Research Article

Linearity Identification for General Partial Linear Single-Index Models

Shaogao Lv and Luhong Wang

Center of Statistics, Southwestern University of Finance and Economics, Chengdu, China

Correspondence should be addressed to Shaogao Lv; lvsg716@swufe.edu.cn

Received 5 April 2016; Accepted 25 August 2016

Academic Editor: Alessio Merola

Copyright © 2016 S. Lv and L. Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Partial linear models, a family of popular semiparametric models, provide us with an interpretable and flexible assumption for modelling complex data. One challenging question in partial linear models is the structure identification for the linear components and the nonlinear components, especially for high dimensional data. This paper considers the structure identification problem in the general partial linear single-index models, where the link function is unknown. We propose two penalized methods based on a modern dimension reduction technique. Under certain regularity conditions, we show that the second estimator is able to identify the underlying true model structure correctly. The convergence rate of the new estimator is established as well.

1. Introduction

Partially linear models (PLMs), containing both linear and nonlinear additive components, provide a useful class of tools for modelling complex data. PLMs have wide applications in practice due to their flexibility and interpretability.

Given the data set $\{y_i, \mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}\}_{i=1}^n$, where y_i is the response and $\mathbf{x}_i^{(1)} = (x_{i1}, \dots, x_{ip})$ and $\mathbf{x}_i^{(2)} = (x_{i(p+1)}, \dots, x_{i(p+q)})$ are vectors of covariates, the basic PLM has the following form:

$$y_i = b + \beta_o' \mathbf{x}_i^{(1)} + \sum_{j=1}^q f_{o,j}(x_{i(p+j)}) + \varepsilon_i, \quad (1)$$

where b is the intercept, β_o is a vector of unknown parameters for linear terms, each $f_{o,j}$ is an unknown nonlinear function from $\mathbb{R}^q$ to $\mathbb{R}$, and ε_i 's are i.i.d. random errors with zero conditional mean and a finite variance.

Given that the linear term and nonlinear term are known in advance, estimation and inference for various PLMs have been well studied in literature, such as [1–3]. Under the same settings, motivated by sparse models and the Lasso idea developed rapidly in recent years, there exists a substantial work on variable selection and estimation for sparse PLMs, including [4–6] and others. However, faced with complicated

data-generating processes, a key and challenging question to use PLM is how to decide which features should be assigned to be linear and which are nonlinear, especially for high dimensional problems such as text categorization and biology. Furthermore, the difficulty level of model search increases dramatically as the data dimension grows due to the curse of dimensionality. In the last several years, a number of methods have been proposed to address various aspects of this problem. Two commonly used methods are the screening and hypothesis testing procedures. The screening method is sometimes useful in practice but lacks theoretical justifications, while, for hypothesis testing, it is often hard to construct proper test statistics and the tests may have a poor performance when the number of covariates is large. To handle estimation and component identification problems efficiently for PLM, Zhang et al. [7] originally proposed a penalized method called LAND for automatically identifying linear and nonlinear terms for PLM. This method is based on the orthogonal decomposition of Sobolev space of order 2, consisting of linear subspace and nonlinear subspaces. Another main stream of research is based on spline approximation for nonparametric functions and penalized likelihood. For example, Huang et al. [8] start with a nonparametric additive model but use spline series expansions to approximate the component functions. This

spline series approximation approach allows them to prove selection consistency for more general functions. Lian et al. [9] proposed a doubly SCAD-penalized approach for partial linear structure identification problems of nonpolynomial (NP) dimensionality. These above methods can be easily extended to be generalized partial linear models; that is, the added link function is known in advance. If the link function is unknown, a standard technique in the literature is the two-step estimation. Firstly given that the component functions are fixed, a rough estimator of the link function is obtained; then the derived link function is fixed, estimating the component functions again. Repeat the process until convergence. However, this method is unstable and quite sensitive to the choice of penalty parameters. Moreover, this method involves a nonconvex optimization, which can not guarantee satisfactory numerical results.

In this paper we consider a general single-index models with the partial linear form

$$y_i = G\left(\beta_o' \mathbf{x}_i^{(1)} + \sum_{j=1}^q f_{o,j}(x_{i(p+j)}), \varepsilon_i\right), \quad (2)$$

where the error term ε_i is assumed to be independent of covariates, besides satisfying the standard moment conditions mentioned above. The link function $G(\cdot)$ is typically unknown. Note that additive structure of the unknown link function and the error is not assumed in model (2). This model is quite general; it includes the classical generalized partial linear model with the form $y_i = G_1(\beta_o' \mathbf{x}_i^{(1)} + \sum_{j=1}^q f_{o,j}(x_{i(p+j)})) + \varepsilon_i$, as well as the partial linear transformation model $G_2(y_i) = \beta_o' \mathbf{x}_i^{(1)} + \sum_{j=1}^q f_{o,j}(x_{i(p+j)}) + \varepsilon_i$. Clearly, when the link function $G(\cdot)$ is not specified, the partial linear term is identifiable only up to a multiplicative scalar because any location-scale change in $\{\beta_o' \mathbf{x}_i^{(1)} + \sum_{j=1}^q f_{o,j}(x_{i(p+j)})\}$ can be absorbed into the link function.

