Journal of Multivariate Analysis 147 (2016) 183–201

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Penalized empirical likelihood for high-dimensional partially linear varying coefficient model with measurement errors

Guo-Liang Fan [a], Han-Ying Liang [b,*], Yu Shen [b]

[a] School of Mathematics & Physics, Anhui Polytechnic University, Wuhu 241000, PR China

[b] Department of Mathematics, Tongji University, Shanghai 200092, PR China

ARTICLE INFO

Article history:
Received 22 August 2015
Available online 4 February 2016

AMS 2010 subject classifications:
62G08
62G20

Keywords:
Partially linear varying coefficient model
Measurement error
High-dimensional data
Variable selection
Penalized empirical likelihood

ABSTRACT

For the high-dimensional partially linear varying coefficient models where covariates in the nonparametric part are measured with additive errors, we, in this paper, study asymptotic distributions of a corrected empirical log-likelihood ratio function and maximum empirical likelihood estimator of the regression parameter. At the same time, based on penalized empirical likelihood (PEL) approach, the parameter estimation and variable selection of the model are investigated, the proposed PEL estimators are shown to possess the oracle property. Also, we introduce the PEL ratio statistic to test a linear hypothesis of the parameter and prove it follows an asymptotically chi-square distribution under the null hypothesis. Simulation study and real data analysis are undertaken to evaluate the finite sample performance of the proposed methods.

1. Introduction

Flexible and refined statistical models are widely sought in a large array of statistical problems. Much of the work has been devoted to deriving estimators and tests for the unknown parameters or functions. Recently, [7] introduced the following partially linear varying coefficient errors-in-variables (or measurement errors) models (PLVCEVM):

$$\begin{cases} Y_i = Z_i^\top \beta + X_i^\top \alpha(U_i) + \varepsilon_i, \\ W_i = X_i + \upsilon_i, \end{cases} \quad i = 1, \ldots, n, \tag{1.1}$$

where $\{(Y_i; W_i, Z_i, U_i), 1 \le i \le n\}$ is an independent and identical distributed (i.i.d.) random sample, $Z_i \in \mathbb{R}^p$, $U_i \in \mathbb{R}$, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is an unknown p-dimensional parameter vector, $\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_q(\cdot))^\top$ is an unknown q-dimensional vector of coefficient functions, ε_i are random errors and the covariate variables $X_i \in \mathbb{R}^q$ are measured with additive errors and are not directly observable. One can only observe the surrogate variables W_i, and the measurement errors υ_i with mean zero and known covariance Σ_υ are independent of $(X_i, Z_i, U_i, \varepsilon_i)$. When Σ_υ is unknown, one can obtain a consistent and unbiased estimator based on W_i. The specific details can be found in [16].

Model (1.1) is flexible enough to include a variety of models of interest. For example, when $\upsilon_i \equiv 0$, i.e. X_i can be observed exactly and $q = 1, X_i = 1$, model (1.1) reduces to the famous partially linear regression model. When X_i can be observed exactly, model (1.1) reduces to well-known partially linear varying coefficient model (PLVCM) which was introduced by [2] and wildly studied by many authors. For example, [13] proposed a local least-squares method with a kernel weight function

* Corresponding author.
E-mail address: hyliang@tongji.edu.cn (H.-Y. Liang).

http://dx.doi.org/10.1016/j.jmva.2016.01.009
0047-259X/

and established the consistency and asymptotic normality of estimator; [31] developed the procedures for estimation of the linear and the nonparametric parts based on local polynomial fitting method; [30] employed the empirical likelihood (EL) approach for deriving the confidence regions of the unknown parametric vector; more references can be found in [32,4,17,28] among others. For the PLVCEVM where the measurement errors emerged in parametric part, a lot of investigative effort has been dedicated to this case in the literature. For example, [29] proposed estimators for the parametric and nonparametric component respectively and derived their asymptotic properties; [23] applied EL method to obtain confidence regions for the parametric component; [25] proposed modified profile least-squares estimator for the parametric component and established the asymptotic normality of the estimator. A test was also developed for the validity of the constraints on the parametric component. For the PLVCEVM (1.1), that is the measurement errors emerged in nonparametric part, we can see, so far, that [7] proposed two types of local bias-corrected restricted profile least squares estimators of the parameter and varying coefficient functions and discussed their asymptotic properties, they also compared the efficiency of the two kinds of parameter estimators under the criterion of Loẅner ordering and developed a linear hypothesis test for the parametric component; [6] employed the EL approach to construct confidence intervals/regions both for the parametric and nonparametric components.

