Specification LASSO and an Application in Financial Markets

Chaohua Dong*

Zhongnan University of Economics and Law, China

Shaoran Li†

University of Cambridge, UK

May 31, 2021

Abstract

This paper proposes the method of Specification-LASSO in a flexible semi-parametric regression model that allows for the interactive effects between different covariates. Specification-LASSO extends LASSO and Adaptive Group LASSO to achieve both relevant variable selection and model specification. Specification-LASSO also gives preliminary estimates that facilitate the estimation of the regression model. Monte Carlo simulations show that the Specification-LASSO can accurately specify partially linear additive models with interactive regressors. Finally, the proposed methods are applied in an empirical study, which examines the topic proposed by Freyberger et al. (2020), which argues that firms' sizes may have interactive effects with other security-specific characteristics, which can explain the stocks excess returns together.

KEYWORDS: Variable Selection; Model Selection; Interaction;

JEL CLASSIFICATION: C14; G12.

*Electronic address: chaohuadong@outlook.com

†Electronic address: sl736@cam.ac.uk

1 Introduction

In a data-rich era, researchers are more likely to suffer both "variable selection" and "specification" challenges. "Variable selection" problem is incurred due to the ease of data attainability, so vast of data are available when researchers intend to model. This seems to be trivial if the number of observations n is relatively large compared with the number of potential covariates P. However, in recent empirical studies that have large P and small n, which causes the classical analysis tool failing to work. Therefore, it is crucial to determine which subset of candidate variables should be considered. Meanwhile, another challenge comes from the model specification, as one may be dazzled to choose a suitable model from a model zoo. In general, all parametric analyses have the risk of misspecification. Thus, nonparametric analysis is introduced to relax the functional form restrictions. Although this helps to increase the model flexibility, the "curse of dimensionality" causes the extremely low convergence rate of estimation when the dimension of independent variables is more than three.

Suppose we observe a sample of data $\{(Y_i, \boldsymbol{P}_i) : 1 \leqslant i \leqslant n\}$, where i represents the i^{th} individual. $\boldsymbol{P}_i$ is a $P \times 1$ large dimensional vector of potential covariates where only the $Q \times 1$ subset $\boldsymbol{Q}_i$ contains relevant regressors to explain or predict the variation of Y_i, which presents a sparse model if $Q << P$.

We suppose:

$$E(Y_i|\boldsymbol{P}_i) = \theta_i + h(\boldsymbol{Q_i}), \quad i = 1, 2, \ldots, n, \tag{1}$$

where θ_i is the intercept whereas $h(\boldsymbol{Q}_i)$ is an unknown multi-variate function of $\boldsymbol{Q}_i$. Most researchers specify an additive semi-parametric structure on $h(\boldsymbol{Q}_i)$ as:

$$h(\boldsymbol{Q}_i) = \sum_{q=1}^{Q} f_q(X_{iq}), \tag{2}$$

where $f_q(X_{iq})$ is an unknown uni-variate function. Models like Equation 2 are called additive nonparametric regressions and are widely discussed by Hastie and Tibshirani (1990), Linton (1997), Linton (2000), and Linton and Härdle (1996).

The Equation 2 avoids the curse of dimensionality by imposing an additive structure, but can be inefficient if some of the relevant covariates only have linear effects as the rate of convergence for nonparametric function $f_q(X_{qi})$ is slower than $O(n^{-1/2})$.

Therefore, a partially linear additive semi-parametric model is proposed to take advantages of linear effects as:

$$h(\boldsymbol{Q_i}) = \theta + \sum_{l=1}^{L} \beta_l X_{il} + \sum_{q=L+1}^{Q} f_q(X_{iq}), \tag{3}$$

where we distinguish L linear effects from $\boldsymbol{Q}_i$, and the coefficients of linear part can be estimated at the rate of convergence $O(n^{-1/2})$, as discussed in Wang et al. (2007) and Ma and Yang (2011). Similar models of Equation 3 are also studied by Li (2000), Fan and Li (2003) and Liang et al. (2008).

Unfortunately, both additive models omit potential interactions between covariates. Pairwise interactions between covariates are quite common in both economic and financial studies.

