Linearity Identification for General Partial Linear Single-Index Models

Shaogao Lv

[1]Statistical College, Southwestern University of Finance and Economics, ChengDu, 611130, China;

Abstract

Partial linear models, a family of popular semi-parametric models, provides us 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. Furthermore, we propose an iterative algorithm to implement the procedure and illustrate its performance by simulated and real examples.

Key Words and Phrases:

1 Introduction

Partially linear models(PLM), containing both linear and nonlinear additive components, provide a useful class of tools for modeling complex data. PLM 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, $\mathbf{x}_i^{(1)} = (x_{i1}, ..., x_{ip})$ and $\mathbf{x}_i^{(2)} = (x_{i(p+1)}, ..., x_{i(p+q)})$ are vectors of covariates, the basic PLM has the following form

$$y_i = b + \boldsymbol{\beta}_o' \mathbf{x}_i^{(1)} + \sum_{j=1}^{q} f_{o,j}(x_{i(p+j)}) + \varepsilon_i \tag{1.1}$$

where b is the intercept, $\boldsymbol{\beta}_o$ 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 $\varepsilon_i^{,}$s are i.i.d. random errors with zero conditional mean and a finite variance.

Given the linear term and nonlinear term are known in advance, estimation and inference for various PLMs have been well studied in literature, such as (Stone, 1985; Opsomer et al., 1997; Liang et al., 2008). Under the same settings, motivated by sparse models and the Lasso idea developed rapidly in recent years, there exist a substantial work on variable selection and estimation for sparse PLMs, including (Wang et al., 2011; Liu et al., 2011; Wang et al., 2014) and others. However, faced with complicated data-generating processes, a key and challenging question to use PLM is that 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 and Liu (2011) 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 are based on spline approximation for nonparametric functions and penalized likelihood. For example, Huang et al. (2012) starts 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. (2014) proposed a doubly SCAD-penalized approach for partial linear structure identification problems of non-polynomial (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 non-convex optimization, which can not guarantee satisfactory numerical results.

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

$$y_i = G(\boldsymbol{\beta}_o'\mathbf{x}_i^{(1)} + \sum_{j=1}^{q} f_{o,j}(x_{i(p+j)}), \varepsilon_i), \tag{1.2}$$

where the error term ε_i's 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 this model (1.2). This model is quite general; it includes the classical generalized partial linear model with the form: $y_i = G_1(\boldsymbol{\beta}_o'\mathbf{x}_i^{(1)} + \sum_{j=1}^{q} f_{o,j}(x_{i(p+j)})) + \varepsilon_i$, as well as the partial linear transformation model: $G_2(y_i) = \boldsymbol{\beta}_o'\mathbf{x}_i^{(1)} + \sum_{j=1}^{q} f_{o,j}(t_{i(p+j)}) + \varepsilon_i$. 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 $\{\boldsymbol{\beta}_o'\mathbf{x}_i^{(1)} + \sum_{j=1}^{q} f_{o,j}(x_{i(p+j)})\}$ 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 (1.2). The proposed method is formulated by the theory of sufficient dimension reduction and two different variant forms of eigenvalues problems. Specially, we 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 (2013). This method shows some specific advantages compared with the above mentioned techniques, for example, the gKDR method can handle any type of variables 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 two-fold Lasso-type penalties. The first algorithm is a non-convex 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 another method, a semi-definite 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 solved in a polynomial time algorithm by the ADMM.

The rest of the article is organized as follows. In Section 2 we transform the model (1.2) 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 two variants of eigenvalues problems. Statistical properties of the new estimators, including its convergence rate and selection consistency, are established in Section 3. Section 4 contains simulated and real examples to illustrate finite sampling performance of the proposed methods. 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 $\langle A, B\rangle := tr(A^T B)$ 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 the (a, b)-norm defined to be the ℓ_b norm of the vector of rowwise ℓ_a norms of A. We denote by $A \preceq B$, meaning that $B - A$ is semi-definite positive.