By making use of the cubic spline approximation to smooth functions and the regularization technique, we develop a new penalized method to identify model structure for the general partial linear models (2). The proposed method is formulated by the theory of sufficient dimension reduction and two different variant forms of eigenvalues problems. We specially use a modern sufficient dimension technique for estimating the linear and nonlinear terms, called gradient-based kernel dimension reduction (gKDR), proposed by Fukumizu and Leng [10]. This method shows some specific advantages compared with the above-mentioned techniques; for example, the gKDR method can handle any type of variable for y including multivariate or nonvectorial one in the same way. Besides, the linear condition and constant variance conditions are avoided when the gKDR is adopted.

Based on the gKDR and two variant forms for the eigenvalues problem, we design two different penalized approaches by adding twofold Lasso-type penalties. The first algorithm is a nonconvex optimization problem with equality constraint, which can be solved efficiently by an *alternating direction method of multipliers* (ADMM). To provide better statistical

properties, a convex relaxation of the eigenvalues problem is used to formulate our other method, a semidefinite programme. Using this convex approach, we can establish the estimation consistency and support recovery for the general partial linear models. The convex optimization problem can also be solved in a polynomial time algorithm by the ADMM.

The rest of the article is organized as follows. In Section 2 we transform model (2) into a surrogate model by the cubic spline approximation. Next we introduce the gKDR in the literature of dimension reduction. Then we propose two different penalized approaches based on two variants of eigenvalues problems. Statistical properties of the new estimators, including its convergence rate and selection consistency, are established in Section 3. All the proofs are relegated to the Appendix.

2. Problem Formulation

Notation. We collect here some standard notation used throughout the paper. For matrices A, B with the same dimension $\langle A, B \rangle := \text{tr}(A^T B)$ is the Frobenius inner product, and $\|A\|_F := \sqrt{\langle A, A \rangle}$ is the square Frobenius norm. $\|x\|_q$ is the usual norm in the Euclidean spaces. $\|A\|_{a,b}$ is (a, b) -norm defined to be ℓ_b norm of the vector of rowwise ℓ_a norms of A . We denote $A \leq B$, meaning that $B - A$ is semidefinite positive.

We mainly consider the problem of linear identification of the general single-index model (2) and estimate relevant variables to the response. For nonparametric components in model (2), one often uses spline or wavelet approximation as finite representations, because of their good approximation capabilities to functions in various nonparametric classes and computational advantages. In our specific setting, we adopt the cubic spline approximation, because two important properties of cubic spline are critical for our analysis. One property is that all the its bases are linearly independent of each other. Another property is that the function x is contained in the cubic spline bases. We refer the reader to [11] for a detailed description for spline functions. Thus, benefited from the two properties of cubic spline, each function of this spline space has a unique finite representation, and the coefficient of the base function x corresponds to the linear part.

Let $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ be the total covariate defined on $\mathcal{X} \subset \mathbb{R}^{p+q}$, and $\mathcal{Y} \in \mathbb{R}$ is the response. For each covariate x_j , $j = 1, \dots, p+q$, let $\phi(x_j) = [\phi_1(x_j), \dots, \phi_M(x_j)]$ be an M -dimensional basis expansion, where $\phi_1(x_j) = x_j$. For any vector $\beta_{nj} = (\beta_{j1}, \dots, \beta_{jM})$, we define $f_j(x) = \beta_{nj}' \phi(x)$ as cubic spline-based function we consider. With this spline finite representation, each $f_{o,j}$ in our model (2) can be approximated well by this form of f_j above. Let $\Phi_n(\mathbf{x}) = (\phi(x_1), \dots, \phi(x_{p+q}))$ be $M \times (p+q)$ vector-valued base function for any $\mathbf{x} = (x_1, \dots, x_{p+q}) \in \mathcal{X}$. To represent our model by a unified form, we denote $\beta_{sj} = (\beta_{o,j}, 0, \dots, 0)$ as the coefficient representation of the j th linear-form variable. Let $\beta_p = (\beta_{s1}', \dots, \beta_{sp}')'$ and $\beta_q = (\beta_{o,1}', \beta_{o,q}')'$, where $\beta_{o,j}$ ($j = 1, \dots, q$) is the coefficient corresponding to the j th true nonlinear component. Thus, let $\theta_o = (\beta_p', \beta_q')'$ be

$M \times (p + q)$ approximating coefficient, our focus on the original model (2) is now transformed to the following one:

$$y_i = G(\theta'_o \Phi_n(\mathbf{x}_i), \varepsilon_i). \quad (3)$$

This semiparametric model (3) indicates essentially that the response y depends solely on the generalized predictor vector Φ_n through a linear combination $\theta'_o \Phi_n$, or, equivalently, y is independent of the original covariate $\mathbf{x}$ when $\theta'_o \Phi_n$ is given. In the literature, the theory of sufficient dimension reduction provides an effective starting point to estimate θ'_o without loss of regression information of y on $\mathbf{x}$ and without assuming the specific link function. The popular methods for sufficient dimension reduction include sliced inversion regression [12], sliced average variance estimation [13], contour regression, directional regression [14], and references therein.

In this article, we use the gKDR formulated by the reproducing kernel techniques. This method shows some specific advantages compared with the above-mentioned techniques; for example, the gKDR method can handle any type of variable for y including multivariate or nonvectorial one in the same way. Besides, the nonparametric nature of the kernel method avoids making strong assumptions on the distribution of covariates, the response, or the conditional probability, which are required usually in many classical dimension reduction methods such as SIR, pHd, and contour regression.

