Neurocomputing 195 (2016) 202–210

Contents lists available at ScienceDirect

Neurocomputing

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

Feature screening and variable selection for partially linear models with ultrahigh-dimensional longitudinal data

Jingyuan Liu

Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, 422 Siming South Road, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 19 February 2015
Received in revised form
30 June 2015
Accepted 17 September 2015
Available online 4 February 2016

Keywords:
Partially linear model
Ultrahigh dimensionality
Longitudinal data
Partial residual two-stage approach
Sure screening property

ABSTRACT

This paper is concerned with longitudinal partially linear models (LPLM) with ultrahigh-dimensional covariates and predictors. As flexible extension of linear regression models by allowing nonparametric intercept function to capture the overall trend over time, the LPLM are expected to be highly potential statistical models for analyzing high-dimensional longitudinal data such as longitudinal genetic data and functional magnetic resonance image data. Feature screening and variable selection are indispensable for LPLM in the presence of ultrahigh-dimensional covariates such as genetic markers and all pixels in image data. This paper proposes a two-stage variable selection procedure that consists of a quick screening stage and a post-screening refining stage, for the ultrahigh dimensional longitudinal partially linear models. The proposed approach is based on the partial residual method for dealing with the nonparametric baseline function. We establish the sure screening property of the proposed screening procedure in the first stage. Simulation results demonstrate the validity of this two-stage method. We further demonstrate the proposed methodology by an empirical analysis of a real data set collected in a soybean plant longitudinal genetic study.

1. Introduction

The partially linear models, as flexible extension of linear models, have been systematically studied in recent years [1]. The advantage of these models over linear models lies in the fact that they allow the intercept to vary with certain covariate, such as time, instead of being fixed as a constant. For instance, to study the gene effect on certain phenotype of plants, the dynamic time effect has to be taken into consideration as any phenotypic trait always experiences a process of development to form its final outcome. Another application is in functional magnetic resonance imaging (fMRI) research, where voxels are also measured over time but the underlying dynamic pattern is unknown. Under such circumstances, the longitudinal partially linear models (LPLM) are usually applied to capture the nonparametric time effect.

However, with the rapid development of data collecting technologies, the LPLM often face the challenge of ultrahigh dimensionality. In the longitudinal genetic study, millions of genes or markers are often studied simultaneously, although only a small amount might be truly important to the phenotype. In the fMRI research, the number of voxels under consideration is also fairly large. Thus a natural question is how to identify the truly relevant predictors in such ultrahigh dimensional LPLM.

In this paper, this question is addressed on the basis of a two-stage procedure: (1) reducing the dimension from ultrahigh to moderate using a fast and efficient independence screening procedure specifically for LPLM and (2) choosing significant variables from the screened submodel by a modified variable selection technique. For the first stage, [2] first advocated the sure independence screening (SIS) method for the ultrahigh dimensional linear models based on the Pearson correlation learning. The sure screening property was proved, meaning that with an overwhelming probability, the true predictors would not be missed by this SIS. Furthermore, due to the fact that the screening procedures rely on the model assumption to a large extent, various methods have been proposed for various models. See [3–7], among others. However, less has been studied for the LPLM. Given the non-linear complexity of the baseline function in these models, we propose a more robust approach, called partial residual sure independence screening (PRSIS), to reduce dimensionality by incorporating the partial residual technique to the nonparametric part. In the second stage, the profile variable selection approach is modified for LPLM to refine the screened submodel from the first stage and to determine the most significant predictors in the LPLM, or more specifically, the most significant genes for complex dynamic traits.

The rest of the paper is organized as follows. The two-stage approach for ultrahigh dimensional LPLM is systematically discussed in Section 2, including the detailed methodology along

E-mail address: jingyuan@xmu.edu.cn

http://dx.doi.org/10.1016/j.neucom.2015.09.122
0925-2312/

with the sure screening property of the first screening stage, and the integrated two-stage algorithm. Then the validity of this method is illustrated through simulation studies in Section 3, and the analysis of the functional mapping in soybean plants is presented in Section 4 to demonstrate its usage. We conclude the paper in Section 5, and provide the proof for the sure screening property of the first screening stage in the Appendix.

2. Methodology

In this section, we present a new two-stage approach, called the partial residual two-stage approach, for the ultrahigh dimensional longitudinal partially linear models (LPLM). A fast screening procedure is proposed in the first stage specifically for LPLM to reduce dimensionality, referred to as partial residual sure independence screening (PRSIS). In the second stage, a modified penalized regression is discussed by cooperating the profile least squared technique to further select important variables.

