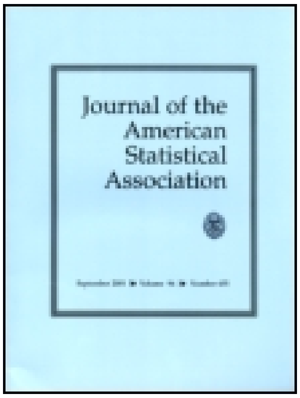

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models

Hao Helen Zhang[a], Guang Cheng[a] & Yufeng Liu[a]

[a] Hao Helen Zhang is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695, and Associate Professor, Department of Mathematics, University of Arizona, Tucson, AZ 85721. Guang Cheng is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906. Yufeng Liu is Associate Professor, Department of Statistics and Operations Research, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599. The authors are supported in part by NSF grants DMS-0645293 (Zhang), DMS-0906497 (Cheng), and DMS-0747575 (Liu), NIH grants NIH/NCI R01 CA-085848 (Zhang), NIH/NCI R01 CA-149569 (Liu), and NIH/NCI P01 CA142538 (Zhang and Liu). The authors thank the editor, the associate editor, and two reviewers for their helpful comments and suggestions which led to a much improved presentation.
Published online: 24 Jan 2012.

To cite this article: Hao Helen Zhang, Guang Cheng & Yufeng Liu (2011) Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models, Journal of the American Statistical Association, 106:495, 1099-1112, DOI: 10.1198/jasa.2011.tm10281

To link to this article: http://dx.doi.org/10.1198/jasa.2011.tm10281

Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models

Hao Helen ZHANG, Guang CHENG, and Yufeng LIU

Partially linear models provide a useful class of tools for modeling complex data by naturally incorporating a combination of linear and nonlinear effects within one framework. One key question in partially linear models is the choice of model structure, that is, how to decide which covariates are linear and which are nonlinear. This is a fundamental, yet largely unsolved problem for partially linear models. In practice, one often assumes that the model structure is given or known and then makes estimation and inference based on that structure. Alternatively, there are two methods in common use for tackling the problem: hypotheses testing and visual screening based on the marginal fits. Both methods are quite useful in practice but have their drawbacks. First, it is difficult to construct a powerful procedure for testing multiple hypotheses of linear against nonlinear fits. Second, the screening procedure based on the scatterplots of individual covariate fits may provide an educated guess on the regression function form, but the procedure is ad hoc and lacks theoretical justifications. In this article, we propose a new approach to structure selection for partially linear models, called the LAND (Linear And Nonlinear Discoverer). The procedure is developed in an elegant mathematical framework and possesses desired theoretical and computational properties. Under certain regularity conditions, we show that the LAND estimator is able to identify the underlying true model structure correctly and at the same time estimate the multivariate regression function consistently. The convergence rate of the new estimator is established as well. We further propose an iterative algorithm to implement the procedure and illustrate its performance by simulated and real examples. Supplementary materials for this article are available online.

KEY WORDS: Model selection; RKHS; Semiparametric regression; Shrinkage; Smoothing splines.

1. INTRODUCTION

Linear and nonparametric models are two important classes of modeling tools for statistical data analysis and both have their unique advantages. Linear models are simple, easy to interpret, and the estimates are most efficient if the linear assumption is valid. Nonparametric models are less dependent on the model assumption and hence able to uncover nonlinear effects hidden in data. Partially linear models, a class of models between linear and nonparametric models, inherit advantages from both sides by allowing some covariates to be linear and others to be nonlinear. Partially linear models have wide applications in practice due to their flexibility.

Given the observations $(y_i, \mathbf{x}_i, \mathbf{t}_i)$, $i = 1, \ldots, n$, where y_i is the response, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $\mathbf{t}_i = (t_{i1}, \ldots, t_{iq})^{\mathrm{T}}$ are vectors of covariates, the partially linear model assumes that

$$y_i = b + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + f(\mathbf{t}_i) + \epsilon_i, \tag{1.1}$$

where b is the intercept, $\boldsymbol{\beta}$ is a vector of unknown parameters for linear terms, f is an unknown function from R^q to R, and ϵ_i's are iid random errors with mean zero and variance σ^2. In practice, the most used model for (1.1) is the following special case when $q = 1$:

$$y_i = b + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + f(t_i) + \epsilon_i. \tag{1.2}$$

For example, in longitudinal data analysis, the time covariate T is often treated as the only nonlinear effect. Model estimation and inference for (1.2) have been actively studied under various smooth regression settings, including smoothing splines (Wahba 1984; Engle et al. 1986; Heckman 1986; Rice 1986; Chen 1988; Hong 1991; Green and Silverman 1994; Liang, Hardle, and Carroll 1999), penalized regression splines (Ruppert, Wand, and Carroll 2003; Liang 2006; Wang, Li, and Huang 2008), kernel smoothing (Speckman 1988), and local polynomial regression (Fan and Gijbels 1996; Fan and Li 2004; Li and Liang 2008). Interesting applications include the analysis of city electricity (Engle et al. 1986), household gasoline consumption in the United States (Schmalensee and Stoker 1999), a marketing price-volume study in the petroleum distribution industry (Green and Silverman 1994), the logistic analysis of bioassay data (Dinse and Lagakos 1983), the mouthwash experiment (Speckman 1988), and so on. A recent monograph by Hardle, Liang, and Gao (2000) gave an excellent overview on partially linear models, and a more comprehensive list of references can be found there.

One natural question about the model (1.1) is, given a set of covariates, how one decides which covariates have linear effects and which covariates have nonlinear effects. For example, in the Boston housing data analyzed in the article, the main goals are to identify important covariates, study how each covariate is associated with the house value, and build a highly interpretable model to predict the median house values. The structure selection problem is fundamentally important, as the validity of the fitted model and its inference heavily depends on whether the model structure is specified correctly. Compared to the linear model selection, the structure selection for partially linear models is much more challenging because the models involve multiple linear and nonlinear functions and a model search needs to be conducted within some infinite-dimensional function space.

Hao Helen Zhang is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: hzhang@stat.ncsu.edu), and Associate Professor, Department of Mathematics, University of Arizona, Tucson, AZ 85721 (E-mail: hzhang@math.arizona.edu). Guang Cheng is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906. Yufeng Liu is Associate Professor, Department of Statistics and Operations Research, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599. The authors are supported in part by NSF grants DMS-0645293 (Zhang), DMS-0906497 (Cheng), and DMS-0747575 (Liu), NIH grants NIH/NCI R01 CA-085848 (Zhang), NIH/NCI R01 CA-149569 (Liu), and NIH/NCI P01 CA142538 (Zhang and Liu). The authors thank the editor, the associate editor, and two reviewers for their helpful comments and suggestions which led to a much improved presentation.

Journal of the American Statistical Association
September 2011, Vol. 106, No. 495, Theory and Methods
DOI: 10.1198/jasa.2011.tm10281

Furthermore, the difficulty level of model search increases dramatically as the data dimension grows due to the curse of dimensionality. This may explain why the problem of structure selection for partially linear models is less studied in the literature. Most works we mentioned above assume that the model structure (1.1) is given or known. In practice, data analysts oftentimes have to rely on their experience, historical data, or some screening tools to make an educated guess on the function forms for individual covariates. Two methods in common use are the screening and hypothesis testing procedures. The screening method first conducts univariate nonparametric regression for each covariate or unstructured additive models and then determines linearity or nonlinearity for each term by visualizing the fitted function. This method is useful in practice but lacks theoretical justifications. The second method is to test linear null hypotheses against nonlinear alternatives, sequentially or simultaneously, for each covariate. However, proper test statistics are often hard to construct and the tests may have low power when the number of covariates is large. In addition, these methods handle the structure selection problem and the model estimation separately, making it difficult to study inferential properties of the final estimator. To our knowledge, none of the existing methods can distinguish linear and nonlinear terms for partially linear models automatically and consistently. The main purpose of this article is to fill this gap.

Motivated by the need of an effective and theoretically justified procedure for structure selection in partially linear models, we propose a new approach, called the LAND (Linear And Nonlinear Discoverer), to identify model structure and estimate the regression function simultaneously. By solving a regularization problem in the frame of smoothing spline ANOVA, the LAND is able to distinguish linear and nonlinear terms, remove uninformative covariates from the model, and provide a consistent function estimate. Specifically, we show that the LAND estimator is consistent and establish its convergence rate. Furthermore, under the tensor product design, we show that the LAND is consistent in recovering the correct model structure and estimating both linear and nonlinear function components. An iterative computational algorithm is developed to implement the procedure. The rest of the article is organized as follows. In Section 2 we introduce the LAND estimator. Statistical properties of the new estimator, including its convergence rate and selection consistency, are established in Section 3. We discuss the idea of two-step LAND in Section 4. The computational algorithm and the tuning issue are discussed in Section 5. Section 6 contains simulated and real examples to illustrate finite sampling performance of the LAND. All the proofs are relegated to the Appendix. Due to the space restriction, Appendix 4 is given in online supplementary materials.

2. METHODOLOGY

2.1 Model Setup

From now on, we use $\mathbf{x}_i \in R^d$ instead of $(\mathbf{x}_i, \mathbf{t}_i)$ to represent the entire covariate vector, as we do not assume the knowledge of linear or nonlinear form for each covariate. Without loss of generality, all covariates are scaled to $[0, 1]$. Let $\{\mathbf{x}_i, y_i\}, i = 1, \ldots, n$, be an independent and identically distributed sample. The underlying true regression model has the form

$$y_i = b + \sum_{j \in I_L} x_{ij}\beta_j + \sum_{j \in I_N} f_j(x_{ij}) + \sum_{j \in I_O} 0(x_{ij}) + \epsilon_i, \tag{2.1}$$

where b is an intercept, I_L, I_N, I_O are the index sets for nonzero linear effects, nonzero nonlinear effects, and null effects, respectively. Let the total index set be $I = \{1, \ldots, d\}$; then $I = I_L \cup I_N \cup I_O$ and the three subgroups are mutually exclusive. The model (2.1) can be regarded as a hypothetical model, since I_L, I_N, I_O are generally unknown in practice. Since nonlinear functions embrace linear functions as special cases, we need to impose some restrictions on f's to assure the identifiability of terms in (2.1). This issue will be carefully treated later.

The model (2.1) is a special case of the additive model

$$y_i = b + g_1(x_{i1}) + \cdots + g_d(x_{id}) + \epsilon_i. \tag{2.2}$$

Without loss of generality, we assume that the function components in (2.2) satisfy some smoothness conditions, say, differentiable up to a certain order. In particular, we let $g_j \in \mathcal{H}_j$, the second-order Sobolev space on $\mathcal{X}_j = [0, 1]$, that is, $\mathcal{H}_j = \{g : g, g'\text{are absolutely continuous}, g'' \in L^2[0, 1]\}$. Using the standard theory in functional analysis, one can show that $\mathcal{H}_j$ is a reproducing kernel Hilbert space (RKHS), when equipped with the following norm:

$$\|g_j\|^2_{\mathcal{H}_j} = \left\{\int_0^1 g_j(x)\,dx\right\}^2 + \left\{\int_0^1 g_j'(x)\,dx\right\}^2 + \int_0^1 \{g_j''(x)\}^2\,dx.$$

The reproducing kernel (RK) associated with $\mathcal{H}_j$ is $R(x, z) = R_0(x, z) + R_1(x, z)$ with $R_0(x, z) = k_1(x)k_1(z)$ and $R_1(x, z) = k_2(x)k_2(z) - k_4(x - z)$, where $k_1(x) = x - \frac{1}{2}$, $k_2(x) = \frac{1}{2}\{k_1^2(x) - \frac{1}{12}\}$, and $k_4(x) = \frac{1}{24}\{k_1^4(x) - \frac{1}{2}k_1^2(x) + \frac{7}{240}\}$. See the works of Wahba (1990) and Gu (2002) for more details. Furthermore, the space $\mathcal{H}_j$ has the following orthogonal decomposition:

$$\mathcal{H}_j = \{1\} \oplus \mathcal{H}_{0j} \oplus \mathcal{H}_{1j}, \tag{2.3}$$

where $\{1\}$ is the mean space, $\mathcal{H}_{0j} = \{g_j : g_j''(x) \equiv 0\}$ is the linear contrast subspace, and $\mathcal{H}_{1j} = \{g_j : \int_0^1 g_j(x)\,dx = 0, \int_0^1 g_j'(x)\,dx = 0, g_j'' \in \mathcal{L}_2[0, 1]\}$ is the nonlinear contrast space. Both $\mathcal{H}_{0j}$ and $\mathcal{H}_{1j}$, as subspaces of $\mathcal{H}_j$, are also RKHS and respectively associated with the reproducing kernels R_0 and R_1. Based on the space decomposition (2.3), any function $g_j \in \mathcal{H}_j$ can be correspondingly decomposed into the linear part and nonlinear part

$$g_j(x_j) = b_{0j} + \beta_j\bigl(x_j - \tfrac{1}{2}\bigr) + g_{1j}(x_j), \tag{2.4}$$

where the term $k_1(x_j) = \beta_j(x_j - \frac{1}{2}) \in \mathcal{H}_{0j}$ is the "linear" component and $g_{1j}(x_j) \in \mathcal{H}_{1j}$ is the "nonlinear" component. The fact that $\mathcal{H}_{0j}$ and $\mathcal{H}_{1j}$ are orthogonal to each other assures the uniqueness of this decomposition.

The function $g(\mathbf{x}) = b + g_1(x_{i1}) + \cdots + g_d(x_{id})$ is then estimated in the tensor sum of $\mathcal{H}_j$'s, that is, $\mathcal{H} = \bigoplus_{j=1}^d \mathcal{H}_j$. The decomposition in (2.3) leads to an orthogonal decomposition of $\mathcal{H}$:

$$\begin{aligned}\mathcal{H} = \bigoplus_{j=1}^{d} \mathcal{H}_j &= \{1\} \oplus \bigoplus_{j=1}^{d} \mathcal{H}_{0j} \oplus \bigoplus_{j=1}^{d} \mathcal{H}_{1j} \\ &= \{1\} \oplus \mathcal{H}_0 \oplus \mathcal{H}_1,\end{aligned} \tag{2.5}$$

where $\mathcal{H}_0 = \bigoplus_{j=1}^d \mathcal{H}_{0j}$ and $\mathcal{H}_1 = \bigoplus_{j=1}^d \mathcal{H}_{1j}$. In the next section, we propose a new regularization problem to estimate $g \in \mathcal{H}$ by imposing some penalty on function components, which facilitates the structure selection for the fitted function.

2.2 New Regularization Method: LAND

Throughout the article, we regard a function $g(x)$ as a zero function, that is, $g \equiv 0$, if and only if $g(x) = 0, \forall x \in \mathcal{X}$. With the above setup, we say X_j is a linear covariate if $\beta_j \neq 0$ and $g_{1j} \equiv 0$, and X_j is a nonlinear covariate if $g_{1j}(x_j)$ is not zero. In other words, we can describe the three index sets in the model (2.1) in a more explicit manner:

$$\text{Linear index set:} \quad I_L = \{j = 1, \ldots, d : \beta_j \neq 0, g_{1j} \equiv 0\},$$

$$\text{Nonlinear index set:} \quad I_N = \{j = 1, \ldots, d : g_{1j} \neq 0\},$$

$$\text{Null index set:} \quad I_O = \{j = 1, \ldots, d : \beta_j = 0, g_{1j} \equiv 0\}.$$

Note that the nonlinear index set I_N can be further decomposed as $I_N = I_{PN} \cup I_{LN}$, where $I_{PN} = \{\beta_j = 0, g_{1j} \neq 0\}$ is the index for purely nonlinear terms and $I_{LN} = \{\beta_j \neq 0, g_{1j} \neq 0\}$ is the index for covariates whose linear and nonlinear terms are both nonzero.

