Journal of Multivariate Analysis 139 (2015) 329–346

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semiparametric estimation with missing covariates

Francesco Bravo

Department of Economics, University of York, York YO10 5DD, UK

ARTICLE INFO

Article history:
Received 17 March 2014
Available online 20 April 2015

This paper is dedicated to Spritz

AMS 2013 subject classifications:
62G05
62G20
62G35

Keywords:
Inverse probability weighting
Local linear approximation
Missing at random
Robust estimation

ABSTRACT

This paper considers estimation in semiparametric models when some of the covariates are missing at random. The paper proposes an iterative estimator based on inverse probability weighting and local linear estimation of the nonparametric component. The resulting estimator is very general and can be used in the context of semiparametric maximum likelihood, quasi likelihood and robust estimation. The paper establishes the asymptotic normality of the estimator using both nonparametric and parametric estimation of the unknown probability weights. Two general examples illustrate the theory and Monte Carlo simulations show that the proposed estimator has good finite sample properties.

1. Introduction

In this paper we consider estimation of a general class of semiparametric models which include as special cases the generalised partially linear varying coefficients model of Lu [26], the generalised partially linear single index model of Carroll et al. [10] and the robust generalised partially linear model of Boente et al. [8]. These semiparametric models have been applied in many different areas such as economics, finance, biostatistics and medical statistics, and can be estimated using different methods and nonparametric estimators: for example Ahmad et al. [1] and Fan and Huang [14] consider semiparametric least squares estimation with, respectively, nonparametric series and kernels, Carroll et al. [10] and Lu [26] use semiparametric quasi maximum likelihood with kernels, while Boente et al. [8], Bianco et al. [4] and Hu and Cui [19] consider robust semiparametric estimation with, respectively, kernels and sieves. All of these estimators are derived under the assumption that all of the data in the sample are observable. However data are frequently missing, especially in medical and biostatistics, and ignoring this fact or simply excluding the missing data, the so-called complete case analysis, may result in inconsistent and/or inefficient estimators, with possibly a great loss of information.

In this paper we propose an estimator for generalised partially linear *index structure* models. We define an index structure a smooth real valued known function that relates the nonparametric parameter to a set of covariates and possibly a set of additional unknown finite dimensional parameters. This structure nests together and generalises the single index, multiple single index and varying coefficient specifications that are widely used in the semiparametric literature. New examples of index structures are the nonlinear varying coefficient model and the partially parametric interaction model given, respectively, in (4) and (6). The resulting estimator can be used to estimate the unknown parameters of a large number of semiparametric models including all of those mentioned above. More importantly it can be used when outliers are present

E-mail address: francesco.bravo@york.ac.uk.
URL: https://sites.google.com/a/york.ac.uk/francescobravo.

http://dx.doi.org/10.1016/j.jmva.2015.03.012
0047-259X/

and/or when some of the covariates are missing. As noted by Boente et al. [8], Bianco et al. [4] and others, outliers (both in terms of large deviations of the response from its (conditional) mean – as measured for example by the Pearson residuals – and of outlying values of the covariates) can negatively affect the estimation of the nonparametric component and thus the estimation of the parametric component as well. To deal with potential outliers we use a real valued function that downweights high leverage covariates and allow for a robustified objective function, such as that considered by Cantoni and Ronchetti [9] or by Bianco et al. [5], which yields estimators with bounded influence functions. To deal with missing covariates we assume that they are missing at random (MAR henceforth). MAR is commonly assumed in many statistical models with missing data – see Little and Rubin [25] for a comprehensive review – and it specifies that the probability of missing – often called selection probability – depends on variables that are always observed. To deal with the MAR covariates we use the inverse probability weighting (IPW henceforth) method [18], which has been used in a number of statistical models with missing data including regressions, see for example Robins et al. [34] and Robins and Rotnitzky [33], treatment effect estimation, see for example Hirano et al. [17], and nonclassical measurement error models, see for example Chen et al. [11].

The semiparametric estimator we propose is a two-step iterative one based on an IPW objective function. In the first step we use the local linear estimator (see Fan and Gijbels [13]) to estimate the nonparametric component. In the second step we use the estimates obtained in the first step to replace the unknown nonparametric parameter and estimate the parametric components globally. These two steps are then iterated until convergence. This type of iterative estimation is often used in the semiparametric literature, see for example Carroll et al. [10], Liang [23], Lu [26] among others. The resulting estimator is fairly general and can be applied in the context of semiparametric maximum likelihood, quasi likelihood and general M estimation, as well as robust semiparametric estimation including robust deviance and robust quasi likelihood. For example it can be applied to semiparametric regression models (see for example Ruppert et al. [36]), to robust semiparametric generalised linear models (see for example Boente et al. [8]) and to semiparametric misspecified likelihood based linear models (i.e. models in which the second Bartlett identity does not hold) with MAR covariates.