Example 1.1. In macroeconomics, most production functions specify a interactive term of capital and labour inputs such as:

$$\text{Cobb-Douglas: } Y = \Gamma X_C^{\alpha} X_L^{\beta} + \epsilon$$

Example 1.2. In microeconomics, Deaton and Muellbauer (1980) document the utility model of a household (Y) containing interactions between eating and drinking (X_E, X_D) for foodstuffs, housing and fuel (X_H, X_F) for shelters, and television and sports (X_T, X_S) for entertainment.

$$Y = m_{ED}(X_E, X_D) + m_{HF}(X_H, X_F) + m_{TS}(X_T, X_S) + \epsilon$$

Example 1.3. In environment studies, Dong et al. (2019) study effects of CO_2 and solar irradiance (SI) on the global sea level (Y_{SL}) rise. They specify the model as:

$$Y_{SL} = m(X_{CO_2}, X_{SI}) + \epsilon,$$

and they verify the interactive effects between CO_2 (X_{CO_2}) and solar irradiance (X_{SI}) through empirical results.

Example 1.4. In finance, Freyberger et al. (2020) argue that assets returns at time t is predictable by stock characteristics, such as capitalization and book-to-market ratio, at $t-1$ as

$$\boldsymbol{Y}_t = \boldsymbol{\theta}_t + \overbrace{\sum_{q \neq s}^{Q} m_{qs}(\boldsymbol{X}_{qt-1} \cdot \boldsymbol{X}_{st-1})}^{\text{interaction with firm size Xs}} + \overbrace{\sum_{q=1}^{Q} m_q(\boldsymbol{X}_{qt-1})}^{\text{uni-variate}},$$

and they find significant effects of interactions between firms' sizes and other characteristics. In this paper, we will revisit this study using our methods.

Interactions among covariates refer to the circumstance that marginal effects of the j^{th} variable X_j on Y are determined by other relevant covariates. Sperlich et al. (2002) illustrate the importance of interactions in the additive model, and propose a marginal integration style estimation and test methods to solve the potential interactions in the model. However, their methods cannot be applied to a high-dimensional case, not only due to the enormous workload but also the failure of estimation when $P > n$. From the above examples and Sperlich et al. (2002), we can conclude that higher-order interactions are barely discussed due to both the curse of dimensionality and interpretation issues. In this paper, we mainly discuss pairwise interactions among variables, although our methods can be easily extended to higher-order interactions.

Based on the aforementioned research and examples, it is more reasonable to expand $h(\boldsymbol{Q}_i)$ in Equation 2 to three components, including linear, nonlinear and pairwise interactive parts.

Compared with specifying the structure of an unknown multivariate function $h(\boldsymbol{Q}_i)$, selecting relevant variables under a high-dimensional setting is more widely discussed. The most popular way for achieving this goal is LASSO (Least Absolute Shrinkage and Selection Operator) style variables selection methods. Tibshirani (1996) proposes this method to perform both variable selection and regularization in the linear model under high-dimensional cases.

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{W} \alpha_j X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{W} |\alpha_j|, \tag{4}$$

In Equation 4, λ_n is a data driving tuning parameter, and the attractive property of LASSO is that it can achieve initial selection by shrinking some $\alpha = 0$ and estimation even if $P >> n$. A necessary condition for consistent selection of LASSO is discussed by Zhao and Yu (2006) and Zou (2006), which is called irrepresentable condition (discussed in subsection 4.1). This condition restricts the correlation between relevant and irrelevant components to be relatively small.

To relax this condition, Zou (2006) proposes Adaptive LASSO, which can achieve consistent selection under mild conditions:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{W} \beta_j X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{W} \hat{w}_j |\beta_j|, \tag{5}$$

where the weight $\hat{w}_j$ is data-dependent and typically chosen as $\hat{w}_j = |\hat{\alpha}_j|^{-\gamma}$ for some $\gamma > 0$, and $\hat{\alpha}_j$ is a preliminary consistent estimate in Equation 4. X_j with a smaller estimate $\hat{\alpha}_j$ will be penalized more severely, and the variable with $\alpha = 0$ will be smoothed out.

As for selecting nonparametric functions, Lin and Zhang (2006) introduce COSSO (COmponent Selection and Smoothing Operator), where they consider the model selection in a general setting of the smoothing spline analysis of variance (SS-ANOVA) framework, shown as:

$$h(\boldsymbol{X}_i) = b + \sum_{j=1}^{d} f_j(X_i^{(j)}) + \sum_{j<k} f_{jk}(X_i^{(j)}, X_i^{(k)}) + \dots.$$

This model can provide large flexibility in terms of the form of nonparametric functions, such as higher dimensional functions. However, in COSSO, it only works under $P < n$, which means variables considered are not allowed to exceed the number of observations. Furthermore, Lin and Zhang (2006) do not give a detailed discussion of the selection of the linear part. Finally, this selection model is not facilitated with the initial estimation. All of these issues will be solved by our method.