We mainly consider the problem of linear identification of the general single-index model (1.2), as well as estimate relevant variables to the response. For nonparametric components in the model (1.2), 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 is critical for our analysis. One is that all the its bases are linearly independent of each others; Another property is that the function x is contained in the cubic spline bases. We refer the reader to (Schumaker, 1981) 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, $j = 1, ..., p + q$, let $\boldsymbol{\phi}(x_j) = [\phi_1(x_j), ..., \phi_M(x_j)]$ be an M-dimensional basis expansion, where $\phi_1(x_j) = x_j$. For any vector $\boldsymbol{\beta}_{nj} = (\beta_{j1}, ..., \beta_{jM})$, we define $f_j(x) = \boldsymbol{\beta}'_{nj}\boldsymbol{\phi}(x)$ as cubic spline-based function we consider. With this spline finite representation, each $f_{o,j}$ in our model (1.2) can be approximated well by this form of f_j above. Let $\Phi_n(\mathbf{x}) = (\boldsymbol{\phi}(x_1), ..., \boldsymbol{\phi}(x_{p+q}))$ be the $M \times (p + q)$ vector-vauled base function for any $\mathbf{x} = (x_1, ..., x_{p+q}) \in \mathcal{X}$. To represent our model by a unified form, we denote $\beta_{sj} = (\beta_{o,j}, 0, ..., 0)$ as the coefficient representation of the j-the linear-form variable. Let $\boldsymbol{\beta}_p = (\beta'_{s1}, ..., \beta'_{sp})'$ and $\boldsymbol{\beta}_q = (\boldsymbol{\beta}'_{o,1}, \boldsymbol{\beta}'_{o,q})'$, where $\boldsymbol{\beta}_{o,j}(j = 1, ..., q)$ is the coefficient corresponding to the j-th true nonlinear component. Thus, let $\boldsymbol{\theta}_o = (\boldsymbol{\beta}'_p, \boldsymbol{\beta}'_q)'$ be the $M \times (p + q)$ approximating coefficient, our focus on the original model (1.2) is now transformed to the following one

$$y_i = G(\boldsymbol{\theta}_o' \Phi_n(\mathbf{x}_i), \varepsilon_i). \tag{2.1}$$

This semiparametric model (2.1) indicates essentially that the response y depends solely on the generalized predictor vector Φ_n through a linear combination $\boldsymbol{\theta}_o'\Phi_n$, or equivalently, y is independent of the original covariate $\mathbf{x}$ when $\boldsymbol{\theta}_o'\Phi_n$ is given. In the literature, the theory of sufficient dimension reduction provides an effective starting point to estimate $\boldsymbol{\theta}_o'$ 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 (Li, 1991), sliced average variance estimation(Cook and Weisberg, 1991), contour regression, directional regression (Li and Wang, 2007) 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 variables 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, contour regression and so on.

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 $\Omega \times \Omega$, i.e., a symmetric function $k : \Omega \times \Omega \to \mathbb{R}$ satisfying the following semi-definite positive condition: $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ for any $x_1, ..., x_n$ in Ω and $c_1, ..., c_n \in \mathbb{R}$. It is known (Aronszajn, 1950) 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 $\mathcal{Z} := \{\Phi_n(\mathbf{x}), , \mathbf{x} \in \mathcal{X}\}$ as a subset of $\mathbb{R}^{M(p+q)}$. Let $(\mathcal{Z}, B_{\mathcal{Z}}, \mu_{\mathcal{Z}})$ and $(\mathcal{Y}, B_{\mathcal{Y}}, \mu_{\mathcal{Y}})$ be measurable spaces, and (Z, Y) be a random variable on $\mathcal{Z} \times \mathcal{Y}$ with probability P. Let $k_{\mathcal{Z}}$ and $k_{\mathcal{Y}}$ be positive-definite kernels on $\mathcal{Z}$ and $\mathcal{Y}$ respectively, and the corresponding RKHSs denoted by $\mathcal{H}_{\mathcal{Z}}$ and $\mathcal{H}_{\mathcal{Y}}$. It is assumed that $E(k_{\mathcal{Z}}(Z, Z))$ 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 $\boldsymbol{\theta}_o$ in (2.1) can be established. The *cross-covariance* *operator* $C_{YZ} : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{Y}}$ is defined as the operator as follows.

