Journal of Multivariate Analysis 143 (2016) 56–70

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

# Separation of linear and index covariates in partially linear single-index models[☆]

Heng Lian [a], Hua Liang [b,*]

[a] School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW 2052, Australia

[b] Department of Statistics, George Washington University, Washington, DC 20052, USA

ARTICLE INFO

*Article history:*
Received 18 September 2012
Available online 11 September 2015

*AMS 2000 subject classification:*
primary 62G08
secondary 62G10
62G20
62J02
62F12

*Keywords:*
Estimating equation
Identifiability constraint
Single-index model
Structure identification

ABSTRACT

Motivated to automatically partition predictors into a linear part and a nonlinear part in partially linear single-index models (PLSIM), we consider the estimation of a partially linear single-index model where the linear part and the nonlinear part involves the same set of covariates. We use two penalties to identify the nonzero components of the linear and index vectors, which automatically separates covariates into the linear and nonlinear part, and thus solves the difficult problem of model structure identification in PLSIM. We propose an estimation procedure and establish its asymptotic properties, which takes into account constraints that guarantee identifiability of the model. Both simulated and real data are used to illustrate the estimation procedure.



## 1. Introduction

The partially linear single-index model (PLSIM) is a very flexible class of semiparametric models given by

$$Y = g(W^T\alpha) + Z^T\beta + \epsilon, \tag{1}$$

where $W$ and $Z$ are two different sets of covariates, and $\epsilon$ is the noise with $E(\epsilon|W, Z) = 0$, $E(\epsilon^2|W, Z) = \sigma^2(W, Z)$. The popularity of (1) can be attributed to its dimension reduction ability to avoid fitting a multivariate nonparametric regression function. It contains single-index models (when $\beta = 0$) [8,6,12] and partially linear models (when $W$ is a scalar) [3,13,7] as special cases. However, in practice, PLSIM met with the problem of having to determine a priori which covariates make up $W$ and which make up $Z$. In many works, including Carroll et al. [1] and Xia and Härdle [17], what has been done is to put discrete predictors in $Z$ and continuous ones in $W$. While it might be a reasonable attempt when no other obvious alternatives exist, this practice is worrisome with no formal justification. Although specification of a linear part in PLSIM can result in more efficient estimation (root-$n$ convergence rate) and a more easily interpretable model, which are the main

☆ Liang's research was partially supported by NSF grants DMS 1440121 and DMS-1418042, Award Number 11529101, made by National Natural Science Foundation of China.

* Corresponding author.
*E-mail address:* hliang@gwu.edu (H. Liang).

http://dx.doi.org/10.1016/j.jmva.2015.08.017

motivations of using PLSIM, this practical difficulty has prevented its applicability in many situations. Thus it is important to investigate whether there is a principled way to deal with the model specification problem.

An extended version of PLSIM was first proposed in [18] which uses $W$ and $Z$ together in both the linear part and the nonlinear part of (1), resulting in

$$Y = g(X^T\alpha) + X^T\beta + \epsilon, \tag{2}$$

where $X = (W^T, Z^T)^T = (X_1, \ldots, X_p)^T$ contains all $p$ predictors. Compared to the single-index model, the extra linear component $X^T\beta$ makes the model more flexible. It is also a special case of additive-index models with only two indices and one of the components is assumed to be linear. The two constraints $\|\alpha\| = 1$ with its first nonzero component positive, and $\alpha^T\beta = 0$ are sufficient to make the model identifiable [18,11], although it causes some additional burden in estimation as we see below. However, (2) gets rid of the headache of partitioning the covariates. When $\alpha = (\alpha^{(1)T}, 0)^T$ and $\beta = (0, \beta^{(2)T})^T$ (here $\alpha^{(1)}$ is $s$-dimensional, say, and $\beta^{(2)}$ is $(p-s)$-dimensional), then the extended PLSIM reduces to (1) (note that we automatically have $\alpha^T\beta = 0$ in this case). The key observation is that if the true model is indeed in the form of (1) and if we can identify the zero components in $\alpha$ and $\beta$ in (2), the predictors are automatically separated into a linear part and a nonlinear part. More formally, we say a covariate $X_j$ has linear effect if $X_j$ only appears in the linear part and nonlinear effect otherwise (in particular, if $X_j$ appears in both nonlinear and linear part we still say it has nonlinear effect).