Moreover, Huang et al. (2010) introduce a selection and estimation method of an additive nonparametric model inspired by both group LASSO as in Yuan and Lin (2006) and adaptive group LASSO as in Wang and Leng

(2008). They use a linear combination of B-splines basis $\phi_k, 1 \geqslant k \geqslant m_n$ to approximate any potential unknown function as:

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk}\phi_k(x).$$

Next, they consider the penalized least squares criterion

$$L_n(\mu, \beta_n) = \sum_{i=1}^{n}[Y_i - \mu - \sum_{j=1}^{P}\sum_{k=1}^{m_n}\beta_{jk}\phi_k(X_{ij})]^2 + \lambda_n \sum_{j=1}^{P}\hat{w}_{nj}\|\boldsymbol{\beta}_{nj}\|_2,$$

where λ_n is a tuning parameter while $\|\beta_{nj}\|_2$ is the L^2 norm of the j^{th} coefficient vector $\boldsymbol{\beta}_{nj} = (\beta_{j1}, \ldots, \beta_{jm_n})^\intercal$, and

$$\hat{w}_{nj} = \begin{cases} \|\tilde{\beta}_{nj}\|_2^{-1} & \text{for} \quad \|\tilde{\beta}_{nj}\|_2 > 0 \\ \infty & \text{for} \quad \|\tilde{\beta}_{nj}\|_2 = 0 \end{cases},$$

where $\tilde{\beta}_{nj}$ is an initial and consistent estimate. Huang et al. (2010) also compare adaptive group LASSO model with COSSO by Lin and Zhang (2006), concluding that, when the number of observations is small, adaptive group LASSO has much higher accuracy in terms of selecting relevant variables in the semi-parametric additive model.

This paper proposes a Specification-LASSO (S-LASSO) for both variables selection and model specification of a **partially linear additive semi-parametric model with interactions**, which can be applied when $P > n$. S-LASSO can achieve variables selection, model specification, and initial estimation at the same time. S-LASSO firstly use levels, B-splines bases and pairwise tensor products of all potentially relevant variables to approximate linear, nonlinear and interactive effects, respectively, and then it extends a two-step procedure to give consistent selection.

In the first step, S-LASSO uses ordinary LASSO to consider all bases indifferently to attain the initial selection and estimates. In the second step, S-LASSO clusters these bases into different groups according to linear, non-linear and interactive parts, and then an adaptive group LASSO is applied to give a final selection and estimation results. The estimates from the first step help the second step to set group-specific penalty-weighting parameters, which leads to the consistency of selection.

In the empirical work, we employ S-LASSO to study a characteristics-based asset pricing model. In Freyberger et al. (2020), they assume assets excess returns can be predicted by security-relevant characteristics and their interaction with the firm's size:

$$\boldsymbol{Y}_t = \boldsymbol{\theta_t} + \overbrace{\sum_{q\neq s}^{Q} m_{qs}(\boldsymbol{X}_{qt-1} \cdot \boldsymbol{X}_{st-1})}^{\text{interaction with firm size Xs}} + \overbrace{\sum_{q=1}^{Q} m_q(\boldsymbol{X}_{qt-1})}^{\text{uni-variate}},$$

where $\boldsymbol{Y}_t$ is a $n \times 1$ vector of assets excess returns at time t while $\boldsymbol{X}_{jt-1}$ is a $n \times 1$ vector of asset-specific characteristic at time $t-1$. However, they fail to consider the potential linear effects of characteristics, which

have a quicker convergence rate and less computational burden. Furthermore, they analyse interactive effects by specifying the form of pairwise interaction as $\boldsymbol{X}_{qt-1} \cdot \boldsymbol{X}_{st-1}$ (elementwise product), which is quite restrictive since $m_{qs}(\boldsymbol{X}_{qt-1} \cdot \boldsymbol{X}_{st-1}) \neq m_{qs}(\boldsymbol{X}_{qt-1}, \boldsymbol{X}_{st-1})$ generally. S-LASSO can overcome this limitation by considering the linear effect and not restricting the form of interactions. We will illustrate these through both simulation and empirical studies.