The model selection problem for (2.2) is therefore equivalent to the problem of identifying I_L, I_N, I_O. To achieve this, we propose to solve the following regularization problem:

$$\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} [y_i - g(\mathbf{x}_i)]^2 + \lambda_1 \sum_{j=1}^{d} w_{0j} \|\mathcal{P}_{0j} g\|_{\mathcal{H}_0} + \lambda_2 \sum_{j=1}^{d} w_{1j} \|\mathcal{P}_{1j} g\|_{\mathcal{H}_1}, \quad (2.6)$$

where $\mathcal{P}_{0j}$ and $\mathcal{P}_{1j}$ are the projection operators respectively from $\mathcal{H}$ to $\mathcal{H}_{0j}$ and $\mathcal{H}_{1j}$. The regularization term in (2.6) consists of two parts: $\|\mathcal{P}_{0j} g\|_{\mathcal{H}_0} = |\beta_j|$ is equivalent to L_1 penalty on linear coefficients (Tibshirani 1996), and $\|\mathcal{P}_{1j} g\|_{\mathcal{H}_1}$ is the RKHS norm of g_j in $\mathcal{H}_{1j}$. In the context of second-order Sobolev space, we have $\|\mathcal{P}_{1j} g\|_{\mathcal{H}_1} = \{\int_0^1 [g''_{1j}(x)]^2 \, dx\}^{1/2}$. Our theoretical results show that this penalty combination enables the proposed procedure to distinguish linear and nonlinear components automatically. Two tuning parameters (λ_1, λ_2) are used to control overall shrinkage imposed on linear and nonlinear terms. As shown in Section 3, when (λ_1, λ_2) are chosen properly, the resulting estimator is consistent in both structure selection and model estimation. The choices of weights w_{0j} and w_{1j} in (2.6) are discussed in the end of this subsection. We call the new procedure *linear and nonlinear discoverer* (LAND) and denote the solution to (2.6) by $\hat{g}$. The model structure selected by the LAND is defined as

$$\hat{I}_L = \{j : \hat{\beta}_j \neq 0, \hat{g}_{1j} \equiv 0\}, \qquad \hat{I}_N = \{j : \hat{g}_{1j} \neq 0\},$$

$$\hat{I}_O = I \setminus \{\hat{I}_L \cup \hat{I}_N\}.$$

We note that the penalty proposed in (2.6) is related to the COSSO penalty for nonparametric model selection proposed by Lin and Zhang (2006) and Zhang and Lin (2006). The following remark reveals the link and difference between the new penalty and the COSSO penalty.

Remark 1. Denote $J_l(g) = \sum_{j=1}^d \|P_{0j} g\|_{\mathcal{H}_0}$ and $J_n(g) = \sum_{j=1}^d \|\mathcal{P}_{1j} g\|_{\mathcal{H}_1}$. We also denote the COSSO penalty term as $J_c(g) = \sum_{j=1}^d \|\mathcal{P}_j g\|_{\mathcal{H}}$, where $\mathcal{P}_j$ is the projection operator from $\mathcal{H}$ to $\bar{\mathcal{H}}_j = \mathcal{H}_{0j} \oplus \mathcal{H}_{1j}$ and $\| \cdot \|_{\mathcal{H}}$ is the previously defined RKHS norm. Based on $\|\mathcal{P}_j g\|_{\mathcal{H}} = \sqrt{\|\mathcal{P}_{0j} g\|^2_{\mathcal{H}_0} + \|\mathcal{P}_{1j} g\|^2_{\mathcal{H}_1}}$, the Cauchy–Schwarz inequality implies that

$$\frac{J_l(g) + J_n(g)}{\sqrt{2}} \leq J_c(g) \leq J_l(g) + J_n(g)$$

for any $g \in \mathcal{H}$.

The above remark implies that the penalty term in (2.6) includes the COSSO penalty as a special case when equal weights and smoothing parameters are used for regularization. The LAND is much more flexible than the COSSO by employing different weights and smoothing parameters, which makes it possible to distinguish linear and nonlinear components effectively.

The weights w_{0j} and w_{1j} are not tuning parameters as they need to be prespecified by data. We propose to choose the weights adaptively such that unimportant components are assigned with large penalties and important components are given small penalties. In this way, nonzero function components are protectively preserved in the selection process, while insignificant components are shrunk more toward zero. This adaptive selection idea has been employed for linear models in various contexts (Zou 2006; Wang, Li, and Jiang 2007; Zhang and Lu 2007) and SS-ANOVA models (Storlie et al. 2011), and it was found to be able to greatly improve performance of nonadaptive shrinkage methods if the weights are chosen properly. Assume $\widetilde{g}$ is a consistent estimator of g in $\mathcal{H}$. We propose to construct the weights as follows:

$$w_{0j} = \frac{1}{|\widetilde{\beta}_j|^{\alpha}}, \qquad w_{1j} = \frac{1}{\|\widetilde{g}_{1j}\|_2^{\gamma}} \quad \text{for } j = 1, \ldots, d, \qquad (2.7)$$

where $\widetilde{\beta}_j, \widetilde{g}_{1j}$ are the decomposition of $\widetilde{g}$ according to (2.4), $\| \cdot \|_2$ represents the L_2 norm, and $\alpha > 0$ and $\gamma > 0$ are some positive constants. We will discuss how to decide α and γ in Section 3. A natural choice of $\widetilde{g}$ is the standard SS-ANOVA solution, which minimizes the least squares in (2.6) subject to the roughness penalty. Other consistent initial estimators should also work.

Remark 2. The implementation of the LAND procedure requires an initial weight estimation. We point out this two-step process has a different nature from that of classical stepwise selection procedures. In forward or backward selection, variable selection is done sequentially and involves multiple decisions. At each step, the decision is made on whether a covariate should be included or not. These decisions are generally myopic, so the selection errors at previous steps may accumulate and affect later decisions. This explains instability and inconsistency of these stepwise procedures in general. By contrast, the model selection of the LAND is not a sequential decision. It conducts model selection by solving (2.6) once, where all the terms are penalized and shrunken toward zero simultaneously. The initial weights are used to assure the selection consistency of the LAND, which is similar to the adaptive LASSO in linear models.

3. THEORETICAL PROPERTIES

In this section, we first establish the convergence rates of the LAND estimator. Then under the tensor product design, we show that the LAND can identify the correct model structure asymptotically, that is, $\hat{I}_L \to I_L$, $\hat{I}_N \to I_N$, $\hat{I}_O \to I_O$ with probability tending to 1.

To facilitate the presentation, we now define some notations and state the technical assumptions used in our theorems. First, we assume the true partially linear regression is

$$\begin{aligned} y_i &= g_0(\mathbf{x}_i) + \epsilon_i, \\ g_0(\mathbf{x}_i) &= b_0 + \sum_{j \in I_L} x_{ij}\beta_{0j} + \sum_{j \in I_N} f_{0j}(x_{ij}) + \sum_{j \in I_O} 0(x_{ij}), \end{aligned} \tag{3.1}$$

where b_0 is the true intercept, β_{0j}'s are the true coefficients for nonzero linear effects and f_{0j}'s are the true nonzero functions for nonlinear effects. For any $g \in \mathcal{H}$, we decompose $g(\cdot)$ in the framework of function ANOVA:

$$g(\mathbf{x}) = b + \sum_{j=1}^{d} \beta_j k_1(x_j) + \sum_{j=1}^{d} g_{1j}(x_j),$$

where $g_{1j} \in \mathcal{H}_{1j}$. For the purpose of identifiability, we assume that each component has mean zero, that is, $\sum_{i=1}^{n} \beta_j k_1(x_{ij}) + \sum_{i=1}^{n} g_{1j}(x_{ij}) = 0$ for each $j = 1, \ldots, d$. For the final estimator $\hat{g}$, the initial estimator $\widetilde{g}$, and the true function g_0, their ANOVA decomposition can also be expressed in terms of the projection operators. For example, $g_{1j}^0 = \mathcal{P}_{1j} g_0$ for $j = 1, \ldots, d$.

Given data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, for any function $g \in \mathcal{H}$, we denote its function values evaluated at the data points by the n-vector $\mathbf{g} = (g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n))$. Similarly, we define $\mathbf{g}_0$ and $\widetilde{\mathbf{g}}$. Also, define the empirical L_2 norm $\| \cdot \|_n$ and inner product $\langle \cdot, \cdot \rangle_n$ in R^n as

$$\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} g^2(\mathbf{x}_i), \qquad \langle g, h \rangle_n = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i) h(\mathbf{x}_i);$$

and thus $\|y - g\|_n^2 = (1/n) \sum_{i=1}^{n} \{y_i - g(\mathbf{x}_i)\}^2$. For any sequence $r_n \to 0$, we denote $\lambda \sim r_n$ when there exists an $M > 0$ so that $M^{-1} r_n \le \lambda \le M r_n$.

We will establish our theorems for fixed d under the following regularity conditions:

(C1) ϵ is assumed to be independent of $\mathbf{X}$, and has the subexponential tail, that is, $E[\exp(|\epsilon|/C_0)] \le C_0$ for some $0 < C_0 < \infty$

(C2) $\sum_{i=1}^{n} (\mathbf{x}_i - 1/2)(\mathbf{x}_i - 1/2)'/n$ converges to some nonsingular matrix

(C3) the density for $\mathbf{X}$ is bounded away from zero and infinity.

3.1 Asymptotic Properties of the LAND

The choices of weights w_{0j}'s and w_{1j}'s are essential to the LAND procedure. In Section 2, we suggest using the weights constructed from the SS-ANOVA solution $\widetilde{g}$: $w_{0j} = |\widetilde{\beta}_j|^{-\alpha}$ and $w_{1j} = \|\widetilde{g}_{1j}\|_2^{-\gamma}$ for $j = 1, \ldots, d$. The standard smoothing ANOVA $\widetilde{g}$ is obtained by solving

$$\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} [y_i - g(\mathbf{x}_i)]^2 + \lambda \sum_{j=1}^{d} \|\mathcal{P}_{1j} g\|_{\mathcal{H}_1}^2. \tag{3.2}$$

In the following theorem, we show that the LAND estimator has a rate of convergence $n^{-2/5}$ if the tuning parameters are chosen appropriately.

Theorem 1. Under the regularity conditions (C1) and (C2) and the weights stated in (2.7) and (3.2), if $\lambda_1, \lambda_2 \sim n^{-4/5}$ and $\alpha \ge 3/2$, $\gamma \ge 3/2$, then the LAND estimator in (2.6) satisfies:

$$\|\hat{g} - g_0\|_n = O_P\big(n^{-2/5}\big) \quad \text{if } g_0 \text{ is not a constant function}$$

and

$$\|\hat{g} - g_0\|_n = O_P\big(n^{-1/2}\big) \quad \text{if } g_0 \text{ is a constant function.}$$

Remark 3. Theorem 1 is consistent with corollary 1 in the COSSO article (Lin and Zhang 2006) since we assume the same order of two smoothing parameters λ_1 and λ_2. It is worth pointing out that we do not have the optimal parametric rate when the nonparametric component of g is zero. This is not surprising because we still apply the standard nonparametric estimation method, which yields $n^{-2/5}$-rate, even when the true function g is purely linear.

3.2 Selection Consistency

To illustrate the selection consistency of our LAND procedure, we give an instructive analysis in the special case of a tensor product design with a smoothing spline ANOVA model built from the second-order Sobolev spaces of periodic functions. For simplicity, we assume that the error ϵ's in the regression model are independent with the distribution $\mathrm{N}(0, \sigma^2)$ here. The space of periodic functions on $[0, 1]$ is denoted by $\mathcal{H}_{per} = \{1\} \oplus \bigoplus_{j=1}^{d} \mathcal{H}_{0j} \oplus \bigoplus_{j=1}^{d} \mathcal{S}_{per,j}$, where $\mathcal{S}_{per,j}$ is the functional space $\mathcal{S}_{per}$ on x_j, and

$$\mathcal{S}_{per} = \Bigg\{ f : f(t) = \sum_{v=1}^{\infty} a_v \sqrt{2} \cos(2\pi v t) + \sum_{v=1}^{\infty} b_v \sqrt{2} \sin(2\pi v t), \\ \text{with } \sum_{v=1}^{\infty} (a_v^2 + b_v^2)(2\pi v t)^4 < \infty \Bigg\}.$$

We also assume that the observations come from a tensor product design, that is,

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d\},$$

where $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n_j,j})'$ and $x_{i,j} = i/n_j$, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$. Without loss of generality, we assume that n_j equals some number m for any $j = 1, \ldots, d$.

Theorem 2. Assume a tensor product design and $g_0 \in \mathcal{H}_{per}$. Under the regularity conditions (C1) to (C3), assume that (i) $n^{1/5} \lambda_1 w_{0j} \to \infty$ for $j \in I \setminus I_L$; (ii) $n^{3/20} \lambda_2^2 w_{1j}^2 \to \infty$ for $j \in I \setminus I_N$, we have $\hat{I}_L = I_L$, $\hat{I}_N = I_N$, $\hat{I}_O = I_O$ with probability tending to 1 as $n \to \infty$.

Remark 4. To achieve the structure selection consistency and convergence rate in Theorem 1 simultaneously, we require that $\lambda_1, \lambda_2 \sim n^{-4/5}$, $\alpha > 3$, $\gamma > 29/8$, by considering the assumptions in Theorems 1 and 2 and Lemma A.1 in the Appendix if we use the weight of the form (2.7).

Remark 5. The proof of the selection consistency requires detailed investigation on eigen-properties of the reproducing kernel, which is generally intractable. In Theorem 2, we assume that the function belongs to the class of periodic functions and $\mathbf{x}$ has a tensor product design. This makes our derivation more tractable, since the eigenfunctions and eigenvalues of the RK for $\mathcal{H}_{per}$ have particularly simple forms. Results for this specific design are often instructive for general designs, as suggested by Wahba (1990). We conjecture that the LAND is still selection consistent in general cases. This is also supported by numerical results in Section 5, where neither the tensor product design nor the periodic function is assumed in the examples. Note that the special design condition is not required for the convergence rate results in Theorem 1.

4. TWO–STEP LAND ESTIMATOR

As shown in Section 3, the LAND estimator can consistently identify the true structure of partially linear models. In other words, the selected model would be correct as the sample size goes to infinity. In finite sample situations, if the selected model is correct or approximately correct, it is natural to ask whether refitting data based on the selected model would improve model estimation. This leads to the two-step LAND procedure: at step I, we identify the model structure using the LAND, and at step II we refit data by using the selected model from step I. In particular, we fit the following model at the second step:

$$y_i = b + \sum_{j \in \hat{I}_L} \beta_j k_1(x_{ij}) + \sum_{j \in \hat{I}_N} g_{1j}(x_{ij}) + \sum_{j \in \hat{I}_O} 0(x_{ij}) + \epsilon_i, \quad (4.1)$$

where $(\hat{I}_L, \hat{I}_N, \hat{I}_O)$ are the index sets identified by $\hat{g}$. Denote the *two-step* LAND solution by $\hat{g}^*$. The rationale behind the two-step LAND is: if the selection in step I is very accurate, then the estimation of $\hat{g}^*$ can be thought of as being based on a (approximately) correct model. This two-step procedure thus will yield better estimation accuracy as shown in the next paragraph.

Let $\Omega_n = \{I_L = \hat{I}_L \text{ and } I_N = \hat{I}_N\}$. In the first step, we estimate I_L and I_N consistently, that is, $P(\Omega_n) \to 1$, according to Theorem 2. In the second step, we fit a partial smoothing spline in (4.1). Denote the solution as $\hat{\boldsymbol{\beta}}^*$ and $\hat{g}^*_{1j}$. Within the event Ω_n, that is, $\hat{I}_L = I_L$ and $\hat{I}_N = I_N$, we know that, by the standard partial smoothing spline theory (Mammen and van de Geer 1997),

$$\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\| = O_P\big(n^{-1/2}\big), \quad (4.2)$$

$$\|\hat{g}^*_{1j} - g^0_{1j}\|_2 = O_P\big(n^{-2/5}\big), \quad (4.3)$$

under regularity conditions. In addition, we know that $\hat{\boldsymbol{\beta}}^*$ is also asymptotically normal within the event Ω_n. Since Ω_n is shown to have probability tending to 1, we can conclude that (4.2) and (4.3) hold asymptotically. Moreover, comparing (4.2)–(4.3) with Theorem 1, we conclude that the convergence rates of both linear and nonlinear components can be further improved to their optimal rates by implementing the above two-step procedure.

In Section 6, we find that the LAND and two-step LAND perform similarly in many cases. If the LAND does a good job in recovering the true model structure correctly, say in strong signal cases, then the additional refitting step can improve the model estimation accuracy. However, if the selection result is not good, say, in weak signal cases, the refitting result is not necessarily better.

5. COMPUTATION ALGORITHMS

5.1 Equivalent Formulation

We first show that the solution to (2.6) lies in a finite-dimensional space. This is an important result for nonparametric modeling, since the LAND estimator involves solving an optimization problem in an infinite-dimensional space $\mathcal{H}$. The finite representer property is known to hold for standard SS-ANOVA models (Kimeldorf and Wahba 1971) and partial splines (Gu 2002).

Lemma 1. Let $\hat{g}(\mathbf{x}) = \hat{b} + \sum_{j=1}^d \hat{\beta}_j k_1(x_j) + \sum_{j=1}^d \hat{g}_{1j}(x_j)$ be a minimizer of (2.6) in $\mathcal{H}$, with $\hat{g}_{1j} \in \mathcal{H}_{1j}$ for $j = 1, \ldots, d$. Then $\hat{g}_{1j} \in \text{span}\{R_{1j}(x_i, \cdot), i = 1, \ldots, n\}$, where $R_{1j}(\cdot, \cdot)$ is the reproducing kernel of the space $\mathcal{H}_{1j}$.

