Journal of Multivariate Analysis 149 (2016) 13–29

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Confidence intervals for high-dimensional partially linear single-index models

Thomas Gueuning, Gerda Claeskens *

ORSTAT and Leuven Statistics Research Center, Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 9 July 2015
Available online 2 April 2016

AMS subject classifications:
62F25
62F12
62G08
62J07

Keywords:
High-dimensional data
Single-index model
Regularized estimation
Sparsity
Asymptotic normality
Confidence interval

ABSTRACT

We study partially linear single-index models where both model parts may contain high-dimensional variables. While the single-index part is of fixed dimension, the dimension of the linear part is allowed to grow with the sample size. Due to the addition of penalty terms to the loss function in order to provide sparse estimators, such as obtained by lasso or smoothly clipped absolute deviation, the construction of confidence intervals for the model parameters is not as straightforward as in the classical low-dimensional data framework. By adding a correction term to the penalized estimator a desparsified estimator is obtained for which asymptotic normality is proven. We study the construction of confidence intervals and hypothesis tests for such models. The simulation results show that the method performs well for high-dimensional single-index models.

1. Introduction

A partially linear single-index model is a semi-parametric model which can be written as $Y = \eta(Z^\top\alpha) + X^\top\beta + \epsilon$, where η is a one-dimensional unknown function and Z, X are covariate vectors of dimension p and q, respectively, such that the mean zero random error ϵ is independent of (Z, X). The underlying idea is that when the linearity assumption may not be valid for all covariates the introduction of an unknown function allows to overcome this problem. We focus on the high-dimensional case where the number of covariates $p + q$ may exceed the sample size n. In this paper, we consider p as fixed and allow q to grow with n.

We show how to construct a desparsified version of a penalized estimator of the high-dimensional parameters (α, β) of the partially linear single-index model, and we establish the asymptotic distribution of this desparsified estimator. The main purpose of desparsifying a penalized estimator that is obtained under the assumption of sparsity is that the construction of confidence intervals becomes thus possible for *all* of the components of the parameter vector, also for the ones asymptotically consistently estimated by zero due to the penalized estimation and which result in a point mass at zero in the asymptotic distribution. van de Geer et al. [23] introduce such desparsified lasso estimators in the context of linear and generalized linear models and obtain their asymptotically normal distribution. This is done by writing the Karush–Kuhn–Tucker conditions which define the lasso estimator. It is then possible to find a de-biased lasso estimator and to characterize its asymptotic distribution under suitable conditions including a stricter sparsity condition. The construction

* Corresponding author.
E-mail addresses: thomas.gueuning@kuleuven.be (T. Gueuning), gerda.claeskens@kuleuven.be (G. Claeskens).

http://dx.doi.org/10.1016/j.jmva.2016.03.007
0047-259X/

of confidence intervals for the model parameters is then straightforward. Note that for linear models, their method is the same one as that from [31]. Javanmard and Montanari [10] also provide a desparsified estimator in the context of the linear model. The main difference with the approach of van de Geer et al. [23] is that they do not make any assumption on the sparsity level of the precision matrix.

Waldorp [24] used the desparsified lasso for comparing high-dimensional graphical models. He obtained desparsified estimators for the coefficients of the precision matrices of the graphical models and constructed hypothesis tests based on the asymptotic distribution of these estimators. Lu et al. [16] also used the desparsifying idea for constructing confidence bands for a class of nonparametric sparse additive models.

The use of a single index as opposed to a full p-variate nonparametric function estimation effectively circumvents the curse of dimensionality. In a low dimensional setting, the (partially linear) single-index model has been studied by Carroll et al. [1], Horowitz [7], Xia et al. [28], Yu and Ruppert [29] and Xia and Härdle [27], among others. The fitting process involves estimation of the parameters and of the unknown function η. Different fitting methods have been introduced, most of them use kernel smoothing. Liang et al. [14] and Wang et al. [25] used the local linear regression technique introduced by Fan and Gijbels [4] to estimate η. The resulting estimators have good theoretical properties regarding consistency and convergence rates. To deal with high dimensional covariate vectors, Liang et al. [14] used a smoothly clipped absolute deviation penalty (SCAD, see [5]). They obtain a profile least-squares function to minimize, similarly to the linear model case. Ma and Zhu [17] study such models under heteroscedasticity.

