Journal of Multivariate Analysis 101 (2010) 1102–1122

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Minimum Hellinger distance estimation in a two-sample semiparametric model

Jingjing Wu [a], Rohana Karunamuni [b,*], Biao Zhang [c]

[a] Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4

[b] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1

[c] Department of Mathematics, University of Toledo, Toledo, OH 43606-3390, USA

ARTICLE INFO

Article history:
Received 3 December 2008
Available online 22 January 2010

AMS 2000 subject classifications:
Primary 62F10
62E20
secondary 60F05

Keywords:
Asymptotic normality
Hellinger distance
Kernel estimator
Two-sample semiparametric model

ABSTRACT

We investigate the estimation problem of parameters in a two-sample semiparametric model. Specifically, let $X_1, \ldots, X_n$ be a sample from a population with distribution function G and density function g. Independent of the X_i's, let $Z_1, \ldots, Z_m$ be another random sample with distribution function H and density function $h(x) = \exp[\alpha + r(x)\beta]g(x)$, where α and β are unknown parameters of interest and g is an unknown density. This model has wide applications in logistic discriminant analysis, case-control studies, and analysis of receiver operating characteristic curves. Furthermore, it can be considered as a biased sampling model with weight function depending on unknown parameters. In this paper, we construct minimum Hellinger distance estimators of α and β. The proposed estimators are chosen to minimize the Hellinger distance between a semiparametric model and a nonparametric density estimator. Theoretical properties such as the existence, strong consistency and asymptotic normality are investigated. Robustness of proposed estimators is also examined using a Monte Carlo study.

1. Introduction

Semiparametric models have continued to receive increasing attention over the years from both practical and theoretical point of views due in large part to the fact that semiparametric models arise frequently in many areas, primarily in biostatistics and econometrics. The well-known semiparametric models include the Cox proportional hazard model in survival analysis, econometric index models, regression models and errors-in-variables models, among many others. More examples and theory on semiparametric models can be found in the monographs [1,2] and in the review articles [3,4].

In this paper, we consider the following two-sample semiparametric model: Let $X_1, \ldots, X_n$ be a random sample from a population with distribution function G and density function g. Independent of the X_i's, let $Z_1, \ldots, Z_m$ be another random sample from a population with distribution function H and density function h. The two unknown density functions g and h are linked by an "exponential tilt" $\exp[\alpha + r(x)\beta]$. Thus, we have

$$\begin{aligned} X_1, \ldots, X_n &\overset{\text{i.i.d.}}{\sim} g(x) \\ Z_1, \ldots, Z_m &\overset{\text{i.i.d.}}{\sim} g(x)\exp[\alpha + r(x)\beta], \end{aligned} \tag{1.1}$$

where $r(x) = (r_1(x), \ldots, r_p(x))$ is a $1 \times p$ vector of functions of x, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ parameter vector, and α is a normalizing parameter that makes $g(x)\exp[\alpha + r(x)\beta]$ integrate to 1. Various choices of $r(x)$ for some conventional distributions are discussed in [5]. In most applications $r(x) = x$ or $r(x) = (x, x^2)$. Note also that the test of equality of G

* Corresponding author.
E-mail address: R.J.Karunamuni@ualberta.ca (R. Karunamuni).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.01.006

and H can be regarded as a special case of model (1.1) with $\alpha = \beta = 0$. We are interested in the estimation problem of parameters α and β when g is unknown (nuisance parameter).

For $r(x) = x$, model (1.1) encompasses many common distributions, including two exponential distributions with different means and two normal distributions with common variance but different means. Furthermore, model (1.1) with $r(x) = x$ or $r(x) = (x, x^2)$ has wide applications in the logistic discriminant analysis [6,7] and in case-control studies [5,8]. Model (1.1) can also be viewed as a biased sampling model with weight function $\exp[\alpha + r(x)\beta]$ depending on the unknown parameters α and β, see [9]. In [10], a goodness-of-fit test is considered for a logistic regression model based on case-control data by employing the maximum semiparametric likelihood estimator of G to test the validity of model (1.1) with $r(x) = x$. In [11], quantiles of G are estimated under model (1.1).

