{ij}\\ &= \frac{1}{3n}\sum_{i=1}^{n}[(a_{ij}-\bar{W}_{(j)})^2+(a_{ij}-\bar{W}_{(j)})(b_{ij}-\bar{W}_{(j)})+(b_{ij}-\bar{W}_{(j)})^2].\\ &= \frac{1}{3n}\sum_{i=1}^{n}[a_{ij}^2+a_{ij}b_{ij}+b_{ij}^2]-\bar{W}_{(j)}^2.\end{aligned}$$

which equals the S^2 in Equation (2.4).

Billard (2007) shows that the total sum of squares of interval-valued observartions $\boldsymbol{X}_i$ for $i = 1, \ldots, n$ can be decomposed into the sum of the internal variation and the external variation called the within sum of squares and the between sum of squares, respectively. Formally,

$$nS^2 = SST = SSB + SSW \tag{2.6}$$

where

$$SSB = \sum_{i=1}^{n} (\bar{W}_{ij} - \bar{W}_{(j)})^2 \tag{2.7}$$

and

$$SSW = \frac{1}{3} \sum_{i=1}^{n} [(a_{ij} - \bar{W}_{ij})^2 + (a_{ij} - \bar{W}_{ij})(b_{ij} - \bar{W}_{ij}) + (b_{ij} - \bar{W}_{ij})^2]. \tag{2.8}$$

Since $(b_{ij} - \bar{W}_{ij}) = (\bar{W}_{ij} - a_{ij}) = \frac{1}{2}(b_{ij} - a_{ij})$,

$$\begin{aligned} SSW &= \frac{1}{3} \sum_{i=1}^{n} \left[(a_{ij} - \bar{W}_{ij})^2 + (a_{ij} - \bar{W}_{ij})(b_{ij} - \bar{W}_{ij}) + (b_{ij} - \bar{W}_{ij})^2 \right] \\ &= \frac{1}{3} \sum_{i=1}^{n} \left[\frac{(b_{ij} - a_{ij})}{2} \right]^2 \\ &= \frac{1}{12} \sum_{i=1}^{n} (b_{ij} - a_{ij})^2. \end{aligned}$$

This result is consistent with the assumption that for each observation $\boldsymbol{X}_i$, W_{ij} is uniformly distributed in the interval $[a_{ij}, b_{ij}]$. That is, for $W_{ij} \sim U(a_{ij}, b_{ij})$,

$$Var(W_{ij}) = \frac{(b_{ij} - a_{ij})^2}{12}. \tag{2.9}$$

With n observations $\boldsymbol{X}_i, i = 1, \ldots, n$, the total variance is the sum of n variances defined in Equation (2.9)

Analogously, when $j \neq j'$ the total sum of products (SPT) is the sum of the internal and the external sum of products, respectively (SPW) and (SPB), is given by,

$$nS_{jj'} = SPT_{jj'} = SPB_{jj'} + SPW_{jj'} \tag{2.10}$$

where

$$SPB_{jj'} = \sum_{i=1}^{n} (\bar{W}_{ij} - \bar{W}_{(j)})(\bar{W}_{ij'} - \bar{W}_{(j')}) \tag{2.11}$$

and

$$SPW_{jj'} = \frac{1}{12} \sum_{i=1}^{n} (a_{ij} - b_{ij})(a_{ij'} - b_{ij'}). \tag{2.12}$$

It is worth noting that Equations (2.6) and (2.10) can be extended to numerically modal-valued data such as histogram-valued data or data with non-uniform internal distribution. In these cases, the formula for the SSB and the SPB as defined in Equations (2.7) and (2.11) remain the same. However, the SSW and SPW vary depending on the internal distribution of the data. This presents a potential direction for future research.

2.1.3 Histogram-valued data

Let $\boldsymbol{X}_i, i = 1, \ldots, n$, be a random sample. Let the j^{th} variable, $X_{(j)}$, be a histogram-valued variable. Let ξ_{ij} be a realization of X_{ij}. Then, ξ_{ij} takes a histogram of values, i.e.,

$$\xi_{ij} = \left\{ [a_{ij}^1, b_{ij}^1), p_{ij}^1; [a_{ij}^2, b_{ij}^2), p_{ij}^2; \ldots; [a_{ij}^{s_{ij}}, b_{ij}^{s_{ij}}], p_{ij}^{s_{ij}} \right\}, \tag{2.13}$$

where $[a_{ij}^l, b_{ij}^l)$ is called the l^{th} subinterval of ξ_{ij} and p_{ij}^l is the relative frequency associated with the l^{th} subinterval. Let s_{ij} denote the number of subintervals in histogram ξ_{ij}. Then, $a_{ij}^l \leq b_{ij}^l$ for all $l = 1, 2, \ldots, s_{ij}$ and $\sum_{l=1}^{s_{ij}} p_{ij}^l = 1$.

Billard and Diday (2003) extends the empirical distribution function derived by Bertrand and Goupil (2000) for interval-valued data to a distribution function for histogram-valued data. Based on the assumption that all values within each subinterval $[a_{ij}^l, b_{ij}^l)$ are uniformly distributed, Billard and Diday (2003) defines the distribution of a point W^l within subinterval $[a_{ij}^l, b_{ij}^l]$ as,

$$P\{W^l \leq \xi | \boldsymbol{x} \in vir(\boldsymbol{d}_i)\} = \begin{cases} 0, & \xi < a_{ij}^l, \\ \frac{\xi - a_{ij}^l}{b_{ij}^l - a_{ij}^l}, & a_{ij}^l \leq \xi < b_{ij}^l, \\ 1, & b_{ij}^l \leq \xi. \end{cases}$$

Billard and Diday (2003) further proposes an empirical distribution function for a random variable W of $X_{(j)}$ in a similar manner to the empirical distribution function of an interval-valued variable as described in Subsection 2.1.2. The empirical distribution function of W is given by

$$F_W(\xi) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l:\xi\in\xi_{ij}^l} p_{ij}^l\left(\frac{\xi - a_{ij}^l}{b_{ij}^l - a_{ij}^l}\right) + \sum_{l:\xi\geq b_{ij}^l} p_{ij}^l\right) \tag{2.14}$$

where $\xi_{ij}^l = [a_{ij}^l, b_{ij}^l)$. Taking the derivative of F_W in Equation (2.14) with respect to ξ gives the following empirical density function of W,

$$f_W(\xi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{l:\xi\in\xi_{ij}^l} p_{ij}^l\left(\frac{1}{b_{ij}^l - a_{ij}^l}\right). \tag{2.15}$$

The symbolic sample mean and the symbolic sample variance derived from the density function defined in Equation (2.15) are, respectively,

$$\bar{W} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{s_{ij}} p_{ij}^l(a_{ij}^l + b_{ij}^l) \tag{2.16}$$

and

$$S^2 = \frac{1}{3n}\sum_{i=1}^{n}\sum_{l=1}^{s_{ij}} p_{ij}^l[(a_{ij}^l)^2 + a_{ij}^l b_{ij}^l + (b_{ij}^l)^2] - \frac{1}{4n^2}\left[\sum_{i=1}^{n}\sum_{l=1}^{s_{ij}} p_{ij}^l(a_{ij}^l + b_{ij}^l)\right]^2. \tag{2.17}$$

Billard and Diday (2006) extends the variance in Equation (2.17) to the bivariate case. Let $S_{jj'}$ be the covariance of $W_{(j)}$ in $X_{(j)}$ and $W_{(j')}$ in $X_{(j')}$ where $j, j' = 1, 2, \ldots, p$. The empirical symbolic covariance for $W_{(j)}$ and $W_{(j')}$ analogous to S^2 of Equation (2.17) is given by

$$S_{jj'} = \frac{1}{3n}\sum_{i=1}^{n}\left(G_{ij}G_{ij'}\sum_{l_j=1}^{s_{ij}}\sum_{l_{j'}=1}^{s_{ij'}} p_{ij}^{l_j} p_{ij'}^{l_{j'}}[Q_{ij}^{l_j} Q_{ij'}^{l_{j'}}]^{1/2}\right) \tag{2.18}$$

where

$$\begin{aligned} Q_{ij}^{l_j} &= (a_{ij}^{l_j} - \bar{W}_{(j)})^2 + (a_{ij}^{l_j} - \bar{W}_{(j)})(b_{ij}^{l_j} - \bar{W}_{(j)}) + (b_{ij}^{l_j} - \bar{W}_{(j)})^2, \\ G_{ij} &= \begin{cases} -1, & \bar{W}_{ij} \leq \bar{W}_{(j)}, \\ 1, & \bar{W}_{ij} > \bar{W}_{(j)}, \end{cases} \end{aligned}$$

and $\bar{W}_{ij} = \frac{1}{2}\sum_{l_j=1}^{s_{ij}} p_{ij}^{l_j}(a_{ij}^{l_j} + b_{ij}^{l_j})$.

2.2 Principal Component Analysis

Principal component analysis (PCA) is a popular multivariate method in classical data analysis. A PCA computes uncorrelated linear combinations of the original variables so that the first linear combination has the largest variance, the second linear combination has the second largest variance, and so on. These linear combinations are called the principal components. Situations where PCA is most applicable include problems with high dimensional data and problems with highly correlated data. Suppose a dataset has p variables where p is very large. Then, the principal components can be computed and the first k ($< p$) principal components which contain most of the variation from the original variables are selected for further analysis. Then, analysis can be performed on these components without much loss of information. In this situation, PCA reduces the dimension of the data, hence reducing the computing power required to solve the problem. In another situation where correlation exists among the variables, PCA creates uncorrelated variables, hence eliminates the problem of collinearity. A PCA is typically used as an exploratory tool. After the principal components are computed, analysis and interpretation can then be performed on the principal components instead of the original variables. Due to its wide applicability, theories of PCA are well developed. Most textbooks in multivariate methods include a chapter on PCA, for example Anderson (2002), Mardia et al. (1979), Johnson and Wichern (2002). Jolliffe (2004) is a full text dedicated to principal component analysis.

Let $\boldsymbol{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(p)})$ be a p-variate random variable from a distribution with mean $\boldsymbol{\mu} = (\mu_{(1)}, \mu_{(2)}, ..., \mu_{(p)})$ and variance-covariance $\boldsymbol{\Sigma} = [\sigma_{jj'}]$ for $j, j' = 1, 2, \ldots, p$. The first principal component is the linear combination $Y_{(1)}$ of $\boldsymbol{X}$ such that $Y_{(1)} = \boldsymbol{\alpha}_1'\boldsymbol{X}$ has the largest variance under the constraint that $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$. The $\boldsymbol{\alpha}_1$ that maximizes $Var(Y_{(1)})$ under these conditions can be found using a Lagrange multiplier as follows. Since $Var(Y_{(1)}) =$

$Var(\boldsymbol{\alpha}_1'\boldsymbol{X}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$, let

$$\begin{aligned}\phi_1 &= Var(Y_{(1)}) - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1) \\ &= \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1).\end{aligned}$$

Differentiate ϕ_1 with respect to $\boldsymbol{\alpha}_1$ and set the derivative equal to zero,

$$\frac{\partial \phi_1}{\partial \boldsymbol{\alpha}_1} = 2(\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1) = 0.$$

Hence,

$$(\boldsymbol{\Sigma} - \lambda\boldsymbol{I})\boldsymbol{\alpha}_1 = 0. \tag{2.19}$$

Then, take the derivative of ϕ_1 with respect to λ and set this derivative to zero,

$$\frac{\partial \phi_1}{\partial \lambda} = (\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1).$$

This gives

$$\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1. \tag{2.20}$$

The λ and $\boldsymbol{\alpha}_1$ that satisfy both Equations (2.19) and (2.20) are an eigenvalue-eigenvector pair of $\boldsymbol{\Sigma}$. Moreover, since

$$Var(Y_{(1)}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \lambda$$

and $Y_{(1)}$ has the largest variance, then $\lambda = \lambda_1$ where λ_1 is the largest eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ the eigenvector corresponding to λ_1. Therefore, the first principal component is a linear combination of $\boldsymbol{X}$ whose coefficients are the elements of the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$.

The second principal component is a linear combination $Y_{(2)} = \boldsymbol{\alpha}_2'\boldsymbol{X}$ with the constraint $\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = 1$. Moreover, $Y_{(2)}$ has the second largest variance and it is uncorrelated to $Y_{(1)}$, i.e., the covariance, $Cov(Y_{(1)}, Y_{(2)}) = 0$. Again, a Lagrange multiplier is used to find $\boldsymbol{\alpha}_2$. Since $Var(Y_{(2)}) = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ and

$$Cov(Y_{(1)}, Y_{(2)}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = 0, \tag{2.21}$$

$$\begin{aligned}\phi_2 &= Var(Y_{(2)}) - \lambda(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \delta(\lambda_1\boldsymbol{\alpha}_2)'\boldsymbol{\alpha}_1 \\ &= \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \delta(\lambda_1\boldsymbol{\alpha}_2)'\boldsymbol{\alpha}_1. \qquad (2.22)\end{aligned}$$

Differentiate ϕ_2 of Equation (2.22)with respect to $\boldsymbol{\alpha}_2$ and set it equal to zero,

$$\frac{\partial\phi_2}{\partial\boldsymbol{\alpha}_2} = 2(\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2) - \delta\lambda_1\boldsymbol{\alpha}_1 = 0. \qquad (2.23)$$

Multiplying the left hand side of Equation (2.23) by $\boldsymbol{\alpha}_1$ and then using Equation (2.21), we obtain

$$2\boldsymbol{\alpha}_1'(\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2) - \delta\lambda_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \delta\lambda_1 = 0.$$

Since $\lambda_1 \neq 0$, $\delta = 0$. Hence, Equation (2.23) becomes

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = 0.$$

Similar to the solution for the first principal component, $\boldsymbol{\alpha}_2$ is the eigenvector corresponding to the second largest eigenvalue, λ_2. That is, the second principal component, $Y_{(2)}$, is a linear combination of $\boldsymbol{X}$ whose coefficients are the elements of the eigenvector corresponding to the second largest eigenvalue of $\boldsymbol{\Sigma}$. The rest of the principal components are found the same way. More formally, the k^{th} principal component is the linear combination

$$Y_{(k)} = \boldsymbol{\alpha}_k'\boldsymbol{X}$$

where $\boldsymbol{\alpha}_k$ is the eigenvector corresponding to the k^{th} eigenvalue of $\boldsymbol{\Sigma}$ and $Cov(Y_{(k)}, Y_{(k')}) = 0$ for $k \neq k'$.

Vector $\boldsymbol{\alpha}_k$ is the vector of coefficients of principal component $Y_{(k)}$. The magnitude of element j of $\boldsymbol{\alpha}_k$, denoted by α_{jk}, indicates the significance of variable $X_{(j)}$ to principal component $Y_{(k)}$. Coefficient α_{jk} is also proportional to the correlation between $X_{(j)}$ and $Y_{(k)}$. Let $\rho_{X_{(j)},Y_{(k)}}$ be the correlation between $X_{(j)}$ and $Y_{(k)}$. Then,

$$\rho_{X_{(j)},Y_{(k)}} = \frac{\alpha_{jk}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}}. \qquad (2.24)$$

The correlation defined in Equation (2.24) provides another avenue to understand the contribution of variable $X_{(j)}$ to principal component $Y_{(k)}$. Unlike α_{jk} which indicates the importance of variable $X_{(j)}$ to principal component $Y_{(k)}$ when other variables are included, $\rho_{X_{(j)},Y_{(k)}}$ measures the importance of variable $X_{(j)}$ to principal component $Y_{(k)}$ individually.

In situations where it is appropriate to standardize the variance, principal components can be found using the correlation coefficient matrix $\boldsymbol{\rho} = [\rho_{jj'}]$ where

$$\rho_{jj'} = \frac{\sigma_{jj'}}{\sigma_{(j)}\sigma_{(j')}}$$

for $j, j' = 1, 2, \ldots, p$ where $\sigma_{(j)} = \sqrt{\sigma_{jj}}$. In this case, $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ are the eigenvalues and $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$ are the eigenvectors of the correlation matrix $\boldsymbol{\rho}$. The correlation measure between $X_{(j)}$ and $Y_{(k)}$ of Equation (2.24) becomes

$$\rho_{X_{(j)},Y_{(k)}} = \alpha_{jk}\sqrt{\lambda_k} \tag{2.25}$$

because element ρ_{jj} of $\boldsymbol{\rho}$ is one for all $j = 1, \ldots, p$.

When the population distribution is not known, estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are used to compute the principal components. In the sample context, let $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ be the i^{th} observation where $i = 1, 2, \ldots, n$ and p is the number of variables. Let $\boldsymbol{S}$ be the sample variance-covariance matrix of

$$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n).$$

Let

$$\boldsymbol{X}^* = (\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_n^*)$$

be the centered data matrix where

$$\boldsymbol{X}_i^* = (X_{i1} - \bar{X}_{(1)}, X_{i2} - \bar{X}_{(2)}, \ldots, X_{ip} - \bar{X}_{(p)})'$$

and

$$\bar{X}_{(j)} = \frac{\sum_{i=1}^{n} X_{ij}}{n}.$$

Then, $\boldsymbol{S}$ can be computed easily as,

$$\boldsymbol{S} = \frac{(\boldsymbol{X}^*)'(\boldsymbol{X}^*)}{n-1}.$$

Let $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_p$ be the eigenvalues of $\boldsymbol{S}$ and $\boldsymbol{\nu}_k$ be the eigenvector corresponding to $\hat{\lambda}_k$ for $k = 1, \ldots, p$. Then, it turns out that the coefficients vector for the k^{th} principal component, $Y_{(k)}$, based on the data matrix $\boldsymbol{X}$ is $\boldsymbol{\nu}_k$.

The sampling distributions of $\hat{\lambda}_k$ and $\boldsymbol{\nu}_k$ are difficult to derive. Anderson (1963) and Girshick (1939) derive asymptotic results for the sampling distributions of $\hat{\lambda}_k$ and $\boldsymbol{\nu}_k$ under the following assumptions: observations $\boldsymbol{X}_i, i = 1, \ldots, n$, is a random sample from a normal distribution and the eigenvalues of the population covariance matrix $\boldsymbol{\Sigma}$ are distinct and positive. Johnson and Wichern (1984) gives the following summary of these results:

1. Let $\boldsymbol{\Lambda}$ be the diagonal matrix of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $\boldsymbol{\Sigma}$; then, $\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ is approximately $N_p(\boldsymbol{0}, 2\boldsymbol{\Lambda}^2)$.

2. Let
$$\boldsymbol{A}_{(k)} = \lambda_k \sum_{j=1, j\neq k}^{p} \frac{\lambda_j}{(\lambda_j - \lambda_k)^2} (\boldsymbol{\alpha}_{(j)} \boldsymbol{\alpha}'_{(j)});$$
then, $\sqrt{n}(\boldsymbol{\nu}_{(k)} - \boldsymbol{\alpha}_{(k)})$ is approximately $N_p(\boldsymbol{0}, \boldsymbol{A}_{(k)})$.

3. Each $\hat{\lambda}_k$ is distributed independently of the elements of the associated $\boldsymbol{\nu}_{(k)}$.

Based on these results, asymptotic estimation and inference for the eigenvalues and the eigenvectors of $\boldsymbol{\Sigma}$ can be performed.

Furthermore, PCA can also be computed based on the sample correlation coefficient vector, $\boldsymbol{R}$, as in the population case. See, e.g., Anderson(2002), Mardia et al. (1979), Wichern and Johnson (2002), and Joliffe (2004) for details.

2.3 Current PCA Methods for Interval-valued Data

Since principal component analysis is a popular multivariate method, naturally it is necessary to extend PCA to application of symbolic data. Some extensions of PCA to interval-valued

data currently found in the literature include the centers and the vertices method (Cazes et al. (1997), Chouakria et al. (2000), and most extensively, Billard et al. (2007)); the symbolic object approach by Lauro and Palumbo (2000), the midpoints and radii method by Palumbo and Lauro (2003), and the interval algebra approach by Gioia and Lauro (2006).

Before presenting these methods, let us define an interval-valued data matrix. Let $\boldsymbol{X}$ be an $n \times p$ data matrix. Then,

$$\boldsymbol{X} = \begin{bmatrix} \xi_{11} & \xi_{12} & \dots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \dots & \xi_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \xi_{n1} & \xi_{n2} & \dots & \xi_{np} \end{bmatrix}. \tag{2.26}$$

If $\boldsymbol{X}$ is an interval-valued data matrix, then $\boldsymbol{X}$ of Equation (2.26) has the following form,

$$\boldsymbol{X} = \begin{bmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & \dots & [a_{1p}, b_{1p}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & \dots & [a_{2p}, b_{2p}] \\ \vdots & \vdots & \vdots & \vdots \\ [a_{n1}, b_{n1}] & [a_{n2}, b_{n2}] & \dots & [a_{np}, b_{np}] \end{bmatrix} \tag{2.27}$$

where $a_{ij} \leq b_{ij}$ for all $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$.

2.3.1 Centers method

Let the matrix of centers corresponding to the data matrix defined in Equation (2.27) be

$$\boldsymbol{X}^c = \begin{bmatrix} X_{11}^c & X_{12}^c & \dots & X_{1p}^c \\ X_{21}^c & X_{22}^c & \dots & X_{2p}^c \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1}^c & X_{n2}^c & \dots & X_{np}^c \end{bmatrix} \tag{2.28}$$

where

$$X_{ij}^c = \frac{a_{ij} + b_{ij}}{2}. \tag{2.29}$$

Note that for $i = 1, \dots, n$ and for $j = 1, \dots, p$, X_{ij}^c is a single point. Therefore, the matrix of centers in Equation (2.28) is a classical data matrix.

In this approach, classical PCA is performed on the centers matrix $\boldsymbol{X}^c$. The resultant k^{th} centers principal component is

$$\boldsymbol{Y}^c_{(k)} = \boldsymbol{X}^c \boldsymbol{\nu}^c_k$$

where $\boldsymbol{\nu}^c_k$ is the eigenvector corresponding to the k^{th} eigenvalue of the classical sample variance-covariance matrix of $\boldsymbol{X}^c$ of Equation (2.28). For observation $i = 1, 2, \ldots, n$, the k^{th} interval-valued principal component is reconstructed as follows. Let $Y^c_{ik} = [y^{lo}_{ik}, y^{up}_{ik}]$ be the interval-valued principal component. Then, its lower and upper endpoints are formed by

$$\begin{aligned} y^{lo}_{ik} &= \sum_{j \in J^-_c} (b_{ij} - \bar{X}_{(j)}) \nu^c_{kj} + \sum_{j \in J^+_c} (a_{ij} - \bar{X}_{(j)}) \nu^c_{kj} \\ y^{up}_{ik} &= \sum_{j \in J^-_c} (a_{ij} - \bar{X}_{(j)}) \nu^c_{kj} + \sum_{j \in J^+_c} (b_{ij} - \bar{X}_{(j)}) \nu^c_{kj} \end{aligned}$$

where $J^-_c = \{j | \nu^c_{kj} < 0\}$ and $J^+_c = \{j | \nu^c_{kj} \geq 0\}$.

2.3.2 Vertices method

Before principal component analysis is performed in this method, the data matrix $\boldsymbol{X}$ is first transformed into a matrix of vertices $\boldsymbol{X}^v$. Again, let $\boldsymbol{X}$ be the data matrix defined in Equation (2.27). Then, for $i = 1, 2, \ldots, n$,

$$\boldsymbol{X}_i = ([a_{i1}, b_{i1}], [a_{i2}, b_{i2}], \ldots, [a_{ip}, b_{ip}]).$$

Define the matrix of vertices for observation i as,

$$\boldsymbol{X}^v_i = \begin{bmatrix} a_{i1} & a_{i2} & \ldots & a_{ip} \\ a_{i1} & a_{i2} & \ldots & b_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ b_{i1} & b_{i2} & \ldots & a_{ip} \\ b_{i1} & b_{i2} & \ldots & b_{ip} \end{bmatrix}. \tag{2.30}$$

Again, note that each element of $\boldsymbol{X}^v_i$ is a single point, i.e., $\boldsymbol{X}^v_i$ is a classical matrix. For each p-variate observation $\boldsymbol{X}_i$, $\boldsymbol{X}^v_i$ is a $(2^{m_i} \times p)$ matrix where m_i is the number of variables in

observation i such that

$$a_{ij} \neq b_{ij}.$$

Each row of $\boldsymbol{X}_i^v$ represents the coordinate of a vertex of the hyper-rectangle formed by observation i in a p-dimensional space. An interval $[a_{ij}, b_{ij}]$ is said to be trivial if it reduces to a single point $a_{ij} = b_{ij}$. Thus, if $[a_{ij}, b_{ij}]$ is trivial for all $j = 1, 2, \ldots, p$, then $\boldsymbol{X}_i^v = [a_{i1}, a_{i2}, \ldots, a_{ip}]$ reduces to one single point in a p-dimensional space which is a classical data point.

The matrix of vertices for the full dataset $\boldsymbol{X}$ is

$$\boldsymbol{X}^v = \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{11} & b_{12} & \ldots & b_{1p} \end{bmatrix} \\ \begin{bmatrix} a_{21} & a_{22} & \ldots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{21} & b_{22} & \ldots & b_{2p} \end{bmatrix} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \begin{bmatrix} a_{n1} & a_{n2} & \ldots & a_{np} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \ldots & b_{np} \end{bmatrix} \end{bmatrix} \end{bmatrix}. \tag{2.31}$$

If $m_i = p$ for all i, $\boldsymbol{X}^v$ has dimension $n2^p \times p$.

Next, classical PCA is performed on $\boldsymbol{X}^v$ of Equation (2.31). The k^{th} principal component of $\boldsymbol{X}^v$ is

$$\boldsymbol{Y}_k^v = \boldsymbol{X}^v \boldsymbol{\nu}_k^v$$

where $\boldsymbol{\nu}_k^v$ is the eigenvector corresponding to the k^{th} eigenvalue of the sample variance-covariance matrix of $\boldsymbol{X}^v$. The k^{th} interval-valued principal component for observation i based on this method is $Y_{ik}^v = [y_{ik}^{lo}, y_{ik}^{up}]$ where

$$y_{ik}^{lo} = min_{\eta \in L_i}\{y_{\eta k}^v\} \tag{2.32}$$

$$y_{ik}^{up} = max_{\eta \in L_i}\{y_{\eta k}^v\} \tag{2.33}$$

where L_i is the set of rows in $\boldsymbol{X}^v$ that belongs to observation i. That is, for $N_i = 2^{m_i}$,

$$L_i = \left\{ \sum_{m=1}^{i-1} N_m + 1, \sum_{m=1}^{i-1} N_m + 2, \ldots, \sum_{m=1}^{i-1} N_m + N_i \right\}. \tag{2.34}$$

Equivalently, Equation (2.32) and Equation (2.33) can be computed by

$$\begin{aligned} y_{ik}^{lo} &= \sum_{j \in J_v^-} (b_{ij} - \bar{X}_{(j)}^v)\nu_{kj}^v + \sum_{j \in J_v^+} (a_{ij} - \bar{X}_{(j)}^v)\nu_{kj}^v \\ y_{ik}^{up} &= \sum_{j \in J_v^-} (a_{ij} - \bar{X}_{(j)}^v)\nu_{kj}^v + \sum_{j \in J_v^+} (b_{ij} - \bar{X}_{(j)}^v)\nu_{kj}^v \end{aligned}$$

where $\bar{X}_{(j)}^v$ is the mean of the j^{th} column of $\boldsymbol{X}^v$, $J_v^- = \{j | \nu_{kj}^v < 0\}$, and $J_v^+ = \{j | \nu_{kj}^v \geq 0\}$.

See Billard et al. (2007) for details and examples. For each observation $i = 1, 2, \ldots, n$, the interval $[y_{ik}^{lo}, y_{ik}^{up}]$ includes all possible values of $\boldsymbol{X}_i$ transformed by $\boldsymbol{\nu}_k^v$. Moreover, for $k, k' = 1, 2, \ldots, p$ and $k \neq k'$, the rectangle formed by two interval-valued principal components $[y_{ik}^{lo}, y_{ik}^{up}]$ and $[y_{ik'}^{lo}, y_{ik'}^{up}]$ is called the *maximum covering area rectangle* (MCAR). This method treats all vertices as independent observations.

2.3.3 SO-PCA: A MIXED STRATEGY

Lauro and Palumbo (2000) proposes a multiple stage approach based on 2 separate procedures, the symbolic-object PCA and the range-transformation PCA. A review of symbolic-object and range-transformation PCA methods follows.

1. **Symbolic-object PCA (SO-PCA)**

   The SO-PCA modifies the vertices method by introducing a boolean matrix $\boldsymbol{A}$ of dimension $n2^p \times n$ where the q^{th} element of the column vector $\boldsymbol{A}_i$ indicates if the q^{th} vertex of $\boldsymbol{X}^v$ as defined in Equation (2.31) belongs to the i^{th} observation. That is,

   $$A_{qi} = \begin{cases} 1, & q \in L_i, \\ 0, & q \notin L_i. \end{cases}$$

   Let $\boldsymbol{Z}^v$ be the standardized version of $\boldsymbol{X}^v$. In this method, the vector of coefficients $\boldsymbol{\nu}_k$ for the k^{th} principal component Y_k is the eigenvector corresponding to the k^{th}

eigenvalue of the matrix

$$\frac{1}{N}\boldsymbol{Z}^{v\prime}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Z}^{v}.$$

The interval-valued principal component $[y_{ik}^{l}, y_{ik}^{u}]$ is formed in a similar manner to its counterpart in the vertices method, i.e., Equations (2.32) and (2.33), respectively.

2. **Range transformation method (RT-PCA)**

   The second procedure proposed by Lauro and Palumbo (2000) uses only the range of the data intervals. This is the same as translating the observed hyper-rectangles so that the vertices closest to the origin are aligned at the origin. This method is used mainly to analyze the size and shape of the interval-valued observations. Let $\boldsymbol{X}_i^R$ be the p-variate vector whose elements represent the ranges of the i^{th} observation. That is,

$$\begin{aligned}\boldsymbol{X}_i^R &= (X_{i1}^R, X_{i2}^R, \ldots, X_{ip}^R) \\ &= ((b_{i1}-a_{i1}), (b_{i2}-a_{i2}), \ldots, (b_{ip}-a_{ip})). \qquad (2.35)\end{aligned}$$

   The range matrix of the full dataset is $\boldsymbol{X}^R = (\boldsymbol{X}_1^R, \boldsymbol{X}_2^R, \ldots, \boldsymbol{X}_n^R)'$. Classical PCA is then performed on $\boldsymbol{X}^R$. For ease of reference later, call the k^{th} principal component of $\boldsymbol{X}^R$, $\boldsymbol{T}_k$. If only the size and shape of the observed data are of interest, range transformation PCA can be used alone. Otherwise, it can be coupled with SO-PCA to form a multi-steps process as described in Lauro and Palumbo (2000).

The mixed strategy is based on three steps:

1. Perform RT-PCA on the data matrix $\boldsymbol{X}$ to extract the principal components matrix $\boldsymbol{T}$,
2. Transform $\boldsymbol{Z}^v$ into $\hat{\boldsymbol{Z}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Z}^v$, and
3. Perform classical PCA on $\boldsymbol{P_T}\hat{\boldsymbol{Z}}$ where $\boldsymbol{P_T}$ is a projection matrix of $\boldsymbol{T}$.

The results of this method depend on the choice of $\boldsymbol{P_T}$. The projection $\boldsymbol{P_T}$ is chosen to reflect the importance of size and shape of the observed hyper-rectangles.

2.3.4 Midpoints and radii method

Another extension of PCA to interval-valued data was proposed by Palumbo and Lauro (2003) which uses the centers and radii of the observed data. The idea for this method is based on interval algebra. A brief introduction of interval algebra follows. To learn more about interval algebra, see Moore (1966), Alefeld and Herzberger (1983), Neumaier (1990), and Kearfott and Kreinovich (1996). Let $[a, b]$ be an interval of real values, i.e.,

$$[a, b] = \{x | a \leq x \leq b\}.$$

The arithmetic operations on $[a, b]$ are defined as follows:

$$\begin{aligned}
[a, b] + [c, d] &= [a + c, b + d], \\
[a, b] - [c, d] &= [a - d, b - c], \\
[a, b] \times [c, d] &= [min(ac, ad, bc, bd), max(ac, ad, bc, bd)],
\end{aligned}$$

and finally, if $c = d \neq 0$ then the ratio of $[a, b]$ and $[c, d]$ is defined as

$$[a, b] \div [c, d] = [a, b] \times \left[\frac{1}{c}, \frac{1}{d}\right].$$

Now, the mean interval $\bar{X}^I_{(j)}$ of the j^{th} variable is defined as

$$\bar{X}^I_{(j)} = \frac{1}{n} \sum_{i=1}^{n} [a_{ij}, b_{ij}]$$

and the distance between X_{ij} and $X_{i'j}$ for $i, i' = 1, 2, \ldots, n$, is defined as

$$d(X_{ij}, X_{i'j}) = |X^c_{ij} - X^c_{i'j}| + |X^r_{ij} - X^r_{i'j}|$$

where the midpoint X^c_{ij} is as defined in Equation (2.29) and the radius $X^r_{ij} = \frac{1}{2}(b_{ij} - a_{ij})$. Note that $X^r_{ij} = \frac{1}{2} X^R_{ij}$ where X^R_{ij} is defined in Equation (2.35). The variance of $X_{(j)}$ is

subsequently defined as

$$\begin{aligned}
\sigma^2_{(j)} &= \frac{1}{n}\sum_{i=1}^{n}\left[d(X_{ij},\bar{X}_{(j)})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(X^c_{ij}-\bar{X}^c_{(j)})^2+2|X^c_{ij}-\bar{X}^c_{(j)}||X^r_{ij}-\bar{X}^r_{(j)}|+(X^r_{ij}-\bar{X}^r_{(j)})^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}(X^c_{ij}-\bar{X}^c_{(j)})^2+\frac{2}{n}\sum_{i=1}^{n}[|X^c_{ij}-\bar{X}^c_{(j)}||X^r_{ij}-\bar{X}^r_{(j)}|]+\frac{1}{n}\sum_{i=1}^{n}(X^r_{ij}-\bar{X}^r_{(j)})^2 \\
&= (\sigma^c_{(j)})^2+2\sigma^{cr}_{(j)}+(\sigma^r_{(j)})^2 \qquad (2.36)
\end{aligned}$$

where $(\sigma^c_{(j)})^2$ is the sample variance of the midpoints, $(\sigma^r_{(j)})^2$ is the sample variance of the radii, and $\sigma^{cr}_{(j)}$ is called the *inter-connection* between the midpoints and radii. Compared to the symbolic variance defined in Equation (2.4) and Equation (2.6), $(\sigma^c_{(j)})^2$ equals to the variance between variables, SSB/n. Thus, comparing the symbolic variance S^2 to $\sigma^2_{(j)}$ is equivalent to comparing SSW/n to $[(\sigma^r_{(j)})^2+2\sigma^{cr}_{(j)}]$. The relation between these two quantities depends on the data. The inter-connection, $\sigma^{cr}_{(j)}$, is always positive unless $\boldsymbol{X}^c_{(j)}$ and $\boldsymbol{X}^r_{(j)}$ are uncorrelated. In the event they are uncorrelated then, $S^2 \geq \sigma^2_{(j)}$ if the average variance within the ranges is larger than the variation between the ranges, i.e., if the data consist of wide intervals of approximately the same size. Otherwise, $S^2 \leq \sigma^2_{(j)}$. Generalizing Equation (2.36) to define the variance-covariance matrix $\boldsymbol{V}$ for the interval-valued data matrix $\boldsymbol{X}$ as,

$$\begin{aligned}
\boldsymbol{V} &= \frac{1}{n}(\boldsymbol{X}^{*c})'(\boldsymbol{X}^{*c})+\frac{1}{n}(\boldsymbol{X}^{*r})'(\boldsymbol{X}^{*r})+\frac{1}{n}[|(\boldsymbol{X}^{*c})'(\boldsymbol{X}^{*r})|+|(\boldsymbol{X}^{*r})'(\boldsymbol{X}^{*c})|] \\
&= \boldsymbol{V}^c+\boldsymbol{V}^r+2\boldsymbol{V}^{cr}
\end{aligned}$$

where the element of $\boldsymbol{X}^{*c}$ is $X^{*c}_{ij}=(X^c_{ij}-\bar{X}^c_{(j)})$ and the element of $\boldsymbol{X}^{*r}$ is $X^{*r}_{ij}=(X^r_{ij}-\bar{X}^r_{(j)})$. Hence, the correlation matrix $\boldsymbol{R}$ is

$$\begin{aligned}
\boldsymbol{R} &= (\boldsymbol{Z}^c)'(\boldsymbol{Z}^c)+(\boldsymbol{Z}^r)'(\boldsymbol{Z}^r)+|(\boldsymbol{Z}^c)'(\boldsymbol{Z}^r)|+|(\boldsymbol{Z}^r)'(\boldsymbol{Z}^c)| \\
&= \boldsymbol{R}^c+\boldsymbol{R}^r+2\boldsymbol{R}^{cr}
\end{aligned}$$

where the elements of $\boldsymbol{Z}^c$ and $\boldsymbol{Z}^r$ are $Z^c_{ij}=\frac{(X^c_{ij}-\bar{X}^c_{(j)})}{\sigma_{(j)}}$ and $Z^r_{ij}=\frac{(X^r_{ij}-\bar{X}^r_{(j)})}{\sigma_{(j)}}$, respectively.

To understand the positions or the shapes and sizes of the data independently, partial PCA can be performed on the midpoints matrix $\boldsymbol{X}^c$ or the radii matrix $\boldsymbol{X}^r$, respectively.

Let $(\lambda_k^c, \boldsymbol{\nu}_k^c)$ be the k^{th} eigenvalue-eigenvector pair of $\boldsymbol{R}^c$ and similarly $(\lambda_k^r, \boldsymbol{\nu}_k^r)$ be the k^{th} eigenvalue-eigenvector pair of $\boldsymbol{R}^r$. Then, the k^{th} principal component of the i^{th} observation for the midpoints and the radii are, respectively,

$$y_{ik}^c = \boldsymbol{Z}_i^c \boldsymbol{\nu}_k^c,$$
$$y_{ik}^r = \boldsymbol{Z}_i^r \boldsymbol{\nu}_k^r.$$

However, to account for all three components of the variance structure in Equation (2.36), Palumbo and Lauro (2003) proposes constructing the interval-valued principal components by superimposing the principal components of the radii on the principal components of the midpoints and then rotating the radii proportionally to the *inter-connection*, $\sigma_{(j)}^{cr}$. The authors suggested *Procrustes* rotation (Gower (1975)) be used to maximize the connection between the midpoints and radii. The rotation is defined by a rotation matrix $\boldsymbol{A}$ that minimizes the trace,

$$trace\left((\boldsymbol{X}^{*c} - \boldsymbol{X}^{*r}\boldsymbol{A})(\boldsymbol{X}^{*c} - \boldsymbol{X}^{*r}\boldsymbol{A})'\right). \tag{2.37}$$

Since Equation (2.37) equals to

$$tr\left((\boldsymbol{X}^{*c})(\boldsymbol{X}^{*c})'\right) + tr\left((\boldsymbol{X}^{*r})(\boldsymbol{X}^{*r})'\right) - 2tr\left((\boldsymbol{X}^{*c})'\boldsymbol{X}^{*r}\boldsymbol{A}\right),$$

this is equivalent to finding $\boldsymbol{A}$ that maximizes

$$trace\left((\boldsymbol{X}^{*c})'\boldsymbol{X}^{*r}\boldsymbol{A}\right). \tag{2.38}$$

Mardia et al. (1979) shows that the solution to Equation (2.38) is $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{P}'$ where $\boldsymbol{Q}$ and $\boldsymbol{P}$ are the solution to the singular value decomposition

$$(\boldsymbol{X}^{*c})'\boldsymbol{X}^{*r} = \boldsymbol{P}\boldsymbol{\Lambda}^{cr}\boldsymbol{Q}.$$

As a result, the rotated *radii* are then

$$y_{ik}^{cr} = \boldsymbol{X}^{*r}{}_i \boldsymbol{\nu}_k^r \boldsymbol{a}_i$$

where $\boldsymbol{a}_i$ is the i^{th} vector of $\boldsymbol{A}$. The k^{th} interval-valued principal component for the i^{th} observation obtained by this method is

$$[y_{ik}^{lo}, y_{ik}^{up}] = [y_{ik}^{c} - y_{ik}^{cr}, y_{ik}^{c} + y_{ik}^{cr}].$$

Although this PCA approach to interval-valued data accounts for all the variability in the data, the interval arithmetic used in this method only works for very narrow intervals. This poses a limitation on the applicability of this method.

2.3.5 Interval-algebra method

Another approach using interval algebra was proposed by Gioia and Lauro (2006). This method is referred to as interval principal component analysis (IPCA). Given a data matrix $\boldsymbol{X}$ as defined in Equation (2.26), the problem of finding the principal components of $\boldsymbol{X}$ as seen in Section 2.2 is reduced to finding the eigen system of the sample variance-covariance matrix $\boldsymbol{S}$. This method uses the sample interval-valued variance-covariance matrix defined as

$$\boldsymbol{S}^I = (\boldsymbol{Z}^I)'(\boldsymbol{Z}^I) = \{\boldsymbol{U}\boldsymbol{W} | \boldsymbol{U} \in (\boldsymbol{Z}^I)', \boldsymbol{W} \in \boldsymbol{Z}^I\}$$

where $\boldsymbol{Z}^I$ is the standardized version of $\boldsymbol{X}$. The k^{th} interval eigenvalue and eigenvector of $\boldsymbol{S}^I$ are

$$\begin{aligned} \lambda_k^I &= \{\lambda | \boldsymbol{S}\boldsymbol{\nu} = \lambda\boldsymbol{\nu}, \forall \boldsymbol{S} \in \boldsymbol{S}^I\}, \\ \boldsymbol{\nu}_k^I &= \{\boldsymbol{\nu} | \boldsymbol{S}\boldsymbol{\nu} = \lambda\boldsymbol{\nu}, \forall \boldsymbol{S} \in \boldsymbol{S}^I\}. \end{aligned}$$

More explicitly, λ_k^I is the set of all k^{th} eigenvalues of all matrices $\boldsymbol{S} \in \boldsymbol{S}^I$ and $\boldsymbol{\nu}_k^I$ is the set of all eigenvectors corresponding to λ_k^I. As a consequence of the interval solutions, the principal components computed by this method are much larger than appropriate. Gioia and Lauro (2006) also proposes a fix for this problem by including only the eigen solutions to

$$\boldsymbol{\Gamma}^I = \{\boldsymbol{Z}'\boldsymbol{Z} | \boldsymbol{Z} \in (\boldsymbol{Z}^I)\}$$

instead of $\boldsymbol{S}^I$. Since $\boldsymbol{\Gamma}^I \subseteq \boldsymbol{S}^I$, the resulting interval-valued principal components will be narrower. However, similar to the midpoints-radii approach, the interval solutions only work with small intervals whose ratio between radius and coordinate of the center is approximately 2-3%. Therefore, the applicability of this method is also very limited.

2.4 References

Alefeld, G. and Herzberger, J. (1983). *Introduction to Interval Computations*. Academic Press, New York.

Anderson, T.W (1963). Asymptotic Theory for Principal Components Analysis. *Annals of Mathematical Statistics*, **34**, 122-148.

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

Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 106-124.

Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-valued Data. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, G. Cucumel, P. Bertrand and F. de Carvalho). Springer-Verlag, Berlin, 3-12.

Billard, L., Chouakia-Douzal, A., and Diday, E. (2007). Symbolic Principal Component Analysis for Interval-valued Observations. *Journal of the American Statistical Association*, pending acceptance.

Billard, L. and Diday E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, **98**, 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, New York.

Bock, H.-H. and Diday, E. (eds.) (2000). *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.

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*, **XLV** (3). 5-24.

Chouakria, A., Cazes, P., and Diday, E. (2000). Symbolic Principal Component Analysis. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 200-212.

Diday, E. (1987). Introduction à l'Approache Symbolique en Analyse des Données. *Premières Journées Symbolic-Numérique*, CEREMADE, Université Paris, 21-56.

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

Girshick, M.A. (1939). On the Sampling Theory of Roots of Determinantal Equations. *Annals of Mathematical Statistics*, **10**, 203-224.

Gower, J.C. (1975). Generalized Procrustes Analysis. *Psychometrika*, **40**, 33-51.

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

Jolliffe, I.T. (2004). *Principal Component Analysis*, 2nd edition. Springer, New York.

Kearfott, R.B. and Kreinovich, V. (eds.) (1996). *Applications of Interval Computations*. Kluwer Academic Publishers.

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

Lauro, C. and Palumbo, F. (2003). Some Results and New Perspectives in Principal Component Analysis for Interval Data. *CLADAG '03 Book of Short Papers* 237-244.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, New York.

Moore, R. (1966). *Interval Analysis*. Prentice-Hall, New Jersey.

Neumaier, A. (1990). *Interval Methods for Systems of Equations*. Cambridge University Press, Cambridge.

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

Chapter 3

Principal Component Analysis for Interval-Valued Data

Section 2.3 of Chapter 2 describes five extensions of classical principal component analysis (PCA) to interval-valued data. However, all methods proposed up to this point have drawbacks. The centers, the vertices, and the symbolic object methods only account for part of the variance of interval-valued observations whereas the midpoints-radii and the interval algebra methods only work for very narrow intervals. In this chapter, we propose a method of PCA that takes into account the total variance of interval-valued observations, and that works for intervals of all sizes. We also propose a method to construct interval-valued observations in the principal components space to reflect their true internal structure.

In classical PCA, an observation remains a point in a principal components space. From Section 2.2, the k^{th} principal component for observation i is simply the inner product of the vector of coefficients for the k^{th} principal component and the data vector representing observation i. However, a symbolic observation has an internal structure that does not exist in classical data. The internal structure of a symbolic observation depends on its data type. For example, a p-variate interval-valued observation is represented by a hyper-rectangle in a p-dimensional space. The shape of a hyper-rectangle in a principal components space may be different from its original shape in the sample space. Therefore, an interval-valued observation must be reconstructed in a principal components space to reflect its internal structure. Concepts of the proposed method are described in Section 3.2. A detailed algorithm to compute the coefficients as well as the principal components is presented in Section 3.3. Two applications using real datasets illustrate our method in Section 3.4.

3.1 Preliminaries

Section 2.1.2 describes an interval-valued variable and gives the derivation of its mean and its variance-covariance. Some notation and results of interval-valued variable necessary for the development of our proposed method are restated in this section without further details. Refer to Section 2.1.2 for more information. For an even more extensive treatment of interval-valued data, refer to Bertrand and Goupil (2000) and Billard and Diday (2003, 2006).

An interval-valued data matrix as defined in Equation (2.27) has the following form,

$$\boldsymbol{X} = \begin{bmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & \dots & [a_{1p}, b_{1p}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & \dots & [a_{2p}, b_{2p}] \\ \vdots & \vdots & \vdots & \vdots \\ [a_{n1}, b_{n1}] & [a_{n2}, b_{n2}] & \dots & [a_{np}, b_{np}] \end{bmatrix}$$

where $a_{ij} \leq b_{ij}$ for all $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$. The empirical density function of a point $W \in \xi_{ij} = [a_{ij}, b_{ij}]$ given in Equation (2.2) is

$$f_W(\xi) = \frac{1}{n} \sum_{i:\xi \in \xi_{ij}} \left(\frac{1}{b_{ij} - a_{ij}}\right).$$

The symbolic sample mean and the symbolic sample variance of variable W as defined in Equations (2.3) and (2.4) are, respectively,

$$\bar{W} = \frac{1}{2n} \sum_{i=1}^{n} (a_{ij} + b_{ij})$$

and

$$S^2 = \frac{1}{3n} \sum_{i=1}^{n} (a_{ij}^2 + a_{ij} b_{ij} + b_{ij}^2) - \frac{1}{4n^2} \left[\sum_{i=1}^{n} a_{ij} + b_{ij} \right]^2.$$

By extending Equation (2.4) to the bivariate case, the empirical symbolic covariance for $W_{(j)}$ in $X_{(j)}$ and $W_{(j')}$ in $X_{(j')}$ where $j, j' = 1, 2, \dots, p$ is given in Equation (2.5) as

$$S_{jj'} = \frac{1}{3n} \sum_{i=1}^{n} G_{ij} G_{ij'} [Q_{ij} Q_{ij'}]^{1/2}$$

where

$$\begin{aligned} Q_{ij} &= (a_{ij} - \bar{W}_{(j)})^2 + (a_{ij} - \bar{W}_{(j)})(b_{ij} - \bar{W}_{(j)}) + (b_{ij} - \bar{W}_{(j)})^2, \\ G_{ij} &= \begin{cases} -1, & \bar{W}_{ij} \leq \bar{W}_{(j)}, \\ 1, & \bar{W}_{ij} > \bar{W}_{(j)}, \end{cases} \end{aligned}$$

and $\bar{W}_{ij} = (a_{ij} + b_{ij})/2$. The sample symbolic variance-covariance resulting from S^2 and $S_{jj'}$ for $j, j' = 1, 2, \ldots, p$ as defined in Equations (2.4) and (2.5) is,

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1p} \\ S_{21} & S_{22} & \ldots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \ldots & S_{pp} \end{bmatrix}. \tag{3.1}$$

Billard (2007) shows that the total sum of squares of an interval-valued data matrix $\boldsymbol{X}$ can be decomposed into the sum of the internal variation and the external variation called the between sum of squares in Equation (2.7) and the within sum of squares in Equation (2.8). Formally,

$$nS^2 = SST = SSB + SSW$$

When $j \neq j'$ the total sum of products is the sum of the internal and the external sum of products and is given by,

$$nS_{jj'} = SPT_{jj'} = SPB_{jj'} + SPW_{jj'}$$

where $SPB_{jj'}$ is defined in Equation (2.11) and $SPW_{jj'}$ is defined in Equation (2.12).

Moreover, given an interval-valued data matrix $\boldsymbol{X}$ of Equation (2.27), $\boldsymbol{X}$ can be expressed in terms of its vertices as a matrix of vertices $\boldsymbol{X}^v$ defined in Equation (2.31). That is, for $i = 1, 2, \ldots, n$, the matrix of vertices for observation i is the $(2^{m_i} \times p)$ matrix $\boldsymbol{X}_i^v$ of Equation (2.30) where m_i is the number of variables in observation i such that $a_{ij} \neq b_{ij}$. Each row of $\boldsymbol{X}_i^v$ represents the coordinate of a vertex of the hyper-rectangle formed by observation i in the p-dimensional sample space. An interval $[a_{ij}, b_{ij}]$ is said to be trivial if it reduces to

a single point $a_{ij} = b_{ij}$. Therefore, if $[a_{ij}, b_{ij}]$ is nontrivial for all $j = 1, 2, \ldots, p$, then $\boldsymbol{X}_i^v$ consists of 2^p rows and it is given by,

$$\boldsymbol{X}_i^v = \begin{bmatrix} a_{i1} & a_{i2} & \ldots & a_{ip} \\ a_{i1} & a_{i2} & \ldots & b_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ b_{i1} & b_{i2} & \ldots & a_{ip} \\ b_{i1} & b_{i2} & \ldots & b_{ip} \end{bmatrix}. \tag{3.2}$$

If, for example, the interval $[a_{ip}, b_{ip}]$ is trivial, then the last two lines of $\boldsymbol{X}_i^v$ in Equation (3.2) reduce to one line only. If $[a_{ij}, b_{ij}]$ is trivial for all $j = 1, 2, \ldots, p$, then $\boldsymbol{X}_i^v = [a_{i1}, a_{i2}, \ldots, a_{ip}]$ reduces to one single point in a p-dimensional space which is a classical data point.

The matrix of vertices for the full dataset $\boldsymbol{X}$ is defined in Equation (2.31) as

$$\boldsymbol{X}^v = \begin{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{11} & b_{12} & \ldots & b_{1p} \end{bmatrix} \\ \begin{bmatrix} a_{21} & a_{22} & \ldots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{21} & b_{22} & \ldots & b_{2p} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} a_{n1} & a_{n2} & \ldots & a_{np} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \ldots & b_{np} \end{bmatrix} \end{bmatrix}.$$

3.2 Methodology

This section lays the foundation for our proposed PCA method for interval-valued observations. It explains the theoretical framework upon which our proposed method is built. Detailed descriptions to construct the principal components based on this framework are

given in Section 3.3 along with the algorithm. This section is divided into two subsections. The first, Subsection 3.2.1, describes our approach to finding coefficients for the principal components of an interval-valued dataset. The second, Subsection 3.2.2, explores the structure of interval-valued observations in a principal components space and provides the theoretical basis for our proposed construction of interval-valued observations in the principal components space. In this subsection, we further propose two representations of the principal components for this type of data.

3.2.1 Finding the coefficients

Let $\boldsymbol{X}$ represent the data matrix defined in Equation (2.27). Based on the classical PCA methodology reviewed in Section 2.2, the coefficients of the uncorrelated linear combinations of $\boldsymbol{X}$ with maximum variances are the eigenvectors of the sample symbolic variance-covariance matrix. For an interval-valued data matrix $\boldsymbol{X}$, the sample variance-covariance is the matrix $\boldsymbol{S}$ of Equation (3.1). Let $\hat{\lambda}_1^S > \hat{\lambda}_2^S > \ldots > \hat{\lambda}_p^S$ be the eigenvalues of $\boldsymbol{S}$ and $\boldsymbol{\nu}_1^S, \boldsymbol{\nu}_2^S, \ldots, \boldsymbol{\nu}_p^S$ be their corresponding eigenvectors. By analogy, the vector of coefficients, also called the loadings, for the k^{th} principal component of $\boldsymbol{X}$ is $\boldsymbol{\nu}_k^S$.

Analogous to the classical PCA, the magnitude of the j^{th} element of $\boldsymbol{\nu}_k^S$, denoted by ν_{jk}^S, indicates the contribution of variable $X_{(j)}$ to the principal component $Y_{(k)}$. The measure of correlation between an individual $X_{(j)}$ and $Y_{(k)}$ is given by

$$\rho_{X_{(j)},Y_{(k)}} = \frac{\nu_{jk}^S \sqrt{\hat{\lambda}_k^S}}{\sqrt{S_{jj}}}. \tag{3.3}$$

Moreover, assume $\boldsymbol{X}_i, i = 1, \ldots, n$, is a random sample and n is large. Let $W_{(j)}$ be a point from $X_{(j)}$ for $j = 1, 2, \ldots, p$. In addition, assume $\boldsymbol{W} = (W_{(1)}, W_{(2)}, \ldots, W_{(p)})$ is from a normal population and the eigenvalues of the population covariance matrix $\boldsymbol{\Sigma}$ are distinct and positive. Then, the asymtotic results regarding $\hat{\lambda}_k$ and $\boldsymbol{\nu}_k$ stated in Section 2.2 carry through for $\hat{\lambda}_k^S$ and $\boldsymbol{\nu}_k^S$ for all $k = 1, 2, \ldots, p$.

In situations where PCA based on the sample correlation is more appropriate, coefficients for the k^{th} principal component of $\boldsymbol{X}$ are then the eigenvector corresponding to the k^{th} eigenvalue of the sample correlation matrix $\boldsymbol{R}$. The jj' element of $\boldsymbol{R}$, denoted by $\rho_{jj'}$, is

$$\rho_{jj'} = \frac{S_{jj'}}{\sqrt{S_{jj}S_{j'j'}}}. \tag{3.4}$$

Measures of contribution for the principal components based on the correlation matrix mirror the results for the principal components based on the sample covariance matrix.

