Journal of Multivariate Analysis 143 (2016) 440–452

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On the closure of relational models

Anna Klimova [a,*], Tamás Rudas [b]

[a] Institute of Science and Technology Austria, Am Campus 1, A-3400 Klosterneuburg, Austria
[b] Eötvös Loránd University, Pázmány Péter sétány 1/A, H-1117 Budapest, Hungary

ARTICLE INFO

Article history:
Received 3 January 2015
Available online 17 October 2015

AMS subject classifications:
62H12
62H17

Keywords:
Algebraic variety
Bregman divergence
Contingency table
Extended MLE
Iterative scaling
Relational model

ABSTRACT

Relational models for contingency tables are generalizations of log-linear models, allowing effects associated with arbitrary subsets of cells in a possibly incomplete table, and not necessarily containing the overall effect. In this generality, the MLEs under Poisson and multinomial sampling are not always identical. This paper deals with the theory of maximum likelihood estimation in the case when there are observed zeros in the data. A unique MLE to such data is shown to always exist in the set of pointwise limits of sequences of distributions in the original model. This set is equal to the closure of the original model with respect to the Bregman information divergence. The same variant of iterative scaling may be used to compute the MLE whether it is in the original model or in its closure.

1. Introduction

The existence of maximum likelihood estimates under log-linear models for contingency tables has been thoroughly studied, see [14,2,3,18], among others. It was established that the maximum likelihood estimates of the cell parameters always exist if the observed table has only positive cell counts, and may exist if some of the observed counts are zero. The patterns of zero cells that lead to the non-existence of the MLE were described in several forms (cf. [14,11]).

Within the extended log-linear model class, all data sets have an MLE, irrespective of the pattern of zeros. An extended log-linear model may be obtained as the closure of the original model in the topology of pointwise convergence (cf. [18]), or the closure with respect to the Kullback–Leibler divergence (cf. [9]), or as the aggregate exponential family [7].

The contribution of this paper is motivated by statistical problems in which models more general than log-linear need to be considered. To illustrate, suppose that the management of a large supermarket classifies all goods in stock into one of three mutually exclusive and exhaustive categories, say, food (F), non-food household (N) and other (O), and wishes to study how the daily sales of each group are related. This is a standard task in market basket analysis (cf. [6]). The first model of interest, routinely, is independence, but the usual model of independence of the three indicator variables is not applicable in this case: if p_F, p_N and p_O denote the probabilities that a purchase (a basket) contains at least one item from the F, N and O groups, respectively, then the probability of an empty purchase would be $(1-p_F)(1-p_N)(1-p_O)$, which has to be positive, in spite of the fact that there are no purchases which do not contain any items.

One alternative independence concept to apply is the AS-independence of the three variables [1]. The indicator variables F, N, and O are said to be AS-independent if

$$p_{FN} = p_F p_N, \qquad p_{FO} = p_F p_O, \qquad p_{NO} = p_N p_O, \qquad p_{FNO} = p_F p_N p_O. \tag{1}$$

Relational models introduced by Klimova, Rudas, and Dobra [17] contain model (1) and many other models of association.

* Corresponding author.
E-mail addresses: aklimova25@gmail.com (A. Klimova), rudas@tarki.hu (T. Rudas).

http://dx.doi.org/10.1016/j.jmva.2015.10.005

A relational model on a contingency table is generated by a class of non-empty subsets of cells and can be specified in the form:

$$\log \boldsymbol{\delta} = \mathbf{A}'\boldsymbol{\beta}. \tag{2}$$

Here, $\boldsymbol{\delta}$ denotes the vector of cell parameters, probabilities or intensities, and $\mathbf{A}$ is the 0–1 matrix whose rows are the indicators of generating subsets. A hierarchical log-linear model (cf. [4]) applies to a table which is a Cartesian product, and the model is generated by a collection of cylinder sets corresponding to marginals of the table and thus is a special case of a relational model. If the row space of $\mathbf{A}$ contains the vector $\mathbf{1}' = (1, \ldots, 1)$, as in the case of hierarchical log-linear models, then the model is said to include the overall effect. A model with the overall effect can be parameterized to include a common parameter in every cell, often called the normalizing constant. The models without the overall effect cannot be parameterized in such a way. The peculiar property of relational models without the overall effect is that models for probabilities (appropriate under multinomial sampling) and models for intensities (appropriate for Poisson sampling) are different and lead to different MLEs. Let $\boldsymbol{y}$ denote the observed frequency distribution. Then, when the overall effect is not present, the MLE for probabilities does not preserve the sufficient statistics $\mathbf{A}\boldsymbol{y}$, and, for intensities, it does not preserve the observed total $\mathbf{1}'\boldsymbol{y}$, see Example 2.1.