Our goal is to provide confidence intervals for the high-dimensional parameter vector in the partially linear single index model and to determine which conditions (design, sparsity, etc.) are necessary to make this construction possible, the challenging part being to be able to tackle the presence of the unknown function η. In a fixed dimension framework, Zhu and Xue [32] introduced the empirical likelihood to construct confidence regions. Further, Zhang et al. [30] developed a dimension reduction approach for estimation in a partially linear single-index model (PLSIM) with diverging number of parameters in both the linear and the single-index part but needed the strong condition $\max(p, q) = o(n^{1/3})$ excluding the $p + q > n$ case. Our method for constructing confidence intervals and regions works in the high-dimensional framework $p + q > n$ with p fixed (potentially larger than n) and q growing with n.

First, in Section 2, we describe a method for estimating a partially linear single-index model. Our main results are in Section 3 where we construct a desparsified estimator and study its asymptotic distribution. Section 4 describes the construction of confidence intervals and regions together with ways to perform hypothesis testing following from the theoretical study. Section 5 deals with computational choices. Further, Section 6 reports on simulation studies to assess the finite sample performance of the desparsified estimator. In Section 7 we illustrate our method on a dataset to study the Bardet–Biedl disease in a rat population. This disease is linked to genetic mutations and also affects humans, provoking several dysfunctions. The dataset comprises of 120 observations and 200 variables. Section 8 concludes. All proofs are contained in Section 9.

2. Estimation for partially linear single-index model

Let $\{(Y_i, Z_i, X_i), i = 1, \ldots, n\}$ be a sample generated by the partially linear single-index model

$$Y = \eta(Z^\top \alpha_0) + X^\top \beta_0 + \epsilon,$$

where $\eta : \mathbb{R} \to \mathbb{R}$ is an unknown differentiable function, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\alpha_0 \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}^q$ are the regression parameters and Z and X are, respectively, p-dimensional and q-dimensional covariate vectors. The error term ϵ and the covariates Z, X are independent. For identifiability reason, we assume that $\|\alpha_0\| = 1$ and that the first non-zero entry of α_0 is positive. We consider p as fixed and allow q to grow with n. The case $p + q > n$ corresponds to the high-dimensional data framework. We denote by $\xi = (\alpha^\top, \beta^\top)^\top$ the $p + q-$ dimensional vector of parameters.

Several estimation approaches have been introduced in the literature. In this paper, we use the profile least-squares procedure presented in [14]. For the paper to be self-contained, we here summarize the main ingredients. This approach uses the local linear regression technique to estimate η, that is, an estimator of $\eta(u)$ is obtained by the minimization of

$$\sum_{i=1}^{n} \{a + b(Z_i^\top \alpha - u) + X_i^\top \beta - Y_i\}^2 K_h(Z_i^\top \alpha - u), \tag{1}$$

with respect to a and b, where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function and h is a bandwidth. The minimizer $(\hat{a}, \hat{b})$ of (1) is an estimator of $(\eta(u), d\eta(u)/du)$. It can be shown (see [4]) that

$$\hat{\eta}(u, \xi) = \hat{a} = \frac{K_{20}(u, \xi)K_{01}(u, \xi) - K_{10}(u, \xi)K_{11}(u, \xi)}{K_{00}(u, \xi)K_{20}(u, \xi) - K_{10}^2(u, \xi)},$$

where $K_{jl}(u, \xi) = \sum_{i=1}^{n} K_h(Z_i^\top \alpha - u)(Z_i^\top \alpha - u)^j (X_i^\top \beta - Y_i)^l$ for $j = 0, 1, 2$ and $l = 0, 1$.

Now, for every data point, we have an estimator $\hat{\eta}(Z_i^\top \alpha; \xi)$ of $\eta(Z_i^\top \alpha)$ and, in the low-dimensional case where $p + q < n$, we can obtain a profile least-squares estimator $\hat{\xi} = (\hat{\alpha}, \hat{\beta})$ by the minimization of

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \{Y_i - \hat{\eta}(Z_i^\top \alpha; \xi) - X_i^\top \beta\}^2, \tag{2}$$