Journal of Statistical Planning and Inference 162 (2015) 1–19

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Spline estimation and variable selection for single-index prediction models with diverging number of index parameters

Guannan Wang [a,*], Li Wang [b]

[a] Department of Statistics, The University of Georgia, United States

[b] Department of Statistics and the Statistical Laboratory, Iowa State University, United States

ARTICLE INFO

Article history:
Received 9 May 2013
Received in revised form 12 August 2014
Accepted 22 January 2015
Available online 20 February 2015

Keywords:
B-spline
Diverging parameters
SCAD
Semiparametric regression
Weakly dependent data

ABSTRACT

Single-index models are useful and fundamental tools for handling "curse of dimensionality" problems in nonparametric regression. Along with that, variable selection also plays an important role in such model building process when the index vectors are high-dimensional. Several procedures have been developed for estimation and variable selection for single-index models when the number of index parameters is fixed. In many high-dimensional model selection problems, the number of parameters is increasing along with the sample size. In this work, we consider weakly dependent data and propose a class of variable selection procedures for single-index prediction models, which are robust against model misspecifications. We apply polynomial spline basis function expansion and smoothly clipped absolute deviation penalty to perform estimation and variable selection in the framework of a diverging number of index parameters. Under stationary and strong mixing conditions, the proposed variable selection method is shown to have the "oracle" property when the number of index parameters tends to infinity as the sample size increases. A fast and efficient iterative algorithm is developed to estimate parameters and select significant variables simultaneously. The finite sample behavior of the proposed method is evaluated with simulation studies and illustrated by the river flow data of Iceland.

1. Introduction

For the past two decades, high dimensional problem is becoming increasingly popular in many scientific areas including biostatistics, medicine, economics and financial econometric. When the dimension of covariates is getting higher, one unavoidable issue is the "curse of dimensionality", which refers to the poor convergence rate. Lots of efforts have been devoted to tackle of this difficulty. As an attractive dimension reduction method, single-index models (SIMs) play a useful and fundamental role for handling "curse of dimensionality" problems. Various intelligent estimators of the single-index coefficients have been derived by lots of researchers. Examples can be found in Powell et al. (1989), Härdle and Stoker (1989), Carroll et al. (1997), Xia and Li (1999) and Hristache et al. (2001). Xia et al. (2002) introduced the minimum average variance estimation (MAVE) for several index vectors. Wang and Yang (2009) proposed the polynomial spline estimator for the single-index prediction model which is more robust against deviations from SIMs. Chang et al. (2010) studied the

* Corresponding author.
E-mail addresses: guannan@uga.edu (G. Wang), lilywang@iastate.edu (L. Wang).

http://dx.doi.org/10.1016/j.jspi.2015.01.007
0378-3758/

SIMs with heteroscedastic errors and recommended an estimating equation method in terms of transferring restricted least squares to un-restricted least squares. Zhang et al. (2010) derived inference for the index parameters by the local linear method. Cui et al. (2011) suggested an estimating function method to study the SIMs.

Along with the SIMs, when the index vectors are high-dimensional, variable selection for significant predictors is very practical in such model building process. For example, in time series modeling, we often need to select significant explanatory lagged variables. Most traditional variable selection procedures, such as Akaike's information criterion (AIC), Mallow's C_p and the Bayesian information criterion (BIC), use a fixed penalty on the size of a model. To overcome the inefficiency of traditional variable selection procedures, Fan and Li (2001) proposed a unified approach via non-concave penalized likelihood and demonstrated that penalized likelihood estimators are asymptotically as efficient as the ideal "oracle" estimator for certain penalty functions, such as the smoothly clipped absolute deviation (SCAD) penalty. Fan and Peng (2004) further extended the method to the situation with a diverging number of parameters, which substantially enlarges the scope of applicability of the shrinkage methods. We refer to Fan and Peng (2004), Huang et al. (2008) and Wang et al. (2012) for more works in the high-dimensional framework where the number of covariates increases with the sample size.