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

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

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

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

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

Reference (Fukumizu and Leng, 2013) has shown that the eigenvector corresponding to the largest eigenvalue of the $M(\mathbf{z})$ is equal to the true vector $\boldsymbol{\theta}_o$ for any $\mathbf{z}$, provided that the maximal eigenvalue is simple. To be precise, the following equality holds

$$\boldsymbol{\theta}_o = \vec{v}\text{max}(\mathbb{E}[M(Z)]), \tag{2.2}$$

where $\vec{v}\text{max}(A)$ is denoted to be the maximum eigenvector of the matrix A. Thus, the problem of estimating $\boldsymbol{\theta}_o$ of (2.1) 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 $\boldsymbol{\theta}_o$ based on available sample. Denote the $n \times n$ Gram matrices $(k_{\mathcal{Z}}(\mathbf{z}_i, \mathbf{z}_j))$ and $(k_{\mathcal{Y}}(y_i, y_j))$ by $G_{\mathcal{Z}}$ and $G_{\mathcal{Y}}$ respectively. Let $\nabla \mathbf{k}_{\mathcal{Z}}(\mathbf{z}) = (\partial k_{\mathcal{Z}}(\mathbf{z}_1, \mathbf{z})/\partial \mathbf{z}, ..., \partial k_{\mathcal{Z}}(\mathbf{z}_n, \mathbf{z})/\partial \mathbf{z}) \in \mathbb{R}^{n \times M(q+p)}$. Then, the gKDR proposes the eigenvectors of the $M(p+q) \times M(p+q)$ symmetric matrix

$$M_n(\mathbf{z}) := \nabla \mathbf{k}_{\mathcal{Z}}(\mathbf{z})' (G_{\mathcal{Z}} + n\epsilon_n I_n)^{-1} G_{\mathcal{Y}} (G_{\mathcal{Z}} + n\epsilon_n I_n)^{-1} \nabla \mathbf{k}_{\mathcal{Z}}(\mathbf{z}), \tag{2.3}$$

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

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

as an empirical estimator of $\boldsymbol{\theta}_o$. In 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 Non-convex Variational Algorithm

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

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^{M(p+q)}} \boldsymbol{\theta}'\mathbb{E}[M(Z)]\boldsymbol{\theta}, \quad s.t.\, \|\boldsymbol{\theta}\|_2^2 = 1. \tag{2.4}$$

To identify the linear and nonlinear terms in model (2.1), we define an estimator by adding two-fold convex penalties to (2.4) such that some sparse estimators might be generated. To be precise, let $\boldsymbol{\theta} = (\theta_1, ..., \theta_{p+q})$, where each $\theta_j \in \mathbb{R}^M$. The penalized estimation method is defined as follows:

$$\hat{\boldsymbol{\theta}}_n = \arg \min_{\boldsymbol{\theta}\in\mathbb{R}^{M(p+q)}} \left\{ -\boldsymbol{\theta}'\widehat{M}_n\boldsymbol{\theta} + \lambda \sum_{j=1}^{p+q} \big(\alpha\|\theta_{j,1}\|_2 + (1-\alpha)\|\theta_{j,-1}\|_2\big) \right\}, \quad s.t.\, \|\boldsymbol{\theta}\|_2^2 = 1, \tag{2.5}$$

where $\theta_{j,1}$ is the first element of θ_j, and $\theta_{j,-1}$ is the last $M-1$ component of θ_j, $j = 1, ..., p+q$. 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 others. 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 (2.5), the proposed method can identify both the linear and nonlinear terms, of course, also find the relevant feature simultaneously.

