Journal of Statistical Planning and Inference 173 (2016) 47–63

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Exploring dependence between categorical variables: Benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms

Michail Papathomas [a,*], Sylvia Richardson [b]

[a] School of Mathematics and Statistics, University of St Andrews, The Observatory, Buchanan Gardens, St Andrews, KY16 9LZ, UK
[b] MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK

ARTICLE INFO

Article history:
Received 14 October 2014
Received in revised form 10 December 2015
Accepted 3 January 2016
Available online 15 January 2016

Keywords:
Bayesian model selection
Sparse contingency tables
Graphical models

ABSTRACT

This manuscript is concerned with relating two approaches that can be used to explore complex dependence structures between categorical variables, namely Bayesian partitioning of the covariate space incorporating a variable selection procedure that highlights the covariates that drive the clustering, and log-linear modelling with interaction terms. We derive theoretical results on this relation and discuss if they can be employed to assist log-linear model determination, demonstrating advantages and limitations with simulated and real data sets. The main advantage concerns sparse contingency tables. Inferences from clustering can potentially reduce the number of covariates considered and, subsequently, the number of competing log-linear models, making the exploration of the model space feasible. Variable selection within clustering can inform on marginal independence in general, thus allowing for a more efficient exploration of the log-linear model space. However, we show that the clustering structure is not informative on the existence of interactions in a consistent manner. This work is of interest to those who utilize log-linear models, as well as practitioners such as epidemiologists that use clustering models to reduce the dimensionality in the data and to reveal interesting patterns on how covariates combine.

1. Introduction

Detecting high-order interactions is becoming increasingly important for investigators in many fields of research. It is now understood that covariates may combine to affect the probability of an outcome, and that the effect of a particular covariate may only be important in the presence of other covariates. For example, in epidemiology it is of interest to examine the presence of interactions between smoking, environmental pollutants and dietary habits (Bingham and Riboli, 2004). In genetic association studies, it is of interest to detect gene–gene and gene-environment interactions in high dimensional data (Wakefield et al., 2010).

In this manuscript, we examine and discuss the relation between variable selection within Bayesian partitioning on one hand and log-linear modelling with interactions on the other, and the extend to which this relation can be explored in log-linear model search. Log-linear modelling is the most popular approach when searching for interactions, used by

* Corresponding author. Tel.: +44 1334461818.
E-mail address: M.Papathomas@st-andrews.ac.uk (M. Papathomas).

http://dx.doi.org/10.1016/j.jspi.2016.01.002

statisticians as well as practitioners in substantive applications. In a classical setting, attempting to fit a linear model with a large number of parameters sometimes requires an impractically large vector of observations to produce valid inferences (Burton et al., 2009). Within the Bayesian framework, the use of prior distributions alleviates identifiability or maximum likelihood estimation difficulties; see Dobra and Massam (2010). However, the space of competing models becomes vast, and model search algorithms like the Reversible Jump approach (Green, 1995) require a large number of iteration before they converge and produce reliable posterior model probabilities (Clyde and George, 2004; Dobra, 2009). With regard to contingency tables, the number of cells and possible graphical log-linear models that explain the cell counts increases exponentially with the number of covariates. For example, considering 20 covariates with 3 levels implies 3^{20} cells and approximately 1.5×10^{57} possible models.

Due to the difficulties associated with searching for interactions within a linear modelling framework, alternative approaches were adopted focusing on the reduction of the dimensionality in the data. Clustering is often the tool used to reduce dimensionality (see, for example Zhang et al., 2010), sometimes combined with a variable selection step (Chung and Dunson, 2009). Whilst log-linear modelling is a standard mathematical construction, there are many different clustering modelling approaches. For the purposes of this manuscript, we choose to focus on Bayesian clustering based on the Dirichlet process. The Dirichlet process produces flexible partitioning, allowing for the evaluation of the uncertainty with regard to the clustering of the subjects. We use a combination of Dirichlet process modelling and variable selection, implementing the modified variable selection step described in Papathomas et al. (2012), so that the covariates that contribute substantially to the clustering are identified.

