Semiparametric regression models with additive nonparametric components and high dimensional parametric components [1]

Pang DU, Guang CHENG and Hua LIANG

Summary

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.

KEY WORDS: Additive models; Backfitting; Model selection; Partial smoothing splines; SCAD; Sparsity.

Short Title: Semiparametric additive models

[1] Corresponding Author: Pang Du (pangdu@vt.edu). Pang Du (E-mail: pangdu@vt.edu) is Assistant Professor, Department of Statistics, Virginia Tech, Blacksburg, Virginia. Guang Cheng (E-mail: chengg@purdue.edu) is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, Indiana. Hua Liang (E-mail:hliang@bst.rochester.edu) is Professor, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York 14642. The authors gratefully thank the supports from the NSF grants DMS-0906497 (Cheng), DMS-1007126 (Du), DMS-0806097 and DMS-1007167 (Liang).

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 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. (2010) 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 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 $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{iq})' \in [0,1]^q$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \in [0,1]^p$ and $y_i \in \mathbb{R}$. Here the predictor domain $[0,1]$ is chosen for simplicity. 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.

Our initial additive partial smoothing spline estimate is defined as the following:

$$(\widetilde{\boldsymbol{\beta}}, \widetilde{f}_1, \ldots, \widetilde{f}_q) = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \sum_{k=1}^{q} f_k(t_{ik}) \right)^2 + \lambda \sum_{k=1}^{q} \|f_k\|^2 \right\}, \tag{2.2}$$

where $\|f\| = (\int_0^1 [f^{(m)}(t)]^2 dt)^{1/2}$ is the Sobolev norm of f. For example, $m = 2$ defines the well-known cubic smoothing splines. In the following proposition, we establish the convergence rate of the partial smoothing spline estimate defined in (2.2). Note that we allow the dimension p of parametric components to diverge.

PROPOSITION 2.1. *Suppose that Conditions (C1)-(C6) in the appendix hold. If* $\lambda \sim n^{-2m/(2m+1)}$ *and* $p_n = o(n^{1/2})$*, then the initial solution (2.2) has the following convergence rates:*

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\sqrt{p_n/n}) \tag{2.3}$$

and

$$\|\widetilde{f}_k - f_{0k}\|_2 = O_P(\sqrt{p_n/n} \vee n^{-m/(2m+1)} p_n^2) \tag{2.4}$$

for any $1 \le k \le q$*. Thus,* $\widetilde{f}_k$ *is consistent if we further assume* $p_n = o(n^{m/(4m+2)})$.

2.2 Joint Variable Selection in Nonparametric and Parametric Parts

We first introduce the adaptive SpAM procedure to estimate the function f given the coefficient vector $\boldsymbol{\beta}$. Next we describe the application of the SCAD procedure to the estimation of $\boldsymbol{\beta}$.

Denote the empirical norm of a function as $\|\cdot\|_n$, i.e., $\|f_j\|_n^2 = \sum_{i=1}^n f_j(t_{ij})^2$, $\|y\|_n^2 = \sum_{i=1}^n y_i^2$, and so on. Define f as a shorthand notation for the additive component functions f_j, which are estimated as the minimizer of

$$l_{\boldsymbol{\beta}}(f) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{q} f_j(t_{ij}) - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{q} w_j \|f_j\|_n, \tag{2.5}$$

for $f_j \in \mathcal{H}_j$, where $\mathcal{H}_j$'s are Hilbert spaces of functions, and w_j's are weights ideally chosen in a data-adaptive way. When an initial estimator $\widetilde{f}_j$ is available, a choice of w_j could be $w_j = \|\widetilde{f}_j\|_n^{-\gamma}$ for some $\gamma > 0$. Note that when $w_j = 1, \forall j$ and $\boldsymbol{\beta} = \mathbf{0}$, (2.5) reduces to the SpAM model proposed in Ravikumar et al. (2009). The following proposition provides the motivation for Step 2 of our algorithm to be introduced later.

PROPOSITION 2.2. *The minimizer $f_j \in \mathcal{H}_j$ of (2.5) satisfies*

$$f_j = \left[1 - \frac{\lambda w_j}{\|P_j\|_n}\right]_+ P_j \quad a.s.,$$

where $[\cdot]_+$ denotes the positive part and P_j is the projection of the residual $R_j = Y - \sum_{k\neq j} f_k(\boldsymbol{t}_k) - X'\boldsymbol{\beta}$ onto $\mathcal{H}_j$.

The proof of Proposition 2.2 is a straightforward extension to Theorem 1 in Ravikumar et al. (2009) and thus omitted here. In Ravikumar et al. (2009), the SpAM procedure was implemented by a backfitting algorithm where the P_j's are estimated by smoothing with an orthogonal series smoother. We will use a similar algorithm to be introduced below, but with the smoother replaced by smoothing splines .

Given an estimate $\widehat{f}$ of f, we propose to estimate $\boldsymbol{\beta}$ by minimizing the penalized profile

least squares

$$l_{\widehat{f}}(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n}(y_i - \sum_{j=1}^{q}\widehat{f}_j(t_{ij}) - \mathbf{x}_i'\boldsymbol{\beta})^2 + n\sum_{k=1}^{d_n} p_{\theta_k}(|\beta_k|), \tag{2.6}$$

where $p_{\theta_k}(|\cdot|)$ is the SCAD penalty on $\boldsymbol{\beta}$ (Fan and Li, 2001).

Thus the complete algorithm for our semiparametric variable selection and estimation procedure is as follows.

Step 1. Start with the initial estimate $\widehat{\boldsymbol{\beta}}^{(0)} = \widetilde{\boldsymbol{\beta}}$.

Step 2. Let $\widehat{\boldsymbol{\beta}}^{(k-1)}$ be the estimate of $\boldsymbol{\beta}$ before the kth iteration. Plug $\widehat{\boldsymbol{\beta}}^{(k-1)}$ into (2.5) and solve for f by solving the adaptive SpAM problem of minimizing $l_{\widehat{\boldsymbol{\beta}}^{(k-1)}}(f)$. Let $\widehat{f}^{(k)}$ be the estimate thus obtained.

Step 3. Plug $\widehat{f}^{(k)}$ into (2.6) and solve for $\boldsymbol{\beta}$ by maximizing the penalized profile least square $l_{\widehat{f}^{(k)}}(\boldsymbol{\beta})$. Let $\widehat{\boldsymbol{\beta}}^{(k)}$ be the estimate thus obtained.

Step 4. Replace $\widehat{\boldsymbol{\beta}}^{(k-1)}$ in Step 2 by $\widehat{\boldsymbol{\beta}}^{(k)}$ and repeat Steps 2 and 3 until convergence to obtain the final estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{f}$.

Our experience shows that the algorithm usually converges quickly within a few iterations. This echoes with a theoretical result in Cheng (2011), which states that a few iterations are sufficient to obtain the efficient and sparse estimate in the semiparametric models.

Step 2 solves an adaptive SpAM problem. We use the following backfitting algorithm modified from Ravikumar et al. (2009).

Initialize $\widehat{f}_k = 0$ *for* $k = 1, \ldots, q$ *and then iterate the following steps until convergence.*

For each $k = 1, \ldots, q$:

(2a) Compute the residuals: $R_{ik} = y_i - \sum_{j\neq k}\widehat{f}_j(t_{ij}) - \mathbf{x}_i'\boldsymbol{\beta}$

(2b) Estimate the projection P_k by smoothing: $\widehat{P}_k = S_k R_k$, where S_k is the smoothing matrix obtained from fitting a nonparametric smoothing spline regression model (Gu, 2002) to the "data" $(t_{ik}, R_{ik}), i = 1, \ldots, n$.

(2c) Estimate the norm: $\widehat{s}_k^2 = \frac{1}{n}\sum_{i=1}^n \widehat{P}_{ik}^2$.