3.2.2 Constructing the principal components

Having determined the coefficients of the principal components, we next reconstruct the observations in the symbolic principal components space. This section is divided into two parts. In the first part, we discuss the structure of an interval-valued observation in a principal components space and propose a geometric representation of the observations in this space. This representation can be used for visualization and data exploration. Since it is not possible at the present time to perform statistical analysis of geometric objects, we propose another representation which gives the principal components numerical values for further analysis. In part two, we propose a method to construct histogram-valued principal components. Again, this section only provides the concept of our proposal. Details of the construction are presented in Section 3.3 along with the algorithm.

In current approaches to PCA for interval-valued data, a principal component for observation i is constructed as an interval formed by the minimum and the maximum transformed values of all vertices belonging to observation i. Interval-valued representation of the principal components presents two drawbacks which are explained in the following paragraphs.

First, in a plot of principal component k_1 versus principal component k_2 (PCk_1 $\times$ PCk_2), an observation is visually represented by a rectangle bounded by the lower and the upper endpoints of PCk_1 and the lower and the upper endpoints of PCk_2 where the lower (upper) endpoint is the minimum (maximum) PCk_1 and PCk_2, respectively, values of all vertices belonging to that observation. This rectangle is called the maximum covering area rectangle

(MCAR). This rectangle covers an area larger than the projection of observation i onto the PCk_1 $\times$ PCk_2 plane. When all observations are included in a PCk_1 $\times$ PCk_2 plot, overage from these rectangles creates unnecessary overlap among observations. As a result, it can be difficult to visually distinguish clusters of observations.

Secondly, an assumption for interval-valued data is that all values between the interval endpoints are uniformly distributed. However, values of the principal component for an observation are not necessarily uniformly distributed between the minimum and maximum transformed values. Uniformity only occurs as a special case when at least one of the principal components is completely correlated with a variable in the dataset. A principal component interval may not reflect the true distribution of values within the principal component. Therefore, when interval-valued principal components are used for further analysis, use of the lower and upper endpoints may lead to a wrong conclusion about the data.

In this section, we propose representations of principal components which are an improvement over the interval-valued principal components constructed in the centers and vertices methods proposed by Cazes et al. (1997) and Chouakria et al. (2000). First, we propose constructing the true structure of the observations in a principal components space. The geometric representation of the observations resulting from their true structure do not produce overlaps which do not exist in the observations. As a result, observations that belong to different groups can be visually identified more easily. Secondly, we propose constructing histograms as another representation of the principal components. Values within a principal component are not necessarily uniformly distributed. Unlike an interval-valued variable, a histogram-valued variable allows variability in the relative frequencies of values within the histogram endpoints. Therefore, a histogram-valued principal component can be constructed to reflect most of the internal variation of an observation in a principal components space. Hence, results from statistical analysis using histogram-valued principal components reflect most of the internal structure of interval-valued data. Details of the proposed methods are given below.

Geometric representation of interval-valued observations in a principal components space

In the original sample space, an interval-valued observation is represented by a hyper-rectangle which is a convex hull of its vertices. Let H_i denote the hyper-rectangle representing observation i in the sample space. Before describing the structure of the linearly transformed H_i in a principal components space, it is necessary to define some geometric terms which will be used in the remainder of this section. Ziegler (1995) gives the following definitions.

Definition 3.2.1. A point set $K \subseteq \boldsymbol{R}^d$ is *convex* if, with any two points $\boldsymbol{x}, \boldsymbol{y} \in K$, it also contains the straight line segment $[\boldsymbol{x}, \boldsymbol{y}] = \{\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} | 0 \leq \alpha \leq 1\}$ between them.

Definition 3.2.2. For any $K \subseteq \boldsymbol{R}^d$, the smallest convex set containing K, called the *convex hull* of K, can be constructed as the intersection of all convex sets that contain K:

$$conv(K) := \bigcap \{K' \subseteq \boldsymbol{R}^d | K \subseteq K',\ K' \text{ convex}\}.$$

Equivalently, H_i is also called a polytope in $\boldsymbol{R}^d$. There are two mathematically equivalent definitions of polytope and they are stated in the following.

Definition 3.2.3. A *$\mathcal{V}$-polytope* is a convex hull of a finite set of points in some $\boldsymbol{R}^d$.

An *$\mathcal{H}$-polyhedron* is an intersection of finitely many closed halfspaces in some $\boldsymbol{R}^d$. An *$\mathcal{H}$-polytope* is an $\mathcal{H}$-polyhedron that is *bounded* in the sense that it does not contain a ray $\{\boldsymbol{x} + t\boldsymbol{y} | t \geq 0\}$ for any $\boldsymbol{y} \neq \boldsymbol{0}$ where a ray is a line originated at one point and extended infinitely in one direction.

A *polytope* is a point set $P \subseteq \boldsymbol{R}^d$ which can be presented either as a $\mathcal{V}$-polytope or as an $\mathcal{H}$-polytope.

For a proof of the equivalence between a $\mathcal{V}$-polytope and a $\mathcal{H}$-polytope, refer to Ziegler (1995).

With a formal definition of a convex set, we can now state and prove the following theorem.

Theorem 3.2.4. *A linear transformation of a convex set is convex.*

Proof. Let $P_d \subseteq \boldsymbol{R}^d$ be a convex set. Let $T : \boldsymbol{R}^d \to \boldsymbol{R}^e$ be a linear transformation. Let $P_e = T(P_d)$ be the image set of P_d, i.e., $P_e = \{\boldsymbol{\beta} | \boldsymbol{\beta} = T(\boldsymbol{b}) \ \forall \boldsymbol{b} \in P_d\}$. Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be two points in P_e. Then, $\boldsymbol{\beta}_1 = T(\boldsymbol{b}_1)$ and $\boldsymbol{\beta}_2 = T(\boldsymbol{b}_2)$ for some $\boldsymbol{b}_1, \boldsymbol{b}_2 \in P_d$, respectively. Then, for $0 \leq \alpha \leq 1$,

$$\begin{aligned} \alpha\boldsymbol{\beta}_1 + (1-\alpha)\boldsymbol{\beta}_2 &= \alpha T(\boldsymbol{b}_1) + (1-\alpha)T(\boldsymbol{b}_2) \\ &= T(\alpha\boldsymbol{b}_1 + (1-\alpha)\boldsymbol{b}_2). \end{aligned}$$

Since P_d is a convex set and $\boldsymbol{b}_1, \boldsymbol{b}_2 \in P_d$, then $\boldsymbol{b}_3 = \alpha\boldsymbol{b}_1 + (1-\alpha)\boldsymbol{b}_2$ is in P_d. Hence, $T(\boldsymbol{b}_3) \in P_e$. Therefore, P_e is convex. □

To establish a framework for our proposed method, we need to define formally the interior and the boundary of a convex set. Davidson and Donsig (2002) gives the following definitions.

Definition 3.2.5. The *interior* of a convex set K, denoted by $int(K)$, is the largest open set contained inside K. That is, a point $a \in int(K)$ if and only if there exist an $\epsilon > 0$ such that $B_\epsilon(a) \subset K$ where $B_\epsilon(a) = \{b | |b - a| < \epsilon\}$.

The *boundary* of K, denoted by $bd(K)$, is defined as the set $\bar{K} \setminus int(K)$ where $\bar{K}$, the closure of K, is the smallest closed set containing K, and where the notation $A \setminus B$ is the set of points in A not containing the points in B.

The theorem that plays the most crucial role in our proposed construction of the principal components involves bijective linear mapping of a convex set. Davidson and Donsig (2000) gives the following definition of a bijection.

Definition 3.2.6. A linear operator T maps a set A *onto* a set B or is *surjective* if for each $b \in B$ there is at least one $a \in A$ such that $T(a) = b$.

A linear operator T that maps a set A to a set B is *one-to-one* or *injective* if $T(a_1) = T(a_2)$ implies that $a_1 = a_2$ for all $a_1, a_2 \in A$.

A linear operator T that is both one-to-one and onto is *bijective*.

Next, we need the following relationship between the interior of a convex set and its linear transformation to prove our theorem.

Theorem 3.2.7. *If K is a convex set in $\boldsymbol{R}^d$ and T is a linear map from $\boldsymbol{R}^d$ to $\boldsymbol{R}^e$, then $T(int(K)) = int(T(K))$.*

Proof. See Davidson and Donsig (2002) for a proof of this theorem. □

Using the result of Theorem 3.2.7, we can now prove the following theorem.

Theorem 3.2.8. *If K is a convex set in $\boldsymbol{R}^d$ and $T : \boldsymbol{R}^d \to \boldsymbol{R}^d$ is a bijective linear operator, then $T(bd(K)) = bd(T(K))$.*

Proof. First, we will show that $T(bd(K)) \subset bd(T(K))$. Let $\boldsymbol{b} \in bd(K)$ and let $\boldsymbol{\beta} = T(\boldsymbol{b})$. Then, $\boldsymbol{\beta} \in T(bd(K)))$. For contradiction, assume that $\boldsymbol{\beta} \in int(T(K))$. By Theorem 3.2.7, $\boldsymbol{\beta} \in T(int(K))$. That is, there must exist some $\boldsymbol{b}_1 \in int(K)$ such that $\boldsymbol{\beta} = T(\boldsymbol{b}_1)$. Since, T is one-to-one, $\boldsymbol{b}_1 = \boldsymbol{b}$ but $\boldsymbol{b} \in bd(K)$ so $\boldsymbol{b} \notin int(K)$. Therefore, by contradiction $\boldsymbol{\beta} \in bd(T(K))$. Hence,

$$T(bd(K)) \subset bd(T(K)). \tag{3.5}$$

Next, we will show that $bd(T(K)) \subset T(bd(K))$. Since T is bijective, there must exist a bijective linear operatior T^{-1} such that $T(T^{-1}) = T^{-1}(T) = I$ where I is the operator that maps an element to itself. Let $\boldsymbol{\beta} \in bd(T(K))$ and let $\boldsymbol{b} = T^{-1}(\boldsymbol{\beta})$, i.e., $\boldsymbol{b} \in T^{-1}(bd(T(K)))$. Again for contradiction, assume that $\boldsymbol{b} \in int(K)$. Since,

$$int(T^{-1}(T(K)) = T^{-1}(int(T(K))) \quad (\text{by Theorem } 3.2.7),$$

$\boldsymbol{b} \in T^{-1}(int(T(K)))$. That is, there must exist some $\boldsymbol{\beta}_1 \in int(T(K))$ such that $\boldsymbol{b} = T^{-1}(\boldsymbol{\beta}_1)$. Since T^{-1} is bijective, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$. However, $\boldsymbol{\beta} \in bd(T(K)) \Rightarrow \boldsymbol{\beta} \notin int(T(K))$. Therefore, by contradiction $\boldsymbol{b} \in bd(K)$. Since $\boldsymbol{b} = T^{-1}(\boldsymbol{\beta}) \Leftrightarrow \boldsymbol{\beta} = T(\boldsymbol{b})$, then $\boldsymbol{\beta} \in T(bd(K))$. Therefore,

$$bd(T(K)) \subset T(bd(K)). \tag{3.6}$$

From Equation (3.5) and Equation (3.6), we conclude that

$$T(bd(K)) = bd(T(K)).$$

□

Since observation i in a principal components space is a linear transformation of H_i from the original sample space, let P_i denote the polytope representing observation i in the principal component space. Then, $P_i = T(H_i)$. From Theorem 3.2.8, P_i is bounded by the transformed boundary of H_i. The boundary of H_i is formed by its vertices. Therefore, the boundary of P_i can be reconstructed from the transformed vertices of H_i.

Therefore, let $\boldsymbol{X}_i^v$ of Equation (2.30) be the matrix of vertices of H_i and let $\boldsymbol{Y}_i^v$ be the $(N_i \times p)$ matrix of transformed vertices of observation i in the principal components space which is the space spanned by eigenvectors $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_p$. Then,

$$\boldsymbol{Y}_i^v = (\boldsymbol{X}_i^v)\boldsymbol{\nu}^S \tag{3.7}$$

is the matrix of vertices of P_i where $\boldsymbol{\nu}^S = [\boldsymbol{\nu}_1^S, \boldsymbol{\nu}_2^S, \ldots, \boldsymbol{\nu}_p^S]$. Each row of $\boldsymbol{X}_i^v$ is the coordinate of a vertex of the hyper-rectangle H_i and each row of $\boldsymbol{Y}_i^v$ is the coordinate of a vertex of the polytope P_i. Moreover, the vertex of P_i represented by row r_v in matrix $\boldsymbol{Y}_i^v$ is the transformed vertex of the vertex of H_i represented by row r_v in matrix $\boldsymbol{X}_i^v$. That is, there exists a one-to-one correspondence between the rows of $\boldsymbol{X}_i^v$ and the rows of $\boldsymbol{Y}_i^v$. Hence, the edges of P_i can be reconstructed by reconnecting the rows of $\boldsymbol{Y}_i^v$ which were connected in $\boldsymbol{X}_i^v$.

Note that an edge of a d-dimensional hyper-rectangle H_i is a line connecting two vertices whose coordinates have $d-1$ identical elements. That is, let $\boldsymbol{b}_1 = (b_{11}, b_{12}, \ldots, b_{1d})$ and $\boldsymbol{b}_2 = (b_{21}, b_{22}, \ldots, b_{2d})$ be two vertices of H_i. Then, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ form an edge if and only if $b_{1k} \neq b_{2k}$ at exactly one k for $k = 1, 2, \ldots, d$. As a consequence, each vertex of H_i is connected to d other vertices.

Now we use this fact to create a matrix $\boldsymbol{C}_i$ to store indices of the vertices of hyper-rectangle H_i that are connected. Matrix $\boldsymbol{C}_i$ will be used to connect vertices of polytope

P_i in the algorithm of Section 3.3. Let $\boldsymbol{C}_i$ be an $(N_i \times (d+1))$ matrix where N_i is the number of vertices of hyper-rectangle H_i. The first column of $\boldsymbol{C}_i$ consists of numbers 1 through N_i in ascending order. If we identify a vertex of H_i by the row number of the row of $\boldsymbol{X}_i^v$ representing that vertex, then the first column of $\boldsymbol{C}_i$ consists of the vertex number in ascending order. The last d columns of $\boldsymbol{C}_i$ keep indices of vertices connected to the vertices of the first column. More specifically, row r_v of the matrix $\boldsymbol{C}_i$ consists of $d+1$ elements. The first element identifies the vertex number, the $2^{nd}, 3^{rd}, \ldots, (d+1)^{th}$ elements are index of d vertices connected to vertex r_v of H_i.

In addition, note that there are many ways to construct a matrix of vertices for an observation X_i, and that $\boldsymbol{C}_i$ is dependent on the construction of $\boldsymbol{X}_i$. Without loss of generality, assume $d = p$, i.e., assume H_i has full dimension p. The matrix $\boldsymbol{X}_i^v$ of Equation (2.30) is constructed by permuting the variables in ascending order starting from variable p working back to variable 1. Since $\boldsymbol{X}_i^v$ was constructed in the same manner for all $i = 1, 2, \ldots, n$, $\boldsymbol{C}_i$ are identical for all i. That is, we can write $\boldsymbol{C}_i = \boldsymbol{C}, \forall i$. Examples to illustrate matrix $\boldsymbol{C}$ for the case of $p = 2$ and the case of $p = 3$ follow.

For $p = 2$, let the vector of observed intervals be $\boldsymbol{X}_i = ([a_{i1}, b_{i1}][a_{i2}, b_{i2}])$. If $a_{ij} < b_{ij}$ for all $j = 1, 2$, then the rectangle H_i representing $\boldsymbol{X}_i$ has four vertices. The matrix of vertices based on Equation (2.30) is

$$\boldsymbol{X}_i^v = \begin{bmatrix} a_{i1} & a_{i2} \\ a_{i1} & b_{i2} \\ b_{i1} & a_{i2} \\ b_{i1} & b_{i2} \end{bmatrix}. \tag{3.8}$$

Therefore, Vertex 1 has coordinate (a_{i1}, a_{i2}), Vertex 2 has coordinate (a_{i1}, b_{i2}), Vertex 3 has coordinate (b_{i1}, a_{i2}), and Vertex 4 has coordinate (b_{i1}, b_{i2}). Figure 3.1 shows the rectangle H_i representing $\boldsymbol{X}_i$ with the vertices labeled according to its row position in $\boldsymbol{X}_i^v$ of Equation (3.8). Check that the vertices are connected as indicated in matrix $\boldsymbol{C}$ as follows,

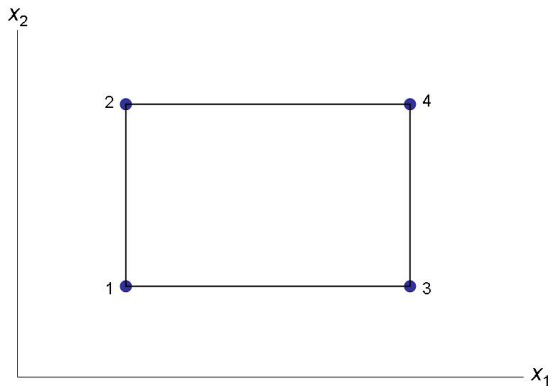

Figure 3.1: Connected Vertices of a Two-Dimensional Rectangle

$$\boldsymbol{C} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 1 & 4 \\ 4 & 2 & 3 \end{bmatrix} \tag{3.9}$$

That is, Vertex 1 is connected to Vertex 2 and Vertex 3, Vertex 2 is connected to Vertex 1 and Vertex 4, Vertex 3 is connected to Vertex 1 and Vertex 4, and Vertex 4 is connected to Vertex 2 and Vertex 3.

For $p = 3$, the vector of observed intervals is $\boldsymbol{X}_i = ([a_{i1}, b_{i1}][a_{i2}, b_{i2}][a_{i3}, b_{i3}])$. If $a_{ij} < b_{ij}$ for all $j = 1, 2, 3$, then the hyper-rectangle H_i representing $\boldsymbol{X}_i$ has eight vertices. The matrix of vertices based on Equation (2.30) is

$$\boldsymbol{X}_i^v = \begin{bmatrix} a_{i1} & a_{i2} & a_{i3} \\ a_{i1} & a_{i2} & b_{i3} \\ a_{i1} & b_{i2} & a_{i3} \\ a_{i1} & b_{i2} & b_{i3} \\ b_{i1} & a_{i2} & a_{i3} \\ b_{i1} & a_{i2} & b_{i3} \\ b_{i1} & b_{i2} & a_{i3} \\ b_{i1} & b_{i2} & b_{i3} \end{bmatrix}. \tag{3.10}$$

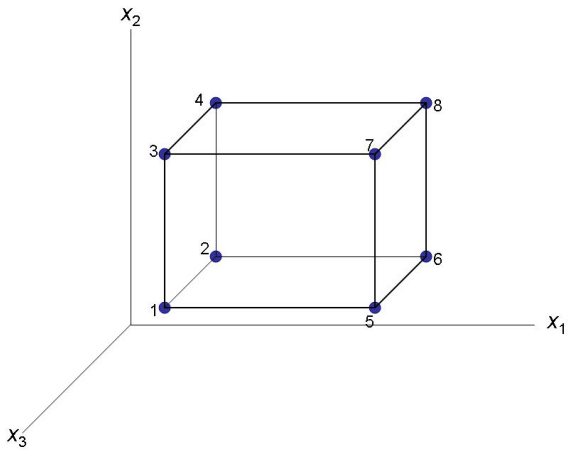

Figure 3.2: Connected Vertices of a Three-Dimensional Hyper-Rectangle

Therefore, Vertex 1 has coordinate (a_{i1}, a_{i2}, a_{i3}), Vertex 2 has coordinate (a_{i1}, a_{i2}, b_{i3}), Vertex 3 has coordinate (a_{i1}, b_{i2}, a_{i3}), Vertex 4 has coordinate (a_{i1}, b_{i2}, b_{i3}), Vertex 5 has coordinate (b_{i1}, a_{i2}, b_{i3}), and so on. Figure 3.2 shows the hyper-rectangle H_i representing $\boldsymbol{X}_i$ with the vertices labeled according to its row position in $\boldsymbol{X}_i^v$ of Equation (3.10). Check that the vertices are connected as indicated in matrix $\boldsymbol{C}$ as follows,

$$\boldsymbol{C} = \begin{bmatrix} 1 & 2 & 3 & 5 \\ 2 & 1 & 4 & 6 \\ 3 & 1 & 4 & 7 \\ 4 & 2 & 3 & 8 \\ 5 & 1 & 6 & 7 \\ 6 & 2 & 5 & 8 \\ 7 & 3 & 5 & 8 \\ 8 & 4 & 6 & 7 \end{bmatrix} \tag{3.11}$$

That is, Vertex 1 is connected to Vertex 2, Vertex 3, and Vertex 5; Vertex 2 is connected to 1, 4, and 6; Vertex 3 is connected to 1, 4, and 7; Vertex 4 is connected to 2, 3, and 8, Vertex 5 is connected to 1, 6, and 7; Vertex 6 is connected to 2, 5, and 8; Vertex 7 is connected to 3, 5, and 8; and Vertex 8 is connected to 4, 6, 7.

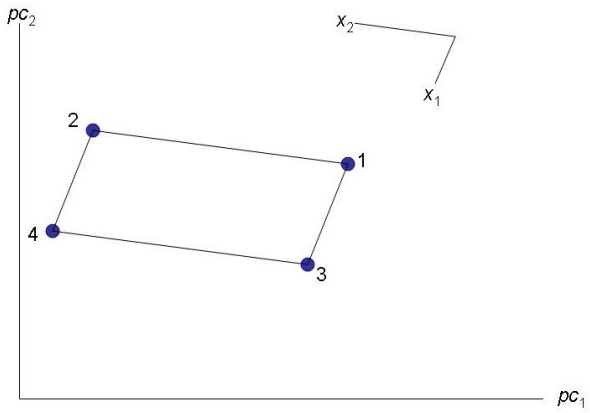

Figure 3.3: Connected Vertices of a Transformed Rectangle

Now, by applying the information from matrix C of Equations (3.9) and (3.11) for $p = 2$ and $p = 3$, respectively, transformed vertices in $\boldsymbol{Y}_i^v$ can be connected to form polytope P_i corresponding to H_i. The polytope P_i constructed this way represents the true structure of observation i in the principal components space. Figure 3.3 shows a polygon resulting from connecting the transformed vertices using the matrix $\boldsymbol{C}$ of Equation (3.9). Note that

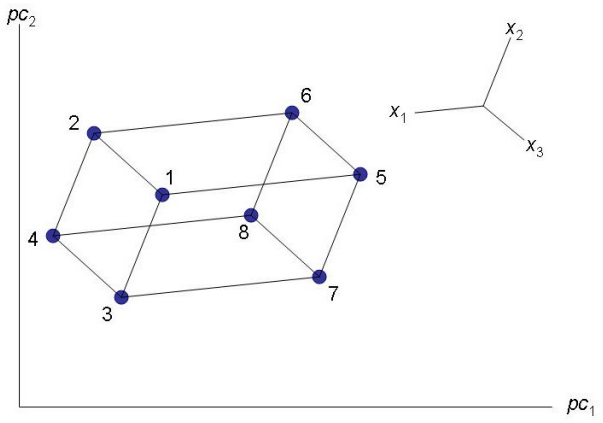

Figure 3.4: Connected Vertices of a Transformed Three-Dimensional Hyper-Rectangle

the resulting polygon in Figure 3.3 is a linear tranformation of the rectangle in Figure 3.1. Figure 3.4 shows a polytope resulting from connecting the transformed vertices using the matrix $\boldsymbol{C}$ of Equation (3.11). Again, note that the resulting polytope in Figure 3.4 is a linear tranformation of the hyper-rectangle in Figure 3.2. A detailed algorithm to construct these polytopes is given in Section 3.3 which also includes the procedure to construct the connected matrix $\boldsymbol{C}$.

Furthermore, a polytope of dimension $p > 3$ is difficult to visualize. Our algorithm in Section 3.3 also creates projections of the polytopes onto a two and a three dimensional space. Because the first few principal components explain most of the variation in the data, projection of the observations onto the first two or three principal components space can convey significant patterns in the observed data.

One of the common graphics generated in PCA is a plot of PCk_1 $\times$ PCk_2 where $k_1 \neq k_2$ and $k_1, k_2 = 1, 2, \ldots, p$. Figure 3.5 shows two representations of the same observation on a PC1 $\times$ PC2 plot. The observation shown in this plot come from a sample space with dimension $p = 6$. The green rectangle in this plot represents the maximum covering area

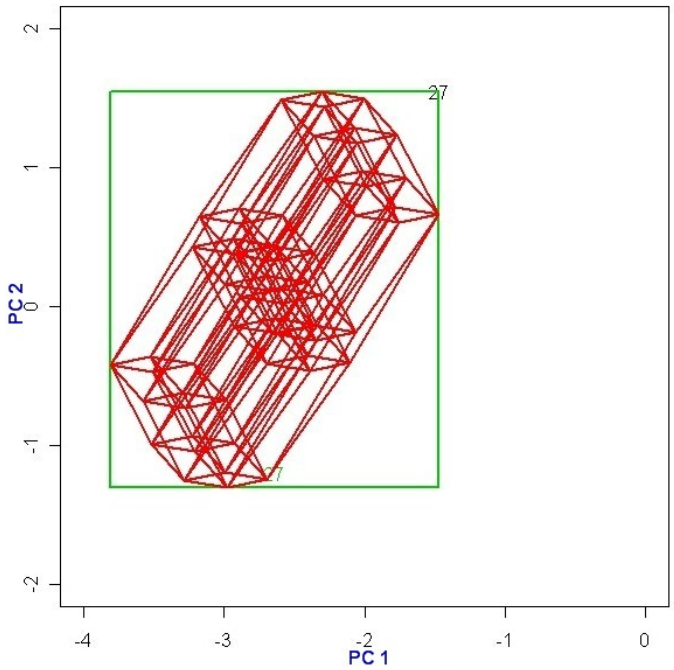

Figure 3.5: True Projection of Interval-Valued Observation versus Maximum Covering Area Rectangle

rectangle (MCAR) for this observation resulting from the vertices and the centers methods proposed by Cazes et al. (1997) and Chouakria et al. (2000). The red polygon represents the true projection of the observation onto the PC1 $\times$ PC2 plane as we propose. Clearly, the green rectangle covers an area much larger than the projection of the polytope resulting from the linear transformation of an interval-valued observation which is shown in red. Therefore, when n observations are shown on the same plot, the overage from MCAR may produce unnecessary overlaps between observations when these observations may be completely detached from each other in their true form. Our proposed visualization provides a true projection of the observed data. It does not produce the overage seen in MCAR.

Histogram-valued principal components

In addition to its role as a data exploration tool, PCA is also used as an intermediate step in data analysis. In classical PCA, the principal components are inputs in models as uncorrelated variables for further analysis. To treat the principal components of interval-valued observations as independent variables in a model, it is necessary to give numerical values to these principal components. In this dissertation we propose a method to translate the polytopes resulting from the linear transformation of the observed data into histograms of values. Symbolic data analysis methods can then be applied to histogram-valued principal components.

Since the original variables $X_{(j)}, j = 1, 2, \ldots, p$, take intervals of values, the resulting principal components have multiple values. Therefore, they are symbolic variables. To determine the internal structure of each principal component, it is necessary to examine the distribution of all values belonging to that principal component.

In the case of interval-valued observations, the k^{th} principal component of observation i, denoted Y_{ik}, takes values in the set Υ_{ik} where

$$\Upsilon_{ik} = \{y | y = \boldsymbol{x}'\boldsymbol{\nu}_k^v, \ \forall \boldsymbol{x} \in \boldsymbol{X}_i)\}. \tag{3.12}$$

The polytope $P_i = T(H_i)$ is the convex hull of the transformed vertices of H_i as established in Subsection 3.2.2. An equivalent expression of Υ_{ik} in Equation (3.12) in terms of a point in the linearly transformed hyper-rectangle is

$$\Upsilon_{ik} = \{y_k | \boldsymbol{y} = T(\boldsymbol{x}), \ \forall \boldsymbol{x} \in H_i\}. \tag{3.13}$$

In Equation (3.13), $\boldsymbol{y}$ is a p-vector and y_k is the k^{th} element of vector $\boldsymbol{y}$. The frequency that y_k takes value y inside P_i is proportional to the volume of the $(p-1)$-dimensional cross-section of P_i at $y_k = y$. Let $\gamma(y)$ be the volume of the $(p-1)$-dimensional cross-section of the polytope P_i at $y_k = y$. Then, the relative frequency or the density of each value $y \in \Upsilon_{ik}$ is the ratio of $\gamma(y)$ and the total volume of P_i. Let $V(P_i)$ be the volume of the polytope P_i.

Then,

$$V(P_i) = \int_{y_{min}}^{y_{max}} \gamma(y)dy$$

where y_{min} is the minimum value of the set Υ_{ik} and y_{max} is the maximum value of Υ_{ik}. Therefore, for $k = 1, 2, \ldots, p$, the density of the k^{th} principal component at y has the following form,

$$P(Y_{ik} = y) = \begin{cases} \gamma(y)/V(P_i), & y_{min} \leq y \leq y_{max}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.14}$$

Obviously, $0 \leq P(Y_{ik} = y) \leq 1$ and $\int_{-\infty}^{\infty} \gamma(y)/V(P_i)dy = 1$. Thus, the quantity $P(Y_{ik} = y)$ as defined in Equation (3.14) satisfies the definition of a density function. Although the density for all values within Υ_{ik} can be conceptually defined in Equation (3.14), computing the density $\gamma(y)$ presents some challenges. Two of these challenges are discussed in the following paragraph along with our proposed solutions.

First, it is impossible to compute volume for infinitely many cross-sections. Therefore, instead of trying to recreate a distribution for each value within the range of the principal component, we propose creating a histogram with subinterval endpoints coinciding with the vertices of the polytope. That is, instead of computing the volume of infinitely many $(p-1)$-dimensional polytopes, the problem is reduced to computing the volume of at most $(2^p - 1)$ p-dimensional polytopes.

Even with only a finite number of polytopes, the second challenge persists. Finding the volume of a polytope is a computationally complex problem as explained in the following. Volumes of polytopes can be computed by triangulation methods, signed decomposition methods or a mixture of both approaches. The efficiency of each approach depends on the exact shape of the polytopes. That is, in our case, for each observation i, we need to know the exact shape of up to $2^p - 1$ polytopes. This information is impractical to obtain. Even if the exact shape of each polytope was known and the most efficient approach was applied to the polytope, computing volume still requires a lot of memory and time when p is large. In addition, the triangulation method is based on the $\mathcal{V}$-*polytope* representation whereas the

signed decomposition method uses the $\mathcal{H}$-*polytope* representation. Since the shape of a polytope determines the method of computation, it determines if the polytope must be presented as a $\mathcal{V}$-*polytope* or a $\mathcal{H}$-*polytope*. Converting between the $\mathcal{V}$ and the $\mathcal{H}$ representation for a polytope can be more intensive than volume computation itself. Based on the computational cost involved in computing volume of a polytope, it seems counterproductive to use volume as a measure of density. One of the main purposes of PCA is to reduce the dimensionality of a dataset with large p to reduce computational cost associated with its analysis. Considering the total cost of all components involved in calculating volume of $2^p - 1$ polytopes, spending resources to calculate volume in these situations seems counterproductive.

In this disseration we propose an approach that is more efficient and still accounts for the internal variability of an interval-valued observation in a principal components space. We propose using the area of the polygon formed by projecting polytope P_i onto the PC1 $\times$ PCk plane for $k = 2, \ldots, p$ and onto the PC1 $\times$ PC2 plane for $k = 1$ instead of using the volume of the polytope to calculate the relative frequency. Computing area of a polygon is much more manageable than computing volume of a polytope. Of course, using the area of a polygon to compute the relative frequency of values within the polytope P_i only accounts for internal variation of the polytope in one dimension. However, the first principal component accounts for the largest amount of variation in the data. Including PC1 in computing the relative frequency means the relative frequency created by our proposed method reflects the most significant source of variation of values inside polytope P_i along the PCk-axis.

Therefore, let P_{2i} be the polygon resulting from projecting the p dimensional polytope P_i onto the PC1 $\times$ PCk plane. The subscript 2 in P_{2i} signifies that the polygon P_{2i} is a two-dimensional projection of the polytope P_i where P_i has dimension p. Let $\boldsymbol{Y}_i^v$ be the matrix of vertices of P_i. Let $\boldsymbol{Y}_{2i}$ be the submatrix of $\boldsymbol{Y}_i^v$ consisting of the k^{th} column and the 1^{st} column of $\boldsymbol{Y}_i^v$, respectively. Then the rows of matrix $\boldsymbol{Y}_{2i}$ represent the coordinates of points that belong to polygon P_{2i}.

Let T be a linear map from $\boldsymbol{R}^p$ to $\boldsymbol{R}^2$ and that if $\boldsymbol{a} = (a_1, a_2, \ldots, a_k, \ldots, a_p)$ is a point in $\boldsymbol{R}^p$ then $T(\boldsymbol{a}) = (a_k, a_1)$. Then, polygon P_{2i} can be thought of as a result of applying linear operator T to polytope P_i, i.e., $P_{2i} = T(P_i)$. Since T is a linear operator and P_i is a convex hull, it follows from Theorem 3.2.4 that P_{2i} is a convex hull. Furthermore, based on Theorem 3.2.7, all interior points of polytope P_i remain in the interior of polygon P_{2i}. Because T is not a bijective transformation, Theorem 3.2.8 does not apply to this transformation. Therefore, a boundary point of polytope P_i may not necessarily remain on the boundary of the polygon P_{2i}, i.e., some of the vertices of P_i may become an interior point of P_{2i} under transformation $T : \boldsymbol{R}^p \rightarrow \boldsymbol{R}^2$. Therefore, P_{2i} is a convex hull of the points represented by matrix $\boldsymbol{Y}_{2i}$. If we let $\boldsymbol{V}$ be a matrix whose rows are the coordinates of vertices of polygon P_{2i}, then the number of rows of matrix $\boldsymbol{V}$ is less than or equal the number of rows of matrix $\boldsymbol{Y}_{2i}$, i.e., polygon P_{2i} has at most the same number of vertices as polytope P_i.

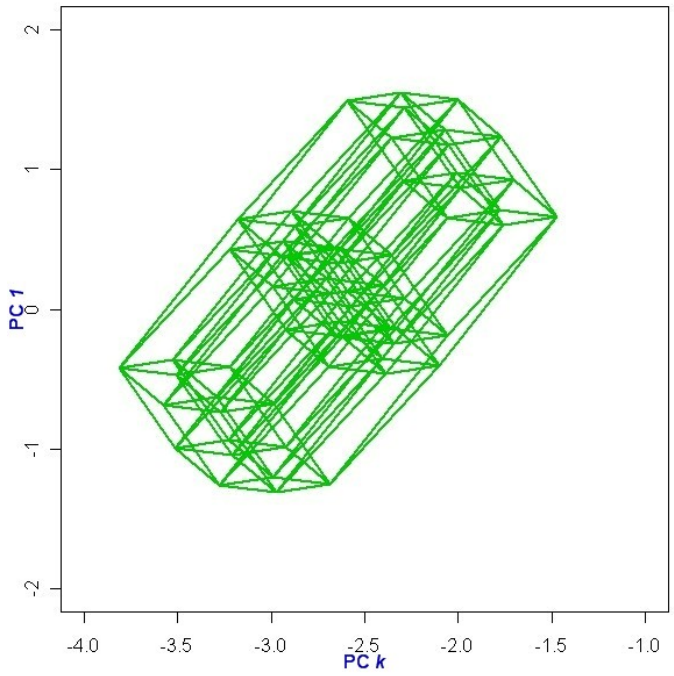

Figure 3.6: Two-Dimensional Projection of a Six-Dimensional Polytope

Figure 3.6 shows a projection of a 6-dimensional polytope onto a PC1 $\times$ PCk plane. The lines shown in green are the edges which are line segments connecting vertices of the

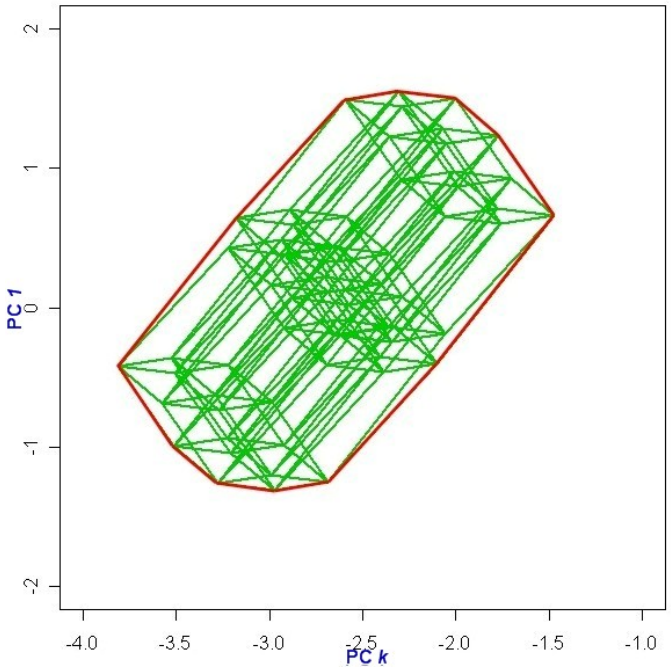

Figure 3.7: Polygon Formed by Convex Hull of Transformed Vertices

polytope. Figure 3.7 shows the same projection as in Figure 3.6. The red line segments in Figure 3.7 outline the polygon which is the convex hull of the vertices of the 6-dimensional polytope. Some vertices of the green polytope are inside the red polygon. All vertices of the red polygon are vertices of the green polytope. The red polygon shown in Figure 3.7 has twelve vertices while the green polytope has 64 vertices. The number of vertices for the polygon is much smaller than those for the polytope in this example.

Let pv denote the number of vertices of P_{2i}, i.e., pv is the number of rows of matrix $\boldsymbol{V}$. Elements of the first column of $\boldsymbol{V}$ are values of the vertices of polygon P_{2i} along principal component k and elements of the second column of $\boldsymbol{V}$ are values of the vertices of polygon P_{2i} along principal component 1. For the k^{th} principal component of observation i, denoted by Y_{ik}, we propose constructing a histogram with $pv-1$ subintervals. The relative frequencies for subintervals of the histogram are computed using area of the polygon P_{2i}. We propose

dividing the polygon P_{2i} into $pv-1$ pieces along the PCk-axis where the dividing lines coincide with the vertices of the polygon along the PCk-axis. Figure 3.8 illustrates the division of the polygon P_{2i}. This figure shows the same green polytope and the red polygon of Figure 3.7. The number of vertices for the red polygon is 12, i.e., $pv = 12$. The blue vertical line segments in Figure 3.8 divides the red polygon into $pv - 1 = 12 - 1 = 11$ pieces along the PCk axis. The vertices of the polygon in Figure 3.8 are labeled based on the order of their PCk values where 1 is the smallest and 11 is the largest.

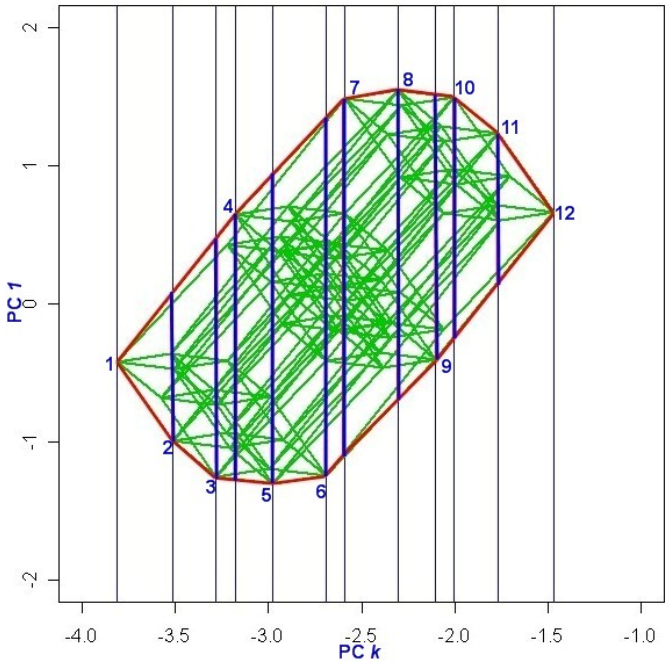

Figure 3.8: Subinterval Endpoints of Principal Component Histogram

Now, the endpoints for subinterval l of the histogram representing Y_{ik} are, respectively, the l^{th} and the $(l+1)^{th}$ lowest values of the first column of matrix $\boldsymbol{V}$, which consists of the PCk values of the vertices. The relative frequency of the l^{th} subinterval of Y_{ik} is the ratio of the area of the l^{th} piece of polygon P_{2i} along the PCk-axis, denoted by P_l, and the total area of the polygon P_{2i}. Figure 3.9 shows part of the polygon representing the first subinterval. The part of the polygon bounded between the first and the second vertices, i.e., P_1, is a triangle outlined in red in Figure 3.9. Values of the endpoints of this subinterval are the PCk values of the points labeled as 1 and 2, respectively, in Figure 3.9. The relative frequency of

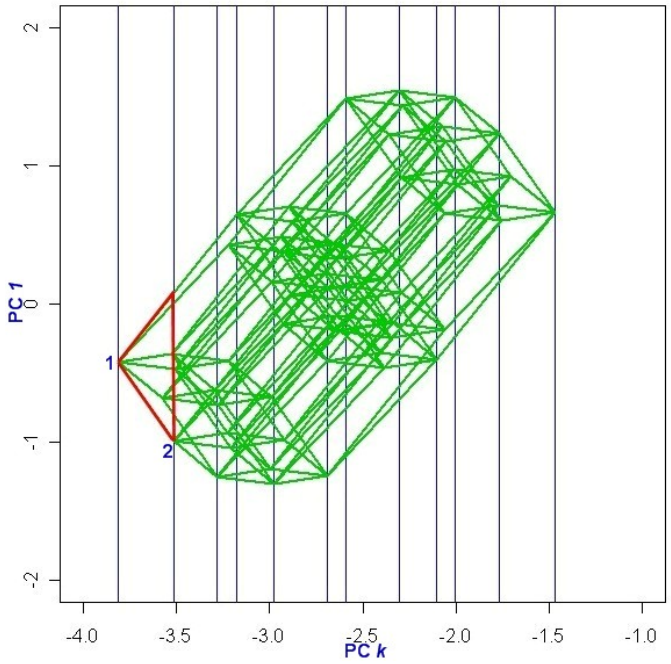

Figure 3.9: Part of Polygon Representing First Subinterval of Principal Component Histogram

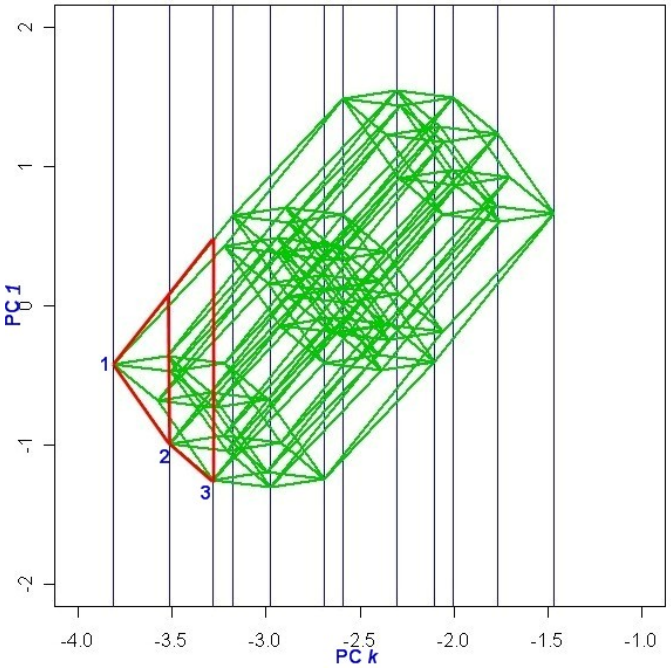

Figure 3.10: Parts of Polygon Representing First and Second Subintervals of Principal Component Histogram

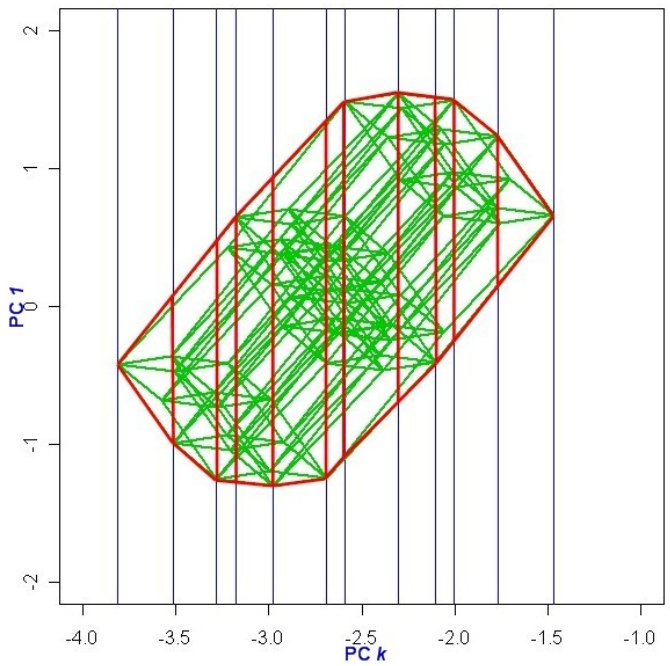

Figure 3.11: Parts of Polygon Representing All Subintervals of Principal Component Histogram

this subinterval is the ratio of the area of triangle P_1 and the area of polygon P_{2i}. In addition to P_1, Figure 3.10 shows the second piece of the polygon P_{2i}, i.e., P_2, representing the second subinterval which is the red trapezoid outlined in red. Values of the endpoints of the second subinterval are the PCk values of the points labeled 2 and 3, respectively, in Figure 3.10. The relative frequency of the second subinterval is the ratio of the area of trapezoid P_2 and the area of polygon P_{2i}. The process is continued until the vertex of polygon P_{2i} which has the largest PCk value has been reached as shown in Figure 3.11. We can also see from Figure 3.11 that part of the polygon bounded between subinterval endpoints is either a triangle or a trapezoid. Computing the area of a triangle or a trapezoid is much simpler than computing the volume of a p-dimensional polytope. Therefore, computing histogram-valued principal components based on the area of a projection of the polytope is much more efficient than computing them based on the volume of a p-dimensional polytope. Our proposed histogram-

valued principal components reflect the largest source of variation in the observations by including the first principal component in the construction of the histograms.

3.3 Algorithm

This section includes detailed descriptions of two algorithms. The first algorithm, presented in Subsection 3.3.1, builds n polytopes representing n observations in the principal components space. This algorithm also plots the polygons which are the 2-dimensional projections of the observed hyper-rectangles onto a PCk_1 $\times$ PCk_2 plane where $k_1 \neq k_2$ and $k_1, k_2 = 1, 2, \ldots, p$. The second algorithm, described in Subsection 3.3.2, constructs histogram-valued principal components based on the polygons formed by projecting the polytopes constructed from the algorithm of Subsection 3.3.1 onto a PC1 $\times$ PCk plane for $k = 2, 3, \ldots, p$.

The following algorithms are based on the assumption that coefficients for the principal components had been determined from the proposed method described in Subsection 3.2.1. The coefficients of the principal components are $\boldsymbol{\nu}^S = [\boldsymbol{\nu}_1^S, \boldsymbol{\nu}_2^S, \ldots, \boldsymbol{\nu}_p^S]$ where $\boldsymbol{\nu}_k^S, k = 1, \ldots, p$, are the eigenvectors of the covariance matrix $\boldsymbol{S}$ as defined in Equation (3.1) if the PCA is based on the covariance structure of the data, and $\boldsymbol{\nu}_k^S, k = 1, \ldots, p$, are the eigenvectors of the correlation matrix $\boldsymbol{R}$ whose elements are defined in Equation (3.4) if the PCA is based on the correlation structure of the data.

Due to the large number of vectors and matrices involved in the algorithms, subscripts are used in naming some vectors and matrices. Sometimes a computation only applies to one element or a subset of a matrix. In these situations, we use an indexing convention similar to the system used in the R language to specify the elements of a matrix. Indices of elements of a subset are specified by a pair of square brackets placed next to the matrix (vector) name. For a subset of a matrix, two numbers appear in the brackets separated by a comma. The first number corresponds to the row number and the second to the column number. For a subset of a vector, only one number appears in the bracket. It specifies the position of the element in the vector. To specify a whole row of a matrix, the first number is left blank; and

when a whole column is used, the second number is left blank. A set of consecutive numbers in the brackets is denoted by a colon between the smallest and the largest integer. Following are some examples to illustrate these notations. Let $\boldsymbol{M}$ be an $n \times p$ matrix. The 2^{nd} column of $\boldsymbol{M}$ is denoted by $\boldsymbol{M}[, 2]$. The notation $\boldsymbol{M}[1 : 15, 5]$ denotes the vector consisting of the first 15 elements of column five. In addition, let $\boldsymbol{v}$ be a vector of length n; then $\boldsymbol{v}[4]$ denotes the 4^{th} element of $\boldsymbol{v}$.

Another note on notation, since the algorithms in this section compute principal components of the observed data, we use a lower case letter to represent an observed data matrix to distinguish it from a random data matrix. For example, $\boldsymbol{x}$ identifies the observed version of the random data matrix $\boldsymbol{X}$ of Equation (2.26), $\boldsymbol{x}_i^v$ represents the observed version of the matrix of vertices $\boldsymbol{X}_i^v$ of Equation (3.2), and $\boldsymbol{y}_i^v$ is the observed version of the matrix of transformed vertices $\boldsymbol{Y}_i^v$ defined in Equation (3.7).

3.3.1 Interval-valued Observations in a Principal Components Space

This subsection is divided into two parts: part one constructs the p-dimensional polytopes representing interval-valued observations in the principal components space and part two includes steps to make 2-dimensional and 3-dimensional projections of the polytopes created from part one.

Constructing the polytopes

Let $\boldsymbol{x}$ represent the observed data matrix to be analyzed. Construction of the polytopes representing the observed data involves four steps:

Step 1. First, construct the matrix of vertices $\boldsymbol{x}^v$ defined in Equation (2.31) for the data matrix $\boldsymbol{x}$ using the following steps:

1. Initialize $\boldsymbol{x}^v$ by letting $\boldsymbol{x}^v$ be a $(n2^p \times p)$ matrix of zeros.

2. Update the elements of $\boldsymbol{x}^v$ by doing the following for each observation $i = 1, 2, \ldots, n$:

(a) For $j = 1, 2, \ldots, p$, do

- For $j_1 = 1, \ldots, 2^{(p-j)}$,
  - set $\boldsymbol{x}^v[(i-1)2^p + j_1, j] = \boldsymbol{x}[i, 2(j-1)]$
  - set $\boldsymbol{x}^v[(i-1)2^p + 2^{(p-j)} + j_1, j] = \boldsymbol{x}[i, 2j]$.

(b) For $j = 2, \ldots, p$, do

- For $j_1 = j, \ldots, p$, do
  - For $j_2 = 1, \ldots, 2^{(p+j-j_1-1)}$,
    set $\boldsymbol{x}^v[(i-1)2^p + 2^{(p+j-j_1-1)} + j_2, j_1] = \boldsymbol{x}^v[(i-1)2^p + j_2, j_1]$.

Step 2. Next, create the matrix of vertices $\boldsymbol{y}^v$ for n polytopes representing the n observations, one polytope for each observation, in the principal components space by transforming $\boldsymbol{x}^v$. This step involves the following two cases:

Case one. If the PCA is based on the symbolic covariance matrix, then $\boldsymbol{\nu}^S$ is the eigen matrix of the symbolic covariance matrix $\boldsymbol{S}$ in Equation (3.1) and $\boldsymbol{y}^v = (\boldsymbol{x}^v)(\boldsymbol{\nu}^S)$.

Case two. If the PCA is based on the symbolic correlation matrix, then $\boldsymbol{\nu}^S$ is the eigen matrix of the symbolic correlation matrix $\boldsymbol{R}$ whose elements are defined in Equation (3.4) and $\boldsymbol{y} = (\boldsymbol{z}^v)(\boldsymbol{\nu}^S)$ where $\boldsymbol{z}^v$ is the standardized matrix of vertices. That is, element mj of $\boldsymbol{z}^v$ comes from

$$\boldsymbol{z}^v[m, j] = \frac{\boldsymbol{x}^v[m, j] - \bar{w}_{(j)}}{\sqrt{s_{jj}}}$$

where $\bar{w}_{(j)}$ is the mean of the j^{th} variable defined in Equation (2.3) and s_{jj} is the variance of the j^{th} variable defined in Equation (2.4).

Step 3. Then, construct the matrix of connected vertices $\boldsymbol{C}$ associated with $\boldsymbol{x}^v$ as follows:

1. Initialize $\boldsymbol{C}$ by letting $\boldsymbol{C}$ be a $2^p \times p$ matrix of zeros.

2. Update $\boldsymbol{C}$ by doing the following step for $j = 1, 2, \ldots, p$:

   - For $j_1 = 0, 1, \ldots, 2^{(j-1)} - 1$, do

- For $j_2 = ((2^{(p-j+1)})j_1 + 1), \ldots, ((2^{(p-j+1)})j_1 + 2^{(p-j)})$,
  set $\boldsymbol{C}[j_2, j] = j_2 + 2^{(p-j)}$.
- 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 $\boldsymbol{C}[j_2, j] = j_2 - 2^{(p-j)}$.

Step 4. Conceptually, a p-dimensional plot of the polytopes can be constructed in the principal components space by the following substeps:

1. Make a scatter plot of $\boldsymbol{y}^v$.
2. Connect the vertices of each polytope by doing the following for $i = 1, 2, \ldots, n$:
   (a) Set $v = 2^p(i-1)$
   (b) For $v_1 = 1, 2, \ldots, 2^p$, do
   for $j_1 = 2, 3, \ldots, p+1$,
   set $v_2 = \boldsymbol{C}[v_1, j_1]$, and
   draw a line between points $\boldsymbol{y}^v[v + v_1,]$ and $\boldsymbol{y}^v[v + v_2,]$.

At the end of Step 4, we obtain a plot of n polytopes representing the n observations of data matrix $\boldsymbol{x}$ in the principal components space.

Making two and three dimensional plots

As mentioned in Subsection 3.2.2, it is difficult to visualize plots with dimension greater than 3. Therefore, 2-dimensional plots of PC$k_1 \times$ PCk_2 are commonly used in PCA. Replace Substeps 4.1 and 4.2 of the algorithm described in the first part of this subsection by the following steps:

1. Let $\boldsymbol{y}_2$ be an $n2^p \times 2$ matrix whose first and second columns are the ${k_1}^{th}$ and ${k_2}^{th}$ column of $\boldsymbol{y}^v$, respectively.
2. Make a scatter plot of $\boldsymbol{y}_2$.

3. Connect corresponding points of $\boldsymbol{y}_2$ by following Substep 4.2 of the algorithm described in the first part of this subsection with $\boldsymbol{y}_2$ in place of $\boldsymbol{y}^v$; now $p = 2$.

The same algorithm can be used to construct a 3-dimensional plot of PCk_1 $\times$ PCk_2 $\times$ PCk_3 with the following modifications:

1. Let $\boldsymbol{y}_3$ be an $n2^p \times 3$ matrix whose first, second, and third columns are columns k_1, k_2 and k_3 of $\boldsymbol{y}^v$, respectively.