We observe that, optimization of (2.5) is complicated, since it belongs to a class of constrained nonconvex programmes. In this article, we consider an alternative formulation of(2.5), precisely, we use the ADMM (Boyd et al., 2010). 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 $\rho(\boldsymbol{\theta}) = \sum_{j=1}^{p+q} \big(\alpha\|\theta_{j,1}\|_2 + (1-\alpha)\|\theta_{j,-1}\|_2\big)$, and by introducing an additional constraint, we can reformulate the algorithm (2.5) as

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

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

$$L(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\nu}, \tau) = \infty \times \Pi_1(\boldsymbol{\gamma}) - \boldsymbol{\theta}' \widehat{M}_n \boldsymbol{\theta} + \lambda \rho(\boldsymbol{\theta}) + \boldsymbol{\nu}'(\boldsymbol{\theta} - \boldsymbol{\gamma}) + \tau/2 \|\boldsymbol{\theta} - \boldsymbol{\gamma}\|_2^2,$$

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

$$L(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\nu}, \tau) = \infty \times \Pi_1(\boldsymbol{\gamma}) - \boldsymbol{\theta}' \widehat{M}_n \boldsymbol{\theta} + \lambda \rho(\boldsymbol{\theta}) + \tau/2 \|\mathbf{r} + \mathbf{u}\|_2^2 - \tau/2 \|\mathbf{u}\|_2^2,$$

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

$$\begin{align} \boldsymbol{\theta}^{k+1} &= \arg\min_{\boldsymbol{\theta}} \left\{ \lambda \rho(\boldsymbol{\theta}) - \boldsymbol{\theta}' \widehat{M}_n \boldsymbol{\theta} + \tau/2 \|\boldsymbol{\theta} - \boldsymbol{\gamma}^k + \mathbf{u}^k\|_2^2 \right\}, \tag{2.6} \\ \boldsymbol{\gamma}^{k+1} &= \Pi_{\mathbb{B}_2}\big(\boldsymbol{\theta}^{k+1} + \mathbf{u}^k\big), \tag{2.7} \\ \mathbf{u}^{k+1} &= \mathbf{u}^k + \boldsymbol{\theta}^{k+1} - \boldsymbol{\gamma}^{k+1}, \tag{2.8} \end{align}$$

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 $L(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\nu}, \tau)$ with respect to $\boldsymbol{\gamma}$, minimizing $L(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\nu}, \tau)$ with respect to $\boldsymbol{\theta}$, and then updating the dual variable $\mathbf{u}$. The $\boldsymbol{\theta}$-update involves iterative evaluation of the proximal operator associated with

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

To this end, we use the *block coordinate descent* (BCD) (Tseng and Yun, 2009) for ℓ_1/ℓ_2-regularization to compute $\boldsymbol{\theta}^{k+1}$ of (2.6). Specially, at t-th iteration, the BCD solves a problem of the form

$$\min_{\boldsymbol{\theta}_g} \left\{ \nabla_g F(\boldsymbol{\theta}_t^k)'(\boldsymbol{\theta}_g - \boldsymbol{\theta}_{g,t}^k) + \frac{1}{2}(\boldsymbol{\theta}_g - \boldsymbol{\theta}_{g,t}^k)' H_{gg} (\boldsymbol{\theta}_g - \boldsymbol{\theta}_{g,t}^k) + \mu \|\boldsymbol{\theta}_g\|_2 \right\}, \tag{2.9}$$

where $\boldsymbol{\theta}_g$ is the subelement of $\boldsymbol{\theta}$ with group $g \in \mathcal{G}$ forming a partition of $\{1, ..., M(p+q)\}$, and H_{gg} equals or approximates $\nabla_{gg}^2 F(\boldsymbol{\theta}_t^k)$. 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 $F(\boldsymbol{\theta})$ yields that