(2d) Soft-threshold: $\widehat{f}_k = [1 - \lambda w_k/\widehat{s}_k]_+ \widehat{P}_k$, where $w_k = \|\widetilde{f}_k\|_n^{-\gamma}$.

(2e) Center: update $\widehat{f}_k$ by $\widehat{f}_k - \text{mean}(\widehat{f}_k)$.

Step 3 involves the nonconcave SCAD penalty, whose optimization can't be handled by standard procedures. We use the following one-step approximation procedure modified from Zou and Li (2008). Write the design matrix as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. Let $A = \{j : p'_{\theta_j}(|\widehat{\beta}_j^{(k-1)}|) = 0\}$ and $B = \{j : p'_{\theta_j}(|\widehat{\beta}_j^{(k-1)}|) > 0\}$. Decompose $\mathbf{X}$ and the new estimate $\widehat{\boldsymbol{\beta}}^{(k)}$ accordingly such that $\mathbf{X} = [\mathbf{X}_A, \mathbf{X}_B]$ and $\widehat{\boldsymbol{\beta}}^{(k)} = (\widehat{\boldsymbol{\beta}}_A^{(k)'}, \widehat{\boldsymbol{\beta}}_B^{(k)'})'$.

(3a) Let $\mathbf{y}^* = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k-1)}$ and create the matrix $\mathbf{X}^*$ by replacing the j-th column of $\mathbf{X}$ with $\mathbf{x}_j^* = \mathbf{x}_j \frac{\theta_j}{p'_{\theta_j}(|\widehat{\beta}_j^{(k-1)}|)}$ for each $j \in B$.

(3b) Let $H_A = \mathbf{X}_A^*(\mathbf{X}_A^{*'}\mathbf{X}_A^*)^{-1}\mathbf{X}_A^{*'}$ be the projection matrix to the column space of $\mathbf{X}_A^*$. Compute $\mathbf{y}^{**} = \mathbf{y}^* - H_A\mathbf{y}^*$ and $\mathbf{X}_B^{**} = \mathbf{X}_B^* - H_A\mathbf{X}_B^*$.

(3c) Apply the LARS algorithm in Efron et al. (2004) to solve

$$\widehat{\boldsymbol{\beta}}_B^* = \arg\min_{\boldsymbol{\beta}}\{\frac{1}{2}\|y^{**} - \mathbf{X}_B^{**}\boldsymbol{\beta}\|^2 + n\sum_{j\in B}\theta_j|\beta_j|\}.$$

(3d) Compute $\widehat{\boldsymbol{\beta}}_A^* = (\mathbf{X}_A^{*'}\mathbf{X}_A^*)^{-1}\mathbf{X}_A^{*'}(\mathbf{y}^* - \mathbf{X}_B^*\widehat{\boldsymbol{\beta}}_B^*)$ to obtain $\widehat{\boldsymbol{\beta}}^* = (\widehat{\boldsymbol{\beta}}_A^{*'}, \widehat{\boldsymbol{\beta}}_B^{*'})'$.

(3e) For $j \in A$, set $\widehat{\beta}_j^{(k)} = \widehat{\beta}_j^*$. For $j \in B$, set $\widehat{\beta}_j^{(k)} = \widehat{\beta}_j^* \frac{\theta_j}{p'_{\theta_j}(|\widehat{\beta}_j^{(k-1)}|)}$.

2.3 Miscellaneous Issues

2.3.1 Regularization parameter selection

We now address the issue of selecting the regularization parameters λ and $\boldsymbol{\theta}$. We select $\boldsymbol{\theta}$ by a K-fold cross validation as in Zou and Li (2008). Particularly, we take $K = 5$ in this

paper. The parameter λ is selected by minimizing the GCV score

$$\mathrm{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{k=1}^{q}\widehat{f}_j(t_{ik}) - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})^2}{(1 - \frac{1}{n}\sum_{k=1}^{q}\mathrm{trace}(S_k)I(\|\widehat{f}_k\|_n \neq 0))^2}.$$

We also modified the C_p criterion proposed in Ravikumar et al. (2009) for selecting λ. Since our simulations show similar results, we choose not to present it here.

2.3.2 Refit after variable selection

As demonstrated in many other settings, penalized variable selection procedures can over-shrink component estimates. To account for this, a common practice is performing a refit using standard non-penalized estimation procedures after the variable selection step. In our case, this corresponds to a partial smoothing spline fit using the selected variables. Our simulations in Section 3 show such refitting indeed improves the final estimation performance.

3 Simulation Studies

In the simulations, we generated our data from the semiparametric regression in (2.1) with $p = 30$ and $q = 10$. For the parametric part, we set the first 8 entries of $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_1 = (1, 0.8, 1.4, 0.6, 1.2, 0.9, 1.1, 1.2, \mathbf{0}_{22}')'$ and $\boldsymbol{\beta}_2 = (0.6, 0, 0, 1, 0, 0, 0.8, 0, \mathbf{0}_{22}')'$, where $\mathbf{0}_{22}$ is the vector of zeros with length 22. The $\mathbf{x}_i$'s were generated from the multivariate normal distribution with zero mean and $\mathrm{Cov}(x_{ij}, x_{ik}) = 0.5^{|j-k|}$. For the nonparametric part, the true functions were set to be

$$f_1(t) = -2\sin(2\pi t), f_2(t) = 12x^2 - 11x + 1.5, f_3(t) = 2x - 1, f_4(t) = 9e^{-(x-0.3)^2} - 8.03,$$

and $f_k(t) \equiv 0$ for $k \geq 5$. Note that all the f_k's integrate to 0 on $[0, 1]$ for the identifiable purpose. The t_{ik}'s were generated independently from the uniform distribution on $[0, 1]$. The random errors ϵ_i were generated from the standard normal distribution. For either of the above two settings of $\boldsymbol{\beta}$, we simulated 100 data sets, each with the sample size $n = 200$ or $n = 500$.

For a prediction procedure $\mathcal{M}$ and the estimator $(\hat{\boldsymbol{\beta}}_{\mathcal{M}}, \hat{f}_{\mathcal{M}})$ obtained from the procedure, an appropriate measure of prediction performance is the empirical prediction error $\mathrm{PE}(\hat{\boldsymbol{\beta}}_{\mathcal{M}}, \hat{f}_{\mathcal{M}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \sum_{j=1}^{q}\hat{f}_{\mathcal{M},j}(t_{ij}))^2$. Here $\{(y_i, \mathbf{x}_i, \mathbf{t}_i) : i = 1, \ldots, N\}$ were a test data set independently generated from the true model. We used $N = 1000$ for all the reported simulations. The relative model error (RPE) of $\mathcal{M}_1$ versus $\mathcal{M}_2$ is defined as the ratio $\mathrm{PE}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_1}, \hat{\eta}_{\mathcal{M}_1})$ /$\mathrm{PE}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_2}, \hat{\eta}_{\mathcal{M}_2})$. The oracle procedure $\mathcal{M}_0$ is used as a benchmark. In $\mathcal{M}_0$, the contributing predictors in both nonparametric and parametric parts are known and a partial smoothing spline model is fitted with such supplied information. Note that $\mathcal{M}_0$ can be implemented only in simulation and is unrealistic in practice. We compare performance of the following procedures through their RPEs versus $\mathcal{M}_0$:

$\mathcal{M}_A$: the proposed semiparametric selection procedure;

$\mathcal{M}_B$: the proposed semiparametric selection procedure followed by a refitting step where a standard partial smoothing spline model with the covariates selected by $\mathcal{M}_A$ is used;

$\mathcal{M}_C$: procedure with all the covariate effects assumed to be of linear form. A SCAD procedure is applied to all coefficients;

$\mathcal{M}_D$: procedure $\mathcal{M}_C$ followed by a refitting step where a standard linear regression model with the covariates selected by $\mathcal{M}_C$ is used;

Procedures $\mathcal{M}_C$ and $\mathcal{M}_D$ misspecify the effects of $\mathbf{t}$ to be linear. We intend to show that it is important to correctly specify the effects of all the covariates. In all the procedures, five-fold cross validation was used to select a common $\theta = \theta_1 = \cdots = \theta_{d_n}$ for the SCAD penalty that minimizes the mean sum of squared prediction errors on a grid of $\log(\theta) = -5$ to 5 by 0.1. In procedures $\mathcal{M}_A$ and $\mathcal{M}_B$, the parameter λ was selected to minimize the GCV score in Section 2.3.1 on a grid of $\log(\lambda) = -4$ to 0 by 0.5. The partial smoothing spline model was fitted by the `ssanova` function provided in the contributed R package `gss`.