To facilitate the LAND implementation, we give an equivalent but more convenient formulation to (2.6). Define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\mathrm{T}}$. Consider the optimization problem:

$$\min_{\boldsymbol{\theta} \geq \mathbf{0}, g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n [y_i - g(\mathbf{x}_i)]^2 + \lambda_1 \sum_{j=1}^d w_{0j} \|\mathcal{P}_{0j} g\|$$
$$+ \tau_0 \sum_{j=1}^d \theta_j^{-1} w_{1j} \|\mathcal{P}_{1j} g\|^2_{\mathcal{H}_1} + \tau_1 \sum_{j=1}^d w_{1j} \theta_j,$$
$$\text{subject to} \quad \theta_j \geq 0, j = 1, \ldots, d, \quad (5.1)$$

where τ_0 is a constant that can be fixed at any positive value, and (λ_1, τ_1) are tuning parameters. The following lemma shows that there is a one-to-one correspondence between the solutions to (2.6) [for all possible pairs (λ_1, λ_2)] and those to (5.1) [for all (λ_1, τ_1) pairs].

Lemma 2. Set $\tau_1 = \lambda_2^2/(4\tau_0)$. (i) If $\hat{g}$ minimizes (2.6), set $\hat{\theta}_j = \tau_0^{1/2} \tau_1^{-1/2} \|\mathcal{P}_{1j} \hat{g}\|$; then the pair $(\hat{\boldsymbol{\theta}}, \hat{g})$ minimizes (5.1). (ii) If $(\hat{\boldsymbol{\theta}}, \hat{g})$ minimizes (5.1), then $\hat{g}$ minimizes (2.6).

In practice, we choose to solve (5.1) since its objective function can be easily handled by standard quadratic programming (QP) and linear programming (LP) techniques. The nonnegative θ_j's can be regarded as scaling parameters and they are interpretable for the purpose of model selection. If $\theta_j = 0$, the minimizer of (5.1) is taken to satisfy $\|\mathcal{P}_{1j} g\| = 0$, which implies that the nonlinear component of g_j vanishes.

With $\boldsymbol{\theta}$ fixed, solving (5.1) is equivalent to fitting a partial spline model in some RKHS space. By the representer theorem, the solution to (5.1) has the following form:

$$\hat{g}(\mathbf{x}) = \hat{b} + \sum_{j=1}^d \hat{\beta}_j k_1(x_j) + \sum_{j=1}^d \hat{\theta}_j w_{1j}^{-1} \sum_{i=1}^n \hat{c}_i R_{1j}(x_{ij}, x_j). \quad (5.2)$$

The expression (5.2) suggests that the linearity or nonlinearity of g_j is determined by the fact whether $\hat{\beta}_j = 0$ or $\hat{\theta}_j = 0$ or not. Therefore, we can define the three index sets as:

$$\hat{I}_L = \{j : \hat{\beta}_j \neq 0, \hat{\theta}_j = 0\}, \qquad \hat{I}_N = \{j : \hat{\theta}_j \neq 0\},$$
$$\hat{I}_O = \{j : \hat{\beta}_j = 0, \hat{\theta}_j = 0\}.$$

5.2 Algorithms

In the following, we propose an iterative algorithm to solve (5.1). Define the vectors $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{g} = (g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n))^{\mathrm{T}}$, $\boldsymbol{\beta} = (b, \beta_1, \ldots, \beta_d)^{\mathrm{T}}$, and $\mathbf{c} = (c_1, \ldots, c_n)^{\mathrm{T}} \in R^n$. With some abuse of notations, let $\mathbf{R}_{1j}$ also stand for the $n \times n$ matrix $\{R_{1j}(x_{ij}, x_{i'j})\}$, for $i, i' = 1, \ldots, n; j = 1, \ldots, d$, and $\mathbf{R}_{\mathbf{w}_1, \boldsymbol{\theta}} = \sum_{j=1}^{d} \theta_j w_{1j}^{-1} \mathbf{R}_{1j}$ be the Gram matrix associated with the weighted kernel. Let $\mathbf{T}$ be the $n \times (1 + d)$ matrix with $t_{i1} = 1$ and $t_{ij} = k_1(x_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Then $\mathbf{g} = \mathbf{T}\boldsymbol{\beta} + \mathbf{R}_{\mathbf{w}_1,\boldsymbol{\theta}}\mathbf{c}$, and (5.1) can be expressed as

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{c}} \frac{1}{n}\big(\mathbf{y} - \mathbf{T}\boldsymbol{\beta} - \mathbf{R}_{\mathbf{w}_1,\boldsymbol{\theta}}\mathbf{c}\big)^{\mathrm{T}}\big(\mathbf{y} - \mathbf{T}\boldsymbol{\beta} - \mathbf{R}_{\mathbf{w}_1,\boldsymbol{\theta}}\mathbf{c}\big) + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| + \tau_0 \mathbf{c}^{\mathrm{T}}\mathbf{R}_{\mathbf{w}_1,\boldsymbol{\theta}}\mathbf{c} + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j,$$
$$\text{s.t.} \quad \theta_j \geq 0,\ j = 1, \ldots, d. \tag{5.3}$$

To solve (5.3), we suggest an iterative algorithm to alternatively update $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\mathbf{c}$.

On one hand, with $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ fixed at their current values, we update $\mathbf{c}$ by the following ridge-type problem: define $\mathbf{z} = \mathbf{y} - \mathbf{T}\hat{\boldsymbol{\beta}}$ and solve

$$\min_{\mathbf{c}} \frac{1}{n}\big(\mathbf{z} - \mathbf{R}_{\mathbf{w}_1,\hat{\boldsymbol{\theta}}}\mathbf{c}\big)^{\mathrm{T}}\big(\mathbf{z} - \mathbf{R}_{\mathbf{w}_1,\hat{\boldsymbol{\theta}}}\mathbf{c}\big) + \tau_0 \mathbf{c}^{\mathrm{T}}\mathbf{R}_{\mathbf{w}_1,\hat{\boldsymbol{\theta}}}\mathbf{c}. \tag{5.4}$$

On the other hand, when $\hat{\mathbf{c}}$ is fixed at their current values, we can update $(\boldsymbol{\beta}, \boldsymbol{\theta})$ by solving a quadratic programming (QP) problem. Define $\mathbf{v}_j = w_{1j}^{-1}\mathbf{R}_{1j}\hat{\mathbf{c}}$ for $j = 1, \ldots, d$ and let $\mathbf{V}$ be the $n \times d$ matrix with the jth column being $\mathbf{v}_j$. Then we obtain the following problem:

$$\min_{\boldsymbol{\theta} \geq \mathbf{0},\boldsymbol{\beta}} \frac{1}{n}(\mathbf{y} - \mathbf{T}\boldsymbol{\beta} - \mathbf{V}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - \mathbf{T}\boldsymbol{\beta} - \mathbf{V}\boldsymbol{\theta}) + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| + \tau_0 \hat{\mathbf{c}}^{\mathrm{T}}\mathbf{V}\boldsymbol{\theta} + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j. \tag{5.5}$$

Further, we can write $|\beta_j| = \beta_j^+ + \beta_j^-$ and $\beta_j = \beta_j^+ - \beta_j^-$ for each j, where β_j^+ and β_j^- are respectively the positive and negative part of β_j. Define $\mathbf{w}_0 = (w_{01}, \ldots, w_{0d})^{\mathrm{T}}$. Then (5.5) can be equivalently expressed as

$$\min_{\boldsymbol{\theta},\boldsymbol{\beta}^+,\boldsymbol{\beta}^-} \frac{1}{n}(\mathbf{y} - T\boldsymbol{\beta}^+ + T\boldsymbol{\beta}^- - \mathbf{V}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - T\boldsymbol{\beta}^+ + T\boldsymbol{\beta}^- - \mathbf{V}\boldsymbol{\theta}) + \lambda_1 \mathbf{w}_0^{\mathrm{T}}(\boldsymbol{\beta}^+ + \boldsymbol{\beta}^-) + \tau_0 \hat{\mathbf{c}}^{\mathrm{T}}\mathbf{V}\boldsymbol{\theta},$$
$$\text{subject to} \quad \sum_{j=1}^{d} w_{1j}\theta_j \leq M, \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\beta}^+ \geq \mathbf{0}, \boldsymbol{\beta}^- \geq \mathbf{0} \tag{5.6}$$

for some $M > 0$. Given any (λ_1, M), the following is a complete algorithm to compute $\hat{g}$.

Algorithm.

Step 0: Obtain the initial estimator $\widetilde{g}$ by fitting a standard SS-ANOVA model. Derive $\widetilde{\beta}_j, \widetilde{g}_{1j}, j = 1, \ldots, d$, and compute the weights $w_{0j}, w_{1j}, j = 1, \ldots, d$, using (2.7).

Step 1: Initialize $\hat{\boldsymbol{\theta}} = \mathbf{1}_d$ and $\hat{\beta}_j = \widetilde{\beta}_j, j = 1, \ldots, d$.

Step 2: Fixing $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ at their current values, update $\mathbf{c}$ by solving (5.4).

Step 3: Fixing $\hat{\mathbf{c}}$ at their current values, update $(\boldsymbol{\theta}, \boldsymbol{\beta})$ by solving (5.6).

Step 4: Go to step 2 until the convergence criterion meets.

6. NUMERICAL STUDIES

In this section, we demonstrate the empirical performance of the LAND estimators in terms of their estimation accuracy and model selection. We compare the LAND with GAM, SS-ANOVA, COSSO, and the two-step LAND (2LAND). Note that LAND and 2LAND procedures give identical performance for model selection. The GAM and COSSO fits were obtained using the R packages "gam" and "cosso," respectively. The built-in tuning procedures in R packages are used to tune the associated tuning parameters.

The following functions on [0, 1] are used as building blocks of functions in simulations:

$$h_1(x) = x, \qquad h_2(x) = \cos(2\pi x),$$
$$h_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x)),$$
$$h_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3(\sin(2\pi x))^2 + 0.4(\cos(2\pi x))^3 + 0.5(\sin(2\pi x))^3,$$
$$h_5(x) = (3x - 1)^2.$$

For each function, we can examine whether it is a pure linear, pure nonlinear, or both linear and nonlinear function based on its functional ANOVA decomposition in (2.4). Simple calculation shows that h_1 is a pure linear function, h_2, h_3, and h_4 are pure nonlinear functions, and h_5 contains both nonzero linear and nonlinear terms.

For the simulation design, we consider four different values of theoretical R^2 as $R^2 = 0.95, 0.75, 0.55, 0.35$, providing varying signal-to-noise ratio (SNR) settings. For the input $\mathbf{x}$, we consider both uncorrelated and correlated situations, corresponding to $\rho = \mathrm{corr}(X_i, X_j) = 0, 0.5, 0.8$ for all $i \neq j$. The combination of four levels of R^2 and three levels of ρ produces twelve unique SNR settings.

To evaluate the model estimation performance of the estimator $\hat{g}$, we report its integrated squared error $\mathrm{ISE} = E_{\mathbf{X}}\{g(\mathbf{X}) - \hat{g}(\mathbf{X})\}^2$. The ISE is calculated via a Monte Carlo integration with 1000 points. For each procedure, we report the average ISEs over 100 realizations and the corresponding standard errors (in parentheses). To evaluate performance of the LAND in structure selection, we summarize the frequency of getting the correct model structure (power) and an incorrect model structure (Type I error) over 100 Monte Carlo simulations. In particular, the power related measures include:

(i) the number of correct linear effects identified (denoted as "corrL")

(ii) the number of correct nonlinear effects identified (denoted as "corrN")

(iii) the number of correct linear and nonlinear effects identified (denoted as "corrLN")

(iv) the number of correct zero coefficients identified (denoted as "corr0").

The Type I error related measures include:

(i) the number of linear effects incorrectly identified as nonlinear effects (denoted as "LtoN")

(ii) the number of nonlinear effects incorrectly identified as linear effects (denoted as "NtoL")

(iii) the number of linear or nonzero effects incorrectly identified as zero (denoted as "LNto0").

The selection of tuning parameters is an important issue. Our empirical experience suggests that the performance of the LAND procedures is not sensitive to γ and α. We recommend to use $\gamma = \alpha = 4$ based on Remark 4 and they work well in our examples. The choices of (λ_1, λ_2) [or (λ_1, M), equivalently] are important, as their magnitude directly controls the amount of penalty and the model sparsity. The numerical results are quite sensitive to λ's. Therefore, we recommend to select the optimal parameters using cross-validation or some information criteria. In our simulation, we generate a validation set of size n from the same distribution of the training set. For each pair of tuning parameters, we implement the procedure and evaluate its prediction error on the validation set. We select the pair of λ_1 and λ_2 (or M) which corresponds to the minimum validation error.

6.1 Example 1

We generate Y from the model

$$Y = 3h_1(X_1) + 2h_2(X_2) + 2h_5(X_3) + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$. The pairwise correlation $\mathrm{corr}(X_j, X_k) = \rho$ for any $j \neq k$. We consider three cases: $\rho = 0, 0.5, 0.8$. In this model, there are one purely linear effect, one purely nonlinear effect, one linear-nonlinear effect, and $d - 3$ noise variables. We consider $d = 10$ and $d = 20$, and the number of noise variables increases as d increases.

Table 1 summarizes the ISEs of all the procedures in twelve settings. To set a baseline for comparison, we also include the oracle model which fits the data using the true model structure. The 2LAND consistently produces smaller ISEs than GAM and SS-ANOVA in all the settings. The LAND is better than GAM and SS-ANOVA in most settings. We also note that the LAND and 2LAND perform similarly in the independent case. When the covariates are correlated at some degree, 2LAND tends to give better ISEs than the LAND as long as the signal is not too weak. The comparison between the LAND methods and COSSO is quite interesting. When R^2 is moderately large, say 0.75 and 0.95, the 2LAND overall gives smaller or comparable ISEs; if R^2 is small, say 0.55 and 0.35, the COSSO gives smaller errors. This pattern is actually not surprising, as the COSSO and LAND aim to tackle different problems. The COSSO can distinguish zero and nonzero components, while the LAND can distinguish zero, linear, and nonlinear components. Since the LAND methods are designed to discover a more detailed model structure than the COSSO, they generally estimate the function better if they can correctly separate different terms, which often require relatively stronger signals in data. The main advantage of the LAND methods is to produce more interpretable models by automatically separating linear and nonlinear terms, while other methods can not achieve this.