2. Make a scatter plot of $\boldsymbol{y}_3$.

3. Follow Substep 4.2 of the algorithm described in the first part of this subsection with $\boldsymbol{y}_3$ in place of $\boldsymbol{y}^v$; now $p = 3$.

Figure 3.6 shows an example of the projection of a 6-dimensional polytope onto a 2-dimensional plane resulting from the algorithm described in Subsection 3.3.1.

3.3.2 Constructing histogram-valued principal components

As discussed in Subsection 3.2.2, principal components constructed as histograms reflect the internal variation of interval-valued observations. In Subsection 3.2.2, we propose constructing histograms of the principal components based on area of the polygons resulting from projecting the polytopes onto a PC1 $\times$ PCk plane. The following algorithm creates histogram-valued principal components. For each observation $i = 1, 2, \ldots, n$, the histogram representing the first principal component is based on the PC1 $\times$ PC2 plot as stated in Subsection 3.2.2 and for principal component k for $k = 2, \ldots, p$, the histogram is based on the PC1 $\times$ PCk plot. The following algorithm describes steps to construct histograms for principal component k in general. Therefore, the computation is based on a polygon from the PC1 $\times$ PCk plot. However, when computing a histogram for the first principal component, keep in mind that the computation is based on a polygon from the PC1 $\times$ PC2 plot.

This algorithm uses the fact that the interior angle formed by two edges of a polygon is less than 180^o. Therefore, given a set of three vertices $\{\boldsymbol{y}_{a_1}, \boldsymbol{y}_{a_0}, \boldsymbol{y}_{a_2}\}$ of the polygon, the

angle inside the polygon formed by line segments $\overline{\boldsymbol{y}_{a_1}\boldsymbol{y}_{a_0}}$ and $\overline{\boldsymbol{y}_{a_0}\boldsymbol{y}_{a_2}}$ is unique where $\overline{\boldsymbol{y}_{a_1}\boldsymbol{y}_{a_0}}$ is the line segment connecting point $\boldsymbol{y}_{a_1}$ to point $\boldsymbol{y}_{a_0}$. The formula used in the algorithm to compute the angle formed by the line segments $\overline{\boldsymbol{y}_{a_1}\boldsymbol{y}_{a_0}}$ and $\overline{\boldsymbol{y}_{a_0}\boldsymbol{y}_{a_2}}$, denoted by $\angle_{a_1a_0a_2}$, is

$$\angle_{a_1a_0a_2} = arccosine\left(\frac{\langle(\boldsymbol{y}_{a_1} - \boldsymbol{y}_{a_0}), (\boldsymbol{y}_{a_2} - \boldsymbol{y}_{a_0})\rangle}{||\boldsymbol{y}_{a_1} - \boldsymbol{y}_{a_0}|| * ||\boldsymbol{y}_{a_2} - \boldsymbol{y}_{a_0}||}\right) \tag{3.15}$$

where $\langle(a_1, a_2), (b_1, b_2)\rangle = a_1b_1 + a_2b_2$ and $||(a_1, a_2)|| = \sqrt{\langle(a_1, a_2), (a_1, a_2)\rangle} = \sqrt{a_1^2 + a_2^2}$. Figure 3.12 shows the angle $\angle_{a_1a_0a_2}$ formed by line segments $\overline{\boldsymbol{y}_{a_1}\boldsymbol{y}_{a_0}}$ and $\overline{\boldsymbol{y}_{a_0}\boldsymbol{y}_{a_2}}$.

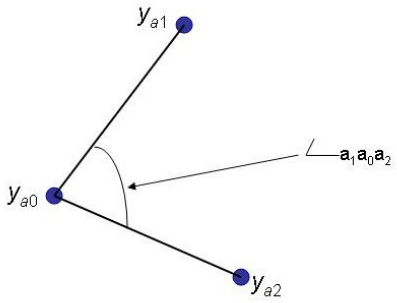

Figure 3.12: Angle Formed by Line $\overline{\boldsymbol{y}_{a_1}\boldsymbol{y}_{a_0}}$ and Line $\overline{\boldsymbol{y}_{a_0}\boldsymbol{y}_{a_2}}$

As explained in Subsection 3.2.2, we propose using area of the polygon P_{2i} to compute the relative frequency for subintervals of the principal component histograms. It is further illustrated in Figure 3.8 that part of the polygon that belongs to subinterval l, denoted by P_l, is either a triangle or a trapezoid. We need to compute the area of P_l. Let $\boldsymbol{p}_l$ be the matrix of vertices of P_l. Then, $\boldsymbol{p}_l$ is a (4×2) matrix whose rows are coordinates of the vertices. That is, when P_l is a trapezoid, then

$$\boldsymbol{p}_l = \begin{bmatrix} p_{11} & p_{12} \\ p_{11} & p_{22} \\ p_{21} & p_{32} \\ p_{21} & p_{42} \end{bmatrix}. \tag{3.16}$$

When P_l is a triangle, $p_{22} = p_{12}$, then

$$\boldsymbol{p}_l = \begin{bmatrix} p_{11} & p_{12} \\ p_{11} & p_{12} \\ p_{21} & p_{32} \\ p_{21} & p_{42} \end{bmatrix}. \tag{3.17}$$

Therefore, the area of P_l, denoted by $\omega(P_l)$, is simply

$$\omega(P_l) = \frac{1}{2}(|p_{12} - p_{22}| + |p_{32} - p_{42}|)(p_{21} - p_{11}) \tag{3.18}$$

for a trapezoid, and it is

$$\omega(P_l) = \frac{1}{2}(|p_{32} - p_{42}|)(p_{21} - p_{11}) \tag{3.19}$$

for a triangle.

Construction of the histogram-valued principal component k involves three steps. Details of these steps along with illustrations are given as follows. The illustrating example shows an interval-valued observation with $p = 3$. It is the same hyper-rectangle shown in Figure 3.4. The algorithm is easier to follow accompanied by an example using the hyper-rectangle representing an observation with a low dimension.

Step 1. Let $\boldsymbol{y}_2$ be an $(n2^p \times 2)$ matrix whose first and second columns are the k^{th} and the first column of $\boldsymbol{y}^v$, respectively.

Step 2. Let $\boldsymbol{pc}_k$ be an $(n \times (3 * 2^p))$ matrix to hold histogram values for principal component k. The i^{th} row of $\boldsymbol{pc}_k$ contains values of the histogram representing the k^{th} principal component for observation i. Initialize matrix $\boldsymbol{pc}_k$ by letting it be a matrix of zeros.

Step 3. Step 3 includes eleven substeps. Each complete execution of Step 3 builds a histogram representing one observation in the dataset. Therefore, Step 3 must be performed n times for n observations. For $i = 1, 2, \ldots, n$, do the following:

1. Set up the following matrices of zeros with specified dimension. These matrices store information to compute the frequency of subinterval l for $l = 1, 2, \ldots, (pv - 1)$ where $(pv - 1)$ is the number of subintervals belonging to the histogram representing the k^{th}

principal component of observation i. Therefore, these matrices are updated each time a new subinterval is computed.

(a) Matrix $\boldsymbol{p_l}$ has dimension (4×2). It consists of vertices of polygon P_l where P_l is the part of polygon P_{2i} belonging to subinterval l as described in Equation (3.16) and Equation (3.17).

(b) Matrix $\boldsymbol{p_d}$ has dimension (2×3). It holds coordinates of two connected vertices which make a line segment that will be used in subsequent subintervals. The first row of $\boldsymbol{p_d}$ represents a point of P_l that will be used in computing frequency of the subinterval $(l + 1)$. The first element of the first row of $\boldsymbol{p_d}$ gives the label of this point where the last two elements give its coordinate. The second row of matrix $\boldsymbol{p_d}$ stores information for the point connected to the point represented by the first row of $\boldsymbol{p_d}$. The point represented by the second row of $\boldsymbol{p_d}$ is a vertex of the polygon P_{2i}. Again, the first element of the second row gives the label for this point and the last two elements give its coordinate.

(c) Vector $\boldsymbol{s_i}$ has three elements. The first two elements of $\boldsymbol{s_i}$ hold the endpoints of subinterval l and the last element stores the frequency of subinterval l.

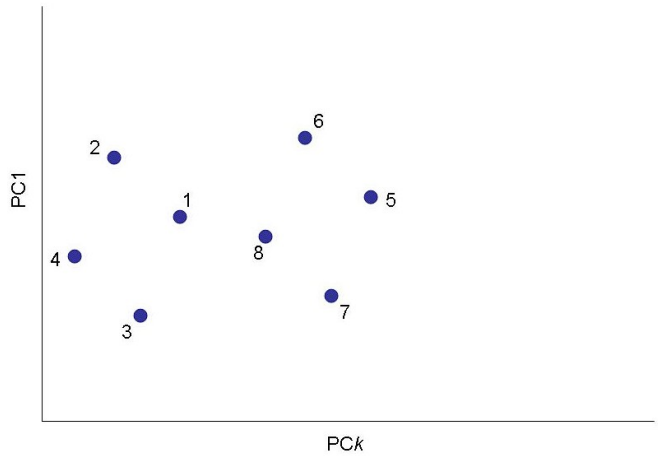

Figure 3.13: Points of Transformed Vertices on Principal Component Plane

2. Let $\boldsymbol{y}_{2i}$ be a $(2^p \times 2)$ matrix where $\boldsymbol{y}_{2i} = \boldsymbol{y}_2[(2^p(i-1)+1) : (2^p i),]$. That is, $\boldsymbol{y}_{2i}$ is a matrix of the k^{th} and the first coordinates of all vertices belonging to observation i. Matrix $\boldsymbol{y}_{2i}$ is the observed version of matrix $\boldsymbol{Y}_{2i}$ defined in Subsection 3.2.2. Figure 3.13 shows the points represented by rows of matrix $\boldsymbol{y}_{2i}$ plotted on the PC1 $\times$ PCk plane with PCk along the horizontal axis. The points in Figure 3.13 are labeled according to their row number in matrix $\boldsymbol{y}_{2i}$. As discussed in Subsection 3.2.2, the row numbers of matrix $\boldsymbol{y}_{2i}$ are identical to the elements of the first column of the connected matrix $\boldsymbol{C}$. The connected matrix $\boldsymbol{C}$ associated with the observation shown in Figure 3.13 is defined in Equation (3.11).

3. Let $y_{01} = min\{\boldsymbol{y}_{2i}[, 1]\}$, i.e., y_{01} is the minimum value of the elements of the first column of $\boldsymbol{y}_{2i}$. Equivalently, y_{01} is the minimum value along the PCk-axis of the vertices belonging to observation i.

4. Let m be the number of rows of $\boldsymbol{y}_{2i}$ whose first element equals y_{01}. For the first subinterval, i.e., when $l = 1$, the lower endpoint of the subinterval is the lowest value of the elements of the first column of $\boldsymbol{y}_{2i}$ which is y_{01}, and the upper endpoint of the subinterval is the second lowest value of the elements of the first column of $\boldsymbol{y}_{2i}$. Moreover, P_1 is the part of the polygon P_{2i} bounded between these lower and upper endpoints. Polygon P_1 can take one of two possible shapes. The shape that P_1 takes depends on the number of rows of matrix $\boldsymbol{y}_{2i}$ whose value for the first column equals y_{01}. There exist three possible cases as follows.

   If there is only one vertex of P_{2i} whose value along the PCk-axis equals to y_{01}, i.e., if m of Substep 3.4 is one, then P_1 is a triangle. This constitutes Case one of the next substep which is Substep 5. Figure 3.14 shows the points represented by matrix $\boldsymbol{y}_{2i}$. In the example shown in Figure 3.14, Vertex 4 has the minimum value along the PCk-axis. The vertical line intersecting Vertex 4 indicates the line representing the lower endpoint of the first subinterval. Moreover, Vertex 4 is the only vertex whose PCk

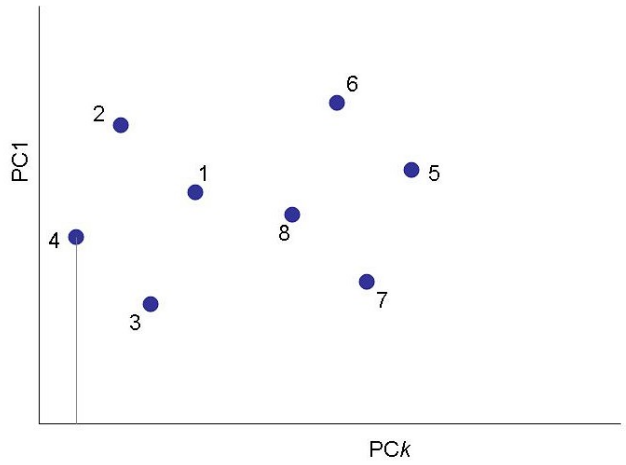

Figure 3.14: Unique Starting Point for First Subinterval

value equals to the minimum value y_{01}. That is, $m = 1$, and P_1 is a triangle in Figure 3.14.

If there is more than one vertex of P_{2i} whose value along the PCk-axis equals y_{01}, i.e., if $m > 1$, then two possible cases result from this situation. If all the vertices of P_{2i} whose value along the PCk-axis equals to y_{01} have the same value along the PC1-axis, i.e., these vertices are identical, then P_1 is still a triangle. This makes up Case two of Substep 5. Lastly, if $m > 1$ and the vertices whose values along the PCk-axis equal y_{01} have different values along the PC1-axis, i.e., these vertices are not all identical, then P_1 is a trapezoid. This constitutes Case three of Substep 5. Figure 3.15 shows an example where there exist two distinct vertices of P_{2i} whose PCk value equals y_{01}. In the example shown in Figure 3.15, Vertex 2 and Vertex 4 have the same value of y_{01} along the PCk-axis and they are not identical points. Therefore, polygon P_1 in this example is a trapezoid. The vertical line intersecting Vertex 2 and Vertex 4 indicates the line representing the lower endpoint of the first subinterval.

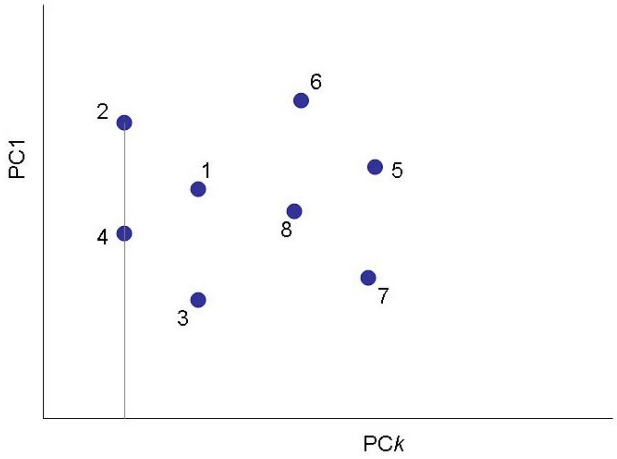

Figure 3.15: Multiple Starting Points for First Subinterval

5. The first subinterval can now be computed based on the three possible cases explained in Substep 4. The three cases follow:

   Case one. If $m = 1$, then:

   (a) Let a_0 be the row number of the vertex whose first element equals to y_{01}. In the example of Figure 3.14, $a_0 = 4$ because Vertex 4 has PCk value equal to y_{01}. The two points connected to $\boldsymbol{y}_{2i}[a_0,]$ that form the largest angle at $\boldsymbol{y}_{2i}[a_0,]$ form two edges of the triangle P_1. These points are found in the steps that follow.

   (b) Set $\boldsymbol{cr} = \boldsymbol{C}[a_0,]$ which is the list of indices of the points connected to the point $\boldsymbol{y}_{2i}[a_0,]$. In the example of Figure 3.14, $\boldsymbol{cr} = \boldsymbol{C}[4,] = (4, 2, 3, 8)$ where $(4, 2, 3, 8)$ is obtained from row four of the connected matrix $\boldsymbol{C}$ of Equation (3.11).

   (c) Let $\boldsymbol{p_d}[1, 1] = a_0$, i.e., Vertex a_0 is stored as a starting point of a line segment that will be used again in computation of the second subinterval.

(d) Set $cvert = a_0$ where $cvert$ is the set of all indices of vertices which had been used in computing the frequency of the first subinterval and should not be considered again in subsequent subintervals.

(e) Set $\boldsymbol{s_i}[1] = y_{01}$ which is the lower endpoint of the first subinterval.

(f) Set $\boldsymbol{p}_l[1,] = \boldsymbol{y}_{2i}[a_0,]$ and $\boldsymbol{p}_l[2,] = \boldsymbol{y}_{2i}[a_0,]$. Since the polygon bounded in the first subinterval is a triangle in this case, coordinates for the first two rows of $\boldsymbol{p}_l$ are identical as defined in Equation (3.17).

(g) Let $\boldsymbol{ang}$ be a 3-vector of zeros. This vector stores the angle and the indices of the set of vertices connected to Vertex a_0 and they form the largest angle at Vertex a_0.

(h) Go through all possible pairs of vertices connected to $\boldsymbol{y}_{2i}[a_0,]$ by:

For $j = 2, 3, \ldots, p$, do

- Set $a_1 = \boldsymbol{cr}[j]$.
- For $k = j + 1, \ldots, p + 1$, do
  - Let $a_2 = \boldsymbol{cr}[k]$.
  - Let $ang = \angle_{a_1 a_0 a_2}$ using the formula shown in equation Equation (3.15).
  - If $ang > \boldsymbol{ang}[1]$, let $\boldsymbol{ang} = (ang, a_1, a_2)$.

At the end of this loop, the second and the third elements of $\boldsymbol{ang}$ are indices of the two vertices connected to $\boldsymbol{y}_{2i}[a_0,]$ and form the largest angle at $\boldsymbol{y}_{2i}[a_0,]$. Figure 3.16 shows three possible pairs of vertices connected to Vertex 4 in this example. They are pair 2 and 3, pair 2 and 8, and pair 3 and 8. Of all the angles formed by these pairs, the angle formed by the pair of Vertices 2 and 3 is the largest. Thus, at the end of this step, the second and the third elements of $\boldsymbol{ang}$ are 2 and 3, respectively. Figure 3.17 shows the angle at Vertex 4 formed by the line segment connecting Vertex 2 to Vertex 4 and the line segment connecting Vertex 3 to Vertex 4.

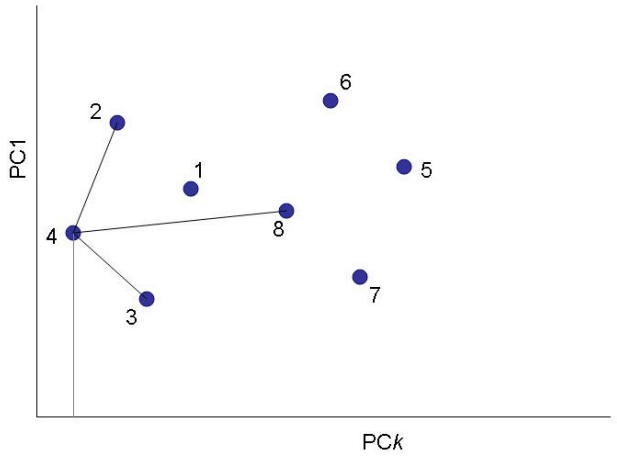

Figure 3.16: Angles Formed by Vertices Connected to Lower Endpoint of First Subinterval

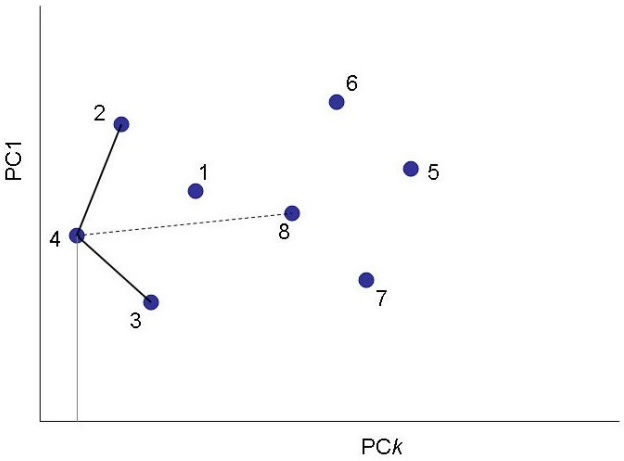

Figure 3.17: Largest Angle at Lower Endpoint of First Subinterval

(i) Next, we find the upper endpoint of the first subinterval by finding the vertex of polygon P_{2i} which has the second smallest PCk value. Let $y_{11} = \boldsymbol{y}_{2i}[\boldsymbol{ang}[2], 1]$ and $y_{12} = \boldsymbol{y}_{2i}[\boldsymbol{ang}[3], 1]$. That is, y_{11} and y_{12} are PCk values of the two vertices of polygon P_{2i} that form two edges of the triangle P_1. Comparing the values of y_{11} and y_{12} gives three possibilities, (a1) $y_{11} < y_{12}$, (b1) $y_{11} > y_{12}$, and (c1) $y_{11} = y_{12}$.
Case (a1). If $y_{11} < y_{12}$, then:

i. Let $a_3 = \boldsymbol{ang}[2]$ and $a_4 = \boldsymbol{ang}[3]$. That is, Vertex a_3 has a smaller PCk value than Vertex a_4. In the example of Figure 3.17, $a_3 = 2$ and $a_4 = 3$ because the PCk value of Vertex 2 is smaller than the PCk value of Vertex 3.

ii. Set $\boldsymbol{p_d}[2,] = (a_4, \boldsymbol{y}_{2i}[a_4,])$, i.e., $\boldsymbol{y}_{2i}[a_4,]$ is stored in the second row of matrix $\boldsymbol{p_d}$ to be used for the next subinterval. In the example of Figure 3.17, Vertex 3 is stored as the second row of matrix $\boldsymbol{p_d}$.

iii. Let $\boldsymbol{p}_l[3,] = (\boldsymbol{y}_{2i}[a_3,])$, i.e., $\boldsymbol{y}_{2i}[a_3,]$ is another vertex of triangle P_1.

iv. Let $\boldsymbol{s_i}[2] = \boldsymbol{y}_{2i}[a_3, 1]$, i.e., the upper endpoint for the first subinterval is the vertex whose PCk value is the second smallest among the PCk values of all vertices belonging to polygon P_{2i}. The vertical line intersecting Vertex a_3 forms the third edge of triangle P_1. Figure 3.18 shows that the line segment perpendicular to the PCk-axis intersecting Vertex 2 makes up the third edge of triangle P_1 and the coordinate of Vertex 2 becomes the third row of matrix $\boldsymbol{p}_l$.

v. Next, we find the last vertex of triangle P_1 by finding the point along the line segment $\overline{\boldsymbol{y}_{2i}[a_0,]\boldsymbol{y}_{2i}[a_4,]}$ intersecting the vertical line which is the third edge of triangle P_1. That is, set $\boldsymbol{p}_l[4, 1] = \boldsymbol{y}_{2i}[a_3, 1]$ and
set $\boldsymbol{p}_l[4, 2] = \boldsymbol{y}_{2i}[a_0, 2] + (\boldsymbol{p_d}[2, 3] - \boldsymbol{y}_{2i}[a_0, 2])(\boldsymbol{s_i}[2] - \boldsymbol{s_i}[1])/(\boldsymbol{p_d}[2, 2] - \boldsymbol{s_i}[1])$.
In Figure 3.18, the last vertex of the triangle is the intersection of the line segment connecting Vertex 4 to Vertex 3 and the vertical line intersecting Vertex 2.

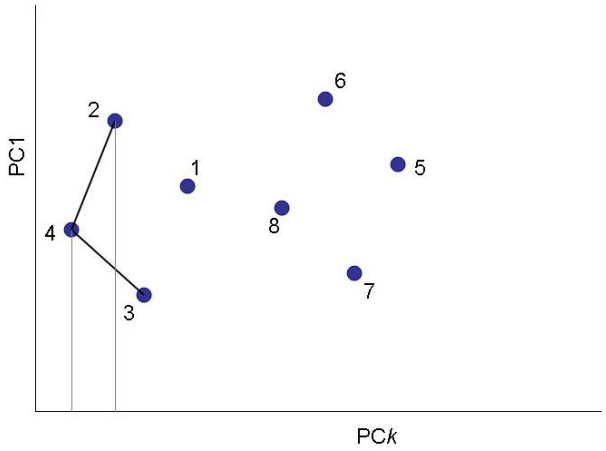

Figure 3.18: Triangle Belonging to First Subinterval of Principal Component Histogram

vi. Set $\boldsymbol{p_d}[1,] = (a_0, \boldsymbol{p}_l[4,])$. Vertex a_0 and the coordinate of the last vertex of triangle P_1 is stored as the first point in matrix $\boldsymbol{p_d}$ to be used in the next subinterval.

vii. Since $\boldsymbol{y}_{2i}[a_0,]$ forms a line with $\boldsymbol{y}_{2i}[a_3,]$ to make an angle for the next subinterval, set $a_1 = a_0$.

End of Case (a1).

Case (b1). If $y_{11} > y_{12}$, then:

i. Let $a_3 = \boldsymbol{ang}[3]$ and $a_4 = \boldsymbol{ang}[2]$.

ii. Follow Steps ii-vii as described in Case (a1).

End of Case (b1).

Case (c1) If $y_{11} = y_{12}$, then:

i. Set $a_3 = \boldsymbol{ang}[2]$, $a_4 = \boldsymbol{ang}[3]$.

ii. Set $\boldsymbol{p}_l[3,] = \boldsymbol{y}_{2i}[a_3,]$ and $\boldsymbol{p}_l[4,] = \boldsymbol{y}_{2i}[a_4,]$. Since both line segments $\overline{\boldsymbol{y}_{2i}[a_3,]\boldsymbol{y}_{2i}[a_0,]}$ and $\overline{\boldsymbol{y}_{2i}[a_4,]\boldsymbol{y}_{2i}[a_0,]}$ end at y_{11}, $\boldsymbol{y}_{2i}[a_3,]$ and $\boldsymbol{y}_{2i}[a_4,]$ make up the last two vertices of triangle P_1.

iii. Set $\boldsymbol{s_i}[2] = \boldsymbol{y}_{2i}[a_3, 1]$, i.e., the first element of $\boldsymbol{y}_{2i}[a_3,]$ becomes the upper endpoint of the subinterval.

iv. Let $\boldsymbol{p_d}[1,] = (a_4, \boldsymbol{y}_{2i}[a_4,])$, i.e., $\boldsymbol{y}_{2i}[a_4,]$ is used as a starting point for the next subinterval.

v. Since $\boldsymbol{y}_{2i}[a_0,]$ forms a line with $\boldsymbol{y}_{2i}[a_4,]$ to make the angle for the next step, set $a_1 = a_0$.

vi. To determine the second row of $\boldsymbol{p_d}$, find the vertex connected to $\boldsymbol{y}_{2i}[a_4,]$ that forms the largest angle with the line segment $\overline{\boldsymbol{y}_{2i}[a_4,]\boldsymbol{y}_{2i}[a_0,]}$ at Vertex a_4. Set $\boldsymbol{cr} = \boldsymbol{C}[a_4,]$, i.e., $\boldsymbol{cr}$ contains indices of the points connected to Vertex a_4.

vii. Delete elements of $\boldsymbol{cr}$ that belong to *cvert* to avoid multiplicity and let *nc* be the number of elements of $\boldsymbol{cr}$.

viii. Let $\boldsymbol{ag}$ be a 2-vector of zeros.

For $k = 2, \ldots, nc$,

set $a_2 = \boldsymbol{cr}[k]$, let $ang = \angle_{a_1 a_4 a_2}$, and if $ang > \boldsymbol{ag}[1]$ set $\boldsymbol{ag} = (ang, a_2)$.

ix. Set $\boldsymbol{p_d}[2,] = (\boldsymbol{ag}[2], y_{\boldsymbol{ag}[2]})$.

x. Set $a_1 = a_0$.

End of Case (c1).

(j) Let $area = \omega(P_1)$ where P_1 is the triangle whose vertices are the rows of $\boldsymbol{p}_l$. Having found the vertices of triangle P_1, apply the formula for $\omega(P_l)$ to matrix $\boldsymbol{p}_l$ as defined in Equation (3.19) to give the area of triangle P_1.

(k) Set $\boldsymbol{s_i}[3] = area$ and $\boldsymbol{hist} = \boldsymbol{s_i}$. Thus, the first row of $\boldsymbol{hist}$ represents the first subinterval endpoints and the area of triangle P_1.

End of Case one.

Case two. If $m > 1$ and the vertices are identical, then:

(a) Add indices of all those identical vertices to the *cvert* list.

(b) Let one of the vertices whose PCk value equals to y_{01}, say the first one, be a_0.

(c) Proceed with Steps (b)-(k) described in Case one.

End of Case two.

Case three. If $m > 1$ and these vertices have different PC1 values, then:

(a) Add indices of all vertices whose PCk value equals y_{01} to *cvert*.

(b) Let a_0 be the index of the vertex among these vertices which has the smallest PC1 value, and let a_1 be the one with the largest PC1 value. In the example shown in Figure 3.15, Vertices 2 and 4 have PCk value equal to y_{01} and Vertex 4 has a smaller PC1 value than Vertex 2. Therefore, $a_0 = 4$ and $a_1 = 2$.

(c) Set $\boldsymbol{s_i}[1] = y_{01}$, i.e., y_{01} is the lower endpoint of the first subinterval.

(d) Set $\boldsymbol{p}_l[1,] = \boldsymbol{y}_{2i}[a_0,]$ and $\boldsymbol{p}_l[2,] = \boldsymbol{y}_{2i}[a_1,]$, i.e., Vertex a_0 is the first vertex and Vertex a_1 is the second vertex of the trapepozoid P_1. In the example of Figure 3.15, Vertex 4 and Vertex 2 are, respectively, the first and the second vertices of trapezoid P_1.

(e) Set $\boldsymbol{cr} = \boldsymbol{C}[a_0,]$ and exclude all vertices belonging to *cvert* from $\boldsymbol{cr}$ and then, let nc be the length of $\boldsymbol{cr}$. In the example illustrated by Figure 3.15, $\boldsymbol{cr} = \boldsymbol{C}[4,] = (4, 2, 3, 8)$ and since Vertices 2 and 4 are in *cvert* by Step (a) in this case because they have PCk value equal to y_{01}, $\boldsymbol{cr} = (3, 8)$. Now, $nc = 2$ in this example.

(f) Next, find the vertex among the vertices in $\boldsymbol{cr}$ that forms the largest angle at $\boldsymbol{y}_{2i}[a_0,]$ with the line segment $\overline{\boldsymbol{y}_{2i}[a_1,]\boldsymbol{y}_{2i}[a_0,]}$ by setting $\boldsymbol{ag}$ to be a 2-vector of zeros. Then:

For $k = 1, 2, \ldots, nc$,

let $a_2 = \boldsymbol{cr}[k]$, let $ang = \angle_{a_1 a_0 a_2}$, and if $ang > \boldsymbol{ag}[1]$ set $\boldsymbol{ag} = (ang, a_2)$. Figure 3.19 illustrates this step. The line segment connecting Vertex 3 to Vertex 4 forms the largest angle with the line segment connecting Vertex 2 to Vertex 4.

(g) Set $a_{20} = \boldsymbol{ag}[2]$. Therefore, $a_{20} = 3$ in the example of Figure 3.19.

(h) Let $y_{11} = \boldsymbol{y}_{2i}[a_{20}, 1]$. Value y_{11} is a potential upper endpoint for the first subinterval which will be determined in Step (m) of this case.

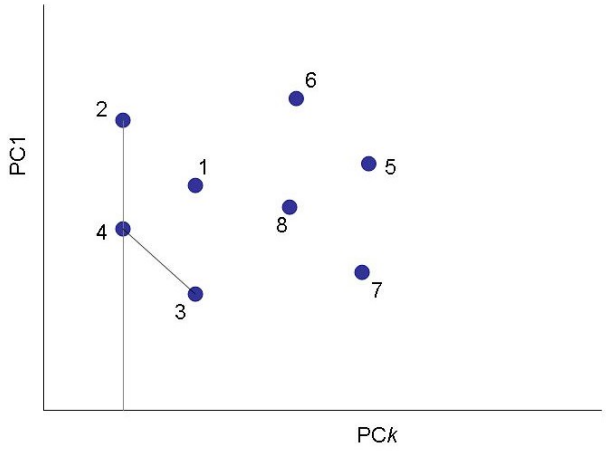

Figure 3.19: Largest Angle at First Vertex of Trapezoid Representing First Subinterval of Principal Component Histogram

(i) Set $\boldsymbol{cr} = \boldsymbol{C}[a_1,]$ and again exclude all vertices belonging to *cvert* from $\boldsymbol{cr}$ and let nc be the length of $\boldsymbol{cr}$. In the example illustrated by Figure 3.19, $\boldsymbol{cr} = \boldsymbol{C}[2,] = (2, 1, 4, 6)$ and since Vertices 2 and 4 are already in *cvert*, $\boldsymbol{cr} = (1, 6)$. Now, $nc = 2$ in this example. .

(j) Find the vertex among the vertices of $\boldsymbol{cr}$ that forms the largest angle at $\boldsymbol{y}_{2i}[a_1,]$ with the line segment $\overline{\boldsymbol{y}_{2i}[a_0,]\boldsymbol{y}_{2i}[a_1,]}$ by setting $\boldsymbol{ag}$ to be a 2-vector of zeros. Then:

For $k = 1, 2, \ldots, nc$,

let $a_2 = \boldsymbol{cr}[k]$, let $ang = \angle_{a_0 a_1 a_2}$, and if $ang > \boldsymbol{ag}[1]$ set $\boldsymbol{ag} = (ang, a_2)$. Figure

3.20 illustrates this step. The line segment connecting Vertex 6 to Vertex 2 forms the largest angle with the line segment connecting Vertex 4 to Vertex 2.

(k) Set $a_{21} = \boldsymbol{ag}[2]$. Therefore, $a_{21} = 6$ in the example of Figure 3.20.

(l) Let $y_{12} = \boldsymbol{y}_{2i}[a_{21}, 1]$. Value y_{11} is a potential upper endpoint for the first subinterval which will be determined in the following Step (m).

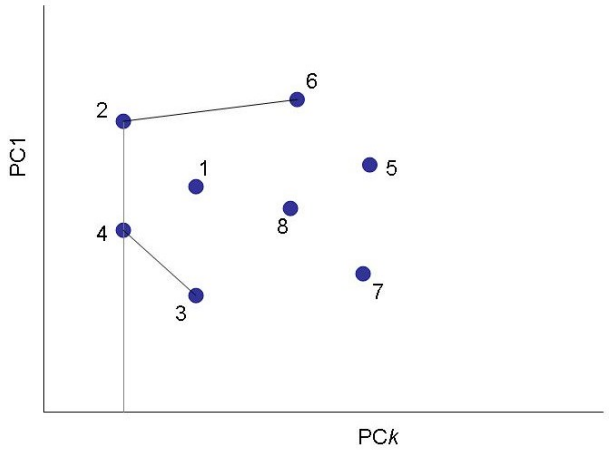

Figure 3.20: Largest Angle at Second Vertex of Trapezoid Representing First Subinterval of Principal Component Histogram

(m) Next, compare y_{11} and y_{12} to determine the subinterval upper endpoint. There are three possible cases, (a2) $y_{11} < y_{12}$, (b2) $y_{11} > y_{12}$, and (c2) $y_{11} = y_{12}$.

Case (a2). If $y_{11} < y_{12}$, then:

i. Let $a_3 = a_{20}$ and $a_4 = a_{21}$, i.e., Vertex a_3 has a smaller PC1 value than Vertex a_4. In Figure 3.21, Vertex 3 has a smaller PC1 value than Vertex 6. Therefore, $a_3 = 3$ and $a_4 = 6$.

ii. Set $\boldsymbol{s_i}[2] = \boldsymbol{y}_{2i}[a_3, 1]$, i.e., the PC$k$ value of Vertex a_3 is the upper endpoint of the first subinterval.

iii. Let $\boldsymbol{p}_l[3,] = \boldsymbol{y}_{2i}[a_3,]$, i.e., Vertex a_3 becomes the third vertex of trapezoid P_l. In Figure 3.21, Vertex 3 becomes the third vertex of trapezoid P_1.

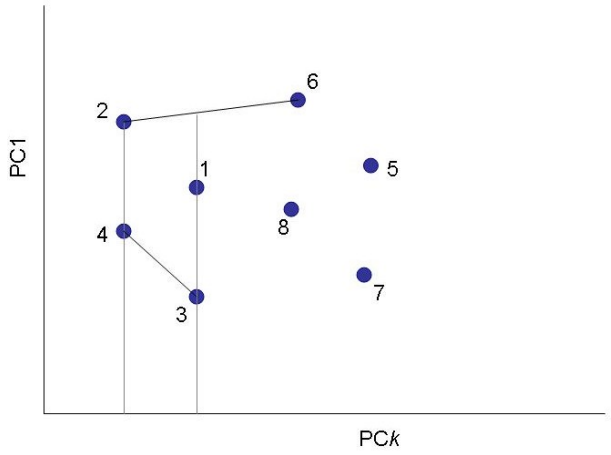

Figure 3.21: Trapezoid Representing First Subinterval of Principal Component Histogram

iv. Let $\boldsymbol{p_d}[2,] = (a_4, \boldsymbol{y}_{2i}[a_4,])$, i.e., Vertex a_4 is stored as the second row of matrix $\boldsymbol{p_d}$ to be used in the second subinterval. In the example of Figure 3.21, Vertex 6 is stored as the second row of matrix $\boldsymbol{p_d}$.

v. Set $\boldsymbol{p}_l[4,1] = \boldsymbol{y}_{2i}[a_3,1]$ and
set $\boldsymbol{p}_l[4,2] = \boldsymbol{y}_{2i}[a_1,2] + (\boldsymbol{p_d}[2,3] - \boldsymbol{y}_{2i}[a_1,2])(\boldsymbol{s_i}[2] - \boldsymbol{s_i}[1])/(\boldsymbol{p_d}[2,2] - \boldsymbol{s_i}[1])$, i.e., the fourth vertex of trapezoid P_1 is the intersection of the line segment connecting Vertices a_1 and a_4 and the vertical line intersecting Vertex a_3. In Figure 3.21, the fourth vertex of trapezoid P_1 is the intersection of the line segment connecting Vertices 2 and 6 and the vertical line intersecting Vertex 3.

vi. Let $\boldsymbol{p_d}[1,] = (a_1, \boldsymbol{p}_l[4,])$, i.e., the fourth vertex of trapezoid P_1 is stored as the first row of matrix $\boldsymbol{p_d}$ for use in the second subinterval.

vii. In the 2^{nd} subinterval, a_0 becomes an endpoint of an angle. Therefore, in preparation, set $a_1 = a_0$.

End of Case (a2).

Case (b2). If $y_{11} > y_{12}$, then:

i. Let $a_3 = a_{21}$ and $a_4 = a_{20}$.

ii. Perform Steps ii-iv described in Case (a2).

iii. Set $\boldsymbol{p}_l[4,2] = \boldsymbol{y}_{2i}[a_0, 2] + (\boldsymbol{p_d}[2,3] - \boldsymbol{y}_{2i}[a_0, 2])(\boldsymbol{s_i}[2] - \boldsymbol{s_i}[1])/(\boldsymbol{p_d}[2,2] - \boldsymbol{s_i}[1])$.

iv. Let $\boldsymbol{p_d}[1,] = (a_0, \boldsymbol{p}_l[4,])$.

End of Case (b2).

Case (c2). If $y_{11} = y_{12}$, then:

i. Set $a_3 = a_{20}$ and $a_4 = a_{21}$.

ii. Let $\boldsymbol{p}_l[3,] = \boldsymbol{y}_{2i}[a_3,]$, i.e., Vertex a_3 is the third vertex of trapezoid P_1.

iii. Let $\boldsymbol{p}_l[4,] = \boldsymbol{y}_{2i}[a_4,]$, i.e., Vertex a_4 is the fourth vertex of trapezoid P_1.

iv. Let $\boldsymbol{sub}[2] = \boldsymbol{y}_{2i}[a_3, 1]$, i.e., the PC$k$ value of Vertex a_3 is the upper endpoint of the first subinterval.

v. Let $\boldsymbol{p_d}[1,] = (a_4, \boldsymbol{y}_{2i}[a_4,])$, i.e., Vertex a_4 is stored as the first row of matrix $\boldsymbol{p_d}$ to be used in the second subinterval.

vi. Next, to determine the vertex for the second row of matrix $\boldsymbol{p_d}$, set $\boldsymbol{cr} = \boldsymbol{C}[a_4]$, remove all vertices belonging to *cvert* from $\boldsymbol{cr}$, and let nc be the length of $\boldsymbol{cr}$.

vii. Find the vertex among the vertices of $\boldsymbol{cr}$ that forms the largest angle at Vertex a_4 with the line segment $\overline{\boldsymbol{y}_{2i}[a_1,]\boldsymbol{y}_{2i}[a_4,]}$ at $\boldsymbol{y}_{2i}[a_4,]$ by letting $\boldsymbol{ag}$ be a 2-vector of zeros. Then:

For $k = 1, 2, \ldots, nc$,

set $a_2 = \boldsymbol{cr}[k]$, let $ang = \angle_{a_1 a_4 a_2}$, and if $ang > \boldsymbol{ag}[1]$ set $\boldsymbol{ag} = (ang, a_2)$.

viii. Set $\boldsymbol{p_d}[2,] = (\boldsymbol{ag}[2], \boldsymbol{y}_{2i}[\boldsymbol{ag}[2],])$.

End of Case (c2).

(n) Set $area = \omega(P_1)$ and let $\boldsymbol{p}_1$ be the matrix consisting of vertices of trapezoid P_1. Then, the area of trapezoid P_1 can be computed by applying Equation (3.18) to the matrix $\boldsymbol{p}_1$.

(o) Set $\boldsymbol{s_i}[3] = area$.

(p) Set $\boldsymbol{hist} = \boldsymbol{s_i}$.

End of Substep 5. At the end of Substep 5, we obtain the endpoints and the frequency for the first subinterval, i.e., in Substep 5, $l = 1$. Now, to find the endpoints and the frequency for subsequent subintervals, proceed with Substep 6.

6. This substep computes the endpoints and the frequency for subsequent subintervals. The first time this substep is executed, it computes the endpoints and the frequency for the second subinterval.

(a) Let $l = l + 1$.

(b) Since Vertex a_3 of subinterval $l-1$ becomes Vertex a_0 of this subinterval l, set $a_0 = a_3$. For the example in Figure 3.22, the new Vertex a_0 for the second subinterval, i.e., $l = 2$, is Vertex 2.

(c) Let $\boldsymbol{s_i}[1] = \boldsymbol{s_i}[2]$ and move $\boldsymbol{p}_l[3 : 4,]$ to $\boldsymbol{p}_l[1 : 2,]$, i.e., the upper endpoint of subinterval $l - 1$ becomes the lower endpoint of subinterval l and the last two vertices of trapezoid P_{l-1} (P_{l-1} could be a triangle if $l = 2$ and the first subinterval falls into Case one or two of Substep 5) become the first two vertices of trapezoid P_l. Figure 3.22 shows an example for $l = 2$ where P_1 is a triangle.

(d) For $l_1 = 1, 2, \ldots, 2^p$, if $\boldsymbol{y}_{2i}[l_1, 1] = \boldsymbol{y}_{2i}[a_3, 1]$ add l_1 to *cvert*, i.e., add all vertices of the polygon P_{2i} whose PCk value equals the lower endpoint of subinterval l to *cvert* because these points need not be used again for subsequent subintervals.

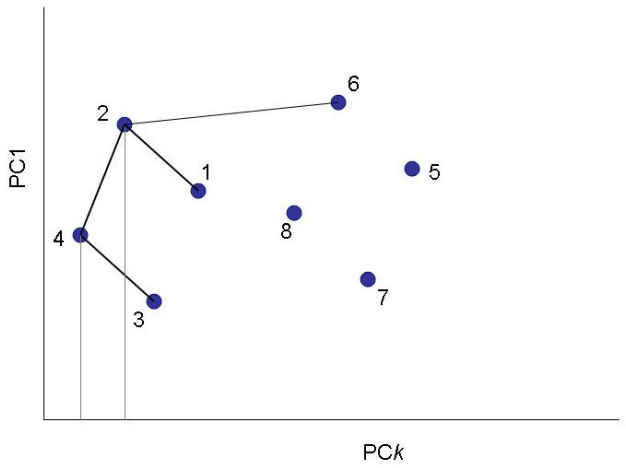

Figure 3.22: Angles Formed by Vertices Connected to Lowest Vertex of Second Subinterval

(e) Set $\boldsymbol{cr} = \boldsymbol{C}[a_0,]$ and exclude vertices belonging to *cvert* from $\boldsymbol{cr}$ and let nc be the length of $\boldsymbol{cr}$. In the example of Figure 3.22, $\boldsymbol{cr} = \boldsymbol{C}[2,] = (2, 1, 4, 6)$ and since Vertex 2 and 4 have been used already, then $\boldsymbol{cr} = (1, 6)$.

(f) Find the vertex in $\boldsymbol{cr}$ that forms the largest angle with the line segment connecting Vertices a_1 and a_0 at Vertex a_0 by letting $\boldsymbol{ag}$ be a 2-vector of zeros. Then:

For $k = 1, 2, \ldots, nc$,

set $a_2 = \boldsymbol{cr}[k]$, let $ang = \angle_{a_1 a_0 a_2}$, and if $ang > \boldsymbol{ag}[1]$ set $\boldsymbol{ag} = (ang, a_2)$. In the example shown in Figure 3.22, the vertex that forms the largest angle with the line segment connecting Vertices 4 and 2 at Vertex 2 is Vertex 6.

(g) Update $\boldsymbol{p_d}$, $\boldsymbol{p}_l$, and $\boldsymbol{s_i}$ by first comparing $\boldsymbol{y}_{2i}[\boldsymbol{ag}[2], 1]$ to $\boldsymbol{p_d}[2, 2]$.

Case (a3). If $\boldsymbol{y}_{2i}[\boldsymbol{ag}[2], 1] > \boldsymbol{p_d}[2, 2]$, then:

Figure 3.23 shows an example of Case (a3) where Vertex 3 has a smaller PCk value than Vertex 6.

i. Set $a_1 = \boldsymbol{p_d}[1, 1]$ and $a_3 = \boldsymbol{p_d}[2, 1]$, i.e., Vertex a_3 has a smaller PCk value than Vertex a_1.

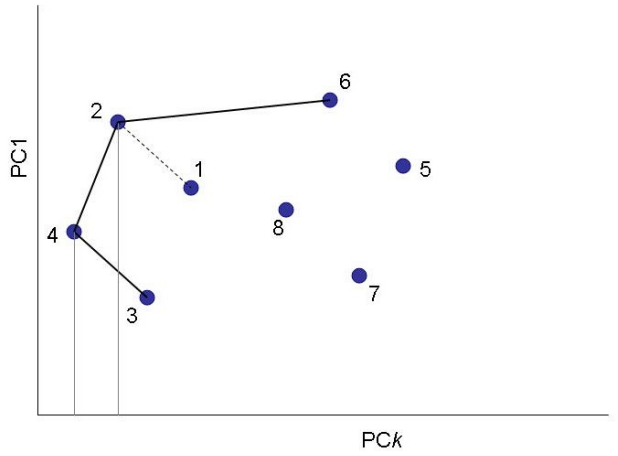

Figure 3.23: Largest Angle at Lowest Vertex of Second Subinterval

ii. Set $\boldsymbol{s_i}[2] = \boldsymbol{p_d}[2,2]$, i.e., the PC$k$ value of Vertex a_3 is the upper endpoint of subinterval l.

iii. Set $\boldsymbol{p}_l[3,] = \boldsymbol{p_d}[2, 2:3]$, i.e., the point represented by the second row of matrix $\boldsymbol{p_d}$ is the third vertex for trapezoid P_l.

iv. Update the second row of matrix $\boldsymbol{p_d}$ by letting $\boldsymbol{p_d}[2,] = (\boldsymbol{ag}[2], \boldsymbol{y}_{2i}[\boldsymbol{ag}[2],])$.

v. Find the new coordinate of the fourth vertex of trapezoid P_l by letting $\boldsymbol{p}_l[4,1] = \boldsymbol{p_d}[2,2]$ and
$\boldsymbol{p}_l[4,2] = \boldsymbol{y}_{2i}[a_0,2] + (\boldsymbol{p_d}[2,3] - \boldsymbol{y}_{2i}[a_0,2])(\boldsymbol{s_i}[2] - \boldsymbol{s_i}[1])/(\boldsymbol{p_d}[2,2] - \boldsymbol{s_i}[1])$. Figure 3.24 shows the trapezoid P_l for $l = 2$ in our example.

vi. Update the first row of $\boldsymbol{p_d}$ by setting $\boldsymbol{p_d}[1,1] = a_0$ and set $\boldsymbol{p_d}[1, 2:3] = \boldsymbol{p}_l[4,]$.

vii. Add a_0 to *cvert* since Vertex a_0 is used in this subinterval and will not be used again in subsequent intervals.

End of Case (a3).

Case (b3). If $\boldsymbol{y}_{2i}[\boldsymbol{ag}[2], 1] < \boldsymbol{p_d}[2,2]$, then:

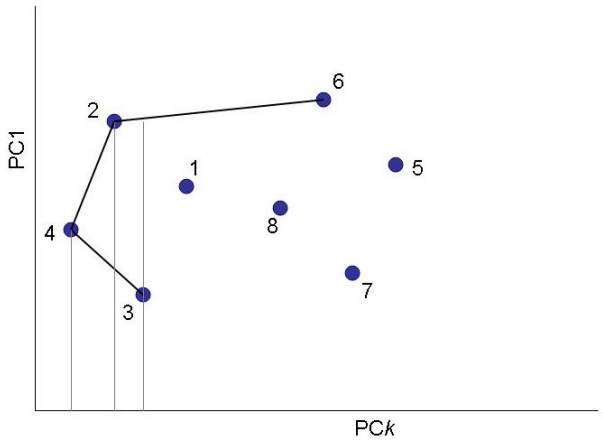

Figure 3.24: Trapezoid Belonging to Second Subinterval of Principal Component Histogram

i. Set $a_3 = \boldsymbol{ag}[2]$.

ii. Set $\boldsymbol{s_i}[2] = \boldsymbol{y}_{2i}[a_3, 1]$.

iii. Update the third vertex of trapezoid P_l by letting $\boldsymbol{p}_l[3,] = (\boldsymbol{y}_{2i}[a_3,])$.

iv. Update the fourth vertex of trapezoid P_l by letting $\boldsymbol{p}_l[4,1] = \boldsymbol{y}_{2i}[a_3, 1]$ and $\boldsymbol{p}_l[4,2] = \boldsymbol{p_d}[1,3] + (\boldsymbol{p_d}[2,3] - \boldsymbol{p_d}[1,3])(\boldsymbol{s_i}[2] - \boldsymbol{s_i}[1])/(\boldsymbol{p_d}[2,2] - \boldsymbol{s_i}[1])$.

v. Set $a_1 = a_0$.

vi. Add a_0 to *cvert*.

End of Case (b3).

Case (c3). If $\boldsymbol{y}_{2i}[\boldsymbol{ag}[2], 1] = \boldsymbol{p_d}[2,2]$, then:

i. Set $a_3 = \boldsymbol{ag}[2]$, $a_4 = \boldsymbol{p_d}[2,1]$ and $a_1 = \boldsymbol{p_d}[1,1]$.

ii. Add a_4 to *cvert*.

iii. Set $\boldsymbol{p}_l[3,] = \boldsymbol{y}_{2i}[a_3,]$ and $\boldsymbol{p}_l[4,] = \boldsymbol{y}_{2i}[a_4,]$.

iv. Set $\boldsymbol{s_i}[2] = \boldsymbol{y}_{2i}[a_3, 1]$.

v. Let $\boldsymbol{p_d}[1,] = (a_4, \boldsymbol{y}_{2i}[a_4,])$.

vi. Find the vertex connected to Vertex a_4 that forms the largest angle with the line segment $\overline{\boldsymbol{y}_{2i}[a_1,]\boldsymbol{y}_{2i}[a_4,]}$ at $\boldsymbol{y}_{2i}[a_4,]$ by setting $\boldsymbol{cr} = \boldsymbol{C}[a_4,]$ and removing all vertices of $\boldsymbol{cr}$ belonging to $cvert$ and let nc be the length of $\boldsymbol{cr}$.

vii. Let $\boldsymbol{ag}$ be a 2-vector of zeros.

For $k = 1, 2, \ldots, nc$,

set $a_2 = \boldsymbol{cr}[k]$, set $ang = \angle_{a_1 a_4 a_2}$, and if $ang > \boldsymbol{ag}[1]$ set $\boldsymbol{ag} = (ang, a_2)$.

viii. Set $\boldsymbol{p_d}[2,] = (\boldsymbol{ag}[2], \boldsymbol{y}_{2i}[\boldsymbol{ag}[2],])$.

End of Case (c3).

(h) Let $area = \omega(P_l)$ of Equation (3.18).

(i) Set $\boldsymbol{s_i}[3] = area$, i.e., the area of P_l is the frequency of subinterval l.

(j) Add $\boldsymbol{s_i}$ as a row to the end of $\boldsymbol{hist}$.

(k) Since polygon P_{2i} is symmetric about its centroid, we only need to compute subintervals for the first half of the polygon, subinterval frequency for the subintervals belonging to the second half of the polygon can be immediately deduced from their counterparts in the first half. Substep 7 computes subinterval endpoints and frequency using this property of polygon P_{2i}. If $\boldsymbol{s_i}[2] < md$ where $md = (min\{\boldsymbol{y}_{2i}[,1]\} + max\{\boldsymbol{y}_{2i}[,1]\})/(2)$ is the average PCk value of the points belonging to polygon P_{2i}, then, repeat Substep 6. If $\boldsymbol{x_i}[2] \geq md$, then go to Substep 7. That is, if the upper endpoint of subinterval l is still in the first half of polygon P_{2i}, then compute the next subinterval using Substep 6. If the upper endpoint of subinterval l has reached the second half of polygon P_{2i}, then the subinterval endpoints and frequency for subsequent subintervals can be obtained by applying Substep 7. Figure 3.25 shows trapezoid P_3 belonging to the first half of polygon P_{2i} of our example. Since the upper endpoint of the third subinterval in this example is larger than the midpoint md, we can go to Substep 7.

End of Substep 6.

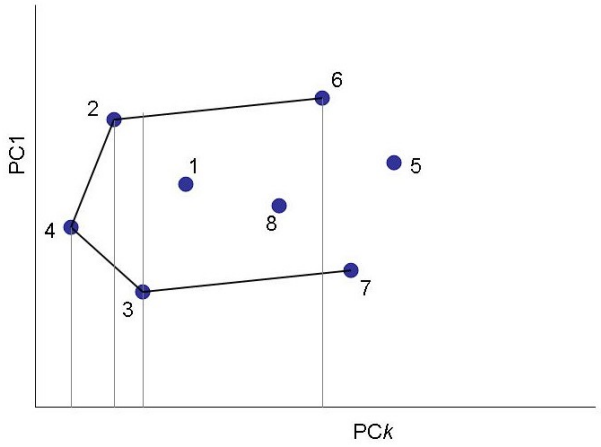

Figure 3.25: Trapezoid Belonging to Third Subinterval of Principal Component Histogram

7. This substep is divided into two different cases. The first case works for polygon P_{2i} which has an odd number of subintervals, and the second case works for an even number of subintervals. Let nrw be the number of rows in ***hist*** resulting from Substep 6.