An iterative scaling procedure based on Bregman divergence can be used to compute the MLE under relational models [16]. The Bregman divergence between two distributions is a generalization of the Kullback–Leibler divergence, but, unlike the latter, stays non-negative whether or not the two distributions have the same total. This property is essential for relational models for intensities without the overall effect as these models may include distributions with different totals.

If the observed frequencies are positive and the model matrix is of full row rank, the MLE under relational models can be computed using algorithms for convex optimization, in particular the algorithms of [1,10], or the Newton–Raphson algorithm. Apparently, these algorithms were not formulated to deal with the case when the MLE does not exist in the model because of zero observed frequencies. In addition, instability may occur near the boundary of the parameter space, see the related discussion in [10]. These difficulties can be handled by the application of iterative scaling type algorithms. A detailed discussion of the relative advantages and disadvantages of the latter kind of algorithms was given in [16]. The contribution of the present paper is the investigation of cases when there are observed zero frequencies in the data, and of the closure of relational models under which such data will always admit an MLE. Of course, if only three groups of goods, as in the example above, are investigated, one cannot expect to see an observed zero, but if 1000 groups of goods are investigated, out of the resulting $2^{1000} - 1$ combinations, many will be empty. As it turns out, the pattern of observed zeros has far reaching implications on the existence and kind of MLE obtained.

A necessary and sufficient condition for the existence of the maximum likelihood estimates of the cell parameters under relational models is obtained in Section 2. The MLE for $\boldsymbol{y}$ exists if and only if there is a positive vector $\boldsymbol{z}$ such that $\mathbf{A}\boldsymbol{z} = \mathbf{A}\boldsymbol{y}$. This is literally the same condition as the one that applies to log-linear models.

In Section 3, extended relational models are studied. The extended relational model is defined as the set of distributions parameterized by the elements of an algebraic variety associated with the model matrix of the original relational model. It is shown that this set is equal to the closure of the original model with respect to both the pointwise convergence and the Bregman divergence.

In Section 4, a polyhedral condition for the existence of the MLE in the original or the extended relational model is formulated. If the vector of the sufficient statistics, $\mathbf{A}\boldsymbol{y}$, of the observed distribution is not contained in any of the faces of the polyhedral cone associated with the model matrix, the MLE exists in the original model, and otherwise, it does in the extended model. This condition is the same as for the log-linear case, but the proof is very different. The multiplicative representation of the distributions in the extended model and the existence of the MLEs of the model parameters are also discussed in this section. Finally, the generalized iterative proportional fitting procedure suggested in [16] is extended to the case of observed zeros.

While the conditions of the existence of the MLE in the generality considered in this paper may be formulated to coincide with the known conditions for the case of log-linear models, the proofs turn out to be more involved. Also, the algorithm proposed to obtain the MLEs is more complex. The additional complications come from properties of the MLE when the overall effect is not present. In fact, Lauritzen [18, p. 75] mentioned the existence of models without the overall effect, which he called the "constant function", but to avoid difficulties did not consider them. On the other hand, such models have been used in practice, see references in [17,16].

2. MLE under relational models

Let $Y_1, \ldots, Y_K$ be discrete random variables with finite ranges, and the vector $\mathcal{I}$ of length $|\mathcal{I}|$ be their joint sample space. Here, $\mathcal{I}$ may also be a proper subset of the Cartesian product of the ranges of the variables. A distribution on $\mathcal{I}$ is parameterized by the cell parameters $\boldsymbol{\delta} = \{\delta_i, \text{ for } i \in \mathcal{I}\}$, and, to simplify notation, is identified with $\boldsymbol{\delta}$. The components of $\boldsymbol{\delta}$ are either probabilities: $\delta_i \equiv p_i \in (0, 1)$, with $\sum_{i\in\mathcal{I}} p_i = 1$, or intensities: $\delta_i \equiv \lambda_i > 0$, for all $i \in \mathcal{I}$. Let $\mathcal{P}$ denote the set of positive distributions, $\boldsymbol{\delta} > \mathbf{0}$, on $\mathcal{I}$.