Figure 1 plots the estimated function components by the SS-ANOVA and the 2LAND in one typical realization of Example 1. For illustration, we plot the first four function components. In each panel, the solid, dashed, dotted lines respectively

Table 1. Average ISEs (and standard errors in parentheses) for 100 runs in Example 1

ρ	d	R^2	GAM	SS-ANOVA	COSSO	LAND	2LAND	Oracle
0	10	0.95	0.17 (0.01)	0.11 (0.01)	0.11 (0.01)	0.05 (0.01)	0.06 (0.00)	0.06 (0.00)
		0.75	0.91 (0.05)	0.56 (0.03)	0.48 (0.03)	0.35 (0.03)	0.39 (0.02)	0.27 (0.02)
		0.55	2.17 (0.12)	1.31 (0.07)	1.07 (0.07)	1.28 (0.10)	1.12 (0.08)	0.61 (0.05)
		0.35	4.73 (0.28)	2.95 (0.17)	2.44 (0.15)	3.28 (0.18)	2.94 (0.18)	1.34 (0.11)
	20	0.95	0.50 (0.01)	0.19 (0.01)	0.18 (0.01)	0.05 (0.01)	0.07 (0.01)	0.06 (0.00)
		0.75	2.48 (0.07)	1.04 (0.04)	0.82 (0.04)	0.46 (0.03)	0.60 (0.03)	0.25 (0.02)
		0.55	5.92 (0.17)	2.46 (0.09)	2.01 (0.11)	1.81 (0.11)	1.89 (0.10)	0.55 (0.05)
		0.35	13.29 (0.39)	5.60 (0.22)	4.18 (0.17)	5.18 (0.22)	5.10 (0.22)	1.17 (0.11)
0.5	10	0.95	0.16 (0.01)	0.11 (0.01)	0.10 (0.01)	0.09 (0.01)	0.06 (0.01)	0.06 (0.00)
		0.75	0.87 (0.05)	0.57 (0.03)	0.45 (0.03)	1.06 (0.06)	0.48 (0.03)	0.27 (0.02)
		0.55	2.08 (0.11)	1.35 (0.06)	1.07 (0.07)	1.94 (0.09)	1.33 (0.08)	0.61 (0.04)
		0.35	4.68 (0.27)	2.97 (0.14)	2.44 (0.15)	3.32 (0.14)	2.99 (0.15)	1.37 (0.09)
	20	0.95	0.41 (0.01)	0.19 (0.01)	0.16 (0.01)	0.24 (0.03)	0.07 (0.01)	0.06 (0.00)
		0.75	2.27 (0.05)	1.02 (0.04)	0.74 (0.04)	1.06 (0.07)	0.67 (0.04)	0.24 (0.02)
		0.55	5.48 (0.13)	2.46 (0.09)	1.86 (0.09)	2.42 (0.10)	2.08 (0.08)	0.55 (0.02)
		0.35	12.36 (0.28)	5.46 (0.20)	3.60 (0.15)	4.97 (0.20)	5.11 (0.20)	1.21 (0.12)
0.8	10	0.95	0.16 (0.01)	0.11 (0.01)	0.10 (0.01)	0.28 (0.01)	0.07 (0.01)	0.06 (0.00)
		0.75	0.89 (0.05)	0.56 (0.03)	0.47 (0.03)	1.07 (0.05)	0.52 (0.03)	0.27 (0.02)
		0.55	2.15 (0.11)	1.35 (0.06)	1.06 (0.07)	1.87 (0.10)	1.35 (0.07)	0.61 (0.04)
		0.35	4.74 (0.25)	2.98 (0.14)	2.29 (0.13)	3.33 (0.14)	2.94 (0.14)	1.36 (0.09)
	20	0.95	0.42 (0.01)	0.19 (0.01)	0.16 (0.01)	0.22 (0.03)	0.07 (0.05)	0.06 (0.01)
		0.75	2.24 (0.05)	1.04 (0.04)	0.76 (0.04)	1.11 (0.07)	0.76 (0.05)	0.24 (0.02)
		0.55	5.40 (0.12)	2.50 (0.10)	1.88 (0.09)	2.61 (0.11)	2.25 (0.11)	0.55 (0.05)
		0.35	12.17 (0.27)	5.59 (0.22)	3.47 (0.14)	5.31 (0.19)	5.49 (0.22)	1.20 (0.12)

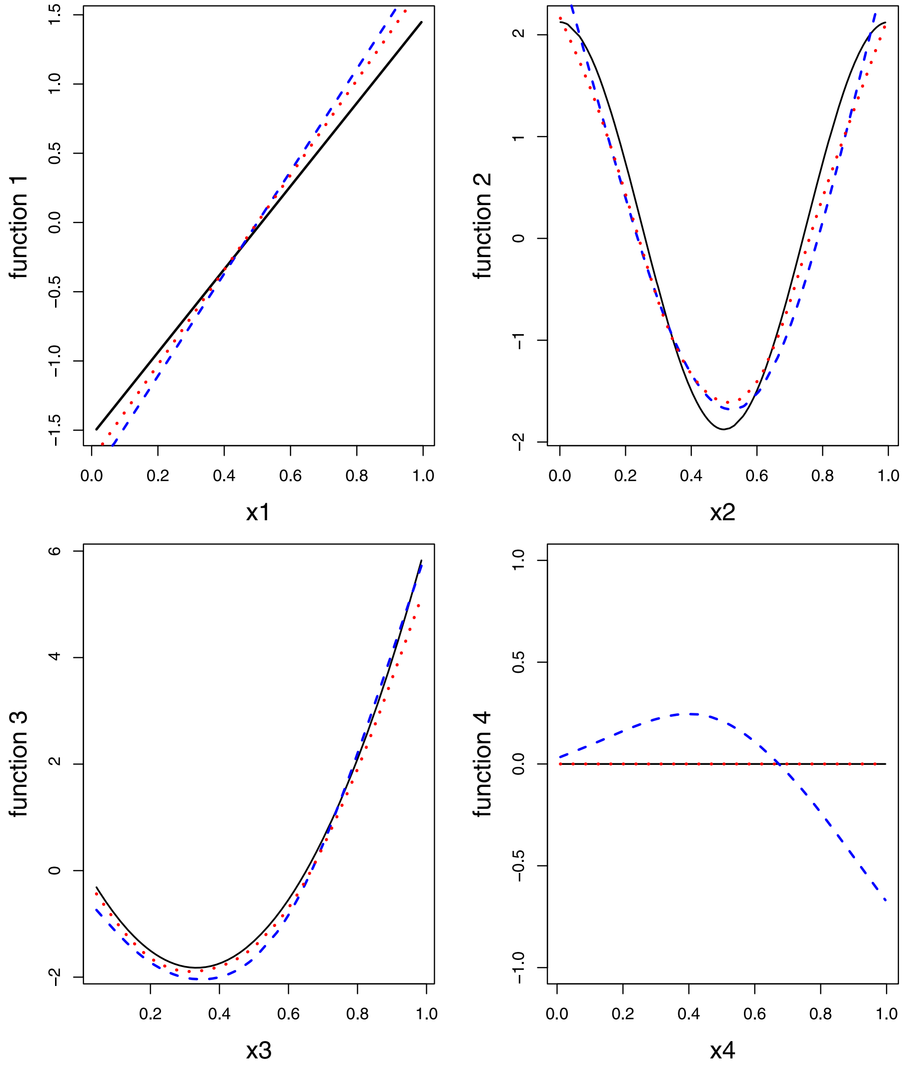

Figure 1. True and estimated function components for Example 1: True function (solid line), SS-ANOVA estimator (dashed line), and 2LAND estimator (dotted line). The online version of this figure is in color.

represent the true function, the fit by SS-ANOVA, and the fit by 2LAND. We observe that both SS-ANOVA and 2LAND perform well in the first three panels, and 2LAND shows better accuracy in estimation than SS-ANOVA by producing a sparse model. In the last panel where the true function is zero, the 2LAND successfully removes it from the final model while the SS-ANOVA provides a nonzero fit.

Table 2 reports the selection performance of the LAND under different settings. Note that the 2LAND is identical to the LAND for model selection. We observe that the LAND shows effective performance in terms of both power and Type-I error measures in all the settings. When the signal is moderately strong, the LAND is able to identify the correct model with high frequency since the “corrL,” “corrN,” “corrLN,” and “corr0” are all close to their true values and the incorrectly selected terms are close to zero. Except in weak signal cases, the frequency of missing any important variable or treating linear terms as nonlinear is low. In more challenging cases, with a small R^2 or a large number of noise variables, the LAND selection gets worse as expected but still performs reasonably well, considering that the sample size $n = 100$ is small.

6.2 Example 2

We modify Example 1 into a more challenging example, which contains a larger number of input variables and a more complex structure for the underlying model. In particular, let $d = 20$. Similarly to Example 1, we consider uncorrelated covariates, correlated covariates with pairwise correlations $\rho = 0, 0.5, 0.8$ respectively. The response Y is generated from the following model:

$$Y = 3h_1(X_1) - 4h_1(X_2) + 2h_1(X_3) + 2h_2(X_4) + 3h_3(X_5) + (5h_4(X_6) + 2h_1(X_6)) + 2h_5(X_7) + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$. In this case, the first three covariates X_1, X_2, and X_3 have purely linear effects, the covariates X_4 and X_5 have purely nonlinear effects, and the covariates X_6 and X_7 have nonzero linear and nonlinear terms. There are $d - 7$ noise variables, and let $n = 250$.

Table 3 summarizes the prediction errors of various estimators under different settings. We consider four different values of theoretical R^2 as $R^2 = 0.95, 0.75, 0.55, 0.35$, which provide different signal-to-noise ratio (SNR) values and hence varying

Table 2. Average selection results (standard errors in parentheses) for 100 runs in Example 1

		R^2	corrlin	cornon	corrlnn	corr0	LNto0	LtoN	NtoL
ρ	d	oracle:	1	1	1	$d-3$	0	0	0
0	10	0.95	0.99 (0.01)	0.90 (0.03)	1.00 (0.00)	6.34 (0.18)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)
		0.75	0.99 (0.01)	0.71 (0.05)	1.00 (0.00)	5.23 (0.20)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)
		0.55	0.99 (0.01)	0.51 (0.05)	0.97 (0.02)	3.97 (0.19)	0.02 (0.01)	0.01 (0.01)	0.05 (0.02)
		0.35	0.92 (0.03)	0.33 (0.05)	0.74 (0.04)	2.33 (0.17)	0.10 (0.03)	0.05 (0.02)	0.43 (0.06)
	20	0.95	1.00 (0.00)	0.94 (0.02)	1.00 (0.00)	16.24 (0.27)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		0.75	1.00 (0.00)	0.71 (0.05)	1.00 (0.00)	13.38 (0.29)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		0.55	0.97 (0.02)	0.53 (0.05)	0.94 (0.02)	9.30 (0.03)	0.04 (0.02)	0.01 (0.01)	0.07 (0.03)
		0.35	0.92 (0.03)	0.31 (0.05)	0.67 (0.05)	5.80 (0.31)	0.10 (0.03)	0.02 (0.01)	0.49 (0.06)
0.5	10	0.95	1.00 (0.00)	0.92 (0.03)	0.95 (0.02)	6.49 (0.14)	0.00 (0.00)	0.00 (0.00)	0.05 (0.02)
		0.75	1.00 (0.00)	0.67 (0.05)	0.64 (0.05)	4.78 (0.18)	0.03 (0.02)	0.00 (0.00)	0.37 (0.05)
		0.55	0.95 (0.02)	0.43 (0.05)	0.48 (0.05)	3.23 (0.17)	0.13 (0.04)	0.01 (0.01)	0.61 (0.05)
		0.35	0.88 (0.03)	0.19 (0.04)	0.39 (0.05)	2.24 (0.15)	0.20 (0.04)	0.06 (0.02)	0.82 (0.06)
	20	0.95	1.00 (0.00)	0.91 (0.03)	0.97 (0.02)	16.23 (0.23)	0.00 (0.00)	0.00 (0.00)	0.03 (0.02)
		0.75	0.99 (0.01)	0.66 (0.05)	0.66 (0.05)	12.32 (0.29)	0.03 (0.02)	0.00 (0.00)	0.36 (0.05)
		0.55	0.90 (0.03)	0.48 (0.05)	0.46 (0.05)	8.14 (0.34)	0.12 (0.03)	0.02 (0.01)	0.65 (0.06)
		0.35	0.85 (0.04)	0.16 (0.04)	0.30 (0.05)	5.40 (0.28)	0.25 (0.04)	0.03 (0.02)	0.96 (0.07)
0.8	10	0.95	1.00 (0.00)	0.88 (0.03)	0.94 (0.02)	6.41 (0.13)	0.00 (0.00)	0.00 (0.00)	0.06 (0.02)
		0.75	1.00 (0.00)	0.61 (0.05)	0.66 (0.05)	4.44 (0.20)	0.05 (0.02)	0.00 (0.00)	0.35 (0.05)
		0.55	0.92 (0.03)	0.36 (0.05)	0.48 (0.05)	3.08 (0.17)	0.18 (0.04)	0.00 (0.00)	0.62 (0.05)
		0.35	0.91 (0.03)	0.16 (0.04)	0.42 (0.05)	1.88 (0.14)	0.15 (0.04)	0.03 (0.02)	0.88 (0.06)
	20	0.95	1.00 (0.00)	0.94 (0.02)	0.97 (0.02)	16.15 (0.23)	0.00 (0.00)	0.00 (0.00)	0.03 (0.02)
		0.75	0.95 (0.02)	0.64 (0.05)	0.72 (0.05)	11.20 (0.29)	0.07 (0.03)	0.00 (0.00)	0.31 (0.05)
		0.55	0.88 (0.03)	0.41 (0.05)	0.43 (0.05)	7.22 (0.31)	0.17 (0.04)	0.02 (0.01)	0.67 (0.06)
		0.35	0.81 (0.04)	0.17 (0.04)	0.26 (0.04)	4.78 (0.26)	0.29 (0.06)	0.04 (0.02)	1.05 (0.07)

signal strength. We have similar observations as in Example 1. The LAND and 2LAND give similar performance, and both of them consistently produce smaller ISEs than GAM and SS-ANOVA in all the settings. The ISEs of the LAND and 2LAND are significantly better than that of the COSSO in all the cases except $R^2 = 0.35$, where the signal is very weak. In Table 4, we report the structure selection performance of the LAND under different settings. Overall, the LAND gives an effective performance as long as the signal is not too weak.

In Figure 2, we plot the estimated functions given by SS-ANOVA and 2LAND for one typical realization of Example 2. Again, with the feature of automatic selection, 2LAND delivers overall better estimation than SS-ANOVA. In the last panel, the SS-ANOVA provides a nonzero fit to a zero component function, while the 2LAND successfully detects the variable as unimportant.

In Table 4, we report the structure selection performance of the LAND under different settings. Similarly to Example 1, we observe that the LAND overall gives effective performance in all the settings. When the signal is moderately strong, the LAND procedure is able to identify the correct model with a high frequency and the incorrectly selected terms are close to zero. When the signal becomes quite weak, the LAND performance gets worse but is still reasonable.

Table 3. Average ISEs (and standard errors in parentheses) for 100 runs in Example 2

ρ	R^2	GAM	SS-ANOVA	COSSO	LAND	2LAND	Oracle
0	0.95	1.07 (0.01)	0.16 (0.01)	0.25 (0.01)	0.07 (0.00)	0.08 (0.00)	0.07 (0.00)
	0.75	2.16 (0.05)	0.87 (0.03)	1.10 (0.05)	0.43 (0.03)	0.46 (0.03)	0.33 (0.02)
	0.55	4.14 (0.10)	2.07 (0.08)	2.19 (0.09)	1.51 (0.09)	1.54 (0.10)	0.72 (0.05)
	0.35	8.47 (0.22)	4.54 (0.17)	4.13 (0.17)	3.90 (0.17)	3.83 (0.18)	1.58 (0.10)
0.5	0.95	0.91 (0.01)	0.18 (0.01)	0.22 (0.01)	0.09 (0.01)	0.10 (0.00)	0.10 (0.00)
	0.75	1.95 (0.04)	0.93 (0.04)	1.07 (0.04)	0.81 (0.05)	0.71 (0.04)	0.42 (0.02)
	0.55	3.83 (0.09)	2.13 (0.07)	2.10 (0.07)	2.02 (0.09)	1.99 (0.02)	0.89 (0.05)
	0.35	7.92 (0.20)	4.31 (0.15)	3.74 (0.13)	4.16 (0.15)	4.12 (0.15)	1.66 (0.10)
0.8	0.95	0.93 (0.01)	0.19 (0.01)	0.24 (0.01)	0.10 (0.00)	0.11 (0.01)	0.10 (0.00)
	0.75	2.02 (0.04)	0.98 (0.03)	1.18 (0.04)	0.85 (0.04)	0.80 (0.04)	0.47 (0.02)
	0.55	3.96 (0.10)	2.29 (0.07)	2.26 (0.07)	2.15 (0.09)	2.16 (0.14)	0.99 (0.05)
	0.35	8.16 (0.21)	4.61 (0.15)	3.82 (0.13)	4.19 (0.15)	4.38 (0.17)	1.85 (0.10)

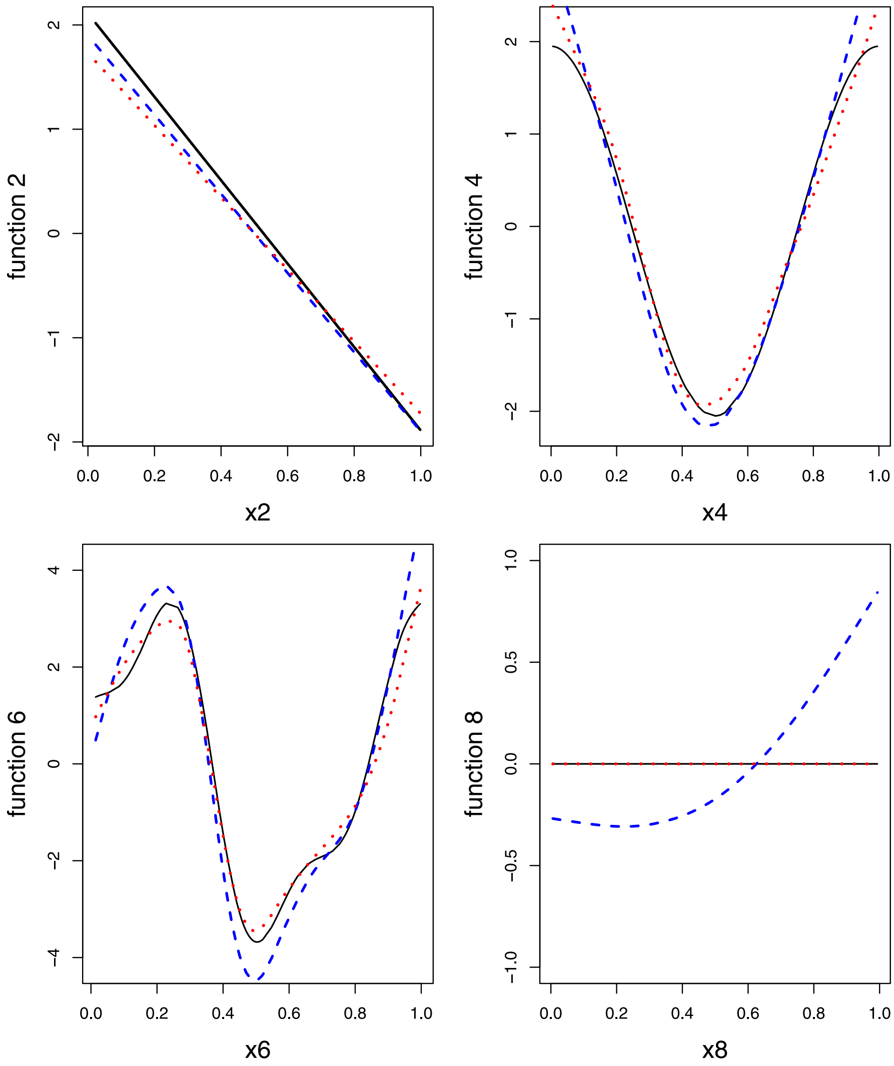

Figure 2. True and estimated function components for Example 1: True function (solid line), SS-ANOVA estimator (dashed line), and 2LAND estimator (dotted line). The online version of this figure is in color.