2.1. Model setting

Suppose the random sample $\{(t_{ik}, \mathbf{x}(t_{ik}), y(t_{ik})),\ i=1,\ldots,n;\ k=1,\ldots,T_i\}$ is from LPLM

$$y(t)=\alpha(t)+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}(t)+\varepsilon(t) \tag{1}$$

where n is the sample size, T_i is the number of observations for subject i, t_{ik} is the time point for the kth observation of the ith subject, $y(t)$ is the response variable, $\mathbf{x}(t)=(x_1(t),\ldots,x_p(t))^{\mathrm{T}}$ is the p-dim covariate vector at time t, $\alpha(t)$ is an unspecified baseline function of t, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^{\mathrm{T}}\in\mathbb{R}^p$ is the ultrahigh p-dim coefficient vector independent of t, and $\varepsilon(t)$ is the stochastic random noise with mean 0. In the genetic study, $\alpha(t)$ captures the dynamic phenotype growth, and the covariate vector $\mathbf{x}(t)$ can either or not depend on the time t, but often not if it refers to the SNP effect vector. Define $x_j(t)$ (or x_j), $j=1,\ldots,p$, as a relevant or important predictor if $\beta_j\neq 0$. The true model is denoted as $\mathcal{M}$, i.e.

$$\mathcal{M}=\{j:\beta_j\neq 0, 1\le j\le p\}. \tag{2}$$

2.2. Partial residual sure independence screening (PRSIS)

We first present a new screening approach, called partial residual sure independence screening (PRSIS) for LPLM in the first stage, to reduce the ultrahigh dimension p of the predictors to a moderate scale $d<n$. Notice that without the nonparametric baseline function $\alpha(t)$, (1) is simply a linear model with longitudinal data structure, and the Pearson correlation can be utilized as a screening criterion [2]. However, as will be shown in the simulation studies, ignoring $\alpha(t)$, if its dynamic pattern indeed exists, would miss some important predictors and thus decrease the power. Therefore, the PRSIS modifies the Pearson correlation in the following fashion. Consider the sample version of the partially linear model (1):

$$y(t_{ik})=\alpha(t_{ik})+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}(t_{ik})+\varepsilon(t_{ik}),\quad i=1,\ldots,n,\ \ k=1,\ldots,T_i. \tag{3}$$

Although the repeated measurements within each of the n individuals are correlated, which is referred to as the within-subject correlation, it can be neglected in the screening stage aimed to reduce dimensionality by implementing a fast and efficient algorithm. Consequently, the data is pooled into the following format:

$$\begin{aligned}
&\mathbf{t}=(t_i)_{N\times 1}=\{t_{11},\ldots,t_{1T_1},t_{21},\ldots,t_{2T_2},\ldots,t_{n1},\ldots,t_{nT_n}\}^{\mathrm{T}}\\
&\mathbf{y}=(y_i)_{N\times 1}=\{y(t_{11}),\ldots,y(t_{1T_1}),\ldots,y(t_{n1}),\ldots,y(t_{nT_n})\}^{\mathrm{T}}\\
&\mathbf{X}=(\mathbf{x}_i^{\mathrm{T}})_{N\times p}=\{\mathbf{x}(t_{11})^{\mathrm{T}},\ldots,\mathbf{x}(t_{1T_1})^{\mathrm{T}},\ldots,\mathbf{x}(t_{n1})^{\mathrm{T}},\ldots,\mathbf{x}(t_{nT_n})^{\mathrm{T}}\}^{\mathrm{T}}
\end{aligned} \tag{4}$$

where $N=\sum_{i=1}^{n}T_i$ is the pooled sample size, t_i and y_i are the ith elements of $\mathbf{t}$ and $\mathbf{y}$, and $\mathbf{x}_i^{\mathrm{T}}$ is the ith row of the pooled design matrix $\mathbf{X}$. Therefore, the model becomes

$$y_i=\alpha(t_i)+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i+\varepsilon_i,\quad i=1,\ldots,N. \tag{5}$$

Inspired by [8,9], we take conditional means for both sides of (5) given t_i,

$$E(y|t_i)=\alpha(t_i)+\boldsymbol{\beta}^{\mathrm{T}}E(\mathbf{x}|t_i)+E(\varepsilon|t_i), \tag{6}$$

where y, $\mathbf{x}$ and ε refer to the generic random variables corresponding to y_i, $\mathbf{x}_i$ and ε_i. Then the nonparametric component $\alpha(t_i)$ is canceled by subtracting (6) from (5):

$$y_{i0}^{*}=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{i0}^{*}+\varepsilon_i^{*},\quad i=1,\ldots,N, \tag{7}$$