2.1. gKDR Method. To give the gKDR method, we need some technical notations. For a compact set Ω in some metric space, we denote a *positive-definite kernel* on $\Omega \times \Omega$, that is, a symmetric function $k : \Omega \times \Omega \rightarrow \mathbb{R}$ satisfying the following semidefinite positive condition: $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$ for any $x_1, \dots, x_n$ in Ω and $c_1, \dots, c_n \in \mathbb{R}$. It is known [15] that a positive-definite kernel on Ω is uniquely associated with a Hilbert space $\mathcal{H}$ consisting of functions on Ω , and $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS) associated with the kernel k .

For short, we denote $\mathcal{X} := \{\Phi_n(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ as a subset of $\mathbb{R}^{M(p+q)}$. Let $(\mathcal{X}, B_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $(\mathcal{Y}, B_{\mathcal{Y}}, \mu_{\mathcal{Y}})$ be measurable spaces and (Z, Y) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with probability P . Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be positive-definite kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and the corresponding RKHSs denoted by $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. It is assumed that $E(k_{\mathcal{X}}(Z, Z))$ and $E(k_{\mathcal{Y}}(Y, Y))$ are finite. Now we introduce two integral operators, so that an appropriate variant of the problem of estimating θ_o in (3) can be established. The *cross-covariance operator* $C_{YZ} : \mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}$ is defined as the operator as follows:

$$(C_{YZ}f)(y) = \int k(y, \tilde{y}) f(\tilde{z}) dP(\tilde{z}, \tilde{y}), \quad \forall f \in \mathcal{H}_{\mathcal{X}}. \quad (4)$$

Similarly, C_{ZZ} denotes the *self-covariance operator* on $\mathcal{H}_{\mathcal{X}}$; that is,

$$(C_{ZZ}f)(\mathbf{z}) = \int k(\mathbf{z}, \tilde{\mathbf{z}}) f(\tilde{\mathbf{z}}) dP_{\mathcal{X}}(\tilde{\mathbf{z}}), \quad \forall f \in \mathcal{H}_{\mathcal{X}}. \quad (5)$$

Remark that these definitions are natural extensions of the ordinary covariance metrics on Euclidean spaces. Although

C_{YZ} and C_{ZZ} depend on the kernels, we omit the dependence for notational simplicity.

With these definitions, we now define the population $M(p+q) \times M(p+q)$ matrix $M(\mathbf{z}) = (M_{ij}(\mathbf{z}))$ by

$$M_{ij}(\mathbf{z}) = \left\langle C_{YZ} C_{XX}^{-1} \frac{\partial k_{\mathcal{X}}(\cdot, \mathbf{z})}{\partial \mathbf{z}^i}, C_{YZ} C_{ZZ}^{-1} \frac{\partial k_{\mathcal{Y}}(\cdot, \mathbf{z})}{\partial \mathbf{z}^j} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}, \quad (6)$$

$$i, j = 1, \dots, M(p+q).$$

Reference [10] has shown that the eigenvector corresponding to the largest eigenvalue of the $M(\mathbf{z})$ is equal to the true vector θ_o for any $\mathbf{z}$, provided that the maximal eigenvalue is simple. To be precise, the following equality holds:

$$\theta_o = \tilde{v}_{\max}(\mathbb{E}[M(\mathbf{Z})]), \quad (7)$$

where $\tilde{v}_{\max}(A)$ is denoted to be the maximum eigenvector of the matrix A . Thus, the problem of estimating θ_o of (3) is equivalent to the eigenvalue problem of solving $M(\mathbf{z})$.

We now turn to the empirical version of $M(\mathbf{z})$, so as to estimate θ_o based on available sample. Denote $n \times n$ Gram matrices $(k_{\mathcal{X}}(\mathbf{z}_i, \mathbf{z}_j))$ and $(k_{\mathcal{Y}}(y_i, y_j))$ by $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$, respectively. Let $\nabla \mathbf{k}_{\mathcal{X}}(\mathbf{z}) = (\partial k_{\mathcal{X}}(\mathbf{z}_1, \mathbf{z})/\partial \mathbf{z}, \dots, \partial k_{\mathcal{X}}(\mathbf{z}_n, \mathbf{z})/\partial \mathbf{z}) \in \mathbb{R}^{n \times M(p+q)}$. Then, the gKDR proposes the eigenvectors of $M(p+q) \times M(p+q)$ symmetric matrix:

$$M_n(\mathbf{z}) := \nabla \mathbf{k}_{\mathcal{X}}(\mathbf{z})' (G_{\mathcal{X}} + n\epsilon_n I_n)^{-1} G_{\mathcal{Y}} (G_{\mathcal{X}} + n\epsilon_n I_n)^{-1} \cdot \nabla \mathbf{k}_{\mathcal{X}}(\mathbf{z}), \quad (8)$$

where ϵ_n is a regularization parameter in Thikonov-type regularization. Then, we define

$$\widehat{M}_n := \frac{1}{n} \sum_{i=1}^n M_n(\mathbf{z}_i) \quad (9)$$

as an empirical estimator of θ_o . In the next two subsections, we introduce two variational forms with respect to the eigenvalue problems, so that our proposed methods based on Lasso idea can be formulated naturally.