The rest of the paper is organized as follows. Section 2 presents the model that S-LASSO is working on; Section 3 provides procedures for S-LASSO to work; Section 3 illustrates the theoretical results; Section 4 gives simulated experiments; Section 5 demonstrates an empirical study; Section 6 concludes the paper. All proofs and other materials are arranged in the Appendix.

2 Model Setup

Suppose we observe a sample data $(\boldsymbol{Y}, \boldsymbol{P})$, where $\boldsymbol{Y}$ presents the $n \times 1$ vector of dependent variables while $\boldsymbol{P}$ denotes the $n \times P$ matrix of potential covariates $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_P)$, allowing for $P > n$.

We assume there is an $n \times Q$ matrix $\boldsymbol{Q} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_Q)$ that is relevant to explain or predict the variation of $\boldsymbol{Y}$ and $\boldsymbol{Q} \subset \boldsymbol{P}$. We restrict that Q is fixed, whereas P is diverging as sample size $n \to \infty$. We propose a sparse structure by assuming Q is relatively small as:

$$\boldsymbol{Y} = \boldsymbol{\theta} + h(\boldsymbol{Q}) + \boldsymbol{U},$$

$$E(\boldsymbol{Y}|\boldsymbol{P}) = \boldsymbol{\theta} + h(\boldsymbol{Q}), \tag{6}$$

where $\boldsymbol{U}$ is an $n \times 1$ vector of idiosyncratic errors ϵ_i with $E(\boldsymbol{U}|\boldsymbol{P}) = \boldsymbol{0}$; $h(\boldsymbol{Q})$ is a multi-variate unknown function.

We also specify a partially linear additive semi-parametric model with interactive terms on $h(\boldsymbol{Q})$ as:

$$E(\boldsymbol{Y}|\boldsymbol{P}) = \boldsymbol{\theta} + h(\boldsymbol{Q}) = \boldsymbol{\theta} + \overbrace{\sum_{1 \leqslant s < s' \leqslant S}^{S} m_{ss'}(\boldsymbol{X}_s, \boldsymbol{X}_{s'})}^{interactive} + \overbrace{\sum_{q=1}^{Q} m_q(\boldsymbol{X_q})}^{uni-variate} \tag{7}$$

$$= \boldsymbol{\theta} + \overbrace{\sum_{1 \leqslant s < s' \leqslant S}^{S} m_{ss'}(\boldsymbol{X}_s, \boldsymbol{X}_{s'})}^{interactive} + \overbrace{\sum_{r=1}^{R} m_r(\boldsymbol{X_r})}^{nonlinear} + \overbrace{\sum_{l=1}^{L} \beta_l \boldsymbol{X}_l}^{linear}, \tag{8}$$

where $\boldsymbol{X}_j$ denotes the vector of the j^{th} covariate. L, R and S are cardinal numbers of three sets corresponding to linear effects variables, non-linear effects variables and interactive variables, respectively, which will be estimated later. The complement of $\boldsymbol{X}$ that does not appear in Equation 7 are regarded as irrelevant variables, which should be smoothed out.

Here we have Q relevant variables in total and S of them have interactive effects with $S \leqslant Q$. Similarly, R of them have uni-variate effects with $R \leqslant Q$. Finally, L out of Q covariates have linear effects, namely, $R + L \leqslant Q$, which means we may have some covariates having only interactive effects with others. s and s' ($s < s'$) is the s^{th} pair of relevant covariates that has interaction.

Meanwhile, $m_{ss'}(\boldsymbol{X}_s, \boldsymbol{X}_{s'})$ is an unknown bivariate nonparametric function of the s^{th} pair of relevant variables; $m_r(\boldsymbol{X}_r)$ is an uni-variate unknown function of the r^{th} relevant variable; β_l is the coefficient of the l^{th} relevant variable.