We note that recently Qin et al. [32] proposed a semiparametric estimator based on the same IPW objective function as that used in this paper. They considered a generalised partially linear model and showed the asymptotic normality of the finite dimensional parameter using sieve estimation (regression splines). The use of sieve estimation in the context of the more general model considered here would be an interesting alternative to the estimation procedure of this paper.

In this paper we make the following contributions: First we establish the asymptotic normality of the estimators of the nonparametric and parametric components considering both a parametric and a nonparametric specification for selection probabilities. The resulting distributions are characterised by a complicated covariance matrix, which however can be consistently estimated using a weighted bootstrap procedure, that is well suited for the type of data considered in this paper. Second we show that estimation of the selection probabilities leads to more efficient estimators. However for the estimator of the nonparametric component, efficiency gains are possible only if the selection probabilities are estimated nonparametrically. Third we illustrate the main results of the paper by considering two general examples: semiparametric quasi likelihood and semiparametric (robust deviance) estimation with missing covariates. Fourth we use simulations to assess the finite sample properties of the proposed estimation method. Finally we show that when all the covariates are observables and no outliers are present the proposed estimator is semiparametric efficient in the sense of Bickel et al. [7]. These results generalise and/or complement a number of results including those obtained by Carroll et al. [10], Liang et al. [24], Fan and Huang [14], Boente et al. [8], Liang [23], Lu [26], Croux and Haesbroeck [12], Hu and Cui [19], Qin et al. [32] and others.

The rest of the paper is structured as follows: Section 2 introduces the statistical model and the estimators. Section 3 contains the main results of the paper; Section 4 presents the two main examples while Section 5 reports the results of the simulation study. All the proofs and further simulations evidence can be found in the supplemental material (see Appendix A).

The following notation is used throughout the paper: "$'$" and "*diag*" denote, respectively, transpose and block diagonal matrix, "$\otimes$" and "*vec*" are the standard Kronecker and the vec operator, $\|\cdot\|$ is the Euclidean norm and finally for any vector $v v^{\otimes 2} = v v'$.

2. The model and the estimator

We consider a statistical model where the response variable Y is related to a set of covariates X, Z, and U by a generalised partially linear index structure $\eta\left(X'\beta + \iota\left(Z, \alpha\left(U, \theta\right)\right)\right)$ where $\eta : \mathcal{X} \times \mathcal{Z} \times \mathcal{U} \times B \times A \times \Theta \to \mathbb{R}$ is a *known* smooth function, $\iota : \mathcal{Z} \times \mathcal{U} \times A \times \Theta \to \mathbb{R}$ represents a *known* index structure discussed below, $\mathcal{X} \subset \mathbb{R}^k$, $\mathcal{Z} \subset \mathbb{R}^p$, $\mathcal{U} \subset \mathbb{R}^q$, β and θ are, respectively, a k and an l dimensional vectors of unknown parameters, $\alpha\left(U, \theta\right)$ is a p dimensional vector of unknown functions depending on the covariates U and possibly θ. To explicitly emphasise the generalised partially linear index structure of the statistical model let

$$\zeta\left(Y, \eta\left(X'\beta + \iota\left(Z, \alpha\left(U, \theta\right)\right)\right)\right)\omega\left(X, Z\right) \tag{1}$$

denote a real valued objective function, where $\omega\left(\cdot\right)$ is a real valued function that is equal to 1 in case of standard estimation or downweights high leverage covariates in case of robust estimation, see for example (18). To simplify the notation let $\eta\left(X'\beta + \iota\left(Z, \alpha\left(U, \theta\right)\right)\right) := \eta\left(\beta, \alpha, \theta\right)$ and $\zeta\left(\cdot\right)\omega\left(X, Z\right) = \zeta_\omega\left(\cdot\right)$.

In the absence of outliers, possible specifications of the objective function $\zeta\left(\cdot\right)$ include a conditional (log) density of the response given the covariates or a quasi-likelihood function. For example $\zeta\left(\cdot\right)$ could be a member of the canonical