2.2. A Nonconvex Variational Algorithm. In order to employ the popular Lasso idea, we need to transform the eigenvalue problem (7) to an optimization scheme. Recall the following *Courant-Fischer variational representation* of the maximal eigenvalue and eigenvector:

$$\begin{aligned} \max_{\theta \in \mathbb{R}^{M(p+q)}} \quad & \theta' \mathbb{E}[M(\mathbf{Z})] \theta, \\ \text{s.t.} \quad & \|\theta\|_2^2 = 1. \end{aligned} \quad (10)$$

To identify the linear and nonlinear terms in model (3), we define an estimator by adding twofold convex penalties to (10) such that some sparse estimators might be generated. To be precise, let $\theta = (\theta_1, \dots, \theta_{p+q})$, where each $\theta_j \in \mathbb{R}^M$. The penalized estimation method is defined as follows:

$$\begin{aligned} \hat{\theta}_n = \arg \min_{\theta \in \mathbb{R}^{M(p+q)}} & \left\{ -\theta' \widehat{M}_n \theta + \lambda \sum_{j=1}^{p+q} (\alpha \|\theta_{j,1}\|_2 + (1-\alpha) \|\theta_{j,-1}\|_2) \right\}, \\ \text{s.t.} \quad & \|\theta\|_2^2 = 1, \end{aligned} \quad (11)$$

where $\theta_{j,1}$ is the first element of θ_j and $\theta_{j,-1}$ is the last $M-1$ component of θ_j , $j = 1, \dots, p+q$. Note that λ is the penalty parameter and α is a tuning parameter, typically, $\alpha = M/(1+M)$ for normalization. Our penalized method is motivated by the following fact. For any given function $f(x) = \sum_{k=1}^M \beta_k \phi_k(x)$, by the cubic spline representation, as mentioned earlier, the first spline base is the single function x and all the bases are linearly independent of each other. Consequently, $\beta_1 = 0$ and $\beta_k \neq 0$ with $k > 1$ if and only if f is a nonlinear function. Based on this and the lasso constraint in (11), the proposed method can identify both the linear and nonlinear terms, of course, also finding the relevant feature simultaneously.

We observe that optimization of (11) is complicated, since it belongs to a class of constrained nonconvex programmes. In this article, we consider an alternative formulation of (11); precisely, we use the ADMM [16]. ADMM is an algorithm that is intended to blend the decomposability of dual ascent with the superior convergence properties of the method of multipliers. Let $\rho(\theta) = \sum_{j=1}^{p+q} (\alpha \|\theta_{j,1}\|_2 + (1-\alpha) \|\theta_{j,-1}\|_2)$, and by introducing an additional constraint, we can reformulate algorithm (11) as

$$\begin{aligned} \min_{\theta, \gamma \in \mathbb{R}^{M(p+q)}} & \left\{ \infty \times \Pi_2(\gamma) - \theta' \widehat{M}_n \theta + \lambda \rho(\theta) \right\}, \\ \text{s.t.} \quad & \theta = \gamma, \end{aligned} \quad (12)$$

where Π_2 is the 0-1 indicator function for the unit ℓ_2 -ball surface in $\mathbb{R}^{M(p+q)}$ and we adopt the convention $\infty \times 0 = 0$. As in the method of multipliers, we form the augmented Lagrangian with a parameter τ :

$$\begin{aligned} L(\gamma, \theta, \nu, \tau) = \infty \times \Pi_1(\gamma) - \theta' \widehat{M}_n \theta + \lambda \rho(\theta) \\ + \nu'(\theta - \gamma) + \frac{\tau}{2 \|\theta - \gamma\|_2^2}, \end{aligned} \quad (13)$$

where (ν, τ) is the Lagrangian multipliers. Defining the residual $\mathbf{r} = \theta - \gamma$, we have

$$\begin{aligned} L(\gamma, \theta, \nu, \tau) = \infty \times \Pi_1(\gamma) - \theta' \widehat{M}_n \theta + \lambda \rho(\theta) \\ + \frac{\tau}{2 \|\mathbf{r} + \mathbf{u}\|_2^2} - \frac{\tau}{2 \|\mathbf{u}\|_2^2}, \end{aligned} \quad (14)$$

where $\mathbf{u} = \nu/\tau$ is called the scaled dual variable. Using the scaled dual variable, we can express ADMM as

$$\begin{aligned} \theta^{k+1} \\ = \arg \min_{\theta} \left\{ \lambda \rho(\theta) - \theta' \widehat{M}_n \theta + \frac{\tau}{2 \|\theta - \gamma^k + \mathbf{u}^k\|_2^2} \right\}, \end{aligned} \quad (15)$$

$$\gamma^{k+1} = \Pi_{\mathbb{B}_2}(\theta^{k+1} + \mathbf{u}^k), \quad (16)$$

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \theta^{k+1} - \gamma^{k+1}, \quad (17)$$

where $\Pi_{\mathbb{B}_2}$ is Euclidean projection onto the unit ℓ_2 -ball surface.

