Computational Statistics and Data Analysis 56 (2012) 2006–2017

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric regression models with additive nonparametric components and high dimensional parametric components

Pang Du [a,*], Guang Cheng [b], Hua Liang [c]

[a] Department of Statistics, Virginia Tech, Blacksburg, VA, United States

[b] Department of Statistics, Purdue University, West Lafayette, IN, United States

[c] Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, United States

ARTICLE INFO

Article history:
Received 22 June 2011
Received in revised form 11 October 2011
Accepted 7 December 2011
Available online 24 December 2011

Keywords:
Additive models
Backfitting
Model selection
Partial smoothing splines
SCAD
Sparsity

ABSTRACT

This paper concerns semiparametric regression models with additive nonparametric components and high dimensional parametric components under sparsity assumptions. To achieve simultaneous model selection for both nonparametric and parametric parts, we introduce a penalty that combines the adaptive empirical L_2-norms of the nonparametric component functions and the SCAD penalty on the coefficients in the parametric part. We use the additive partial smoothing spline estimate as the initial estimate and establish its convergence rate with diverging dimensions of parametric components. Our simulation studies reveal excellent model selection performance of the proposed method. An application to an economic study on Canadian household gasoline consumption reveals interesting results.

Published by Elsevier B.V.

1. Introduction

The last decade has seen the emergence of large data sets with big sets of variables that are more and more commonly collected in modern research studies. This has stimulated vast developments in efficient procedures that can perform variable selection on such large data sets. The first wave of developments focus on parametric models. Examples include the most well-known LASSO estimator (Tibshirani, 1996) and its adaptive version (Zou, 2006), the SCAD estimator (Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007), the sure independence screening (Fan and Lv, 2008), and numerous others. A common feature in these papers is that their models all assume a linear relationship between response and predictors.

Recognizing this limitation for parametric variable selection procedures, various nonparametric models have been used and the associated model selection procedures have been developed in the past years. For example, Lin and Zhang (2006) have proposed the COmponent Selection and Smoothing Operator (COSSO) which can be viewed as a functional generalization of LASSO using the Sobolev norm as the penalty. Taking advantage of the smoothing spline ANOVA framework, the COSSO can perform model selection on non-additive models as well as the additive ones. An adaptive version of it is developed recently by Storlie et al. (2011). Huang et al. (2010) studied variable selection in nonparametric additive models when the number of additive components may diverge with the sample size. Another generalization of the LASSO to nonparametric regression is the sparse additive models (SpAM) proposed in Ravikumar et al. (2009) where the empirical L_2-norm of each additive component function is used. Radchenko and James (2010) later extends the SpAM to incorporate non-additive models with the heredity constraint enforced. Xue (2009) also considered penalizing the empirical norm of

* Corresponding author.
E-mail addresses: pangdu@vt.edu (P. Du), chengg@purdue.edu (G. Cheng), hliang@bst.rochester.edu (H. Liang).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2011.12.007

each component in additive models but used a penalty that generalizes the SCAD instead. Meier et al. (2009) and Koltchinskii and Yuan (2010), despite the difference in the forms of their penalties, both proposed penalties combining the empirical L_2-norm and the usual roughness norm to enforce both sparsity and smoothness. Fan et al. (2011) generalized the sure independence screening for ultra-high dimension regression problems to nonparametric models. In these methods, the predictors are often assumed to be continuous. Although discrete predictors can be included as indicator variables, their corresponding nonparametric effects are essentially of parametric form. Treating them as nonparametric components increases the computational cost and leads to efficiency loss in theory. This motivates us to look into variable selection within the framework of semiparametric models.

In this article, we focus on variable selection in partially linear additive models (PLAM), which are more flexible than parametric models and more efficient than nonparametric models. See Härdle et al. (2004) for a comprehensive survey for PLAM and Härdle et al. (2000) for a survey of partially linear models (PLM), a special case of PLAM, in which there is only one nonparametric component. A lot of efforts have been devoted to variable selection in this area with examples like Liang and Li (2009) for PLM with measurement errors, and Ni et al. (2009) and Xie and Huang (2009) for high-dimensional PLM. Cheng and Zhang (2011) considered similar models in the partial smoothing spline framework with diverging dimension of parametric components but focused only on variable selection in the parametric part. Ma and Yang (2011) proposed a spline-backfitted kernel smoothing method for partially linear additive models which also contain a single nonparametric component and there is no sparsity enforced on the parametric part. Liu et al. (2011) studied variable selection in PLAM when the number of linear covariate is fixed. Wei et al. (2011) considered variable selection in varying coefficient models where coefficient functions are expanded as linear combinations of basis functions and the group LASSO penalty (Yuan and Lin, 2006) was used to enforce sparsity among the coefficient functions. Compared with the existing semiparametric methods, the method we propose innovates in the following aspects: (i) it can perform estimation and variable selection simultaneously on both the nonparametric and parametric components; (ii) the parametric part can have dimensions diverging with the sample size; and (iii) the nonparametric part can have a large but fixed number of additive components.

The initial estimate we use to compute the adaptive weights in the nonparametric part of penalty is the partial smoothing spline estimate with additive nonparametric components. Under certain conditions, we establish the convergence rates of such partial smoothing spline estimates. This extends the classical result in Heckman (1986) that dealt with the case of a single nonparametric component and fixed-dimension parametric components.

To achieve variable selection, we apply double penalties to enforce sparsity in both parametric and nonparametric components. For the parametric components, we use the SCAD penalty (Fan and Li, 2001). For the nonparametric components, we consider an adaptive version of the empirical L_2-norm penalty proposed in the SpAM model of Ravikumar et al. (2009). We choose the SpAM penalty because of its sparsistency and persistence shown in Ravikumar et al. (2009) and the simplicity of its practical implementation. Our adaptive extension of the SpAM penalty computes the weights using the consistent initial estimate mentioned above. This idea is borrowed from the adaptive LASSO of Zou (2006) where a weighted l_1 norm is used to ensure the oracle property for variable selection procedures.

The rest of the article is organized as follows. Section 2 describes the details of the method, including the initial partial smoothing spline estimator and its theoretical properties, the joint variable selection procedures and the computational algorithms, and issues like tuning parameter selection and refitting after variable selection. Section 3 presents simulation studies investigating the performance of the proposed method in prediction, variable selection, and estimation accuracy. Section 4 analyzes the data from an economic study on the Canadian household gasoline consumption.

2. Method

Suppose the observed data are $(\boldsymbol{t}_1, \mathbf{x}_1, y_1), \ldots, (\boldsymbol{t}_n, \mathbf{x}_n, y_n)$, where $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{iq})' \in \mathbb{R}^q$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. We consider the following semiparametric regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sum_{k=1}^{q} f_k(t_{ik}) + \epsilon_i, \tag{2.1}$$

where $\boldsymbol{\beta}$ is an unknown coefficient vector, f_k is an unknown smooth function belonging to the Sobolev space of order m, and ϵ_i is a mean zero error term. For identifiability purpose, assume $\int_0^1 f_k(t)dt = 0$ for each k. In practice, when there is no prior information, an straightforward way to separate the predictors is to treat every continuous variable as one nonparametric component and every categorical variable, or essentially the dummies it generates, as parametric components. More discussion on this topic is in Section 5.

2.1. Initial estimate by additive partial smoothing splines

When $q = 1$ in (2.1), Heckman (1986) has proposed a partial smoothing spline estimate. In this section, we extend her result by defining an additive partial smoothing spline estimate that will be used as the initial estimate for our variable selection procedure to be introduced in Section 2.2.