We had four combinations with $n = 200, 500$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$. For each combination, we simulated 100 data replicates and computed the following quantities: the mean and standard

deviation of the 100 RPEs of the oracle procedure $\mathcal{M}_0$ versus procedures $\mathcal{M}_A$ to $\mathcal{M}_D$, the proportion of being selected for each of the nonzero linear covariates, the average number of noise linear variables selected, the proportion of being selected for each of the nonzero nonlinear covariates, the average number of noise nonlinear variables selected.

The prediction results are summarized in Table 1. From Table 1, we can see that when the nonlinear effects are misspecified as linear, the prediction performances of the procedures $\mathcal{M}_C$ and $\mathcal{M}_D$ are quite poor when compared with the procedures $\mathcal{M}_A$ and $\mathcal{M}_B$ which have correctly specified effects. Also, when model is misspecified, the prediction performance does not improve with an additional round of refitting or even an increase in sample size. But there are notable improvements from $\mathcal{M}_A$ to $\mathcal{M}_B$ under all the four settings, and from the settings with $n = 200$ to the settings with $n = 500$.

For variable selection performance, only procedures $\mathcal{M}_A$ to $\mathcal{M}_C$ are relevant. The variable selection results for the parametric components are summarized in Table 2. A notable fact is that the procedure $\mathcal{M}_C$ with misspecified effects actually has competitive performance to $\mathcal{M}_A$ in selecting the right set of parametric components in all the four settings. This echoes with a similar finding in Ma and Du (2011). We also note that the procedure $\mathcal{M}_A$ tends to include fewer noise variables than $\mathcal{M}_C$ when $n = 500$. The variable selection results for the nonparametric component are summarized in Table 3. Recall that in procedure $\mathcal{M}_C$, the nonlinear effects were misspecified to be linear. Clearly, Procedure $\mathcal{M}_A$ consistently performs better than $\mathcal{M}_C$ in picking up the signal variables more frequently and including fewer noise variables. Also, there is a notable improvement when sample size increases from 200 to 500. Another interesting pattern shown here is that $\mathcal{M}_C$ actually does a decent job in picking up $f_1(t_1)$, $f_3(t_3)$ and $f_4(t_4)$, despite poor selection frequencies of $f_2(t_2)$. An examination of the shapes of f_k's in Figure 1 reveals a possible reason for this. The function f_3 is designed to be linear. Although f_1 and f_4 are nonlinear, the closest linear functions to them should show significant positive and negative slopes respectively. On the other hand, the closest linear function to f_2 is quite flat. Hence, a linear variable selection procedure would have no problem in picking up f_1, f_3 and f_4 but would have difficulty in detecting f_2.

We then investigated the estimation performance of the procedures $\mathcal{M}_A$ and $\mathcal{M}_b$. Table 4 contains the mean and standard deviation of the 100 $\boldsymbol{\beta}$ estimates for each setting. Figure 1 plots the connected point-wise means, and 0.05 and 0.95 quantiles of the function estimates based on the 100 fits for the setting of $n = 200$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_1$. The plots of the other settings show similar patterns and hence not presented here. In general, the estimation performance of both procedures are pretty good with notable improvements in $\mathcal{M}_B$ over $\mathcal{M}_A$ and when sample size increases.

4 Real Data Example: Household Gasoline Demand in Canada

An economic study in Yatchew and No (2001), hereafter YN, investigated the household gasoline demand in Canada. The data were obtained from the National Private Vehicle Use Survey conducted by Statistics Canada between October 1994 and September 1996 and contain household-based information on vehicles, fuel consumption patterns, and demographics. The study focused on households with a nonzero number of licensed drivers, vehicles, and distance driven. After removing 29 observations that belong to categories containing a very small number of observations, we were left with 6201 observations. Following YN, the dependent variable we used in our analysis was the logarithm of total distance driven by each household, which loyally represented the pattern shown in gasoline consumption. The continuous predictors that formed the nonparametric components included: age of a representative household member (`age`), logarithm of the household income (`linc`), total fuel expenses in a month (`paid`), and monthly average price of gasoline per liter (`price`). Thirty one dummy variables, which served as the parametric components, were created from the following categorical variables with the last category always being the baseline: number of licensed driver in the household (range from 1 to 4, `driver1` to `driver3`), number of vehicles in the household (range from 1 to 3, `veh1` and `veh2`), size of the household (range from 1 to 6, `hhsize1` to `hhsize5`), month of the year (`month1` to `month11`), type of fuel (4 types, `fueltype1` to

`fueltype3`), urban/rural indicator (`urban`), province of the household (5 levels, respectively, Atlantic provinces, Quebec, Ontario, the Prairie provinces, and British Columbia, `prov1` to `prov4`), and year of the observation (1994 to 1996, `year94` and `year95`).

We fitted the proposed method without and with refitting to this data set. At the variable selection step, the parameter θ in the SCAD penalty was selected by the five-fold cross validation on a grid of $\log(\theta) = -1$ to 1 by 0.01, and the parameter λ was selected on a grid of $\log(\lambda) = -7$ to -3 by 0.5. Table 5 contains the coefficient estimates for the parametric components. The function estimates for the nonparametric component are plotted in Figure 2. Based on our simulation results in Section 3, the estimates from the method with refitting are more accurate and reliable, although the estimates from the method without refitting can give a rough idea of the patterns to be expected in the refitted estimates. Hence, our discussion below focus only on the refitted estimates. Also, whenever applicable we compare our findings with those in YN.

Let's first look at the nonparametric effect estimates shown in Figure 2. The estimated age effect agrees with the finding in YN. distance driven is the highest in the early 20's, declining steadily to age 40 and then remains approximately flat afterwards. It's interesting to see that the effect of log(INCOME) actually shows a linear pattern. Nevertheless, its increasing trend agrees with the positive coefficient estimate in YN where log(INCOME) was a linear predictor. Distance driven shows a general increasing trend against total fuel expenses. The increase slows down a bit when the total fuel expenses hit 200 Canadian dollars, possibly indicating that people spending more on gasoline tend to purchase more gasoline than needed. The dropout of the price effect is not completely a surprise in the view of the relatively flat albeit decreasing trend shown in YN.

Next, let's examine the parametric effect estimates in Table 5. Clearly, having more than one driver in a household increases the total distance driven. This agrees with the finding in YN who used the logarithm of number of drivers as a linear predictor and obtained a positive coefficient estimate. Moreover, the zero coefficients for `driver2` and `driver3` indicate that having more than two drivers does not result in any further increase. It might

be a bit surprising to see that households with a single vehicle actually covers more distance than those with multiple vehicles. However, this probably provides another footnote on our finding in the age effect: young people, most of whom are single and possess only one car, tend to drive the most distance. The coefficient estimates for the household size dummies agrees with the small positive coefficient estimate in YN where the logarithm of household size was used as a linear predictor. The monthly effects also agree with those in Figure 3 of YN. Peak driving months occur during July and August and the winter months, from November to February, sees the lowest drives. The coefficient estimates for the fuel type dummies basically tell the well-known fact that the higher degree gasoline is more efficient. The `urban` coefficient estimate also agrees with that in YN, indicating that families residing in cities drive less than their rural counterparts. While the distance driven by households in Quebec and the Prairie provinces is similar to that by households in British Columbia, households in the Atlantic provinces and Ontario tend to drive more distance. Distance driven in 1995 is slightly more than that in 1994 and 1996, which are similar to each other.