There are many convergence results for ADMM discussed in the literature. ADMM consists of iteratively minimizing $L(\gamma, \theta, \nu, \tau)$ with respect to γ , minimizing $L(\gamma, \theta, \nu, \tau)$ with respect to θ , and then updating the dual variable $\mathbf{u}$. The θ -update involves iterative evaluation of the proximal operator associated with

$$F(\theta) := -\theta' \widehat{M}_n \theta + \frac{\tau}{2 \|\theta - \gamma^k + \mathbf{u}^k\|_2^2}. \quad (18)$$

To this end, we use the *block coordinate descent* (BCD) [17] for ℓ_1/ℓ_2 -regularization to compute θ^{k+1} of (15). Specially at t th iteration, the BCD solves a problem of the form

$$\min_{\theta_g} \left\{ \nabla_g F(\theta_t^k)' (\theta_g - \theta_{g,t}^k) + \frac{1}{2} (\theta_g - \theta_{g,t}^k)' H_{gg} (\theta_g - \theta_{g,t}^k) + \mu \|\theta_g\|_2 \right\}, \quad (19)$$

where θ_g is the subelement of θ with group $g \in \mathcal{G}$ forming a partition of $\{1, \dots, M(p+q)\}$ and H_{gg} equals or approximates $\nabla_{gg}^2 F(\theta_t^k)$. Note that μ is taken as $\lambda\alpha$ or $\lambda(1-\alpha)$ according to the specified group. Obviously, the BCD can be viewed as a generalization of the classical coordinate descent for

ℓ_1 -regularization. In our specific setting, a direct computation on $F(\theta)$ yields that

$$\begin{aligned} \nabla_g F(\theta) = -2 (\theta' \widehat{M}_{n,g_1}, \dots, \theta' \widehat{M}_{n,g_k}) \\ + \tau (\theta + \mathbf{u}^k - \gamma^k)_g, \end{aligned} \quad (20)$$

$$[\nabla^2 F(\boldsymbol{\theta})]_{gg} = -2(\widehat{M}_n)_{gg} + \tau I_{|g|}, \quad (21)$$

where we write $g = (g_1, \dots, g_k)$ and $\widehat{M}_{n,g_l}$ is the l th column of $\widehat{M}_n$, $l = 1, \dots, k$. I_{gg} is the identity and $(\widehat{M}_n)_{gg}$ is the submatrix of $\widehat{M}_n$ with the index g . Since $(\widehat{M}_n)_{gg}$ in (21) is usually not a multiple of the identity, the solution minimizing (19) has no closed form. We here use a simple but efficient replacement; that is, $H_{gg} = h_{gg}I_{|g|}$, where h_{gg} is the largest eigenvalue of $-2(\widehat{M}_n)_{gg} + \tau I_{|g|}$, which can be computed efficiently, since $-2(\widehat{M}_n)_{gg} + \tau I_{|g|}$ is only a submatrix indexed by g and independent of $\boldsymbol{\theta}$. In this case, the iterative solution $\boldsymbol{\theta}_{t+1}^k$ of (19) is obtained by *group soft thresholding* of the Newton step:

$$\boldsymbol{\theta}_{g,t+1}^k = \text{Prox}_{\mu\|\cdot\|_2}(\boldsymbol{\theta}_{g,t}^k - h_{gg}^{-1}\nabla_g F(\boldsymbol{\theta}_t^k)), \quad (22)$$

where $\text{Prox}_{\mu\|\cdot\|_2}$ is called the soft thresholding operator associated with $\|\cdot\|_2$, defined as $\text{Prox}_{\mu\|\cdot\|_2}(\boldsymbol{\theta}_g) = (1 - \mu/\|\boldsymbol{\theta}_g\|_2)_+ \boldsymbol{\theta}_g$ for any $\boldsymbol{\theta}_g$. Under mild conditions, $\boldsymbol{\theta}_{t+1}^k$ converges to some vector as t goes to infinity, denoted by $\boldsymbol{\theta}^{k+1}$. It is worthwhile to note that, by (20), we find that the iteration in (22) can be computed parallel between groups.

With regard to projection computation of (16), it is known that $\boldsymbol{\gamma}^{k+1} = (\boldsymbol{\theta}^{k+1} + \mathbf{u}^k)/\|\boldsymbol{\theta}^{k+1} + \mathbf{u}^k\|_2$ if $\|\boldsymbol{\theta}^{k+1} + \mathbf{u}^k\|_2 \neq 0$; otherwise, $\boldsymbol{\gamma}^{k+1}$ is set to any given value on ℓ_2 -ball surface. Thus, An efficient implementation for ADMM is obtained; to be precise, the overall procedure can be stated as in Algorithm 1.

2.3. A Semidefinite Programming Relaxation. This section introduces a convex relaxation of (10). Since estimate of individual eigenvectors of $\boldsymbol{\theta}_o$ in (7) is unstable if the gap between their eigenvalues is small, it seems reasonable to instead focus on their span, that is, the principal subspace of variation. Let $\Lambda_o = \boldsymbol{\theta}_o \boldsymbol{\theta}_o'$ as the projection onto the subspace spanned by $\boldsymbol{\theta}_o$; a lesser known but equivalent variational representation of (10) is of the convex program:

$$\max_{\Theta \in \mathbb{S}_+^{M(p+q)}, \text{tr}(\Theta)=1} \text{tr}(\mathbb{E}[M(Z)]\Theta), \quad (23)$$

where $\mathbb{S}_+^{M(p+q)} = \{\Theta \in \mathbb{R}^{M(p+q) \times M(p+q)} \mid 0 \leq \Theta \leq I\}$. The optimization problem (23) is a semidefinite program (SDP), which can be solved exactly in polynomial time. If the maximal eigenvalue is simple, the optimum always achieved at a rank-one matrix, denoted by $\widehat{\Theta}$, corresponds to the maximal eigenvector; that is, $\widehat{\Theta} = \boldsymbol{\theta}_o \boldsymbol{\theta}_o'$. Motivated by this nice conclusion, we can estimate the sparse vector $\boldsymbol{\theta}_o$ by adding ℓ_1 -type regularization,

$$\begin{aligned} \widehat{\Theta} &:= \arg \max_{\Theta \in \mathbb{S}_+^{M(p+q)}} \text{tr}(\widehat{M}_n \Theta) - \lambda \|\Theta\|_{1,1}, \\ \text{s.t.} \quad &\text{tr}(\Theta) = 1, \end{aligned} \quad (24)$$

and computing the maximal eigenvector $\widehat{\boldsymbol{\theta}} = \widehat{\mathbf{v}}_{\max}(\widehat{\Theta})$.

