Symbolic Principal Components for Interval-Valued Data

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

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

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

1 Introduction

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

2 Background

2.1 Standard Principal Component Analysis

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