5 Discussion

Taking advantage of recent developments for variable selection in parametric and nonparametric models, we have proposed a flexible semiparametric method that have parametric components of diverging dimensions and nonparametric part with multiple additive components and can performance simultaneous variable selection on both parametric and nonparametric parts. In practice, although a categorical variable is usually included in the linear part as dummies, it is hard to determine whether the effect of a continuous predictor is linear or nonlinear. As shown in our numerical examples, the proposed method can often recover the linear pattern if the true effect is linear. However, a more formal treatment in point is the linear and nonlinear discover (LAND) developed in Zhang et al. (2011), which automatically distinguish linear or nonlinear effects in an additive model through the application of a novel penalty. Their method can surely serve as a preliminary analysis tool before the application

of our method.

Appendix

In this Appendix, we present the conditions, prepare several preliminary results, and give the proofs of the main results.

Let $\boldsymbol{\beta}_0$ and f_{0k} be respectively the true coefficient vector for linear effects and the true functions for nonlinear effects. For any $r_n \to 0$, $\lambda \sim r_n$ means that $\lambda = O(r_n)$ and $\lambda^{-1} = O(r_n^{-1})$. Let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimal and maximal eigenvalue of the square matrix M, respectively.

A.1 Conditions and A Lemma

We assume the following regularity conditions:

(C1) ϵ is assumed to be independent of $\mathbf{x}$, and has the sub-Exponential tail, i.e. $E[\exp(|\epsilon|/C_0)] \leq C_0$ almost surely for some $0 < C_0 < \infty$;

(C2) $\sum_{i=1}^n \phi_i \phi_i'/n$ converges to some non-singular matrix with

$$\phi_i = [1, t_{i1}, \ldots, t_{i1}^{m-1}, \ldots, t_{iq}, \ldots, t_{iq}^{m-1}, x_{i1}, \ldots, x_{ip}]',$$

and $\lambda_{\min}(\sum_{i=1}^n \phi_i \phi_i'/n) \geq c_1 > 0$ for any n;

(C3) Assume that $E\mathbf{x} = 0$ and $\sum_i \mathbf{x}_i \mathbf{x}_i'/n$ converges to some nonsingular matrix Σ with $0 < c_2 \leq \lambda_{\min}(\sum_i \mathbf{x}_i \mathbf{x}_i'/n) \leq \lambda_{\max}(\sum_i \mathbf{x}_i \mathbf{x}_i'/n) \leq c_3 < \infty$ for any n;

(C4) Let $h_2(\mathbf{x}, \mathbf{T}) = \mathbf{x} - \sum_{k=1}^q E(\mathbf{x}|T_k)$. Assume that $\Sigma \equiv E h_2^{\otimes 2}$ is nonsingular;

(C5) Let $h_{1k}(t_k) = E(\mathbf{x}|T_k = t_k)$. We assume that $\|h_{1k}\| < \infty$ for any $k = 1, \ldots, q$;

(C6) The random covariate $\mathbf{T} \equiv (T_1, \ldots, T_q)'$ on $[0,1]^q$ is (i) pairwise independent; and (ii) $E f_k(T_k) = 0$ for all $k = 1, \ldots, q$;

The above conditions are very mild and commonly used in the literature. For example, the sub-exponential condition C1 is satisfied if the error is Gaussian distribution. If $\mathbf{x}$ are $\mathbf{T}$ are independent, Condition (C5) is trivially satisfied, and (C4) is simply implied by (C3).

We provide three useful matrix inequalities and one lemma for preparing the proof of Proposition 2.1. Given any $n \times m$ matrix $\mathbf{A}$ and symmetric strictly positive definite matrix $\mathbf{B}$, $n \times 1$ vector $\mathbf{s}$ and $\mathbf{z}$, and $m \times 1$ vector $\mathbf{w}$, we have

$$|\mathbf{s}'\mathbf{A}\mathbf{w}| \leq \|\mathbf{s}\|\|\mathbf{A}\|\|\mathbf{w}\| \tag{A.1}$$

$$|\mathbf{s}'\mathbf{B}\mathbf{z}| \leq |\mathbf{s}'\mathbf{B}\mathbf{s}|^{1/2}|\mathbf{z}'\mathbf{B}\mathbf{z}|^{1/2} \tag{A.2}$$

$$|\mathbf{s}'\mathbf{z}| \leq \|\mathbf{s}\|\|\mathbf{z}\| \tag{A.3}$$

where $\|\mathbf{A}\|^2 = \sum_j \sum_i a_{ij}^2$. (A.2) follows from the Cauchy-Schwartz inequality.

Since smoothing spline is a linear smoother, the solution to (2.2) can be written as

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'(I - A(\lambda))\mathbf{X})^{-1}\mathbf{X}'(I - A(\lambda))\mathbf{y}, \quad \widetilde{\mathbf{f}} = A(\lambda)(\mathbf{y} - \mathbf{X}\widetilde{\beta}),$$

where $A(\lambda)$ is the smoothing matrix as defined in Section 3.1 of Gu (2002) and $\widetilde{\mathbf{f}} = (\sum_{k=1}^q \widetilde{f}_k(t_{1k}), \ldots, \sum_{k=1}^q \widetilde{f}_k(t_{nk}))'$. It typically holds that $tr(A(\lambda)) = O(\lambda^{-1/2m})$ and $tr(A^2(\lambda)) = O(\lambda^{-1/2m})$ as $\lambda \to 0$ and $n\lambda^{1/2m} \to \infty$. Our next lemma on the properties of the smoothing matrix $A(\lambda)$ extend the results for one-dimensional t in Heckman (1986).

LEMMA A.1. *Let* $\mathbf{f}_0 = (\sum_{k=1}^q f_0(t_{1k}), \ldots, \sum_{k=1}^q f_0(t_{nk}))'$. *If* $\lambda \to 0$, $n\lambda^{1/2m} \to \infty$ *and* $p_n = o(n^{1/2} \wedge n\lambda^{1/2m})$, *then*

$$n^{-k/2} \sum_{i=1}^n |[(I - A(\lambda))\mathbf{f}_0]_i|^k = O(\lambda^{k/2}), k = 2, 3, \ldots, \tag{A.4}$$

$$[\mathbf{X}'A(\lambda)\epsilon]_i = O_P(\lambda^{-1/4m}), \tag{A.5}$$

$$[\mathbf{X}'((I - A(\lambda))\mathbf{f}_0 + \epsilon)]_i = O_P(n^{1/2}), \tag{A.6}$$

$$[\mathbf{X}'(I - A(\lambda))\mathbf{X}/n]_{ij} = \Sigma_{ij} + O_P(n^{-1/2} \vee n^{-1}\lambda^{-1/2m}), \tag{A.7}$$

$$\|\mathbf{X}'(I - A(\lambda))\mathbf{X}/n - \Sigma\| = o_P(1). \tag{A.8}$$

Proof of Lemma A.1. For $k = 2$ in (A.4), note that $A(\lambda)\mathbf{f}_0$ is the vector of values of the solution function, call it f^*, to the problem: Find $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(f_0(\boldsymbol{t}_i) - f(\boldsymbol{t}_i))^2 + \lambda\sum_{k=1}^{q}\|f_k\|^2.$$