Using the ADMM again, we rewrite (11) as the equivalent problem with equality constrained

$$\begin{aligned} \min \quad &\infty \times \mathbf{1}_{\mathcal{F}_1}(\Theta) - \langle \widehat{M}_n, \Theta \rangle + \lambda \|\Delta\|_{1,1} \\ \text{s.t.} \quad &\Theta - \Delta = 0, \end{aligned} \quad (25)$$

where $\mathcal{F}_1 = \{\Theta \in \mathbb{S}^{M(p+q)}, \text{tr}(\Theta) = 1\}$ and $\mathbf{1}_{\mathcal{F}_1}$ is the 0–1 indicator function for $\mathcal{F}_1$. The augmented Lagrangian associated with (25) has the form

$$\begin{aligned} \mathcal{L}_\tau(\Theta, \Delta, U) &:= \infty \times \mathbf{1}_{\mathcal{F}_1}(\Theta) - \langle \widehat{M}_n, \Theta \rangle + \lambda \|\Delta\|_{1,1} \\ &\quad + \frac{\tau}{2} (\|\Theta - \Delta + U\|_F^2 - \|U\|_F^2). \end{aligned} \quad (26)$$

Θ and Δ updates mainly involve computing the projection operator and the proximal operator, respectively; that is,

$$\begin{aligned} \Pi_{\mathcal{F}_1} \left(\Delta - U + \frac{\widehat{M}_n}{\tau} \right) \\ := \arg \min_{\Theta \in \mathcal{F}_1} \frac{1}{2} \left\| \Theta - \left(\Delta - U + \frac{\widehat{M}_n}{\tau} \right) \right\|_F^2, \end{aligned} \quad (27)$$

$$\mathcal{S}_{\lambda/\tau}(\Theta + U) := \arg \min_{\Delta} \frac{\lambda}{\tau} \|\Delta\|_{1,1} + \frac{1}{2} \|\Theta + U - \Delta\|_F^2,$$

where $\mathcal{S}_{\lambda/\tau}$ is the elementwise soft thresholding operator defined as

$$\mathcal{S}_{\lambda/\tau}(x) = \text{sign}(x) \max\left(|x| - \frac{\lambda}{\tau}, 0\right). \quad (28)$$

On the other hand, $\Pi_{\mathcal{F}_1}$ is the Euclidean projection onto $\mathcal{F}_1$ and has a closed form as follows.

Lemma 1 (Fantope projection). *If $U = \sum_j \gamma_j u_j u_j'$ is a spectral decomposition of U , then $\Pi_{\mathcal{F}_1}(U) = \sum_j \gamma_j^+(\theta) u_j u_j'$, where $\gamma_j^+(\theta) = \min(\max(\gamma_j - \theta, 0), 1)$ and θ satisfies the equation $\sum_j \gamma_j^+(\theta) = 1$.*

We notice that computing of $\Pi_{\mathcal{F}_1}(U)$ involves an eigen-decomposition of U and then modifying the eigenvalues by solving a monotone, piecewise linear equation.

3. Statistical Properties

Before giving statistical results of the estimator (24) in terms of consistency and model selection properties, we recall the existing error bound [10] for $M_n(\mathbf{z}) - M(\mathbf{z})$ in Frobenius norm for any $\mathbf{z}$.

Recall that the following technique conditions are required to derive the gradient-based estimator. For short-hand, let $m = M(p+q)$.

- (i) $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are separable.
- (ii) $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are measurable, and $\mathbb{E}[k_{\mathcal{X}}(Z, Z)] < \infty$, $\mathbb{E}[k_{\mathcal{Y}}(Y, Y)] < \infty$.

given: parameters $\lambda, \alpha, \epsilon_n, \tau$, the group indexes g and the kernels $k_{\mathcal{X}}, k_{\mathcal{Y}}$.
initialize: $\theta^0 = 0, \gamma^0 = 0, \mathbf{u}^0 = 0$,
for $k \geq 0$
repeat
 for $t \geq 0$ **repeat**
 $\theta_{g,t+1}^k = \text{Prox}_{\mu\|\cdot\|_2}(\theta_{g,t}^k - h_{gg}^{-1} \nabla_g F(\theta_t^k))$, for each group g ,
 $t \leftarrow t + 1$
 until $\theta_{g,t}^k$ converges to some θ_g^*
 $\theta_g^{k+1} \leftarrow \theta_g^*$, then compute $\gamma^{k+1} = \Pi_{\mathbb{B}_2}(\theta^{k+1} + \mathbf{u}^k)$,
 and $\mathbf{u}^{k+1} = \mathbf{u}^k + \theta^{k+1} - \gamma^{k+1}$
 then $k \leftarrow k + 1$
until θ^{k+1} converges to $\hat{\theta}_n$.
then return $(\hat{\theta}_n)$

ALGORITHM 1: ADMM algorithm.