Several procedures have been developed for estimation and variable selection for SIMs when the number of index parameters is fixed. Examples include the dissected cross-validation (DCV) method in Kong and Xia (2007), the profile least squares (PrLS) estimation procedure in Liang et al. (2010), the adaptive lasso with kernel smoothing in Zhu et al. (2011), the penalized least squares method in Peng and Huang (2011), and the lasso with local linear smoothing method in Zeng et al. (2012). Unfortunately, in practice, many variables can be introduced to reduce possible modeling biases. In many high-dimensional model selection problems, the number of introduced variables depends on the sample size, which reflects the ensilability of the parametric problem. For example, when running regressions on time-series data, it is often important to include many lagged values of the dependent variable as predictor variables. Sometimes, to capture the persistence of a time series, the lag length can be very long, or even close to the length of time series.

When a diverging number of predictors are involved in SIM, Zhu and Zhu (2009) proposed a method based on slice inverse regression (SIR) to select variables. However, the SIR based method imposes a strong assumption on the predictors: the distribution of the covariates need to be elliptically symmetric distributions. In time series analysis, usually, the covariates are the lagged values of a time series. As discussed in Xia et al. (2002), the elliptical symmetry of the covariates implies the time series itself is time reversible (Tong, 1990), which is an exception feature in time series analysis, therefore, their method would not work for many time series data; see the discussions in Xia et al. (2002) and Peng and Huang (2011).

In this work, we consider weakly dependent data and focus on variable selection and estimation for single-index prediction models introduced by Wang and Yang (2009), which are robust against model misspecifications. We apply the SCAD penalty and polynomial spline basis function expansion to perform variable selection and estimation simultaneously in the framework of a diverging number of index parameters. Under a mixing condition and some other regularity conditions, the proposed variable selection method is shown to have the "oracle" property when the number of parameters diverges as the sample size increases. A fast and efficient algorithm is developed to estimate parameters and select significant variables simultaneously. Our method is applicable to selecting significant variables when modeling time series data which may include endogenous variables (lagged variables) as well as exogenous variables.

The rest of the paper is organized as follows. Section 2 first provides the background of the single-index prediction model, then introduce the polynomial spline smoothing and the penalized SCAD estimators. Section 3 shows the main theoretical results in the framework of a diverging number of index parameters. Section 4 presents an algorithm to implement the proposed method. Section 5 reports our findings in three simulation studies. The proposed method is applied in Section 6 to the river flow data of Iceland. Section 7 provides concluding remarks and discussion. All technical proofs are given in the Appendix.

2. Methodologies

2.1. Single-index prediction model

Let $\{X_i, Y_i\}_{i=1}^n$ be a length n realization of a $(d+1)$-dimensional (strictly) stationary process with $X_i = \{X_{i,1}, \ldots, X_{i,d}\}$ being R^d valued ($d \geq 1$) and Y_i being real valued. In particular, X_i may consist of lagged values of Y_i, and X_i can also include some exogenous variables. Let $m(x) = E(Y_i|X_i = x)$, $x \in R^d$, be the d-variate regression function. We assume $\{X_i, Y_i\}_{i=1}^n$ follow the heteroscedastic model

$$Y_i = m(X_i) + \sigma(X_i)\,\varepsilon_i, \quad m(X_i) = E(Y_i|X_i), \; i = 1, 2, \ldots, n,$$

in which $E(\varepsilon_i|X_i) = 0$, $E\left(\varepsilon_i^2|X_i\right) = 1$. The function $\sigma(\cdot)$ is an unknown standard deviation of the response Y_i conditional on the predictor vector X_i. In what follows, let $\left(X^T, Y, \varepsilon\right)$ have the stationary distribution of $\left(X_i^T, Y_i, \varepsilon_i\right)$.

It is well known that nonparametric estimation suffers from the "curse of dimensionality". One way to overcome the difficulty is to impose some structure on the unknown regression function m. For example, the single-index models assume that $m(x) = g(x^T\theta_0)$. If the model is misspecified, i.e., m is not a genuine single-index function, the estimation of θ_0 might be biased and a goodness-of-fit test is often needed in this case. In this paper, instead of presuming that underlying true