6.3 Real Example

We apply the LAND to analyze the Boston housing data, which are available at the UCI Data Repository and can be loaded in R. The data are for 506 census tracts of Boston from the 1970 census, containing twelve continuous covariates and one binary covariate. These covariates are per capita crime rate

Table 4. Average selection results (standard errors in parentheses) for 100 runs in Example 2

	R^2	corrlin	corrnon	corrlnn	corr0	LNto0	LtoN	NtoL
ρ	oracle:	3	2	2	13	0	0	0
0	0.95	3.00 (0.00)	1.85 (0.04)	1.99 (0.01)	12.79 (0.09)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.75	2.95 (0.02)	1.66 (0.06)	1.91 (0.03)	11.98 (0.11)	0.05 (0.02)	0.00 (0.00)	0.01 (0.01)
	0.55	2.77 (0.05)	1.31 (0.07)	1.73 (0.05)	10.60 (0.23)	0.24 (0.05)	0.01 (0.01)	0.26 (0.05)
	0.35	2.31 (0.09)	0.90 (0.07)	1.39 (0.07)	9.29 (0.27)	0.88 (0.12)	0.06 (0.02)	0.55 (0.06)
0.5	0.95	2.97 (0.02)	1.81 (0.04)	1.98 (0.01)	12.55 (0.12)	0.00 (0.00)	0.03 (0.02)	0.00 (0.00)
	0.75	2.82 (0.04)	1.29 (0.07)	1.58 (0.05)	11.33 (0.18)	0.17 (0.04)	0.01 (0.01)	0.31 (0.05)
	0.55	2.50 (0.07)	0.98 (0.07)	1.07 (0.08)	9.66 (0.24)	0.57 (0.08)	0.02 (0.01)	0.90 (0.08)
	0.35	1.72 (0.11)	0.51 (0.06)	0.65 (0.07)	9.47 (0.31)	1.80 (0.17)	0.03 (0.02)	1.35 (0.09)
0.8	0.95	2.99 (0.01)	1.78 (0.05)	1.95 (0.02)	12.48 (0.18)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)
	0.75	2.62 (0.06)	1.30 (0.07)	1.50 (0.06)	11.07 (0.22)	0.38 (0.06)	0.00 (0.00)	0.34 (0.05)
	0.55	2.29 (0.08)	0.94 (0.06)	1.12 (0.07)	9.14 (0.27)	0.76 (0.09)	0.04 (0.02)	0.94 (0.07)
	0.35	1.46 (0.11)	0.53 (0.06)	0.63 (0.07)	9.70 (0.35)	2.34 (0.19)	0.06 (0.02)	1.05 (0.09)

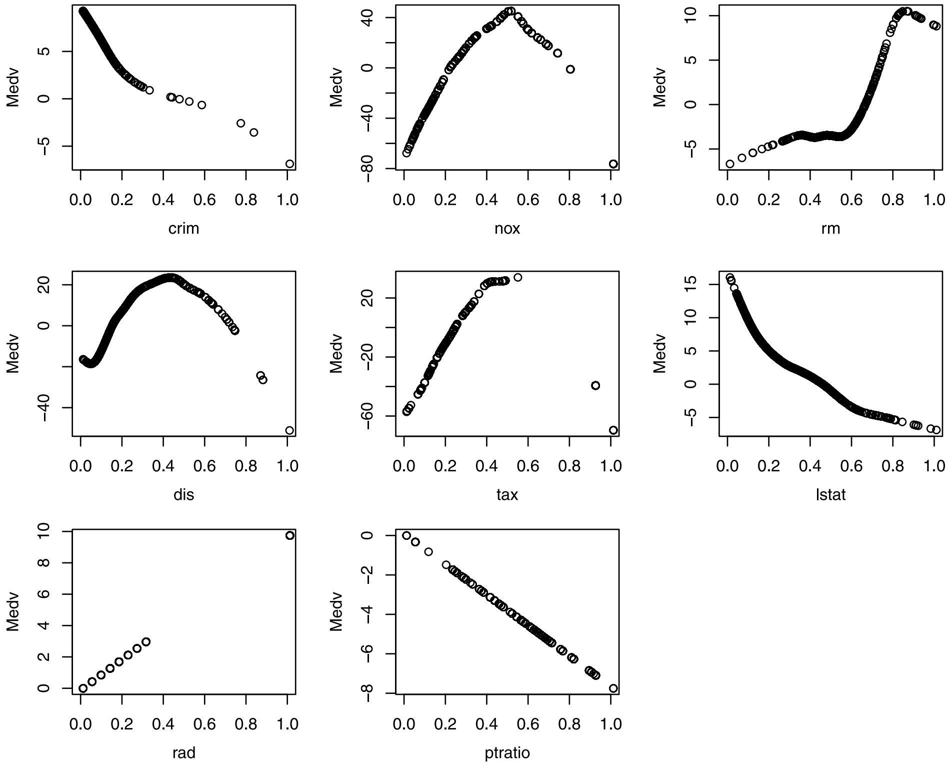

Figure 3. The selected components and their fits by 2LAND for Boston Housing data.

by town (crime), proportion of residential land zoned for lots over 25,000 sq.ft (zn), proportion of non-retail business acres per town (indus), Charles River dummy variable (chas), nitric oxides concentration (nox), average number of rooms per dwelling (rm), proportion of owner-occupied units built prior to 1940 (age), weighted distances to five Boston employment centers (dis), index of accessibility to radial highways (rad), full-value property-tax rate per USD 10,000 (tax), pupil-teacher ratio by town (ptratio), $1000(B-0.63)^2$ where B is the proportion of blacks by town (b), and percentage of lower status of the population (lstat). The response variable is the median value of owner-occupied homes in USD 1000's (medv).

We scale all the covariates to [0, 1] and fit the 2LAND procedure. The parameters are tuned using 5-fold cross-validation. From the thirteen covariates, the 2LAND identifies two linear effects: rad and ptratio, and six nonlinear effects: crime, nox, rm.dis, tax, and lstat. The remaining five covariates: zn, indus, chas, age, b, are removed from the final model as unimportant covariates. For comparison, we also fit the additive model in R with the function gam, which identify four covariates as insignificant at level $\alpha = 0.05$: zn, chas, age, and b. Figure 3 plots the fitted function components provided by the 2LAND estimator. The first six panels are for nonlinear terms and the last two are for linear terms.

7. DISCUSSION

Partially linear models are widely used in practice, but none of the existing methods can consistently distinguish linear and nonlinear terms for the models. This work aims to fill this gap with a new regularization framework in the context of smoothing spline ANOVA models. Rates of convergence of the proposed estimator were established. With a proper choice of tuning parameters, we have shown that the proposed estimator is consistent in both structure selection and model estimation. The methods were shown to be effective through numerical examples. An iterative algorithm was proposed for solving the optimization problem. Compared with existing approaches, the LAND procedure is developed in a unified mathematical framework and well-justified in theory.

In this article, we consider classical settings where d is fixed. It would be interesting to extend the LAND to high-dimensional data, with a diverging d or $d \gg n$. For ultrahigh-dimensional data, we suggest to combine the LAND procedures with dimension reduction techniques such as Sure Independence Screening (Fan and Jinchi 2008; Fan, Feng, and Song 2011). Alternatively, we can first implement the variable selection procedures for high-dimensional additive models, using SpAM (Ravikumar et al. 2009) or the adaptive group LASSO (Huang, Horowitz, and Wei 2010). These procedures are consistent in variable selection for high-dimensional data, but they cannot distinguish linear and nonlinear terms. After variable screening is performed in the first step, the LAND can be applied to discover the more subtle structure of the reduced model.

Additive models are a rich class of models and provide greater flexibility than linear models. The possible model misspecification associated with additive models is to overlook the potential interactions between variables. The LAND can be naturally extended to two-way functional SS-ANOVA models and conduct selection for both main effects and interactions. Interestingly, this extension makes it possible to detect subtle structures for interaction terms, such as linear-linear, linear-nonlinear, and nonlinear-nonlinear interactions between two variables.

APPENDIXES

Some Notations. Recall that the ANOVA decomposition of any $g \in \mathcal{H}$ is $g(\mathbf{x}) = b + \sum_{j=1}^d \beta_j k_1(x_j) + \sum_{j=1}^d g_{1j}(x_j)$. Then we define $h_j(x_j) = \beta_j k_1(x_j) + g_{1j}(x_j)$, $H_0(\mathbf{x}) = \sum_{j=1}^d \beta_j k_1(x_j)$, $H_1(\mathbf{x}) = \sum_{j=1}^d g_{1j}(x_j)$, and $H(\mathbf{x}) = \sum_{j=1}^d h_j(x_j)$. The same notational rule also applies to $\hat{g}$, $\widetilde{g}$, and g_0.

Appendix 1. Important Lemmas: Convergence Rates of $\widetilde{g}$

We derive the convergence rate of $\widetilde{g}$ in Lemma A.1.

Lemma A.1. Suppose Conditions (C1)–(C3) hold. If we set $\lambda \sim n^{-4/5}$, then the initial solution (3.2) is proved to have the following convergence rates, for any $1 \le j \le d$:

$$\|\widetilde{g}_{1j} - g_{1j}^0\|_2 = O_P\big(n^{-1/5}\big),$$

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P\big(n^{-1/5}\big),$$

where $\|\cdot\|$ is the Euclidean norm.

Proof. We first prove that $\|\widetilde{H} - H_0\|_2 = O_P(n^{-2/5})$. Denote $J_i(H) = \sum_{j=1}^d \|\mathcal{P}_{1j} H\|_{\mathcal{H}_1}^2$. Since $\widetilde{H}$ minimizes $H \mapsto \|H_0 + \epsilon - H\|_n^2 + \lambda J_i(H)$, we have the following inequality:

$$\|\widetilde{H} - H_0\|_n^2 + \lambda J_i(\widetilde{H}) \le 2\langle \widetilde{H} - H_0, \epsilon\rangle_n + \lambda J_i(H_0), \tag{A.1}$$

$$\begin{aligned}\|\widetilde{H} - H_0\|_n^2 &\le 2\|\epsilon\|_n \|\widetilde{H} - H_0\|_n + \lambda J_i(H_0)\\ &\le O_P(1)\|\widetilde{H} - H_0\|_n + o_P(1)\end{aligned}$$

by the Cauchy–Schwarz inequality and the subexponential tail assumption of ϵ. The above inequality implies that $\|\widetilde{H} - H_0\|_n = O_P(1)$ so that $\|\widetilde{H}\|_n = O_P(1)$. By the Sobolev embedding theorem, we can decompose $H(\mathbf{x})$ as $H_0(\mathbf{x}) + H_1(\mathbf{x})$, where $\|H_1\|_\infty \le \sum_{j=1}^d \|g_{1j}\|_\infty \le J_n(H)$. Similarly, we can write $\widetilde{H} = \widetilde{H}_0 + \widetilde{H}_1$, where $\widetilde{H}_0(\mathbf{x}) = \sum_{j=1}^d \widetilde{\beta}_j k_1(x_j)$ and $\|\widetilde{H}_1\|_\infty \le J_n(\widetilde{H})$. We shall now show that $\|\widetilde{H}\|_\infty/(1 + J_n(\widetilde{H})) = O_P(1)$ as follows. First, we have

$$\frac{\|\widetilde{H}_0\|_n}{1 + J_n(\widetilde{H})} \le \frac{\|\widetilde{H}\|_n}{1 + J_n(\widetilde{H})} + \frac{\|\widetilde{H}_1\|_n}{1 + J_n(\widetilde{H})} = O_P(1). \tag{A.2}$$

Combining with Condition (C2), (A.2) implies that $\|\widetilde{\boldsymbol{\beta}}\|/(1 + J_n(\widetilde{H})) = O_P(1)$. Since $\mathbf{x} \in [0,1]^d$, $\|\widetilde{H}_0\|_\infty/(1 + J_n(\widetilde{H})) = O_P(1)$. So we have proved that $\|\widetilde{H}\|_\infty/(1 + J_n(\widetilde{H})) = O_P(1)$ by the triangular inequality and the Sobolev embedding theorem. Thus, according to Birman and Solomjak (1967), we know the entropy number for the below constructed class of functions:

$$H\left(\delta, \left\{\frac{H - H_0}{1 + J_n(H)} : \frac{\|H\|_\infty}{1 + J_n(H)} \le C\right\}, \|\cdot\|_\infty\right) \le M_1 \delta^{-1/2},$$

where M_1 is some positive constant. Based on theorem 2.2 in the article by Mammen and van de Geer (1997) about the continuity modulus of the empirical processes $\{\sum_{i=1}^n \epsilon_i (H - H_0)(\mathbf{x}_i)\}$ indexed by H and (A.1), we can establish the following set of inequalities:

$$\begin{aligned}\lambda J_i(\widetilde{H}) \le \big[&\|\widetilde{H} - H_0\|_n^{3/4}(1 + J_n(\widetilde{H}))^{1/4} \vee (1 + J_n(\widetilde{H}))n^{-3/10}\big]\\ &\times O_P\big(n^{-1/2}\big) + \lambda J_i(H_0),\end{aligned}$$

and

$$\begin{aligned}\|\widetilde{H} - H_0\|_n^2 \le \big[&\|\widetilde{H} - H_0\|_n^{3/4}(1 + J_n(\widetilde{H}))^{1/4} \vee (1 + J_n(\widetilde{H}))n^{-3/10}\big]\\ &\times O_P\big(n^{-1/2}\big) + \lambda J_i(H_0).\end{aligned}$$

Considering $J_n^2/d \le J_i \le J_n^2$, we can solve the above two inequalities to obtain $\|\widetilde{H} - H_0\|_n = O_P(n^{-2/5})$ and $J_i(\widetilde{H}) = O_P(1)$ given $\lambda \sim n^{4/5}$. Theorem 2.3 in the article by Mammen and van de Geer (1997) further implies that

$$\|\widetilde{H} - H_0\|_2 = O_P\big(n^{-2/5}\big). \tag{A.3}$$

Recall that $\widetilde{H}(\mathbf{x}) = \sum_{j=1}^d \widetilde{h}_j(x_j) = \sum_{j=1}^d \widetilde{\beta}_j k_1(x_j) + \widetilde{g}_{1j}(x_j)$ and $(\boldsymbol{\beta}_0, g_{1j}^0(\cdot))$ is the true value of $(\boldsymbol{\beta}, g_{1j}(\cdot))$. We next prove $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/5})$ and $\|\widetilde{g}_{1j} - g_{1j}^0\|_2 = O_P(n^{-1/5})$ for any $j = 1, \ldots, d$ based on (A.3). We first take a differentiation approach to get the convergence rate for $\widetilde{\beta}_j$. Since the density for $\mathbf{X}$ is bounded away from zero and $\int_0^1 h_j(u)\,du = 0$, (A.3) implies

$$\max_{1\le j\le d} \int_0^1 (\widetilde{h}_j(u) - h_j^0(u))^2\,du = O_P\big(n^{-4/5}\big). \tag{A.4}$$