Furthermore, we define variable sets as follows:

$$\mathcal{L} = \{\boldsymbol{X}_l \in \boldsymbol{Q} : \boldsymbol{X}_l \text{ has linear effects on } \boldsymbol{Y}\},$$

$$\mathcal{R} = \{\boldsymbol{X}_r \in \boldsymbol{Q} : \boldsymbol{X}_r \text{ has nonlinear effects on } \boldsymbol{Y}\}$$

$$\mathcal{S} = \{\boldsymbol{X}_s, \boldsymbol{X}_{s'} \in \boldsymbol{Q} : \boldsymbol{X}_s, \boldsymbol{X}_{s'} \text{ have interactive effects on } \boldsymbol{Y}\}.$$

The cardinality for each set are: $|\mathcal{L}| = L$, $|\mathcal{R}| = L$ and $|\mathcal{S}| = S$. Each set above is unknown to researchers and can be empty.

Equation 7 avoids the curse of dimensionality with fewer restrictions. Compared with conventional additive models where components are uni-variate, we allow potential covariates to interact with each other to provide more information and flexibility. We also allow for a linear part since it has a better convergence rate and less computational burden. Therefore, practitioners do not bother employing nonparametric techniques when simpler parametric methods work. The decomposition in Equation 7 gives considerable adaptability to mitigate possible model misspecification. We do not include higher-order interactions among covariates, but our methods can be extended accordingly.

Based on the model above, our methodology focuses on:

1. Selecting the relevant variables subset $\boldsymbol{Q}$ from $\boldsymbol{P}$;
2. Specifying the form of decomposition in Equation 7;
3. Giving initial estimates of Equation 7.

3 Methodology

This section provides the detailed procedures to select relevant variables, decompose and estimate of $h(\boldsymbol{X})$.

3.1 Variables and Model Selection by Specification-LASSO

Without external knowledge and other information, it is hard for us to determine relevant variables and the form of Equation 7. Therefore, all forms of entire covariates and their interactive effects should be considered, and then, a proper variable selection model can be applied to filter all possibilities. After analyzing selection results, one can examine whether the function form of each covariate is linear or not, and whether some of them have interactive effects.

To develop our methods and theoretical results, we introduce some notations and definitions. First, we illustrate spline spaces.

Similar to Schumaker (1981) and Huang et al. (2010), we suppose that the j^{th} potential covariates $\boldsymbol{X}_j$, where $\boldsymbol{X}_j$ is a $n \times 1$ vector taking values in $[a, b]$ as:

$$\boldsymbol{X}_j = (X_{1j}, X_{2j}, \ldots, X_{nj})^\intercal, \quad \boldsymbol{X}_j \in \boldsymbol{P} \text{ and } j = 1, 2, \ldots, P.$$

Furthermore, a, b are finite with $a < b$. Let $\boldsymbol{K} = \{a = \underbrace{\kappa_0 = \kappa_0 = \ldots = \kappa_0}_{g} < \kappa_1 < \kappa_2 < \cdots < \kappa_{k_n} < \underbrace{\kappa = \kappa = \ldots = \kappa}_{g} = b\}$ be a sequence of knots partitioning the interval $[a, b]$ into subintervals, where $k_n = [n^v]$ with $0 < v < 0.5$ being a positive integer whereas g is the order of bases used. Let $K_n = k_n + g$, which denotes the total number of bases. For the i^{th} individual of $\boldsymbol{X}_j$, where $j = 1, 2, \ldots, P$ and $i = 1, 2, \ldots, n$, a set of B-splines can be built in the L^2 space $\Omega_n[\boldsymbol{K}]$ as $\Phi_{\boldsymbol{K}}(\boldsymbol{X}_j) = \{\phi_1(\boldsymbol{X}_j), \phi_2(\boldsymbol{X}_j), \ldots, \phi_{K_n}(\boldsymbol{X}_j)\}$. Next, we define a B-splines matrix:

$$\Phi_{\boldsymbol{K}}(\boldsymbol{X}_j) = \begin{Bmatrix} \phi_1(X_{j1}) & \phi_2(X_{j1}) & \ldots & \phi_{K_n}(X_{j1}) \\ \phi_1(X_{j2}) & \phi_2(X_{j2}) & \ldots & \phi_{K_n}(X_{j2}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(X_{jn}) & \phi_2(X_{jn}) & \ldots & \phi_{K_n}(X_{jn}) \end{Bmatrix},$$

Definition 3.1. Define spline space $\mathcal{K}_{g,\boldsymbol{K}}$ as linear combination of B-splines by:

$$\mathcal{K}_{g,\boldsymbol{K}} = span\{\phi_{\boldsymbol{K},k}, 1 \leqslant k \leqslant K_n\} = \{\sum_{k=1}^{K_n} \beta_k \phi_{\boldsymbol{K},k} | \beta_k \in \mathbb{R} \text{ for } 1 \leqslant k \leqslant K_n\},$$

where g is the degree of those bases and $\boldsymbol{K}$ is the knots sequence, and β_k is the k^{th} B-spline coefficient. To simplify the notation without causing confusion, we drop the sequence subscript $\boldsymbol{K}$ henceforth.