where $y_{i0}^{*}=y_i-E(y|t_i)$, $\mathbf{x}_{i0}^{*}=\mathbf{x}_i-E(\mathbf{x}|t_i)$ and $\varepsilon_i^{*}=\varepsilon_i-E(\varepsilon|t_i)$. Therefore, the partially linear model (5) is transformed to the linear model (7), then the marginal importance of the jth predictor of $\mathbf{x}$, denoted by x_j, can be represented by $\rho_j=\mathrm{corr}(x_j-E(x_j|t), y-E(y|t))$. Thus the dynamic effect of t, which is overlooked by the ordinary Pearson correlation $\mathrm{corr}(x_j,y)$ between the original x_j and y, is taken into consideration. In practice, the unknown conditional expectations are estimated by the kernel smoothing method [10]: for any given $t>0$,

$$\widehat{E}(y|t)=\sum_{s=1}^{N}\omega_s(t)y_s,\quad \widehat{E}(x_j|t)=\sum_{s=1}^{N}\omega_s(t)x_{sj}, \tag{8}$$

where $\omega_s(t)=K_h(t_s-t)/\sum_{s=1}^{N}K_h(t_s-t)$ with $K_h(\cdot)=h^{-1}K(\cdot/h)$ and $K(\cdot)$ is a symmetric kernel function. In this paper, the Epanechnikov kernel function is used for simplicity, i.e. $K(t)=0.75(1-t^2)I(|t|<1)$, where $I(E)$ is the indicator function taking value 1 if the statement E is true and 0 otherwise. The bandwidth h is chosen by the plug-in method [11]. In this fashion, $E(y|t_i)$ and $E(\mathbf{x}|t_i)$ are estimated, and hence y_{i0}^{*} and $\mathbf{x}_{i0}^{*}$. Denote the estimates as y_i^{*} and $\mathbf{x}_i^{*}$. The marginal Pearson correlations $\widehat{\rho}_j$'s can then be calculated with the modified sample points y_i^{*} and $\mathbf{x}_i^{*}$, $i=1,\ldots,N$:

$$\widehat{\rho}_j=\frac{(\mathfrak{x}_j^{*}-\overline{\mathfrak{x}}_j^{*})^{\mathrm{T}}(\mathbf{y}^{*}-\overline{\mathbf{y}}^{*})}{\sqrt{(\mathfrak{x}_j^{*}-\overline{\mathfrak{x}}_j^{*})^{\mathrm{T}}(\mathfrak{x}_j^{*}-\overline{\mathfrak{x}}_j^{*})(\mathbf{y}^{*}-\overline{\mathbf{y}}^{*})^{\mathrm{T}}(\mathbf{y}^{*}-\overline{\mathbf{y}}^{*})}},\quad j=1,\ldots,p, \tag{9}$$

where

$$\begin{aligned}
&\mathbf{y}^{*}=(y_i^{*})_{N\times 1}=\left\{y_1-\hat{E}(y|t_{11}),\ldots,y_{T_1}-\hat{E}(y|t_{1T_1}),\ldots,y_N-\hat{E}(y|t_{nT_n})\right\}^{\mathrm{T}}\\
&(\mathbf{x}_i^{*\mathrm{T}})_{N\times p}=(\mathfrak{x}_1^{*},\ldots,\mathfrak{x}_p^{*})=\left\{\mathbf{x}_1^{\mathrm{T}}-\hat{E}(\mathbf{x}^{\mathrm{T}}|t_{11}),\ldots,\mathbf{x}_N^{\mathrm{T}}-\hat{E}(\mathbf{x}^{\mathrm{T}}|t_{nT_n})\right\}^{\mathrm{T}}
\end{aligned} \tag{10}$$

and $\overline{\mathfrak{x}}_j^{*}$ and $\overline{\mathbf{y}}^{*}$ denote the sample means of the N pooled samples.

By ranking $\widehat{\rho}_j$ calculated for all p x-variables, we select the top d predictors and the screened submodel index set is

$$\widehat{\mathcal{M}}=\{j:1\le j\le p, |\widehat{\rho}_j| \text{ ranks among the top } d\}. \tag{11}$$

The submodel size d is taken to be the hard threshold following [12], i.e. $d=\lfloor N^{4/5}/\log(N^{4/5})\rfloor$. And we can always take more conservative d to be $\nu\lfloor N^{4/5}/\log(N^{4/5})\rfloor$ in practice, where ν is an integer larger than 1, to enlarge the probability of selecting all the relevant predictors. One can also adopt the soft threshold in practice. See [7,13] for details. This screening technique is referred as partial residual sure independence screening (PRSIS).

Next, we discuss the sure screening property of PRSIS for the ordinary partially linear models, based on the following two technical conditions that are standard and commonly used in the ultrahigh dimensional literature.

(A1) $E(y|t)$ and $E(\mathbf{x}|t)$, where $\mathbf{x}=(x_1(t),\ldots,x_p(t))^{\mathrm{T}}$, have finite first and second order derivatives over the support of t which is denoted by $\mathbb{U}$.