Journal of Multivariate Analysis 140 (2015) 302–316

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Identifiability of a model for discrete frequency distributions with a multidimensional parameter space

Marica Manisera, Paola Zuccolotto *

University of Brescia, C.da S. Chiara 50, 25122 Brescia, Italy

ARTICLE INFO

Article history:
Received 25 July 2014
Available online 6 June 2015

AMS subject classifications:
62-07
62F99

Keywords:
Identifiability
Mixture distributions
Likert scales
Categorical ordinal variables
Rating data
Nonlinear CUB

ABSTRACT

This paper is concerned with the identifiability of models depending on a multidimensional parameter vector, aimed at fitting a probability distribution to discrete observed data, with a special focus on a recently proposed mixture model. Starting from the necessary and sufficient condition derived by the definition of identifiability, we describe a general method to verify whether a specific model is identifiable or not. This procedure is then applied to investigate the identifiability of a recently proposed mixture model for rating data, Nonlinear CUB, which is an extension of a class of mixture models called CUB (Combination of Uniform and Binomial). Formal proofs and a numerical study show that some sufficient conditions for identifiability of Nonlinear CUB are always satisfied, provided that in the estimation procedure one quantity is fixed at a relatively small value.

1. Introduction

In the statistical literature, there are several theoretical models aimed at fitting a probability distribution to discrete observed data, from the classical statistical distributions (e.g., the Binomial or the Poisson model) to the mixture models combining several distributions in a very flexible framework. Among the properties of these models, this paper focuses on the identifiability issue, fundamental for both specification and inference. In the literature, there exist some results on the identifiability of mixture distributions, starting from the seminal works by Teicher [1] and Yakowitz and Spragins [2] who investigated identifiability for large families of finite mixtures. Further results have then been derived by Chandra [3], Al-Hussaini and Ahmad [4] and, more recently, McLachlan and Peel [5], Atienza et al. [6], Allman et al. [7] and Coretto and Hennig [8,9]. However these results, although being valid for a wide set of possible finite mixtures (e.g. mixtures of the following families of distributions: Gaussian, Gamma, rectangular, noncentral χ^2, logistic, generalized logistic, generalized hyperbolic-secant, inverse Gaussian distributions and several others, as well), cannot be applied to all the models proposed in the literature, especially those fitting probability distributions to discrete observed data. As a matter of fact, as observed by Ljung and Glad [10], the issue of model identifiability is often treated on a case by case basis, starting from the set of non-linear equations for the model parameters deriving from the definition of identifiability itself, which can be examined with different tools, sometimes analytical.

The aim of this paper is twofold, so it can be viewed as ideally divided into two parts. In the first part, we formalize the above mentioned general approach for investigating identifiability on a case by case basis, restricting the attention to models conceived for fitting discrete frequency distributions in the case of multidimensional parameter space. It is noteworthy that

* Corresponding author.
E-mail addresses: marica.manisera@unibs.it (M. Manisera), paola.zuccolotto@unibs.it (P. Zuccolotto).

http://dx.doi.org/10.1016/j.jmva.2015.05.011
0047-259X/

the procedure we describe here does not provide general conditions for the identifiability of a particular class of models, but illustrates an approach for proving identifiability for a specific model.

For the formalization of this general procedure, we drew inspiration from Iannario [11], whose study on the identifiability of a particular class of finite mixture models implicitly follows the approach which we describe here in a generalized way. The mixture models analyzed by Iannario [11] are called CUB (the acronym standing for Combination of a shifted Binomial and a discrete Uniform random variables) [12,13], and constitute an original framework addressed to model ordinal data as an alternative to the best-established approaches in the literature (for example, [14,15]).

In the second part of the paper, the general approach formerly described is applied to investigate the identifiability of Nonlinear CUB (NLCUB) models, recently introduced in the literature as a possible generalization in the class of CUB models for modeling rating data [16]. In fact, the study of Iannario [11] on the identifiability of CUB models cannot be directly exploited in the NLCUB framework, because of a more complex formulation.

The paper is organized as follows: Section 2 describes a general approach which can be followed in order to prove the identifiability of specific parametric models for discrete probability distributions. Section 3 briefly describes CUB and NLCUB models, with Section 3.1 focused on the two-step procedure proposed to estimate the parameters of an NLCUB model. In Section 4, by means of the procedure described in Section 2 and a numerical study, we derive the conditions under which an NLCUB model is identifiable. Some open issues are discussed at the end of the section. Section 5 presents some concluding remarks.