Agmon (1965) showed that there exists a constant $C > 0$ such that for all $0 \le k \le 2$, $0 < \rho < 1$ and for all functions $\gamma : \mathbb{R} \mapsto \mathbb{R}$:

$$\int_0^1 \big(\gamma^{(k)}(x)\big)^2\,dx \le C\rho^{-2k} \int_0^1 \gamma^2(x)\,dx + C\rho^{4-2k} \int_0^1 \big(\gamma^{(2)}(x)\big)^2\,dx. \tag{A.5}$$

Having proved that $J_i(\widetilde{H}) = O_P(1)$, we can apply the above interpolation inequality (A.5) to (A.4) with $k = 1$, $\rho = \lambda^{1/4}$, and $\gamma(x) = \widetilde{h}_j(x) - h_j^0(x)$. Thus we conclude that

$$\max_{1\le j\le d} \int \big((\partial/\partial u)\widetilde{h}_j(u) - (\partial/\partial u)h_j^0(u)\big)^2\,du = O_P\big(n^{-2/5}\big). \tag{A.6}$$

Note that we can write $(\partial/\partial u)\widetilde{h}_j(u) = \widetilde{\beta}_j + (\partial/\partial u)\widetilde{g}_{1j}(u)$ and $(\partial/\partial u)h_j^0(u) = \beta_j^0 + (\partial/\partial u)g_{1j}^0(u)$, respectively. Thus (A.6) becomes

$$\begin{aligned}O_P\big(n^{-2/5}\big) &= (\widetilde{\beta}_j - \beta_j^0)^2\\ &\quad + 2(\widetilde{\beta}_j - \beta_j^0) \int_0^1 \left(\frac{\partial}{\partial u}\widetilde{g}_{1j}(u) - \frac{\partial}{\partial u}g_{1j}^0(u)\right) du\\ &\quad + \int_0^1 \left(\frac{\partial}{\partial u}\widetilde{g}_{1j}(u) - \frac{\partial}{\partial u}g_{1j}^0(u)\right)^2 du\\ &= (\widetilde{\beta}_j - \beta_j^0)^2 + \int_0^1 \left(\frac{\partial}{\partial u}\widetilde{g}_{1j}(u) - \frac{\partial}{\partial u}g_{1j}^0(u)\right)^2 du \end{aligned}\tag{A.7}$$

for any $1 \le j \le d$, where the second equality follows from the definition of $\mathcal{H}_{1j}$ in RKHS. Obviously, (A.7) implies that $\widetilde{\beta}_j - \beta_j^0 = O_P(n^{-1/5})$.

We next prove the convergence rate for $\widetilde{g}_{1j}$ by decomposing the function $g_j(x_j)$ in another form; see example 9.3.2 in the book by van de Geer (2000). We can write $g_j(x_j) = (b/d) + \beta_j(x_j - 1/2) + g_{1j}(x_j) = g_{0j}(x_j) + g_{1j}(x_j)$, where $g_{1j}(x_j) = \int_0^1 g_j^{(2)}(u)\psi_u(x_j)\,du$ and $\psi_u(x_j) = (x_j - u)1\{u \le x_j\}$. Let $\bar{\psi}_u(x_j) = \alpha_{0,u}^j + \alpha_{1,u}^j x_j$ be the projection in terms of empirical L_2-norm of $\psi_u(x_j)$ on the linear space spanned by $\{1, x_j\}$. Let $\widetilde{\psi}_u(x_j) = \psi_u(x_j) - \bar{\psi}_u(x_j)$. Then, we can further decompose

$$\begin{aligned}g_j(x_j) &= [(b/d) + \beta_j(x_j - 1/2)] + \left[\int_0^1 g_j^{(2)}(u)\alpha_{0,u}^j\,du\right.\\ &\quad \left. + x_j \int_0^1 g_j^{(2)}(u)\alpha_{1,u}^j\,du\right] + \int_0^1 g_j^{(2)}(u)\widetilde{\psi}_u(x_j)\,du\\ &= g_{0j}(x_j) + g_{1j,1}(x_j) + g_{1j,\mathrm{nl}}(x_j),\end{aligned}$$

where $g_{1j,1}$ and $g_{1j,\mathrm{nl}}$ are the (orthogonal) linear and nonlinear components of g_{1j}, respectively. We define $(\widetilde{g}_{0j}, \widetilde{g}_{1j,1}, \widetilde{g}_{1j,\mathrm{nl}})$ and $(g_{0j}^0, g_{1j,1}^0, g_{1j,\mathrm{nl}}^0)$ as the initial estimators and true values of $(g_{0j}, g_{1j,1}, g_{1j,\mathrm{nl}})$, respectively. By corollary 10.4 in the work by van de Geer (2000), we have

$$\|\widetilde{g}_{0j} + \widetilde{g}_{1j,1} - g_{0j}^0 - g_{1j,1}^0\|_n = O_P\big(n^{-1/2}\big) \qquad \text{and}$$

$$\|\widetilde{g}_{1j,\mathrm{nl}} - g_{1j,\mathrm{nl}}^0\|_n = O_P\big(n^{-2/5}\big).$$

By the triangle inequality and the result obtained previously, that is, $\widetilde{\beta}_j - \beta_j^0 = O_P(n^{-1/5})$, we have $\|\widetilde{g}_{1j,1} - g^0_{1j,1}\|_n = O_P(n^{-1/5})$. Then combining the fact that $g_{1j} = g_{1j,1} + g_{1j,\mathrm{nl}}$, we have shown $\|\widetilde{g}_{1j} - g^0_{1j}\|_n = O_P(n^{-1/5})$ by applying the triangle inequality again. We further obtain the L_2-rate for $\widetilde{g}_{1j}$, that is, $\|\widetilde{g}_{1j} - g^0_{1j}\|_2 = O_P(n^{-1/5})$, by applying theorem 2.3 from the book by Mammen and van de Geer (1997). This completes the whole proof.

Appendix 2. Proof of Theorem 1

Denote $J_l^w(H) = \sum_{j=1}^d \omega_{0j}|\beta_j|$ and $J_n^w(H) = \sum_{j=1}^d \omega_{1j}\|\mathcal{P}_{1j}H\|_{\mathcal{H}_1}$. We first rewrite (2.6) as

$$\frac{1}{n}\sum_{i=1}^{n}(b_0 + H_0(\mathbf{x}_i) + \epsilon_i - b - H(\mathbf{x}_i))^2 + \lambda_1 J_l^w(H) + \lambda_2 J_n^w(H).$$

Since we assume that $\sum_{i=1}^n h_j(x_{ij}) = 0$, the terms involving b in the above equation are $(b_0 - b)^2 + 2(b_0 - b)\sum_{i=1}^n \epsilon_i/n$. Therefore, we obtain that $\hat{b} = b_0 + \sum_{i=1}^n \epsilon_i/n$ which implies that

$$\hat{b} - b_0 = O_P\big(n^{-1/2}\big). \tag{A.8}$$

Recall that $J(H) \equiv \sum_{j=1}^d |\beta_j| + \sum_{j=1}^d \|\mathcal{P}_{1j}H\|_{\mathcal{H}_1}$. It remains to prove that $\|\hat{H} - H_0\|_n = O_P(n^{-2/5})$ when $J(H_0) > 0$ and $\|\hat{H} - H_0\|_n = O_P(n^{-1/2})$ when $J(H_0) = 0$ as follows.

The definition of $\hat{H}$ implies the following inequality:

$$\begin{aligned}
&\|\hat{H} - H_0\|_n^2 + \lambda_1 J_l^w(\hat{H}) + \lambda_2 J_n^w(\hat{H})\\
&\quad\le 2\langle\epsilon, \hat{H} - H_0\rangle_n + \lambda_1 J_l^w(H_0) + \lambda_2 J_n^w(H_0), \qquad (\text{A.9})\\
&\|\hat{H} - H_0\|_n^2\\
&\quad\le 2\|\epsilon\|_n\|\hat{H} - H_0\|_n + \lambda_1 J_l^w(H_0) + \lambda_2 J_n^w(H_0)\\
&\quad\le O_P(1)\|\hat{H} - H_0\|_n + o_P(1),
\end{aligned}$$

where the second inequality follows from the Cauchy–Schwarz inequality, and the third one follows from the subexponential tail of ϵ. Hence, we can prove $\|\hat{H} - H_0\|_n = O_P(1)$ so that $\|\hat{H}\|_n = O_P(1)$. Now we consider two different cases that $J(H_0) > 0$ and $J(H_0) = 0$.

Case I: $J(H_0) > 0$.

We first prove

$$\frac{\|\hat{H}\|_\infty}{J(H_0) + J(\hat{H})} = O_P(1) \tag{A.10}$$

by the Sobolev embedding theorem. The Sobolev embedding theorem implies that $\|g_{1j}(x_j)\|_\infty \le \|\mathcal{P}_{1j}H\|_{\mathcal{H}_1}$, and thus we can establish that

$$\begin{aligned}
\frac{\|\hat{H}_0\|_n}{J(H_0) + J(\hat{H})} &\le \frac{\|\hat{H}\|_n}{J(H_0) + J(\hat{H})} + \frac{\|\hat{H}_1\|_n}{J(H_0) + J(\hat{H})}\\
&\le O_P(1) + \frac{\sum_{j=1}^d |\mathcal{P}_{1j}\hat{H}\|_{\mathcal{H}_1}}{J(H_0) + J(\hat{H})} \le O_P(1).
\end{aligned}$$

Combining the above result with Condition (C2), we have $\|\hat{\boldsymbol{\beta}}\|/(J(H_0) + J(\hat{H})) = O_P(1)$ which further implies that $\|\hat{H}_0\|_\infty/(J(H_0) + J(\hat{H})) = O_P(1)$ by the assumption that $\mathbf{x} \in [0, 1]^d$. Again, by the Sobolev embedding theorem, we have proved (A.10). By theorem 2.4 in the book by van de Geer (2000), we know the bracket-entropy number for the below class of constructed functions is

$$H_B\left(\delta, \left\{\frac{H - H_0}{J(H_0) + J(H)} : H \equiv \sum_{j=1}^d h_j, \text{ where } h_j \in \mathcal{G}\right\}, \|\cdot\|_\infty\right) \le M_2\delta^{-1/2},$$

where $\mathcal{G} = \{h_j(x) = (x - 1/2)\beta_j + g_{1j}(x) : \|g_{1j}\|_{\mathcal{H}_1} < \infty\}$ and M_2 is some positive constant. Based on lemma 8.4 from the work of van de Geer (2000) about the continuity modulus of the empirical processes $\langle H - H_0, \epsilon\rangle_n$ indexed by H in (A.9), we can establish the following set of inequalities:

$$\begin{aligned}
&\|\hat{H} - H_0\|_n^2 + \lambda_1 J_l^w(\hat{H}) + \lambda_2 J_n^w(\hat{H})\\
&\quad\le \big[\|\hat{H} - H_0\|_n^{3/4}(J(H_0) + J(\hat{H}))^{1/4}\big]O_P\big(n^{-1/2}\big)\\
&\qquad + \lambda_1 J_l^w(H_0) + \lambda_2 J_n^w(H_0).
\end{aligned} \tag{A.11}$$

Note that the sub-Gaussian tail condition in lemma 8.4 of the book by van de Geer (2000) can be relaxed to the assumed subexponential tail condition; see discussions on page 168 of that book. In the following, we will analyze (A.11) for the cases $J(\hat{H}) \le J(H_0)$ and $J(\hat{H}) > J(H_0)$. If $J(\hat{H}) \le J(H_0)$, then $J(\hat{H}) = O_P(1)$. Thus, (A.11) implies that

$$\begin{aligned}
\|\hat{H} - H_0\|_n^2 &\le \|\hat{H} - H_0\|_n^{3/4} J(H_0)^{1/4} O_P\big(n^{-1/2}\big)\\
&\quad + \lambda_1 J_l^w(H_0) + \lambda_2 J_n^w(H_0).
\end{aligned} \tag{A.12}$$

Since $\lambda_1, \lambda_2 \sim n^{-4/5}$, we have $\|\hat{H} - H_0\|_n = O_P(n^{-2/5})$ based on (A.12). We next consider the case that $J(\hat{H}) > J(H_0) > 0$. In this case, (A.11) becomes

$$\begin{aligned}
&\|\hat{H} - H_0\|_n^2 + \lambda_1 J_l^w(\hat{H}) + \lambda_2 J_n^w(\hat{H})\\
&\quad\le \|\hat{H} - H_0\|_n^{3/4} J(\hat{H})^{1/4} O_P\big(n^{-1/2}\big) + \lambda_1 J_l^w(H_0) + \lambda_2 J_n^w(H_0),
\end{aligned}$$

which implies either

$$\begin{aligned}
&\|\hat{H} - H_0\|_n^2 + \lambda_1 J_l^w(\hat{H}) + \lambda_2 J_n^w(\hat{H})\\
&\quad\le \|\hat{H} - H_0\|_n^{3/4} J(\hat{H})^{1/4} O_P\big(n^{-1/2}\big)
\end{aligned} \tag{A.13}$$

or

$$\|\hat{H} - H_0\|_n^2 + \lambda_1 J_l^w(\hat{H}) + \lambda_2 J_n^w(\hat{H}) \le \lambda_1 J_l^w(H_0) + \lambda_2 J_n^w(H_0). \tag{A.14}$$

Note that

$$\begin{aligned}
\lambda_1 J_l^w(\hat{H}) + \lambda_2 J_n^w(\hat{H}) &\ge \lambda_1 w_0^* \sum_{j=1}^d \|P_{0j}\hat{H}\|_{\mathcal{H}_0} + \lambda_2 w_1^* \sum_{j=1}^d \|P_{1j}\hat{H}\|_{\mathcal{H}_1}\\
&\ge r_n J(\hat{H}),
\end{aligned} \tag{A.15}$$

where $r_n = \lambda_1 w_0^* \wedge \lambda_2 w_1^*$, $w_0^* = \min\{w_{01}, \ldots, w_{0d}\}$, and $w_1^* = \min\{w_{11}, \ldots, w_{1d}\}$. Thus solving (A.13) gives

$$\|\hat{H} - H_0\|_n \le r_n^{-1/3} O_P\big(n^{-2/3}\big), \tag{A.16}$$

$$J(\hat{H}) \le r_n^{-5/3} O_P\big(n^{-4/3}\big). \tag{A.17}$$

Because of the conditions on λ_1, λ_2, w_{0j}, and w_{1j}, we know $r_n^{-1} = O_P(n^{4/5})$. Hence (A.16) and (A.17) imply that $J(\hat{H}) = O_P(1)$ and $\|\hat{H} - H_0\|_n = O_P(n^{-2/5})$. By similar logic, we can show that (A.14) also implies $J(\hat{H}) = O_P(1)$ and $\|\hat{H} - H_0\|_n = O_P(n^{-2/5})$.

So far, we have proved $\|\hat{H} - H_0\|_n = O_P(n^{-2/5})$ and $J(\hat{H}) = O_P(1)$ given that $J(H_0) > 0$. Next we consider the trivial case that $J(H_0) = 0$.

Case II: $J(H_0) = 0$.

Based on (2.7) and Lemma A.1, we know that $w_{0j}^{-1} = O_P(n^{-\alpha/5})$ and $w_{1j}^{-1} = O_P(n^{-\gamma/5})$ given that $J(H_0) = 0$. Thus we have w_{0j}^{-1}, $w_{1j}^{-1} = O_P(n^{-3/10})$ based on the assumption that $\alpha \ge 3/2$, $\gamma \ge 3/2$. Then we know that $r_n^{-1} = O_P(n^{1/2})$. From (A.16) and (A.17), we can get $\|\hat{H} - H_0\|_n = O_P(n^{-1/2})$ and $J(\hat{H}) = O_P(n^{-1/2}) = o_P(1)$.

Appendix 3. Proof of Lemmas 1 and 2

The proof of Lemma 1 is similar to those of lemmas 1 and 4 in COSSO, and the proof of lemma 2 is similar to that of lemma 2 in the COSSO article.

SUPPLEMENTARY MATERIALS

Appendix: The proof of Theorem 2 is given in Appendix 4, which is provided as online supplementary materials for this article. (Supplement.pdf)

[Received May 2010. Revised January 2011.]

