Bull. Aust. Math. Soc. **95** (2017), 516–517
doi:10.1017/S000497271700003X

REGRESSION CLUSTERING USING GIBBS SAMPLER AND OPTIMAL CLUSTER NUMBER ESTIMATION

LING DING

(Received 28 December 2016; first published online 2 March 2017)

2010 *Mathematics subject classification*: primary 62H30; secondary 91C20.

Keywords and phrases: cluster analysis, multiple regression, Gibbs sampler.

Regression clustering, or cluster regression, integrates cluster analysis and multiple regression to develop a new method for data mining. This thesis contains three parts related to two critical topics in cluster regression, including choosing a clustering method and estimating the optimal cluster number.

In the first part, we propose an efficient cluster selection procedure using the Gibbs sampler, a Markov chain Monte Carlo algorithm, by which we can find the best partition with high probability and efficiency without the need to compare all candidate partitions one by one. We apply this procedure on the ordinary regression clustering to achieve data partition and parameter estimation simultaneously in a cohesive way. We also list some asymptotic results related to the proposed methods as well as algorithms. In addition, we present simulation studies and real data applications to illustrate and verify the advantage of this method.

A fundamental problem with most clustering approaches is that the number of clusters needs to be pre-specified before clustering, and the clustering results heavily depend on the number of clusters. So it is vital to have efficient methods to determine the optimal cluster number in order to achieve appropriate clustering results. Hence, the second topic in this thesis is to propose some criteria for estimating the optimal cluster number. We introduce the idea of a global likelihood function and use it to propose a class of extended model selection criteria to better meet the needs of cluster number selection in regression clustering. We provide simulation studies that cover a wide range of data sets to illustrate the performance of the newly proposed criteria.

In the third part, we generalise the ordinary regression clustering to orthogonal regression clustering (ORC), which is a new way to detect different linear structures in a data set. In contrast to most of the existing clustering methods, ORC aims to detect

Thesis submitted to the University of Melbourne in July 2016, degree approved on 8 August 2016, supervisor Guoqi Qian.

functional linear relationships by using orthogonal regression. ORC has the advantage over other regression clustering methods that a response variable is not needed, which makes it a useful tool in statistics and exploratory data analysis for finding interesting linear patterns.

LING DING, School of Mathematics and Statistics,
University of Melbourne, Victoria 3010, Australia
e-mail: lingdmani@gmail.com

https://doi.org/10.1017/S000497271700003X Published online by Cambridge University Press