As an immediate consequence of our asymptotic results later, if the true model is of the form (1), we will be able to recover it with probability approaching one by starting from the extended model (2) and adding variable selection mechanism. In this respect, the purpose of the current study is reminiscent of that of Zhang et al. [21], where structure identification in partially linear additive models is the goal. In that paper, the authors started with general additive models and attempted to shrink the component functions both to zero functions and to linear functions (using two penalties respectively), with the latter achieved by decomposing each component function as a sum of a linear function and another nonlinear function orthogonal to the linear function. Shrinking the nonlinear function to zero will make the final estimate linear. However, in partially linear single-index models, as we discussed above, identification of partially linear structure can be achieved in a much more direct way. Superficially, the formulated problem is very similar to the problem studied in [10]. However, the approach taken by Liang et al. [10] requires correct specification of the linear and nonlinear part. On the contrary, our approach can overcome this limitation by uncovering the true partially linear structure automatically. Note that only variable selection in extended PLSIM can produce the effect of model structure identification as a by-product, while variable selection for the usual PLSIM (as previously considered in [10]) does not achieve this goal obviously. Mathematically, the constraint $\alpha^T\beta = 0$ need to be taken into account for theoretical analysis and empirical implementation, which makes our analysis more complicated.

Variable selection via penalization has become very popular in recent years, starting from the pioneering work of Tibshirani [14] which introduced lasso. In this paper, we use the smoothly clipped absolute deviation (SCAD) penalty which possesses the oracle property. Other penalties could be used such as adaptive lasso penalty [22] or MCP [20], which is expected to possess similar theoretical and numerical properties. The rest of the article is organized as follows. In Section 2, we consider estimation procedure of the penalized extended PLSIM. We also present asymptotic theoretical properties of our estimator. Computational issues are discussed in Section 3. In Section 4, simulations and a real data application are presented to illustrate the numerical property of the proposed model. We conclude our study in Section 5.

## 2. Penalized extended PLSIM

We assume $X = (X_1, \ldots, X_p)$ can be partitioned into four groups $X = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)})^T$ and the true model is

$$Y_i = g(X_i^{(1,2)T}\alpha) + X_i^{(1,3)T}\beta + \epsilon_i,$$

where $X_i^{(1,2)} = (X_i^{(1)T}, X_i^{(2)T})^T$ and $X_i^{(1,3)T} = (X_i^{(1)T}, X_i^{(3)T})^T$. This setup is actually more general than model (1), and thus if the true model is really (1), separation of the linear and index covariates will be achieved by trying to recover the true model. For notational simplicity, we assume the partition obeys the original order of the list of predictors. That is, $X^{(1)}$ is composed of the first $p_1$ predictors in $X$, $X^{(2)}$ the next $p_2$ elements in $X$, etc., with of course $\sum_{j=1}^4 p_j = p$. Here $\alpha = (\alpha^{(1)T}, \alpha^{(2)T})^T$ is $(p_1+p_2)$-dimensional and $\beta = (\beta^{(1)T}, \beta^{(3)T})^T$ is $(p_1+p_3)$-dimensional. Thus in the true model, predictors in $X^{(4)}$ are not related to responses. $X^{(2)}$ only appears in the nonlinear part, $X^{(3)}$ only appears in the linear part, and $X^{(1)}$ appears in both. If $X^{(1)}$ is empty, then the true model is indeed in the form of (1). Thus the true model we consider is actually more general than that presented in (1).

In general, the true model structure (that is how the predictors are partitioned) is unknown. So one must start from the extended PLSIM (2). The two constraints $\|\alpha\| = 1$ (with first nonzero component positive) and $\alpha^T\beta = 0$ in effect reduce the number of parameters from $2p$ to $2p-2$. To take into account the first constraint, we use the popular "delete-one-component" method [19,15,2]. Without loss of generality, we assume the first component of $\alpha$, $\alpha_1$ is positive, and thus we can write $\alpha = ((1-\|\tilde{\alpha}\|^2)^{1/2}, \alpha_2, \ldots, \alpha_p)^T$ where $\tilde{\alpha} = (\alpha_2, \ldots, \alpha_p)^T$ is $\alpha$ without the first component. Similarly we denote $\tilde{\beta} = (\beta_2, \ldots, \beta_p)^T$. Due to the second constraint, we have $\beta_1 = \alpha_1^{-1}(\sum_{j=2}^p \alpha_j\beta_j) = (1-\|\tilde{\alpha}\|^2)^{-1/2}(\sum_{j=2}^p \alpha_j\beta_j)$. Thus $(\alpha, \beta)$