Accordingly, the r^{th} unknown uni-variate function can be approximated as:

$$m_r(\boldsymbol{X}_r) = \Phi(\boldsymbol{X}_r)\boldsymbol{\beta}_r + \boldsymbol{\xi}_r,$$

where $\boldsymbol{\beta}_r = (\beta_{r1}, \beta_{r2}, \ldots, \beta_{rK_n})^\intercal$, and $\boldsymbol{\xi}_r$ is the approximation error.

Similar to spline space $\mathcal{K}_{g,\boldsymbol{K}}$, we construct another spline space $\mathcal{D}_{g,\boldsymbol{D}}$ using knot sequence $\boldsymbol{D}$ in interval $[a', b']$.

Definition 3.2. Define the **tensor product** of spline spaces $\mathcal{K}_{g,\boldsymbol{K}} \bigotimes \mathcal{D}_{g,\boldsymbol{D}}$ as the famlily of all functions of the form:

$$f(\boldsymbol{x}_p, \boldsymbol{x}_{p'}) = \sum_{k=1}^{K_n} \sum_{d=1}^{D_n} \beta_{kd} \phi_k(\boldsymbol{x}_p) \mu_d(\boldsymbol{x}_{p'}), \text{ where } 1 < 2 < \cdots < p < p' < \cdots < P$$

where coefficients β_{kd} can be any real numbers.

Accordingly, for any covariates $\boldsymbol{X}_a, \boldsymbol{X}_b \in \boldsymbol{P}$, their potential interactive effects can be approximated as:

$$m_{ab}(\boldsymbol{X}_a, \boldsymbol{X}_b) = \sum_{k=1}^{K_n} \sum_{d=1}^{D_n} \beta_{abkd} \phi_k(\boldsymbol{X}_a) \mu_d(\boldsymbol{X}_b) + \boldsymbol{\xi}_{ab}, \quad 1 \leqslant a < b \leqslant P,$$

where $\boldsymbol{\xi}_{ab}$ is the approximation error.

Equivalently, let

$$\Phi_{\boldsymbol{K}}(X_{ia}) = (\phi_1(X_{ia}), \phi_2(X_{ia}), \ldots, \phi_{K_n}(X_{ia}))^\intercal,$$

$$\mu_{\boldsymbol{D}}(X_{ib}) = (\mu_1(X_{ib}), \mu_2(X_{ib}), \ldots, \mu_{D_n}(X_{ib}))^\intercal.$$

Equivalently:

$$\Phi_{\boldsymbol{K}}(X_{ia}) \otimes \mu_{\boldsymbol{D}}(X_{ib}) = Vec \left(\begin{bmatrix} \phi_1(X_{ia})\mu_1(X_{ib}) & \phi_1(X_{ia})\mu_2(X_{ib}) & \ldots & \phi_1(X_{ia})\mu_{D_n}(X_{ib}) \\ \phi_2(X_{ia})\mu_1(X_{ib}) & \phi_2(X_{ia})\mu_2(X_{ib}) & \ldots & \phi_2(X_{ia})\mu_{D_n}(X_{ib}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{K_n}(X_{ia})\mu_1(X_{ib}) & \phi_{K_n}(X_{ia})\mu_2(X_{ib}) & \ldots & \phi_{K_n}(X_{ia})\mu_{D_n}(X_{ib}) \end{bmatrix} \right)^\intercal$$

$$= (\phi_1(X_{ia})\mu_1(X_{ib}), \phi_1(X_{ia})\mu_2(X_{ib}), \ldots, \phi_1(X_{ia})\mu_{D_n}(X_{ib}), \ldots, \phi_{K_n}(X_{ia})\mu_{D_n}(X_{ib})).$$

Then:

$$\boldsymbol{\Phi}_{\boldsymbol{K}}(\boldsymbol{X}_a) \otimes \boldsymbol{\mu}_{\boldsymbol{D}}(\boldsymbol{X}_b) = \begin{bmatrix} \Phi_{\boldsymbol{K}}(X_{1a}) \otimes \mu_{\boldsymbol{D}}(X_{1b}) \\ \Phi_{\boldsymbol{K}}(X_{2a}) \otimes \mu_{\boldsymbol{D}}(X_{2b}) \\ \vdots \\ \Phi_{\boldsymbol{K}}(X_{na}) \otimes \mu_{\boldsymbol{D}}(X_{nb}) \end{bmatrix}.$$

To simplify the notation without causing any confusion, we drop the sequence subscript $\boldsymbol{K}$ and $\boldsymbol{D}$ henceforth.

