Statistics and Probability Letters 83 (2013) 2601–2606

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Using thresholding difference-based estimators for variable selection in partial linear models

June Luo *, Patrick Gerard

Department of Mathematical Sciences, Clemson University, United States

ARTICLE INFO

Article history:
Received 27 June 2013
Received in revised form 23 August 2013
Accepted 23 August 2013
Available online 28 August 2013

Keywords:
Semi-parametric model
Differencing method
Asymptotic
High dimension
Variable selection

ABSTRACT

A commonly used semiparametric model is considered. We adopt two difference based estimators of the linear component of the model and propose corresponding thresholding estimators that can be used for variable selection. For each thresholding estimator, variable selection in the linear component is developed and consistency of the variable selection procedure is shown. We evaluate our method in a simulation study and implement it on a real data set.

1. Introduction

Semiparametric models have received considerable attention in statistics. For example, a semiparametric approach was discussed in Mukherjee and Pozo (2011) for economics data and was used in Huang et al. (2005) for microarray data. In these models, some of the relations are believed to be of certain parametric form while others are not easily parameterized. In this paper, we consider the following semiparametric model:

$$Y_i = X_i\beta + f(U_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $X_i \in R^{p_n}$, $U_i \in R^1$, $\beta = (\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_{p_n})$ is a vector of parameters, $f(\cdot)$ is an unknown function and the ϵ_i's are independent and identically distributed random error with mean 0 and variance σ^2. We also assume that ϵ_i's are independent of (X_i', U_i).

There have been several methods, such as these discussed in Ahn and Powell (1993), Liang et al. (1999), Fan and Huang (2001) and Lam and Fan (2008), developed to construct estimators of the linear component in (1). A penalized least-squares method was used in Wahba (1984), Engle et al. (1986) and Chen and Shiau (1991). A partial residual method was proposed in Cuzick (1992). By using higher order differences, Yatchew (1997, 2000) showed that the bias induced from the presence of the nonparametric component can be essentially eliminated, and thus the linear component can be estimated without requiring an estimator of the unknown function f. This differencing method has been used widely in the analysis of semiparametric models. For example, a difference-based ridge estimator was discussed in Tabakan and Akdeniz (2010) and Luo (2012). In particular, the asymptotic distribution of a difference-based estimator was extensively discussed in Wang et al. (2011).

Following the literature of variable selection in high dimensions, we assume that β vector is sparse. That is the number of non-zero components in β is finite. We will extend the ideas of the differencing method in Wang et al. (2011) for the purpose

* Corresponding author.
E-mail addresses: jluo@clemson.edu, juneluo1110@gmail.com (J. Luo).

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.08.011

of variable selection. A thresholding estimator is proposed to identify significant variables in the linear part of model (1). Our method is shown to be asymptotically efficient in the sense that all significant variables will be detected and all insignificant variables will be detected.

Even though Wang et al. (2011) obtained an optimal convergence rate of their difference-based estimator, Tabakan and Akdeniz (2010) argued that a difference-based ridge estimator could have smaller mean square error than the estimator in Wang et al. (2011) under certain conditions. We will also define a thresholding estimator using a difference-based ridge estimator. This thresholding estimator can also be used for variable selection and is shown to be asymptotically efficient.

In summary, given the asymptotic distribution of the difference-based estimator in Wang et al. (2011) and the discussion in Tabakan and Akdeniz (2010), we further propose two thresholding estimators of the linear component in model (1). The proposed estimators can identify the non-zero coefficients in the linear part of model (1). We will show consistency of the variable selection methods and demonstrate their usefulness through a real data and a simulation study.