Therefore,

$$\begin{aligned}
&\frac{1}{n}\|(I - A(\lambda))\mathbf{f}_0\|^2 + \lambda\sum_{k=1}^{q}\|f_k^*\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(f_0(\boldsymbol{t}_i) - f^*(\boldsymbol{t}_i))^2 + \lambda\sum_{k=1}^{q}\|f_k^*\|^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}(f_0(\boldsymbol{t}_i) - f_0(\boldsymbol{t}_i))^2 + \lambda\sum_{k=1}^{q}\|f_{0k}\|^2 = \lambda\sum_{k=1}^{q}\|f_{0k}\|^2 = O(\lambda).
\end{aligned}$$

Hence (A.4) is valid for $k = 2$. For a general k, we can apply induction on k. Assume $n^{-(k-1)/2}\sum_{i=1}^{n}|[(I - A(\lambda))\mathbf{f}_0]_i|^{k-1} = O(\lambda^{(k-1)/2})$. Then

$$\begin{aligned}
&n^{-k/2}\sum_{i=1}^{n}|[(I - A(\lambda))\mathbf{f}_0]_i|^k \\
\le&\{n^{-1/2}\max_{1\le i\le n}|[(I - A(\lambda))\mathbf{f}_0]_i|\}\{n^{-(k-1)/2}\sum_{i=1}^{n}|[(I - A(\lambda))\mathbf{f}_0]_i|^{k-1}\} \\
\le&\{n^{-1}\sum_{i=1}^{n}|[(I - A(\lambda))\mathbf{f}_0]_i|^2\}^{1/2}O(\lambda^{(k-1)/2}) = O(\lambda^{k/2}).
\end{aligned}$$

Hence (A.4) is proved.

(A.5) is straightforward from the fact that $Var((\mathbf{X}'A(\lambda)\epsilon)_i) = \sigma^2\Sigma_{ii}\mathrm{tr}(A^2(\lambda))$.

To prove (A.6), write the left hand side as $\sqrt{n}\sum_{j=1}W_{ij}$, where $W_{ij} = n^{-1/2}X_{ij}(\epsilon_j + [(I - A(\lambda))\mathbf{f}_0]_j)$. The W_{ij}'s are independent mean zero random variables with

$$\sum_{j=1}^{n}Var(W_{ij}) = \Sigma_{ii}(\sigma^2 + n^{-1}\sum_{j=1}^{n}\{[(I - A(\lambda))\mathbf{f}_0]_j\}^2) \to \Sigma_{ii}\sigma^2,$$

where (A.4) is used in the last approximation. Also,

$$\begin{aligned}
\sum_{j=1}^{n}E|W_{ij}|^3 &= n^{-3/2}E|X_{i1}|^3\sum_{j=1}^{n}E|\epsilon_j + [(I - A(\lambda))\mathbf{f}_0]_j|^3 \\
&\le 3n^{-3/2}\{\sum_{j=1}^{n}E|\epsilon_j|^3 + \sum_{j=1}^{n}|[(I - A(\lambda))\mathbf{f}_0]_j|^3\} \to 0
\end{aligned}$$

by the sub-exponential tail condition on ϵ_j's and (A.4). Hence, Liapounov's central limit theorem implies (A.6).

To see (A.7), we note that $E|[\mathbf{X}'A(\lambda)\mathbf{X}]_{ii}| = \Sigma_{ii}\mathrm{tr}(A(\lambda))$ and for $i \neq j$,

$$E\{[\mathbf{X}'A(\lambda)\mathbf{X}]_{ij}\}^2 = \Sigma_{ii}^2(\mathrm{tr}(A(\lambda)))^2 + \{\Sigma_{ii}\Sigma_{jj} + \Sigma_{ij}^2\}\mathrm{tr}(A^2(\lambda)) + \{E[(X_{1i}X_{1j})^2] - 2\Sigma_{ij}^2 - \Sigma_{ii}\Sigma_{jj}\}\Sigma_r[A(\lambda)]_{rr}^2.$$

Hence (A.7) follows from the order bound on $\mathrm{tr}(A(\lambda))$ and $\mathrm{tr}(A^2(\lambda))$.

(A.7) implies that $\|\mathbf{X}'(I - A(\lambda))\mathbf{X}/n - \Sigma\| = O_P(p_n n^{-1/2} \vee p_n n^{-1}\lambda^{-1/2m})$. Thus (A.8) follows from the assumed dimension condition $p_n = o(n^{1/2} \wedge n\lambda^{1/2m})$. □

A.2 Proof of Proposition 2.1

We first prove (2.3). Based on the definition on $\widetilde{\boldsymbol{\beta}}$, we have the below inequality:

$$\frac{1}{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'(I - A)\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \frac{2}{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'(I - A)(\mathbf{f}_0 + \boldsymbol{\epsilon}) \leq 0.$$

Let $\delta_n = n^{-1/2}[\mathbf{X}'(I - A)\mathbf{X}]^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\omega_n = n^{-1/2}[\mathbf{X}'(I - A)\mathbf{X}]^{-1/2}\mathbf{X}'(I - A)(\mathbf{f}_0 + \boldsymbol{\epsilon})$. Then the above inequality can be rewritten as $\|\delta_n\|^2 - 2\omega_n'\delta_n \leq 0$, i.e. $\|\delta_n - \omega_n\|^2 \leq \|\omega_n\|^2$. By Cauchy-Schwartz inequality, we have $\|\delta_n\|^2 \leq 2(\|\delta_n - \omega_n\|^2 + \|\omega_n\|^2) \leq 4\|\omega_n\|^2$. Examine $\|\omega_n\|^2 = K_{1n} + K_{2n} + K_{3n}$, where

$$K_{1n} = n^{-1}\boldsymbol{\epsilon}'(I - A)\mathbf{X}[\mathbf{X}'(I - A)\mathbf{X}]^{-1}\mathbf{X}'(I - A)\boldsymbol{\epsilon}$$
$$K_{2n} = 2n^{-1}\boldsymbol{\epsilon}'(I - A)\mathbf{X}[\mathbf{X}'(I - A)\mathbf{X}]^{-1}\mathbf{X}'(I - A)\mathbf{f}_0(t)$$
$$K_{3n} = n^{-1}\mathbf{f}_0(T)'(I - A)\mathbf{X}[\mathbf{X}'(I - A)\mathbf{X}]^{-1}\mathbf{X}'(I - A)\mathbf{f}_0(t)$$

Applying (A.5), (A.6) and (A.7) to the above three terms, we can conclude that all of them are of the order $O_P(p_n n^{-1})$ by considering the matrix inequalities (A.1)-(A.3). Thus we have proved (2.3) by considering (A.8).

We next prove (2.4). Define $g_0(\mathbf{x}, \boldsymbol{t}) = \mathbf{x}'\boldsymbol{\beta}_0 + \sum_{k=1}^q f_{0k}(t_k)$ and $\widetilde{g}_n = \mathbf{x}'\widetilde{\boldsymbol{\beta}} + \sum_{k=1}^q \widetilde{f}_k(t_k)$. Since $\widetilde{g}_n$ is defined to minimize $g \mapsto \|g_0 + \epsilon - g\|_n^2 + \lambda \sum \|f_k\|^2$, we thus establish the following

inequality

$$\|\widetilde{g}_n - g_0\|_n^2 + \lambda \sum_{k=1}^{q} \|\widetilde{f}_k\|^2 \leq 2 < \widetilde{g}_n - g_0, \epsilon >_n + \lambda \sum_{k=1}^{q} \|f_{0k}\|^2 \tag{A.9}$$

$$\|\widetilde{g}_n - g_0\|_n^2 \leq 2\|\epsilon\|_n \|\widetilde{g}_n - g_0\|_n + \lambda \sum_{k=1}^{q} \|f_{0k}\|^2$$