We focus on categorical variables and log-linear models, as this is the standard framework for modelling interactions. In fact, for a set of categorical variables, where at least one is binary, there is a correspondence between log-linear and logistic regression modelling, and under certain conditions it is valid to translate inferences from the log-linear framework to the logistic one, regarding the presence of main effects and interactions; see Agresti (2002) and Papathomas (2015).

We explore the relation between log-linear modelling and clustering for two reasons. First, practitioners such as epidemiologists often use clustering in order to explore the manner in which covariates combine to affect the risk for disease; see Papathomas et al. (2011b). They frequently question if the clustering structures may inform in some way on the existence of interactions in associated log-linear models, and our investigation aims to provide some answers. Second, we aim to explore if any relation between log-linear modelling and clustering can be utilized to assist the exploration of large log-linear model spaces and the search for high-order interactions. The intuitive idea is that models that combine clustering and variable selection do not select covariates in accordance with the size of their marginal effect. Covariates are selected because they work together and combine with each other to create distinct groups of subjects. Consequently, this type of modelling may be able to inform on covariates that combine to describe the structure in the data, rather than covariates with a strong marginal signal.

In this manuscript, we are not concerned with the large-p problem, where thousands or hundreds of thousands of covariates are considered; see, for example, Hans et al. (2007), Richardson et al. (2010), or Cho and Fryzlewicz (2012) for a comprehensive review. Although our discussion is relevant to data sets of higher dimension, we focus on a relatively modest number of categorical variables, say one hundred or fewer, with fewer than twenty involved in interaction terms.

We demonstrate that inferences from clustering can potentially reduce the number of factors considered, by determining covariates that are independent of all others. Subsequently, the number of competing log-linear models is reduced, making the exploration of the model space feasible. This is crucial when analysing data that form large sparse contingency tables. We introduce a novel model search approach for a log-linear model space, informed by results from variable selection within clustering. We demonstrate that this model search algorithm can identify parts of the model space that contain models of low probability (thus helping to locate the highest probability model in less iterations, on average, compared to a less informed approach), especially in the presence of covariates that are independent of all other factors. With regard to limitations, first we show that there is no dependable correspondence between the covariate profile of the generated clusters and the log-linear model that best describes the data. More importantly, using simulated and real data, we show that variable selection within Bayesian clustering does not consistently detect marginal independence between covariates when the independent covariates form interaction terms with other factors.

Studies on the relation between the two different modelling approaches are not commonplace. In Dunson and Xing (2009), a Dirichlet process mixture of product multinomial distributions defines the prior on a set of categorical variables. Bhattacharya and Dunson (2012) model the joint distribution of categorical variables using simplex factor models. In contrast to our approach, variable selection switches are not considered in the aforementioned manuscripts, and no direct connection is made with log-linear model search. We are aware of three recent manuscripts that utilize clustering. The first is Marbac et al. (2014), where the clustering is applied to the covariates. This is different to the clustering we consider, widely used by practitioners, where the partitioning is applied to the subjects of the study. The second, Johndrow et al. (2014), has some connection to our work. In this preprint, the authors examine situations where the joint distribution implied by a sparse log-linear model has a low-rank tensor factorization. Relevant to our work is also the third, Zhou et al. (2015). This manuscript introduces and utilizes the idea that marginally independent variables reduce the dimensionality of the problem. This approach, central also to our work, was conceived and developed independently in parallel in our manuscript. The modelling in Zhou et al. (2015) with regard to marginal independence has similarities with the one we adopt, and significant differences. Our focus is different from Zhou as we utilize results from clustering to accelerate Bayesian log-linear graphical model selection with the Reversible Jump, a novel approach in log-linear model determination. We come back to these points