   Case one: If $\boldsymbol{s_i}[2] > md$, then add $(nrw - 1)$ rows of zeros to ***hist***, i.e., if P_{2i} has an odd number of subintervals the midpoint md is located at the center of trapezoid P_{nrw}. Therefore, we need to add $(nrw - 1)$ more subintervals. Figure 3.26 shows an example where the vertical line intersecting the midpoint md cuts through the center of the trapezoid P_3. Therefore, there are only $(nrw - 1)$ subintervals remaining in the second half of the polygon P_{2i}.

   For $l = 1, 2, \ldots, (nrw - 1)$,

   set $\boldsymbol{hist}[nrw + l, 1:2] = 2 * md - \boldsymbol{hist}[nrw - l, 2:1]$ and

   set $\boldsymbol{hist}[nrw + l, 3] = \boldsymbol{hist}[nrw - l, 3]$.

   End of Case one.

   Case two: If $\boldsymbol{s_i}[2] = md$, then add nrw rows of zeros to ***hist***, i.e., if P_{2i} has an even number of subintervals the midpoint md is located at the upper end of trapezoid P_{nrw}.

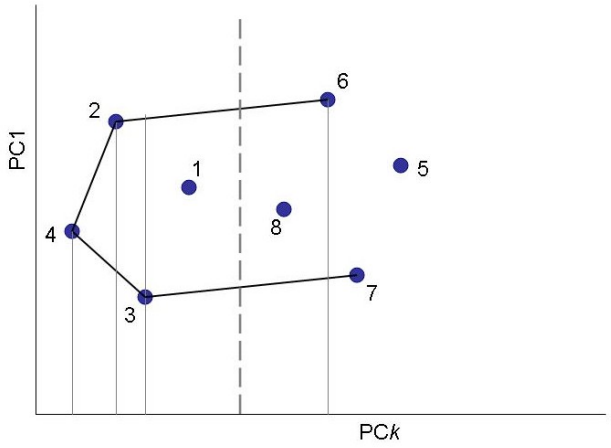

Figure 3.26: Symmetry of Polygon Formed by Transformed Vertices

Therefore, we need to add nrw more subintervals.

For $l = 1, 2, \ldots, nrw$,

set $\boldsymbol{hist}[nrw + l, 1:2] = 2 * md - \boldsymbol{hist}[nrw - l + 1, 2:1]$ and

set $\boldsymbol{hist}[nrw + l, 3] = \boldsymbol{hist}[nrw - l + 1, 3]$.

End of Case two.

8. Update nrw by letting nrw be the number of rows of $\boldsymbol{hist}$.

9. The elements of $\boldsymbol{hist}[, 3]$ correspond to areas of the polygons bounded between the subinterval endpoints. To obtained the relative frequency, do the following:

   - Add a column to the end of $\boldsymbol{hist}$.
   - Let $sum = \sum_{k=1}^{nrw} \boldsymbol{hist}[k, 3]$.
   - For $l = 1, 2, \ldots, nrw$,

     set $\boldsymbol{hist}[l, 4] = (\boldsymbol{hist}[l, 3])/(sum)$.

10. Finally, transfer the information from matrix $\boldsymbol{hist}$ to the row i of matrix $\boldsymbol{pc}_k$ which is the matrix set up to keep the endpoints and relative frequencies for the histograms representing principal component k for observations 1 through n.

    For $l = 1, 2, \ldots, nrw$,

    set $\boldsymbol{pc}_k[i, (3s-2)] = \boldsymbol{hist}[s, 1]$,

    set $\boldsymbol{pc}_k[i, (3s-1)] = \boldsymbol{hist}[s, 2]$, and

    set $\boldsymbol{pc}_k[i, (3s)] = \boldsymbol{hist}[s, 4]$.

11. At the end of Substep 10, we obtain a histogram representing principal component k for observation i. Information for the histogram for observation i is stored in row i of the matrix $\boldsymbol{pc}_k$. If $i < n$, let $i = i + 1$ then repeat Step 3. If $i = n$, quit.

End of Step 3.

End of algorithm. At the end of this algorithm, we obtain matrix $\boldsymbol{pc}_k$ whose rows contain histograms representing the k^{th} principal component of n observations in the data matrix $\boldsymbol{X}$. This algorithm must be executed for each principal component k of interest.

3.4 Applications

In this section, we apply our proposed symbolic covariance PCA method to two datasets. The first example includes a famous dataset known as the Fisher's Iris data. There are two reasons this dataset was chosen as our first example. The first reason is due to the fact that as it is one of the most widely used examples in multivariate analysis, it gives a good frame of reference to the effectiveness of our proposed method. Additionally, with only four variables, it illustrates the internal structure of interval-valued observations more effectively than using data with higher dimension. In the second example, we use a realistic application of symbolic data analysis by using the interval-valued face measurements of a dataset collected by Leroy et al. (1996). Each observation in this Face Recognition dataset is a compilation of sequences

of images. Therefore, the resulting measurements for each observation cover a range of values. They naturally have interval values.

3.4.1 Iris example

The data

In this example we apply our proposed PCA method for interval observations to the Iris dataset. Although the Iris data were collected by Edgar Anderson to determine geographic variation of Iris flowers (Anderson (1935)), it is known as the Fisher's Iris dataset due to an article published by R.A. Fisher in 1936. In Fisher (1936), the data were used as an application of discriminant analysis. The Iris dataset gives measurements of 50 Iris flowers each from three species: setosa, versicolor, and virginica. The measurements for each flower include sepal length, sepal width, petal length, and petal width.

The original dataset consists of 150 classical observations. Suppose that each set of five consecutive flowers listed in the original dataset came from the same location. Further suppose that we are interested in the characteristics of groups of flowers by location instead of features of individual flowers. Then, each set of five classical observations can be aggregated into one interval-valued observation. The resulting interval-valued dataset consists of ten observations of the setosa species, ten of the versicolor species, and ten of the virginica species. There are four variables in this dataset: $X_{(1)}$ = Sepal Length, $X_{(2)}$ = Sepal Width, $X_{(3)}$ = Petal Length, and $X_{(4)}$ = Petal Width. Table 3.1 shows the interval-valued Iris data obtained from Billard and Diday (2006).

We first computed the symbolic variance-covariance matrix as defined in Equation (3.1) of Section 3.2.1 for the data in Table 3.1. The resulting covariance matrix is given by

$$\boldsymbol{S} = \begin{bmatrix} 0.6007 & -0.1275 & 1.2369 & 0.5271 \\ & 0.1675 & -0.4471 & -0.1854 \\ & & 2.9937 & 1.2711 \\ & & & 0.5592 \end{bmatrix}. \tag{3.20}$$

Table 3.1: Interval-Valued Iris Data

i	**Species**	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
1	S1	[4.6, 5.1]	[3.0, 3.6]	[1.3, 1.5]	[0.2, 0.2]
2	S2	[4.4, 5.4]	[2.9, 3.9]	[1.4, 1.7]	[0.1, 0.4]
3	S3	[4.3, 5.8]	[3.0, 4.0]	[1.1, 1.6]	[0.1, 0.2]
4	S4	[5.1, 5.7]	[3.5, 4.4]	[1.3, 1.7]	[0.3, 0.4]
5	S5	[4.6, 5.4]	[3.3, 3.7]	[1.0, 1.9]	[0.2, 0.5]
6	S6	[4.7, 5.2]	[3.0, 3.5]	[1.4, 1.6]	[0.2, 0.4]
7	S7	[4.8, 5.5]	[3.1, 4.2]	[1.4, 1.6]	[0.1, 0.4]
8	S8	[4.4, 5.5]	[3.0, 3.5]	[1.2, 1.5]	[0.1, 0.2]
9	S9	[4.4, 5.1]	[2.3, 3.8]	[1.3, 1.9]	[0.2, 0.6]
10	S10	[4.6, 5.3]	[3.0, 3.8]	[1.4, 1.6]	[0.2, 0.3]
11	Ve1	[5.5, 7.0]	[2.3, 3.2]	[4.0, 4.9]	[1.3, 1.5]
12	Ve2	[4.9, 6.6]	[2.4, 3.3]	[3.3, 4.7]	[1.0, 1.6]
13	Ve3	[5.0, 6.1]	[2.0, 3.0]	[3.5, 4.7]	[1.0, 1.5]
14	Ve4	[5.6, 6.7]	[2.2, 3.1]	[3.9, 4.5]	[1.0, 1.5]
15	Ve5	[5.9, 6.4]	[2.5, 3.2]	[4.0, 4.9]	[1.2, 1.8]
16	Ve6	[5.7, 6.8]	[2.6, 3.0]	[3.5, 5.0]	[1.0, 1.7]
17	Ve7	[5.4, 6.0]	[2.4, 3.0]	[3.7, 5.1]	[1.0, 1.6]
18	Ve8	[5.5, 6.7]	[2.3, 3.4]	[4.0, 4.7]	[1.3, 1.6]
19	Ve9	[5.0, 6.1]	[2.3, 3.0]	[3.3, 4.6]	[1.0, 1.4]
20	Ve10	[5.1, 6.2]	[2.5, 3.0]	[3.0, 4.3]	[1.1, 1.3]
21	Vi1	[5.8, 7.1]	[2.7, 3.3]	[5.1, 6.0]	[1.8, 2.5]
22	Vi2	[4.9, 7.6]	[2.5, 3.6]	[4.5, 6.6]	[1.7, 2.5]
23	Vi3	[5.7, 6.8]	[2.5, 3.2]	[5.0, 5.5]	[1.9, 2.4]
24	Vi4	[6.0, 7.7]	[2.2, 3.8]	[5.0, 6.9]	[1.5, 2.3]
25	Vi5	[5.6, 7.7]	[2.7, 3.3]	[4.9, 6.7]	[1.8, 2.3]
26	Vi6	[6.1, 7.2]	[2.8, 3.2]	[4.8, 6.0]	[1.6, 2.1]
27	Vi7	[6.1, 7.9]	[2.6, 3.8]	[5.1, 6.4]	[1.4, 2.2]
28	Vi8	[6.0, 7.7]	[3.0, 3.4]	[4.8, 6.1]	[1.8, 2.4]
29	Vi9	[5.8, 6.9]	[2.7, 3.3]	[5.1, 5.9]	[1.9, 2.5]
30	Vi10	[5.0, 6.7]	[2.5, 3.4]	[5.0, 5.4]	[1.8, 2.3]

Matrix $\boldsymbol{S}$ of Equation (3.20) shows that Petal Length, $X_{(3)}$, has the largest variance (2.9937) whereas Sepal Width, $X_{(2)}$, has the smallest variance (0.1675). The variance for Sepal Length, $X_{(1)}$, and variance for Petal Width, $X_{(4)}$, are about the same. The symbolic correlation matrix as defined in Equation (3.4) of section 3.2.1 for the Iris data is

$$\boldsymbol{R} = \begin{bmatrix} 1 & -0.4019 & 0.9224 & 0.9095 \\ & 1 & -0.6313 & -0.6056 \\ & & 1 & 0.9825 \\ & & & 1 \end{bmatrix}. \tag{3.21}$$

The elements of the symbolic correlation matrix of Equation (3.21) indicate a strong correlation between Sepal Length, Petal Length, and Petal Width. The coefficient of correlation between Sepal Length and Petal Length is 0.9224. The coefficient of correlation between Sepal Length and Petal Width is 0.9095. Petal Length has an almost perfect correlation to Petal Width with a correlation coefficient of 0.9825. The correlation matrix of Equation (3.21) also shows some negative correlation between Sepal Width and Petal Length and Petal Width with coefficients of -0.6313 and -0.6056, respectively. With knowledge of the covariance structure, we can find the principal components of the data of Table 3.1. The symbolic PCA results are presented in the following subsection.

Analysis results

For the first part of this analysis, we computed the coefficients for the principal components of the interval-valued Iris data using the correlation structure as described in Section 3.2.1. These coefficients are shown in Table 3.2 along with the proportion of total variance explained by each principal component. The first principal component is mainly composed of Sepal Length, Petal Length, and Petal Width. These three variables give approximately equal contribution to the first principal component with coefficient of 0.5072, 0.5473, and 0.5423, respectively. The first principal component explains 81.84% of the total variation in the data. The second principal component is composed of mostly Sepal Width with a coefficient of

Table 3.2: Principal Component Coefficients and Variance Proportion of Iris Data Based on Symbolic Covariance Method

Variable	PC1	PC2	PC3	PC4
Sepal Length	0.5072	0.4372	0.7231	0.1694
Sepal Width	-0.3861	0.8855	-0.2460	-0.0788
Petal Length	0.5473	0.0968	-0.2573	-0.7905
Petal Width	0.5423	0.1239	-0.5919	0.5833
Proportion of Variance	0.8184	0.1620	0.0160	0.0036
Cumulative Proportion	0.8184	0.9804	0.9964	1.0000

0.885. The second principal component explains another 16.20% of the total variation in the data. With the last two principal components explaining less than 2% of the total variation, they can be excluded from further analysis without causing significant loss of information to the data. Thus, the dimension of the Iris dataset can be reduced from four to two.

Next, we plot the observations onto the space spanned by the first two principal components as described in the algorithm of Section 3.3.1. A plot of the observations along the first and the second principal component axis is shown in Figure 3.27. In this plot, observations are colored according to species. Black represents setosa, red versicolor, and green virginica.

Looking along the first principal component (PC1) axis, observations belonging to the species setosa clearly stand away from the other species. They form a distinct cluster of observations with small PC1 values. Since Sepal Length, Petal Length, and Petal Width are the major contributors to the first principal component, all with positive coefficient, it means iris flowers with short sepal and short and narrow petal have small PC1 values whereas flowers with long sepal and long and wide petal have large PC1 values. That is, irises of the setosa species generally have shorter sepal, shorter and narrower petal than irises of the versicolor and virginica species. A closer look at the red and green polygons along the PC1-axis shows some overlap between these two groups. However, the red polygons generally have smaller

PC1 values than the black polygons. That means, irises of species versicolor are generally smaller than those of virginica.

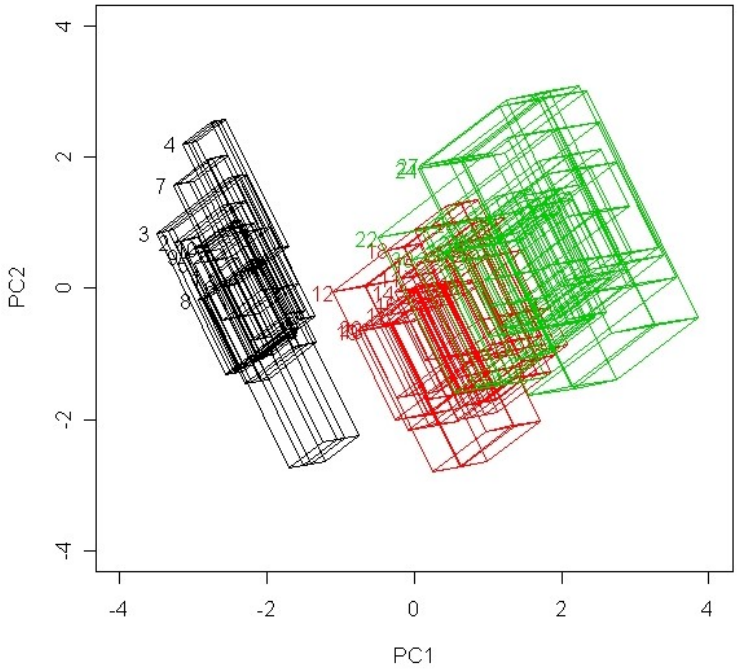

Figure 3.27: Plot of PC1 × PC2 for Iris Data Based on Symbolic Covariance Method (Color Represents Species)

Studying the observations along the second principal component (PC2) axis shows no distinct clusters. The irises of species setosa, represented by black polygons, cover a wide range of PC2 values. Since PC2 is composed of mainly Sepal Width, it means sepal width of setosa irises varies. Moreover, the red polygons generally have smaller PC2 values than do the green polygons. Since Sepal Width contributes a positive coefficient to the second principal component, it means irises belonging to species versicolor generally have narrower sepal width than do those that belong to the species virginica.

The correlation between a random variable $X_{(j)}$ and a principal component $Y_{(k)}$ for $j, k = 1, \ldots, p$, as defined in Equation (2.25) provides another way to understand the relationship between $X_{(j)}$ and $Y_{(k)}$. Table 3.3 shows the relationship between $X_{(j)}$ and $Y_{(k)}$ in the absence of all other variables.

Table 3.3: Correlation between Principal Components and Random Variables of Iris Data Based on Symbolic Covariance Method

Variable	PC1	PC2	PC3	PC4
Sepal Length	0.9177	0.3519	0.1831	0.0204
Sepal Width	-0.6986	0.7128	-0.0623	-0.0095
Petal Length	0.9903	0.0779	-0.0651	-0.0951
Petal Width	0.9812	0.0997	-0.1499	0.0702

Tables 3.3 indicates strong correlation between PC1 and Sepal Length, Petal Length, and Petal Width with correlation coefficients of 0.9177, 0.9903, and 0.9812, respectively. These correlations agree with the observations based solely on the coefficients of the principal components shown in Table 3.2. There also exists a negative correlation between Sepal Width and the first principal component with a coefficient of -0.6986. In addition, Sepal Width is positively correlated with PC2 with a coefficient of 0.7128. Although the magnitude of the coefficient of Sepal Width for the first principal component is 0.3861 compared to its coefficient for the second principal component of 0.8855, the magnitude of the correlation between Sepal Width and PC1 is almost equal to its correlation to PC2 when other variables are ignored. This is due to the fact that PC1 explains almost 82% of the total variation in the data whereas PC2 only explains 16% of the total data variation.

Now, if numerical values for the principal components are required for further analysis, histogram-valued principal components of the Iris dataset can be computed by the algorithm described in Section 3.3.2. Table 3.4 gives histograms for the first principal component of the Iris data of Table 3.1. Table 3.5 gives histograms for the second principal component of the Iris data. The relative frequencies of the histograms are computed based on area of the polygons shown in Figure 3.27. Since the first two principal components take care of 98% of the variation in the data, these histograms reflect almost all of the internal variation of the observations in the principal components space.

Table 3.4: Histogram for the First Principal Component of Iris Data Based on Symbolic Covariance Method

i	**Frequency Histogram for the First Principal Component**
1	$\{[-2.80,-2.47),0.25;[-2.47,-2.41),0.10;[-2.41,-2.23),0.30;[-2.23,-2.17),0.10;$ $[-2.17,-1.84],0.25\}$
2	$\{[-3.25,-2.60),0.24;[-2.60,-2.38),0.18;[-2.38,-2.31),0.07;[-2.31,-2.28),0.02;$ $[-2.28,-2.21),0.07;[-2.21,-2.00),0.18;[-2.00,-1.34],0.24\}$
3	$\{[-3.51,-2.56),0.39;[-2.56,-2.52),0.03;[-2.52,-2.45),0.06;[-2.45,-2.40),0.04;$ $[-2.40,-2.33),0.06;[-2.33,-2.29),0.03;[-2.29,-1.35],0.39\}$
4	$\{[-3.15,-2.76),0.16;[-2.76,-2.69),0.06;[-2.69,-2.56),0.13;[-2.56,-2.30),0.29;$ $[-2.30,-2.18),0.13;[-2.18,-2.10),0.06;[-2.10,-1.71],0.16\}$
5	$\{[-2.99,-2.61),0.18;[-2.61,-2.46),0.14;[-2.46,-2.32),0.14;[-2.32,-2.24),0.08;$ $[-2.24,-2.11),0.14;[-2.11,-1.96),0.14;[-1.96,-1.58],0.18\}$
6	$\{[-2.60,-2.28),0.22;[-2.28,-2.13),0.23;[-2.13,-2.13),0.00;[-2.13,-2.07),0.11;$ $[-2.07,-2.07),0.00;[-2.07,-1.92),0.23;[-1.92,-1.60],0.22\}$
7	$\{[-3.27,-2.81),0.14;[-2.81,-2.60),0.16;[-2.60,-2.53),0.06;[-2.53,-2.23),0.28;$ $[-2.23,-2.17),0.06;[-2.17,-1.95),0.16;[-1.95,-1.50],0.14\}$
8	$\{[-2.94,-2.47),0.26;[-2.47,-2.37),0.11;[-2.37,-2.30),0.08;[-2.30,-2.22),0.10;$ $[-2.22,-2.14),0.08;[-2.14,-2.05),0.11;[-2.05,-1.58],0.26\}$
9	$\{[-3.12,-2.66),0.09;[-2.66,-2.37),0.14;[-2.37,-2.18),0.12;[-2.18,-1.70),0.32;$ $[-1.70,-1.51),0.12;[-1.51,-1.22),0.14;[-1.22,-0.76],0.09\}$
10	$\{[-2.95,-2.50),0.24;[-2.50,-2.42),0.08;[-2.42,-2.36),0.08;[-2.36,-2.20),0.21;$ $[-2.20,-2.14),0.08;[-2.14,-2.06),0.08;[-2.06,-1.60],0.24\}$
11	$\{[-0.18,0.67),0.30;[0.67,0.80),0.09;[0.80,0.95),0.11;[0.95,0.96),0.01;$ $[0.96,1.10),0.11;[1.10,1.23),0.09;[1.23,2.08],0.30\}$

Continued on next page

Table 3.4 – continued from previous page

i	Frequency Histogram for the First Principal Component
12	$\{[-1.10, -0.26), 0.21; [-0.26, 0.01), 0.13; [0.01, 0.19), 0.09; [0.19, 0.44), 0.13;$ $[0.44, 0.62), 0.09; [0.62, 0.89), 0.13; [0.89, 1.74], 0.21\}$
13	$\{[-0.69, 0.03), 0.19; [0.03, 0.25), 0.13; [0.25, 0.39), 0.09; [0.39, 0.63), 0.16;$ $[0.63, 0.77), 0.09; [0.77, 0.99), 0.13; [0.99, 1.71], 0.19\}$
14	$\{[-0.27, 0.45), 0.24; [0.45, 0.58), 0.09; [0.58, 0.77), 0.15; [0.77, 0.81), 0.03;$ $[0.81, 1.00), 0.15; [1.00, 1.13), 0.09; [1.13, 1.85], 0.24\}$
15	$\{[0.01, 0.34), 0.08; [0.34, 0.67), 0.24; [0.67, 0.77), 0.09; [0.77, 0.96), 0.17;$ $[0.96, 1.06), 0.09; [1.06, 1.39), 0.24; [1.39, 1.72], 0.08\}$
16	$\{[-0.23, 0.14), 0.10; [0.14, 0.49), 0.20; [0.49, 0.62), 0.08; [0.62, 0.99), 0.24;$ $[0.99, 1.13), 0.08; [1.13, 1.47), 0.20; [1.47, 1.85], 0.10\}$
17	$\{[-0.37, 0.03), 0.11; [0.03, 0.20), 0.12; [0.20, 0.46), 0.20; [0.46, 0.64), 0.14;$ $[0.64, 0.90), 0.20; [0.90, 1.08), 0.12; [1.08, 1.47], 0.11\}$
18	$\{[-0.37, 0.42), 0.25; [0.42, 0.64), 0.15; [0.64, 0.67), 0.03; [0.67, 0.86), 0.15;$ $[0.86, 0.89), 0.03; [0.89, 1.11), 0.15; [1.11, 1.90], 0.25\}$
19	$\{[-0.76, -0.10), 0.23; [-0.10, -0.04), 0.04; [-0.04, 0.25), 0.21; [0.25, 0.32), 0.04;$ $[0.32, 0.61), 0.21; [0.61, 0.67), 0.04; [0.67, 1.33], 0.23\}$
20	$\{[-0.71, -0.24), 0.18; [-0.24, 0.01), 0.20; [0.01, 0.15), 0.12; [0.15, 0.17), 0.01;$ $[0.17, 0.32), 0.12; [0.32, 0.56), 0.20; [0.56, 1.04], 0.18\}$
21	$\{[0.63, 1.20), 0.16; [1.20, 1.49), 0.17; [1.49, 1.49), 0.00; [1.49, 1.99), 0.33;$ $[1.99, 1.99), 0.00; [1.99, 2.28), 0.17; [2.28, 2.84], 0.16\}$
22	$\{[-0.50, 0.54), 0.16; [0.54, 1.20), 0.22; [1.20, 1.27), 0.02; [1.27, 1.78), 0.19;$ $[1.78, 1.85), 0.02; [1.85, 2.51), 0.22; [2.51, 3.55], 0.16\}$
23	$\{[0.70, 1.36), 0.26; [1.36, 1.42), 0.05; [1.42, 1.52), 0.08; [1.52, 1.79), 0.22;$ $[1.79, 1.89), 0.08; [1.89, 1.94), 0.05; [1.94, 2.61], 0.26\}$

Continued on next page

Table 3.4 – continued from previous page

i	**Frequency Histogram for the First Principal Component**
24	$\{[0.04, 1.16), 0.18; [1.16, 1.55), 0.15; [1.55, 1.74), 0.08; [1.74, 2.15), 0.18;$ $[2.15, 2.34), 0.08; [2.34, 2.73), 0.15; [2.73, 3.85], 0.18\}$
25	$\{[0.44, 1.01), 0.11; [1.01, 1.58), 0.24; [1.58, 1.81), 0.12; [1.81, 1.94), 0.06;$ $[1.94, 2.18), 0.12; [2.18, 2.75), 0.24; [2.75, 3.31], 0.11\}$
26	$\{[0.69, 1.06), 0.12; [1.06, 1.41), 0.23; [1.41, 1.44), 0.03; [1.44, 1.77), 0.24;$ $[1.77, 1.80), 0.03; [1.80, 2.15), 0.23; [2.15, 2.52], 0.12\}$
27	$\{[0.07, 1.20), 0.26; [1.20, 1.25), 0.02; [1.25, 1.61), 0.17; [1.61, 1.83), 0.10;$ $[1.83, 2.19), 0.17; [2.19, 2.24), 0.02; [2.24, 3.37], 0.26\}$
28	$\{[0.58, 0.95), 0.08; [0.95, 1.36), 0.20; [1.36, 1.69), 0.19; [1.69, 1.80), 0.07;$ $[1.80, 2.12), 0.19; [2.12, 2.54), 0.20; [2.54, 2.91], 0.08\}$
29	$\{[0.71, 1.27), 0.20; [1.27, 1.43), 0.11; [1.43, 1.53), 0.07; [1.53, 1.86), 0.25;$ $[1.86, 1.96), 0.07; [1.96, 2.12), 0.11; [2.12, 2.68], 0.20\}$
30	$\{[-0.02, 0.83), 0.26; [0.83, 0.96), 0.08; [0.96, 1.10), 0.09; [1.10, 1.32), 0.15;$ $[1.32, 1.46), 0.09; [1.46, 1.59), 0.08; [1.59, 2.44], 0.26\}$

Table 3.5: Histogram for the Second Principal Component of Iris Data Based on Symbolic Covariance Method

i	**Frequency Histogram for the Second Principal Component**
1	$\{[-1.11, -0.82), 0.12; [-0.82, 0.18), 0.76; [0.18, 0.46], 0.12\}$
2	$\{[-1.39, -0.83), 0.16; [-0.83, 0.71), 0.67; [0.71, 1.27], 0.16\}$
3	$\{[-1.27, -0.42), 0.22; [-0.42, 0.85), 0.56; [0.85, 1.70], 0.22\}$
4	$\{[0.30, 0.64), 0.11; [0.64, 2.21), 0.78; [2.21, 2.55], 0.11\}$

Continued on next page

Table 3.5 – continued from previous page

i	Frequency Histogram for the First Principal Component
5	$\{[-0.48,-0.43),0.01;[-0.43,-0.38),0.02;[-0.38,0.07),0.32;[0.07,0.38),0.31;$ $[0.38,0.83),0.32;[0.83,0.88),0.02;[0.88,0.93],0.01\}$
6	$\{[-1.04,-1.01),0.01;[-1.01,-0.73),0.16;[-0.73,0.03),0.66;[0.03,0.31),0.16;$ $[0.31,0.34],0.01\}$
7	$\{[-0.74,-0.34),0.11;[-0.34,1.58),0.78;[1.58,1.98],0.11\}$
8	$\{[-1.22,-0.60),0.31;[-0.60,-0.17),0.37;[-0.17,0.45],0.31\}$
9	$\{[-2.71,-2.64),0.01;[-2.64,-2.25),0.09;[-2.25,0.50),0.82;[0.50,0.90),0.09;$ $[0.90,0.96],0.01\}$
10	$\{[-1.08,-0.69),0.13;[-0.69,0.62),0.73;[0.62,1.01],0.13\}$
11	$\{[-1.71,-0.86),0.25;[-0.86,0.16),0.48;[0.16,1.00],0.25\}$
12	$\{[-1.92,-1.82),0.01;[-1.82,-0.86),0.30;[-0.86,-0.05),0.37;[-0.05,0.90),0.30;$ $[0.90,1.00],0.01\}$
13	$\{[-2.73,-2.65),0.01;[-2.65,-2.03),0.19;[-2.03,-0.64),0.59;[-0.64,-0.02),0.19;$ $[-0.02,0.07],0.01\}$
14	$\{[-1.97,-1.89),0.01;[-1.89,-1.27),0.20;[-1.27,-0.06),0.57;[-0.06,0.56),0.20;$ $[0.56,0.64],0.01\}$
15	$\{[-1.10,-1.00),0.03;[-1.00,-0.72),0.14;[-0.72,0.37),0.66;[0.37,0.65),0.14;$ $[0.65,0.75],0.03\}$
16	$\{[-1.11,-1.02),0.01;[-1.02,-0.91),0.04;[-0.91,-0.29),0.43;[-0.29,-0.24),0.04;$ $[-0.24,0.38),0.43;[0.38,0.50),0.04;[0.50,0.58],0.01\}$
17	$\{[-1.70,-1.62),0.01;[-1.62,-1.52),0.03;[-1.52,-1.18),0.19;[-1.18,-0.40),0.54;$ $[-0.40,-0.06),0.19;[-0.06,0.04),0.03;[0.04,0.12],0.01\}$
18	$\{[-1.70,-1.02),0.17;[-1.02,0.59),0.64;[0.59,1.27],0.17\}$
19	$\{[-2.16,-2.09),0.01;[-2.09,-2.02),0.01;[-2.02,-1.40),0.26;[-1.40,-0.65),0.44;$

Continued on next page

Table 3.5 – continued from previous page

i	**Frequency Histogram for the First Principal Component**
	$[-0.65, -0.03), 0.26; [-0.03, 0.04), 0.01; [0.04, 0.11], 0.01\}$
20	$\{[-1.67, -1.60), 0.01; [-1.60, -1.57), 0.01; [-1.57, -0.95), 0.34; [-0.95, -0.59), 0.28;$ $[-0.59, 0.03), 0.34; [0.03, 0.06), 0.01; [0.06, 0.14], 0.01\}$
21	$\{[-0.56, -0.44), 0.02; [-0.44, 0.29), 0.34; [0.29, 0.69), 0.26; [0.69, 1.42), 0.34;$ $[1.42, 1.54], 0.02\}$
22	$\{[-1.48, -1.35), 0.01; [-1.35, 0.17), 0.37; [0.17, 0.78), 0.22; [0.78, 2.30), 0.37;$ $[2.30, 2.43], 0.01\}$
23	$\{[-1.06, -0.98), 0.01; [-0.98, -0.36), 0.25; [-0.36, 0.43), 0.47; [0.43, 1.05), 0.25;$ $[1.05, 1.13], 0.01\}$
24	$\{[-1.53, -1.40), 0.01; [-1.40, -0.44), 0.18; [-0.44, 1.83), 0.60; [1.83, 2.78), 0.18;$ $[2.78, 2.92], 0.01\}$
25	$\{[-0.74, -0.63), 0.01; [-0.63, -0.55), 0.01; [-0.55, 0.56), 0.46; [0.56, 0.63), 0.04;$ $[0.63, 1.75), 0.46; [1.75, 1.83), 0.01; [1.83, 1.93], 0.01\}$
26	$\{[-0.28, -0.21), 0.01; [-0.21, -0.13), 0.03; [-0.13, 0.49), 0.42; [0.49, 0.59), 0.09;$ $[0.59, 1.21), 0.42; [1.21, 1.29), 0.03; [1.29, 1.36], 0.01\}$
27	$\{[-0.65, -0.52), 0.01; [-0.52, 0.50), 0.25; [0.50, 1.87), 0.48; [1.87, 2.89), 0.25;$ $[2.89, 3.02], 0.01\}$
28	$\{[0.13, 0.21), 0.01; [0.21, 0.31), 0.02; [0.31, 1.00), 0.37; [1.00, 1.26), 0.20;$ $[1.26, 1.96), 0.37; [1.96, 2.06), 0.02; [2.06, 2.13], 0.01\}$
29	$\{[-0.55, -0.45), 0.02; [-0.45, 0.17), 0.30; [0.17, 0.70), 0.36; [0.70, 1.32), 0.30;$ $[1.32, 1.42], 0.02\}$
30	$\{[-1.48, -1.40), 0.01; [-1.40, -0.44), 0.28; [-0.44, 0.45), 0.42; [0.45, 1.41), 0.28;$ $[1.41, 1.49], 0.01\}$

Comparison of symbolic covariance, vertices, and centers methods

To understand the differences between our proposed method and current PCA methods for interval-valued observations, we apply the vertices and the centers PCA methods to the Iris dataset of Table 3.1. Results from these methods and their performance compared to the proposed symbolic covariance method are discussed in this subsection. We divide the comparison into two parts. First, we will examine the differences in the coefficients of the principal components. Then, we will study the differences between the principal components constructed from these coefficients.

Of the current PCA methods for interval-valued observations, the vertices method is the most comparable to our proposed method so we begin with the vertices method. First, we compare the covariance and the correlation matrices of the data because they determine the coefficients for the principal components. In the vertices method, the data matrix $\boldsymbol{X}$ is transformed into a matrix of vertices $\boldsymbol{X}^v$ as defined in Equation (2.31). Then, the classical covariance and correlation matrices of $\boldsymbol{X}^v$ are computed. The covariance matrix and the correlation matrix of the interval-valued Iris data are given in the following,

$$\boldsymbol{S} = \begin{bmatrix} 0.8763 & -0.0928 & 1.0865 & 0.4526 \\ & 0.2910 & -0.3405 & -0.1355 \\ & & 3.1788 & 1.2300 \\ & & & 0.5983 \end{bmatrix} \tag{3.22}$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & -0.1838 & 0.6510 & 0.6251 \\ & 1 & -0.3540 & -0.3247 \\ & & 1 & 0.8919 \\ & & & 1 \end{bmatrix}. \tag{3.23}$$

All diagonal elemements of Equation (3.22) have values larger than their counterpart in Equation (3.20) as expected. This is due to the fact that the vertices method treats all vertices of $\boldsymbol{X}^v$ as independent observations. The covariance matrix in Equation (3.22) gives

the covariance of $(2^p)n = (2^4)30 = (16)(30) = 480$ data points instead of 30 data points as in the symbolic covariance matrix of Equation (3.20). Moreover, the variances of Equation (3.22) are the variances of the vertices of the hyper-rectangles representing observations in the dataset. The vertices are the most extreme points of a hyper-rectangle and variances of the extreme points are expected to be larger than variance of all points belonging to the hyper-rectangle.

The correlation coefficients shown in Equation (3.23) indicate less correlation between variables based on the vertices than the correlation between variables based on the symbolic covariance structure shown in Equation (3.21). The largest correlation occurs between vertices of Petal Length and Petal Width with coefficient of 0.8919. This coefficient is much less than the coefficient of 0.9825 in the symbolic correlation matrix of Equation (3.21). Moreover, the correlation coefficients between Sepal Length and Petal Length and Petal Width are reduced from 0.9224 and 0.9095, respectively, in the symbolic correlation of Equation (3.21) to 0.6510 and 0.6251, respectively, in the vertices correlation of Equation (3.23).

Coefficients of the principal components resulting from the correlation structure of the vertices shown in Equation (3.23) are listed in Table 3.6. Elements whose values are close to zero are left blank in Table 3.6. In the vertices method, the first principal component explains only 65.28% of the total variation in the data compared to 81.84% in the symbolic covariance method. With the second principal component explains an additional 21.56% and the third principal component explains an additional 10.50% of the total variance, the first three principal components of the vertices method explains 97% of the total variation in the data compared to the 98% explained by the first two principal components in the symbolic covariance method.

More specifically, in the vertices method Petal Length and Petal Width contribute a much larger share to the first principal component than does Sepal Length with coefficients of 0.5820 (Petal Length) and 0.5740 (Petal Width) versus 0.4920 (Sepal Length). The share is more comparable in the symbolic covariance method with coefficients of 0.5473 (Petal

Table 3.6: Principal Component Coefficients and Variance Proportion of Iris Data Based on Vertices Method

Variable	PC1	PC2	PC3	PC4
Sepal Length	0.4920	-0.3370	0.8010	
Sepal Width	-0.3000	-0.9310	-0.2060	
Petal Length	0.5820		-0.3520	-0.7280
Petal Width	0.5740	-0.1100	-0.4380	0.6830
Proportion of Variance	0.6528	0.2156	0.1050	0.0267
Cumulative Variance	0.6528	0.8684	0.9733	1.0000

Table 3.7: Correlation between Principal Components and Random Variables of Iris Data Based on Vertices Method

Variable	PC1	PC2	PC3	PC4
Sepal Length	0.7951	-0.3130	0.5193	
Sepal Width	-0.4843	-0.8646	-0.1334	
Petal Length	0.9408		-0.2280	-0.2377
Petal Width	0.9269	-0.1025	-0.2836	0.2232

Length) and 0.5423 (Petal Width) versus 0.5072 (Sepal Length). The reduction in contribution of Sepal Length to the first principal component in the vertices method is probably a result of its reduced correlation to Petal Length and Petal Width.

In the absence of other variables, the correlation coefficient between Sepal Length and the first principal component shown in Table 3.7 is also reduced significantly from 0.9177 in the symbolic covariance method to 0.7951 in the vertices method. This observation agrees with other differences based on the correlation structure and the coefficients of the principal components discussed in the previous paragraph.

Now, we compare the covariance and the correlation structure of the centers method to the symbolic covariance method. In the centers method, the classical covariance is computed using the midpoints of the observations as discussed in Section 2.3.1. The diagonal elements

of the covariance matrix of the centers in Equation (3.24) are smaller then their counterpart in the symbolic covariance of Equation (3.20) as expected;

$$\boldsymbol{S} = \begin{bmatrix} 0.4630 & -0.0928 & 1.0865 & 0.4526 \\ & 0.1058 & -0.3405 & -0.1355 \\ & & 2.9011 & 1.2300 \\ & & & 0.5396 \end{bmatrix}. \tag{3.24}$$

A variance in Equation (3.24) reflects the variation between observations whereas a variance in Equation (3.20) reflects the variation between observations and variation within observation as discussed in Section 2.1.2 of chapter two and detailed in Billard (2007).

The correlation coefficients of the centers shown in the matrix of Equation (3.25) have values almost equivalent to their counterpart in the symbolic correlation matrix of Equation (3.21);

$$\boldsymbol{R} = \begin{bmatrix} 1 & -0.4195 & 0.9375 & 0.9056 \\ & 1 & -0.6147 & -0.5671 \\ & & 1 & 0.9830 \\ & & & 1 \end{bmatrix}. \tag{3.25}$$

Therefore, the correlation structure for the centers and symbolic methods are similar in the Iris dataset.

The coefficients of the principal components based on the centers method are shown in Table 3.8, along with the proportion of variance and the cumulative proportion of variance explained by the principal components. The composition of the principal components as shown in Table 3.2 and Table 3.8 reflects the similarity in the correlation structure of the centers and the symbolic covariance method. The first principal component in the centers method accounts for 81.60% of the total variation compared to 81.84% in the symbolic covariance method. In both methods, the first principal component is composed of Sepal Length, Petal Length, and Petal Width, each with an approximately equal contribution.

Table 3.8: Principal Component Coefficients and Variance Proportion of Iris Data Based on Centers Method

Variable	PC1	PC2	PC3	PC4
Sepal Length	0.5130	-0.3880	0.7200	0.2600
Sepal Width	-0.3790	-0.9020	-0.1780	-0.1030
Petal Length	0.5500	-0.1080	-0.1560	-0.8140
Petal Width	0.5390	-0.1560	-0.6520	0.5100
Proportion of Variance	0.8160	0.1621	0.0201	0.0018
Cumulative Variance	0.8160	0.9781	0.9982	1.0000

Table 3.9: Correlation between Principal Components and Random Variables of Iris Data Based on Centers Method

Variable	PC1	PC2	PC3	PC4
Sepal Length	0.9276	-0.3121	0.2043	0.0217
Sepal Width	-0.6853	-0.7264	-0.0505	-0.0086
Petal Length	0.9929	-0.0867	-0.0444	-0.0682
Petal Width	0.9738	-0.1256	-0.1849	0.0427

The second principal components in both methods are composed of Sepal Width and explain around 16% of the total variation.

The correlations between the random variables and the principal components for the centers method are shown in Table 3.9. A comparison of Table 3.3 and Table 3.9 shows further similarity in the two methods for this dataset. There exists strong correlation between the first principal component and Sepal Length, Petal Length, and Petal Width with coefficients of 0.9177, 0.9903, and 0.9812, respectively, for the symbolic covariance method and 0.9276, 0.9929, and 0.9738, respectively, for the centers method. Furthermore, the correlation between the second principal component and Sepal Width is similar in both methods. The correlation coefficient between PC2 and Sepal Width is 0.7128 for the symbolic covariance method and 0.7264 for the centers method.

Now, we are ready to compare the observations in the principal components space constructed from the proposed symbolic covariance method, the vertices method, and the centers method. As stated in section 2.3.1 and 2.3.2, the principal components resulting from the vertices and the centers methods have interval values. The lower endpoint of the interval representing principal component k of observation i is the minimum value of the transformed vertices of observation i along the k^{th} principal component. The upper endpoint of the interval is the maximum value of the transformed vertices along the PCk-axis. As a result, an observation is represented by a rectangle, called MCAR, as described in Section 3.2.1, on a PCk $\times$ PCk' plot created by the vertices and the centers method.

Figure 3.28 shows a plot of the Iris data along the first and the second principal component based on the vertices method. Figure 3.29 shows a plot of the Iris data along the first and the second principal component resulting from the centers method. Observations in Figures 3.28 and 3.29 are colored according to species similar to the PC1 $\times$ PC2 plot of the Iris data resulting from the symbolic covariance method of Figure 3.27. In these plots, black represents species setosa, red represents versicolor, and green represents virginica.

A pattern of species grouping emerges in the PC1 $\times$ PC2 plots for the vertices and the centers methods. Similar to the pattern reveals in the symbolic covariance PC1 $\times$ PC2 plot of Figure 3.27, observations of species setosa have smaller PC1 value and form their own cluster to the left of observations belonging to the other species. Comparison of Figure 3.28 and Figure 3.29 reveals that along the PC1 axis the cluster formed by the rectangles representing observations of species setosa is more distinct in the vertices method than in the centers method. There exists overlap in PC1 values between observation 9 and observations 12, 13, and 19 in Figure 3.29 of the centers method which does not exist between those observations in the vertices method. Otherwise, the plots of Figures 3.28 and 3.29 show a similar spread of observations along the second principal component axis.

To understand how transformation of the same observations can create overlap along the PC1-axis in one method but not in the other, we need to look at principal component

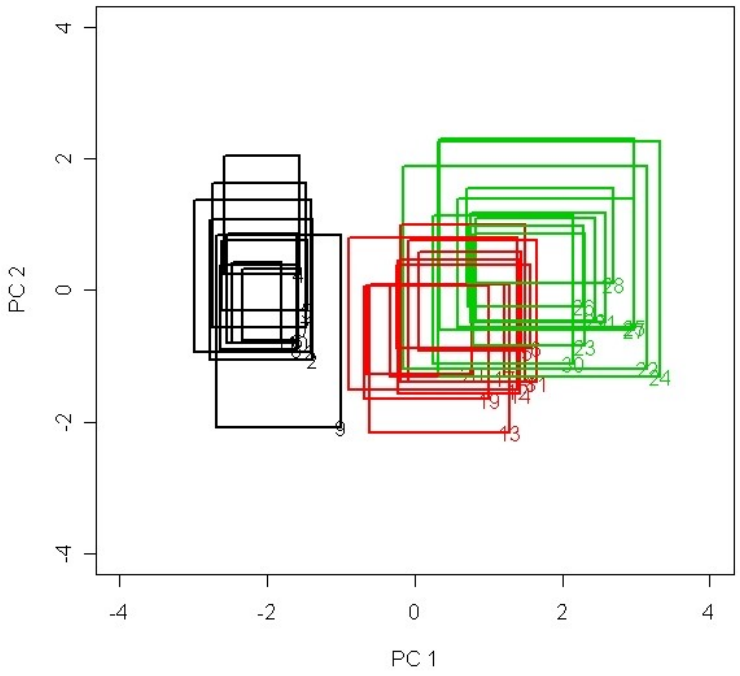

Figure 3.28: Plot of PC1 × PC2 for Iris Data Based on Vertices Method (Color Represents Species)

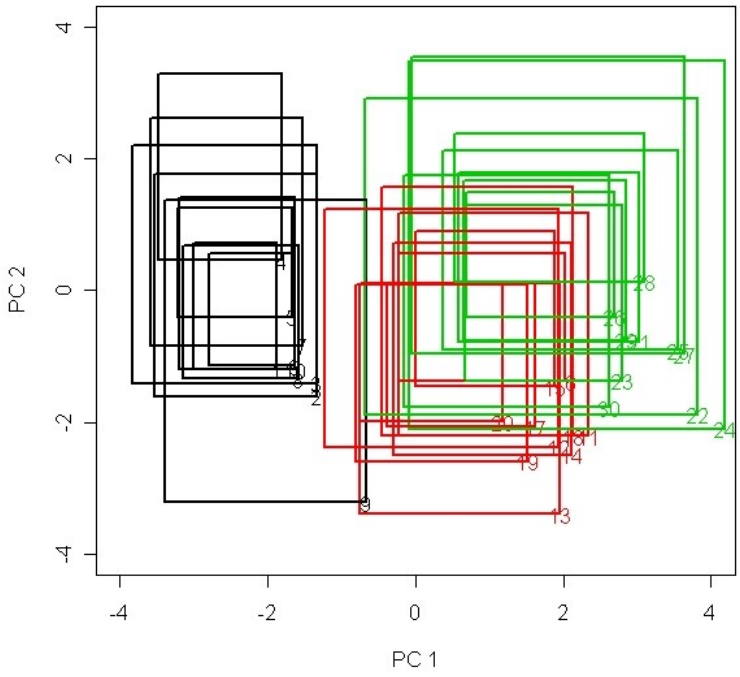

Figure 3.29: Plot of PC1 × PC2 for Iris Data Based on Centers Method (Color Represents Species)

analysis geometrically. A principal components space of a dataset is a coordinate system that maximizes the variances in the data. The principal component axes can be obtained by rotating the axes of the original sample space in such a way that the axis along the first principal component coincides with the direction of the largest variation in the data, the PC2-axis is orthogonal to the PC1-axis and coincides with the direction of the second largest variation, and so on.

The direction of rotation for each principal component axis is determined by the coefficients for that principal component. We discussed earlier in this section that the coefficients for Sepal Length, Petal Length, and Petal Width for PC1 are approximately equal in the centers method whereas the coefficient of Sepal Length is smaller than the coefficients of Petal Length and Petal Width in the vertices method. Therefore, the hyper-rectangles representing observations in the Iris dataset are rotated differently in the principal components spaces resulting from these methods. However, a PC1 $\times$ PC2 plot of an observation based on the vertices and the centers method only shows the MCAR enscribing all vertices of the observation. It does not show the rotation direction of the principal component axes.

The direction of rotation affects the size and shape of the resulting MCAR as illustrated in Figure 3.30. In Figure 3.30 the solid orange and solid green rectangles represent the same interval-valued observation in a two-dimensional space. Two different colors for the rectangles signify two differernt methods of PCA, the vertices and the centers methods, applied to the same observation. The resulting principal components spaces are different for the orange observation and the green observation. Now, if we make a PC1 $\times$ PC2 plot for each method and put one plot on top of the other so that their PC1 and PC2 axes align we obtain the picture in Figure 3.30. We can see that the rectangles have identical size and shape in their respective sample space. That is, if we look at the orange rectangle in the direction of its coordinate axes, annotated in orange and look at the green rectangle in the direction of its coordinate axes, annoted in green; then the shape and size of these two rectangles are the same. Figure 3.30 shows that the rotation applied to the orange rectangle gives it a

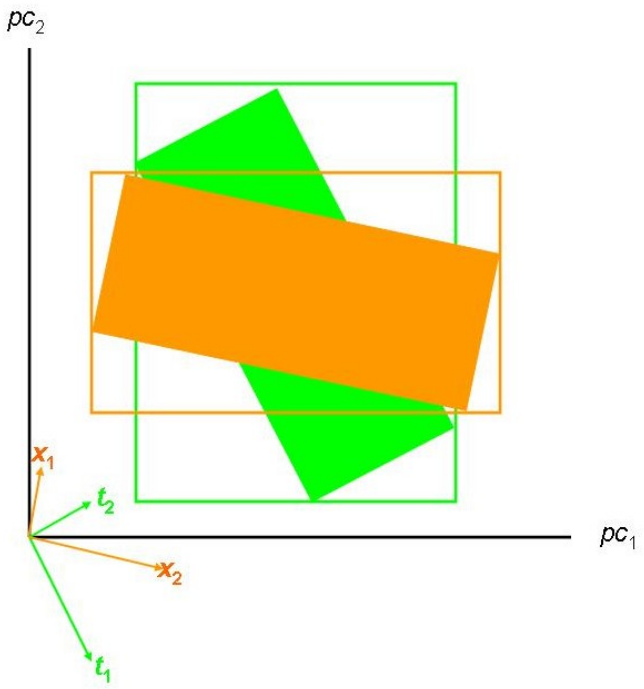

Figure 3.30: Maximum Covering Areas from Two Different Rotations of One Rectangle (Color Represents Original Sample Space)

wider range along the PC1 axis than the rotation applied to the green rectangle. At the same time, the rotation applied to the green rectangle gives it a wider range of values along the PC2 axis than the orange rectangle. The resulting MCAR for the orange rectangle on the PC1 $\times$ PC2 plane, outlined in orange, is wider along the PC1 axis and shorter along the PC2 axis than the MCAR for the green rectangle, outlined in green. The rectangles representing different rotations of the same observation can be significantly different in a principal components space. Therefore, we can see that the overlap between observations in the centers method results from the rotation that stretched the MCAR more along the PC1 axis than the rotation in the vertices method.

The affect of rotation on the shape and the size of the rectangle representing an observation on a PCk_1 $\times$ PCk_2 plane can be minimized when the observation is represented by

a polygon resulting from projecting an observation onto a PCk_1 $\times$ PCk_2 plane as proposed in the symbolic covariance PCA method. In our proposed method, rotating the observations about the origin of the PC1 $\times$ PC2 plane of Figure 3.27 does not change the shape of the polygons. Another drawback of plotting an observation as a rectangle on a PCk_1 $\times$ PCk_2 plane is that, given a maximum covering area rectangle, it is impossible to discern the exact direction of rotation and the shape of the rectangle in its original sample space. Figure 3.31 shows two rectangles that could be rotated to produce the same MCAR outlined in black.

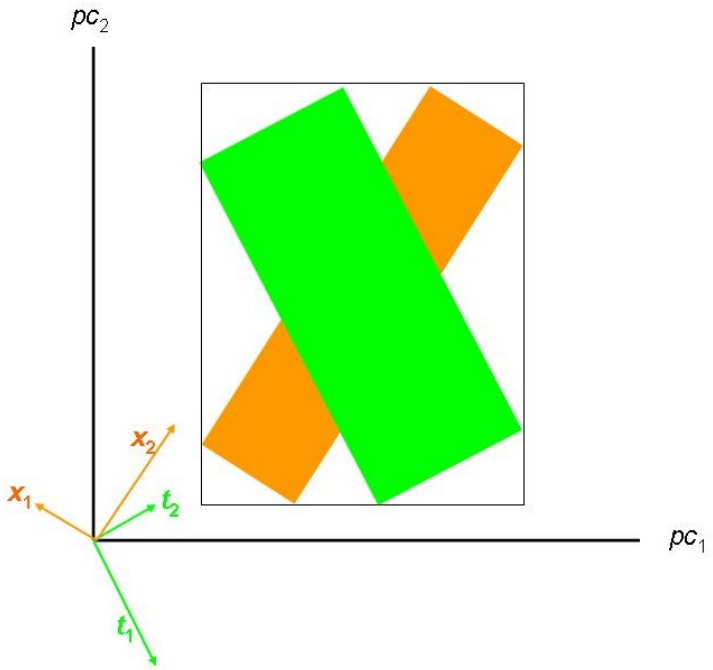

Figure 3.31: Maximum Covering Area from Two Different Rectangles (Color Represents Original Sample Space)

By looking at the rectangles in Figures 3.28 and 3.29, it is impossible to tell the shape and the direction of the observations. However, the polygons constructed by our proposed symbolic covariance method shown in Figure 3.27 give the shape of the observations as well as the direction of rotation of the principal components space in a two-dimensional plane. The plot of Figure 3.27 based on the symbolic covariance method shows clearer clusters of

species and gives better representation of the observations than do the plots of Figures 3.28 and 3.29 resulting from the vertices and the centers methods.

In this subsection, we have shown that the symbolic covariance principal components maximize the total variance of an interval-valued dataset instead of maximizing only part of the variance as in the centers and the vertices methods. We further showed that our proposed symbolic covariance method reconstructs the true structure of the observed data in a principal components space. The PC1 $\times$ PC2 plot of the observations we proposed shows the projection of the true structure of the observations. True projection of the observations eliminates unnecessary overlap between observations caused by the maximum covering area rectangle used in the vertices and the centers methods. Without the unnecessary overlap, our PC1 $\times$ PC2 plot allows the data to show clusters of observations with common features more clearly.

Comparison of symbolic covariance PCA and classical PCA

Having shown the improvements of our proposed method over current PCA methods for interval-valued data, we now compare the symbolic covariance principal components and classical principal components for the Iris data. Prior to the introduction of symbolic data analysis, symbolic observations had to be reduced to a classical data point before analysis could be performed. In the case of interval-valued variables, the midpoint is typically used to represent an interval. Another alternative is to treat the interval endpoints as two independent variables. In this subsection, we will show how these two alternatives do not adequately account for the total structure of interval-valued data.

First, we apply classical PCA to the midpoints of the Iris dataset of Table 3.1. The covariance and the correlation matrix for the midpoints are identical to those of the centers method which are shown in Equations (3.24) and (3.25) because the centers method is based on the covariance structure of the interval midpoints. Similar to the discussion of the centers principal components in the previous subsection, classical principal components using the

midpoints only maximize the variance between observations and ignore the variance within observations.

