Statistics and Probability Letters 106 (2015) 113–120

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Equivalency between vertices and centers-coupled-with-radii principal component analyses for interval data

Chengcheng Hao [a,1,2], Yuli Liang [a,2,3], Anuradha Roy [b,*]

[a] Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

[b] Department of Management Science and Statistics, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form 25 June 2015
Accepted 2 July 2015
Available online 17 July 2015

MSC:
primary 62H25
secondary 62H12

Keywords:
Blocked compound symmetry covariance
Principal component analysis
Interval data

ABSTRACT

Centers and vertices principal component analyses are common methods to explain variations within multivariate interval data. We introduce multivariate equicorrelated structures to vertices' covariance. Assuming the structure, we show equivalence between centers and vertices methods by proving their eigensystems proportional.

1. Introduction

Due to recent developments of cheaper and more manageable ways to store large amounts of digital data, in almost all fields complex data are registered continuously in this big data age. In these situations, it is better to model interval-valued data, which captures the variability of events, rather than classical data; Bock and Diday (2000) describe many applications. Principal component analysis (PCA) based on the interval-valued data is an active field over the last ten years. The first extensions of PCA to the interval-valued data was proposed in Cazes et al. (1997) and Chouakria et al. (1999) as "vertices principal component analysis" (V-PCA) and "centers principal component analysis" (C-PCA). Although there have been many developments in the area, especially the *symbolic sample covariance matrix* (Le-Rademacher and Billard, 2011) for symbolic data, V-PCA, C-PCA and their extensions are still methods that are commonly used to explain the total variances of a set of interval variables. This paper studies the relationship between the centers and the vertices methods.

* Corresponding author. Tel.: +1 210 458 6343; fax: +1 210 458 6350.
E-mail addresses: chengcheng.hao@sjtu.edu.cn (C. Hao), yuli.liang@stat.su.se, yuli.liang@scb.se (Y. Liang), Anuradha.Roy@utsa.edu (A. Roy).

1 Current address: Department of Automation, Shanghai Jiao Tong University, 200240 Shanghai, China.

2 Chengcheng Hao and Yuli Liang contributed equally to this work..

3 Current address: Statistics Sweden, Karlavägen 100, SE-104 51 Stockholm, Sweden.

http://dx.doi.org/10.1016/j.spl.2015.07.005
0167-7152/

Let us consider a dimension reduction problem for p-variate interval-valued data. Denote $I[\boldsymbol{Y}]$ an $n \times p$ interval-valued data matrix (Bock and Diday, 2000) such that

$$I[\boldsymbol{Y}] = \begin{pmatrix} [y_{11}^-, y_{11}^+] & \cdots & [y_{1p}^-, y_{1p}^+] \\ \vdots & \ddots & \vdots \\ [y_{n1}^-, y_{n1}^+] & \cdots & [y_{np}^-, y_{np}^+] \end{pmatrix},$$

where each component is an interval with lower and upper bounds y_{ij}^- and y_{ij}^+, respectively, $i = 1, \ldots, n$, $j = 1, \ldots, p$. The ith row of $I[\boldsymbol{Y}]$ pertains to the ith observation unit that is characterized by p (interval-valued) variables, and treated as a p-dimensional hyper-rectangle with randomly distributed single values inside, $i = 1, \ldots, n$.

The generic interval $I[y]_{ij}$ can be expressed by the couple $\{y_{ij}^c, y_{ij}^r\}$, where $y_{ij}^c = \frac{1}{2}(y_{ij}^- + y_{ij}^+)$ and $y_{ij}^r = \frac{1}{2}(y_{ij}^+ - y_{ij}^-)$ represent the center and radii of the interval $I[y]_{ij}$, respectively. The C-PCA method transforms the interval-valued data matrix into a new single valued matrix—center of the interval at hand. Thus, C-PCA method performs PCA on the $n \times p$ numerical matrix

$$\boldsymbol{Y}_{\text{C-PCA}} = \begin{pmatrix} y_{11}^c & \cdots & y_{1p}^c \\ \vdots & \ddots & \vdots \\ y_{n1}^c & \cdots & y_{np}^c \end{pmatrix}. \tag{1}$$