- (iii) $k_{\mathcal{X}}(\mathbf{z}, \bar{\mathbf{z}})$ is continuously differentiable and $\partial k_{\mathcal{X}}(\cdot, \mathbf{z}) / \partial \mathbf{z}^i \in \mathcal{R}(C_{ZZ})$ for $i = 1, \dots, m$.
- (iv) $\mathbb{E}[k_{\mathcal{Y}}(y, Y) \mid Z = \cdot] \in \mathcal{H}_{\mathcal{X}}$ for any $y \in \mathcal{Y}$.
- (v) Define $\mathbb{E}[h(Y) \mid Z] = \varphi_h(\beta'Z)$ for any given function h . Suppose that φ_h is differentiable with respect to z , and the linear function $h \rightarrow \partial \varphi_h(\mathbf{z}) / \partial \mathbf{z}^a$ is continuous for any $\mathbf{z} \in \mathbb{R}^m$ and $a = 1, \dots, m$.
- (vi) There is $\vartheta_m \geq 0$ and $L_m \geq 0$ such that some $h_{a,\mathbf{z}}^m \in \mathcal{H}_{\mathcal{X}}^m$ satisfies

$$\frac{\partial k_{\mathcal{X}}(\cdot, \mathbf{z})}{\partial \mathbf{z}^a} = C_{ZZ}^{\vartheta_m+1} h_{a,\mathbf{z}}^m, \quad (a = 1, \dots, m), \quad (29)$$

and $\mathbb{E}[\|h_{\mathbf{z}}^a\|_{\mathcal{H}_{\mathcal{X}}}] \leq L_m$.

- (vii) Let $\alpha_m := (\mathbb{E}[k^m(Z, Z)] - \mathbb{E}[k^m(Z, \tilde{Z})])^{1/2}$, where $\tilde{Z}$ is an i.i.d. copy of Z . Then $\alpha_m / \sqrt{n} \rightarrow 0$ ($n \rightarrow \infty$).

Theorem 2. Under Assumptions (i)–(vii), for the choice $\epsilon_n = n^{\max\{1/3, 1/(2\beta+2)\}}$, we have

$$\begin{aligned} & \|M_n(\mathbf{z}) - M(\mathbf{z})\|_F \\ &= O_p \left(mL_m^2 \left(\frac{\alpha_m^2}{n} \right)^{\min\{1/3, (2\vartheta_m+1)/(4\vartheta_m+4)\}} \right), \end{aligned} \quad (30)$$

for every $\mathbf{z} \in \mathcal{X}$ as $n \rightarrow \infty$. If additionally, $mL_m^2 / \sqrt{n} \rightarrow 0$ ($n \rightarrow \infty$), then $\widehat{M}_n$ converges in probability to $\mathbb{E}(M(Z))$ of the order $O_p(mL_m^2 / \sqrt{n} + mL_m(\alpha_m^2/n)^{\min\{1/3, (2\vartheta_m+1)/(4\vartheta_m+4)\}})$ in Frobenius norm.

For simplicity, we write $\delta_o = \vec{v}_1(\mathbb{E}[M(Z)]) - \vec{v}_2(\mathbb{E}[M(Z)])$. We write the second error bound of Theorem 2 by $\Omega(n, m) := O_p(mL_m^2 / \sqrt{n} + mL_m(\alpha_m^2/n)^{\min\{1/3, (2\vartheta_m+1)/(4\vartheta_m+4)\}})$.

Theorem 3. Under the assumptions of Theorem 2, let $s = \|\Lambda_o\|_{2,0}$ and λ is chosen to satisfy $\lambda = \Omega(n, m)$; then

$$\|\widehat{\Theta} - \Lambda_o\|_F \leq \frac{4s}{\delta_o} \Omega(n, m). \quad (31)$$

The proof is shown in the Appendix. The following theorem provides a sufficient condition for support recovery by diagonal thresholding $\widehat{\Theta}$.

Theorem 4. Under the assumptions of Theorem 3, for any $t > 0$,

$$\begin{aligned} & \left| \left\{ j : (\Lambda_o)_{jj} = 0, (\widehat{\Theta})_{jj} \geq t \right\} \right| \\ &+ \left| \left\{ j : (\Lambda_o)_{jj} \geq 2t, (\widehat{\Theta})_{jj} \leq t \right\} \right| \\ &\leq \frac{4s}{\delta_o t} \Omega(n, m). \end{aligned} \quad (32)$$

Therefore, the feature selection procedure $\{j : (\widehat{\Theta})_{jj} \geq t\}$ succeeds with high probability if $\min_{j: (\Lambda_o)_{jj} \neq 0} (\Lambda_o)_{jj} \geq 2t > (8s/\delta_o)\Omega(n, m)$.

The above theorem follows immediately from the conclusion of Theorem 3.

Appendix

The following lemma is critical for our analysis, shown in [18]. It establishes a close relationship between the matrix and the subspace spanned by its some largest eigenvalues.