REFERENCES

Agmon, S. (1965), *Lectures on Elliptic Boundary Value Problems*, Princeton, NJ: van Nostrand. [1110]

Birman, M. S., and Solomjak, M. Z. (1967), “Piecewise-Polynomial Approximation of Functions of the Classes w_p,” *Mathematics of the USSR Sbornik*, 73, 295–317. [1110]

Chen, H. (1988), “Convergence Rates for Parametric Components in a Partly Linear Model,” *The Annals of Statistics*, 16, 136–146. [1099]

Dinse, G. E., and Lagakos, S. W. (1983), “Regression Analysis of Tumour Prevalence Data,” *Journal of the Royal Statistical Society, Ser. C*, 32, 236–248. [1099]

Engle, R., Granger, C., Rice, J., and Weiss, A. (1986), “Semiparametric Estimates of the Raltion Between Weather and Electricity Sales,” *Journal of the American Statistical Association*, 81, 310–386. [1099]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall. [1099]

Fan, J., and Jinchi, L. (2008), “Sure Independence Screening for Ultrahigh Dimensional Feature Space,” *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911. [1109]

Fan, J., and Li, R. (2004), “New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis,” *Journal of the American Statistical Association*, 99, 710–723. [1099]

Fan, J., Feng, Y., and Song, R. (2011), “Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models,” *Journal of the American Statistical Association*, 106, 544–557. [1109]

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman & Hall. [1099]

Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer-Verlag. [1100,1103]

Hardle, W., Liang, H., and Gao, J. (2000), *Partially Linear Models*, Heidelberg: Physica-Verlag. [1099]

Heckman, N. E. (1986), “Spline Smoothing in a Partly Linear Model,” *Journal of the Royal Statistical Society, Ser. B*, 48, 244–248. [1099]

Hong, S. Y. (1991), “Estimation Theory of a Class of Semiparametric Regression Models,” *Sciences in China Ser. A*, 12, 1258–1272. [1099]

Huang, J., Horowitz, J., and Wei, F. (2010), “Variable Selection in Nonparametric Additive Models,” *The Annals of Statistics*, 38, 2282–2313. [1109]

Kimeldorf, G., and Wahba, G. (1971), “Some Results on Tchebycheffian Spline Functions,” *Journal of Mathematical Analysis and Applications*, 33, 82–85. [1103]

Li, R., and Liang, H. (2008), “Variable Selection in Semiparametric Regression Modeling,” *The Annals of Statistics*, 36, 261–286. [1099]

Liang, H. (2006), “Estimation in Partially Linear Models and Numerical Comparisons,” *Computational Statistics and Data Analysis*, 50, 675–687. [1099]

Liang, H., Hardle, W., and Carroll, R. J. (1999), “Estimation in a Semiparametric Partially Linear Errors-in-Variables Model,” *The Annals of Statistics*, 27, 1519–1535. [1099]

Lin, Y., and Zhang, H. H. (2006), “Component Selection and Smoothing in Multivariate Nonparametric Regression,” *The Annals of Statistics*, 34, 2272–2297. [1101,1102]

Mammen, E., and van de Geer, S. (1997), “Penalized Quasi-Likelihood Estimation in Partial Linear Models,” *The Annals of Statistics*, 25, 1014–1035. [1103,1110,1111]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), “Sparse Additive Models,” *Journal of the Royal Statistical Society, Ser. B*, 71, 1009–1030. [1109]

Rice, J. (1986), “Convergence Rates for Partially Spline Models,” *Statistics and Probability Letters*, 4, 203–208. [1099]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press. [1099]

Schmalensee, R., and Stoker, T. M. (1999), “Household Gasoline Demand in the United States,” *Econometrica*, 67, 645–662. [1099]

Speckman, P. (1988), “Kernel Smoothing in Partial Linear Models,” *Journal of the Royal Statistical Society, Ser. B*, 50, 413–436. [1099]

Storlie, C., Bondell, H., Reich, B., and Zhang, H. H. (2011), “Surface Estimation, Variable Selection, and the Nonparametric Oracle Property,” *Statistica Sinica*, 21, 679–705. [1101]

Tibshirani, R. J. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [1101]

van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge: Cambridge University Press. [1110,1111]

Wahba, G. (1984), “Cross Validated Spline Methods for the Estimation of Multivariate Functions From Data on Functions, Statistics: An Appraisal,” in *Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory*, ed. H. A. David, Ames, IA: Iowa State University Press, pp. 205–235. [1099]

——— (1990), *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 59, Philadelphia: SIAM. [1100,1103]

Wang, H., Li, G., and Jiang, G. (2007), “Robust Regression Shrinkage and Consistent Variable Selection via the Lad-Lasso,” *Journal of Business & Economic Statistics*, 20, 347–355. [1101]

Wang, L., Li, H., and Huang, J. (2008), “Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements,” *Journal of the American Statistical Association*, 103, 1556–1569. [1099]

Zhang, H. H., and Lin, Y. (2006), “Component Selection and Smoothing for Nonparametric Regression in Exponential Families,” *Statistica Sinica*, 16, 1021–1042. [1101]

Zhang, H. H., and Lu, W. (2007), “Adaptive-Lasso for Cox’s Proportional Hazards Model,” *Biometrika*, 94, 691–703. [1101]

Zou, H. (2006), “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, 101, 1418–1429. [1101]

Supplementary Materials for Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models

Hao Helen Zhang*, Guang Cheng and Yufeng Liu

The supplementary material contains the proof of Theorem 2.

Appendix 4. Proof of Theorem 2

It suffices to show that, with probability tending to one,

$$\mathcal{P}_{1j}\widehat{g} = 0 \Longleftrightarrow \mathcal{P}_{1j}g_0 = 0, \tag{8.18}$$

$$\widehat{\beta}_j = 0 \Longleftrightarrow \beta_j^0 = 0 \tag{8.19}$$

for $j = 1, \ldots, d$. Without loss of generality, we focus on the case $d = 2$, i.e. $g(x_1, x_2) = b + \beta_1 k_1(x_1) + \beta_2 k_1(x_2) + g_{11}(x_1) + g_{12}(x_2)$, where $g_{1j}(x_j) \in \mathcal{S}_{per,j}$, in the proof. Note that in this case the sample size n is m^2 since we assume $n_1 = n_2 = m$. We have three major steps in the proof.

Step I: Formulation

Let $\mathbf{\Sigma} = \{R_1(x_{i,1}, x_{k,1})\}_{i,k=1}^m$ be the $m \times m$ marginal Gram matrix corresponding to the reproducing kernel for $\mathcal{S}_{per}$. Let $\mathbf{1}_m$ be a vector of m ones. Assuming the observations are permuted appropriately, we can write the $n \times n$ Gram matrix $\mathbf{R}_{11} = \mathbf{\Sigma} \odot (\mathbf{1}_m \mathbf{1}_m')$ and $\mathbf{R}_{12} = (\mathbf{1}_m \mathbf{1}_m') \odot \mathbf{\Sigma}$, where $\odot$ stands for the Kronecker product between two matrices. Let $\{\boldsymbol{\xi}_1 = \mathbf{1}_m, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m\}$ be an orthonormal (with respect to the inner product $< \cdot >_m$ in $\mathcal{R}^m$) eigensystem of $\mathbf{\Sigma}$ with corresponding eigenvalues $m\eta_1, \ldots, m\eta_m$ where $\eta_1 = (720m^4)^{-1}$. Thus, we have

$$< \boldsymbol{\xi}_1, \boldsymbol{\xi}_j >_m = 0 \Longrightarrow \frac{1}{m}\sum_{i=1}^m \xi_{ij} = 0 \text{ for } j \geq 2,$$

$$< \boldsymbol{\xi}_j, \boldsymbol{\xi}_j >_m = 1 \Longrightarrow \frac{1}{m}\sum_{i=1}^m \xi_{ij}^2 = 1 \text{ for } j \geq 1.$$

From Utreras (1983), we know that $\eta_2 \geq \eta_3 \geq \ldots \geq \eta_m$ and $\eta_i \sim i^{-4}$ for $i \geq 2$.

Let Υ be the $m \times m$ matrix with $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m\}$ as its columns. We then define a $n \times n$ matrix $\mathbf{O} = \Upsilon \odot \Upsilon$. It is easy to verify that the columns of $\mathbf{O}$, i.e. $\{\widetilde{\boldsymbol{\xi}}_i : i = 1, 2, \ldots, n\}$, form an eigensystem for each of $\mathbf{R}_{11}$ and $\mathbf{R}_{12}$. We next rearrange the columns of $\mathbf{O}$ to form $\{\boldsymbol{\zeta}_{1j}, \ldots, \boldsymbol{\zeta}_{nj}\}$ so that their first m elements are those corresponding to nonzero eigenvalues for $\mathbf{R}_{1j}$ and the rest $(n-m)$ elements are given by the remaining $\widetilde{\boldsymbol{\xi}}_i$ for $j = 1, 2$. The corresponding eigenvalues are then $\eta_{ij} = n\eta_i$ for $i = 1, \ldots, m$ and zero otherwise. It is clear that $\{\widetilde{\boldsymbol{\xi}}_1, \ldots, \widetilde{\boldsymbol{\xi}}_n\}$ is also an orthonormal basis in $\mathbb{R}^n$ with respect to the inner product $< \mathbf{u}, \mathbf{v} >_n$. Thus we have $\mathbf{O}'\mathbf{O} = n\mathbf{I}$ and $\mathbf{O}\mathbf{O}' = n\mathbf{I}$.

Recall that our estimate $(\widehat{\boldsymbol{\beta}}, \widehat{g}_{11}, \widehat{g}_{12})$ is obtained by minimizing

$$\frac{1}{n}\left(\mathbf{y} - \mathbf{T}\boldsymbol{\beta} - \mathbf{R}_{\mathbf{w}_1,\boldsymbol{\theta}}\mathbf{c}\right)'\left(\mathbf{y} - \mathbf{T}\boldsymbol{\beta} - \mathbf{R}_{\mathbf{w}_1,\boldsymbol{\theta}}\mathbf{c}\right) + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| + \tau_0 \mathbf{c}'\mathbf{R}_{\mathbf{w}_1,\boldsymbol{\theta}}\mathbf{c} + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j, \tag{8.20}$$

over $(\boldsymbol{\beta}, \mathbf{c}, \boldsymbol{\theta})$, see (5.3). For simplicity, we hold $\tau_0 = 1$. By using the special construction of $\mathbf{O}$, i.e. $\mathbf{O}\mathbf{O}' = n\mathbf{I}$, we can rewrite (8.20) as

$$\left(\mathbf{z} - \mathbf{O}'\mathbf{T}\boldsymbol{\beta}/n - \mathbf{D}_\theta \mathbf{s}\right)'\left(\mathbf{z} - \mathbf{O}'\mathbf{T}\boldsymbol{\beta}/n - \mathbf{D}_\theta \mathbf{s}\right) + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| + \mathbf{s}'\mathbf{D}_\theta \mathbf{s} + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j, \tag{8.21}$$

where $\mathbf{z} = (1/n)\mathbf{O}'\mathbf{y}$, $\mathbf{s} = \mathbf{O}'\mathbf{c}$, $\mathbf{D}_\theta = \sum_{j=1}^{d} \theta_j w_{1j}^{-1} \mathbf{D}_j$ and $\mathbf{D}_j = (1/n^2)\mathbf{O}'\mathbf{R}_{1j}\mathbf{O}$ is a diagonal $n \times n$ matrix with diagonal elements η_{ij}. We further write $\mathbf{O}'\mathbf{T}\boldsymbol{\beta}/n = (b, 0, 0, \ldots)' + \mathbf{O}'\boldsymbol{t}_1\beta_1/n + \mathbf{O}'\boldsymbol{t}_2\beta_2/n$, where $\mathbf{T} = (\mathbf{1}_n, \boldsymbol{t}_1, \boldsymbol{t}_2)$ and

$$\boldsymbol{t}_1 = (1/m - 1/2, 2/m - 1/2, \ldots, 1 - 1/2)' \otimes \mathbf{1}_m, \tag{8.22}$$

$$\boldsymbol{t}_2 = \mathbf{1}_m \otimes (1/m - 1/2, 2/m - 1/2, \ldots, 1 - 1/2)'. \tag{8.23}$$

Due to the orthogonality of basis $\{\boldsymbol{\zeta}_{1j}, \ldots, \boldsymbol{\zeta}_{nj}\}$ for any j, we can further write (8.21) as

$$L(\mathbf{s}, \boldsymbol{\beta}, \boldsymbol{\theta}) = (z_{11} - b - t_{1,11}\beta_1 - t_{2,11}\beta_2 - \theta_1\eta_{11}s_{11})^2 + \left(\sum_{i=2}^{m}\sum_{j=1}^{d} + \sum_{i=1}^{1}\sum_{j=2}^{d}\right)$$

$$(z_{ij} - t_{1,ij}\beta_1 - t_{2,ij}\beta_2 - \theta_j\eta_{ij}s_{ij})^2 + \sum_{i=1}^{m}\sum_{j=1}^{d} \eta_{ij}\theta_j w_{1j}^{-1} s_{ij}^2 + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j, \tag{8.24}$$

where $t_{1,ij} = \boldsymbol{\zeta}_{ij}'\boldsymbol{t}_1/n$, $t_{2,ij} = \boldsymbol{\zeta}_{ij}'\boldsymbol{t}_2/n$, $z_{ij} = \boldsymbol{\zeta}_{ij}'\mathbf{y}/n$ and $s_{ij} = \boldsymbol{\zeta}_{ij}'\mathbf{c}$.

Note that our estimate $(\widehat{\boldsymbol{\beta}}, \widehat{g}_{11}, \widehat{g}_{12})$ are related to the minimizer of (8.24), denoted by $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{s}}, \widehat{\boldsymbol{\theta}})$, as shown in (5.2). Thus, we first analyze $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{s}}, \widehat{\boldsymbol{\theta}})$. Straightforward calculation shows that $\widehat{s}_{11} = 0$ and $z_{11} - \widehat{b} - t_{1,11}\widehat{\beta}_1 - t_{2,11}\widehat{\beta}_2 = 0$. Thus, we only need to consider minimizing

$$\begin{aligned} L_1(\mathbf{s}, \beta_1, \beta_2, \boldsymbol{\theta}) = & \left(\sum_{i=2}^{m} \sum_{j=1}^{d} + \sum_{i=1}^{1} \sum_{j=2}^{d} \right) \left[\left(z_{ij} - t_{1,ij}\beta_1 - t_{2,ij}\beta_2 - \theta_j w_{1j}^{-1} \eta_{ij} s_{ij} \right)^2 + \eta_{ij}\theta_j s_{ij}^2 \right] \\ & + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j, \end{aligned} \tag{8.25}$$

We minimize $L_1(\mathbf{s}, \beta_1, \beta_2, \boldsymbol{\theta})$ in two steps. Given fixed $(\beta_1, \beta_2, \boldsymbol{\theta})$, we first minimize L_1 over $\mathbf{s}$. Since L_1 is a convex function in $\mathbf{s}$, we can obtain the minimizer

$$\widehat{s}_{ij}(\beta_1, \beta_2, \boldsymbol{\theta}) = \frac{z_{ij} - t_{1,ij}\beta_1 - t_{2,ij}\beta_2}{1 + \theta_j \eta_{ij}/w_{1j}}. \tag{8.26}$$

Plugging (8.26) into (8.25), we obtain $L_1(\widehat{\mathbf{s}}(\beta_1, \beta_2, \boldsymbol{\theta}), \beta_1, \beta_2, \boldsymbol{\theta})$, denoted as $L_2(\beta_1, \beta_2, \boldsymbol{\theta})$:

$$\begin{aligned} L_2(\beta_1, \beta_2, \boldsymbol{\theta}) = & \left(\sum_{i=2}^{m} \sum_{j=1}^{d} + \sum_{i=1}^{1} \sum_{j=2}^{d} \right) \left[\frac{(z_{ij} - t_{1,ij}\beta_1 - t_{2,ij}\beta_2)^2}{(1 + \theta_j \eta_{ij}/w_{1j})} \right] + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| \\ & + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j \end{aligned} \tag{8.27}$$

Step 2: Prove $\mathcal{P}_{1j}\widehat{g} = 0 \Longleftrightarrow \mathcal{P}_{1j}g_0 = 0$

In this step we consider selection consistency for $\mathcal{P}_{1j}g$. We first verify that $L_2(\beta_1, \beta_2, \boldsymbol{\theta})$ in (8.27) is convex in $\boldsymbol{\theta}$ for any fixed values of β_1 and β_2 by obtaining that

$$\begin{aligned} \frac{\partial^2 L_2(\beta_1, \beta_2, \boldsymbol{\theta})}{\partial \theta_j^2} &= 2 \left(\sum_{i=2}^{m} \sum_{j=1}^{d} + \sum_{i=1}^{1} \sum_{j=2}^{d} \right) \left[\frac{\eta_{ij}^2 (z_{ij} - t_{1,ij}\beta_1 - t_{2,ij}\beta_2)^2}{(1 + \theta_j \eta_{ij}/w_{1j})^3} \right] > 0, \\ \frac{\partial^2 L_2(\beta_1, \beta_2, \boldsymbol{\theta})}{\partial \theta_j \theta_k} &= 0 \text{ for } j \neq k. \end{aligned}$$

By the above convexity, we know $\widehat{\theta}_j = 0$ if and only if

$$\left(\frac{\partial}{\partial \theta_j} |_{\theta_j = 0} \right) L_2(\widehat{\beta}_1, \widehat{\beta}_2, \boldsymbol{\theta}) \geq 0,$$

which is equivalent to

$$U_1 \equiv \sum_{i=2}^{m} \eta_{i1}(z_{i1} - t_{1,i1}\widehat{\beta}_1 - t_{2,i1}\widehat{\beta}_2)^2 \leq \tau_1 w_{11}^2, \tag{8.28}$$

$$U_j \equiv \sum_{i=1}^{m} \eta_{ij}(z_{ij} - t_{1,ij}\widehat{\beta}_1 - t_{2,ij}\widehat{\beta}_2)^2 \leq \tau_1 w_{1j}^2 \quad \text{for } j \geq 2. \tag{8.29}$$

We define $a_{ij} = \boldsymbol{\zeta}_{ij}'\mathbf{G}_1/n$, where $\mathbf{G}_1 = (G_1(\mathbf{x}_1), \ldots, G_1(\mathbf{x}_n))'$ and $G_1(\mathbf{x}_i) = \sum_{j=1}^d g_{1j}^0(x_{ij})$. Combining the fact that $z_{ij} = \boldsymbol{\zeta}_{ij}'\mathbf{y}/n$, we have the following equation:

$$z_{ij} - t_{1,ij}\beta_1^0 - t_{2,ij}\beta_2^0 = a_{ij} + e_{ij}, \tag{8.30}$$

where $e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2/n)$ for $1 \le i \le m$ and $1 \le j \le d$. Thus, (8.28) and (8.29) become

$$U_1 = \sum_{i=2}^m \eta_{i1}\left(t_{1,i1}(\beta_1^0 - \widehat{\beta}_1) + t_{2,i1}(\beta_2^0 - \widehat{\beta}_2) + e_{i1} + a_{i1}\right)^2, \tag{8.31}$$

$$U_j = \sum_{i=1}^m \eta_{ij}\left(t_{1,ij}(\beta_1^0 - \widehat{\beta}_1) + t_{2,ij}(\beta_2^0 - \widehat{\beta}_2) + e_{ij} + a_{ij}\right)^2 \tag{8.32}$$

by considering (8.30).

