International Journal of Innovative Computing, Information and Control
Volume **2**, Number **1**, February **2006**

ICIC International ©2006 ISSN 1349-4198
pp. **69–82**

ON WEIGHTED PRINCIPAL COMPONENT ANALYSIS FOR INTERVAL-VALUED DATA AND ITS DYNAMIC FEATURE

Mika Sato-Ilic and Junya Oshima

Faculty of Systems and Information Engineering
University of Tsukuba
Tsukuba 305-8573, Japan
mika@risk.tsukuba.ac.jp

Received February 2005; revised August 2005

Abstract. *This paper presents a weighted principal component analysis (WPCA) for interval-valued data by using the result of fuzzy clustering. In this method, we consider two data structures. One is a classification structure and the other is a data structure captured by principal components. The classification structure is used for estimating the weights and the data structure captured by principal components is used for self analysis. By considering the two structures, we can reduce the risk of a wrong assumption for the true data structure, when compared with the conventional methods which assume only one data structure. Moreover, we investigate the dynamic feature of the principal components under the assumption of linear transformation from minimum values of the interval-valued data to a weighted combination of minimum and maximum values of the interval-valued data. Several numerical examples show the better performance of the proposed method.*

Keywords: Symbolic data, Fuzzy clustering, Uncertainty, Classification structure

1. **Introduction.** Principal component analysis (PCA) [1], [6] is a well known method in the area of multivariate analysis which can represent the main tendency of an observed data. This method obtains the projection space spanned by several vectors which represent principal components. These principal components show the latent factors of the data. By using the data structure on this projection space, latent features of the data can be captured. Due to this particular feature of PCA, this method has been used when we can not see any significant structure of the data in the space spanned by variables.

In this paper, we focus on interval-valued data. For example, we observe how long people watch TV each day and ask a subject, "How long do you watch TV each day?". If one subject answers 4.5 hours, then this observation has a risk of not showing a realistic situation, compared to an answer that is from 3 to 5 hours. Such an observation should be treated as interval-valued data.

Many conventional analyzes for the interval-valued data have been proposed. [2], [4], [5], [8], [14]. This is because direct representation of the interval-valued data can show the realistic situation of the data and analyze the interval-valued data directly without loss of important information contained in the interval-valued data.