Lemma A.1. Let A be a symmetric matrix and E be the projection onto the subspace spanned by the eigenvectors of A corresponding to its d largest eigenvalues $\nu_1 \geq \nu_2 \geq \dots$. If $\delta_A = \nu_d - \nu_{d+1} > 0$, then

$$\frac{\delta_A}{2} \|E - F\|_F^2 \leq \langle A, E - F \rangle \quad (A.1)$$

for all F satisfying $0 \leq F \leq I$ and $\text{tr}(F) = d$.

Proof of Theorem 3

Proof. Since $\hat{v}_{\max}(\mathbb{E}[M(Z)]) = \theta_o$, recall that $\Lambda_o = \theta_o \theta_o'$ as the projection onto the subspace spanned by θ_o . Letting $E = \Lambda_o$ and $F = \hat{\Theta}$ in Lemma A.1, if $\|\hat{M}_n - \mathbb{E}[M(Z)]\|_{\infty} = \rho$, we have

$$\begin{aligned} \frac{\delta_o}{2} \|\hat{\Theta} - \Lambda_o\|_F^2 &\leq \langle \mathbb{E}[M(Z)], \hat{\Theta} - \Lambda_o \rangle \\ &\leq \rho \|\hat{\Theta} - \Lambda_o\|_{1,1} - \langle \hat{M}_n, \hat{\Theta} - \Lambda_o \rangle \\ &\leq \rho \|\hat{\Theta} - \Lambda_o\|_{1,1} \\ &\quad - \lambda \left(\|\hat{\Theta}\|_{1,1} - \|\Lambda_o\|_{1,1} \right), \end{aligned} \quad (\text{A.2})$$

where the first inequality follows from Lemma A.1, the second inequality follows from the Cauchy-Schwartz inequality, and the last one follows from the definition of $\hat{\Theta}$. Furthermore, with the choice of $\lambda \geq \rho$, it is easy to verify that

$$\begin{aligned} \frac{\delta_o}{2} \|\hat{\Theta} - \Lambda_o\|_F^2 &\leq 2\lambda \left\| (\hat{\Theta} - \Lambda_o)_S \right\|_{1,1} \\ &\leq 2s\lambda \left\| (\hat{\Theta} - \Lambda_o)_S \right\|_F, \end{aligned} \quad (\text{A.3})$$

which implies that

$$\|\hat{\Theta} - \Lambda_o\|_F \leq \frac{4s\lambda}{\delta_o}. \quad (\text{A.3})$$

Plugging the conclusion of Theorem 2 into (A.3), we complete the proof of Theorem 3. $\square$

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The first author's research is supported partially by National Natural Science Foundation of China (Grant no. 11301421), and Fundamental Research Funds for the Central Universities of China (Grants nos. JBK120509 and JBK140507), as well as KLAS-130026507.

References

- [1] C. J. Stone, "Additive regression and other nonparametric models," *The Annals of Statistics*, vol. 13, no. 2, pp. 689–705, 1985.
- [2] J. D. Opsomer and D. Ruppert, "Fitting a bivariate additive model by local polynomial regression," *The Annals of Statistics*, vol. 25, no. 1, pp. 186–211, 1997.
- [3] H. Liang, S. W. Thurston, D. Ruppert, T. Apanasovich, and R. Hauser, "Additive partial linear models with measurement errors," *Biometrika*, vol. 95, no. 3, pp. 667–678, 2008.
- [4] L. Wang, X. Liu, H. Liang, and R. J. Carroll, "Estimation and variable selection for generalized additive partial linear models," *The Annals of Statistics*, vol. 39, no. 4, pp. 1827–1851, 2011.
- [5] X. Liu, L. Wang, and H. Liang, "Estimation and variable selection for semiparametric additive partial linear models," *Statistica Sinica*, vol. 21, no. 3, pp. 1225–1248, 2011.
- [6] L. Wang, L. Xue, A. Qu, and H. Liang, "Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates," *The Annals of Statistics*, vol. 42, no. 2, pp. 592–624, 2014.
- [7] H. H. Zhang, G. Cheng, and Y. Liu, "Linear or nonlinear? Automatic structure discovery for partially linear models," *Journal of the American Statistical Association*, vol. 106, no. 495, pp. 1099–1112, 2011.
- [8] J. Huang, F. Wei, and S. Ma, "Semiparametric regression pursuit," *Statistica Sinica*, vol. 22, no. 4, pp. 1403–1426, 2012.
- [9] H. Lian, P. Du, Y. Li, and H. Liang, "Partially linear structure identification in generalized additive models with NP-dimensionality," *Computational Statistics & Data Analysis*, vol. 80, pp. 197–208, 2014.
- [10] K. J. Fukumizu and C. L. Leng, "Gradient-based kernel dimension reduction for regression," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 359–370, 2014.
- [11] L. Schumaker, *Spline Functions: Basic Theory*, John Wiley & Sons, New York, NY, USA, 1981.
- [12] K.-C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–342, 1991.
- [13] R. D. Cook and S. Weisberg, "Discussion of li," *Journal of the American Statistical Association*, vol. 86, pp. 328–332, 1991.
- [14] B. Li and S. Wang, "On directional regression for dimension reduction," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 997–1008, 2007.
- [15] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.
- [16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.
- [17] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming—Series B*, vol. 117, no. 1, pp. 387–423, 2009.
- [18] V. Q. Vu and J. Lei, "Minimax sparse principal subspace estimation in high dimensions," *The Annals of Statistics*, vol. 41, no. 6, pp. 2905–2947, 2013.

Hindawi

Submit your manuscripts at
<http://www.hindawi.com>