$$\begin{align} \nabla_g F(\boldsymbol{\theta}) &= -2\big(\boldsymbol{\theta}' \widehat{M}_{n,g_1}, ..., \boldsymbol{\theta}' \widehat{M}_{n,g_k}\big) + \tau\big(\boldsymbol{\theta} + \mathbf{u}^k - \boldsymbol{\gamma}^k\big)_g, \tag{2.10} \\ [\nabla^2 F(\boldsymbol{\theta})]_{gg} &= -2(\widehat{M}_n)_{gg} + \tau I_{|g|}, \tag{2.11} \end{align}$$

where we write $g = (g_1, ..., g_k)$ and $\widehat{M}_{n,g_l}$ is the l-th column of $\widehat{M}_n$, $l = 1, ..., k$. I_{gg} is the identity and $(\widehat{M}_n)_{gg}$ is the sub-matrix of $\widehat{M}_n$ with the index g. Since $(\widehat{M}_n)_{gg}$ in (2.11) is usually not a multiple of the identity, the solution minimizing (2.9) 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 sub-matrix indexed by g and independent of $\boldsymbol{\theta}$. In this case, the iterative solution $\boldsymbol{\theta}^k_{t+1}$ of (2.9) is obtained by *group soft-thresholding* of the Newton step:

$$\boldsymbol{\theta}^k_{g,t+1} = \mathrm{Prox}_{\mu\|\cdot\|_2}\big(\boldsymbol{\theta}^k_{g,t} - h_{gg}^{-1}\nabla_g F(\boldsymbol{\theta}^k_t)\big) \tag{2.12}$$

where $\mathrm{Prox}_{\mu\|\cdot\|_2}$ is called the soft thresholding operator associated with $\|\cdot\|_2$, which defined as $\mathrm{Prox}_{\mu\|\cdot\|_2}(\boldsymbol{\theta}_g) = \left(1 - \frac{\mu}{\|\boldsymbol{\theta}_g\|_2}\right)_+ \boldsymbol{\theta}_g$ for any $\boldsymbol{\theta}_g$. Under mild conditions, $\boldsymbol{\theta}^k_{t+1}$ converges to some vector as t goes to infinity, denoted by $\boldsymbol{\theta}^{k+1}$. It is worthwhile to note that, by (2.10), we find that the iteration in (2.12) can be computed parallel between groups.

With regard to projection computation of (2.7), it is known that $\boldsymbol{\gamma}^{k+1} = \frac{\boldsymbol{\theta}^{k+1} + \mathbf{u}^k}{\|\boldsymbol{\theta}^{k+1} + \mathbf{u}^k\|_2}$ if $\|\boldsymbol{\theta}^{k+1} + \mathbf{u}^k\|_2 \neq 0$, otherwise, $\boldsymbol{\gamma}^{k+1}$ is set be to any given value on the ℓ_2-ball surface. Thus, An efficient implementation for ADMM is obtained, to be precise, the overall procedure can be stated as follows.

Algorithm ADMM algorithm

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

2.3 A Semidefinite Programming Relaxation

This section intrduces a convex relaxation of (2.4). Since estimate of individual eigenvectors of $\boldsymbol{\theta}_o$ in (2.2) is unstable if the gap between their eigenvalues is small, it seems reasonable to instead focus on their span, i.e. 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 (2.4) is of the convex program

$$\max_{\Theta\in\mathbb{S}_+^{M(p+q)},\, \mathrm{tr}(\Theta)=1} \mathrm{tr}(\mathbb{E}[M(Z)]\Theta), \tag{2.13}$$