In the below, without loss of generality, we assume that $g_{12}^0(x_{i2}) = 0$ for $i = 1, \ldots, n$. We first show "$\mathcal{P}_{12}g_0 = 0 \Longrightarrow \mathcal{P}_{12}\widehat{g} = 0$". To show $\mathcal{P}_{12}\widehat{g} = 0$, it suffices to show

$$P(U_2 > \tau_1 w_{12}^2) \to 0. \tag{8.33}$$

based on the above analysis and (5.2). Note that $\mathcal{P}_{12}g_0 = 0$ implies $a_{i2} = 0$ for all $1 \le i \le m$. Thus, we have

$$\begin{aligned}
P(U_2 > \tau_1 w_{12}) &= P\left(\sum_{i=1}^m \eta_{i2}\left(t_{1,i2}(\beta_1^0 - \widehat{\beta}_1) + t_{2,i2}(\beta_2^0 - \widehat{\beta}_2) + e_{i2}\right)^2 > \tau_1 w_{12}^2\right)\\
&\le P\left(\sum_{i=1}^m \eta_{i2}\left[t_{1,i2}^2(\widehat{\beta}_1 - \beta_1^0)^2 + t_{2,i2}^2(\widehat{\beta}_2 - \beta_2^0)^2 + e_{i2}^2\right] > \tau_1 w_{12}^2/3\right)\\
&\le \sum_{k=1}^2 P\left(\sum_{i=1}^m \eta_{i2}t_{k,i2}^2(\widehat{\beta}_k - \beta_k^0)^2 > \tau_1 w_{12}^2/9\right) + P\left(\sum_{i=1}^m \eta_{i2}e_{i2}^2 > \tau_1 w_{12}^2/9\right). \quad (8.34)
\end{aligned}$$

The first inequality in the above follows from the Cauchy-Schwarz inequality. For $k = 1, 2$, we have

$$\begin{aligned}
\sum_{i=1}^m \eta_{i2}t_{k,i2}^2 &\le \sqrt{\sum_{i=1}^m \eta_{i2}^2}\sqrt{\sum_{i=1}^m t_{k,i2}^4} \le \sqrt{\sum_{i=1}^m (n\eta_i)^2}\sqrt{\sum_{i=1}^m (\boldsymbol{\zeta}_{i2}'\boldsymbol{t}_k/n)^4}\\
&\le n^{-1} \times \sqrt{\sum_{i=1}^m \|\boldsymbol{\zeta}_{i2}\|^4\|\boldsymbol{t}_k\|^4}\\
&\le n^{-1} \times O(n^{5/4}) = O(n^{1/4})
\end{aligned} \tag{8.35}$$

by considering $\eta_1 = (720m^4)^{-1}$, $\eta_i \sim i^{-4}$ for $i = 2, \ldots, m$, and Holder's inequality. By adapting the arguments in Lemma 8.1, we can show

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/5}). \tag{8.36}$$

Now we focus on the first two probabilities in (8.34). Combining (8.35), (8.36) and the condition that $n^{3/20}\tau_1 w_{12}^2 \to \infty$, we can show that they converge to zero. Let $V_2 = \sum_{i=1}^m \eta_{i2} e_{i2}^2$. Since e_{i2} follows $N(0, \sigma^2/n)$ as discussed above, we have

$$E(nV_2) \sim \sigma^2 \text{ and } Var(nV_2) \sim \sigma^4. \tag{8.37}$$

As for the third probability in (8.34), we have

$$\begin{aligned} P(V_2 > \tau_1 w_{12}^2/9) &\leq P\left(|nV_2 - EnV_2| > n\tau_1 w_{12}^2/9 - EnV_2\right) \\ &\leq \frac{Var(nV_2)}{(n\tau_1 w_{12}^2/9 - EnV_2)^2} \to 0 \end{aligned}$$

where the second inequality follows from the Chebyshev's inequality and the condition that $n\tau_1 w_{12}^2 \to \infty$. This completes the proof of (8.33), thus shows "$\mathcal{P}_{12} g_0 = 0 \Longrightarrow \mathcal{P}_{12}\widehat{g} = 0$".

Next we prove "$\mathcal{P}_{12}\widehat{g} = 0 \Longrightarrow \mathcal{P}_{12} g_0 = 0$" by showing the equivalent statement "$\mathcal{P}_{12} g_0 \neq 0 \Longrightarrow \mathcal{P}_{12}\widehat{g} \neq 0$". To show $\mathcal{P}_{12}\widehat{g} \neq 0$, it suffices to show

$$P(U_2 \leq \tau_1 w_{12}^2) \to 0 \tag{8.38}$$

based on the previous discussions. We first establish the following inequalities:

$$\begin{aligned} P(U_2 \leq \tau_1 w_{12}^2) &\leq P(|U_2 - EW_2| \geq EW_2 - \tau_1 w_{12}^2) \\ &\leq P(|U_2 - W_2| \geq (EW_2 - \tau_1 w_{12}^2)/2) + P(|W_2 - EW_2| \geq (EW_2 - \tau_1 w_{12}^2)/2) \\ &\leq I + II, \end{aligned}$$

where $W_2 = \sum_{i=1}^m \eta_{i2}(e_{i2} + a_{i2})^2$. By the Cauchy-Schwartz inequality, we have

$$|U_2 - W_2| \leq 4W_2 + 3\sum_{k=1}^2 \sum_{i=1}^m \eta_{i2} t_{k,i2}^2 (\widehat{\beta}_k - \beta_k^0)^2.$$

Thus, the term I can be further bounded by

$$\begin{aligned} I &\leq P(W_2 \geq (EW_2 - \tau_1 w_{12}^2)/16) + \sum_{k=1}^2 P\left(\sum_{i=1}^m \eta_{i2} t_{k,i2}^2 (\widehat{\beta}_k - \beta_k^0)^2 \geq (EW_2 - \tau_1 w_{12}^2)/24\right) \\ &\leq I_1 + I_2. \end{aligned}$$

To analyze the order of I_1, I_2 and II, we need to study the order of EW_2 and $VarW_2$. Note that $\mathcal{P}_{12}g_0 \neq 0$ implies $a_{i_0 2} \neq 0$ for some $1 \leq i_0 \leq m$. Thus, we have

$$E(W_2) \geq E(\eta_{i_0 2}(e_{i_0 2} + a_{i_0 2})^2) \geq \eta_{i_0 2} a_{i_0 2}^2, \tag{8.39}$$

$$\begin{aligned} Var(W_2) &= \sum_{i=1}^{m} \eta_{i2}^2 Var((e_{i2} + a_{i2})^2) = \sum_{i=1}^{m} \eta_{i2}^2 (4n^{-1} a_{i2}^2 \sigma^2 + 2n^{-2}\sigma^4) \\ &\leq 4n^{-1}\sigma^2 \sum_{i=1}^{m} a_{i2}^2 + 2n^{-2}\sigma^4 \leq 4n^{-1}\sigma^2 \|\mathcal{P}_{12}g_0\|_2 + 2n^{-2}\sigma^2 = O(n^{-1}) \qquad (8.40) \end{aligned}$$

By (8.39) and Lemma 8.1, we know $(EW_2 - \tau_1 w_{12}^2)$ is bounded away from zero. Then, by Chebyshev's inequality, we have

$$II \lesssim \frac{Var(W_2)}{(EW_2 - \tau_1 w_{12}^2)^2} \to 0$$

by (8.40). As for the term I_2, we can also show it converges to zero by considering (8.35) and (8.36). For the term I_1, we have

$$I_2 = P(16(W_2 - EW_2) \geq -\tau_1 w_{12}^2 - 15EW_2) \lesssim \frac{Var(W_2)}{(\tau_1 w_{12}^2 + 15EW_2)^2} \to 0$$

since $(EW_2 + \tau_1 w_{12}^2)$ is bounded away from zero and $Var(W_2) = O(n^{-1})$.

Step 3: Prove $\widehat{\beta}_j = 0 \Longleftrightarrow \beta_j^0 = 0$

In this step we consider selection consistency for β_j. Without loss of generality, we assume that $\beta_2^0 = 0$. First, we rewrite (8.27) as $Q(\beta_1, \beta_2, \boldsymbol{\theta}) + \lambda_1 \sum_{j=1}^{d} w_{0j}|\beta_j| + \tau_1 \sum_{j=1}^{d} w_{1j}\theta_j$. Applying the Taylor expansion to (8.27), we have

$$\begin{aligned} \frac{\partial L_2(\beta_1, \beta_2, \widehat{\boldsymbol{\theta}})}{\partial \beta_2} &= \frac{\partial Q(\beta_1, \beta_2, \widehat{\boldsymbol{\theta}})}{\partial \beta_2} + \lambda_1 w_{02}\mathrm{sign}(\beta_2) \\ &= \frac{\partial Q(\beta_1^0, \beta_2^0, \widehat{\boldsymbol{\theta}})}{\partial \beta_2} + \frac{\partial^2 Q(\beta_1^0, \beta_2^0, \widehat{\boldsymbol{\theta}})}{\partial \beta_1 \partial \beta_2}(\beta_1 - \beta_1^0) + \frac{\partial^2 Q(\beta_1^0, \beta_2^0, \widehat{\boldsymbol{\theta}})}{\partial \beta_2^2}(\beta_2 - \beta_2^0) \\ &\quad + \lambda_1 w_{02}\mathrm{sign}(\beta_2). \qquad (8.41) \end{aligned}$$

Recall that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/5})$ by (8.36). Thus, in the below, we only consider β_1 and β_2 satisfying $|\beta_1 - \beta_1^0| = O_P(n^{-1/5})$ and $|\beta_2 - \beta_2^0| = O_P(n^{-1/5})$.

By (8.30), the first term in (8.41) can be written as

$$\begin{aligned} &-2\left(\sum_{i=2}^{m}\sum_{j=1}^{d} + \sum_{i=1}^{1}\sum_{j=2}^{d}\right)\left[\frac{(a_{ij} + e_{ij})t_{2,ij}}{1 + \widehat{\theta}_j \eta_{ij}/w_{1j}}\right] \\ &= -2\left(\sum_{i=2}^{m}\sum_{j=1}^{d} + \sum_{i=1}^{1}\sum_{j=2}^{d}\right)\left[\frac{\mathbf{G}_1'\boldsymbol{\zeta}_{ij}\boldsymbol{\zeta}_{ij}'\boldsymbol{t}_2 + \boldsymbol{\epsilon}'\boldsymbol{\zeta}_{ij}\boldsymbol{\zeta}_{ij}'\boldsymbol{t}_2}{n^2(1 + \widehat{\theta}_j \eta_{ij}/w_{1j})}\right] \\ &= O_P(n^{-1/2}), \qquad (8.42) \end{aligned}$$

where the last equality follows from the orthogonality of the constructed $\{\boldsymbol{\zeta}_{ij}\}$ and Lindeberger-Feller theorem. As for the second term of (8.41), we have

$$\begin{aligned}\frac{\partial^2 Q(\beta_1^0,\beta_2^0,\widehat{\boldsymbol{\theta}})}{\partial\beta_1\partial\beta_2}(\beta_1-\beta_1^0) &= 2\left(\left[\sum_{i=2}^{m}\sum_{j=1}^{d}+\sum_{i=1}^{1}\sum_{j=2}^{d}\right]\frac{t_{1,ij}t_{2,ij}}{1+\widehat{\theta}_j\eta_{ij}/w_{1j}}(\beta_1-\beta_1^0)\right)\\ &= 2\left(\left[\sum_{i=2}^{m}\sum_{j=1}^{d}+\sum_{i=1}^{1}\sum_{j=2}^{d}\right]\frac{\boldsymbol{t}_1'\boldsymbol{\zeta}_{ij}\boldsymbol{\zeta}_{ij}'\boldsymbol{t}_2}{n^2(1+\widehat{\theta}_j\eta_{ij}/w_{1j})}(\beta_1-\beta_1^0)\right)\\ &\leq O(n^{-1})O_P(n^{-1/5})=O_P(n^{-6/5}),\end{aligned}$$

where the last inequality follows from the orthogonality of the constructed $\{\boldsymbol{\zeta}_{ij}\}$ and the forms of $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, i.e. (8.22) and (8.23). By applying similar analysis to the third term in (8.41), we know its order is also $O_P(n^{-1/5})$. In summary, we have

$$\frac{\partial L_2(\beta_1,\beta_2,\widehat{\boldsymbol{\theta}})}{\partial\beta_2}=O_P(n^{-1/5})+\lambda_1 w_{02}\text{sign}(\beta_2). \tag{8.43}$$

We first show "$\beta_2^0=0 \Longrightarrow \widehat{\beta}_2=0$". If $\beta_2^0=0$, then the range of β_2 in (8.43) is $(-Cn^{-1/5},Cn^{-1/5})$ for some $C>0$. By the assumed condition that $n^{1/5}\lambda_1 w_{02}\to\infty$, we can conclude that $\partial L_2(\beta_1,\beta_2,\widehat{\boldsymbol{\theta}})/\partial\beta_2<0$ for $\beta_2\in(-Cn^{-1/5},0)$ and $\partial L_2(\beta_1,\beta_2,\widehat{\boldsymbol{\theta}})/\partial\beta_2>0$ for $\beta_2\in(0,Cn^{-1/5})$. In other words, we have

$$L_2(\beta_1,0,\widehat{\boldsymbol{\theta}})=\min_{|\beta_2|\leq Cn^{-1/5}}L_2(\beta_1,\beta_2,\widehat{\boldsymbol{\theta}}) \text{ with probability tending to one,}$$

which implies $\widehat{\beta}_2=0$. We next show "$\widehat{\beta}_2=0\Longrightarrow\beta_2^0=0$" by showing the equivalent statement that "$\beta_2^0\neq 0\Longrightarrow\widehat{\beta}_2\neq 0$". For simplicity, we assume $\beta_2^0=1$ which means that $\beta_2\in(1-Cn^{-1/5},1+Cn^{-1/5})$. Then, by considering the condition $n^{1/5}\lambda_1 w_{02}\to\infty$ in (8.43), we have $\partial L_2(\beta_1,\beta_2,\widehat{\boldsymbol{\theta}})/\partial\beta_2>0$ which implies that $\widehat{\beta}_2>0$. This completes the whole proof. □