$$\leq O_P(1)\|\widetilde{g}_n - g_0\|_n + o_P(1)$$

by the Cauchy-Schwartz inequality and the sub-exponential tail assumption of ϵ. The above inequality implies that $\|\widetilde{g}_n - g_0\|_n = O_P(1)$. Considering the fact that $\|g_0\|_\infty = O(p_n)$, we have $\|\widetilde{g}_n\|_n = O_P(p_n)$. By the Sobolev embedding Theorem, we can decompose $g(\mathbf{x}, \boldsymbol{t})$ as $g_1(\mathbf{x}, \boldsymbol{t}) + g_2(\boldsymbol{t})$, where $g_1(\mathbf{x}, \boldsymbol{t}) = \mathbf{x}'\boldsymbol{\beta} + \sum_{k=1}^{q} \sum_{j=1}^{m} \alpha_{jk} t_k^{j-1}$ and $g_2(\boldsymbol{t}) = \sum_{k=1}^{q} f_{2k}(t_k)$ with $\|g_2(\boldsymbol{t})\|_\infty \leq \sum_{k=1}^{q} \|f_{2k}\|_\infty \leq \sum_{k=1}^{q} \|f_k\|$. Similarly, we can write $\widetilde{g}_n = \widetilde{g}_{1n} + \widetilde{g}_{2n}$, where $\widetilde{g}_{1n} = \mathbf{x}'\widetilde{\boldsymbol{\beta}} + \sum_{k=1}^{q} \sum_{j=1}^{m} \widetilde{\alpha}_{jk} t_k^{j-1} = \widetilde{\delta}'\phi$ and $\|\widetilde{g}_{2n}\|_\infty \leq \sum_{k=1}^{q} \|\widetilde{f}_k\|$. We shall now show that $\|\widetilde{g}_n\|_\infty / (1 + \sum_k \|\widetilde{f}_k\|) = O_P(p_n)$ as follows. First, we have

$$\frac{\|\widetilde{g}_{1n}\|_n}{1 + \sum_k \|\widetilde{f}_k\|} \leq \frac{\|\widetilde{g}_n\|_n}{1 + \sum_k \|\widetilde{f}_k\|} + \frac{\|\widetilde{g}_{2n}\|_n}{1 + \sum_k \|\widetilde{f}_k\|} = O_P(p_n). \tag{A.10}$$

Based on the assumption about $\sum_i \phi_i \phi_i'/n$, (A.10) implies that $\|\widetilde{\delta}\|/(1 + \sum_k \|\widetilde{f}_k\|) = O_P(p_n)$. Since $(\mathbf{x}, \boldsymbol{t}) \in [0,1]^{p+q}$, $\|\widetilde{g}_{1n}\|_\infty / (1 + \sum_k \|\widetilde{f}_k\|) = O_P(p_n)$. So we have proved that $\|\widetilde{g}_n\|_\infty / (1 + \sum_k \|\widetilde{f}_k\|) = O_P(p_n)$.

According to Birman and Solomjak (1967), we know the entropy number for the below constructed class of functions:

$$H\left(\delta, \left\{ \frac{(g - g_0)/p_n}{1 + \sum_k \|f_k\|} : g \in \mathcal{G}, \frac{\|g\|_\infty / p_n}{1 + \sum_k \|f_k\|} \leq C \right\}, \|\cdot\|_\infty \right) \leq M_1 q (\delta/q)^{-1/m},$$

where M_1 is some positive constant and $\mathcal{G} = \{g(x,t) = \mathbf{x}'\boldsymbol{\beta} + \sum_k f_k(t_k) : \boldsymbol{\beta} \in R^d, \sum_k \|f_k\| < \infty\}$. Combining the above entropy result and Theorem 2.2 in Mammen and van de Geer (1997) about the continuity modulus of the empirical processes $\{\sum_{i=1}^{n} \epsilon_i (g - g_0)(\mathbf{x}_i, \boldsymbol{t}_i)/p_n\}$ indexed by g, we can establish the following set of inequalities:

$$\lambda \sum_k \|\widetilde{f}_k\|^2 \leq \left[\|\widetilde{g}_n - g_0\|_n^{1-1/2m} (1 + \sum_k \|\widetilde{f}_k\|)^{1/2m} p_n^{1/2m} \vee (1 + \sum_k \|\widetilde{f}_k\|) p_n n^{-\frac{2m-1}{2(2m+1)}} \right] O_P(n^{-1/2})$$

$$+ \lambda \sum_k \|f_{0k}\|^2, \tag{A.11}$$

and

$$\|\widetilde{g}_n - g_0\|_n^2 \leq \left[\|\widetilde{g}_n - g_0\|_n^{1-1/2m}(1+\sum_k \|\widetilde{f}_k\|)^{1/2m} p_n^{1/2m} \vee (1+\sum_k \|\widetilde{f}_k\|) p_n n^{-\frac{2m-1}{2(2m+1)}} \right] O_P(n^{-1/2}) + \lambda \sum_k \|f_{0k}\|^2 \tag{A.12}$$

based on the inequality (A.9). Let $a_n = \|\widetilde{g}_n - g_0\|_n / [(1+\sum_k \|\widetilde{f}_k\|)p_n]$, then (A.12) becomes

$$\begin{aligned} a_n^2 &\leq O_P(n^{-1/2}) a_n^{1-1/2m} \vee O_P(n^{-2m/(2m+1)}) \vee O_P(\lambda/p_n) \\ &\leq O_P(n^{-1/2}) a_n^{1-1/2m} \vee O_P(n^{-2m/(2m+1)}), \end{aligned} \tag{A.13}$$

where the second inequality follows from $\lambda \sim n^{-2m/(2m+1)}$ and $p_n \to \infty$. Therefore, (A.13) implies that $a_n = O_P(n^{-m/(2m+1)})$. We next analyze (A.11) which can be rewritten as

$$\begin{aligned} \frac{\lambda}{p_n}(\sum_k \|\widetilde{f}_k\| - 1) &\leq O_P(n^{-1/2}) a_n^{1-1/2m} \vee O_P(n^{-2m/(2m+1)}) \\ (\sum_k \|\widetilde{f}_k\| - 1) &\leq \frac{p_n}{\lambda} O_P(n^{-2m/(2m+1)}) \\ \sum_k \|\widetilde{f}_k\| &\leq O_P(p_n). \end{aligned} \tag{A.14}$$

in view of the condition that $\lambda \sim n^{-2m/(2m+1)}$. Combining the result that $a_n = O_P(n^{-m/(2m+1)})$ with (A.14), we have $\|\widetilde{g}_n - g_0\|_n = O_P(n^{-m/(2m+1)} p_n^2)$. Given that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\sqrt{p_n/n})$ shown above, we have $\|\mathbf{x}'(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_n = O_P(\sqrt{p_n/n})$ under the assumption on $\sum \mathbf{x}_i \mathbf{x}_i'/n$. By the triangular inequality, we further have $\|\sum_k \widetilde{f}_k - \sum_k f_{0k}\|_n = O_P(\sqrt{p_n/n} \vee n^{-m/(2m+1)} p_n^2)$. Considering that $\|\sum_k \widetilde{f}_k\|_\infty / p_n = O_P(1)$ as shown above, we can apply Theorem 2.3 in Mammen and van de Geer (1997) to obtain that $\|\sum_k \widetilde{f}_k - \sum_k f_{0k}\|_2 = O_P(\sqrt{p_n/n} \vee n^{-m/(2m+1)} p_n^2)$. Considering Condition (C6), we can show (2.4). This completes the whole proof. □

References

Candes, E. and Tao, T. (2007), "The Dantzig selector: statistical estimation when p is much larger than n (with discussion)," *Ann. Statist.*, 35, 2313–2351.

Cheng, G. (2011), "How many iterations are sufficeint for semiparametric estimation?" Manuscript.

Cheng, G. and Zhang, H. H. (2011), "Sparse and Efficient Estimation for Partial Spline Models with Increasing Dimension," *Scand. J. of Statist.*, invited Revision.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression (with discussion)," *Ann. Statist.*, 32, 407–499.

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric independence screening in sparse ultra-high dimensional additive models," *J. Amer. Statist. Assoc.*, to appear.