Next, we apply classical PCA to the interval endpoints of the Iris dataset. Since the endpoints are treated as two independent variables, the number of variables in this analysis doubled to eight, a lower endpoint and an upper endpoint each for Sepal Length, Sepal Width, Petal Length, and Petal Width. The covariance matrix of the endpoints is given by

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.8174 & -0.3261 & -0.4107 & 0.8535 & 0.8267 & 0.7998 & 0.8121 \\ & 1 & -0.3932 & -0.3224 & 0.9106 & 0.9342 & 0.8875 & 0.9044 \\ & & 1 & 0.6318 & -0.5536 & -0.5758 & -0.4810 & -0.5155 \\ & & & 1 & -0.5330 & -0.5422 & -0.5316 & -0.5093 \\ & & & & 1 & 0.9743 & 0.9776 & 0.9815 \\ & & & & & 1 & 0.9533 & 0.9762 \\ & & & & & & 1 & 0.9811 \\ & & & & & & & 1 \end{bmatrix}. \quad (3.26)$$

The covariance matrix of Equation (3.26) indicates strong correlation between the lower and upper endpoints for all variables. The correlation coefficients between the endpoints are 0.8174 for Sepal Length, 0.6318 for Sepal Width, 0.9533 for Petal Length, and 0.9811 for Petal Width. Principal component coefficients and the proportion of variance explained are shown in Table 3.10. As a result of the extremely high correlation between the endpoints, each set of endpoints contribute an approximately identical ammount to the first two principal components. That is, the coefficients for Sepal Length are 0.347 and 0.369. They are 0.399 and 0.398 for Petal Length and 0.390 and 0.394 for Petal Width. Although there are eight variables in this analysis it takes only two principal components to explain 90.27% of the total variation in the data.

We further observe that the composition of the first two principal components based on the endpoints is similar to the composition based on the midpoints. Variables Sepal Length,

Table 3.10: Principal Component Coefficients and Variance Proportion of Iris Data Based on Classical Method Using Endpoints

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Sepal Length	0.3470	-0.2410	0.3240	0.8170	-0.1630		-0.1530	
	0.3690	-0.2860	-0.1860		0.7740	-0.3010		0.2240
Sepal Width	-0.2440	-0.6460	0.6350	-0.3040			0.1160	
	-0.2430	-0.6450	-0.6650	0.1230	-0.2480			
Petal Length	0.3990				-0.2520	-0.2740	0.8310	
	0.3980				0.1960	0.6270		-0.6190
Petal Width	0.3900			-0.3690	-0.3570	-0.4740	-0.4820	-0.3390
	0.3940			-0.2730	-0.2780	0.4560	-0.1680	0.6670
Proportion	0.7680	0.1347	0.0506	0.0289	0.0125	0.0032	0.0012	0.0010
Cumulative	0.7680	0.9027	0.9533	0.9821	0.9947	0.9978	0.9990	1.0000

Table 3.11: Correlation between Principal Components and Random Variables of Iris Data Based on Classical Method Using Endpoints

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Sepal Length	0.8601	-0.2502	0.2061	0.3927	-0.0516		-0.0148	
	0.9146	-0.2969	-0.1183		0.2450	-0.0479		0.0200
Sepal Width	-0.6048	-0.6706	0.4040	-0.1461			0.0112	
	-0.6023	-0.6695	-0.4231	0.0591	-0.0785			
Petal Length	0.9890				-0.0798	-0.0436	0.0805	
	0.9865				0.0621	0.0998		-0.0553
Petal Width	0.9667			-0.1774	-0.1130	-0.0754	-0.0467	-0.0303
	0.9766			-0.1312	-0.0880	0.0725	-0.0163	0.0596

Petal Length, and Petal Width contribute approximately equal amounts to the first principal component and Sepal Width is the largest contributor to the second principal component.

The similarity between classical PCA using the endpoints and using the midpoints is also seen in the correlation between the principal components and the random variables. In both methods, the first principal component is highly correlated with Sepal Length, Petal Length and Petal Width with an average correlation coefficients for the endpoints being 0.89 versus a correlation coefficient of 0.92 for the midpoints for Sepal Length, 0.99 versus 0.99 for Petal Length, and 0.97 versus 0.97 for Petal Width. The second principal component is highly correlated with Sepal Width with an average correlation coefficients of 0.67 for the endpoints versus 0.72 for the midpoints.

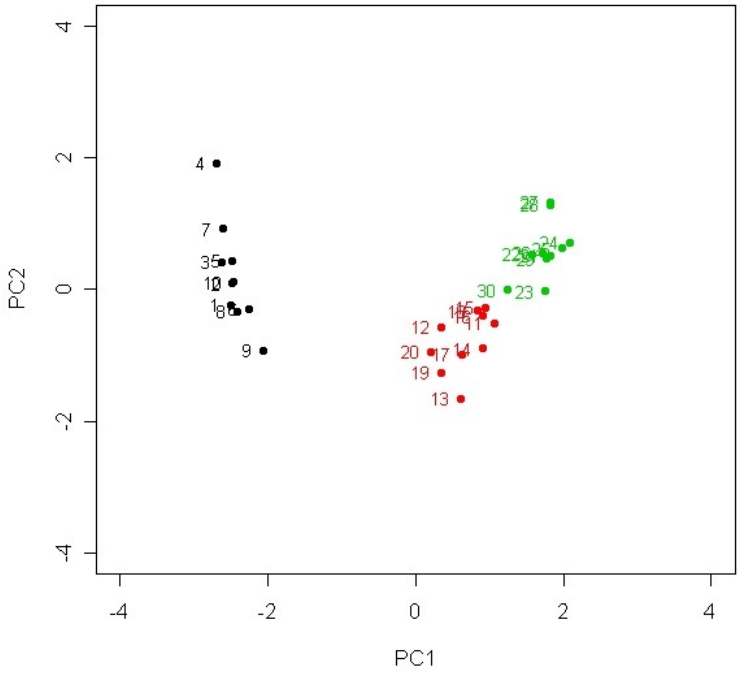

Figure 3.32: Plot of PC1 × PC2 for Iris Data Based on Classical PCA Method Using Midpoints (Color Represents Species)

The midpoints PC1 × PC2 plot of Figure 3.32 and the endpoints PC1 × PC2 plot of Figure 3.33 show similar patterns. Points representing setosa irises have small PC1 values and form a distinct cluster from points representing other species. Because each observation is represented by only one point in classical PCA, the plots of Figures 3.32 and 3.33 do not show

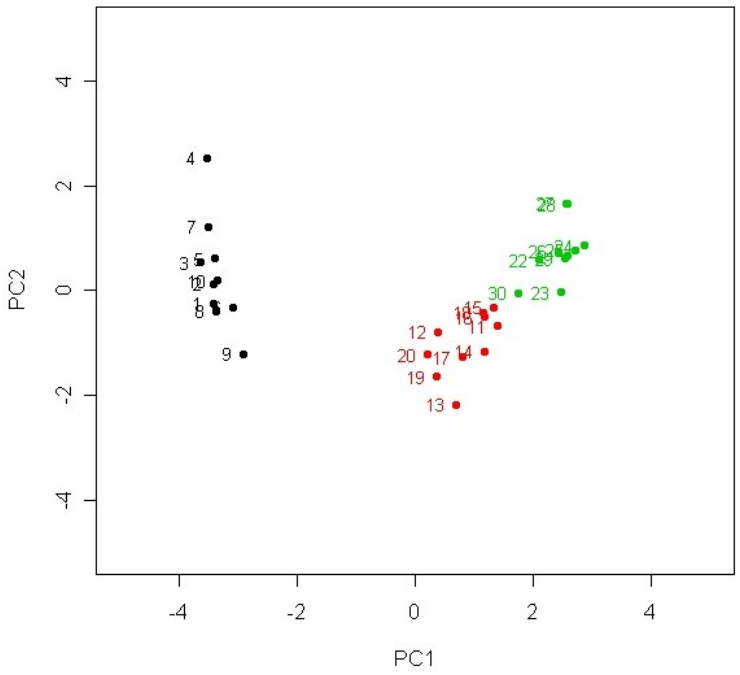

Figure 3.33: Plot of PC1 × PC2 for Iris Data Based on Classical PCA Method Using Endpoints (Color Represents Species)

the structure of interval-valued observations. Therefore, differences in internal variation of an observation can not be detected in these plots. For instance, from the symbolic covariance PC1 × PC2 plot of Figure 3.27 we can see that the size of observation 4 is only half as wide as observation 3 and it is only half as long as observation 9; or observation 22 is much shorter and much wider than observations 24 and 27. Moreover, without the structure of the observation, classical PCA plots can show clusters where they do not exist. For example, Figures 3.32 and 3.33 show no overlap between versicolor and viriginica species along the PC2 axis. However, the symbolic covariance plot of Figure 3.27 reveals that with the structure of interval-valued observations, these two species have much overlap in PC2 values.

In this comparison, we showed that the symbolic covariance PCA method accounts for the total variation of interval-valued observations instead of only part of the variation when

classical PCA is used to anlyze interval-valued data. Plots from our proposed method further show the structure of interval-valued observations which does not exist in classical PCA.

3.4.2 Face Recognition Example

The data

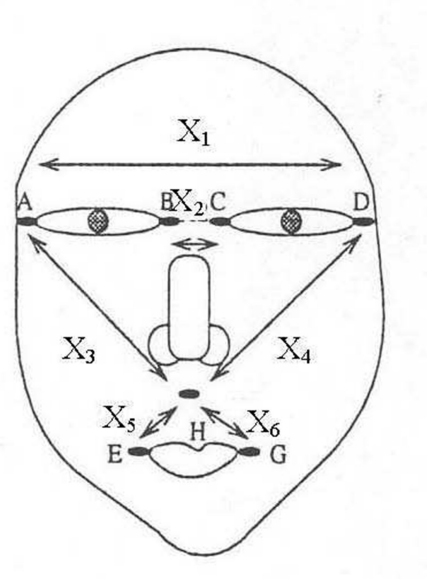

Figure 3.34: Diagram of Variables for Face Recognition Data

In the second example, we apply the proposed symbolic covariance PCA method to the Face Recognition data published by Leroy et al. (1996). This dataset came from a study of face identification using two dimensional images. Leroy et al. (1996) identified points characterizing a person's face. Images of the subjects were taken and distances between these points were measured. The six variables for the Face Recognition data are shown in Figure 3.34, namely, AD, BC, AH, DH, HE, and HG. This dataset gives face measurements of nine subjects, each with three observations. Therefore, the dataset consists of 27 observations. Since measurements for each observation came from a sequence of images, they have interval values. Table 3.12 gives the complete dataset. Each observation is labeled by three letters identifying the subject and a number between 1 and 3 distinguishing observations for the

Table 3.12: Interval-Valued Face Recognition Data

i	Label	AD	BC	AH	DH	EH	GH
1	FRA1	[155, 157]	[58, 61]	[100, 103]	[105, 107]	[61, 66]	[64, 68]
2	FRA2	[154, 160]	[57, 64]	[102, 106]	[104, 107]	[61, 63]	[63, 66]
3	FRA3	[154, 161]	[57, 63]	[99, 106]	[101, 109]	[61, 66]	[60, 66]
4	HUS1	[169, 173]	[59, 63]	[103, 107]	[122, 125]	[57, 61]	[60, 65]
5	HUS2	[170, 175]	[60, 64]	[103, 109]	[120, 125]	[57, 62]	[60, 67]
6	HUS3	[169, 175]	[61, 64]	[104, 107]	[121, 125]	[57, 62]	[58, 67]
7	INC1	[155, 160]	[53, 60]	[96, 98]	[92, 94]	[62, 66]	[59, 63]
8	INC2	[156, 161]	[51, 60]	[96, 99]	[91, 97]	[55, 64]	[54, 62]
9	INC3	[154, 160]	[55, 59]	[94, 99]	[90, 96]	[59, 66]	[56, 66]
10	ISA1	[164, 168]	[55, 60]	[120, 123]	[118, 121]	[54, 57]	[51, 53]
11	ISA2	[163, 170]	[54, 59]	[119, 123]	[117, 120]	[55, 59]	[52, 55]
12	ISA3	[164, 169]	[55, 59]	[117, 123]	[117, 122]	[53, 58]	[52, 55]
13	JPL1	[167, 171]	[61, 65]	[118, 122]	[108, 111]	[64, 68]	[57, 61]
14	JPL2	[169, 173]	[60, 65]	[119, 121]	[109, 113]	[63, 69]	[57, 62]
15	JPL3	[169, 170]	[59, 65]	[116, 121]	[110, 112]	[62, 68]	[59, 63]
16	KHA1	[149, 156]	[54, 59]	[112, 116]	[105, 111]	[54, 58]	[48, 51]
17	KHA2	[149, 155]	[52, 58]	[111, 113]	[105, 111]	[54, 58]	[49, 53]
18	KHA3	[150, 157]	[52, 60]	[109, 113]	[105, 111]	[55, 60]	[49, 53]
19	LOT1	[153, 158]	[51, 56]	[117, 120]	[115, 117]	[55, 60]	[53, 57]
20	LOT2	[155, 158]	[52, 56]	[118, 120]	[114, 117]	[58, 61]	[54, 58]
21	LOT3	[155, 158]	[50, 55]	[118, 120]	[114, 117]	[57, 61]	[55, 58]
22	PHI1	[163, 167]	[66, 68]	[115, 120]	[116, 121]	[61, 65]	[57, 60]
23	PHI2	[164, 168]	[65, 68]	[115, 120]	[115, 121]	[61, 67]	[55, 62]
24	PHI3	[161, 167]	[64, 69]	[117, 119]	[115, 119]	[62, 69]	[57, 60]
25	ROM1	[167, 171]	[64, 68]	[124, 127]	[123, 126]	[51, 55]	[50, 54]
26	ROM2	[168, 172]	[63, 68]	[122, 127]	[124, 127]	[50, 57]	[50, 57]
27	ROM3	[167, 171]	[63, 68]	[122, 127]	[123, 128]	[49, 57]	[51, 60]

same subject. Further discussion of the data can be found in Leroy et al. (1996) and Billard et al. (2007).

We first computed the symbolic covariance matrix for the Face Recognition data of Table 3.12 based on Equation (3.1). The variance-covariance matrix is

$$\boldsymbol{S} = \begin{bmatrix} 46.80 & 21.47 & 23.78 & 41.13 & -0.10 & 5.47 \\ & 21.01 & 11.09 & 19.67 & 4.71 & 5.98 \\ & & 80.67 & 59.28 & -12.18 & -27.67 \\ & & & 88.27 & -17.90 & -12.87 \\ & & & & 18.00 & 13.72 \\ & & & & & 22.87 \end{bmatrix}. \tag{3.27}$$

Matrix $\boldsymbol{S}$ of Equation (3.27) shows that AH and DH have the largest variance, 80.67 and 88.27, respectively. Variables BC, HE, and HG have variance of similar size, 21.01, 18.00, and 22.87, respectively. Equation (3.28) gives the correlation matrix of the Face Recognition dataset using the symbolic covariance matrix of Equation (3.27) as follows,

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.6846 & 0.3871 & 0.6400 & -0.0034 & 0.1672 \\ & 1 & 0.2693 & 0.4567 & 0.2423 & 0.2729 \\ & & 1 & 0.7025 & -0.3197 & -0.6443 \\ & & & 1 & -0.4490 & -0.2865 \\ & & & & 1 & 0.6762 \\ & & & & & 1 \end{bmatrix}. \tag{3.28}$$

The correlation matrix of Equation (3.28) shows strong correlation between AD and BC with a coefficient of 0.6846, AD and and DH with a coefficient of 0.6400. Variable AH is positively correlated with DH with a coefficient of 0.7025 and AH is negatively correlated with HG with a coefficient of -0.6443. In addition, HE and HG are correlated with a coefficient of 0.6762. With information about the correlation structure of the Face data, we proceed with computing the principal components based on the symbolic correlation matrix of the Face Recognition data.

Table 3.13: Principal Component Coefficients and Variance Proportion of Face Recognition Data Based on Symbolic Covariance Method

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AD	-0.3894	0.4425	-0.2516	-0.3450	0.6801	0.0874
BC	-0.2783	0.5368	0.1148	0.7844	-0.0627	-0.0435
AH	-0.5177	-0.1067	0.5796	-0.2317	-0.0565	-0.5726
DH	-0.5491	0.0888	-0.2282	-0.2502	-0.6366	0.4130
HE	0.3165	0.4562	0.6603	-0.2788	-0.0662	0.4166
HG	0.3181	0.5372	-0.3154	-0.2676	-0.3473	-0.5644
Proportion of Variance	0.4621	0.3442	0.0992	0.0491	0.0368	0.0087
Cumulative Proportion	0.4621	0.8063	0.9055	0.9545	0.9913	1.0000

Analysis results

Now, we apply the proposed symbolic covariance PCA method to the Face Recognition data of Table 3.12. The coefficients and the proportion of variation explained by the principal components are listed in Table 3.13. The first principal component explains 46.21% and the second principal component explains another 34.42% of the overall variability in the data. Together they account for more than 80% of the total variation in the data. The biggest contributors to the first principal component include AH and DH with coefficients of -0.5177 and -0.5491, respectively. Variables AH and DH give measurements along the length of a face as shown in Figure 3.34. Therefore, longer faces have smaller PC1 values. The second principal component is composed of AD, BC, HE, and HG with coefficients of 0.4425, 0.5368, 0.4562, and 0.5372, respectively. Variables AD, BC, HE, and HG measure points spread along the width of the face. Therefore, wider faces have larger PC2 values.

With the first two principal components explaining 80% of the total variation in the data, a plot of observations along the PC1 and PC2 axes should reveal important features in the data. A plot of the symbolic PC1 $\times$ PC2 for the Face Recognition data is shown in Figure 3.35. Observations are colored by person. The polygons representing the observations are

labeled by person in the same order as they are listed in Table 3.12. That is, person one is FRA, person two is HUS, person three is INC, and so on.

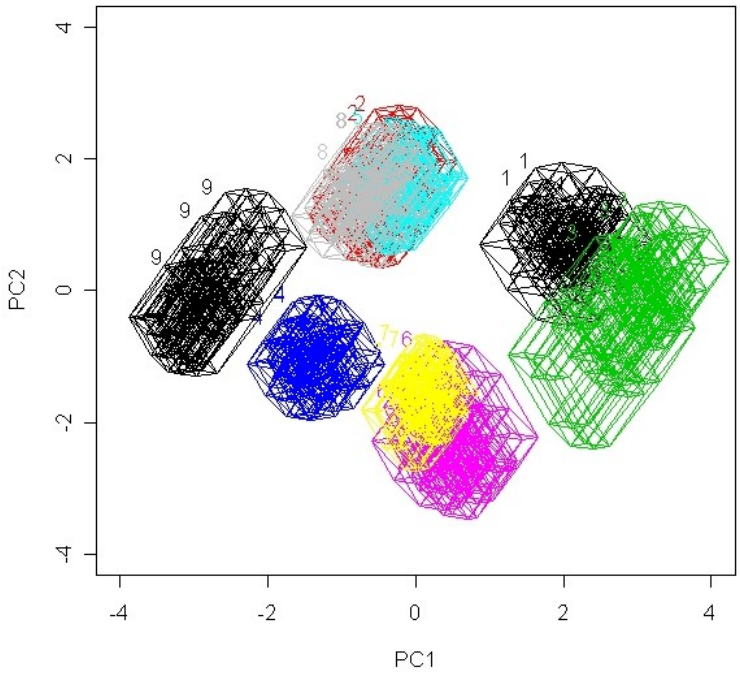

Figure 3.35: Plot of PC1 × PC2 for Face Recognition Fata Based on Symbolic Covariance Method (Color Represents Person)

An immediate observation from Figure 3.35 is that sets of three faces belonging to the same person group together. Faces of the nine subjects form five distint groups. Figure 3.35 indicates that HUS (person 2), JPL (person 5), and PHI (person 8) have similar facial features. All three subjects have wide faces of medium length. Subjects KHA (6) and LOT (7) have similar features. Their faces are narrow with medium length. The person with the longest face seems to be ROM (9). Whereas, ISA (4) has a long and narrow face which is distinct from all others. The last group is formed by FRA (1) and INC (3). Both of these subjects have short faces of medium width. However, measurements of INC's face have higher variability than do measurements of FRA's.

The correlations between the principal components and the random variables are shown in Table 3.14. The correlation coefficients of -0.8620 and -0.9143 indicate a strong negative correlation between PC1 and AH and DH. Again, AH and DH are measurements along the

Table 3.14: Correlation Between Principal Components and Random Variables of Face Recognition Data Based on Symbolic Covariance Method

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AD	-0.6484	0.6359	-0.1941	-0.1872	0.3195	0.0200
BC	-0.4634	0.7715	0.0885	0.4257	-0.0294	-0.0099
AH	-0.8620	-0.1534	0.4471	-0.1257	-0.0266	-0.1307
DH	-0.9143	0.1277	-0.1761	-0.1358	-0.2991	0.0943
HE	0.5270	0.6557	0.5094	-0.1513	-0.0311	0.0951
HG	0.5296	0.7720	-0.2433	-0.1452	-0.1631	-0.1289

length of the face. Therefore, PC1 is correlated to face length. Correlation coefficients of 0.6359 for AD, 0.7715 for BC, 0.6557 for HE, and 0.7720 for HG in the second principal component column indicate strong positive correlation between PC2 and these variables which measure the horizontal span of the face. Hence, PC2 is correlated to face width. The conclusions about the first two principal components drawn from the correlation coefficients of Table 3.14 agree with our conclusions from the coefficients of the principal components shown in Table 3.13.

After exploring the coefficients and the plot of the observations on a PC1 $\times$ PC2 plane, we computed the histogram-valued principal components based on the PC1 $\times$ PC2 plot of the Face data. The principal components in Table 3.15 can be used as input into a model for further analysis.

Table 3.15: Histogram for the First Principal Component of the Face Recognition Data Based on Symbolic Covariance Method

i	**Frequency Histogram for the First Principal Component**
1	$\{[1.74, 1.92), 0.06; [1.92, 1.98), 0.04; [1.98, 2.03), 0.05; [2.03, 2.17), 0.15; [2.17, 2.30), 0.17;$ $[2.30, 2.33), 0.04; [2.33, 2.46), 0.17; [2.46, 2.60), 0.15; [2.60, 2.66), 0.05; [2.66, 2.71), 0.04;$
Continued on next page	

Table 3.15 – continued from previous page

i	Frequency Histogram for the First Principal Component
	$[2.71, 2.89], 0.06\}$
2	$\{[1.13, 1.37), 0.06; [1.37, 1.53), 0.10; [1.53, 1.56), 0.02; [1.56, 1.73), 0.15; [1.73, 1.90), 0.16; [1.90, 1.91), 0.01; [1.91, 2.07), 0.16; [2.07, 2.25), 0.15; [2.25, 2.28), 0.02; [2.28, 2.44), 0.10; [2.44, 2.67], 0.06\}$
3	$\{[0.87, 1.23), 0.06; [1.23, 1.26), 0.01; [1.26, 1.61), 0.17; [1.61, 1.63), 0.01; [1.63, 1.97), 0.21; [1.97, 2.10), 0.08; [2.10, 2.44), 0.21; [2.44, 2.46), 0.01; [2.46, 2.81), 0.17; [2.81, 2.84), 0.01; [2.84, 3.20], 0.06\}$
4	$\{[-1.10, -0.83), 0.08; [-0.83, -0.81), 0.01; [-0.81, -0.58), 0.19; [-0.58, -0.50), 0.08; [-0.50, -0.45), 0.05; [-0.45, -0.29), 0.18; [-0.29, -0.24), 0.05; [-0.24, -0.17), 0.08; [-0.17, 0.06), 0.19; [0.06, 0.08), 0.01; [0.08, 0.35], 0.08\}$
5	$\{[-1.41, -1.16), 0.04; [-1.16, -0.99), 0.07; [-0.99, -0.86), 0.07; [-0.86, -0.61), 0.17; [-0.61, -0.56), 0.04; [-0.56, -0.28), 0.22; [-0.28, -0.23), 0.04; [-0.23, 0.02), 0.17; [0.02, 0.14), 0.07; [0.14, 0.31), 0.07; [0.31, 0.57], 0.04\}$
6	$\{[-1.45, -1.32), 0.01; [-1.32, -0.96), 0.15; [-0.96, -0.85), 0.07; [-0.85, -0.71), 0.11; [-0.71, -0.53), 0.15; [-0.53, -0.52), 0.01; [-0.52, -0.34), 0.15; [-0.34, -0.19), 0.11; [-0.19, -0.09), 0.07; [-0.09, 0.27), 0.15; [0.27, 0.40], 0.01\}$
7	$\{[2.35, 2.63), 0.07; [2.63, 2.78), 0.09; [2.78, 2.91), 0.11; [2.91, 3.06), 0.14; [3.06, 3.07), 0.01; [3.07, 3.22), 0.15; [3.22, 3.23), 0.01; [3.23, 3.38), 0.14; [3.38, 3.51), 0.11; [3.51, 3.66), 0.09; [3.66, 3.94], 0.07\}$
8	$\{[1.25, 1.73), 0.08; [1.73, 1.80), 0.03; [1.80, 2.09), 0.13; [2.09, 2.41), 0.19; [2.41, 2.62), 0.13; [2.62, 2.96), 0.19; [2.96, 3.24), 0.13; [3.24, 3.31), 0.03; [3.31, 3.79], 0.08\}$
9	$\{[1.83, 2.07), 0.03; [2.07, 2.40), 0.11; [2.40, 2.48), 0.04; [2.48, 2.75), 0.15; [2.75, 2.99), 0.15; [2.99, 3.07), 0.05; [3.07, 3.31), 0.15; [3.31, 3.57), 0.15; [3.57, 3.66), 0.04; [3.66, 3.99), 0.11; [3.99, 4.23], 0.03\}$

Continued on next page

Table 3.15 – continued from previous page

i	Frequency Histogram for the First Principal Component
10	$\{[-2.18,-2.02),0.04;[-2.02,-1.88),0.09;[-1.88,-1.79),0.08;[-1.79,-1.65),0.15;$ $[-1.65,-1.63),0.02;[-1.63,-1.45),0.22;[-1.45,-1.43),0.02;[-1.43,-1.29),0.15;$ $[-1.29,-1.20),0.08;[-1.20,-1.06),0.09;[-1.06,-0.90],0.04\}$
11	$\{[-2.00,-1.81),0.03;[-1.81,-1.70),0.06;[-1.70,-1.58),0.08;[-1.58,-1.33),0.21;$ $[-1.33,-1.30),0.03;[-1.30,-1.12),0.17;[-1.12,-1.09),0.03;[-1.09,-0.85),0.21;$ $[-0.85,-0.72),0.08;[-0.72,-0.61),0.06;[-0.61,-0.42],0.03\}$
12	$\{[-2.29,-2.07),0.04;[-2.07,-2.05),0.01;[-2.05,-1.76),0.15;[-1.76,-1.66),0.07;$ $[-1.66,-1.43),0.19;[-1.43,-1.33),0.08;[-1.33,-1.10),0.19;[-1.10,-1.00),0.07;$ $[-1.00,-0.71),0.15;[-0.71,-0.69),0.01;[-0.69,-0.47],0.04\}$
13	$\{[-0.88,-0.65),0.07;[-0.65,-0.41),0.19;[-0.41,-0.35),0.06;[-0.35,-0.24),0.12;$ $[-0.24,-0.14),0.11;[-0.14,-0.03),0.12;[-0.03,0.03),0.06;[0.03,0.26),0.19;$ $[0.26,0.26),0.00;[0.26,0.50],0.07\}$
14	$\{[-1.15,-0.87),0.07;[-0.87,-0.85),0.01;[-0.85,-0.62),0.15;[-0.62,-0.39),0.20;$ $[-0.39,-0.37),0.02;[-0.37,-0.27),0.10;[-0.27,-0.25),0.02;[-0.25,-0.02),0.20;$ $[-0.02,0.21),0.15;[0.21,0.23),0.01;[0.23,0.51],0.07\}$
15	$\{[-0.89,-0.66),0.05;[-0.66,-0.52),0.07;[-0.52,-0.46),0.04;[-0.46,-0.34),0.10;$ $[-0.34,-0.17),0.15;[-0.17,-0.02),0.15;[-0.02,0.15),0.15;[0.15,0.27),0.10;$ $[0.27,0.33),0.04;[0.33,0.47),0.07;[0.47,0.70],0.05\}$
16	$\{[-0.60,-0.45),0.02;[-0.45,-0.30),0.06;[-0.30,-0.15),0.09;[-0.15,0.05),0.17;$ $[0.05,0.07),0.01;[0.07,0.39),0.28;[0.39,0.40),0.01;[0.40,0.61),0.17;$ $[0.61,0.75),0.09;[0.75,0.90),0.06;[0.90,1.06],0.02\}$
17	$\{[-0.35,-0.12),0.04;[-0.12,0.03),0.07;[0.03,0.21),0.13;[0.21,0.32),0.10;$ $[0.32,0.37),0.04;[0.37,0.66),0.26;[0.66,0.70),0.04;[0.70,0.82),0.10;$ $[0.82,1.00),0.13;[1.00,1.15),0.07;[1.15,1.37],0.04\}$

Continued on next page

Table 3.15 – continued from previous page

i	Frequency Histogram for the First Principal Component
18	$\{[-0.43,-0.15),0.04;[-0.15,0.06),0.09;[0.06,0.19),0.07;[0.19,0.40),0.14;$ $[0.40,0.46),0.04;[0.46,0.77),0.23;[0.77,0.83),0.04;[0.83,1.04),0.14;$ $[1.04,1.16),0.07;[1.16,1.38),0.09;[1.38,1.66],0.04\}$
19	$\{[-0.73,-0.49),0.06;[-0.49,-0.44),0.03;[-0.44,-0.19),0.21;[-0.19,-0.15),0.04;$ $[-0.15,-0.02),0.15;[-0.02,0.01),0.03;[0.01,0.15),0.15;[0.15,0.18),0.04;$ $[0.18,0.43),0.21;[0.43,0.49),0.03;[0.49,0.73],0.06\}$
20	$\{[-0.48,-0.24),0.09;[-0.24,-0.23),0.01;[-0.23,-0.06),0.17;[-0.06,-0.02),0.05;$ $[-0.02,0.10),0.16;[0.10,0.12),0.02;[0.12,0.24),0.16;[0.24,0.29),0.05;$ $[0.29,0.46),0.17;[0.46,0.47),0.01;[0.47,0.70],0.09\}$
21	$\{[-0.42,-0.24),0.04;[-0.24,-0.12),0.08;[-0.12,0.05),0.17;[0.05,0.09),0.04;$ $[0.09,0.21),0.16;[0.21,0.34),0.16;[0.34,0.37),0.04;[0.37,0.54),0.17;$ $[0.54,0.67),0.08;[0.67,0.84],0.04\}$
22	$\{[-1.52,-1.39),0.02;[-1.39,-1.33),0.03;[-1.33,-1.17),0.11;[-1.17,-1.01),0.15;$ $[-1.01,-0.88),0.13;[-0.88,-0.76),0.13;[-0.76,-0.63),0.13;[-0.63,-0.47),0.15;$ $[-0.47,-0.30),0.11;[-0.30,-0.24),0.03;[-0.24,-0.12],0.02\}$
23	$\{[-1.68,-1.50),0.03;[-1.50,-1.27),0.09;[-1.27,-0.93),0.21;[-0.93,-0.80),0.10;$ $[-0.80,-0.64),0.13;[-0.64,-0.51),0.10;[-0.51,-0.18),0.21;[-0.18,0.05),0.09;$ $[0.05,0.24],0.03\}$
24	$\{[-1.38,-1.15),0.04;[-1.15,-1.08),0.03;[-1.08,-0.74),0.21;[-0.74,-0.61),0.11;$ $[-0.61,-0.49),0.11;[-0.49,-0.37),0.11;[-0.37,-0.24),0.11;[-0.24,0.10),0.21;$ $[0.10,0.16),0.03;[0.16,0.40],0.04\}$
25	$\{[-3.69,-3.44),0.08;[-3.44,-3.42),0.02;[-3.42,-3.21),0.18;[-3.21,-3.16),0.05;$ $[-3.16,-3.01),0.17;[-3.01,-3.00),0.01;[-3.00,-2.85),0.17;[-2.85,-2.80),0.05;$ $[-2.80,-2.59),0.18;[-2.59,-2.56),0.02;[-2.56,-2.32],0.08\}$

Continued on next page

Table 3.15 – continued from previous page

i	**Frequency Histogram for the First Principal Component**
26	$\{[-3.88,-3.58),0.06;[-3.58,-3.41),0.08;[-3.41,-3.35),0.03;[-3.35,-3.17),0.12;$ $[-3.17,-2.90),0.21;[-2.90,-2.89),0.01;[-2.89,-2.61),0.21;[-2.61,-2.43),0.12;$ $[-2.43,-2.37),0.03;[-2.37,-2.20),0.08;[-2.20,-1.90],0.06\}$
27	$\{[-3.81,-3.51),0.04;[-3.51,-3.28),0.09;[-3.28,-3.18),0.05;[-3.18,-2.97),0.11;$ $[-2.97,-2.69),0.18;[-2.69,-2.59),0.07;[-2.59,-2.31),0.18;[-2.31,-2.10),0.11;$ $[-2.10,-2.00),0.05;[-2.00,-1.77),0.09;[-1.77,-1.47],0.04\}$

Table 3.16: Histogram for the Second Principal Component of Face Recognition Data Based on Symbolic Covariance Method

i	**Frequency Histogram for the Second Principal Component**
1	$\{[0.23,0.34),0.05;[0.34,0.69),0.25;[0.69,1.10),0.38;[1.10,1.45),0.25;$ $[1.45,1.57),0.05\}$
2	$\{[-0.18,0.05),0.07;[0.05,0.20),0.07;[0.20,0.45),0.16;[0.45,1.02),0.39;$ $[1.02,1.27),0.16;[1.27,1.41),0.07;[1.41,1.65),0.07\}$
3	$\{[-0.53,-0.45),0.01;[-0.45,0.00),0.14;[0.00,0.05),0.02;[0.05,0.70),0.33;$ $[0.70,0.72),0.01;[0.72,1.38),0.33;[1.38,1.42),0.02;[1.42,1.87),0.14;$ $[1.87,1.95),0.01\}$
4	$\{[0.40,0.63),0.08;[0.63,0.87),0.12;[0.87,1.20),0.24;[1.20,1.33),0.11;$ $[1.33,1.66),0.24;[1.66,1.89),0.12;[1.89,2.13),0.08\}$
5	$\{[0.61,0.64),0.01;[0.64,0.94),0.09;[0.94,1.24),0.14;[1.24,1.43),0.11;$ $[1.43,1.96),0.31;[1.96,2.15),0.11;[2.15,2.45),0.14;[2.45,2.76),0.09;$ $[2.76,2.79),0.01\}$

Continued on next page

Table 3.16 – continued from previous page

i	Frequency Histogram for the Second Principal Component
6	$\{[0.49, 0.90), 0.13; [0.90, 0.97), 0.03; [0.97, 1.15), 0.09; [1.15, 1.98), 0.47;$ $[1.98, 2.16), 0.09; [2.16, 2.22), 0.03; [2.22, 2.64), 0.13\}$
7	$\{[-0.88, -0.54), 0.09; [-0.54, -0.47), 0.03; [-0.47, 0.00), 0.28; [0.00, 0.28), 0.20;$ $[0.28, 0.76), 0.28; [0.76, 0.83), 0.03; [0.83, 1.16), 0.09\}$
8	$\{[-2.36, -2.04), 0.05; [-2.04, -1.37), 0.19; [-1.37, -0.98), 0.16; [-0.98, -0.57), 0.19;$ $[-0.57, -0.18), 0.16; [-0.18, 0.48), 0.19; [0.48, 0.81), 0.05\}$
9	$\{[-1.38, -1.01), 0.09; [-1.01, -0.65), 0.13; [-0.65, -0.55), 0.04; [-0.55, 0.45), 0.45;$ $[0.45, 0.55), 0.04; [0.55, 0.92), 0.13; [0.92, 1.29), 0.09\}$
10	$\{[-1.92, -1.66), 0.12; [-1.66, -1.59), 0.05; [-1.59, -1.31), 0.22; [-1.31, -1.07), 0.22;$ $[-1.07, -0.80), 0.22; [-0.80, -0.73), 0.05; [-0.73, -0.47), 0.12\}$
11	$\{[-1.79, -1.44), 0.14; [-1.44, -1.35), 0.06; [-1.35, -1.13), 0.16; [-1.13, -0.77), 0.28;$ $[-0.77, -0.55), 0.16; [-0.55, -0.46), 0.06; [-0.46, -0.11), 0.14\}$
12	$\{[-1.97, -1.92), 0.01; [-1.92, -1.60), 0.12; [-1.60, -1.31), 0.17; [-1.31, -1.13), 0.13;$ $[-1.13, -0.95), 0.14; [-0.95, -0.76), 0.13; [-0.76, -0.48), 0.17; [-0.48, -0.17), 0.12;$ $[-0.15, -0.10), 0.01\}$
13	$\{[0.68, 0.92), 0.10; [0.92, 1.10), 0.11; [1.10, 1.39), 0.23; [1.39, 1.50), 0.10;$ $[1.50, 1.79), 0.23; [1.79, 1.97), 0.11; [1.97, 2.22), 0.10\}$
14	$\{[0.58, 0.85), 0.07; [0.85, 1.26), 0.20; [1.26, 1.43), 0.11; [1.43, 1.74), 0.22;$ $[1.74, 1.91), 0.11; [1.91, 2.32), 0.20; [2.32, 2.58), 0.07\}$
15	$\{[0.58, 0.62), 0.01; [0.62, 0.65), 0.01; [0.65, 1.32), 0.33; [1.32, 1.35), 0.02;$ $[1.35, 1.72), 0.26; [1.72, 1.74), 0.02; [1.74, 2.42), 0.33; [2.42, 2.44), 0.01;$ $[2.44, 2.49), 0.01\}$
16	$\{[-3.33, -3.29), 0.01; [-3.29, -2.88), 0.18; [-2.88, -2.86), 0.01; [-2.86, -2.60), 0.18;$ $[-2.60, -2.29), 0.24; [-2.29, -2.03), 0.18; [-2.03, -2.02), 0.01; [-2.02, -1.62), 0.18;$

Continued on next page

Table 3.16 – continued from previous page

i	Frequency Histogram for the Second Principal Component
	$[-1.62, -1.56), 0.01\}$
17	$\{[-3.42, -3.03), 0.13; [-3.03, -2.97), 0.03; [-2.97, -2.59), 0.23; [-2.59, -2.31), 0.20;$ $[-2.31, -1.93), 0.23; [-1.93, -1.87), 0.03; [-1.87, -1.48), 0.13\}$
18	$\{[-3.18, -2.73), 0.13; [-2.73, -2.70), 0.01; [-2.70, -2.23), 0.23; [-2.23, -1.78), 0.25;$ $[-1.78, -1.31), 0.23; [-1.31, -1.28), 0.01; [-1.28, -0.83), 0.13\}$
19	$\{[-2.70, -2.39), 0.11; [-2.39, -2.26), 0.07; [-2.26, -1.86), 0.30; [-1.86, -1.82), 0.03;$ $[-1.82, -1.42), 0.30; [-1.42, -1.29), 0.07; [-1.29, -0.98), 0.11\}$
20	$\{[-2.09, -1.89), 0.08; [-1.89, -1.77), 0.08; [-1.77, -1.42), 0.31; [-1.42, -1.37), 0.05;$ $[-1.37, -1.02), 0.31; [-1.02, -0.89), 0.08; [-0.89, -0.70), 0.08\}$
21	$\{[-2.32, -2.12), 0.07; [-2.12, -1.84), 0.17; [-1.84, -1.55), 0.25; [-1.55, -1.26), 0.25;$ $[-1.26, -0.98), 0.17; [-0.98, -0.79), 0.07\}$
22	$\{[0.71, 0.76), 0.01; [0.76, 1.02), 0.15; [1.02, 1.23), 0.18; [1.23, 1.26), 0.02;$ $[1.26, 1.55), 0.28; [1.55, 1.57), 0.02; [1.57, 1.79), 0.18; [1.79, 2.04), 0.15;$ $[2.04, 2.09), 0.01\}$
23	$\{[0.46, 0.51), 0.01; [0.51, 0.77), 0.09; [0.77, 1.13), 0.18; [1.13, 1.20), 0.04;$ $[1.20, 1.83), 0.38; [1.83, 1.89), 0.04; [1.89, 2.25), 0.18; [2.25, 2.51), 0.09;$ $[2.51, 2.57), 0.01\}$
24	$\{[0.40, 0.78), 0.11; [0.78, 1.15), 0.19; [1.15, 1.36), 0.14; [1.36, 1.55), 0.13;$ $[1.55, 1.76), 0.14; [1.76, 2.13), 0.19; [2.13, 2.51), 0.11\}$
25	$\{[-1.11, -0.85), 0.11; [-0.85, -0.75), 0.06; [-0.75, -0.38), 0.28; [-0.38, -0.29), 0.09;$ $[-0.29, 0.08), 0.28; [0.08, 0.18), 0.06; [0.18, 0.44), 0.11\}$
26	$\{[-1.21, -0.98), 0.05; [-0.98, -0.46), 0.21; [-0.46, -0.40), 0.03; [-0.40, 0.33), 0.39;$ $[0.33, 0.39), 0.03; [0.39, 0.90), 0.21; [0.90, 1.13), 0.05\}$
27	$\{[-1.25, -0.99), 0.05; [-0.99, -0.42), 0.20; [-0.42, -0.40), 0.01; [-0.40, 0.64), 0.46;$

Continued on next page

Table 3.16 – continued from previous page

i	**Frequency Histogram for the Second Principal Component**
	$[0.64, 0.66), 0.01; [0.66, 1.24), 0.20; [1.24, 1.49), 0.05\}$

Comparison of symbolic covariance, vertices, centers, and classical PCA

To compare the results of our proposed method to other PCA methods, we apply the vertices method, the centers method, and the classical PCA method to the Face Recognition data of Table 3.12. In the Iris data example, there was an extensive discussion of the differences in the correlation structure between the symbolic covariance method, the vertices method, the centers method, and the classical PCA using the midpoints and the endpoints. The general differences also apply in this example so we only discus the plots resulting from the principal components of these methods. However, the covariance and the correlation matrices along with tables of coefficients of the principal components for the vertices, the centers, and the classical PCA methods are included in the appendix.

Figure 3.36 shows the MCAR's representing faces on the PC1 $\times$ PC2 plane for the vertices method and Figure 3.37 shows faces for the centers method. These plots show a similar pattern and similar position for the observations. We can also see that sets of observations belonging to a person are close together. However, all groups are connected to some other group. They do not form distinct clusters as revealed in Figure 3.35. This is a result of overlap caused by the maximum covering area rectangles in the vertices and the centers methods. For example, observations belonging to ISA (person 4) have distinct features and form a cluster completely separated from the others in the symbolic covariance plot of Figure 3.35. In contrast, in Figures 3.36 and 3.37, observations belonging to ISA are connected to observations belonging to ROM (9) on their left and observations belonging to KHA (6) and LOT (7) on their right. Moreover, a large intersection between observations of FRA (1) and

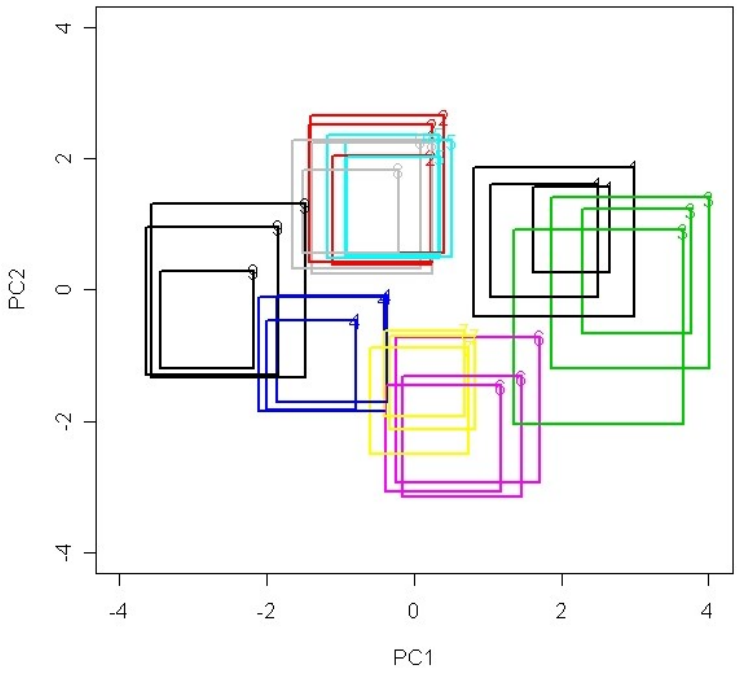

Figure 3.36: Plot of PC1 × PC2 for Face Recognition Data Based on Vertices Method (Color Represents Person)

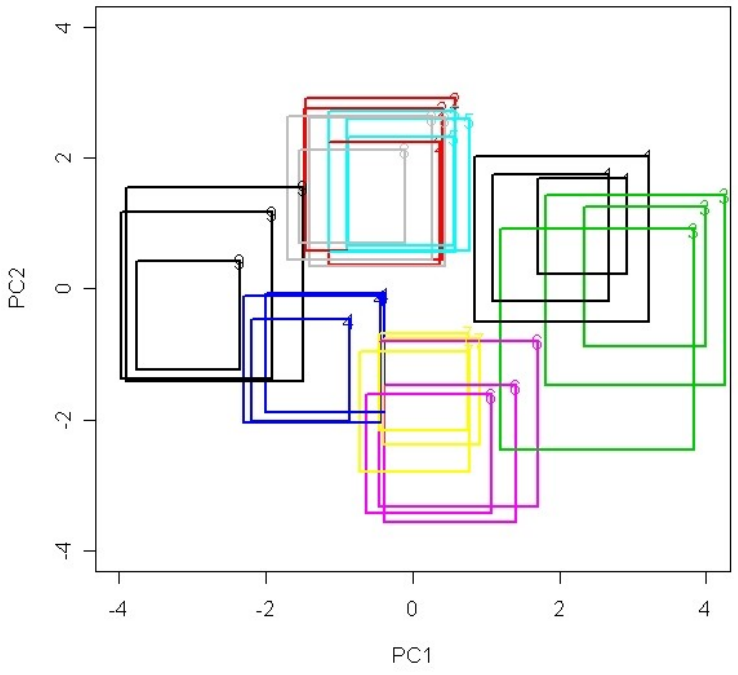

Figure 3.37: Plot of PC1 × PC2 for Face Recognition Data Based on Centers Method (Color Represents Person)

INC (3) in Figure 3.36 and Figure 3.37 turns out to be much less significant in Figure 3.35 which shows the projection of the true observations onto the symbolic PC1 × PC2 plane. Our proposed presentation of interval-valued observations shows more clearly the clusters of observations.

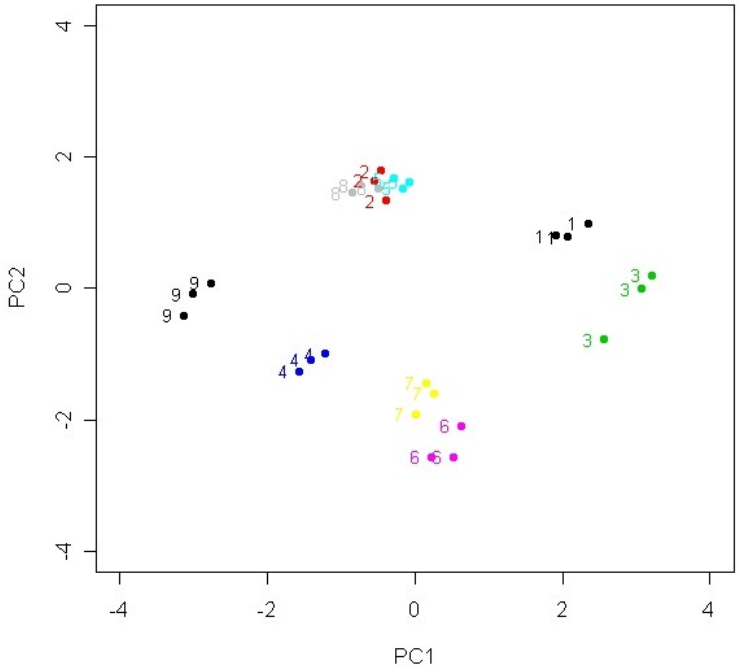

Figure 3.38: Plot of PC1 × PC2 for Face Recognition Data Based on Classical PCA Using Midpoints (Color Represent Person)

Figure 3.38 shows points representing observations along the PC1 and PC2 axes based on the classical PCA method using the midpoints. Figure 3.39 shows points representing observations along the PC1 and PC2 axes based on the classical PCA method using the endpoints. Both plots show identical pattern of grouping of points on slightly different scales. Figures 3.38 and 3.39 clearly show five distinct clusters as seen in Figure 3.35 of the symbolic covariance method. However, a closer look at the group of points with the largest PC1 values indicates this cluster can be further separated into two or even three groups. Without the whole structure of an observation, the points seem more separated. One of the points representing INC (3) is as far away from the other points belonging to INC as points belonging to FRA (1). However, with the shape and direction of rotation for the observations clearly

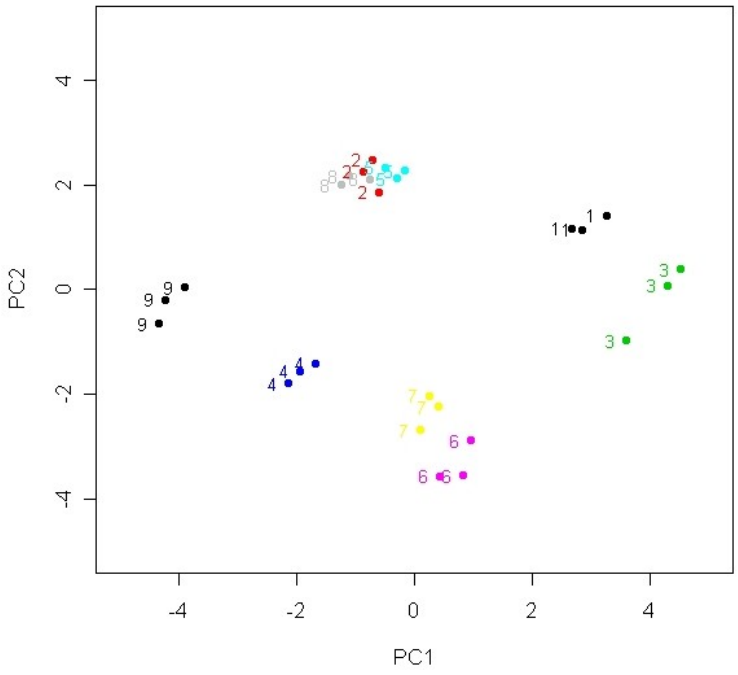

Figure 3.39: Plot of PC1 × PC2 for Face Recognition Data Based on Classical PCA Using Endpoints (Color Represents Person)

shown in the plot of Figure 3.35, we can see that all three observations belonging to INC are mostly intersected while the overlap between observations belonging to FRA and INC is very small.

Another feature that can not be discerned from the plots of classical PCA of Figures 3.38 and 3.39 is the size of the observation. For example, observations belonging to KHA (6) and LOT (7) form a cluster in Figures 3.38 and 3.39. All points in this cluster have the same size. However, the symbolic PC1 × PC2 plot of Figure 3.35 shows that the structure of observations belonging to LOT (7) is much smaller than is the structure of observations belonging to KHA (6). In fact, the size of observations belonging to LOT is half the size of observations belonging to KHA.

Comparison of the symbolic covariance, the vertices, the centers, and the classical PCA methods shows that the principal components constructed from the symbolic covariance

method maximize the total variation in the data. Plots of interval-valued observations on a principal components plane as polygons resulting from projecting the observations onto the PC planes give clearer features of the observations. These plots also show the size, shape, and direction of rotation for the data which can not be discerned on plots based on the vertices, the centers, and the classical PCA methods.

3.5 References

Anderson, E. (1935). The Irises of the Gasp Peninsula. *Bulletin of the American Iris Society*, **59**. 25.

Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 106-124.

Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-Valued Data. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, G. Cucumel, P. Bertrand and F. de Carvalho). Springer-Verlag, Berlin, 3-12.

Billard, L. and Diday E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, **98**, 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, New York.

Billard, L., Chouakia-Douzal, A., and Diday, E. (2007). Symbolic Principal Component Analysis for Interval-Valued Observations. *Journal of the American Statistical Association*, pending acceptance.

Bock, H.-H. and Diday, E. (eds.) (2000). *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.

Bueler, B., Enge A., and Fukuda, K. (2000). Exact Volume Computation for Polytopes: A Practical Study. In: *Polytopes - Combinatorics and Computation*, (eds. G. Kalai and G. M. Ziegler), Birhuser Verlag, Basel, 131-154.

Chouakria, A., Cazes, P., and Diday, E. (2000). Symbolic Principal Component Analysis. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 200-212.

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*, **XLV** (3). 5-24.

Davidson, K.R. and Donsig, A.P. (2002). *Real Analysis with Real Applications*. Prentice Hall, New Jersey.

Diday, E. (1987). Introduction à l'Approache Symbolique en Analyse des Données. *Premières Journées Symbolic-Numérique*, CEREMADE, Université Paris, 21-56.

Fisher, R.A (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**. 179188.

Jolliffe, I.T. (2004). *Principal Component Analysis*, 2nd edition. Springer, New York.

Leroy, B., Chouakria, A., Herlin, I., and Diday, E. (1996). Approche geometrique et classification pour la reconnaissance de visage. *Reconnaissance des Forms et Intelligence Artificelle*, INRIA and IRISA and CNRS, France, 548-557.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, New York.

Ziegler, G.M. (1995). *Lectures on Polytopes*. Springer-Verlag, New York.

Appendix

Tables of covariance, correlation, and coefficients of the principal components resulting from the vertices, the centers, and the classical PCA methods are included in this section.