However, all articles mentioned above are the cases with fixed number of predictors. It is well-known that high-dimensional data analysis arises frequently in many contemporary statistical studies. The emergence of high-dimensional data, such as the gene expression values in microarray and the single nucleotide polymorphism data, brings challenges to many traditional statistical methods and theory. One important aspect of the high-dimensional data under the regression setting is that the number of covariates is diverging. When the number of parameters grows with sample size, much of the work has been devoted to various parametric and semiparametric models. For example, [33] investigated parameter estimation in a semiparametric regression model with diverging number of predictors that are highly correlated; [34] considered the stable direction recovery in single-index models with a diverging number of predictors; [26] investigated asymptotic properties of a family of sufficient dimension reduction estimators when the number of predictors diverges to infinity with the sample size; see also [1,9,35] among others. For high-dimensional PLVCM, i.e., the model (1.1) when X_i can be observed exactly and p diverges, [15] employed the EL method to construct confidence regions of the unknown parameter.

It is well known that variable selection for predictors has caused much attention of many researchers. Various powerful penalization methods have been developed for variable selection. For example, [3] proposed a family of variable selection procedures for parametric models via nonconcave penalized likelihood. [14] developed the nonconcave penalized quasi-likelihood method for variable selection in PLVCM. [10] proposed adaptive penalization methods for variable selection in PLVCM and proved that the methods possess the oracle property. Recently, a new and efficient variable selection approach, penalized empirical likelihood (PEL) introduced for the first time by [21], was applied to analyze mean vector in multivariate analysis and regression coefficients in linear models with diverging number of parameters. As demonstrated in [21], the PEL has merits in both efficiency and adaptivity stemming from a nonparametric likelihood method. Also, the PEL method possesses the same merit of the EL which only uses the data to determine the shape and orientation of confidence regions and without estimating the complex covariance. As far as we know, there are a few papers related to the PEL approach, such as [12] applied the PEL approach to parametric estimation and variable selection for general estimating equations; [27] used the PEL method to study linear regression model with right censored data.

It is worth pointing out that there is no result available in the literature when the number of covariates in PLVCEVM (1.1) is diverging. In this paper we focus on this model. Our contribution includes the following three aspects: (1) utilize the EL method to construct confidence regions of unknown parameter and establish asymptotic normality of maximum empirical likelihood (MEL) estimator of the parameter; (2) employ the PEL approach to build parsimonious and robust models and obtain the oracle property of the PEL estimator; (3) apply the PEL ratio for testing hypotheses and constructing confidence regions of finite-dimensional subsects of unknown parameter.

The organization of the paper is as follows. In Section 2, we construct corrected EL ratio and test statistic as well as define the MEL and PEL estimators of the parameter and give their asymptotic properties; at the same time, we briefly introduce some computational issues. Simulation study and real data analysis are done in Section 3. The proofs of the main results are collected in the Appendix.

2. Methodology and main results

2.1. Corrected empirical likelihood and asymptotic properties

If X_i can be observed exactly and β is known, the model (1.1) can be reduced to the following varying coefficient regression model

$$Y_i - \sum_{j=1}^{p} Z_{ij}\beta_j = \sum_{j=1}^{q} X_{ij}\alpha_j(U_i) + \varepsilon_i, \quad 1 \le i \le n. \tag{2.1}$$

Then, the varying coefficient functions $\{\alpha_j(\cdot), j = 1, \ldots, q\}$ can be estimated by a local linear regression approximation. Specifically, for U in a neighborhood of u, we use a local linear approximation $\alpha_j(U) \approx \alpha_j(u) + \alpha_j'(u)(U - u) \equiv$