2. Identifiability of models for discrete probability distributions

Let us consider a random variable X assuming m discrete real values $x_1, \ldots, x_m$. The assumed probability distribution is $p_x(\boldsymbol{\theta}) = Pr(X = x|\boldsymbol{\theta}), x = x_1, \ldots, x_m$. The parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ has dimension $p < m$. We denote the p-dimensional parameter space by $\Theta = \{\mathcal{D}_{\theta_1} \times \cdots \times \mathcal{D}_{\theta_p}\} \subseteq \mathbb{R}^p$, where $\times$ denotes the cartesian product and $\mathcal{D}_{\theta_j}$, $j = 1, \ldots, p$, is the parameter space of θ_j and consider the case of $p > 1$. The parameter $\boldsymbol{\theta}$ is estimated starting from a random sample of n observations providing the (relative) frequency distribution $f_x, x = x_1, \ldots, x_m$.

In general, the model is identifiable if there is a biunivocal correspondence between the parameter $\boldsymbol{\theta}$ and the probabilities $p_x(\boldsymbol{\theta})$. More specifically, we require that for any two parameter vectors $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}$ such that $\boldsymbol{\theta}^* \neq \tilde{\boldsymbol{\theta}}$, there exists at least one value x in the support of X such that $p_x(\boldsymbol{\theta}^*) \neq p_x(\tilde{\boldsymbol{\theta}})$.

Proposition 1. *A necessary and sufficient condition for identifiability is that, for any parameter vector $\boldsymbol{\theta}^*$, the following system of equations in $\boldsymbol{\theta}$*

$$\begin{cases} p_{x_1}(\boldsymbol{\theta}) = p_{x_1}(\boldsymbol{\theta}^*) \\ p_{x_2}(\boldsymbol{\theta}) = p_{x_2}(\boldsymbol{\theta}^*) \\ \vdots \\ p_{x_{m-1}}(\boldsymbol{\theta}) = p_{x_{m-1}}(\boldsymbol{\theta}^*) \\ p_{x_m}(\boldsymbol{\theta}) = p_{x_m}(\boldsymbol{\theta}^*) \end{cases} \tag{1}$$

admits only one solution in the parameter space Θ (see Appendix A). ◇

The condition in Proposition 1 is sometimes hard to verify in practice. So, we can refer to the more tractable sufficient condition of Proposition 2, motivated by the fact that $p < m$.

Proposition 2. *A sufficient condition for identifiability is that, for any parameter vector $\boldsymbol{\theta}^*$, a system $\mathcal{S}_p$ composed of any p equations of (1) admits only one solution in the parameter space Θ (see Appendix B).* ◇

Without loss of generality, we now restrict to the case $p = 2$, with $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and $\Theta = \{\mathcal{D}_{\theta_1} \times \mathcal{D}_{\theta_2}\}$, the extension to higher dimensional parameter spaces being in general straightforward. So, according to Proposition 2, the sufficient condition for identifiability can be checked by considering the system $\mathcal{S}_2$

$$\begin{cases} p_{x_i}(\boldsymbol{\theta}) = p_{x_i}(\boldsymbol{\theta}^*) \\ p_{x_j}(\boldsymbol{\theta}) = p_{x_j}(\boldsymbol{\theta}^*) \end{cases} \tag{2}$$

where x_i and $x_j, x_i \neq x_j$ are any two values of the domain of X. In some cases, the system (2) can be written in the following form:

$$\begin{cases} \theta_1 = h(\theta_2; \boldsymbol{\theta}^*) \\ f(\theta_2) = k_{\boldsymbol{\theta}^*} \end{cases} \tag{3}$$

where, for a given $\boldsymbol{\theta}^*$, h is a (not multivalued) function of θ_2, f is a function of only θ_2 with domain $\mathcal{D}_f$ and $k_{\boldsymbol{\theta}^*}$ is a real value depending on $\boldsymbol{\theta}^*$ (and varying according to the choice of x_i and x_j).

The function f plays a central role for identifiability, according to the following Propositions 3 and 4. We denote the intersection of the domain of f and the parameter space of θ_2 by $\mathcal{D}_{f\theta_2} \equiv \mathcal{D}_f \cap \mathcal{D}_{\theta_2}$.