Fan, J. and Li, R. (2001), "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties," *J. Amer. Statist. Assoc.*, 96, 1348–1360.

Fan, J. and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *J. Roy. Statist. Soc. Ser. B*, 70, 849–911.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer-Verlag.

Härdle, W., Liang, H., and Gao, J. T. (2000), *Partially Linear Models*, Heidelberg: Springer Physica.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, New York: Springer-Verlag.

Heckman, N. E. (1986), "Spline Smoothing in a Partly Linear Model," *J. Roy. Statist. Soc. Ser. B*, 48, 244–248.

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable selection in nonparametric additive models," *Ann. Statist.*, 38, 2282–2313.

Koltchinskii, V. and Yuan, M. (2010), "Sparsity in Multiple Kernel Learning," *Ann. Statist.*, 38, 3660–3695.

Liang, H. and Li, R. (2009), "Variable Selection and Inference in Partially Linear Error-in-Variable Models," *J. Amer. Statist. Assoc.*, 104, 234–248.

Lin, Y. and Zhang, H. H. (2006), "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models," *Ann. Statist.*, 34, 2272–2297.

Liu, X., Wang, L., and Liang, H. (2010), "Estimation and Variable Selection for Semiparametric Additive Partial Linear Models," *Statist. Sin.*, in press.

Ma, S. and Du, P. (2011), "Variable selection in partly linear regression model with diverging dimensions for right censored data," *Statist. Sin.*, to appear.

Ma, S. and Yang, L. (2011), "Spline-backfitted kernel smoothing of partially linear additive model," *J. Statist. Plann. Inference*, 141, 204 – 219.

Meier, L., van de Geer, S., and Bühlmann, P. (2009), "High-dimensional additive modeling," *Ann. Statist.*, 37, 3779–3821.

Ni, X., Zhang, H. H., and Zhang, D. (2009), "Automatic model selection for partially linear models," *Journal of Multivariate Analysis*, 100, 2100–2111.

Radchenko, P. and James, G. M. (2010), "Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions," *J. Amer. Statist. Assoc.*, 105, 1541–1553.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *J. Roy. Statist. Soc. Ser. B*, 71, 1009–1030.

Storlie, C., Bondell, H., Reich, B., and Zhang, H. H. (2011), "The adaptive COSSO for nonparametric surface estimation and model selection," *Statist. Sin.*, to appear.

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

Wei, F., Huang, J., and Li, H. (2011), "Variable selection and estimation in high-dimensional varying-coefficient models," *Statist. Sin.*, in press.

Xie, H. and Huang, J. (2009), "SCAD-Penalized Regression in High-Dimensional Partially Linear Models," *Ann. Statist.*, 37, 673–696.

Xue, L. (2009), "Consistent variable selection in additive models," *Statist. Sin.*, 19, 1281–1296.

Yatchew, A. and No, J. A. (2001), "Household Gasoline Demand in Canada," *Econometrica*, 69, 1697–1709.

Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *J. Roy. Statist. Soc. Ser. B*, 68, 49–67.

Zhang, H. H., Cheng, G., and Liu, Y. (2011), "Linear or nonlinear? Automatic structure discovery for partially linear models," *J. Amer. Statist. Assoc.*, to appear.

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *J. Amer. Statist. Assoc.*, 101, 1418–1429.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models (with discussion)," *Ann. Statist.*, 36, 1509–1533.

Table 1: Prediction performance comparison by the means and standard deviations (in the brackets) of the RPEs computed from 100 replicates. Here $\boldsymbol{\beta}_1 = (1, 0.8, 1.4, 0.6, 1.2, 0.9, 1.1, 1.2, \mathbf{0}'_{22})'$ and $\boldsymbol{\beta}_2 = (0.6, 0, 0, 1, 0, 0, 0.8, 0, \mathbf{0}'_{22})'$.

$\boldsymbol{\beta}$	n	$\mathcal{M}_A$	$\mathcal{M}_B$	$\mathcal{M}_C$	$\mathcal{M}_D$
$\boldsymbol{\beta}_1$	200	0.736(0.092)	0.909(0.070)	0.312(0.018)	0.304(0.021)
	500	0.896(0.053)	0.984(0.015)	0.316(0.008)	0.311(0.010)
$\boldsymbol{\beta}_2$	200	0.740(0.087)	0.921(0.076)	0.311(0.018)	0.305(0.020)
	500	0.894(0.054)	0.983(0.015)	0.313(0.007)	0.309(0.009)

Table 2: Variable selection frequencies for parametric components. Values are the average numbers of selection for signal variables and average total numbers of selected noise variables computed from 100 replicates.

			Signal Variables (with values of β_js)								
			1.0	**0.8**	**1.4**	**0.6**	**1.2**	**0.9**	**1.1**	**1.2**	**Noise**
$\boldsymbol{\beta} = \boldsymbol{\beta}_1$	$n = 200$	$\mathcal{M}_A$	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	4.96
		$\mathcal{M}_C$	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	5.09
	$n = 500$	$\mathcal{M}_A$	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.65
		$\mathcal{M}_C$	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	6.16
			0.6			**1.0**			**0.8**		**Noise**
$\boldsymbol{\beta} = \boldsymbol{\beta}_2$	$n = 200$	$\mathcal{M}_A$	1.00			1.00			1.00		3.66
		$\mathcal{M}_C$	0.97			1.00			1.00		3.52
	$n = 500$	$\mathcal{M}_A$	1.00			1.00			1.00		1.76
		$\mathcal{M}_C$	1.00			1.00			1.00		5.39

Table 3: Variable selection frequencies for nonparametric components. In procedure $\mathcal{M}_C$, nonlinear effects were misspecified as linear. Values are the average numbers of selection for signal variables and average total numbers of selected noise variables computed from 100 replicates.

			Signal Variables				
			$f_1(t_1)$	$f_2(t_2)$	$f_3(t_3)$	$f_4(t_4)$	**Noise**
$\boldsymbol{\beta} = \boldsymbol{\beta}_1$	$n = 200$	$\mathcal{M}_A$	1.00	0.98	0.96	1.00	1.06
		$\mathcal{M}_C$	1.00	0.41	0.97	1.00	1.76
	$n = 500$	$\mathcal{M}_A$	1.00	1.00	1.00	1.00	0.57
		$\mathcal{M}_C$	1.00	0.63	1.00	1.00	2.74
$\boldsymbol{\beta} = \boldsymbol{\beta}_2$	$n = 200$	$\mathcal{M}_A$	1.00	1.00	0.96	1.00	1.08
		$\mathcal{M}_C$	1.00	0.29	0.94	1.00	1.33
	$n = 500$	$\mathcal{M}_A$	1.00	1.00	1.00	1.00	0.59
		$\mathcal{M}_C$	1.00	0.51	1.00	1.00	2.30

Table 4: Coefficient estimates of signal variables in parametric components for procedures with and without refitting. Values are the means and standard deviations (in bracket) of the coefficient estimates for signal variables computed from 100 replicates.

