Applied Mathematics and Computation 296 (2017) 277–288

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Extracting clusters from aggregate panel data: A market segmentation study☆

Graça Trindade[a], José G. Dias[a,*], Jorge Ambrósio[b]

[a] Business Research Unit, Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[b] LAETA, IDMEC, Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal

ARTICLE INFO

Keywords:
Sequential quadratic programing
Cluster analysis
Panel data
Market segmentation

ABSTRACT

This paper introduces a new application of the Sequential Quadratic Programing (SQP) algorithm to the context of clustering aggregate panel data. The optimization applies the SQP method in parameter estimation. The method is illustrated on synthetic and empirical data sets. Distinct models are estimated and compared with varying numbers of clusters, explanatory variables, and data aggregation.

Results show a good performance of the SQP algorithm for synthetic and empirical data sets. Synthetic data sets were simulated assuming two segments and two covariates, and the correlation between the two covariates was controlled in three scenarios: $\rho = 0.00$ (no correlation), $\rho = 0.25$ (weak correlation), and $\rho = 0.50$ (moderate correlation). The SQP algorithm identifies the correct number of segments for these three scenarios based on all information criteria (AIC, AIC3, and BIC) and retrieves the unobserved heterogeneity in preferences. The empirical case study applies the SQP algorithm to consumer purchase data to find market segments. Results for the empirical data set can provide insights for retail category managers because they are able to compute the impact on the marginal shares caused by a change in the average price of one brand or product.

1. Introduction

Cluster analysis or clustering is the research field that deals with the definition of groups of objects (called clusters or segments) in such a way that members of the same cluster are more similar to each other than to those in other groups. Clustering techniques have been enhanced in many fields of research, such as machine learning, statistics, bioinformatics, and marketing (e.g., market segmentation). In recent years, widespread data collection of longitudinal and stream data has created the need for the identification of unique groups or trajectories in panel data. This type of observations tends to be challenging for any clustering process given data dependency [1]. Despite having attracted much attention in statistics and machine learning, most of the proposals have adapted the clustering of cross-sectional data to longitudinal data [2,3]. Heuristic cluster analysis for time series data may operate directly on the correlation matrix, a common practice in financial econometrics (e.g., [4,5]). More recently, hybrid algorithms have been introduced combining filtering processes that control

☆ The authors would like to thank the editor-in-chief, an associate editor, and four anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

* Correspondence to: Department of Quantitative Methods for Management and Economics, Edifício ISCTE, Av. Forças Armadas, 1649-026 Lisboa, Portugal. Fax: +351 217964710.
E-mail addresses: jose.dias@iscte.pt, jose.g.dias@gmail.com (J.G. Dias).

http://dx.doi.org/10.1016/j.amc.2016.10.012
0096-3003/

the longitudinal structure with heuristic clustering. For example, Sáfadi [6] proposes filtering the time series using independent component analysis and then, based on the coefficients or correlation obtained, time series are clustered by complete linkage. An alternative process of filtering using hidden Markov models prior to heuristic clustering has been suggested to cluster time series [7,8].

Model-based clustering, also known as finite mixture or latent class models, has proven to be a powerful paradigm in many scientific fields as a parametric alternative to heuristic clustering [9,10]. Many applications have been developed with different purposes ranging from outlier detection to density estimation. Nevertheless, cluster analysis has been its main objective by assuming that each component of the mixture is a distinct cluster. In the context of survival or reliability analysis, Razali and Al-Wakeel [11] and Elmahdy [12] model survival data using mixtures of Weibull distributions, whereas Alves and Dias [13] compare distinct specifications of mixtures in the context of behavioral credit scoring analysis. Proposals that accommodate for serial dependencies in clustering can have different definitions given data structure. For instance, a mixture of Markov chains [14] and a mixture of hidden Markov models [15] have been applied to clustering times series data.

Market segmentation is one of the most important applications of clustering. It simplifies a complex market structure by dividing the market into submarkets and provides the foundations for developing specific strategies for each segment. It uses demographic, geographic, or other segmentation criteria that can help uncover behavioral differences associated with specific groups of consumers. As a result of a lack of information about demand, the market analysis is often based on supply-side data. For over a decade, mixture models have been the standard technique for market analysis [16]. The Latent Segment Logit (LSL) model of Zenor and Srivastava [17], which is a generalization of the multinomial logit (MNL) model of McFadden [18], allows heterogeneity through segment-varying parameters. In other words, it retrieves the structure of market segments that is lost due to data aggregation when price is the variable to explain the choice between different products/brands in the same market.

This study proposes a mixture model for clustering panel data that takes unobserved heterogeneity due to aggregation into account. It extends the model proposed in Zenor and Srivastava [17] by taking multiple covariates. Apart from the generalized specification of the model, a deterministic algorithm for model estimation is discussed and applied to a market segmentation case study.

In the next section (Section 2), the model is defined by introducing a general clustering framework that can be used in other contexts besides market segmentation. Section 3 addresses the algorithm for estimating the model. Section 4 discusses inference and model selection. Section 5 explores the model using synthetic data. Section 6 illustrates the use of the model in the context of market segmentation. Two levels of aggregation are discussed: brand- and product-level analyses. The paper ends with a discussion of the main contributions, limitations, and suggestions for further extensions.

2. Definition of the model

Following the model of Zenor and Srivastava [17], there are two sets of latent variables that are not directly observed in the input data which are modeled as (1) the multinomial logit choice of option i, at time t, within cluster s (m_{ist}) and (2) the expected frequencies within each cluster, where the observed choice frequencies of each option are combined in a multinomial logit specification to generate (n_{ist}). These variables are given by

$$
\begin{aligned}
M_{ist} &= \frac{\exp\left(\alpha_{0is} + \sum_v \alpha_{vs} X_{ivt}\right)}{\sum_j \exp\left(\alpha_{0js} + \sum_v \alpha_{vs} X_{jvt}\right)} \\
M_{it} &= \sum_{s=1}^{S} g_s M_{ist} \\
m_{ist} &= E[M_{ist} | \boldsymbol{\alpha}_s] \\
N_{ist} &= N_{it} \frac{g_s m_{ist}}{\sum_r g_r m_{irt}} \\
n_{ist} &= E[N_{ist} | \boldsymbol{\alpha}_s, \mathbf{g}_s]
\end{aligned}
$$

where M_{ist} is the proportion of option i in cluster s at time t; M_{it} is the proportion of option i at time t; thus, m_{ist} is the expected share of option i in cluster s at time t. The intercept parameters in cluster s are α_{0is} and the slope parameters are α_{vs}; g_s is the size of cluster s; X_{ivt} is the explanatory variable v for option i observed at time t; and N_{it} is the total number of counts of option i observed at time t and V is the number of explanatory variables in the model. The expectations of the number of counts of option i in cluster s at time t is given by n_{ist}. To keep the model statistically identified, one of the intercepts in each cluster needs to be fixed at zero.

Assuming that the expected number of counts, n_{ist}, is the product of an independent multinomial, the likelihood function is defined by

$$
L(\boldsymbol{\alpha}, \mathbf{g}) = \prod_t \left[\frac{\left(\sum_i \sum_s n_{ist}\right)!}{\prod_i \prod_s n_{ist}!} \prod_i \prod_s \left(g_s m_{ist}\right)^{n_{ist}} \right].
$$