Table 3.17: Vertices Covariance Matrix for Face Recognition Data

50.91	19.95	22.46	39.26	0.14	5.70
	25.48	11.23	18.69	3.33	4.63
		83.31	58.55	-13.29	-24.64
			91.58	-17.39	-10.25
				22.60	12.41
					27.27

Table 3.18: Vertices Correlation Matrix for Face Recognition Data

1	0.5540	0.3450	0.5750	0.0041	0.1529
	1	0.2438	0.3869	0.1386	0.1757
		1	0.6703	-0.3063	-0.5171
			1	-0.3822	-0.2051
				1	0.4999
					1

Table 3.19: Principal Component Coefficients and Variance Proportion of Face Recognition Data Based on Vertices Method

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AD	-0.403	0.439	-0.214	-0.257	0.723	0.104
BC	-0.306	0.497	0.175	0.779	-0.146	
AH	-0.523	-0.147	0.463	-0.273	-0.248	0.595
DH	-0.557		-0.206	-0.283	-0.423	-0.620
HE	0.297	0.466	0.700	-0.329		-0.311
HG	0.268	0.563	-0.421	-0.253	-0.465	0.392
Proportion	0.427	0.300	0.107	0.079	0.056	0.031
Cumulative	0.427	0.726	0.833	0.913	0.969	1.000

Table 3.20: Correlation between Principal Components and Random Variables of Face Recognition Data Based on Vertices Method

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AD	-0.6445	0.5889	-0.1717	-0.1771	0.4183	0.0453
BC	-0.4903	0.6662	0.1402	0.5375	-0.0845	
AH	-0.8372	-0.1968	0.3707	-0.1883	-0.1437	0.2584
DH	-0.8914		-0.1648	-0.1954	-0.2447	-0.2689
HE	0.4747	0.6248	0.5608	-0.2268		-0.1351
HG	0.4284	0.7555	-0.3377	-0.1746	-0.2689	0.1700

Table 3.21: Centers Covariance Matrix for Face Recognition Data

44.75	19.95	22.46	39.26	0.14	5.70
	18.78	11.23	18.69	3.33	4.63
		79.35	58.55	-13.29	-24.64
			86.62	-17.39	-10.25
				15.70	12.41
					20.67

Table 3.22: Centers Correlation Matrix for Face Recognition Data

1	0.6883	0.3770	0.6306	0.0052	0.1873
	1	0.2910	0.4635	0.1937	0.2351
		1	0.7062	-0.3765	-0.6086
			1	-0.4716	-0.2423
				1	0.6889
					1

Table 3.23: Principal Component Coefficients and Variance Proportion of Face Recognition Data Based on Centers Method

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AD	-0.383	0.453	-0.235	-0.165	0.752	
BC	-0.297	0.515	0.189	0.743	-0.238	
AH	-0.517	-0.115	0.565	-0.331		-0.535
DH	-0.545		-0.277	-0.339	-0.511	0.492
HE	0.334	0.458	0.628	-0.310		0.433
HG	0.299	0.546	-0.345	-0.316	-0.335	-0.530
Proportion	0.465	0.341	0.091	0.054	0.039	0.010
Cumulative	0.465	0.805	0.897	0.951	0.990	1.000

Table 3.24: Correlation between Principal Components and Random Variables of Face Recognition Data Based on Centers Method

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AD	-0.640	0.648	-0.174	-0.094	0.364	
BC	-0.496	0.736	0.140	0.423	-0.115	
AH	-0.863	-0.164	0.418	-0.189		-0.134
DH	-0.910		-0.205	-0.193	-0.247	0.123
HE	0.558	0.655	0.464	-0.177		0.108
HG	0.499	0.781	-0.255	-0.180	-0.162	-0.132

Table 3.25: Classical Correlation Matrix for Face Recognition Data using Endpoints

1	0.98	0.70	0.65	0.37	0.42	0.64	0.62	-0.03	0.07	0.15	0.23
	1	0.69	0.66	0.33	0.37	0.62	0.62	-0.06	0.04	0.11	0.22
		1	0.95	0.30	0.34	0.51	0.52	0.14	0.23	0.19	0.28
			1	0.23	0.27	0.38	0.40	0.12	0.26	0.16	0.26
				1	0.99	0.70	0.69	-0.35	-0.37	-0.56	-0.64
					1	0.71	0.71	-0.37	-0.37	-0.55	-0.61
						1	0.99	-0.43	-0.46	-0.21	-0.22
							1	-0.47	-0.48	-0.27	-0.25
								1	0.92	0.75	0.57
									1	0.69	0.63
										1	0.90
											1

Table 3.26: Principal Component Coefficients and Variance Proportion of Face Recognition Data Based on Classical Method Using Endpoints

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
AD	-0.28	-0.32	0.14		0.51			
	-0.28	-0.31	0.20		0.54	-0.18		
BC	-0.23	-0.36	-0.11	0.39	-0.31			-0.61
	-0.20	-0.36	-0.13	0.57	-0.20		-0.20	0.54
AH	-0.36	0.10	-0.43	-0.21		0.33		
	-0.37		-0.40	-0.18		0.40		
DH	-0.39		0.18	-0.33	-0.31	-0.23		0.15
	-0.39		0.19	-0.24	-0.33	-0.33	0.30	0.11
EH	0.24	-0.31	-0.45	-0.29		-0.43	-0.30	-0.26
	0.22	-0.35	-0.43		0.14	-0.11	0.62	0.33
GH	0.21	-0.38	0.13	-0.40	-0.25	0.25	-0.48	0.24
	0.19	-0.40	0.31	-0.12	-0.14	0.51	0.39	-0.26
Proportion	0.46	0.33	0.09	0.06	0.04	0.01	0.01	0.00
Cumulative	0.46	0.79	0.88	0.93	0.97	0.99	0.99	1.00

Table 3.27: Correlation between Principal Components and Random Variables of Face Recognition Data Based on Classical Method Using Endpoints

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
AD	-0.65	-0.64	0.15		0.35			
	-0.65	-0.62	0.20		0.38	-0.07		
BC	-0.54	-0.72	-0.11	0.33	-0.21			-0.13
	-0.47	-0.71	-0.13	0.48	-0.14		-0.06	0.11
AH	-0.84	0.20	-0.44	-0.18		0.12		
	-0.87		-0.41	-0.15		0.15		
DH	-0.90		0.19	-0.27	-0.21	-0.08		0.03
	-0.92		0.19	-0.20	-0.23	-0.12	0.09	0.02
EH	0.55	-0.61	-0.47	-0.24		-0.16	-0.09	-0.05
	0.51	-0.69	-0.45		0.10	-0.04	0.18	0.07
GH	0.49	-0.75	0.14	-0.33	-0.17	0.09	-0.14	0.05
	0.45	-0.79	0.32	-0.10	-0.09	0.18	0.11	-0.05

Chapter 4

Principal Component Analysis for Histogram-Valued Data

In this chapter we propose a method of principal component analysis (PCA) for histogram-valued observations. This method is the first such methodology developed for a principal component analysis for histogram-valued data. It is a generalization of the PCA method for interval-valued observations proposed in Chapter 3. This chapter is divided into four sections. Some notation and statistics for histogram-valued data are given in Section 4.1. Section 4.2 explains the basis for generalization of interval-valued data to histogram-valued data. Section 4.3 gives an algorithm to compute the principal components for histogram-valued observations along with detailed descriptions of the method. An example illustrating the proposed method is given in Section 4.4.

4.1 Preliminaries

Section 2.1.3 gives a brief introduction to histogram-valued data as well as some descriptive statistics for this type of data. In this section, we restate some notation and results necessary for the development of our proposed method. A more extensive treatment of histogram-valued data can be found in Billard and Diday (2003) and Billard and Diday (2006).

Let $\boldsymbol{X}$ be an $n \times p$ symbolic data matrix

$$\boldsymbol{X} = \begin{bmatrix} \xi_{11} & \xi_{12} & \dots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \dots & \xi_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \xi_{n1} & \xi_{n2} & \dots & \xi_{np} \end{bmatrix}. \tag{4.1}$$

The matrix $\boldsymbol{X}$ is a histogram-valued data matrix if an element ij of $\boldsymbol{X}$ is a histogram ξ_{ij} defined in Equation (2.13), where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$. That is,

$$\xi_{ij} = \left\{ [a_{ij}^1, b_{ij}^1), p_{ij}^1; [a_{ij}^2, b_{ij}^2), p_{ij}^2; \ldots; [a_{ij}^{s_{ij}}, b_{ij}^{s_{ij}}], p_{ij}^{s_{ij}} \right\},$$

where $[a_{ij}^l, b_{ij}^l)$ is called the l^{th} subinterval of ξ_{ij} and p_{ij}^l is the relative frequency associated with the l^{th} subinterval. Let s_{ij} denote the number of subintervals in histogram ξ_{ij}. Then, $a_{ij}^l \leq b_{ij}^l$ for all $l = 1, 2, \ldots, s_{ij}$ and $\sum_{l=1}^{s_{ij}} p_{ij}^l = 1$.

Billard and Diday (2003) gives the following empirical density function of a point $W \in X_{(j)}$ as defined in Equation (2.15),

$$f_W(\xi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l:\xi \in \xi_{ij}^l} p_{ij}^l \left(\frac{1}{b_{ij}^l - a_{ij}^l}\right).$$

The symbolic sample mean and the symbolic sample variance derived from the density function of Equation (2.15) are given in Equations (2.16) and (2.17), respectively, as follow,

$$\bar{W} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{l=1}^{s_{ij}} p_{ij}^l (a_{ij}^l + b_{ij}^l)$$

and

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

Billard and Diday (2006) extends the variance in equation (2.17) to the bivariate case to obtain the empirical symbolic covariance for $W_{(j)}$ in $X_{(j)}$ and $W_{(j')}$ in $X_{(j')}$ defined in Equation (2.18) as

$$S_{jj'} = \frac{1}{3n} \sum_{i=1}^{n} \left(G_{ij} G_{ij'} \sum_{l_1=1}^{s_{ij}} \sum_{l_2=1}^{s_{ij'}} p_{ij}^{l_1} p_{ij'}^{l_2} [Q_{ij}^{l_1} Q_{ij'}^{l_2}]^{1/2} \right)$$

where

$$\begin{aligned} Q_{ij}^{l_j} &= (a_{ij}^{l_j} - \bar{W}_{(j)})^2 + (a_{ij}^{l_j} - \bar{W}_{(j)})(b_{ij}^{l_j} - \bar{W}_{(j)}) + (b_{ij}^{l_j} - \bar{W}_{(j)})^2, \\ G_{ij} &= \begin{cases} -1, & \bar{W}_{ij} \leq \bar{W}_{(j)}, \\ 1, & \bar{W}_{ij} > \bar{W}_{(j)}, \end{cases} \end{aligned}$$

and $\bar{W}_{ij} = \frac{1}{2} \sum_{l_j=1}^{s_{ij}} p_{ij}^{l_j} (a_{ij}^{l_j} + b_{ij}^{l_j})$.

If all subintervals of a histogram have equal relative frequencies, then $p_{ij}^1 = p_{ij}^2 = \ldots = p_{ij}^{s_{ij}}$. Let $p_{ij} = p_{ij}^1$. Then, $\sum_{l=1}^{s_{ij}} p_{ij}^l = s_{ij}p_{ij} = 1$. In this case, the histogram ξ_{ij} as defined in Equation (2.13) can be rewritten as

$$\xi_{ij} = \{[a_{ij}^1, b_{ij}^1), p_{ij}; \ldots; [a_{ij}^{s_{ij}}, b_{ij}^{s_{ij}}], p_{ij}\}.$$

Since the relative frequency is the same for all subintervals, all values within the entire histogram are uniformly distributed. Combining the subintervals into one interval gives

$$\xi_{ij} = \{[a_{ij}^1, b_{ij}^{s_{ij}}], s_{ij}p_{ij}\} = \{[a_{ij}^1, b_{ij}^{s_{ij}}], 1\}$$

or simply $\xi_{ij} = [a_{ij}, b_{ij}]$ which is an interval with lower endpoint $a_{ij} = a_{ij}^1$ and upper endpoint $b_{ij} = b_{ij}^{s_{ij}}$. Therefore, interval-valued data can be thought of as a special case of histogram-valued data. Based on this special relationship between a histogram and an interval, we propose a method of principal component analysis for histogram-valued data by generalizing our proposed PCA method for interval-valued data presented in Chapter 3. In Section 4.2 of this chapter, we present details of this generalization. Section 4.3 includes two algorithms to compute the principal components of a histogram-valued dataset. The first algorithm reconstructs histogram-valued observations in a principal components space and the second algorithm computes histogram-valued principal components for those observations.

4.2 Methodology

For ease of reference, the layout of this section mirrors that of Section 3.2. This section is divided into two subsections. The first subsection, Subsection 4.2.1, explains how the coefficients of the principal components are obtained. The second subsection, Subsection 4.2.2, explains the basis for extending our proposed PCA method for interval-valued data to histogram-valued data. Subsection 4.2.2 is further divided into two parts. Part one of Subsection 4.2.2 explores the geometric structure of a histogram-valued observation and its relation to the structure of an interval-valued observation. It also explains how the algorithm

of Subsection 3.3.1 can be applied to histogram-valued observations. Part two of Subsection 4.2.2 describes the process of creating histogram-valued principal components.

4.2.1 Finding the coefficients of the principal components

In our proposed method, coefficients of the principal components for histogram-valued data matrix $\boldsymbol{X}$ are obtained in the same manner as in classical and in interval-valued PCA. The theory for coefficients of classical and interval-valued PCA carries through for histogram-valued data. Therefore, the approach to finding the coefficients in this subsection mirrors that of Subsection 3.2.2 with one exception; the symbolic covariance matrix, $\boldsymbol{S}$, used in this subsection is the sample variance-covariance matrix of a histogram-valued data matrix $\boldsymbol{X}$ whose elements are as defined in Equation (2.13). That is, the sample variance-covariance of $\boldsymbol{X}$ is given by

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix} \tag{4.2}$$

where $S_{jj'}$ is defined in Equation (2.18) and $S_{jj'} = S_{j'j}$ for $j, j' = 1, 2, \dots, p$. Let $\hat{\lambda}_1^S > \hat{\lambda}_2^S > \dots > \hat{\lambda}_p^S$ be the eigenvalues of $\boldsymbol{S}$ and $\boldsymbol{\nu}_1^S, \boldsymbol{\nu}_2^S, \dots, \boldsymbol{\nu}_p^S$ be their corresponding eigenvectors. By analogy, the vector of coefficients, also called the loadings, for the k^{th} principal component of $\boldsymbol{X}$ is $\boldsymbol{\nu}_k^S$.

Analogous to classical and interval-valued PCA, the magnitude of the j^{th} element of $\boldsymbol{\nu}_k^S$, denoted by ν_{jk}^S, indicates the contribution of the variable $X_{(j)}$ to the principal component $Y_{(k)}$. Therefore, the measure of correlation between an individual variable $X_{(j)}$ and a principal component $Y_{(k)}$ for histogram-valued data is the same as that defined in Equation (3.3) for interval-valued data which is given by,

$$\rho_{X_{(j)},Y_{(k)}} = \frac{\nu_{jk}\sqrt{\hat{\lambda}_k^S}}{\sqrt{S_{jj}}}. \tag{4.3}$$

Again, similar to the PCA method for interval-valued observations, assume $\boldsymbol{X}_i, i = 1, \ldots, n$, is a random sample and n is large. Let $W_{(j)}$ be a point from $X_{(j)}$ for $j = 1, 2, \ldots, p$. In addition, assume $\boldsymbol{W} = (W_{(1)}, W_{(2)}, \ldots, W_{(p)})$ is from a normal population and the eigenvalues of the population covariance matrix $\boldsymbol{\Sigma}$ are distinct and positive. Then, the asymtotic results regarding $\hat{\lambda}_k$ and $\boldsymbol{\nu}_k$ stated in Section 2.2 carry through for $\hat{\lambda}_k^S$ and $\boldsymbol{\nu}_k^S$ for all $k = 1, 2, \ldots, p$.

In situations where PCA based on the sample correlation is more appropriate, coefficients for the k^{th} principal component of $\boldsymbol{X}$ are then the eigenvector corresponding to the k^{th} eigenvalue of the sample correlation matrix $\boldsymbol{R}$. The jj' element of $\boldsymbol{R}$, denoted by $\rho_{jj'}$, is

$$\rho_{jj'} = \frac{S_{jj'}}{\sqrt{S_{jj} S_{j'j'}}}. \tag{4.4}$$

The measures of contribution to the principal components based on the correlation matrix mirror the results for the principal components based on the sample covariance matrix.

4.2.2 Constructing the principal components

Due to the complexity of histogram-valued data, we describe the process for constructing observations in a principal components space by working with one observation at a time. That is, the method proposed in this section works on observation X_i for $i = 1, 2, \ldots, n$, and must be applied n times.

Geometric representation of histogram-valued observations

First, let $X_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{ip})$ be a histogram-valued observation in a p-dimensional space. As defined in Equation (2.13), each variable j, for $j = 1, 2, \ldots, p$, is a histogram,

$$\xi_{ij} = \{([a_{ij}^l, b_{ij}^l), p_{ij}^l) | l = 1, 2, \ldots, s_{ij}\},$$

where s_{ij} is the number of subintervals. Since all points within each subinterval $[a_{ij}^l, b_{ij}^l)$ of ξ_{ij} are assumed to have uniform density, the subinterval itself can be thought of as an interval-valued variable with the relative frequency p_{ij}^l as its weight. Thus, each histogram ξ_{ij} can be thought of as a compilation of s_{ij} observed intervals with associated weights. Based on

this idea, a histogram-valued observation X_i can be divided into r weighted interval-valued observations where

$$r = \prod_{j=1}^{p} s_{ij}. \tag{4.5}$$

As stated in Chapter 2, an interval-valued observation is represented by a hyper-rectangle in a p-dimensional sample space whose density is uniform. Therefore, a histogram-valued observation can be thought of as a hyper-rectangle that is partitioned into r sub-hyperrectangles in the sample space. Each sub-hyperrectangle has uniform density. However, the density may differ from one sub-hyperrectangle to the next.

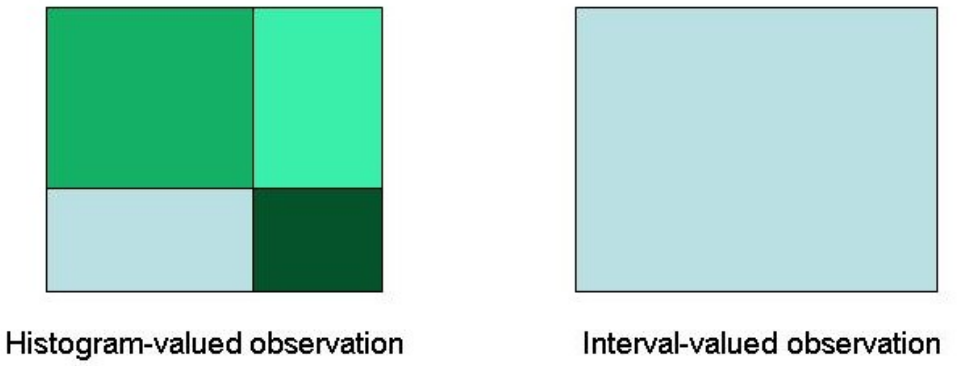

Figure 4.1: Histogram-Valued Observation and Interval-Valued Observation

Figure 4.1 shows a rectangle representing a histogram-valued observation (left) versus one representing an interval-valued observation (right) in a 2-dimensional space. Each variable in the histogram-valued observation shown in this example has two subintervals. Therefore, the rectangle representing this observation contains $r = 4$ sub-rectangles. The differences in density for these sub-rectangles are illustrated by the different colors. Whereas, the rectangle representing the interval-valued observation in this example is made up by only one rectangle, with the uniformity of its density illustrated by a single color.

Having established that a histogram-valued observation can be represented by a hyper-rectangle in the sample space, an observation X_i can be expressed in terms of its vertices. That is, X_i has an equivalent expression as a $(2^p r \times p)$ matrix, say $\boldsymbol{X}_i^{v2}$, where r is the

number of sub-hyperrectangles contained in observation X_i and is as defined in Equation (4.5). Each row of $\boldsymbol{X}_i^{v2}$ is the coordinate of a vertex of a sub-hyperrectangle belonging to X_i. Let E_{ij} be the set of subinterval endpoints for histogram ξ_{ij}. Then,

$$E_{ij} = \{a_{ij}^1, b_{ij}^1, a_{ij}^2, b_{ij}^2, \ldots, a_{ij}^{s_{ij}}, b_{ij}^{s_{ij}}\}. \tag{4.6}$$

The rows of $\boldsymbol{X}_i^{v2}$ include all possible permutations of the elements of $E_{i1}, E_{i2}, \ldots, E_{ip}$.

Without loss of generality, assume $b_{ij}^{(l-1)} = a_{ij}^l$ for $l = 1, \ldots, s_{ij}$. By eliminating redundant subinterval endpoints, we obtain $E_{ij} = \{a_{ij}^1, a_{ij}^2, \ldots, a_{ij}^{s_{ij}}, b_{ij}^{s_{ij}}\}$. For the consistency of notation, let $a_{ij}^{s_{ij}+1} = b_{ij}^{s_{ij}}$. Then,

$$E_{ij} = \{a_{ij}^l | l = 1, 2, \ldots, s_{ij} + 1\}. \tag{4.7}$$

Let $\boldsymbol{X}_i^v$ be the matrix whose rows include all possible permutations of the elements of $E_{i1}, E_{i2}, \ldots, E_{ip}$ defined in Equation (4.7). Then, $\boldsymbol{X}_i^v$ has $N_i = \prod_{j=1}^p (s_{ij} + 1)$ rows whereas $\boldsymbol{X}_i^{v2}$ has $2^p r = 2^p(\prod_{j=1}^p s_{ij})$. The quantity $\prod_{j=1}^p (s_{ij} + 1)$ is less than or equal to $2^p(\prod_{j=1}^p s_{ij})$. The amount of difference between these two quantities depends on the number of subintervals for each histogram ij and the number of variables, p. When $s_{ij} = 1$ for all $j = 1, \ldots, p$,

$$\frac{\prod_{j=1}^p (s_{ij} + 1)}{2^p(\prod_{j=1}^p s_{ij})} = \frac{\prod_{j=1}^p 2}{2^p(\prod_{j=1}^p 1)} = \frac{2^p}{2^p} = 1. \tag{4.8}$$

When s_{ij} is large for most ξ_{ij},

$$\frac{\prod_{j=1}^p (s_{ij} + 1)}{2^p(\prod_{j=1}^p s_{ij})} \to \frac{1}{2^p}. \tag{4.9}$$

For example, when $p = 2$, suppose observation i is the histogram-valued observation shown in Figure 4.2. Both variables of this example take histograms of values consisting of two subintervals. Therefore, the matrix of vertices $\boldsymbol{X}_i^v$ represents the histogram-valued

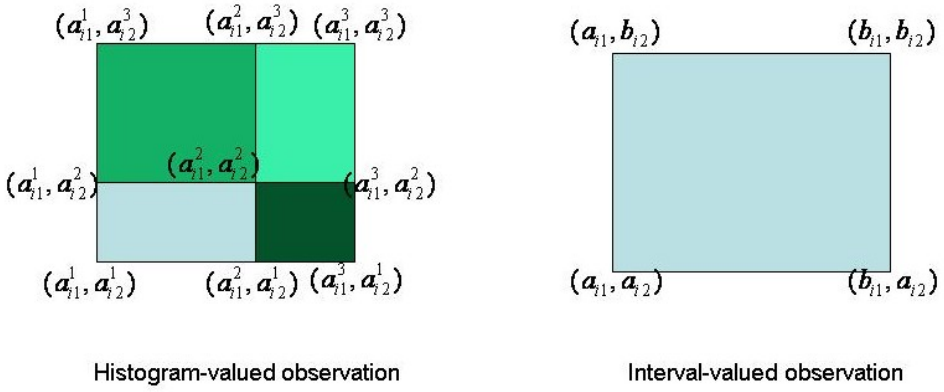

Figure 4.2: Vertices of Histogram-Valued Observation and Interval-Valued Observation

rectangle in Figure 4.2 has $N_i = \prod_{j=1}^{2}(2+1) = 9$ rows and

$$\boldsymbol{X}_i^v = \begin{bmatrix} a_{i1}^1 & a_{i2}^1 \\ a_{i1}^1 & a_{i2}^2 \\ a_{i1}^1 & a_{i2}^3 \\ a_{i1}^2 & a_{i2}^1 \\ a_{i1}^2 & a_{i2}^2 \\ a_{i1}^2 & a_{i2}^3 \\ a_{i1}^3 & a_{i2}^1 \\ a_{i1}^3 & a_{i2}^2 \\ a_{i1}^3 & a_{i2}^3 \end{bmatrix}. \tag{4.10}$$

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

$$\boldsymbol{X}_i^v = \begin{bmatrix} a_{i1}^1 & a_{i2}^1 & \dots & a_{ip}^1 \\ a_{i1}^1 & a_{i2}^1 & \dots & a_{ip}^2 \\ a_{i1}^1 & a_{i2}^1 & \dots & a_{ip}^3 \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1}^1 & a_{i2}^1 & \dots & a_{ip}^{s_{ip}} \\ a_{i1}^1 & a_{i2}^1 & \dots & a_{ip}^{s_{ip}+1} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1}^1 & a_{i2}^2 & \dots & a_{ip}^1 \\ a_{i1}^1 & a_{i2}^2 & \dots & a_{ip}^2 \\ a_{i1}^1 & a_{i2}^2 & \dots & a_{ip}^3 \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1}^1 & a_{i2}^2 & \dots & a_{ip}^{s_{ip}} \\ a_{i1}^1 & a_{i2}^2 & \dots & a_{ip}^{s_{ip}+1} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1}^{s_{i1}+1} & a_{i2}^{s_{i2}+1} & \dots & a_{ip}^1 \\ a_{i1}^{s_{i1}+1} & a_{i2}^{s_{i2}+1} & \dots & a_{ip}^2 \\ a_{i1}^{s_{i1}+1} & a_{i2}^{s_{i2}+1} & \dots & a_{ip}^3 \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1}^{s_{i1}+1} & a_{i2}^{s_{i2}+1} & \dots & a_{ip}^{s_{ip}} \\ a_{i1}^{s_{i1}+1} & a_{i2}^{s_{i2}+1} & \dots & a_{ip}^{s_{ip}+1} \end{bmatrix}. \tag{4.11}$$

Now, the vertices belonging to observation i in the sample space which are represented by the matrix of vertices $\boldsymbol{X}_i^v$ can be transformed into points in a principal components space. Let $\boldsymbol{Y}_i^v$ be the matrix consisting of these transformed vertices of observation i. That is,

$$\boldsymbol{Y}_i^v = \boldsymbol{X}_i^v \boldsymbol{\nu}^S \tag{4.12}$$

where $\boldsymbol{\nu}^S = [\boldsymbol{\nu}_1^S, \boldsymbol{\nu}_2^S, \dots, \boldsymbol{\nu}_p^S]$ is the eigen matrix of the symbolic covariance matrix $\boldsymbol{S}$ of Equation (4.2) if the PCA is based on the covariance structure of the data and $\boldsymbol{\nu}^S$ is the

eigen matrix of the symbolic correlation matrix $\boldsymbol{R}$ whose elements are defined in Equation (4.4) if the PCA is based on the correlation structure of the data.

As explained in Subsection 3.2.2, a hyper-rectangle 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 a principal component space. The vertices of the polytope are the transformed vertices of the sub-hyperrectangle. The algorithm described in Subsection 3.3.1 reconstructs polytopes for a set of n interval-valued observations. To construct the polytopes for a histogram-valued observation i based on this algorithm, we treat the r sub-hyperrectangles belonging to histogram-valued observation i as r interval-valued observations. That is, each histogram-valued observation i is treated as a dataset of r interval-valued observations but with weights, i.e., with densities, say d_i^h, for $h = 1, 2, \ldots, r$.

To apply the algorithm of Subsection 3.3.1 to a histogram-valued observation i, we need to construct a matrix of vertices for each sub-hyperrectangle h belonging to observation i. Let $\boldsymbol{X}_i^h$ be a matrix of vertices for sub-hyperrectangle h of observation i. Then, the rows of $\boldsymbol{X}_i^h$ come from the matrix of vertices $\boldsymbol{X}_i^v$. Detailed construction of matrix $\boldsymbol{X}_i^v$ and matrices $\boldsymbol{X}_i^h$ is given in Subsections 4.3.1 and 4.3.2. Using the coordinates in $\boldsymbol{X}^h$, algorithm 3.3.1 can now be readily applied. To illustrate the construction of matrix $\boldsymbol{X}_i^h$, go back to the example of Figure 4.2. Observation i in this example is composed of four sub-rectangles.
When $h = 1$, i.e., for the first sub-rectangle, the matrix of vertices $\boldsymbol{X}_i^1$ is given by

$$\boldsymbol{X}_i^1 = \begin{bmatrix} a_{i1}^1 & a_{i2}^1 \\ a_{i1}^1 & a_{i2}^2 \\ a_{i1}^2 & a_{i2}^1 \\ a_{i1}^2 & a_{i2}^2 \end{bmatrix} \tag{4.13}$$

which is a matrix consisting of rows 1, 2, 4, and 5 of matrix $\boldsymbol{X}_i^v$ in Equation (4.10). When $h = 2$, i.e., for the second sub-rectangle, the matrix of vertices $\boldsymbol{X}_i^2$ is given by

$$\boldsymbol{X}_i^2 = \begin{bmatrix} a_{i1}^1 & a_{i2}^2 \\ a_{i1}^1 & a_{i2}^3 \\ a_{i1}^2 & a_{i2}^2 \\ a_{i1}^2 & a_{i2}^3 \end{bmatrix} \tag{4.14}$$

which is a matrix consisting of rows 2, 3, 5, and 6 of matrix $\boldsymbol{X}_i^v$ in Equation (4.10). Matrices of vertices for the third and fourth sub-rectangle can be constructed in the same manner.

Although densities (or weights), d_i^h for $h = 1, 2, \ldots, r$, of the sub-hyperrectangles in observation i vary, illustrating this variability presents some challenges. The first challenge is due to the fact that only the surface of a hyper-rectangle is visible in a plot. Interior points of a hyper-rectangle are shielded by the boundary points. Therefore, it is impossible to illustrate the densities of the interior sub-hyperrectangles of a hyper-rectangle. Hence, only densities of the sub-hyperrectangles formed by the first or the last subintervals can be visualized. The second challenge includes computational complexity. If densities are specified by different colors, then all polytopes must be filled with the color associated with their density. Writing a program to automate this process is a time-consuming and a challenging project. This can be a potential future project extended from this dissertation.

With the challenges presented in the previous paragraph in mind, we propose an alternate approach to understand the variability in densities of the sub-hyperrectangles by constructing a matrix of densities associated with the sub-hyperrectangles. Let $\boldsymbol{d}_i$ be an r-vector of densities. Then, the h^{th} element of $\boldsymbol{d}_i$ is the density of the h^{th} sub-hyperrectangle of observation i. For $j = 1, 2, \ldots, p$, and $l_j = 1, 2, \ldots, s_{ij}^{ij}$,

$$d_i^h = \prod_{j=1}^{p} p_{ij}^{l_j} \tag{4.15}$$

where $h = \sum_{j=1}^{p-1} (l_j - 1) s_{i,j+1} + l_p$. That is, e.g., for $h = 1$, the first sub-hyperrectangle of observation i is formed by the first subinterval of histograms ξ_{ij} for all $j = 1, \ldots, p$. Then

$l_j = 1$ for all j. Thus,

$$h = \sum_{j=1}^{p-1} (l_j - 1)s_{i,j+1} + l_p = 0 + 1 = 1.$$

Hence, the first element of $\boldsymbol{d}_i$ is

$$d_i^1 = \prod_{j=1}^{p} p_{ij}^{l_j} = \prod_{j=1}^{p} p_{ij}^1.$$

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

$$h = \sum_{j=1}^{p-1} (l_j - 1)s_{i,j+1} + l_p = 0 + 2 = 2.$$

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

$$d_i^2 = \prod_{j=1}^{p} p_{ij}^{l_j} = \left(\prod_{j=1}^{p-1} p_{ij}^1 \right) p_{ip}^2.$$

Other elements of $\boldsymbol{d}_i$ can be found the same way. Now, the value of d_i^h is the density of the sub-hyperrectangle whose vertices are expressed in matrix $\boldsymbol{X}_i^h$. In addition to their usage in constructing the polytopes for observation i in a principal components space, matrices $\boldsymbol{X}_i^h$ for $h = 1, 2, \ldots, r$ and vector $\boldsymbol{d}_i$ are used in constructing histograms for the principal components as explained in the second part of Subsection 4.2.2 as follows.

Histogram-valued principal components for histogram-valued observations

To construct histogram-valued principal components for histogram-valued observations, we use the same approach used in constructing the polytopes of these observations, see part one of Subsection 4.2.2 immediately above. We propose treating each histogram-valued observation i as itself a dataset. Each sub-hyperrectangle of observation i is treated as an interval-valued observation in this dataset. Now, for the k^{th} principal component, the algorithm detailed in Subsection 3.3.2 can be applied to create r histograms for r sub-hyperrectangles of observation i where r is as defined in Equation (4.5). Since observation i is composed of

r sub-hyperrectangles, we propose combining the r histograms, constructed from the r sub-hyperrectangles, into one histogram representing observation i. This is done in Section 4.3 where Subsection 4.3.2 describes specific steps involved in creating a histogram of histograms.

4.3 Algorithm

This section includes descriptions of two algorithms. The first algorithm, presented in Subsection 4.3.1, constructs matrices of vertices for histogram-valued observations and builds polytopes representing these observations. The second algorithm, presented in Subsection 4.3.2, constructs histograms for the principal components of histogram-valued observations.

Similar to the algorithms proposed for interval-valued observations described in Section 3.3, the following algorithms are based on the assumption that coefficients for the principal components have been determined from the proposed method described in Subsection 4.2.1. The coefficients of the principal components are $\boldsymbol{\nu}^S = [\boldsymbol{\nu}_1^S, \boldsymbol{\nu}_2^S, \ldots, \boldsymbol{\nu}_p^S]$ where $\boldsymbol{\nu}_k^S, k = 1, \ldots, p$, are the eigenvectors of the covariance matrix $\boldsymbol{S}$ as defined in Equation (4.2) if the PCA is based on the covariance structure of the data and $\boldsymbol{\nu}_k^S, k = 1, \ldots, p$, are the eigenvectors of the correlation matrix $\boldsymbol{R}$ whose elements are defined in Equation (4.4) if the PCA is based on the correlation structure of the data.

In the following algorithms, we use the same indexing system as described in the algorithms of Section 3.3. 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 the row, the column, and the matrix, respectively.

For a variable j in a histogram-valued dataset, a histogram representing observation i_1 may have a different number of subintervals from observation i_2 for $i_1 \neq i_2$ and $i_1, i_2 = 1, 2, \ldots, n$. Due to its complexity, the structure and the principal components are constructed

one observation at a time. Therefore, the algorithm described in this section must be executed n times once for each of the n observations. Therefore, the data used in the following algorithm comes from the data vector $\boldsymbol{X}_i$ for $i = 1, 2, \ldots, n$.

Moreover, since the algorithms presented in this section compute principal components of the observed data, we use a lower case letter to represent an observed data matrix to distinguish it from a random data matrix. For example, $\boldsymbol{x}_i^v$ represents the observed version of the matrix of vertices $\boldsymbol{X}_i^v$ of Equation (4.11) and $\boldsymbol{y}_i^v$ is the observed version of the matrix of transformed vertices $\boldsymbol{Y}_i^v$ defined in Equation (4.12). In addition, the matrix of vertices of sub-hyperrectangle h, $\boldsymbol{X}_i^h$, constructed in the second algorithm of Subsection 4.3.1 is denoted by $\boldsymbol{x}^h$. The lower case in $\boldsymbol{x}^h$ signifies that this matrix consists of observed values. Because the algorithm is executed separately for each i, then i is fixed inside the algorithm; we drop the subscript i to keep the notation simple.

Before running the following algorithms, an $n \times 3(sn)$ matrix $\boldsymbol{pc}_k$ is created to store histogram values for principal component k where sn is the number of subintervals desired for principal component k.

4.3.1 Histogram-valued Observations in a Principal Components Space

This subsection is divided into two parts: the first part constructs the matrices of vertices $\boldsymbol{X}_i^v$ for observation $i = 1, 2, \ldots, n$, as defined in Equation (4.11) and the second part builds the polytopes representing observation i for $i = 1, 2, \ldots, n$, in a principal components space.

Constructing the matrix of vertices

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

$$\boldsymbol{x}_{ep} = \left[\begin{array}{cccccccccccc} a_{i1}^1 & a_{i1}^2 & \ldots & a_{i1}^{s_{i1}+1} & a_{i2}^1 & a_{i2}^2 & \ldots & a_{i2}^{s_{i2}+1} & \ldots & a_{ip}^1 & a_{ip}^2 & \ldots & a_{ip}^{s_{ip}+1} \end{array} \right] \quad (4.16)$$

where a_{ij}^l for $l = 1, \ldots, s_{ij} + 1$, and $j = 1, \ldots, p$, are elements of the set E_{ij} as defined in equation (4.7). The vector $\boldsymbol{x}_{rf}$ has $\sum_{j=1}^{p} s_{ij}$ elements and has the following form,

$$\boldsymbol{x}_{rf} = \left[\begin{array}{cccccccccccc} p_{i1}^1 & p_{i1}^2 & \ldots & p_{i1}^{s_{i1}} & p_{i2}^1 & p_{i2}^2 & \ldots & p_{i2}^{s_{i2}} & \ldots & p_{ip}^1 & p_{ip}^2 & \ldots & p_{ip}^{s_{ip}} \end{array} \right] \tag{4.17}$$

where p_{ij}^l is the relative frequency of the l^{th} subinterval of the observed histogram ξ_{ij}. Before creating the matrix of vertices for observation i, a p-vector whose elements are the number of subintervals for ξ_{ij} is also needed. Let $\boldsymbol{ns}$ denote the vector consisting of the number of subintervals of ξ_{ij}. Then,

$$\boldsymbol{ns} = \left[\begin{array}{cccc} s_{i1} & s_{i2} & \ldots & s_{ip} \end{array} \right]. \tag{4.18}$$

With the information in $\boldsymbol{x}_{ep}$, $\boldsymbol{x}_{rf}$, and $\boldsymbol{ns}$, we can proceed with constructing the matrix of vertices $\boldsymbol{x}_i^v$ using the following five steps:

Step 1. Create a $(p+1)$-vector $\boldsymbol{nr}$ whose $(j+1)^{th}$ element, for $j = 1, \ldots, p$, is the number of times that points a_{ij}^l, for $l = 1, \ldots, s_{ij} + 1$, must be repeated in Step 5 below. The first element of $\boldsymbol{nr}$ is the number of rows of the matrix of observed vertices, $\boldsymbol{x}_i^v$.

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

2. Set $\boldsymbol{nr}[p + 1] = 1$.

Step 2. Create a $(p+1)$-vector $\boldsymbol{nr}_p$ 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 $j = 1, 2, \ldots, p$,
   set $\boldsymbol{nr}_p[p - j + 1] = \prod_{l=1}^{j} s_{ij}$.

2. Set $\boldsymbol{nr}_p[p + 1] = 1$.

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

1. Set $\boldsymbol{sp}[1] = 1$.

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

Step 4. Initialize the matrix of observed vertices $\boldsymbol{x}_i^v$ by letting $\boldsymbol{x}_i^v$ 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 $\boldsymbol{x}_i^v$ by

1. For $j = 1, 2, \ldots, p$, do

   (a) Let $nj = \boldsymbol{ns}[j]$.

   (b) Let $rj = \boldsymbol{nr}[j+1]$.

   (c) Let $sj = \boldsymbol{sp}[j]$.

   (d) For $l_n = 0, 1, \ldots, nj$,

   - For $l_r = 1, 2, \ldots, rj$,
     set $\boldsymbol{x}_i^v[l_n(rj) + l_r, j] = \boldsymbol{x}_{ep}[sj + l_n]$.

2. For $j = 2, 3, \ldots, p$, do

   (a) Let $tj = \frac{\boldsymbol{nr}[1]}{\boldsymbol{nr}[j]} - 1$.

   (b) Let $rj = \boldsymbol{nr}[j]$.

   (c) For $l_t = 1, 2, \ldots, tj$,

   - For $l_r = 1, 2, \ldots, r_j$,
     set $\boldsymbol{x}_i^v[l_t(rj) + l_r, j] = \boldsymbol{x}_i^v[l_r, j]$.

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

Constructing the polytopes

We now present the algorithm to construct the polytopes representing the sub-hyperrectangles of observation i in a principal component space. The following algorithm includes seven steps:

Step 1. First, compute the matrix of transformed vertices, $\boldsymbol{y}_i^v$, for the polytope representing observation i in a principal components space based on the two cases:

Case one. If the PCA is based on the symbolic covariance matrix, then $\boldsymbol{\nu}^S$ is the eigen matrix of the symbolic covariance matrix $\boldsymbol{S}$ in Equation (4.2) and $\boldsymbol{y}^v = (\boldsymbol{x}^v)(\boldsymbol{\nu}^S)$.

Case two. If the PCA is based on the symbolic correlation matrix, then $\boldsymbol{\nu}^S$ is the eigen matrix of the symbolic correlation matrix $\boldsymbol{R}$ whose elements are defined in Equation (4.4) and $\boldsymbol{y}^v = (\boldsymbol{z}^v)(\boldsymbol{\nu}^S)$ where $\boldsymbol{z}^v$ is the standardized matrix of vertices. That is, element mj of $\boldsymbol{z}^v$ comes from

$$\boldsymbol{z}^v[m, j] = \frac{\boldsymbol{x}^v[m, j] - \bar{w}_{(j)}}{\sqrt{s_{jj}}}$$

where $\bar{w}_{(j)}$ is the mean of the j^{th} variable defined in Equation (2.16) and s_{jj} is the variance of the j^{th} variable defined in Equation (2.17).

Step 2. Next, create a three-dimensional array $\boldsymbol{x}_a$ to hold the matrices of vertices of the sub-hyperrectangles, $\boldsymbol{x}^h$ for $h = 1, 2, \ldots, r$. The array $\boldsymbol{x}_a$ is a result of combining $r = \prod_{j=1}^{p} s_{ij}$ matrices $\boldsymbol{x}^h$ where $h = 1, 2, \ldots, r$ where each matrix $\boldsymbol{x}^h$ of dimension $(2^p \times p)$ contains coordinates of all vertices belonging to sub-hyperrectangle h.

1. Initialize array $\boldsymbol{x}_a$ by letting $\boldsymbol{x}_a$ be an array of zeros with dimension $(2^p \times p \times r)$.

2. Update the elements of $\boldsymbol{x}_a$ 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 $kr_j = kr_{j-1} + (\boldsymbol{nr}[j+1])l_j$.

ii. Let $ni_j = ni_{j-1} + (\boldsymbol{nr}_p[j+1])l_j$.

iii. For $k = 1, 2, \ldots, \boldsymbol{ns}[p]$,

A. Set $kr = kr_{p-1} + k$.

B. Set $ni = ni_{p-1} + k$.

C. Set $\boldsymbol{x}_a[1, , ni] = \boldsymbol{y}_i[kr,]$.

D. For $o = 1, \ldots, p$, do,

For $r = 1, \ldots, 2^{(o-1)}$,

set $\boldsymbol{x}_a[2^{(o-1)} + r, , ni] = \boldsymbol{y}_i[kr[r] + \boldsymbol{nr}[p - o + 2],]$, and

set $kr = (kr, kr[r] + \boldsymbol{nr}[p - o + 2])$.

End of Step 2.

Step 3. Next, reconstruct the polytopes corresponding to the sub-hyperrectangles of observation i by following Steps 3 and 4 of Section 3.3.1 by replacing $\boldsymbol{y}^v$ by $\boldsymbol{y}_i^v$. Similarly, two and three dimensional plots can also be created for observation i by following all steps described in Subsection 3.3.1 with $\boldsymbol{y}_i^v$ in place of $\boldsymbol{y}^v$.

At the end of Step 3 of the present subsection, polytopes representing observation i in a principal components are plotted. To create the vector of densities for these polytopes, follow the next 4 steps, Steps 4-7.

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

1. Set $\boldsymbol{sp}_p[1] = 1$.

2. For $j = 1, 2, \ldots, p - 1$,

set $\boldsymbol{sp}_p[j+1] = \sum_{l=1}^{j} (s_{ij} + 1)$.

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

Step 6. Update the elements of $\boldsymbol{x}_p^v$ by

1. For $j = 1, 2, \ldots, p$, do

   (a) Let $nj = \boldsymbol{ns}[j]$.

   (b) Let $rj = \boldsymbol{nr}_p[j+1]$.

   (c) Let $sj = \boldsymbol{sp}_p[j]$.

   (d) For $l_n = 0, 1, \ldots, nj - 1$,

   - For $l_r = 1, 2, \ldots, rj$,

     set $\boldsymbol{x}_p^v[l_n(rj) + l_r, j] = \boldsymbol{x}_{rf}[sj + l_n]$.

2. For $j = 2, 3, \ldots, p$, do

   (a) Let $tj = \frac{\boldsymbol{nr}_p[1]}{\boldsymbol{nr}_p[j]} - 1$.

   (b) Let $rj = \boldsymbol{nr}_p[j]$.

   (c) For $l_t = 1, 2, \ldots, tj$

   - For $l_r = 1, 2, \ldots, rj$,

     set $\boldsymbol{x}_p^v[l_t(rj) + l_r, j] = \boldsymbol{x}_p^v[l_r, j]$.

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

At the end of Step 7, we obtain a vector of densities $\boldsymbol{d}_i$ whose h^{th} element is the density of sub-hyperrectangle h of observation i. It is now possible to construct the histograms representing the principal components for histogram-valued observation i; see Section 4.3.2.

4.3.2 Construction of histogram-valued principal components

The steps described in the following algorithm construct a histogram representing the k^{th} principal component for observation i based on the algorithm described in Subsection 3.3.2.

by first computing the histograms representing the principal components for each of the r sub-hyperectangles of observation i, then combining the r histograms to form one histogram representing principal component k for observation i. This algorithm is performed for $k = 1, 2, \ldots, p$. This algorithm includes eight steps as follows:

Step 1. Let h be the index for the r sub-hyperrectangles of observation i, $h = 1, 2, \ldots, r$.

Step 2. Let $\boldsymbol{pc}_{ik}$ be an $(r \times 3(2^p))$ matrix of zeros. The h^{th} row of $\boldsymbol{pc}_{ik}$ contains the endpoints and the relative frequencies of the histogram representing the k^{th} principal component for sub-hyperrectangle h of observation i.

Step 3. Let $\boldsymbol{pc}_s$ be an r-vector of zeros. This vector is used to store the number of subintervals belonging to the histogram representing sub-hyperrectangle h that have relative frequency greater than 0.

Step 4. Execute the following loop, which is Step 3 of the algorithm described in Subsection 3.3.2 with i replaced by h and other modifications, as shown in the following six substeps: For $h = 1, 2, \ldots, r$,

1. Set up the following matrices of zeros, $\boldsymbol{ps}$, $\boldsymbol{pd}$ and $\boldsymbol{s_i}$, as in Step 3.1 of the algorithm in Subsection 3.3.2.

2. Let $\boldsymbol{y}_h$ be an $(2^p \times 2)$ matrix whose first and second columns are the k^{th} and the first column of the matrix of vertices for sub-hyperrectangle h, $\boldsymbol{x}^h$, where $\boldsymbol{x}^h$ is the h^{th} matrix of array $\boldsymbol{x}_a$ constructed in Step 2 of the second algorithm described in Subsection 4.3.2. That is, $\boldsymbol{y}_h = (\boldsymbol{x}_a[, k, h], \boldsymbol{x}_a[, 1, h])$.

3. Follow Substeps 3.3 through 3.9 of the algorithm in Subsection 3.3.2 with $\boldsymbol{y}_h$ in place of $\boldsymbol{y}_{2i}$. At the end of this step, we obtain a matrix $\boldsymbol{hist}$ of dimension $nrw \times 4$ where nrw is the number of subintervals of the histogram representing sub-hyperrectangle h. The first two columns of $\boldsymbol{hist}$ store the subinterval endpoints. The third column stores the areas of the polygons bounded by those endpoints. Finally, the fourth column gives the relative frequencies of the subintervals by dividing the areas in column three by

the total area of the polygon that resulted from projecting polytope h onto the PC1 $\times$ PCk plane.

4. Before combining all histograms belonging to the r sub-hyperrectangles to make one histogram representing observation i, we need to account for the differences in density among the sub-hyperrectangles by multipling the relative frequencies of the subintervals by the densities of the sub-hyperrectangles. Therefore, add a fifth column to the matrix $\boldsymbol{hist}$ and for each element $s = 1, 2, \ldots, nrw$ of this column, set $\boldsymbol{hist}[s, 5] = (\boldsymbol{hist}[s, 4])(\boldsymbol{d}[h])$.

5. Update matrix $\boldsymbol{pc}_s$ which was initialized in Step 3 of this algorithm by letting $\boldsymbol{pc}_s[h] = nrw$.

6. To combine the nrw rows of matrix $\boldsymbol{hist}$ into one row in matrix $\boldsymbol{pc}_{ik}$, do the following which is equivalent to Substep 3.10 of algorithm in Subsection 3.3.2:
   For $s = 1, 2, \ldots, nrw$,
   set $\boldsymbol{pc}_{ik}[h, (3s-2)] = \boldsymbol{hist}[s, 1]$,
   set $\boldsymbol{pc}_{ik}[h, (3s-1)] = \boldsymbol{hist}[s, 2]$, and
   set $\boldsymbol{pc}_{ik}[h, (3s)] = \boldsymbol{hist}[s, 5]$.

End of Step 4.

Step 5. Let c_{pc} be the largest number of non-zero subintervals for each sub-hyperrectangle h. That is, $c_{pc} = max(\boldsymbol{pc}_s)$.

Step 6: Delete the columns $3c_{pc} + 1$ through $3(2^p)$ of $\boldsymbol{pc}_{ik}$ to keep the computation of the following Step 7 more efficient.

Step 7. This step includes seven substeps. Each complete execution of Step 7 makes a histogram representing the k^{th} principal component of observation i out of the r histograms in stored in matrix $\boldsymbol{pc}_{ik}$. The resulting histogram has subintervals with equal width.

1. Let lo and hi be the lowest and the highest endpoints of the histogram representing observation i. Then, $lo = min(\boldsymbol{y}_i[, k])$ and $hi = max(\boldsymbol{y}_i[, k])$.

2. Specify the number of subintervals to create for the histogram representing the k^{th} principal component for observation i by letting sn be the number of subintervals desired.

3. Let sw be the width of the subintervals. Then, $sw = (hi - lo)/sn$.

4. Let $\boldsymbol{hm}$ be an $(sn \times 4)$ transition matrix used to combine r histograms into one histogram for principal component k of observation i. Each row of $\boldsymbol{hm}$ stores values for subinterval s of the histogram representing the k^{th} principal component of observation i. The first two columns of $\boldsymbol{hm}$ store the subinterval endpoints. The third column stores the total relative frequencies of the r subintervals that intersect with the subinterval formed by the endpoints specified in the first two columns of $\boldsymbol{hm}$. Finally, the last column of $\boldsymbol{hm}$ gives the relative frequency for each subinterval by dividing the values of column three by the sum of the elements of column three. Initialize $\boldsymbol{hm}$ by setting its elements to zero.

5. Update $\boldsymbol{hm}$ by doing the following:

   For $s = 1, \ldots, ns$, do

   (a) Set the endpoints of subinterval s by letting $\boldsymbol{hm}[s, 1] = lo + (sw)(s - 1)$ and $\boldsymbol{hm}[s, 2] = lo + (sw)s$.

   (b) Let $\boldsymbol{fr}$ be an $r \times c_{pc}$ matrix whose $h \times q$ element corresponds to the proportion of subinterval q of polytope h that falls within the interval formed by $\boldsymbol{hm}[s, 1]$ and $\boldsymbol{hm}[s, 2]$. Initialize $\boldsymbol{fr}$ by setting its elements to zero.

   - For $h = 1, \ldots, r$, do

     For $q = 1, \ldots, \boldsymbol{pc}_s[h]$, do

     Case a: If $(\boldsymbol{pc}_{ik}[h, 3q - 2] \geq \boldsymbol{hm}[s, 1])$ and $(\boldsymbol{pc}_{ik}[h, 3q - 1] \leq \boldsymbol{hm}[s, 2])$, set $\boldsymbol{fr}[h, q] = \boldsymbol{pc}[h, 3q]$.

     Case b: If $(\boldsymbol{pc}_{ik}[h, 3q - 2] \geq \boldsymbol{hm}[s, 1])$ and $(\boldsymbol{pc}_{ik}[h, 3q - 2] < \boldsymbol{hm}[s, 2])$ and

$\boldsymbol{pc}_{ik}[h, 3q-1] > \boldsymbol{hm}[s, 2]$,

set $\boldsymbol{fr}[h, q] = (\boldsymbol{pc}_{ik}[h, 3q])(\boldsymbol{hm}[s, 2] - \boldsymbol{pc}_{ik}[h, 3q-2])/(\boldsymbol{pc}_{ik}[h, 3q-1] - \boldsymbol{pc}_{ik}[h, 3q-2])$.

Case c: If $(\boldsymbol{pc}_{ik}[h, 3q-2] < \boldsymbol{hm}[s, 1])$ and $(\boldsymbol{pc}_{ik}[h, 3q-1] > \boldsymbol{hm}[s, 1])$ and $(\boldsymbol{pc}_{ik}[h, 3q-1] \leq \boldsymbol{hm}[s, 2])$,

set $\boldsymbol{fr}[h, q] = (\boldsymbol{pc}_{ik}[h, 3q])(\boldsymbol{pc}_{ik}[h, 3q-1] - \boldsymbol{hm}[s, 1])/(\boldsymbol{pc}_{ik}[h, 3q-1] - \boldsymbol{pc}_{ik}[h, 3q-2])$.

Case d: If $(\boldsymbol{pc}_{ik}[h, 3q-2] < \boldsymbol{hm}[s, 1])$ and $(\boldsymbol{pc}_{ik}[h, 3q-1] > \boldsymbol{hm}[s, 2])$,

set $\boldsymbol{fr}[h, q] = (\boldsymbol{pc}_{ik}[h, 3q])(\boldsymbol{hm}[s, 2] - \boldsymbol{hm}[s, 1])/(\boldsymbol{pc}_{ik}[h, 3q-1] - \boldsymbol{pc}_{ik}[h, 3q-2])$.

(c) Let $\boldsymbol{hm}[s, 3] = \sum_{h=1}^{r} \sum_{q=1}^{\boldsymbol{pc}_s[h]} \boldsymbol{fr}[h, q]$.

6. Let $sh = \sum_{s=1}^{ns} \boldsymbol{hm}[s, 3]$.

7. Let $\boldsymbol{hm}[s, 4] = \boldsymbol{hm}[s, 3]/sh$.

End of Step 7. At the end of Step 7, we have all information for the histogram representing the k^{th} principal component for observation i stored in matrix $\boldsymbol{hm}$. This information can be entered into row i of the matrix $\boldsymbol{pc}_k$ by performing Step 8 as follows.

Step 8. For $s = 1, \ldots, ns$, do

set $\boldsymbol{pc}_k[i, 3s-2] = \boldsymbol{hm}[s, 1]$,

set $\boldsymbol{pc}_k[i, 3s-1] = \boldsymbol{hm}[s, 2]$, and

set $\boldsymbol{pc}_k[i, 3s] = \boldsymbol{hm}[s, 4]$.

Step 8 concludes the construction of the histogram for the k^{th} principal component for observation i.

End of algorithm. At the end of this algorithm, we obtain a histogram representing principal component k of observation i. All eight steps of the algorithm described in this subsection must be executed n times to obtain histograms for the k^{th} principal component of all n observations.

4.4 Medical Income Application

As stated in Chapter 1, symbolic data can occur naturally or they can result from aggregating very large datasets. With large databases becoming more common in recent years, symbolic data will become relevant for a wide range of applications in the near future. One area that can benefit tremendously from symbolic data analysis is medicine. A more particular example of applications in this area includes hospitals where records of patient visits are maintained electronically. Hospital databases contain millions of records. These databases provide an enormous amount of information on patient demographics, illnesses, treatments, and treatment results. When the subject of interest is not on individual patients but on groups of patients, individual records in a database may be aggregated; the resulting database is perforce a symbolic-valued dataset. We use an example in this area to illustrate our proposed method of principal component analysis for histogram-valued data.

4.4.1 The data

The dataset used in this section contains ten variables on 40,000 patients. The variables included Age, Race, Diabetes, Income, Glucose, Cholesterol, Hemoglobin, Hematocrit, Red blood, and White blood. This dataset is referred to as the Medical Income dataset. Table 4.1 lists the variables and their possible values. Each observation in the Medical Income dataset describes one patient and each variable takes only one value. That is, this dataset consists of 40,000 classical observations.