			Signal Variables (with values of β_js)							
			1.0	**0.8**	**1.4**	**0.6**	**1.2**	**0.9**	**1.1**	**1.2**
$\boldsymbol{\beta} = \boldsymbol{\beta}_1$	$n = 200$	$\mathcal{M}_A$	0.997	0.783	1.423	0.550	1.237	0.894	1.115	1.191
			(.105)	(.121)	(.109)	(.153)	(.121)	(.124)	(.116)	(.097)
		$\mathcal{M}_B$	0.994	0.794	1.402	0.592	1.209	0.905	1.108	1.192
			(.093)	(.104)	(.099)	(.137)	(.104)	(.100)	(.101)	(.097)
	$n = 500$	$\mathcal{M}_A$	1.007	0.804	1.407	0.567	1.214	0.895	1.111	1.186
			(.052)	(.068)	(.062)	(.080)	(.067)	(.064)	(.059)	(.058)
		$\mathcal{M}_B$	1.003	0.807	1.392	0.602	1.199	0.898	1.110	1.186
			(.050)	(.062)	(.055)	(.060)	(.062)	(.066)	(.054)	(.056)
			0.6			**1.0**			**0.8**	
$\boldsymbol{\beta} = \boldsymbol{\beta}_2$	$n = 200$	$\mathcal{M}_A$	0.522			1.006			0.776	
			(.102)			(.104)			(.104)	
		$\mathcal{M}_B$	0.586			0.999			0.810	
			(.087)			(.100)			(.084)	
	$n = 500$	$\mathcal{M}_A$	0.586			1.001			0.802	
			(.054)			(.047)			(.048)	
		$\mathcal{M}_B$	0.603			1.000			0.806	
			(.048)			(.048)			(.052)	

Table 5: Coefficient estimates for the parametric components in the gasoline demand example of Section 4 for procedures without and with refitting.

Refitting	driver1	driver2	driver3	veh1	veh2	hhsize1	hhsize2
No	-.0009	0	0	.0369	0	0	.0245
Yes	-.0204	-	-	.0450	-	-	.0279
	hhsize3	hhsize4	hhsize5	month1	month2	month3	month4
No	.0285	0	0	0	0	.1267	.0419
Yes	.0607	-	-	-	-	.1355	.0764
	month5	month6	month7	month8	month9	month10	month11
No	.1061	.0970	.2520	.2575	.1500	.1245	0
Yes	.1170	.1185	.2597	.2674	.1563	.1440	-
	fueltype1	fueltype2	fueltype3	urban	prov1	prov2	prov3
No	-.3158	-.4217	-.4398	-.0907	.0341	0	.0305
Yes	-.3293	-.4371	-.4592	-.0923	.0607	-	.0637
	prov4	year94	year95	Intercept			
No	0	0	.0023	7.2233			
Yes	-	-	.0205	8.2466			

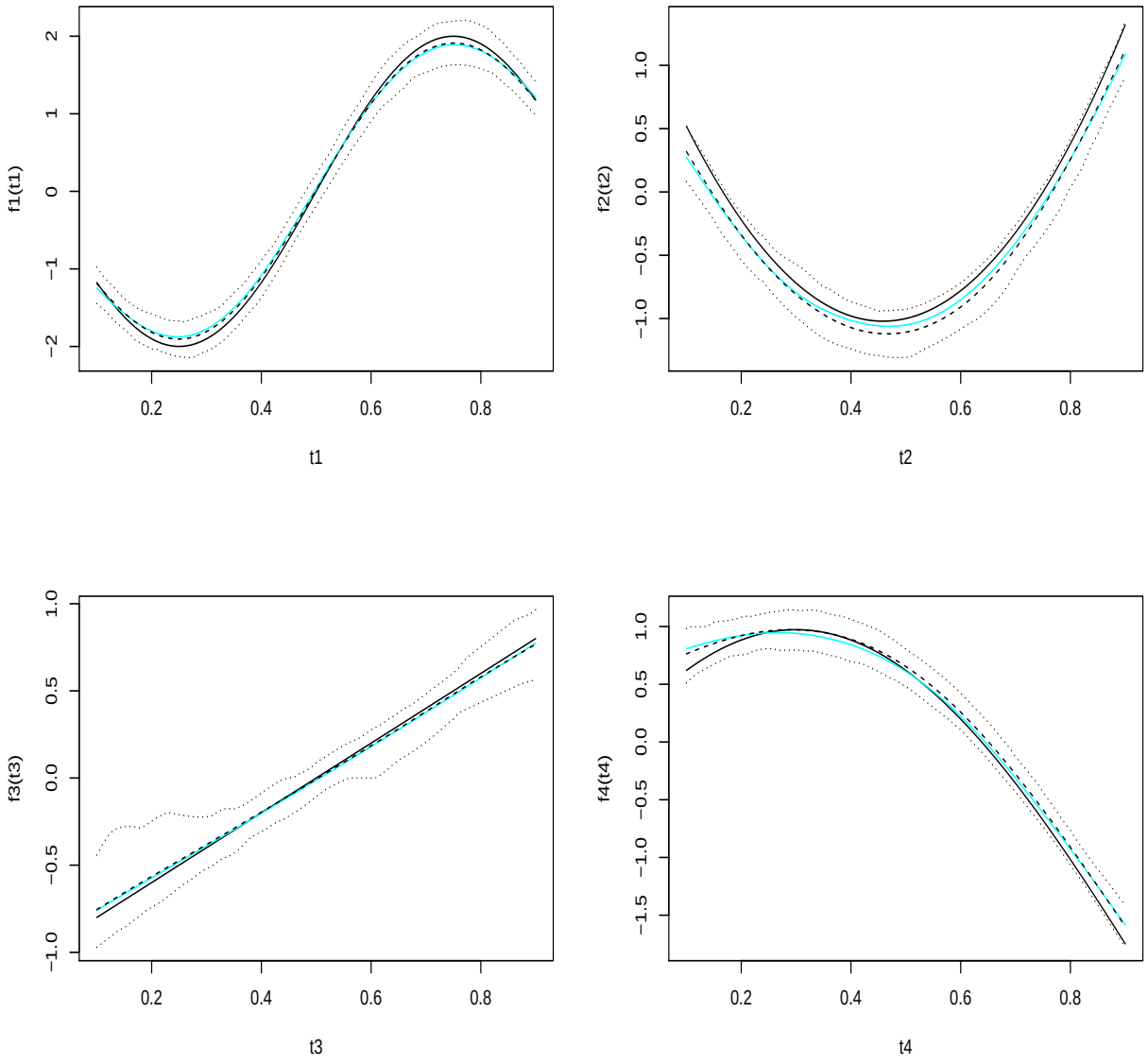

Figure 1: Estimates for Nonparametric Components. Solid lines are the true functions, faded lines are the connected point-wise mean estimates without refitting, dashed lines are the connected point-wise mean estimates after refitting, and dotted lines are the connected 0.05 and 0.95 quantiles of the point-wise estimates after refitting.

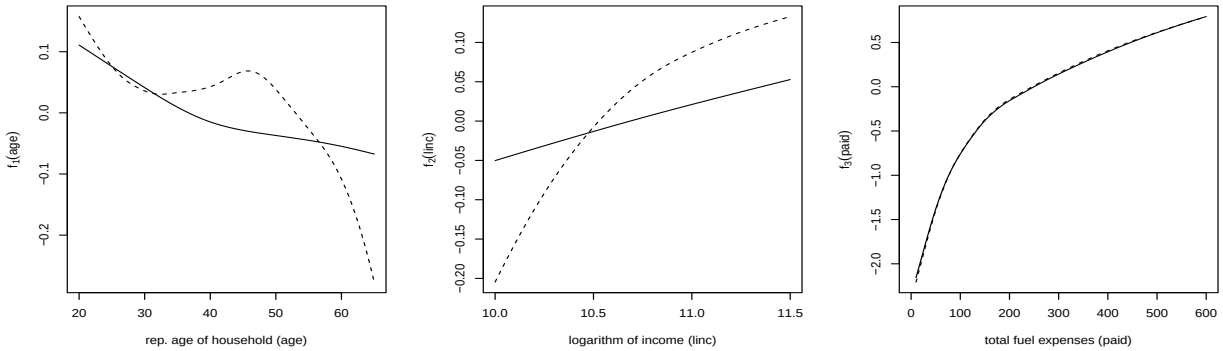

Figure 2: Estimates for nonparametric components in the gasoline demand example of Section 4 for procedures without and with refitting. Dashed lines are estimates without refitting and solid lines are estimates with refitting. Not plotted here is the `price` effect, which was zeroed out at the variable selection step.