Dynamics of Continuous, Discrete and Impulsive Systems
Series B: Applications & Algorithms **17 (2010)** 709-721
 http://www.watam.org

THE CHOICE OF A SIMILARITY MEASURE WITH RESPECT TO ITS SENSITIVITY TO OUTLIERS

A.M. Rubinov[1], N. Sukhorukova[1] and J.Ugon[1]

[1]Center for Informatics and Applied Optimisation, Graduate School of ITMS
Ballarat University, P.O. Box 663, Ballarat, Vic. Australia

Corresponding author email:{j.ugon,n.sukhorukova@ballarat.edu.au}

Abstract. This paper examines differences in the choice of similarity measures with respect to their sensitivity to outliers in clustering problems, formulated as mathematical programming problems. Namely, we are focusing on the study of norms (norm-based similarity measures) and convex functions of norms (function-norm-based similarity measures). The study consists of two parts: the study of theoretical models and numerical experiments. The main result of this study is a criterion for the outliers sensitivity with respect to the corresponding similarity measure. In particular, the obtained results show that the norm-based similarity measures are not sensitive to outliers whilst a very widely used square of the Euclidean norm similarity measure (least squares) is sensitive to outliers.

Keywords. Clustering, dissimilarity, outliers, least squares, optimization.

1 Introduction

The problem of minimisation of the sum of distances is a very important problem in optimisation, as it has many practical applications in location analysis and in data analysis (clustering). The distance, or similarity measure, is used to evaluate the dissimilarity between two points. The aim is to find the solution minimising the average dissimilarity between a centre and a set of point (see [2, 5, 6, 13]).

This formulation of clustering problems leads to difficult optimisation problems, with the objective functions which depend on the choice of similarity measures. In many cases metrics, in particular norms, are the preferable similarity measures. On the other hand, the most popular similarity measure is $\|\cdot\|_2^2$, where $\|\cdot\|_2$ stands for the Euclidean norm. In particular, k-means algorithm for clustering and its generalizations (see [7, 8, 9, 12]) are based on the use of an objective function with this measure.

A substantial amount of research goes in evaluating the similarities and differences between two clustering approaches (see e.g. [10] and references therein). Yet the square of the Euclidean norm is often preferred to the clustering results which are based on norms $\|\cdot\|_p$. The reason is that the