Accepted Manuscript

Variable selection for partially linear models via partial correlation

Jingyuan Liu, Lejia Lou, Runze Li

PII: S0047-259X(17)30556-0
DOI: https://doi.org/10.1016/j.jmva.2018.06.005
Reference: YJMVA 4373

To appear in: *Journal of Multivariate Analysis*

Received date : 13 September 2017

Please cite this article as: J. Liu, L. Lou, R. Li, Variable selection for partially linear models via partial correlation, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.06.005

Variable selection for partially linear models via partial correlation

Jingyuan Liu[a], Lejia Lou[b], Runze Li[c,*]

[a]*Department of Statistics in School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen, Fujian, 361005, China*
[b]*Ernst & Young, 5 Times Square, New York, NY 10036 USA*
[c]*Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802 USA*

Abstract

The partially linear model (PLM) is a useful semiparametric extension of the linear model that has been well studied in the statistical literature. This paper proposes a variable selection procedure for the PLM with ultrahigh dimensional predictors. The proposed method is different from the existing penalized least squares procedure in that it relies on partial correlation between the partial residuals of the response and the predictors. We systematically study the theoretical properties of the proposed procedure and prove its model consistency property. We further establish the root-n convergence of the estimator of the regression coefficients and the asymptotic normality of the estimate of the baseline function. We conduct Monte Carlo simulations to examine the finite-sample performance of the proposed procedure and illustrate the proposed method with a real data example.

Keywords: Model selection consistency, partial faithfulness, semiparametric regression modeling.

1. Introduction

Let y be a response variable, u be a univariate continuous covariate and $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dimensional covariate vector. The partially linear model (PLM) assumes that

$$y = g(u) + \mathbf{x}^\top \boldsymbol{\beta} + \epsilon, \tag{1}$$

where g is an unspecified baseline function, and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. The PLM thus assumes that the regression function linearly depends on the covariates $\mathbf{x}$ while depending nonparametrically on u. This model increases the flexibility of linear models by allowing the intercept to be a nonparametric function of the covariate u. It is one of the most popular semiparametric regression models in the literature [14].

This work aims to develop a variable selection procedure for the PLM with ultrahigh dimensional $\mathbf{x}$, i.e, $p = O\{\exp(n^a)\}$ for some positive constant a, where n is the sample size. PLM estimation has been well studied in the case where p is finite and fixed; see, e.g., [8], [15]. Variable selection procedures have also been developed in this case, e.g., by Fan and Li [6] via penalized least squares, and by Liang and Li [11] who employed the penalized least squares method for variable selection in the PLM in the presence of error in variables. Xie and Huang [18] studied the penalized least squares method with the SCAD penalty [5] for variable selection in the PLM with $p = o(\sqrt{n})$.

In this paper, we propose a new variable selection procedure for PLM. This procedure differs from the aforementioned penalized least squares methods in that it is a partial correlation learning procedure based on the notion of partial faithfulness that was first advocated by Bühlmann et al. [1] for normal linear models and further used for elliptical linear models in [10]. We first utilize partial residual techniques to eliminate the nonparametric baseline function, and then conduct variable selection by recursively testing the partial correlations between the partial residual of the response and that of the linear covariates. That is, we recursively compare the partial correlations with some threshold

*Corresponding author
Email address: rzli@psu.edu (Runze Li)