where $\mathbb{S}_+^{M(p+q)} = \{\Theta \in \mathbb{R}^{M(p+q)\times M(p+q)} \,\big|\, 0 \preceq \Theta \preceq I\}$. The optimization problem (2.13) 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 $\bar{\Theta}$, corresponds to the maximal eigenvector, that is, $\bar{\Theta} = \boldsymbol{\theta}_o\boldsymbol{\theta}_o'$. Motivated by this nice conclusion, we can estimate the sparse vector $\boldsymbol{\theta}_o$ by adding ℓ_1-type regularization,

$$\widehat{\Theta} := \arg\max_{\Theta\in\mathbb{S}_+^{M(p+q)}} \mathrm{tr}(\widehat{M_n}\Theta) - \lambda\|\Theta\|_{1,1}, \quad s.t.\, \mathrm{tr}(\Theta) = 1, \tag{2.14}$$

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

Using the ADMM again, we rewrite (2.5) as the equivalent problem with equality constrained

$$\begin{aligned} \min \quad & \infty \times \mathbf{1}_{\mathcal{F}_1}(\Theta) - \langle \widehat{M_n}, \Theta\rangle + \lambda\|\Delta\|_{1,1} \\ \text{s.t.} \quad & \Theta - \Delta = 0, \end{aligned} \tag{2.15}$$

where $\mathcal{F}_1 = \{\Theta \in \mathbb{S}^{M(p+q)}, \mathrm{tr}(\Theta) = 1\}$ and $\mathbf{1}_{\mathcal{F}_1}$ is the $0-1$ indicator function for $\mathcal{F}_1$. The augmented Lagrangian associated with (2.15) has the form

$$\mathcal{L}_\tau(\Theta, \Delta, U) := \infty \times \mathbf{1}_{\mathcal{F}_1}(\Theta) - \langle \widehat{M_n}, \Theta\rangle + \lambda\|\Delta\|_{1,1} + \frac{\tau}{2}\left(\|\Theta - \Delta + U\|_F^2 - \|U\|_F^2\right).$$

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

$$\Pi_{\mathcal{F}_1}(\Delta - U + \widehat{M_n}/\tau) := \arg\min_{\Theta\in\mathcal{F}_1} \frac{1}{2}\|\Theta - (\Delta - U + \widehat{M_n}/\tau)\|_F^2, \tag{2.16}$$

and

$$\mathcal{S}_{\lambda/\tau}(\Theta + U) := \arg\min_{\Delta} \frac{\lambda}{\tau}\|\Delta\|_{1,1} + \frac{1}{2}\|\Theta + U - \Delta\|_F^2, \tag{2.17}$$

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

$$\mathcal{S}_{\lambda/\tau}(x) = \mathrm{sign}(x)\max(|x| - \lambda/\tau, 0).$$

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 $\Pi_{\mathcal{F}_1}(U)$ involves an eigendecomposition of U, and then modifying the eigenvalues by solving a monotone, piecewise linear equation.

3 Statistical Properties

Before giving statistical results of the estimator (2.14) in terms of consistency and model selection properties, we recall the existing error bound (Fukumizu et al., 2013) for $M_n(\mathbf{z}) - M(\mathbf{z})$ in Frobenius norm for any $\mathbf{z}$.

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

(i) $\mathcal{H}_{\mathcal{Z}}$ and $\mathcal{H}_{\mathcal{Y}}$ and separable.

(ii) $k_{\mathcal{Z}}$ and $k_{\mathcal{Y}}$ are measurable, and $\mathbb{E}[k_{\mathcal{Z}}(Z,Z)] < \infty$, $\mathbb{E}[k_{\mathcal{Y}}(Y,Y)] < \infty$.

(iii) $k_{\mathcal{Z}}(\mathbf{z}, \tilde{\mathbf{z}})$ is continuously differentiable and $\partial k_{\mathcal{Z}}(\cdot, \mathbf{z})/\partial \mathbf{z}^i \in \mathcal{R}(C_{ZZ})$ for $i = 1, ..., m$.