Suppose we want to study medical characteristics of groups of patients categorized by their age, race, and diabetes status. Classical observations of the Medical Income dataset can be aggregated into groups by crossing the first three variables, Age, Race, and Diabetes. Since Age is a continuous variable, we turned it into a categorical variable consisting of seven groups: 15-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, 65-74 years, and over 74 years. Crossing seven categories of age, two categories of race, and three categories of diabetes produces $7 \times 2 \times 3 = 42$ groups. Each of the 42 groups becomes an observation in

Table 4.1: Variables for Classical Medical Income Data

Variable	**Possible Values**
Age	count in years
Race	black and white
Diabetes	no, mild, and yes
Income	amount in dollars
Glucose	concentration in miligram/deciliter
Cholesterol	concentration in miligram/deciliter
Hemoglobin	concentration in gram/deciliter
Hematocrit	percentage of red blood cell volume in one microliter of blood
Redblood	count in million cells in one microliter of blood
Whiteblood	count in thousand cells in one microliter of blood

the symbolic dataset. The aggregated dataset now consists of 42 symbolic observations versus 40,000 classical observations in the original dataset. Table 4.2 gives a list of observations, their label, and characteristics for the aggregated Medical Income dataset.

With three classical variables from the original dataset used in creating symbolic observations where each observation consists of patients belonging to an Age $\times$ Race $\times$ Diabetes group, the aggregated dataset ended up with seven variables. These seven variables are labeled $X_{(j)}, j = 1, 2, \ldots, 7$ as shown in Table 4.3.

Each variable $X_{(j)}$ for $j = 1, \ldots, 7$, of a symbolic observation, ξ_i for $i = 1, 2, \ldots, 42$, in the aggregated dataset contains values from all classical observations in the orginal dataset that belong to ξ_i. A histogram was constructed from these values. That is, the realized values for each variable ξ_i is a histogram. Tables 4.4 through 4.10 gives observed histograms for the first five observations for each variable. The complete dataset is available upon request.

We computed the symbolic variance of Equation (2.17) and the symbolic covariance of Equation (2.18) for the Medical Income dataset shown in Tables 4.4 through 4.10. The

Table 4.2: Observation Labels for Symbolic Medical Income Data

i	**Label**	**Age**		**Race**		**Diabetes**	
		Code	Years	Code	Race	Code	Diabetes
1	100	1	15-24	0	white	0	no
2	101	1	15-24	0	white	1	mild
3	102	1	15-24	0	white	2	yes
4	110	1	15-24	1	black	0	no
5	111	1	15-24	1	black	1	mild
6	112	1	15-24	1	black	2	yes
7	200	2	25-34	0	white	0	no
8	201	2	25-34	0	white	1	mild
9	202	2	25-34	0	white	2	yes
10	210	2	25-34	1	black	0	no
11	211	2	25-34	1	black	1	mild
12	212	2	25-34	1	black	2	yes
13	300	3	35-44	0	white	0	no
14	301	3	35-44	0	white	1	mild
15	302	3	35-44	0	white	2	yes
16	310	3	35-44	1	black	0	no
17	311	3	35-44	1	black	1	mild
18	312	3	35-44	1	black	2	yes
19	400	4	45-54	0	white	0	no
20	401	4	45-54	0	white	1	mild
21	402	4	45-54	0	white	2	yes
22	410	4	45-54	1	black	0	no
23	411	4	45-54	1	black	1	mild
24	412	4	45-54	1	black	2	yes
25	500	5	55-64	0	white	0	no
26	501	5	55-64	0	white	1	mild
27	502	5	55-64	0	white	2	yes
28	510	5	55-64	1	black	0	no
29	511	5	55-64	1	black	1	mild
30	512	5	55-64	1	black	2	yes
31	600	6	65-74	0	white	0	no
32	601	6	65-74	0	white	1	mild
33	602	6	65-74	0	white	2	yes
34	610	6	65-74	1	black	0	no
35	611	6	65-74	1	black	1	mild
36	612	6	65-74	1	black	2	yes
37	700	7	over 74	0	white	0	no
38	701	7	over 74	0	white	1	mild
39	702	7	over 74	0	white	2	yes
40	710	7	over 74	1	black	0	no
41	711	7	over 74	1	black	1	mild
42	712	7	over 74	1	black	2	yes

Table 4.3: Variables for Symbolic Medical Income Data

	Variable
$X_{(1)}$	Income
$X_{(2)}$	Glucose
$X_{(3)}$	Cholesterol
$X_{(4)}$	Hemoglobin
$X_{(5)}$	Hematocrit
$X_{(6)}$	Redblood
$X_{(7)}$	Whiteblood

Table 4.4: Histograms of Income for the First Five Observations

i	**Label**	**Frequency Histogram for Income**
1	100	$\{[6500, 10500), 0.022; [10500, 14500), 0.377; [14500, 18500), 0.412;$ $[18500, 22500), 0.166; [22500, 26500], 0.024\}$
2	101	$\{[6500, 10500), 0.007; [10500, 14500), 0.344; [14500, 18500), 0.437;$ $[18500, 22500), 0.181; [22500, 26500], 0.030\}$
3	102	$\{[6500, 10500), 0.008; [10500, 14500), 0.453; [14500, 18500), 0.331;$ $[18500, 22500), 0.176; [22500, 26500], 0.033\}$
4	110	$\{[2500, 6500), 0.047; [6500, 10500), 0.375; [10500, 14500), 0.428;$ $[14500, 18500), 0.111; [18500, 22500), 0.015; [22500, 26500), 0.012;$ $[26500, 30500), 0.006; [30500, 34500], 0.006\}$
5	111	$\{[2500, 6500), 0.050; [6500, 10500), 0.450; [10500, 14500), 0.300;$ $[14500, 18500], 0.200\}$
⋮	⋮	⋮

Table 4.5: Histograms of Glucose for the First Five Observations

i	**Label**	**Frequency Histogram for Glucose**
1	100	$\{[51, 63), 0.005; [63, 75), 0.034; [75, 87), 0.193; [87, 99), 0.403; [99, 111), 0.239;$ $[111, 123), 0.115; [123, 135), 0.009; [135, 147], 0.001\}$
2	101	$\{[96.25, 103.75), 0.211; [103.75, 111.25), 0.381; [111.25, 118.75), 0.304;$ $[118.75, 126.25), 0.096; [126.25, 133.75), 0.004; [133.75, 141.25], 0.004\}$
3	102	$\{[95, 105), 0.233; [105, 115), 0.331; [115, 125), 0.192; [125, 135), 0.192;$ $[135, 145], 0.053; [1, 2), 0.000\}$
4	110	$\{[63, 75), 0.041; [75, 87), 0.161; [87, 99), 0.349; [99, 111), 0.276; [111, 123), 0.170;$ $[123, 135], 0.003\}$
5	111	$\{[96.25, 103.75), 0.225; [103.75, 111.25), 0.300; [111.25, 118.75), 0.200;$ $[118.75, 126.25], 0.275; [1, 2), 0.000; [1, 2), 0.000\}$
⋮	⋮	⋮

Table 4.6: Histograms of Cholesterol for the First Five Observations

i	Label	Frequency Histogram for Cholesterol
1	100	$\{[62.5, 87.5), 0.007; [87.5, 112.5), 0.058; [112.5, 137.5), 0.244; [137.5, 162.5), 0.362; [162.5, 187.5), 0.252; [187.5, 212.5), 0.070; [212.5, 237.5], 0.007\}$
2	101	$\{[62.5, 87.5), 0.004; [87.5, 112.5), 0.056; [112.5, 137.5), 0.244; [137.5, 162.5), 0.422; [162.5, 187.5), 0.204; [187.5, 212.5), 0.063; [212.5, 237.5], 0.007\}$
3	102	$\{[62.5, 87.5), 0.004; [87.5, 112.5), 0.069; [112.5, 137.5), 0.229; [137.5, 162.5), 0.367; [162.5, 187.5), 0.253; [187.5, 212.5), 0.057; [212.5, 237.5), 0.016; [237.5, 262.5], 0.004\}$
4	110	$\{[87.5, 112.5), 0.021; [112.5, 137.5), 0.144; [137.5, 162.5), 0.370; [162.5, 187.5), 0.340; [187.5, 212.5), 0.120; [212.5, 237.5), 0.003; [237.5, 262.5], 0.003\}$
5	111	$\{[112.5, 137.5), 0.125; [137.5, 162.5), 0.375; [162.5, 187.5), 0.400; [187.5, 212.5], 0.100\}$
⋮	⋮	⋮

Table 4.7: Histograms of Hemoglobin for the First Five Observations

i	Label	Frequency Histogram for Hemoglobin
1	100	$\{[12.25, 12.75), 0.009; [12.75, 13.25), 0.100; [13.25, 13.75), 0.357; [13.75, 14.25), 0.375; [14.25, 14.75), 0.137473831; [14.75, 15.25], 0.022\}$
2	101	$\{[12.25, 12.75), 0.007; [12.75, 13.25), 0.093; [13.25, 13.75), 0.296; [13.75, 14.25), 0.430; [14.25, 14.75), 0.156; [14.75, 15.25), 0.015; [15.25, 15.75], 0.004\}$
3	102	$\{[12.25, 12.75), 0.008; [12.75, 13.25), 0.090; [13.25, 13.75), 0.388; [13.75, 14.25), 0.363; [14.25, 14.75), 0.135; [14.75, 15.25], 0.016\}$
4	110	$\{[12.25, 12.75), 0.009; [12.75, 13.25), 0.132; [13.25, 13.75), 0.346; [13.75, 14.25), 0.375; [14.25, 14.75), 0.123; [14.75, 15.25], 0.015\}$
5	111	$\{[12.75, 13.25), 0.100; [13.25, 13.75), 0.300; [13.75, 14.25), 0.475; [14.25, 14.75), 0.100; [14.75, 15.25], 0.025\}$
⋮	⋮	⋮

Table 4.8: Histograms of Hematocrit for the First Five Observations

i	Label	Frequency Histogram for Hematocrit
1	100	$\{[33.75, 36.25), 0.003; [36.25, 38.75), 0.099; [38.75, 41.25), 0.370; [41.25, 43.75), 0.392; [43.75, 46.25), 0.127; [46.25, 48.75], 0.009\}$
2	101	$\{[36.25, 38.75), 0.096; [38.75, 41.25), 0.396; [41.25, 43.75), 0.385; [43.75, 46.25], 0.122\}$
3	102	$\{[36.25, 38.75), 0.127; [38.75, 41.25), 0.327; [41.25, 43.75), 0.416; [43.75, 46.25), 0.127; [46.25, 48.75], 0.004\}$
4	110	$\{[36.25, 38.75), 0.117; [38.75, 41.25), 0.396; [41.25, 43.75), 0.372; [43.75, 46.25), 0.103; [46.25, 48.75], 0.012\}$
5	111	$\{[36.25, 38.75), 0.050; [38.75, 41.25), 0.300; [41.25, 43.75), 0.575; [43.75, 46.25], 0.075\}$
⋮	⋮	⋮

Table 4.9: Histograms of Red Blood for the First Five Observations

i	Label	Frequency Histogram for Red Blood
1	100	$\{[3.4, 3.8), 0.003; [3.8, 4.2), 0.080; [4.2, 4.6), 0.372; [4.6, 5), 0.409; [5, 5.4), 0.125; [5.4, 5.8), 0.010; [5.8, 6.2], 0.001\}$
2	101	$\{[3.4, 3.8), 0.011; [3.8, 4.2), 0.093; [4.2, 4.6), 0.326; [4.6, 5), 0.400; [5, 5.4), 0.156; [5.4, 5.8], 0.015\}$
3	102	$\{[3.4, 3.8), 0.004; [3.8, 4.2), 0.069; [4.2, 4.6), 0.347; [4.6, 5), 0.441; [5, 5.4), 0.135; [5.4, 5.8], 0.004\}$
4	110	$\{[3.8, 4.2), 0.097; [4.2, 4.6), 0.314; [4.6, 5), 0.460; [5, 5.4), 0.120; [5.4, 5.8], 0.009\}$
5	111	$\{[3.8, 4.2), 0.075; [4.2, 4.6), 0.375; [4.6, 5), 0.450; [5, 5.4), 0.075; [5.4, 5.8], 0.025\}$
⋮	⋮	⋮

Table 4.10: Histograms of White Blood for the First Five Observations

i	Label	Frequency Histogram for White Blood
1	100	$\{[1.5, 3), 0.004; [3, 4.5), 0.043; [4.5, 6), 0.134; [6, 7.5), 0.290; [7.5, 9), 0.307; [9, 10.5), 0.165; [10.5, 12), 0.049; [12, 13.5], 0.008\}$
2	101	$\{[1.5, 3), 0.007; [3, 4.5), 0.052; [4.5, 6), 0.137; [6, 7.5), 0.293; [7.5, 9), 0.296; [9, 10.5), 0.170; [10.5, 12), 0.041; [12, 13.5], 0.004\}$
3	102	$\{[1.5, 3), 0.004; [3, 4.5), 0.053; [4.5, 6), 0.151; [6, 7.5), 0.265; [7.5, 9), 0.269; [9, 10.5), 0.184; [10.5, 12], 0.073\}$
4	110	$\{[1.5, 3), 0.003; [3, 4.5), 0.044; [4.5, 6), 0.152; [6, 7.5), 0.261; [7.5, 9), 0.331; [9, 10.5), 0.147; [10.5, 12), 0.053; [12, 13.5], 0.009\}$
5	111	$\{[4.5, 6), 0.075; [6, 7.5), 0.375; [7.5, 9), 0.325; [9, 10.5], 0.225\}$
⋮	⋮	⋮

symbolic variance-covariance matrix for this dataset as defined in Equation (4.2) is given by

$$\boldsymbol{S} = \begin{bmatrix} 46452198 & 1202 & 33432 & -1331 & -4264 & -436 & -3407 \\ & 206.34 & 13.94 & -0.60 & -2.64 & -0.17 & 1.59 \\ & & 1077.29 & -9.61 & -33.73 & -8.85 & -3.25 \\ & & & 0.60 & 1.70 & 0.11 & 0.35 \\ & & & & 8.13 & 0.38 & 1.28 \\ & & & & & 0.14 & 0.04 \\ & & & & & & 3.64 \end{bmatrix}. \tag{4.19}$$

The symbolic correlation matrix as defined in Equation (4.4) is

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.0123 & 0.1495 & -0.2527 & -0.2194 & -0.1681 & -0.2618 \\ & 1 & 0.0296 & -0.0542 & -0.0645 & -0.0306 & 0.0581 \\ & & 1 & -0.3788 & -0.3603 & -0.7085 & -0.0519 \\ & & & 1 & 0.7721 & 0.3707 & 0.2392 \\ & & & & 1 & 0.3516 & 0.2358 \\ & & & & & 1 & 0.0511 \\ & & & & & & 1 \end{bmatrix}. \tag{4.20}$$

The symbolic correlation matrix in Equation (4.20) reveals the following observations: the correlation coefficients between Glucose ($X_{(2)}$) and other medical variables ($X_{(3)} - X_{(7)}$) are less than 0.1. That is, there is essentially no correlation between Glucose ($X_{(2)}$) and the concentration of red blood and white blood cells; White blood count ($X_{(7)}$) is not correlated with Glucose ($X_{(2)}$), Cholesterol ($X_{(3)}$) and Red blood count ($X_{(6)}$) ; Cholesterol($X_{(3)}$) is negatively correlated to Red blood count ($X_{(6)}$) with a coefficient of -0.7085; and Hemoglobin ($X_{(4)}$) is highly correlated with Hematocrit ($X_{(5)}$) with a coefficient of 0.7721. The negative correlation between Cholesterol and Red blood count makes sense because cholesterol is a lipid circulating in the blood stream. When there is more cholesterol in the blood, there is less room for the red blood cells in the blood stream and vice versa. The high correlation between Hemoglobin and Hematocrit also makes sense because they are both measurements

Table 4.11: Coefficients and Variance Proportion of Principal Components for Medical Income Data Using All Variables

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Income	-0.2609	0.4074	-0.0417	0.6622	0.5693	-0.0248	0.0336
Glucose	-0.0464	-0.1641	0.9151	0.3087	-0.1954	-0.0015	-0.0098
Cholesterol	-0.4429	-0.4359	-0.2059	0.2007	-0.1712	-0.7083	0.0129
Hemoglobin	0.5025	-0.1560	-0.1643	0.3941	-0.1709	-0.0044	0.7153
Hematocrit	0.4920	-0.1550	-0.1884	0.4365	-0.1445	-0.0116	-0.6978
Red blood	0.4414	0.4313	0.2097	-0.2341	0.1510	-0.7054	-0.0071
White blood	0.2103	-0.6186	0.1098	-0.1592	0.7319	-0.0046	0.0050
Proportion of Variance	0.3788	0.1764	0.1451	0.1255	0.1002	0.0416	0.0324
Cumulative Proportion	0.3788	0.5553	0.7004	0.8259	0.9260	0.9676	1.0000

of concentration of red blood cells. Now that we have determined the covariance structure of the dataset, we can proceed with the principal component analysis. The symbolic principal component analysis results for the Medical Income dataset are presented in the following subsection.

4.4.2 Analysis results

PCA using all variables

We first apply our proposed principal component method to all seven variables in the histogram-valued dataset of Tables 4.4 through 4.10. Table 4.1 shows that standard measurements for the variables vary widely in both scales and of measurement units. In datasets with such high variability between the variable measurements, it is more appropriate to perform principal component analysis based on the correlation matrix than based on the covariance matrix. We apply the proposed principal component method to the correlation matrix of Equation (4.20). Table 4.11 gives the coefficients and the proportion of variation explained by the principal components of the Medical Income data when all seven variables are included.

We can see from Table 4.11 that the first principal component is composed of Hemoglobin with a coefficient of 0.5025, Hematocrit with a coeficient of 0.4920, and Red blood with a coefficient of 0.4414. Since Cholesterol is negatively correlated to Red blood counts, it makes sense that Cholesterol contributes a negative amount to the first principal component with a coefficient of -0.4429. Therefore, the first principal component describes the overall red blood concentration and it accounts for 37.88% of the total variation of the data. The next three principal components, each accounts for about $18\%, 15\%$, and 13% of the overall variation. The first four principal components altogether contribute almost 83% of the total variation in the data. A closer look at the coefficients of principal components two, three, and four shows that they are composed of mainly White blood with a coefficient of -0.6186, Glucose with a coefficient of 0.9151, and Income with a coefficient of 0.6622, respectively. Since each of these principal components is mainly composed of one variable and each principal component explains an approximately equal proportion of data variation, it means the variables that make up principal component two, three, and four are relatively uncorrelated with other variables. That is, White blood, Glucose, and Income have very low correlation with other variables.

To help identify observations in each plot, labels for observations are displayed in a corner of the plot where space allows. In each plot, colors of the polygons representing the observations are based on one of the following characteristics of the observations: age, race, and diabetes status. These characteristics are coded in an observation's label as shown in Table 4.2. A specific characteristic for a plot is chosen to illustrate a significant feature of the principal components revealed in the plot. In all plots, a label has the same color as its corresponding observation.

Plots of the observations along the first principal component (PC1) and the second principal component (PC2) are shown in Figure 4.3. In Figure 4.3, color represents age (the first digit of the labels) of the observation. That is, the polygons representing groups whose age fall in the age range of 15-24 is displayed in black, the polygons representing groups of

patients in the age range of 25-34 is colored in red, and so on. See Figure 4.3 for details. At first look, the PC1 × PC2 plot suggests the observations form one cluster. All observations cover a wide range of values along the PC1-axis. A closer look shows a pattern of slight degeneration of red blood concentration for older groups of patients. More specifically, the black polygons representing the youngest group of patients form a cluster and this cluster has the largest first principal component values. As age increases, the polygons associated with these observations slightly move toward the smaller end of the PC1-axis. Examining the observations along the PC2 axis shows the range of white blood concentration are about the same for all groups. There exists no special feature along the second principal component.

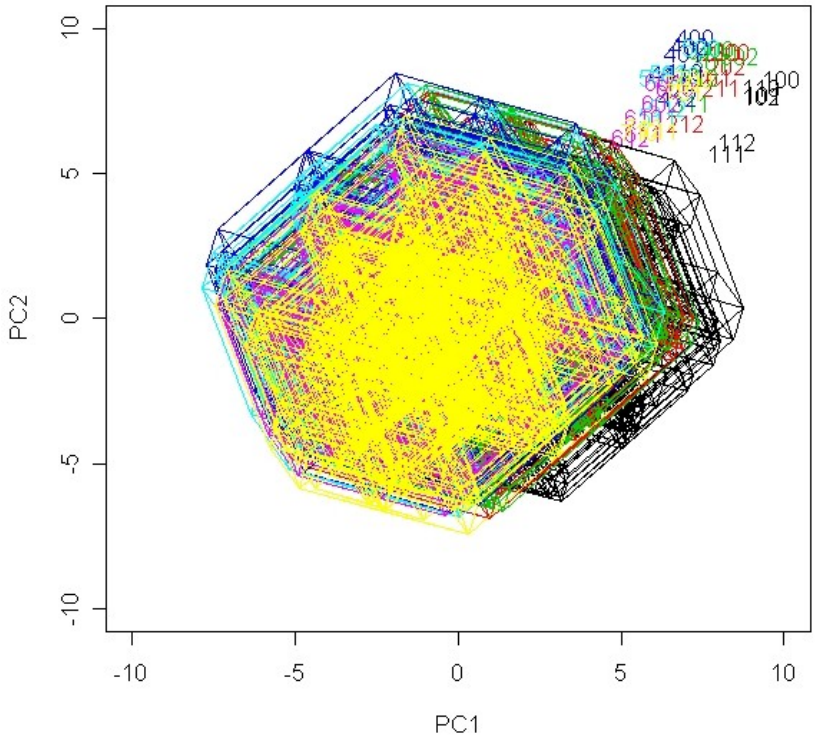

Figure 4.3: Plot of PC1 × PC2 for Medical Income Data Using All Variables (Color Represents Age)

Figure 4.4 shows a plot of principal component one versus principal component three (PC1 × PC3) and Figures 4.5 shows a plot of PC2 × PC3. In both of these plots, a polygon is colored according to diabetes status (the last digit of the labels) of the group it represents. That is, the polygons representing groups of patients who do not have diabetes are colored

in black, the polygons representing groups with mild diabetes are in red, and the polygons representing groups with severe diabetes are in green. The third principal component values for groups with no diabetes (whose labels end with zero) ranges from -8 to 6 whereas PC3 values for groups with diabetes (whose labels end with one or two) range from -3 to 7. The black polygons covers a wide range of values for the third principal component which is mostly composed of Glucose. That means for patients who do not have diabetes, some have a low level of glucose and some have a high level of glucose. The red and the green polygons stand out in a cluster with high PC3 values whose lowest value is much higher than the lowest PC3 values of the group with no diabetes. That is, patients with diabetes consistently have high level of glucose in their blood. The third principal component confirms the relationship between Glucose and Diabetes.

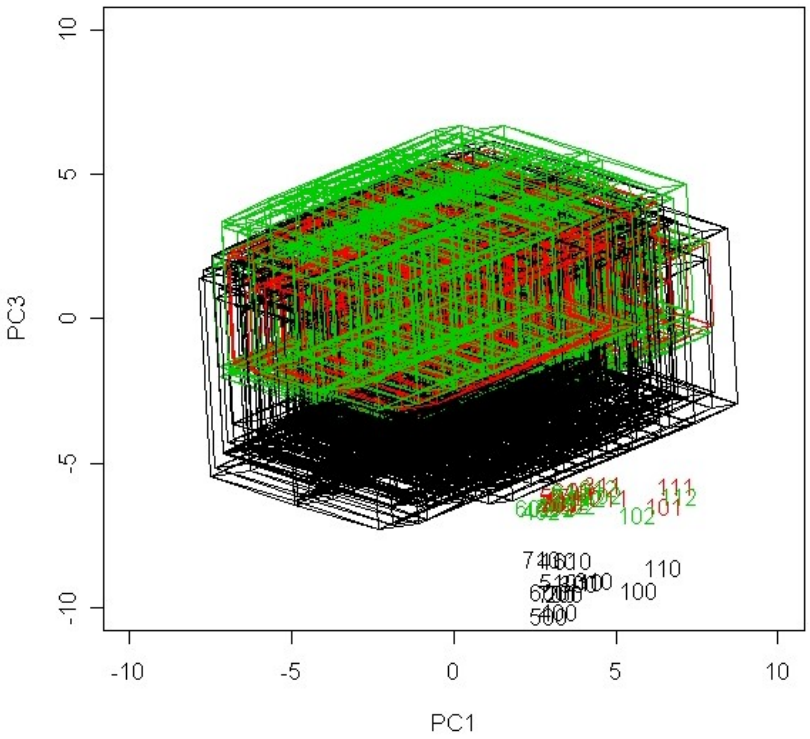

Figure 4.4: Plot of PC1 × PC3 for Medical Income Data Using All Variables (Color Represents Diabetes)

The correlations between a variable $X_{(j)}$ and a principal component $Y_{(k)}$ for $j, k = 1, 2, \ldots, p$, as defined in Equation (2.25) are shown in Table 4.12. The coefficients in Table

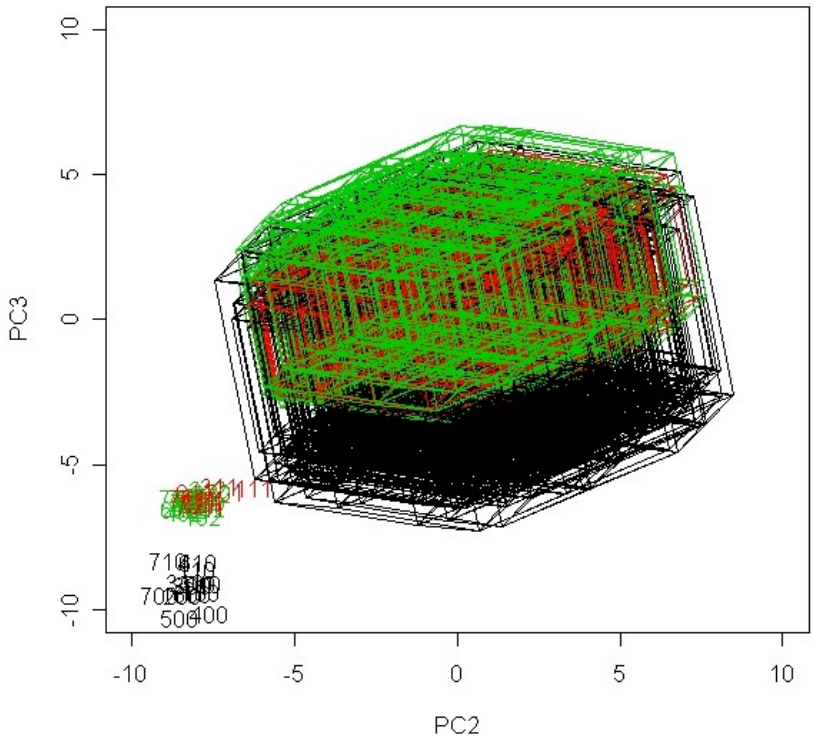

Figure 4.5: Plot of PC2 × PC3 for Medical Income Data Using All Variables (Color Represents Diabetes)

Table 4.12: Correlation between Principal Components and Random Variables of Medical Income Data Using All Variables

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Income	-0.4249	0.4527	-0.0420	0.6206	0.4767	-0.0134	0.0160
Glucose	-0.0755	-0.1823	0.9223	0.2893	-0.1636	-0.0008	-0.0046
Cholesterol	-0.7213	-0.4845	-0.2075	0.1881	-0.1434	-0.3821	0.0061
Hemoglobin	0.8183	-0.1734	-0.1656	0.3694	-0.1431	-0.0024	0.3406
Hematocrit	0.8011	-0.1723	-0.1899	0.4091	-0.1210	-0.0063	-0.3323
Red blood	0.7188	0.4793	0.2113	-0.2194	0.1264	-0.3806	-0.0034
White blood	0.3425	-0.6875	0.1106	-0.1492	0.6128	-0.0025	0.0024

4.12 show the correlation between variable $X_{(j)}$ and principal component $Y_{(k)}$ in the absence of all other variables. The first principal component is highly correlated with Cholesterol, Hemoglobin, Hematocrit, and Red blood with correlation coefficients of -0.7213, 0.8183, 0.8011, and 0.7188, respectively. The second principal component has the highest correlation with White blood whose coefficient is -0.6875. The third principal component is highly correlated with Glucose with a coefficient of 0.9223. These correlation coefficients concur with the conclusions drawn from the coefficients of the principal components as shown in Table 4.11.

After exploring the plots of Figures 4.3, 4.4, and 4.5 as well as the coefficients of Table 4.11 and Table 4.12, if further analysis of the principal components becomes necessary, histogram-valued principal components can be computed. Table 4.13 shows histograms of the first principal component for observations one through five. The number of subintervals for the histograms representing the principal components can be specified in the algorithm. In this example, the number of subinterval was set at eight; however, some of the lowest and highest subintervals have relative frequency less than 0.001 and they are excluded from the histograms shown in Table 4.13. Therefore, the resulting histograms have less than five subintervals.

PCA based on Income, Glucose, Hemoglobin, and Hematocrit

Besides the analysis based on all seven variables, we also apply the proposed PCA method to various combinations of subsets of variables from the Medical Income dataset. Plots for some of these examples are included in the appendix at the end of this chapter. As can be seen in the plots of Figures 4.3 through 4.5, the internal structure of a histogram-valued observation can be complex when the number of variables, p, is high. In the following subsection, we illustrate the proposed method using a subset of four variables from the Medical Income dataset. The internal complexity of a histogram-valued observation is reduced significantly

Table 4.13: Histograms for the First Principal Component of Medical Income Data Using All Variables

i	**Label**	**Frequency Histogram for the First Principal Component**
1	100	$\{[-0.675, 0.893), 0.046; [0.893, 2.461), 0.489; [2.461, 4.029), 0.433; [4.029, 5.598], 0.032\}$
2	101	$\{[-1.939, -0.516), 0.001; [-0.516, 0.906), 0.051; [0.906, 2.328), 0.432; [2.328, 3.751), 0.458; [3.751, 5.173), 0.058; [5.173, 6.595), 0.001\}$
3	102	$\{[-0.792, 0.667), 0.032; [0.667, 2.126), 0.379; [2.126, 3.585), 0.511; [3.585, 5.044), 0.077; [5.044, 6.504), 0.001\}$
4	110	$\{[-0.629, 0.817), 0.039; [0.817, 2.263), 0.417; [2.263, 3.709), 0.480; [3.709, 5.155), 0.062; [5.155, 6.601), 0.001\}$
5	111	$\{[-0.501, 0.548), 0.008; [0.548, 1.597), 0.139; [1.597, 2.646), 0.454; [2.646, 3.694), 0.340; [3.694, 4.743), 0.057; [4.743, 5.792), 0.002\}$
⋮	⋮	⋮

Table 4.14: Coefficients and Variance Proportion of Principal Components of Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit

Variable	PC1	PC2	PC3	PC4
Income	-0.3461	-0.1714	0.9218	-0.0335
Glucose	-0.0914	0.9845	0.1491	0.0092
Hemoglobin	0.6634	0.0335	0.2295	-0.7114
Hematocrit	0.6571	0.0128	0.2746	0.7019
Proportion of Variance	0.4760	0.2488	0.2184	0.0568
Cumulative Proportion	0.4760	0.7248	0.9432	1.0000

from $p = 7$ to $p = 4$. An additional example with a simpler structure gives a clearer picture of our methodology.

In the following example we compute the principal components using Income, Glucose, Hemoglobin, and Hematocrit. The coefficients and the proportion of variance explained by the principal components based on the correlation structure of these variables are shown in Table 4.14.

Hemoglobin and Hematocrit are the major contributors of the first principal component with coefficients of 0.6634 and 0.6571. The size of their contribution is almost identical. The first principal component explains almost 50% of the total variation in the data. The second principal component is made up of mostly Glucose and it explains about 25% of the total variation. The third principal component is composed mostly of Income and contributes another 22% to the total variation. Together, the first three principal components explain 94% of the data variation leaving only 6% to the last principal component. The tiny percentage of the variance contributed by the fourth principal component indicates that one of the variables is almost completely dependent on the other variables. That is, the space spanned by Income, Glucose, Hemoglobin, and Hematocrit has dimension of three. Based on the correlation matrix of Equation (4.20), we know that Hemoglobin and Hematocrit are highly correlated.

The principal components can further be explored in Figures 4.6 through 4.8. Plots of the observations along the first and the second principal components are shown in Figure 4.6. The polygons in this plot are colored based on age group. Visually, the first principal component based on Income, Glucose, Hemoglobin, and Hematocrit gives a general pattern similar to the first principal component based on all variables. We can see that the youngest groups of patients, represented by the black polygons, have the highest values of PC1. Their PC1 values range from about -1 to 6. This pattern also appears along the PC1-axis of Figure 4.7. Polygons representing other age groups cluster toward the smaller end of the PC1-axis. Their values range from approxiamtely -6 to less than 4.5.

Another feature of the second principal component is clearly visible in Figure 4.8. Observations are plotted along the second and third principal components in Figure 4.8. The polygons in this plot are colored based on diabetes status with black represents groups with no diabetes, red for groups with mild diabetes and green for groups with severe diabetes. Values along the PC2-axis range from -6 to 4 for the black polygons, from -2 to 4 for the red polygons, and from -2 to 5 for the green polygons. Since Glucose is the primary contributor

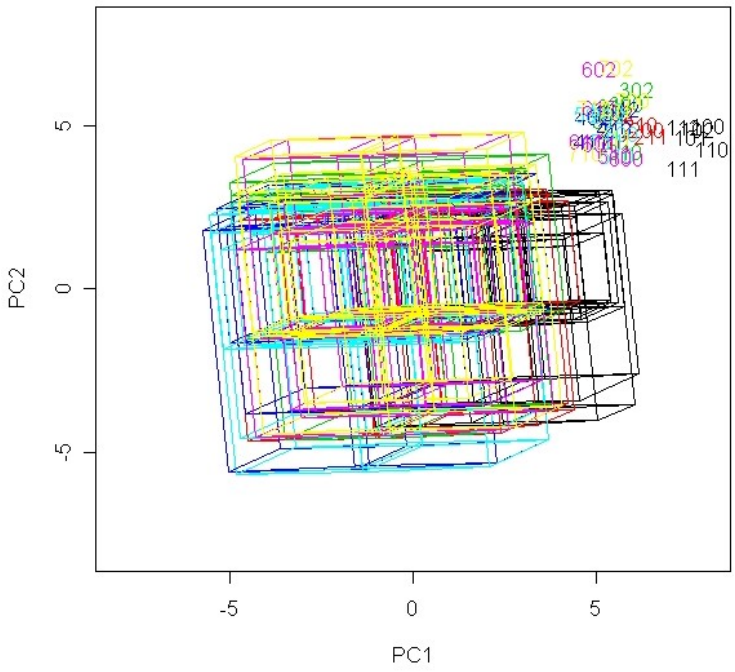

Figure 4.6: Plot of PC1 × PC2 for Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit (Color Represents Age)

to the second principal component, we can say that groups of patients with diabetes have a higher level of glucose in their blood. Groups of patients with severe diabetes have especially high levels of blood glucose compared to the other groups. Again, this confirms the correlationship between glucose and diabetes.

The general pattern for the third principal component is more subtle than for PC1 and PC2. Close examination of Figure 4.7 along the PC3-axis shows that groups of patients in the 45 to 64 age range have the highest PC3 values. The older groups of over 65 and younger groups have lower PC3 values. Further inspection of the polygons within the same age group shows what groups of black patients (whose labels have the second digit of zero) have PC3 values lower than their white counterpart. Since the third principal component is composed of mostly Income, we can conclude that the income is somewhat related to

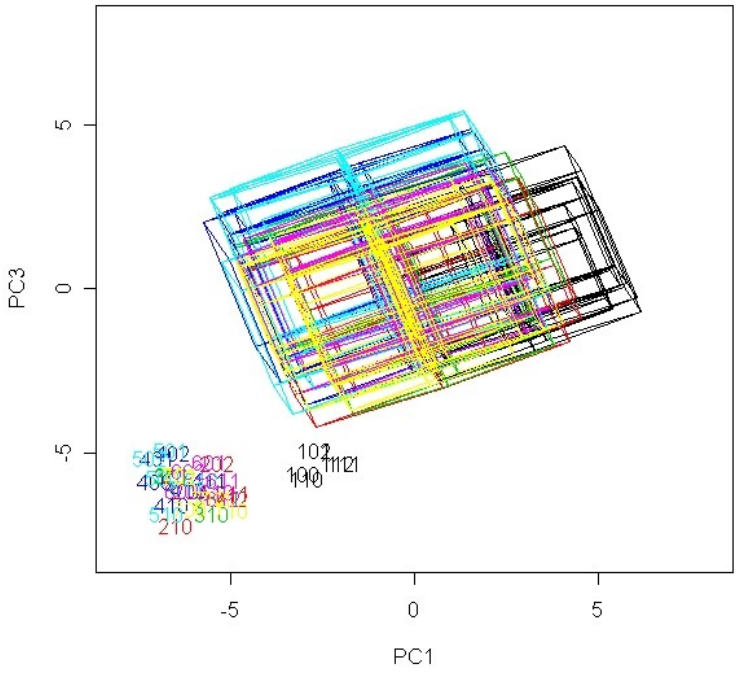

Figure 4.7: Plot of PC1 × PC3 for Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit (Color Represents Age)

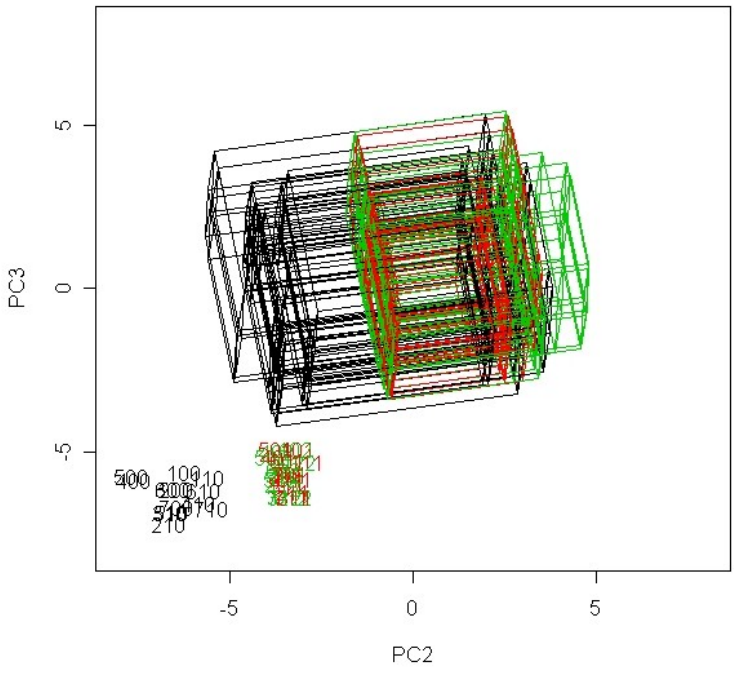

Figure 4.8: Plot of PC2 × PC3 for Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit (Color Represents Diabetes)

Table 4.15: Correlation between Principal Components and Random Variables of Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit

Variable	PC1	PC2	PC3	PC4
Income	-0.4776	-0.1710	0.8616	-0.0160
Glucose	-0.1261	0.9822	0.1394	0.0044
Hemoglobin	0.9154	0.0334	0.2145	-0.3390
Hematocrit	0.9067	0.0128	0.2567	0.3345

both age and race. More specifically, groups of middle-aged white patients have the highest income level. The correlation between a variable $X_{(j)}$ and a principal component $Y_{(k)}$ in the absence of other variables as defined in Equation (2.25) are shown in Table 4.15. The correlation coefficients of the first principal component show a very high correlation between the first principal component and Hemoglobin with a coefficient of 0.9154. The first principal component is also highly correlated to Hematocrit with a coefficient of 0.9067. Table 4.15 also indicates an almost perfect correlationship between the second principal component and Glucose with a coefficient of 0.9822. Income is highly correlated with the third principal component whose coefficient is 0.8616. Again, these correlations concur with the results drawn from the coefficients of the principal components shown in Table 4.14.

If further analysis of the principal components is needed where numerical values are necessary, histogram-valued principal components can be computed. Histograms for the first three principal components for a subset of observations are shown in Tables 4.16, 4.17, and 4.18.

4.4.3 Comparison of symbolic PCA and classical PCA

Without a method to analyze symbolic data, a midpoint would be used to represent all values in a symbolic variable. When only one point represents many points, information including variability and distribution of the points is lost. Principal component analysis

Table 4.16: Histograms for the First Principal Component of Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit

i	Label	Frequency Histogram for the First Principal Component
1	100	$\{[-0.662, 0.295), 0.003; [0.295, 1.252), 0.071; [1.252, 2.210), 0.354;$ $[2.210, 3.167), 0.429; [3.167, 4.124), 0.135; [4.124, 5.082], 0.009\}$
2	101	$\{[-0.180, 0.646), 0.013; [0.646, 1.472), 0.129; [1.472, 2.299), 0.375;$ $[2.299, 3.125), 0.359; [3.125, 3.952), 0.114; [3.952, 4.778], 0.009\}$
3	102	$\{[-0.182, 0.667), 0.019; [0.667, 1.515), 0.161; [1.515, 2.364), 0.393;$ $[2.364, 3.213), 0.332; [3.213, 4.062), 0.089; [4.062, 4.910], 0.006\}$
4	110	$\{[-0.431, 0.511), 0.004; [0.511, 1.454), 0.087; [1.454, 2.396), 0.370;$ $[2.396, 3.339), 0.407; [3.339, 4.281), 0.123; [4.281, 5.223], 0.009\}$
5	111	$\{[-0.076, 0.606), 0.002; [0.606, 1.288), 0.029; [1.288, 1.969), 0.155;$ $[1.969, 2.651), 0.357; [2.651, 3.333), 0.330; [3.333, 4.015), 0.114;$ $[4.015, 4.697], 0.014\}$
6	112	$\{[-0.398, 0.325), 0.001; [0.325, 1.048), 0.027; [1.048, 1.771), 0.148;$ $[1.771, 2.494), 0.317; [2.494, 3.217), 0.314; [3.217, 3.940), 0.158;$ $[3.940, 4.663), 0.033; [4.663, 5.386], 0.001\}$
7	200	$\{[-3.555, -2.419), 0.004; [-2.419, -1.282), 0.116; [-1.282, -0.145), 0.481;$ $[-0.145, 0.991), 0.354; [0.991, 2.128), 0.045; [2.128, 3.265], 0.001\}$
8	201	$\{[-3.631, -2.607), 0.002; [-2.607, -1.583), 0.071; [-1.583, -0.559), 0.373;$ $[-0.559, 0.465), 0.432; [0.465, 1.489), 0.116; [1.489, 2.513], 0.006\}$
9	202	$\{[-3.738, -2.827), 0.001; [-2.827, -1.917), 0.026; [-1.917, -1.007), 0.204;$ $[-1.007, -0.096), 0.429; [-0.096, 0.814), 0.281; [0.814, 1.725), 0.056;$ $[1.725, 2.635], 0.003\}$
10	210	$\{[-2.766, -1.600), 0.030; [-1.600, -0.433), 0.318; [-0.433, 0.733), 0.520;$ $[0.733, 1.899), 0.128; [1.899, 3.065], 0.003\}$
⋮	⋮	⋮
37	700	$\{[-3.669, -2.569), 0.001; [-2.569, -1.470), 0.063; [-1.470, -0.370), 0.396;$ $[-0.370, 0.730), 0.446; [0.730, 1.829), 0.092; [1.829, 2.929], 0.003\}$
38	701	$\{[-3.715, -2.679), 0.001; [-2.679, -1.643), 0.048; [-1.643, -0.607), 0.338;$ $[-0.607, 0.429), 0.460; [0.429, 1.465), 0.142; [1.465, 2.501], 0.011\}$
39	702	$\{[-3.818, -2.766), 0.002; [-2.766, -1.714), 0.055; [-1.714, -0.662), 0.347;$ $[-0.662, 0.389), 0.454; [0.389, 1.441), 0.133; [1.441, 2.493], 0.008\}$
40	710	$\{[-2.641, -1.875), 0.005; [-1.875, -1.109), 0.063; [-1.109, -0.343), 0.258;$ $[-0.343, 0.423), 0.394; [0.423, 1.188), 0.232; [1.188, 1.954), 0.046;$ $[1.954, 2.720], 0.002\}$
41	711	$\{[-2.755, -1.859), 0.011; [-1.859, -0.963), 0.119; [-0.963, -0.068), 0.355;$ $[-0.068, 0.828), 0.360; [0.828, 1.723), 0.136; [1.723, 2.619), 0.019;$ $[2.619, 3.514], 0.001\}$
42	712	$\{[-3.080, -2.219), 0.003; [-2.219, -1.358), 0.052; [-1.358, -0.497), 0.254;$ $[-0.497, 0.364), 0.413; [0.364, 1.224), 0.234; [1.224, 2.085), 0.042;$ $[2.085, 2.946], 0.002\}$

Table 4.17: Histograms for the Second Principal Component of Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit

i	Label	Frequency Histogram for the Second Principal Component
1	100	$\{[-2.610,-1.855),0.104;[-1.855,-1.100),0.262;[-1.100,-0.346),0.332;$ $[-0.346,0.409),0.198;[0.409,1.164),0.092;[1.164,1.918),0.010;$ $[1.918,2.673],0.002\}$
2	101	$\{[-1.252,-0.605),0.071;[-0.605,0.041),0.354;[0.041,0.688),0.406;$ $[0.688,1.335),0.157;[1.335,1.982),0.010;[1.982,2.629],0.002\}$
3	102	$\{[-1.247,-0.582),0.091;[-0.582,0.083),0.262;[0.083,0.747),0.275;$ $[0.747,1.412),0.187;[1.412,2.077),0.147;[2.077,2.742],0.038\}$
4	110	$\{[-2.785,-1.889),0.080;[-1.889,-0.993),0.247;[-0.993,-0.096),0.347;$ $[-0.096,0.800),0.243;[0.800,1.696),0.081;[1.696,2.592],0.001\}$
5	111	$\{[-1.246,-0.724),0.008;[-0.724,-0.201),0.184;[-0.201,0.322),0.285;$ $[0.322,0.844),0.227;[0.844,1.367),0.250;[1.367,1.889],0.046\}$
6	112	$\{[-1.072,-0.456),0.086;[-0.456,0.160),0.222;[0.160,0.777),0.263;$ $[0.777,1.393),0.209;[1.393,2.009),0.180;[2.009,2.625],0.041\}$
7	200	$\{[-3.144,-2.322),0.067;[-2.322,-1.500),0.222;[-1.500,-0.679),0.380;$ $[-0.679,0.143),0.217;[0.143,0.965),0.100;[0.965,1.787),0.013;$ $[1.787,2.609],0.001\}$
8	201	$\{[-1.223,-0.498),0.199;[-0.498,0.228),0.498;[0.228,0.954),0.253;$ $[0.954,1.679),0.046;[1.679,2.405],0.004\}$
9	202	$\{[-1.602,-0.905),0.040;[-0.905,-0.209),0.205;[-0.209,0.488),0.266;$ $[0.488,1.184),0.211;[1.184,1.880),0.239;[1.880,2.577),0.030;$ $[2.577,3.273],0.009\}$
10	210	$\{[-4.226,-3.164),0.007;[-3.164,-2.102),0.079;[-2.102,-1.039),0.367;$ $[-1.039,0.023),0.367;[0.023,1.085),0.168;[1.085,2.147),0.011;$ $[2.147,3.210],0.001\}$
⋮	⋮	⋮
37	700	$\{[-3.685,-2.810),0.003;[-2.810,-1.935),0.036;[-1.935,-1.060),0.189;$ $[-1.060,-0.185),0.302;[-0.185,0.690),0.309;[0.690,1.565),0.136;$ $[1.565,2.440),0.023;[2.440,3.315],0.002\}$
38	701	$\{[-1.184,-0.439),0.117;[-0.439,0.306),0.417;[0.306,1.051),0.353;$ $[1.051,1.796),0.098;[1.796,2.541),0.013;[2.541,3.286],0.002\}$
39	702	$\{[-1.916,-1.145),0.001;[-1.145,-0.374),0.084;[-0.374,0.397),0.213;$ $[0.397,1.168),0.283;[1.168,1.939),0.320;[1.939,2.710),0.088;$ $[2.710,3.481],0.012\}$
40	710	$\{[-3.712,-3.000),0.002;[-3.000,-2.288),0.013;[-2.288,-1.575),0.043;$ $[-1.575,-0.863),0.213;[-0.863,-0.150),0.271;[-0.150,0.562),0.250;$ $[0.562,1.274),0.174;[1.274,1.987],0.034\}$
41	711	$\{[-1.780,-0.986),0.001;[-0.986,-0.193),0.169;[-0.193,0.601),0.455;$ $[0.601,1.395),0.319;[1.395,2.188),0.045;[2.188,2.982],0.012\}$
42	712	$\{[-1.679,-0.828),0.006;[-0.828,0.022),0.103;[0.022,0.873),0.226;$ $[0.873,1.724),0.383;[1.724,2.574),0.240;[2.574,3.425],0.041\}$

Table 4.18: Histograms for the Third Principal Component of Medical Income Data Using Income, Glucose, Hemoglobin, and Hematocrit

i	Label	Frequency Histogram for the Third Principal Component
1	100	$\{[-1.855,-1.100),0.004;[-1.100,-0.346),0.132;[-0.346,0.409),0.455;$ $[0.409,1.164),0.334;[1.164,1.918),0.071;[1.918,2.673],0.004\}$
2	101	$\{[-1.252,-0.605),0.014;[-0.605,0.041),0.203;[0.041,0.688),0.429;$ $[0.688,1.335),0.280;[1.335,1.982),0.068;[1.982,2.629],0.005\}$
3	102	$\{[-1.247,-0.582),0.021;[-0.582,0.083),0.251;[0.083,0.747),0.417;$ $[0.747,1.412),0.240;[1.412,2.077),0.065;[2.077,2.742],0.005\}$
4	110	$\{[-2.785,-1.889),0.002;[-1.889,-0.993),0.116;[-0.993,-0.096),0.530;$ $[-0.096,0.800),0.302;[0.800,1.696),0.036;[1.696,2.592),0.011;$ $[2.592,3.489],0.003\}$
5	111	$\{[-1.769,-1.246),0.015;[-1.246,-0.724),0.130;[-0.724,-0.201),0.329;$ $[-0.201,0.322),0.298;[0.322,0.844),0.179;[0.844,1.367),0.047;$ $[1.367,1.889],0.001\}$
6	112	$\{[-2.305,-1.688),0.002;[-1.688,-1.072),0.038;[-1.072,-0.456),0.216;$ $[-0.456,0.160),0.390;[0.160,0.777),0.259;[0.777,1.393),0.080;$ $[1.393,2.009],0.015\}$
7	200	$\{[-2.322,-1.500),0.013;[-1.500,-0.679),0.197;[-0.679,0.143),0.445;$ $[0.143,0.965),0.293;[0.965,1.787),0.051;[1.787,2.609],0.001\}$
8	201	$\{[-1.949,-1.223),0.018;[-1.223,-0.498),0.194;[-0.498,0.228),0.401;$ $[0.228,0.954),0.302;[0.954,1.679),0.080;[1.679,2.405],0.004\}$
9	202	$\{[-2.298,-1.602),0.001;[-1.602,-0.905),0.044;[-0.905,-0.209),0.264;$ $[-0.209,0.488),0.380;[0.488,1.184),0.239;[1.184,1.880),0.067;$ $[1.880,2.577],0.005\}$
10	210	$\{[-3.164,-2.102),0.022;[-2.102,-1.039),0.392;[-1.039,0.023),0.476;$ $[0.023,1.085),0.103;[1.085,2.147),0.005;[2.147,3.210],0.001\}$
⋮	⋮	⋮
37	700	$\{[-2.810,-1.935),0.001;[-1.935,-1.060),0.073;[-1.060,-0.185),0.444;$ $[-0.185,0.690),0.403;[0.690,1.565),0.076;[1.565,2.440],0.003\}$
38	701	$\{[-1.929,-1.184),0.035;[-1.184,-0.439),0.288;[-0.439,0.306),0.464;$ $[0.306,1.051),0.186;[1.051,1.796),0.025;[1.796,2.541],0.001\}$
39	702	$\{[-1.916,-1.145),0.031;[-1.145,-0.374),0.275;[-0.374,0.397),0.461;$ $[0.397,1.168),0.202;[1.168,1.939),0.030;[1.939,2.710],0.001\}$
40	710	$\{[-3.000,-2.288),0.010;[-2.288,-1.575),0.173;[-1.575,-0.863),0.477;$ $[-0.863,-0.150),0.291;[-0.150,0.562),0.042;[0.562,1.274],0.005\}$
41	711	$\{[-3.367,-2.573),0.001;[-2.573,-1.780),0.073;[-1.780,-0.986),0.443;$ $[-0.986,-0.193),0.381;[-0.193,0.601),0.079;[0.601,1.395),0.020;$ $[1.395,2.188],0.004\}$
42	712	$\{[-3.380,-2.529),0.001;[-2.529,-1.679),0.073;[-1.679,-0.828),0.435;$ $[-0.828,0.022),0.395;[0.022,0.873),0.079;[0.873,1.724),0.014;$ $[1.724,2.574],0.003\}$

using the midpoints only accounts for part of the total variation in the dataset. The variance maximized by the principal components is the variance between observations. These principal components were determined in the absence of the variation within observations as defined in Equation (2.6) through Equation (2.12) of Chapter 2 and detailed in Billard (2007). The classical variance-covariance matrix of the midpoints is given by

$$\boldsymbol{S} = \begin{bmatrix} 26898268 & -1790 & 18636 & -1139 & -3480 & -158 & -172 \\ & 74.57 & 14.62 & -0.41 & -1.19 & -0.21 & 0.11 \\ & & 355.99 & -5.54 & -16.48 & -2.75 & 0.17 \\ & & & 0.32 & 0.96 & 0.05 & 0.02 \\ & & & & 2.89 & 0.14 & 0.05 \\ & & & & & 0.03 & 0.00 \\ & & & & & & 0.02 \end{bmatrix}. \tag{4.21}$$

All variances in Equation (4.21) are much smaller than their counterparts in Equation (4.19). The smaller values of the variances using the midpoints reflect the omission of the internal variation that is inherent in symbolic data.

The classical correlation matrix resulting from the covariance matrix of Equation (4.21) for the Medical Income dataset is

$$\boldsymbol{R} = \begin{bmatrix} 1 & -0.0400 & 0.1905 & -0.3891 & -0.3951 & -0.1834 & -0.2541 \\ & 1 & 0.0897 & -0.0851 & -0.0815 & -0.1436 & 0.0953 \\ & & 1 & -0.5202 & -0.5142 & -0.8770 & 0.0683 \\ & & & 1 & 0.9986 & 0.5098 & 0.2364 \\ & & & & 1 & 0.5059 & 0.2406 \\ & & & & & 1 & -0.0194 \\ & & & & & & 1 \end{bmatrix}. \tag{4.22}$$

Compared to the symbolic correlation matrix of Equation (4.20), the correlation between the midpoints are in general stronger than the correlation between the histograms. In this

example, observations have centers that are close together and the range of values for each observation is large. Therefore, it makes sense that the correlation between the midpoints are stronger than the symbolic correlation because the symbolic correlation accounts for all values belonging to the observations. These values spread out around the midpoints so they weaken the correlation between variables. More particularly, the correlation between the midpoints of Hemoglobin ($X_{(4)}$) and Hematocrit ($X_{(5)}$) is 0.9986 versus 0.7721 in the correlation between histograms of these two variables. Similarly, the correlation between the midpoints of Cholesterol ($X_{(3)}$) and Red blood ($X_{(6)}$) is -0.8770 versus the correlation of -0.7085 between the histograms.