In this paper, we propose MHD estimation for the two-sample semiparametric model (1.1). In fully parametric models, MHD estimators have been shown to achieve efficiency and have excellent robustness properties such as the resistance to outliers and robustness with respect to model misspecification, see [12,13]. Efficiency combined with excellent robustness properties make MHD estimators appealing in practice. For a comparison between MHD estimators with the MLEs and the balance between robustness and efficiency of estimators, see [14]. Moreover, it has been shown that MLE and MHD estimators are members of a larger class of efficient estimators with various second-order efficiency properties [14]. MHD estimation in fully parametric models have been investigated by various authors, including [12,15–22]. MHD estimators for branching processes and for the mixture complexity in a finite mixture model have been studied in [23–25]. However, MHD estimators for semiparametric models have been studied less. A MHD estimator for finite mixtures of Poisson regression models with the distribution of covariates unknown has been investigated in [26]. Recently, a MHD estimator of the mixture parameter for a nonparametric two-component mixture model has been obtained in [27,28]. Apart from the preceding three articles, there has been very little work reported in the literature on the application of the MHD methodology for semiparametric models. In this paper, we extend the implementation of the MHD approach to the two-sample semiparametric model (1.1). Specifically, we construct minimum Hellinger distance estimators of parameters α and β in model (1.1). The proposed estimators are chosen to minimize the Hellinger distance between a semiparametric model and a nonparametric density estimator. Asymptotic properties such as the existence, strong consistency and asymptotic normality of the proposed MHD estimators of α and β are investigated. Robustness of proposed estimators is also examined using a Monte Carlo study.

This paper is organized as follows. In Section 2, we investigate MHD estimators of the parameters α and β and study their existence and strong consistency. In Section 3, we derive the asymptotic distribution of the proposed estimators. Section 4 contains a simulation study where efficiency and robustness properties of the proposed MHD estimators are studied using a Monte Carlo study. A real data example is given in Section 5. A detailed proof of asymptotic normality of the estimators (Theorem 3.2) is given in Section 6.

2. MHD estimators of regression parameters

Define $\theta = (\alpha, \beta^T)^T$, where α and β are as in (1.1). Then the model (1.1) can be written as

$$\begin{aligned} &X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} g(x) \\ &Z_1, \ldots, Z_m \overset{\text{i.i.d.}}{\sim} h_\theta(x), \end{aligned} \tag{2.1}$$

where $h_\theta(x) = g(x)\exp[(1, r(x))\theta]$, $r(x) = (r_1(x), \ldots, r_p(x))$ is a $1 \times p$ vector of continuous functions of x on $\mathbb{R}$, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ parameter vector and α is a normalizing parameter that makes $h_\theta(x)$ integrate to 1. We assume here and in what follows that $\theta \in \Theta$ and Θ is a compact subset of $\mathbb{R}^{p+1}$.

We first define following kernel density estimators of g and h_θ based on the data $X_1, \ldots, X_n$ and $Z_1, \ldots, Z_m$, respectively, of (2.1):

$$g_n(x) = \frac{1}{nb_n}\sum_{i=1}^{n} K_0\left(\frac{x - X_i}{b_n}\right), \tag{2.2}$$

$$h_m(x) = \frac{1}{mb_m}\sum_{j=1}^{m} K_1\left(\frac{x - Z_j}{b_m}\right), \tag{2.3}$$

where K_0 and K_1 are symmetric density functions, bandwidths b_n and b_m are positive constants such that $b_n \to 0$ as $n \to \infty$ and $b_m \to 0$ as $m \to \infty$. We can also employ adaptive kernel density estimators, which use $S_n b_n$ instead of b_n with S_n being a robust scale statistic. Here we use non-adaptive kernel density estimators (2.2) and (2.3) for convenience. The results can be easily extended for adaptive kernel density estimators with some additional conditions on S_n.

Let $\mathcal{H}$ be the set of all densities w.r.t. Lebesgue measure on the real line. For $\phi \in \mathcal{H}$, we define a MHD functional $T_0(\phi)$ as

$$T_0(\phi) = T\left(\{h_\theta\}_{\theta\in\Theta}, \phi\right) = \arg\min_{\theta\in\Theta} \|h_\theta^{1/2} - \phi^{1/2}\|. \tag{2.4}$$