(iv) $\mathbb{E}[k_{\mathcal{Y}}(y, Y) | \, Z = \cdot\,] \in \mathcal{H}_{\mathcal{Z}}$ for any $y \in \mathcal{Y}$.

(v) Define $\mathbb{E}[h(Y)| \, Z] = \varphi_h(\boldsymbol{\beta}' Z)$ for any given function h. Suppose that φ_h is differentiable with respect to z, and the linear function $h \to \partial \varphi_h(\mathbf{z})/\partial \mathbf{z}^a$ is continuous for any $\mathbf{z} \in \mathbb{R}^m$ and $a = 1, ..., m$.

(vi) There is $\vartheta_m \geq 0$ and $L_m \geq 0$ such that some $h^m_{a,\mathbf{z}} \in \mathcal{H}^m_{\mathcal{Z}}$ satisfies

$$\partial k_{\mathcal{Z}}(\cdot, \mathbf{z})/\partial \mathbf{z}^a = C_{ZZ}^{\vartheta_m + 1} h^m_{a,\mathbf{z}}, \quad (a = 1, ..., m),$$

and $\mathbb{E}[\|h^a_Z\|_{\mathcal{H}_{\mathcal{Z}}}] \leq L_m$.

(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 $\frac{\alpha_m}{\sqrt{n}} \to 0 \quad (n \to \infty)$.

Theorem 1. *Under Assumptions (i)-(vii), for the choice* $\varepsilon_n = n^{\max\{\frac{1}{3}, \frac{1}{2\beta+2}\}}$, *we have*

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

for every $\mathbf{z} \in \mathcal{Z}$ *as* $n \to \infty$. *If additionally,* $mL_m^2/\sqrt{n} \to 0$ $(n \to \infty)$, *then* $\widehat{M}_n$ *converges in probability to* $\mathbb{E}(M(Z))$ *of the order* $O_p\left(mL_m^2/\sqrt{n} + mL_m\left(\frac{\alpha_m^2}{n}\right)^{\min\{\frac{1}{3}, \frac{2\vartheta_m+1}{4\vartheta_m+4}\}}\right)$ *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 1 by $\Omega(n,m) := O_p\left(mL_m^2/\sqrt{n} + mL_m\left(\frac{\alpha_m^2}{n}\right)^{\min\{\frac{1}{3}, \frac{2\vartheta_m+1}{4\vartheta_m+4}\}}\right)$.

Theorem 2. *Under the assumptions of Theorem 1. Let* $s = \|\Lambda_o\|_{2,0}$ *and* λ *is chosen to satisfy* $\lambda = \Omega(n,m)$, *then*

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

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

Theorem 3. *Under the assumptions of Theorem 2. For any* $t > 0$

$$\left|\{j : (\Lambda_o)_{jj} = 0,\, (\widehat{\Theta})_{jj} \ge t\}\right| + \left|\{j : (\Lambda_o)_{jj} \ge 2t,\, (\widehat{\Theta})_{jj} \le t\}\right| \le \frac{4s}{\delta_o t}\Omega(n,m).$$

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

The above theorem follows immediately from the conclusion of Theorem 2.

4 Numerical Experiments

4.1 Simulated Examples

4.2 Real Examples

APPENDIX

The following lemma is critical for our analysis, appeared in (Vu et al., 2013). It establishes a close relationship between the matrix and the subspace spanned by its some largest eigenvalues.

Lemma 2. *Let A be a symmetric matrix and E be the projection onto the subspace spanned by the eigenvectors of A coressponding to its d largest eigenvalues $\nu_1 \geq \nu_2 \geq \ldots$. If $\delta_A = \nu_d - \nu_{d+1} > 0$, then*

$$\frac{\delta_A}{2} \|E - F\|_F^2 \leq \langle A, E - F \rangle$$

for all F satisfying $0 \preceq F \preceq I$ and $tr(F) = d$.