We also write tensor product coefficients as vector $\boldsymbol{\beta}_{ab}$ as:

$$\boldsymbol{\beta}_{ab} = (\beta_{ab11}, \beta_{ab12}, \ldots, \beta_{ab1D_n}, \ldots, \beta_{abK_n1}, \beta_{abK_n2}, \ldots, \beta_{abK_nD_n})^\intercal$$

The true model can be approximated as:

$$\boldsymbol{Y} = \boldsymbol{\theta} + \sum_{\boldsymbol{X}_l \in \mathcal{L}} \beta_l \boldsymbol{X}_l + \sum_{\boldsymbol{X}_r \in \mathcal{R}} \boldsymbol{\Phi}(X_r)\boldsymbol{\beta}_r + \sum_{\boldsymbol{X}_s, \boldsymbol{X}_{s'} \in \mathcal{S}} \boldsymbol{\Phi}(\boldsymbol{X}_s) \otimes \boldsymbol{\mu}(\boldsymbol{X}_b)\boldsymbol{\beta}_{ss'} + \boldsymbol{\Xi}_n + \boldsymbol{U},$$

where $\boldsymbol{\Xi}_n$ is the approximation error and $\boldsymbol{U}$ is the $n \times 1$ vector of idiosyncratic error ϵ_i .

Those non-zero coefficients are:

$$\boldsymbol{\beta}_{\mathcal{L}} = (\beta_1, \ldots, \beta_L)^\intercal,$$

$$\boldsymbol{\beta}_{\mathcal{R}} = (\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_R^\intercal)^\intercal,$$

$$\boldsymbol{\beta}_{\mathcal{S}} = (\boldsymbol{\beta}_{11'}^\intercal, \ldots, \boldsymbol{\beta}_{SS'}^\intercal)^\intercal.$$

We define a non-zero coefficient vector:

$$\boldsymbol{\beta}_{P_1} = (\boldsymbol{\beta}_{\mathcal{L}}^\intercal, \boldsymbol{\beta}_{\mathcal{R}}^\intercal, \boldsymbol{\beta}_{\mathcal{S}}^\intercal)^\intercal.$$

Let $dim(\boldsymbol{\beta}_{P_1}) = P_1$, where $dim(\cdot)$ means the dimension of any vector. We also define B-spline bases of relevant covariates as:

$$\boldsymbol{X}_{\mathcal{L}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_l, \ldots, \boldsymbol{X}_L), \quad \boldsymbol{X}_l \in \mathcal{L}.$$

$$\boldsymbol{N}(\boldsymbol{X}_{\mathcal{R}}) = (\boldsymbol{\Phi}(\boldsymbol{X}_1), \ldots, \boldsymbol{\Phi}(\boldsymbol{X}_r), \ldots, \boldsymbol{\Phi}(\boldsymbol{X}_R)), \quad \boldsymbol{X}_r \in \mathcal{R}.$$

$$\boldsymbol{I}(\boldsymbol{X}_{\mathcal{S}}) = (\boldsymbol{\Phi}(\boldsymbol{X}_1) \otimes \boldsymbol{\mu}(\boldsymbol{X}_{1'}), \ldots, \boldsymbol{\Phi}(\boldsymbol{X}_S) \otimes \boldsymbol{\mu}(\boldsymbol{X}_{S'})), \quad \boldsymbol{X}_s \text{ and } \boldsymbol{X}_{s'} \in \mathcal{S}.$$

Recall that for individual i, we observe P potential covariates denoted as a vector $\boldsymbol{P}_i$, and there are Q relevant variables denoted as $\boldsymbol{Q}_i$, $Q \leqslant P$. There are two steps for the S-LASSO to work to select $\boldsymbol{Q}_i$ out of $\boldsymbol{P}_i$ to specify the model as in Equation 7.