However, it indicates that C-PCA studies between-hyper-rectangle variation only. In order to model the additional within-hyper-rectangle variation terms that are summarized in y_{ij}^r (though these terms differ from the correct within variation values, see e.g., Le-Rademacher and Billard, 2012), V-PCA was proposed by using all vertices of the hyper-rectangle defined by the intervals of all variables for each observation. In V-PCA, each interval valued row in $I[\boldsymbol{Y}]$ is transformed to a $(2^p \times p)$-dimensional numerical matrix $\boldsymbol{Y}_{V,i}$ as follows

$$\boldsymbol{Y}_{V,i} = \begin{pmatrix} y_{i1}^- & y_{i2}^- & \cdots & y_{ip}^- \\ y_{i1}^- & y_{i2}^- & \cdots & y_{ip}^+ \\ \vdots & \vdots & \ddots & \vdots \\ y_{i1}^+ & y_{i2}^+ & \cdots & y_{ip}^- \\ y_{i1}^+ & y_{i2}^+ & \cdots & y_{ip}^+ \end{pmatrix}, \tag{2}$$

such that each row in $\boldsymbol{Y}_{V,i}$ refers to one vertex of the ith hyper-rectangle. By stacking one below the other the matrices $\boldsymbol{Y}_{V,i}$'s $i = 1, \ldots, n$, we get the new numerical-valued $(n2^p \times p)$-dimensional data matrix $\boldsymbol{Y}_{\text{V-PCA}}$ as

$$\boldsymbol{Y}_{\text{V-PCA}} = \begin{pmatrix} \boldsymbol{Y}_{V,1}' & \cdots & \boldsymbol{Y}_{V,n}' \end{pmatrix}'. \tag{3}$$

The classical V-PCA (Cazes et al., 1997) performs PCA on the data matrix $\boldsymbol{Y}_{\text{V-PCA}}$ directly, where the matrix is treated as though it represents p-variate data with $n2^p$ independent observations. Unfortunately, this assumption of independent vertices may be problematic. Since the 2^p rows of $\boldsymbol{Y}_{V,i}$ are from the same observation unit, and therefore are correlated, $i = 1, \ldots, n$. In the existing literature, Le-Rademacher and Billard (2012) presented covariance structure for interval-valued observations. Also, assuming that the values within an interval are uniformly distributed across interval, Bertrand and Goupil (2000) derived the sample mean and sample variance, and Billard (2008) derived the covariance for such a covariance matrix. Billard et al. (2011) derived maximum likelihood estimators (MLEs) for bivariate interval-valued data, which provide the theoretical backing for the so-called symbolic covariance matrix. Le-Rademacher and Billard (2011) derived the likelihood functions and some MLEs for symbolic data using the symbolic sample covariance matrix.

No matter whether the dependency between vertices is considered, the dimension of the data in the vertices method increases exponentially with the number of variables. In order to eliminate this '*curse of dimensionality*' that affects V-PCA, Lauro and Palumbo (2000) suggested to map $\boldsymbol{Y}_{\text{V-PCA}}$ into a n-dimensional vector space spanned by the normalized ranges of the intervals. In this way the total number of rows is reduced from $n2^p$ to n. Douzal-Chouakria et al. (2011) proposed to use two surrogate variables for each interval variable, viz., the interval endpoints. That is, the interval variable $[y_{ij}^-, y_{ij}^+] \forall\, i = 1, \ldots, n,\ j = 1, \ldots, p$ is replaced by two variables y_{ij}^- and y_{ij}^+, then a standard principal component analysis can be performed on the resulting $n \times 2p$ classical dataset. Douzal-Chouakria et al. (2011) also proposed another possible way to accommodate intervals of differing lengths by replacing each interval variable via two surrogate variables, viz., the mid point and range variables. See also, Lauro and Palumbo (2000), and Palumbo and Lauro (2003).

We consider the correlation within vertex data and reduce its redundancy in a different manner. Giordani and Kiers (2006, p. 385) mentioned "C-PCA does not exploit all the available information in detecting the underlying structure of the data". One has to accomplish PCA for interval data by taking centers and radii variables together to calculate the percent and percent cumulative eigenvalues. Otherwise, some part of the total variance would be unaccountable. To circumvent this Roy (submitted for publication) proposed a two-stage PCA, called centers-coupled-with-radii PCA (CCR-PCA), using multivariate equicorrelated and jointly equicorrelated covariance structures (Leiva, 2007), also referred to as blocked compound symmetry (BCS) covariance structures, and used adjusted percent eigenvalues and adjusted percent cumulative