Proof of Theorem 2

Proof. Since $\vec{v_{\max}}\big(\mathbb{E}[M(Z)]\big) = \boldsymbol{\theta}_o$, and recall that $\Lambda_o = \boldsymbol{\theta}_o \boldsymbol{\theta}_o'$ as the projection onto the subspace spanned by $\boldsymbol{\theta}_o$. Letting $E = \Lambda_o$ and $F = \widehat{\Theta}$ in Lemma 2, if $\|\widehat{M}_n - \mathbb{E}[M(Z)]\|_\infty = \rho$, we have

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

where the first inequality follows from Lemma 2, the second inequality the Cauchy-Schwartz inequality, and the last one follows from the definition of $\widehat{\Theta}$. Furthermore, with the choice of $\lambda \geq \rho$, it is easy to verify that

$$\frac{\delta_o}{2} \|\widehat{\Theta} - \Lambda_o\|_F^2 \leq 2\lambda \|(\widehat{\Theta} - \Lambda_o)_S\|_{1,1} \leq 2s\lambda \|(\widehat{\Theta} - \Lambda_o)_S\|_F,$$

which implies that

$$\|\widehat{\Theta} - \Lambda_o\|_F \leq \frac{4s\lambda}{\delta_o}. \tag{4.1}$$

Plugging the conclusion of Theorem 1 into (4.1), we complete the proof of Theorem 2. □

References

N., Aronszajn. (1950). Theory of reporudcing kernels. *Tran. Am. Math. Sco.*, *68*, 337–404.

S., Boyd, N., Parikh, E., Chu, B., Peleato and J., Eckstein. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trend. Mach. Learn.*, *3*, 1–122.

R. D., Cook and S., Weisberg. (1991). Discussion of Li (1991). *J. Amer. Stat. Assoc.*, *86*, 328–332.

K. J., Fukumizu and C. L., Leng. (2013). Gradient-based kernel dimension reduction for regression. *J. Amer. Stat. Assoc.*, *109*, 359–370.

B., Li, and S., Wang. On directional regression for dimension reduction. *J Amer. Stat. Assoc.*, *102*, 997–1008.

K. C., Li. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Stat. Assoc.*, *86*, 316–327.

L., Schumaker. (1981). Spline Functions: Basic Theory. *Wiley, New York.*

P., Tseng and S., Yun. (2009). A coodinate gradient descent method for nonsmooth separable minimization. *Math. Program. ser. B*, *117*, 387–423.

V. Q. Vu and J. Lei. (2013). Minimax sparse principal subspace estimation in high dimenisons. *Ann. Statis.*, *41*, 2905–2947.

H. H., Zhang, G., Cheng, and Y., Liu. (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *J. Amer. Stat. Assoc.*, *106*, 1099–1112.

L., Wang, X., Liu, H., Liang and R. J., Carroll. (2011). Estimation and variable selection for generalized additive partial linear models. *Ann. Statist.*, *39*, 1827–1851.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, *13*, 689–705.

J. D., Opsomer, D., Ruppert. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.*, *25*, 186–211.

H. Liang, S., Thurston, D., Ruppert, T., Apanasovich, R., Hauser. (2008). Additive partial linear models with measurement errors. *Biometrika.*, *95*, 667–678.

X., Liu, L., Wang and H., Liang. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Stat Sin.*, *21*, 1225–1248.

L., Wang, L., Xue, A., Qu and H., Liang. (2014). Estimation and model selection in generalized additive partial linear models for high-dimensional correlated data. *Ann. Statist*, *42*, 592–624.

J., Huang, F., Wei, S., Ma. (2012). Semiparametric regression pursuit. *Statist. Sin.*, *22*, 1403–1426.

H., Lian, P., Dub, Y. Z., Li and H. Liang. (2014). Partially linear structure identification in generalized additive models with NP-dimensionality. *Comput. Statist.Data Anal.*, *80*, 197–208.