Having seen the differences in the correlation structure of histogram-valued Medical Income data and classical Medical Income data using the midpoints, we next apply classical PCA to the midpoints based on the correlation matrix of Equation (4.22). The coefficients and the proportion of variance explained by the principal components resulting from the classical PCA using the midpoints are shown in Table 4.19. Coefficients too close to zero are left blank in Table 4.19. The coefficients in Table 4.19 show that the general composition of the principal components based on the midpoints is similar to that based on the histograms, with only slight differences in the magnitudes of the coefficients. However, the first principal component using the midpoints accounts for 45% of the total variation compared to 38% in the case of the histograms. The next two principal components also contribute more to the total variation in the data than those in the histgorams results in Subsection 4.4.2. Together the first three principal components account for a total of 79% of the data variation.

Plots of the observations along the first three principal components are shown in Figures 4.9, 4.10, and 4.11. Color schemes for these plots are the same as those for Figures 4.3, 4.4, and 4.5. That is, observations in PC1 $\times$ PC2 plots are colored according to age group and observations in PC1 $\times$ PC3 and PC2 $\times$ PC3 are colored according to diabetes status. The plot of observations along the PC1 and the PC2 axes of Figure 4.9 shows three distinct clusters of observations. One cluster consists of observations which represent the youngest

Table 4.19: Coefficients and Variance Proportion of Principal Components for Medical Income Data Using Midpoints of All Variables

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Income	-0.2720	0.4090		0.8370	0.2400		
Glucose		-0.3400	0.9200		0.1620		
Cholesterol	-0.4430	-0.3730	-0.2630		0.2950	-0.7130	
Hemoglobin	0.5070	-0.1610	-0.1050	0.1260	0.4350		-0.7070
Hematocrit	0.5060	-0.1680	-0.1040	0.1220	0.4350		0.7070
Red blood	0.4420	0.3680	0.1960		-0.3740	-0.6970	
White blood	0.1220	-0.6230	-0.1540	0.5090	-0.5570		
Proportion of Variance	0.4534	0.1995	0.1351	0.1006	0.0943	0.0170	0.0002
Cumulative Proportion	0.4534	0.6529	0.7880	0.8885	0.9828	0.9998	1.0000

age group. These observations have the highest PC1 values which cluster around 4. The next cluster consists of observations in the second and the third youngest age group. These observations have smaller PC1 values which range from 0 to 1. The last cluster consists of all observations in the older groups with PC values ranging from -2 to -0.5.

The pattern of degeneration in values of the first principal component as age increases described here coincides with the trend of the first principal component based on histogram-valued observations discussed in Subsection 4.4.2. However, in Figure 4.9 observations form distinct clusters along the PC1-axis. The complete separation between observations in this plot reflects a drawback of using the midpoints to represent multiple values of a symbolic observation. With only one point, we can not see that each observation actually covers an area much larger than the point shown in Figure 4.9. We can not see from the plot of Figure 4.9 that along the first principal component groups of patients have more common values than not and that the range of red blood concentration is large for all groups of patients.

Similar conclusions can be drawn from studying plots of the midpoints values along the third principal component in Figure 4.10 and Figure 4.11. Observations in these plots indicate that patients with no diabetes have lower values of PC3. Patients with mild diabetes have

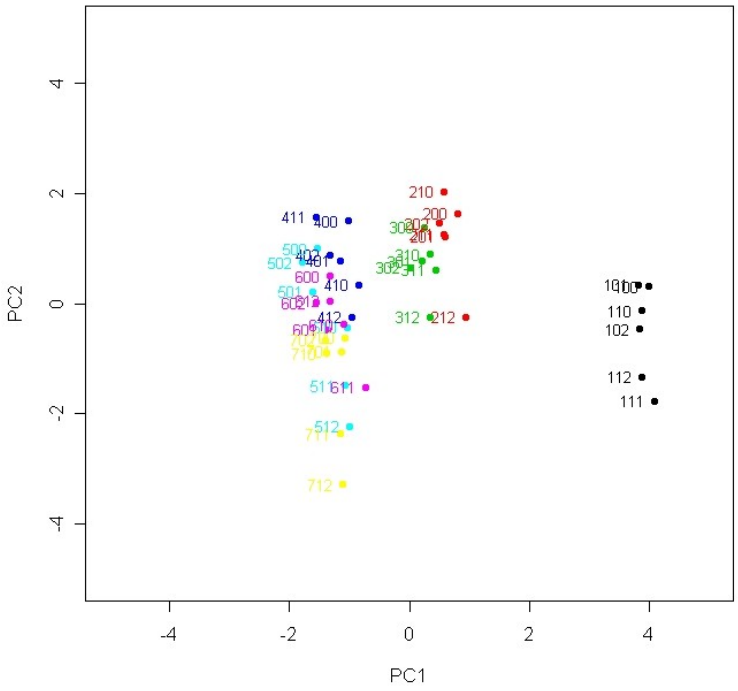

Figure 4.9: Plot of PC1 × PC2 for Medical Income Data Using Midpoints of All Variables

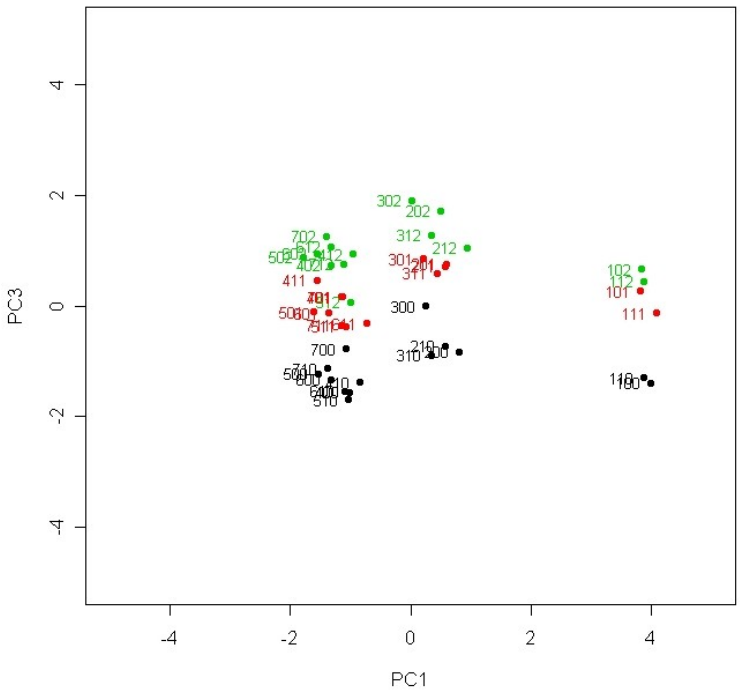

Figure 4.10: Plot of PC1 × PC3 for Medical Income Data Using Midpoints of All Variables

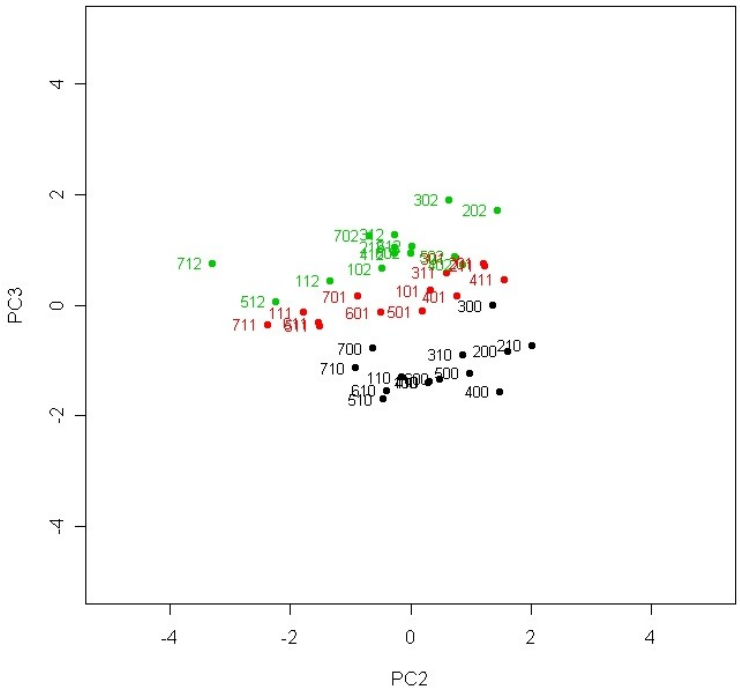

Figure 4.11: Plot of PC2 × PC3 for Medical Income Data Using Midpoints of All Variables

higher level of PC3 and patients with severe diabetes have the highest values in the third principal component. Since the third principal component is mainly composed of Glucose, we can conclude that in general glucose is correlated with diabetes. This conclusion agrees with the general trend of PC3 obtained in Subsection 4.2.2. However, we can not see from Figures 4.10 and 4.11 that the histograms representing groups with no diabetes have a much larger internal variation than those of mild and severe diabetes. About half of each group of patients with no diabetes has glucose level as high as those of diabetes groups. Therefore, using only the midpoints does not capture the full structure of histogram-valued observations.

4.5 References

Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 106-124.

Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-valued Data. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, G. Cucumel, P. Bertrand and F. de Carvalho). Springer-Verlag, Berlin, 3-12.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, New York.

Bock, H.-H. and Diday, E. (eds.) (2000). *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.

Appendix

We also apply the proposed method to the Medical Income data using Glucose, Cholesterol, Hemoglobin, Hematocrit, Red blood, and White blood. The resulting composition for the first three principal components in this analysis is very similar to the analysis using all variables. Two-dimensional plots of the first three principal components for this analysis are shown in Figures 4.12, 4.13, and 4.14. Plots resulting from application of our proposed method on the Medical Income dataset using other subsets of variables are also included in this appendix.

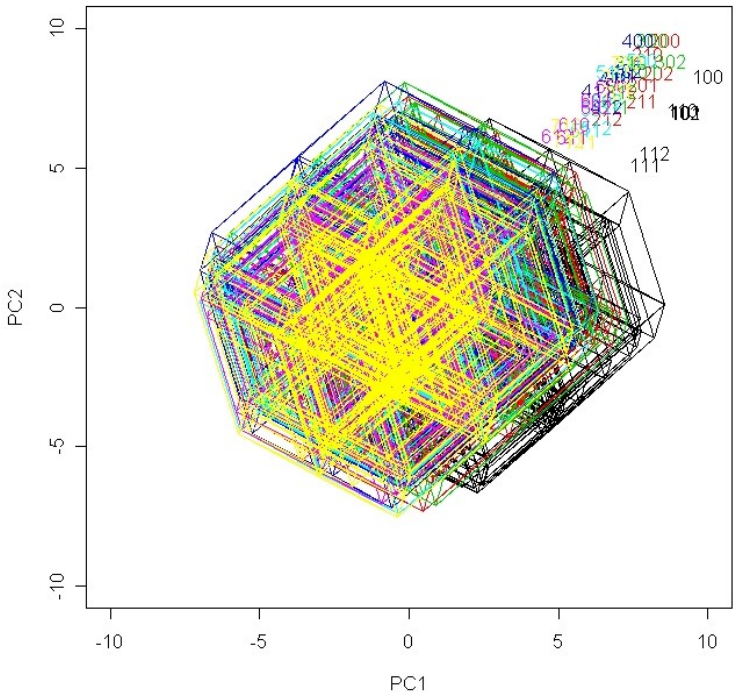

Figure 4.12: Plot of PC1 × PC2 for Medical Income Data Using Glucose, Cholesterol, Hemoglobin, Hematocrit, Red blood, and White blood (Color Represents Age)

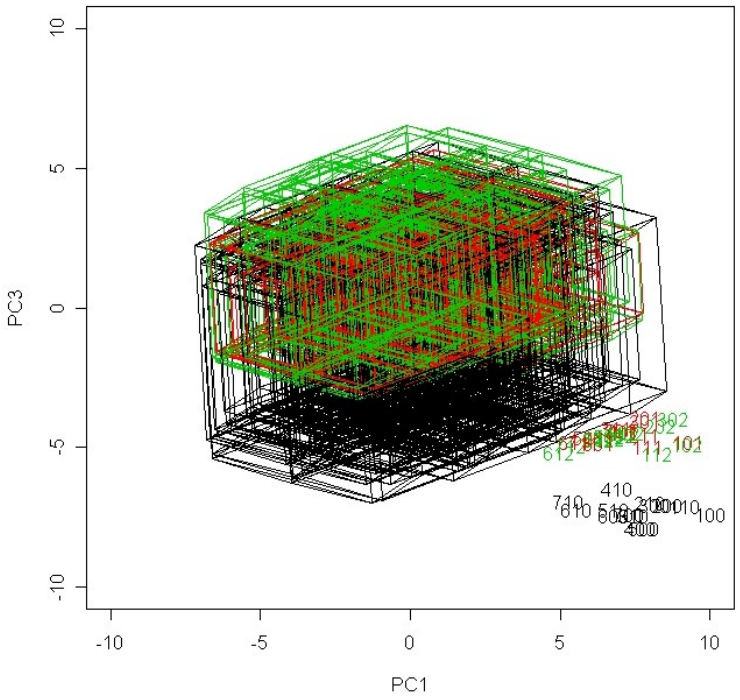

Figure 4.13: Plot of PC1 × PC3 for Medical Income Data Using Glucose, Cholesterol, Hemoglobin, Hematocrit, Red blood, and White blood (Color Represents Diabetes)

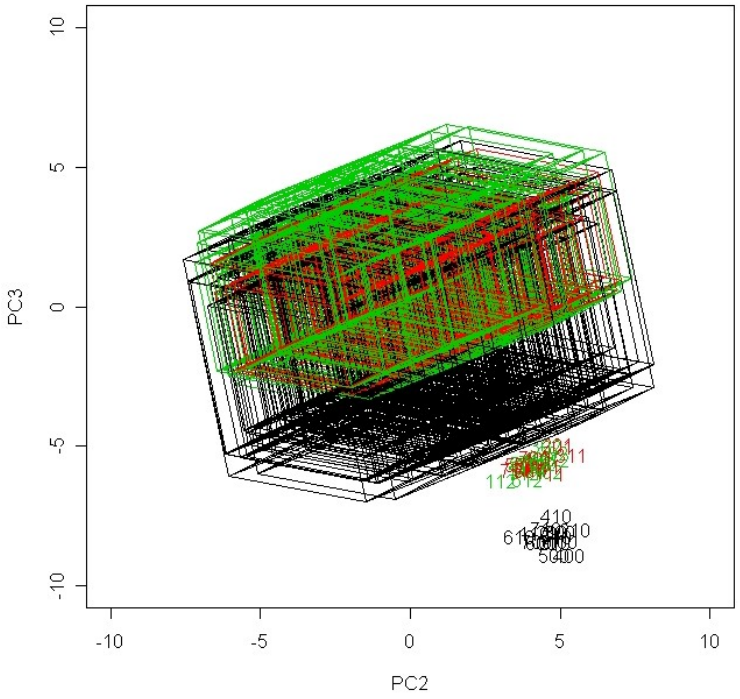

Figure 4.14: Plot of PC2 × PC3 for Medical Income Data Using Glucose, Cholesterol, Hemoglobin, Hematocrit, Red blood, and White blood (Color Represents Diabetes)

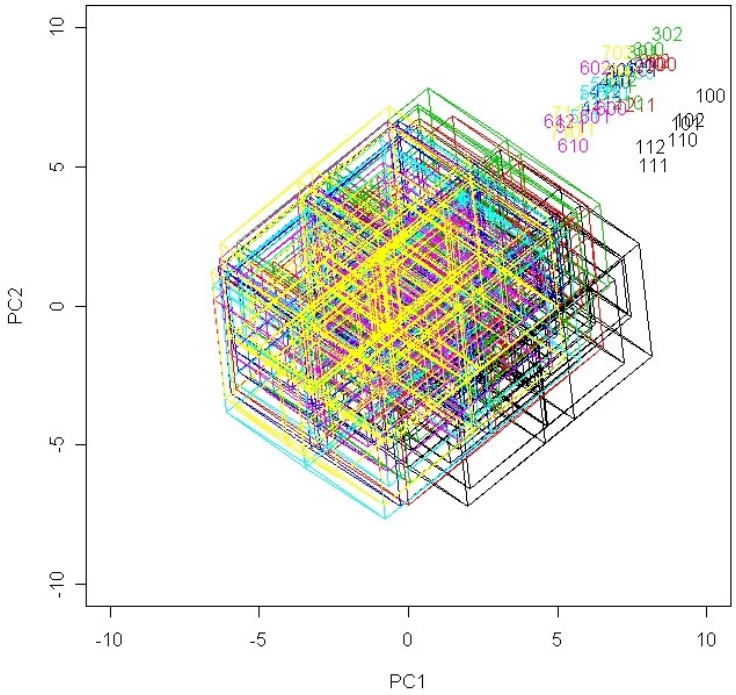

Figure 4.15: Plot of PC1 × PC2 for Medical Income Data Using Glucose, Cholesterol, Hemoglobin, Hematocrit, and Red blood (Color Represents Age)

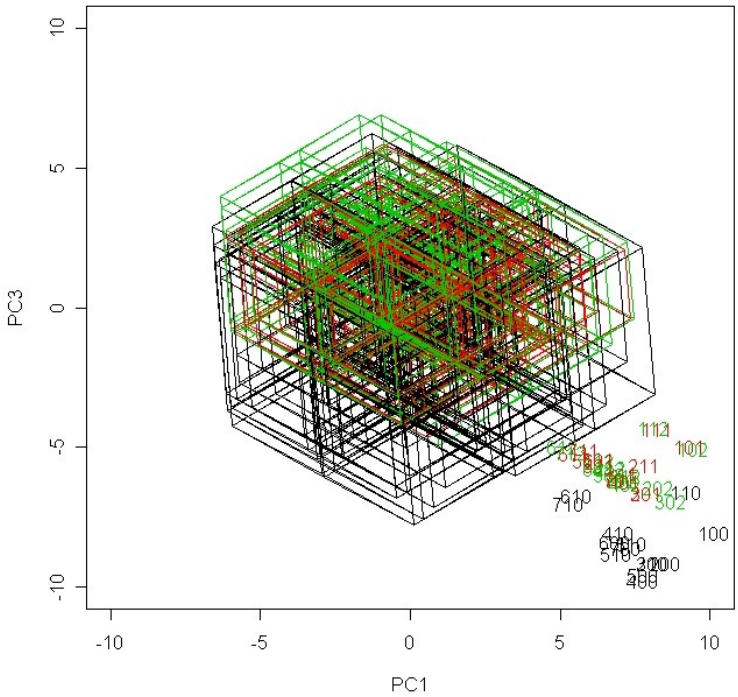

Figure 4.16: Plot of PC1 × PC3 for Medical Income Data Using Glucose, Cholesterol, Hemoglobin, Hematocrit, and Red blood (Color Represents Diabetes)

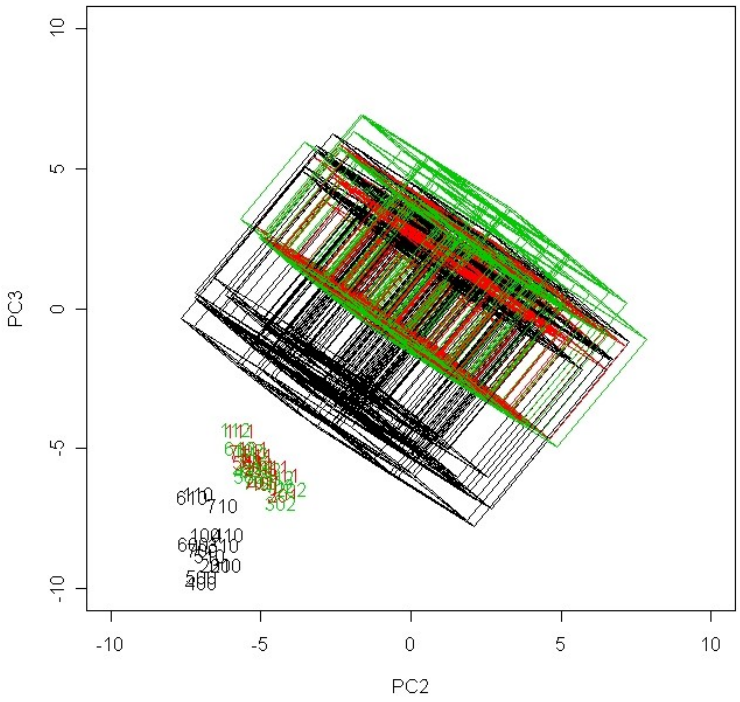

Figure 4.17: Plot of PC2 × PC3 for Medical Income Data Using Glucose, Cholesterol, Hemoglobin, Hematocrit, and Red blood (Color Represents Diabetes)

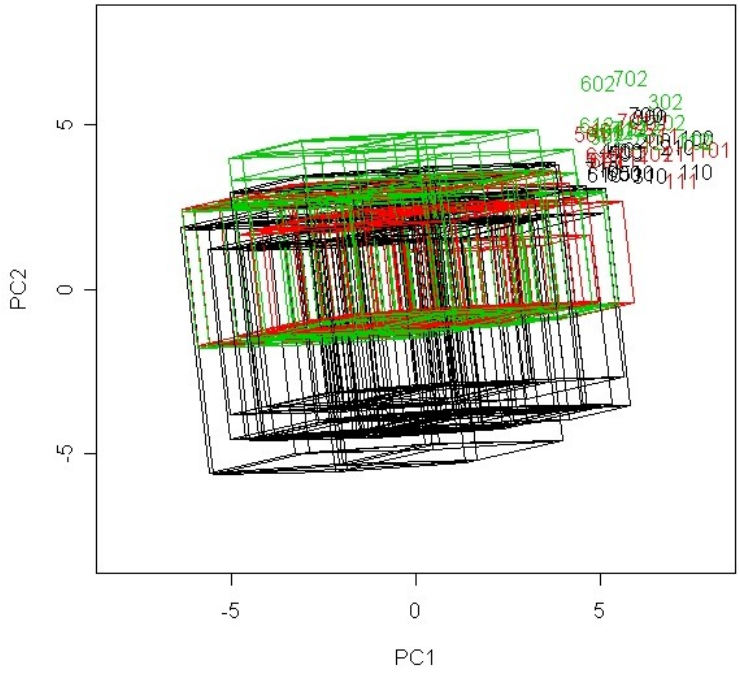

Figure 4.18: Plot of PC1 × PC2 for Medical Income Data Using Income, Glucose, Cholesterol, and Hemoglobin (Color Represents Diabetes)

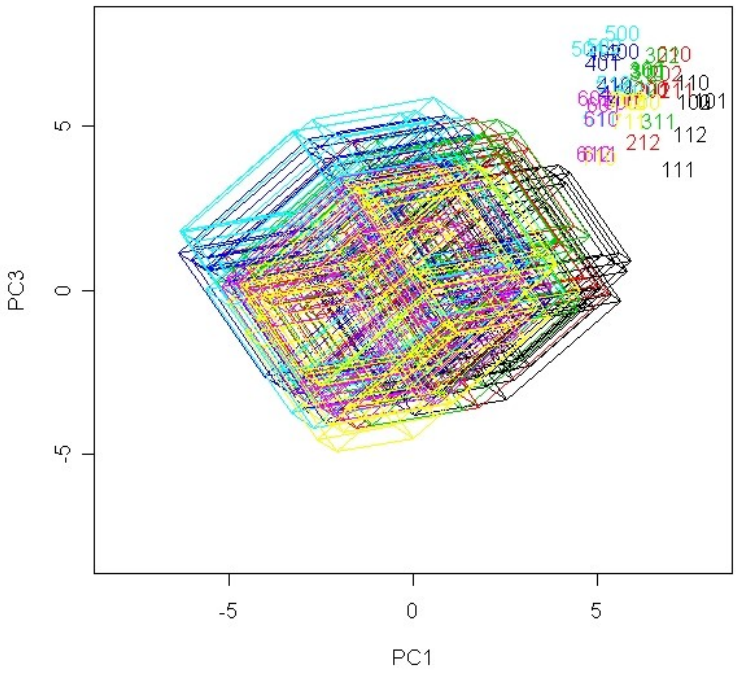

Figure 4.19: Plot of PC1 × PC3 for Medical Income Data Using Income, Glucose, Cholesterol, and Hemoglobin (Color Represents Age)

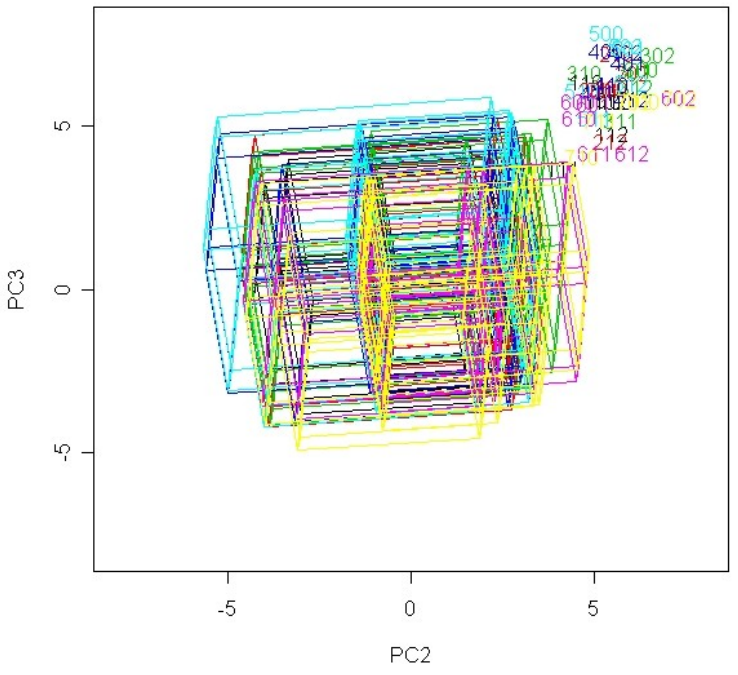

Figure 4.20: Plot of PC2 × PC3 for Medical Income Data Using Income, Glucose, Cholesterol, and Hemoglobin (Color Represents Age)

Chapter 5

Symbolic Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data

In this chapter, we introduce likelihood functions for symbolic random variables. In symbolic data, each observation contains multiple values. A symbolic observation can be a list of values, a range of values, a histogram of values or, more generally, a distribution of values. The proposed likelihood function in Section 5.1 is based on the general case where each observation is a distribution of values. In Section 5.2, we derive maximum likelihood estimators (MLE) for the mean and the variance of three of the most common types of symbolic variables: interval-valued variable, histogram-valued variable, and triangular-ditribution-valued variable.

5.1 Symbolic Likelihood Functions

Let $X_1, X_2, \ldots, X_n$ be a symbolic random sample from a population with distribution H and parameter vector $\boldsymbol{\delta}$. A random variable X from a distribution H with parameter $\boldsymbol{\delta}$ means

$$H_X(x; \boldsymbol{\delta}) = P_{\boldsymbol{\delta}}(X \leq x). \tag{5.1}$$

Let h be the density function corresponding to distribution function H of Equation (5.1). Then,

$$h_X(x; \boldsymbol{\delta}) = P_{\boldsymbol{\delta}}(X = x). \tag{5.2}$$

Now let ξ_i denote a realization of X_i for $i = 1, 2, \ldots, n$. Since X_i is a symbolic random variable, ξ_i consists of a set of values where each value has an associated relative frequency.

That is,

$$\xi_i = \{\xi_{i1}, f(\xi_{i1}); \xi_{i2}, f(\xi_{i2}); \ldots; \xi_{in_i}, f(\xi_{in_i})\}$$

where n_i is the number of values in observation ξ_i and $f(\xi_{il})$ is the relative frequency that ξ_{il} occurs within ξ_i for $l = 1, 2, \ldots, n_i$. When n_i is countable, ξ_i is itself a discrete distribution of values and $\sum_{l=1}^{n_i} f(\xi_{il}) = 1$. When n_i is uncountable, ξ_i is itself a continuous distribution of values and $\int_{-\infty}^{\infty} f(w)dw = 1$ for all real-valued w.

In general, assume X_i consists of a distribution of values. The values within X_i come from a parametric family of distributions with density function f and parameter vector $\boldsymbol{\Theta}_i$. It is important to note that f and $\boldsymbol{\Theta}_i$ are the density function and the parameter vector for values within X_i. We refer to f as the internal distribution of X_i to distinguish it from h which is the distribution of X_i, and refer to $\boldsymbol{\Theta}_i$ as the vector of internal parameters to distinguish it from $\boldsymbol{\delta}$ which is the vector of parameters of X_i associated with the density function h.

Since $\boldsymbol{\Theta}_i$ is an internal parameter vector of X_i and X_i is a random variable, $\boldsymbol{\Theta}_i$ is not a fixed vector but a random vector. Suppose $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_n$ come from a family of distributions with density function g and parameter matrix $\boldsymbol{\tau}$. Then,

$$g(\boldsymbol{\theta}_i; \boldsymbol{\tau}) = P_{\boldsymbol{\tau}}(\boldsymbol{\Theta} = \boldsymbol{\theta}_i). \tag{5.3}$$

Furthermore, given a parametric family, a distribution is uniquely determined by its parameters. Hence, there exists a one-to-one correspondence between X_i and $\boldsymbol{\Theta}_i$ based on the assumption that X_i consists of a distribution of values from a parametric family. That is, a realization ξ_i of X_i can be reconstructed if the realized vector $\boldsymbol{\theta}_i$ of $\boldsymbol{\Theta}_i$ is known and vice versa. Therefore,

$$P_{\boldsymbol{\delta}}(X = \xi_i) = P_{\boldsymbol{\tau}}(\boldsymbol{\Theta} = \boldsymbol{\theta}_i)$$

or equivalently,

$$h(\xi_i; \boldsymbol{\delta}) = g(\boldsymbol{\theta}_i; \boldsymbol{\tau}) \tag{5.4}$$

where h and g are as defined in Equations (5.2) and (5.3). The likelihood funtion of $\boldsymbol{\delta}$ given $\xi_1, \xi_2, \ldots, \xi_n$ is defined as,

$$L(\boldsymbol{\delta}; \xi_1, \xi_2, \ldots, \xi_n) = \prod_{i=1}^{n} h(\xi_i; \boldsymbol{\delta}). \tag{5.5}$$

Given Equation (5.4), the likelihood function of Equation (5.5) can be expressed in terms of $\boldsymbol{\tau}$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ by replacing $h(\xi_i; \boldsymbol{\delta})$ with $g(\boldsymbol{\theta}_i; \boldsymbol{\tau})$ as follows,

$$L(\boldsymbol{\delta}; \xi_1, \xi_2, \ldots, \xi_n) = \prod_{i=1}^{n} h(\xi_i; \boldsymbol{\delta}) = \prod_{i=1}^{n} g(\boldsymbol{\theta}_i; \boldsymbol{\tau}) = L(\boldsymbol{\tau}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n). \tag{5.6}$$

That is, the symbolic likelihood function of $\boldsymbol{\delta}$ given $\xi_1, \xi_2, \ldots, \xi_n$ can be stated as a classical likelihood function of $\boldsymbol{\tau}$ given $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$. Now classical maximum likelihood methods from estimation to regression can be applied to symbolic data via the relationship stated in Equation (5.6). Some examples of estimating the mean and the variance of symbolic data based on the proposed maximum likelihood functions follow.

5.2 Some Maximum Likelihood Estimators based on the Proposed Likelihood Functions

All examples shown in this section are based on the following assumptions. Let $X_1, X_2, \ldots, X_n$ be a symbolic random sample. Let $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_n$ be the corresponding vectors of internal parameters. Assume further that $\boldsymbol{\Theta}_i = (\Theta_{i1}, \Theta_{i2})$ and $\boldsymbol{\Theta}_i \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It is also reasonable to assume that Θ_{i1} and Θ_{i2} are independent. If an indication of dependency exists, the following method can be generalized to the multivariate case. In the future, we plan to extend this method to the multivariate case. Therefore, with the assumption of independence and that $\Theta_{i1} \sim N(\mu_1, \sigma_1^2)$ and $\Theta_{i2} \sim N(\mu_2, \sigma_2^2)$, the joint distribution of $\boldsymbol{\Theta}_i$ becomes

$$g(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = g(\theta_{i1}; \mu_1, \sigma_1^2) g(\theta_{i2}; \mu_2, \sigma_2^2).$$

The likelihood function of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is then

$$L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n) \;=\; \prod_{i=1}^{n} \left[g(\theta_{i1}; \mu_1, \sigma_1^2) g(\theta_{i2}; \mu_2, \sigma_2^2) \right]$$

$$
\begin{aligned}
&= \prod_{i=1}^{n}\prod_{j=1}^{2} g(\theta_{ij};\mu_j,\sigma_j^2) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(\theta_{ij}-\mu_j)^2/2\sigma_j^2}.
\end{aligned}
$$

Now, the estimators of μ_1, μ_2, σ_1^2 and σ_2^2 can be obtained using the classical maximum likelihood method. Thus we find

$$
\begin{aligned}
\hat{\mu}_j &= \frac{1}{n}\sum_{i=1}^{n}\theta_{ij} \\
\hat{\sigma}_j^2 &= \frac{1}{n}\sum_{i=1}^{n}(\theta_{ij}-\hat{\mu}_j)^2. \qquad (5.7)
\end{aligned}
$$

5.2.1 Interval-valued data

Let X_i be an interval-valued random variable. Then the realization of X_i is $\xi_i = [a_i, b_i]$ and for any W in the interval $[a_i, b_i]$, $W \sim U(a_i, b_i)$, i.e.,

$$
f(W = w|\xi_i) = \frac{1}{b_i - a_i}.
$$

Furthermore, let $\Theta_{i1} = E(W|X_i = \xi_i)$ and $\Theta_{i2} = Var(W|X_i = \xi_i)$. Then $\Theta_{i1} = (a_i + b_i)/2$ and $\Theta_{i2} = (b_i - a_i)^2/12$. Now, the overall mean of X_i for $i = 1, 2, \ldots, n$, is

$$
E(W) = E(E(W|X_i = \xi_i)) = E(\Theta_{i1}) = \mu_1 \qquad (5.8)
$$

and its variance is

$$
\begin{aligned}
Var(W) &= E(Var(W|X_i = \xi_i)) + Var(E(W|X_i = \xi_i)) \\
&= E(\Theta_{i2}) + Var(\Theta_{i1}) \\
&= \mu_2 + \sigma_1^2. \qquad (5.9)
\end{aligned}
$$

Note that the estimator for $Var(W)$ in Equation (5.9) is the sum of two components, the mean of the variances within each observation and the variance of the means of the observations. These two components correspond to the mean of squares within observations and the mean of squares between observations shown by Billard (2007).

Now, replacing μ_1, μ_2, σ_1^2 and σ_2^2 by their MLE in (5.7) gives the following estimators for (5.8) and (5.9), respectively,

$$\widehat{E(W)} = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\theta_{i1} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i + b_i}{2} \tag{5.10}$$

and

$$\begin{aligned}
\widehat{Var(W)} &= \hat{\mu}_2 + \hat{\sigma}_1^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\theta_{i2} + \frac{1}{n}\sum_{i=1}^{n}(\theta_{i1} - \hat{\mu}_1)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\theta_{i2} + \theta_{i1}^2 - 2\theta_{i1}\hat{\mu}_1 + \hat{\mu}_1^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\theta_{i2} + \theta_{i1}^2) - \hat{\mu}_1^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(b_i - a_i)^2}{12} + \frac{(a_i + b_i)^2}{4}\right) - \hat{\mu}_1^2 \\
&= \frac{1}{12n}\sum_{i=1}^{n}\left((a_i - b_i)^2 + 3(a_i + b_i)^2\right) - \hat{\mu}_1^2 \\
&= \frac{1}{12n}\sum_{i=1}^{n}\left(a_i^2 - 2a_ib_i + b_i^2 + 3a_i^2 + 6a_ib_i + 3b_i^2\right) - \hat{\mu}_1^2 \\
&= \frac{1}{12n}\sum_{i=1}^{n}4(a_i^2 + a_ib_i + b_i^2) - \hat{\mu}_1^2 \\
&= \frac{1}{3n}\sum_{i=1}^{n}(a_i^2 + a_ib_i + b_i^2) - \hat{\mu}_1^2 \\
&= \frac{1}{3n}\sum_{i=1}^{n}(a_i^2 + a_ib_i + b_i^2) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{a_i + b_i}{2}\right)^2 .
\end{aligned} \tag{5.11}$$

The MLE's in (5.10) and (5.11) match the empirical mean and the empirical variance for interval-valued data as defined in equations (2.3) and (2.4) which were derived by Bertrand and Goupil (2000) based on the empirical density for an interval-valued variable.

5.2.2 Histogram-valued data

Let X_i be a histogram-valued random variable. Then,

$$\xi_i = \{[a_i^1, b_i^1), p_i^1; [a_i^2, b_i^2), p_i^2; \ldots; [a_i^{s_i}, b_i^{s_i}], p_i^{s_i}\}$$

where s_i is the number of subintervals in ξ_i and p_i^l is the relative frequency associated with interval $[a_i^l, b_i^l]$. That is, a histogram-valued variable is a generalized version of an interval-valued variable. Again, let $\Theta_{i1} = E(W|X_i = \xi_i)$ and $\Theta_{i2} = Var(W|X_i = \xi_i)$. Then,

$$\Theta_{i1} = \sum_{l=1}^{s_i} p_i^l \frac{(a_i^l + b_i^l)}{2} \tag{5.12}$$

and

$$\Theta_{i2} = \frac{1}{3}\sum_{l=1}^{s_i} p_i^l[(a_i^l)^2 + a_i^l b_i^l + (b_i^l)^2] - \left(\sum_{l=1}^{s_i} p_i^l \frac{(a_i^l + b_i^l)}{2}\right)^2. \tag{5.13}$$

By analogy, the mean and the variance of X_i for $i = 1, 2, \ldots, n$ here are the same as those of interval-valued data shown in Equation (5.8) and (5.9). Now, replacing μ_1, μ_2, σ_1^2 and σ_2^2 by their MLE's shown in (5.7) with $\mathbf{\Theta}_i$ as defined in Equation (5.12) and (5.13) gives the following estimators for $E(W)$ and $Var(W)$, respectively,

$$\widehat{E(W)} = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} \theta_{i1} = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{s_i} p_i^l \frac{(a_i^l + b_i^l)}{2} \tag{5.14}$$

and

$$\begin{aligned}
\widehat{Var(W)} &= \hat{\mu}_2 + \hat{\sigma}_1^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \theta_{i2} + \frac{1}{n}\sum_{i=1}^{n} (\theta_{i1} - \hat{\mu}_1)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \theta_{i2} + \frac{1}{n}\sum_{i=1}^{n} (\theta_{i1}^2 - 2\theta_{i1}\hat{\mu}_1 + \hat{\mu}_1^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{1}{3}\sum_{l=1}^{s_i} p_i^l[(a_i^l)^2 + a_i^l b_i^l + (b_i^l)^2] - \theta_{i1}^2\right) + \frac{1}{n}\sum_{i=1}^{n} \theta_{i1}^2 - \hat{\mu}_1^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{1}{3}\sum_{l=1}^{s_i} p_i^l[(a_i^l)^2 + a_i^l b_i^l + (b_i^l)^2] - \theta_{i1}^2 + \theta_{i1}^2\right) - \hat{\mu}_1^2 \\
&= \frac{1}{3n}\sum_{i=1}^{n}\sum_{l=1}^{s_i} p_i^l[(a_i^l)^2 + a_i^l b_i^l + (b_i^l)^2] - \hat{\mu}_1^2 \\
&= \frac{1}{3n}\sum_{i=1}^{n}\sum_{l=1}^{s_i} p_i^l[(a_i^l)^2 + a_i^l b_i^l + (b_i^l)^2] - \left(\frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{s_i} p_i^l(a_i^l + b_i^l)\right)^2.
\end{aligned} \tag{5.15}$$

Again, the MLE's in (5.14) and (5.15) match the empirical mean and the empirical variance for histogram-valued data as defined in equations (2.16) and (2.17) which were derived by Billard and Diday (2003).

5.2.3 Triangular-distribution-valued data

In this example, we will derive the MLE's for the mean and the variance of another common type of symbolic data known as the triangular-distribution-valued data (for convenience, it will be refered to as triangular-valued data from here on). Let $X_1, X_2, \ldots, X_n$ be a random sample of triangular-valued variable. Let ξ_i be a realization of X_i where all values of ξ_i fall inside an interval $[a_i, b_i]$. Then for all W in the interval $[a_i, b_i]$,

$$f(W = w|\xi_i) = \begin{cases} \frac{4(w-a_i)}{(b_i-a_i)^2} & \text{for} \quad a_i \leq w < \frac{(a_i+b_i)}{2}, \\ \\ \frac{4(b_i-w)}{(b_i-a_i)^2} & \text{for} \quad \frac{(a_i+b_i)}{2} \leq w \leq b_i. \end{cases} \tag{5.16}$$

That is, the curve $f(w|\xi_i)$ of Equation (5.16) forms an equilateral triangle with the w-axis whose base is the interval $[a_i, b_i]$ and peak located at $(\frac{a_i+b_i}{2}, \frac{2}{b_i-a_i})$.

Now given $X_i = \xi_i$, the expected value of W is

$$\begin{aligned}
E(W|X_i = \xi_i) &= \int_{-\infty}^{\infty} w f(w|\xi_i) dw \\
&= \int_{a_i}^{\frac{a_i+b_i}{2}} w \frac{4(w-a_i)}{(b_i-a_i)^2} dw + \int_{\frac{a_i+b_i}{2}}^{b_i} w \frac{4(b_i-w)}{(b_i-a_i)^2} dw \\
&= \frac{4}{(b_i-a_i)^2} \left(\int_{a_i}^{\frac{a_i+b_i}{2}} w(w-a_i) dw + \int_{\frac{a_i+b_i}{2}}^{b_i} w(b_i-w) dw \right) \\
&= \frac{4}{(b_i-a_i)^2} \left(\int_{a_i}^{\frac{a_i+b_i}{2}} (w^2 - a_i w) dw + \int_{\frac{a_i+b_i}{2}}^{b_i} (b_i w - w^2) dw \right) \\
&= \frac{4}{(b_i-a_i)^2} \left((\frac{w^3}{3} - a_i \frac{w^2}{2}) \Big|_{a_i}^{\frac{a_i+b_i}{2}} + (b_i \frac{w^2}{2} - \frac{w^3}{3}) \Big|_{\frac{a_i+b_i}{2}}^{b_i} \right) \\
&= \frac{4}{(b_i-a_i)^2} \left(\frac{(a_i+b_i)^3}{24} - \frac{a_i(a_i+b_i)^2}{8} + \frac{a_i^3}{6} + \frac{b_i^3}{6} - b_i \frac{(a_i+b_i)^2}{8} + \frac{(a_i+b_i)^3}{24} \right) \\
&= \frac{4}{(b_i-a_i)^2} \left(\frac{(a_i+b_i)^3}{12} - \frac{(a_i+b_i)^3}{8} + \frac{a_i^3+b_i^3}{6} \right) \\
&= \frac{a_i+b_i}{6(b_i-a_i)^2} \left(-(a_i+b_i)^2 + 4(a_i^2 - a_i b_i + b_i^2) \right) \\
&= \frac{a_i+b_i}{6(b_i-a_i)^2} \left(-a_i^2 - 2a_i b_i - b_i^2 + 4a_i^2 - 4a_i b_i + 4b_i^2 \right) \\
&= \frac{a_i+b_i}{6(b_i-a_i)^2} \left(3a_i^2 - 6a_i b_i + 3b_i^2 \right)
\end{aligned}$$

$$= \frac{a_i + b_i}{2}; \tag{5.17}$$

and its variance is

$$\begin{aligned} Var(W|X_i = \xi_i) &= \int_{-\infty}^{\infty} (w - E(W|X_i = \xi_i))^2 f(w|\xi_i) dw \\ &= \int_{-\infty}^{\infty} (w)^2 f(w|\xi_i) dw - (E(W|X_i = \xi_i))^2. \end{aligned} \tag{5.18}$$

Now,

$$\begin{aligned} \int_{-\infty}^{\infty} (w)^2 f(w|\xi_i) dw &= \int_{a_i}^{\frac{a_i+b_i}{2}} w^2 \frac{4(w - a_i)}{(b_i - a_i)^2} dw + \int_{\frac{a_i+b_i}{2}}^{b_i} w^2 \frac{4(b_i - w)}{(b_i - a_i)^2} dw \\ &= \frac{4}{(b_i - a_i)^2} \left(\int_{a_i}^{\frac{a_i+b_i}{2}} (w^3 - a_i w^2) dw + \int_{\frac{a_i+b_i}{2}}^{b_i} (b_i w^2 - w^3) dw \right) \\ &= \frac{4}{(b_i - a_i)^2} \left((\frac{w^4}{4} - a_i \frac{w^3}{3}) \Big|_{a_i}^{\frac{a_i+b_i}{2}} + (b_i \frac{w^3}{3} - \frac{w^4}{4}) \Big|_{\frac{a_i+b_i}{2}}^{b_i} \right) \\ &= \frac{1}{(b_i - a_i)^2} \left(\frac{(a_i + b_i)^4}{16} - \frac{a_i(a_i + b_i)^3}{6} + \frac{a_i^4}{3} + \frac{b_i^4}{3} - \frac{b_i(a_i + b_i)^3}{6} + \frac{(a_i + b_i)^4}{16} \right) \\ &= \frac{1}{(b_i - a_i)^2} \left(\frac{(a_i + b_i)^4}{8} - \frac{(a_i + b_i)^4}{6} + \frac{a_i^4 + b_i^4}{3} \right) \\ &= \frac{1}{24(b_i - a_i)^2} \left(8a_i^4 + 8b_i^4 - (a_i + b_i)^4 \right). \end{aligned} \tag{5.19}$$

Substituting the right handside of Equation (5.19) for $\int_{-\infty}^{\infty} (w)^2 f(w|\xi_i) dw$ in Equation (5.18) gives

$$\begin{aligned} Var(W|X_i = \xi_i) &= \frac{1}{24(b_i - a_i)^2} \left(8a_i^4 + 8b_i^4 - (a_i + b_i)^4 \right) - (\frac{a_i + b_i}{2})^2 \\ &= \frac{1}{24(b_i - a_i)^2} \left(8a_i^4 + 8b_i^4 - (a_i + b_i)^4 - 6(a_i + b_i)^2 (b_i - a_i)^2 \right) \\ &= \frac{1}{24(b_i - a_i)^2} \left(a_i^4 + b_i^4 - 4a_i^3 b_i + 6a_i^2 b_i^2 - 4a_i b_i^3 \right) \\ &= \frac{1}{24(b_i - a_i)^2} (a_i - b_i)^4 \\ &= \frac{(a_i - b_i)^2}{24}. \end{aligned} \tag{5.20}$$

Again, let $\Theta_{i1} = E(W|X_i = \xi_i)$ of Equation (5.17) and $\Theta_{i2} = Var(W|X_i = \xi_i)$ of Equation (5.20). The MLE's for $E(W)$ and $Var(W)$ are

$$\widehat{E(W)} = \hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} \theta_{i1} = \frac{1}{2n} \sum_{i=1}^{n} (a_i + b_i) \tag{5.21}$$

and

$$\begin{aligned}
\widehat{Var(W)} &= \hat{\mu}_2 + \hat{\sigma}_1^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\theta_{i2} + \frac{1}{n}\sum_{i=1}^{n}(\theta_{i1} - \hat{\mu}_1)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\theta_{i2} + \frac{1}{n}\sum_{i=1}^{n}\theta_{i1}^2 - \hat{\mu}_1^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{(a_i - b_i)^2}{24} + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_i + b_i}{2}\right)^2 - \hat{\mu}_1^2 \\
&= \frac{1}{24n}\sum_{i=1}^{n}\left((a_i - b_i)^2 + 6(a_i + b_i)^2\right) - \hat{\mu}_1^2 \\
&= \frac{1}{24n}\sum_{i=1}^{n}(a_i^2 - 2a_ib_i + b_i^2 + 6a_i^2 + 12a_ib_i + 6b_i^2) - \hat{\mu}_1^2 \\
&= \frac{1}{24n}\sum_{i=1}^{n}(7a_i^2 + 10a_ib_i + 7b_i^2) - \hat{\mu}_1^2 \\
&= \frac{1}{24n}\sum_{i=1}^{n}(7a_i^2 + 10a_ib_i + 7b_i^2) - \left(\frac{1}{2n}\sum_{i=1}^{n}(a_i + b_i)\right)^2. \qquad (5.22)
\end{aligned}$$

Let $\bar{w} = \widehat{E(W)}$ of Equation (5.21), Equation (5.22) can be rewritten as,

$$\widehat{Var(W)} = \frac{1}{24n}\sum_{i=1}^{n}\left[7(a_i - \bar{w})^2 + 10(a_i - \bar{w})(b_i - \bar{w}) + 7(b_i - \bar{w})^2\right]. \qquad (5.23)$$

5.3 Summary

The likelihood funtions proposed in Section 5.1 provide a method to perform maximum likelihood analysis of symbolic data such as regressions and inferences. In Section 5.2, we derived the maximum likelihood estimators for the mean and the variance of interval-valued, histogram-valued, and triangular-valued variables assuming their internal mean and internal variance are independent and normally distributed. Maximum likelihood estimators of parameters for other types of symbolic random variable can be derived using the same approach we showed in Section 5.2. This method can be used to estimate the mean and the variance of symbolic data with all combinations of internal and external distributions. Furthermore, when internal parameters are not independently distributed, multivariate distribution functions are used in the likelihood funtions instead of the product of univariate

distribution functions. One potential future project is to extend our proposed likelihood functions to the multivariate case.

5.4 References

Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 106-124.

Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-valued Data. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, G. Cucumel, P. Bertrand and F. de Carvalho). Springer-Verlag, Berlin, 3-12.

Billard, L. and Diday E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, **98**, 470-487.

Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd edition. Duxbury, California.

Lehman, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd edition. Springer, New York.

Chapter 6

Conclusions and Future Research

In this dissertation, we made three contributions to the area of symbolic data analysis. First, we proposed a method of principal component analysis (PCA) for interval-valued data. Next, we extended this proposed method to a PCA method for histogram-valued data. Finally, we introduced a method to construct the likelihood funtions for symbolic data.

Although many extensions of classical PCA had been proposed for interval-valued data, they either account for only part of the variance structure of interval-valued observations or only work for very narrow intervals. In Chapter 3 of this dissertation, we proposed a method to compute the principal components using a so-called symbolic covariance structure of interval-valued data. This symbolic covariance structure accounts for all variation inherent in interval-valued observations. Therefore, our proposed method creates principal components that explain the total variance of interval-valued data.

Moreover, all current methods construct the principal components as intervals. In this work, we showed that the true structure of an interval-valued observation is a polytope in a principal components space. We proposed a method to reconstruct these polytopes in a principal components space. Our method can also be used to make plots of projections of the observations onto a 2-dimensional plane or a 3-dimensional space. Applications of the proposed method using real datasets illustrate that the plots of observations onto a PCk_1 $\times$ PCk_2 resulting from our proposed method reveal clusters of obervations with common features more clearly than plots resulting from other methods. For further analysis where numerical values of the principal components are required, we proposed creating histograms of values to represent the principal components based on our proposed PC1 $\times$ PCk plots.

Histogram-valued principal components reflect most of the internal structure of interval-valued observations in a principal components space.

Next, using the fact that histograms are extensions of intervals, we generalized the proposed PCA method for interval-valued observations to a method for histogram-valued observations. The method presented in Chapter 4 of this dissertation is the first PCA method proposed for histogram-valued data. This method inherits all advantages of the symbolic PCA method of Chapter 3. The principal components resulting from this approach explain the total variance structure of histogram-valued observations. The plots resulting from this method show the true structure of the observations and the histograms representing the principal components also reflect most of the internal variation of histogram-valued observations.

The third part of this dissertation introduced a method to construct likelihood functions for symbolic data. A symbolic random variable is itself a distribution of values. In this method, we assume that a symbolic variable is a parametric distribution belonging to a family of distributions. Using a one-to-one relationship between a parameter and its parametric distribution, we contructed the likelihood function for a symbolic random variable as the classical likelihood function of the variable's internal parameters in Chapter 5. We then, used the proposed likelihood function to derive estimators for the mean and the variance of symbolic variables that take an interval of values, a histogram of values, and a triangular distribution of values. The likelihood function presented in Chapter 5 is the first piece of theoretical work proposed for symbolic data. This likelihood function can serve as a framework to extend classical maximum likelihood estimation or classical maximum likelihood regression methods to symbolic data.

Having been introduced only recently, many areas of symbolic data analysis remain unexplored. Therefore, a wealth of research problems in theory as well as applications exists for symbolic data. Many potential research problems arise from this dissertation alone. Some

future research directions stemmed from the methods proposed in this work are presented in the following paragraphs.

In the symbolic PCA method for interval-valued data, we proposed constructing histograms to represent the principal components of the interval-valued observations. As explained in Subsection 3.2.2, the true relative frequency of a subinterval of the histogram representing a p-dimensional interval-valued observation in a principal components space is the volume of the $(p-1)$-cross-sectional polytope bounded between the subinterval endpoints. Subsection 3.2.2 further explains that the cost of computing volume for a polytope can be prohibitive when p is large. Therefore, we proposed computing the relative frequency based on the area of the polygon resulting from projecting an interval-valued observation onto a PC1 $\times$ PCk plane. The PC1-axis is included in computing the relative frequency to ensure that the largest variability in the data is accounted for. One potential extension from this method is to compute the relative frequency of a subinterval based on the volume of a 3 (or 4) dimensional polytope resulting from projecting an obervation onto a PC1 $\times$ PC2 $\times$ PCk ($\times$ PC3) space. Volume computation for polytopes with 3 (or 4) dimensions is more manageable than it is for polytopes with higher dimension. In situations where it takes three (or four) principal components to explain a reasonable amount of variation in the data, relative frequency for the histogram based on the volume of a 3 (or 4) dimensional polytope may reflect the internal variability in the data better than does the relative frequency based on the area of a polygon as proposed in our method.

Another potential future problem comes from the PCA method proposed for histogram-valued data. A histogram-valued observation can be represented as a hyper-rectangle partitioned into sub-hyperrectangles. Each sub-hyperrectangle resulting from this partition has uniform density. However, the density may differ from one sub-hyperrectangle to the next. Refer to Subsection 4.2.2 for more details. To depict density differences between sub-hyperrectangles in a plot, we propose using color to represent density. However, writing a

program to implement this proposal requires time and extensive computing effort. It makes for a challenging future project.

Yet, more future work can be generated from the likelihood function proposed in Chapter 5 than from the other methods proposed in this dissertation. Likelihood functions serve as the theoretical foundation for so many statistical applications. Based on the proposed likelihood function, we derived estimators for the mean and variance of three common types of symbolic data. Estimators for parameters of symbolic data of all possible combinations of internal and external distributions can be derived following the approach of Chapter 5. The likelihood function proposed in this dissertation is based on the assumption that the internal parameters of a symbolic variable are independent. It is possible to extend this likelihood function to include the case where the internal parameters are dependent by applying theory of multivariate statistics. Finally, many applications such as estimation, modelling, and regression methods based on maximum likelihood can be extended in the future.