2. The thresholding estimator

We consider a fixed design version of the semiparametric model (1) with $U_i = i/n$. The covariates X_i in the linear component are assumed to be random. Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip_n})'$ be p_n-dimensional independent random vectors with the covariance matrix Σ_X. We will start from reviewing the estimator in Wang et al. (2011). Define a Lipschitz ball $\Lambda^\alpha(M)$ as

$$\Lambda^\alpha(M) = \{g : \text{ for all } 0 \le x, y \le 1, k = 0, \ldots, \lfloor\alpha\rfloor - 1, |g^{(k)}(x)| \le M, \text{ and } |g^{(\lfloor\alpha\rfloor)}(x) - g^{(\lfloor\alpha\rfloor)}(y)| \le |x-y|^{\alpha - \lfloor\alpha\rfloor}\}$$

where $\lfloor\alpha\rfloor$ is the largest integer less than α. Let $f \in \Lambda^\alpha(M)$ for some $\alpha > 0$.

Suppose a difference sequence $d_1, d_2, \ldots, d_{m+1}$ satisfies

$$\sum_{i=1}^{m+1} d_i = 0 \quad \text{and} \quad \sum_{i=1}^{m+1} d_i^2 = 1. \tag{2}$$

Such a sequence is called an mth order difference sequence. Let $c_k = \sum_{i=1}^{m+1-k} d_i d_{i+k}$ and define $\lambda_m = \sum_{k=1}^{m} c_k^2$. We now consider the difference-based estimator of β. Let

$$D_i = \sum_{t=1}^{m+1} d_t Y_{i+m+1-t}, \quad i = 1, 2, \ldots, n-m-1.$$

Then

$$D_i = Z_i\beta + \delta_i + \omega_i, \quad i = 1, 2, \ldots, n-m-1,$$

where $Z_i = \sum_{t=1}^{m+1} d_t X_{i+m+1-t}$, $\delta_i = \sum_{t=1}^{m+1} d_t f(U_{i+m+1-t})$, and $\omega_i = \sum_{t=1}^{m+1} d_t \epsilon_{i+m+1-t}$. When written in the matrix form, this becomes

$$D = Z\beta + \delta + \omega$$

where $D = (D_1, D_2, \ldots, D_{n-m-1})'$, $\omega = (\omega_1, \omega_2, \ldots, \omega_{n-m-1})'$ and Z is a matrix whose ith row is given by Z_i. In Wang et al. (2011), when $f \in \Lambda^\alpha(M)$, ignoring both the deterministic errors δ_i and correlation among the random errors, ω_i, they estimated β by ordinary least squares as

$$\hat\beta = (Z'Z)^{-1}Z'D. \tag{3}$$

The proof of Theorem 1 in Wang et al. (2011) shows that

$$\sqrt{n}(\hat\beta - \beta) \sim N(0, \sigma^2(1 + 2\lambda_m)\Sigma_X^{-1}) \tag{4}$$

if $\alpha > 0$, $m \to \infty$ and $m/n \to 0$ for any $p_n < n - m$. Since $n \to \infty$, the dimension p_n is allowed to increase with n so long as $p_n < n - m$.

As an alternative choice, a difference-based ridge estimator $\hat\beta(k) = (Z'Z + kI)^{-1}Z'D$ was discussed in Tabakan and Akdeniz (2010). Theorem 3.1 in Tabakan and Akdeniz (2010) shows that $\hat\beta(k)$ is MSE superior over the $\hat\beta$ in (3) if and only if

$$\beta' W^{-1}\beta \le \sigma^2, \tag{5}$$

where

$$W = (Z'Z)^{-1}(T'Z)'(T'Z)(Z'Z)^{-1} + \frac{1}{k}((Z'Z)^{-1}(T'Z)'(T'Z) + (T'Z)'(T'Z)(Z'Z)^{-1}), \quad \text{with}$$

$$T = \begin{bmatrix} d_1 & d_2 & \ldots & d_{m+1} & & 0 & \ldots & 0 \\ 0 & d_1 & d_2 & \ldots & d_{m+1} & 0 & \ldots & 0 \\ \vdots & & & & & & & \\ 0 & \ldots & 0 & 0 & d_1 & d_2 & \ldots & d_{m+1} \end{bmatrix}_{(n-m)\times n}.$$