In the next step, our job is to put all possible linear, nonlinear and interactive forms of all potential covariates in a selection model. S-LASSO can achieve at least three goals, namely, to select all the relevant variables, to specify the model and to obtain the preliminary estimates.

Step 1. Substitute all possible forms of each variable and pairwise interactive terms in $\boldsymbol{P}$ into LASSO selection:

$$\min_{\boldsymbol{\beta}_l, \boldsymbol{\beta}_r, \boldsymbol{\beta}_{ab}} ||\boldsymbol{Y} - \boldsymbol{\theta} - \sum_{l=1}^{P} \beta_l \boldsymbol{X}_l - \sum_{r=1}^{P} \boldsymbol{\Phi}(\boldsymbol{X}_r)\boldsymbol{\beta}_r - \sum_{a=1}^{P-1} \sum_{b>a}^{P} \boldsymbol{\Phi}(\boldsymbol{X}_a) \otimes \boldsymbol{\mu}(\boldsymbol{X}_b)\boldsymbol{\beta}_{ab}||_2^2$$
$$+\lambda_n \left(\sum_{l=1}^{P} |\beta_l| + \sum_{r=1}^{P} |\boldsymbol{\beta}_r| + \sum_{a=1}^{P-1} \sum_{b>a}^{P} |\boldsymbol{\beta}_{ab}| \right)$$

where $|\beta|$ and $|\boldsymbol{\beta}_n|$ are l_1 norms and $||\boldsymbol{\beta}||_2 \equiv (\sum_{n=1}^{N} |\beta_n|^2)^{1/2}$ denotes the l_2 norm of any $n \times 1$ vector $\boldsymbol{\beta}$. $\lambda_n > 0$ is a data driven tuning parameter. This step provides us with preliminary information after the initial selection. However, one drawback of LASSO process is that it may leave plenty of small but non-zero coefficients. Nonetheless, the first step provide crucial hints which are helpful for discriminatory penalty in the next step.

Step 2. Use step 1 estimates to construct penalty weighting coefficients and substitute all bases into adaptive group LASSO:

$$\hat{\omega}_l = \begin{cases} \sqrt{N_{\mathcal{L}}} |\tilde{\beta}_l|^{-1}, & \text{if } |\tilde{\beta}_l| > 0 \\ \infty, & \text{if } \tilde{\beta}_l = 0. \end{cases}$$

$$\hat{\omega}_r = \begin{cases} \sqrt{N_{\mathcal{R}}}||\tilde{\boldsymbol{\beta}}_r||_2^{-1}, & \text{if } ||\tilde{\boldsymbol{\beta}}_r||_2 > 0 \\ \infty, & \text{if } ||\tilde{\boldsymbol{\beta}}_r||_2 = 0. \end{cases}$$

$$\hat{\omega}_{ab} = \begin{cases} \sqrt{N_{\mathcal{S}}}||\tilde{\boldsymbol{\beta}}_{ab}||_2^{-1}, & \text{if } ||\tilde{\boldsymbol{\beta}}_{ab}||_2 > 0 \\ \infty, & \text{if } ||\tilde{\boldsymbol{\beta}}_{ab}||_2 = 0. \end{cases}$$

$N_{\mathcal{L}} = L$, $N_{\mathcal{R}} = R \times K_n$ and $N_{\mathcal{S}} = \frac{S(S-1)}{2} \times (K_n)^2$ are the number of coefficients within each group as our group sizes are significantly different. We use group cardinality to control the strength of the penalty.

To eliminate the noise from step 1, we consider the adaptive group LASSO which can select variables in a group manner.

$$L(\theta, \beta_l, \boldsymbol{\beta}_r, \boldsymbol{\beta}_{ab}) = ||\boldsymbol{Y} - \boldsymbol{\theta} - \sum_{l=1}^{P} \beta_l \boldsymbol{X}_l - \sum_{r=1}^{P} \boldsymbol{\Phi}(\boldsymbol{X}_r)\boldsymbol{\beta}_r - \sum_{a=1}^{P-1}\sum_{b>a}^{P} \boldsymbol{\Phi}(\boldsymbol{X}_a) \otimes \boldsymbol{\mu}(\boldsymbol{X}_b)\boldsymbol{\beta_{ab}}||_2$$
$$+\tilde{\lambda}_n \left(\sum_{l=1}^{P} \hat{\omega}_l|\beta_l| + \sum_{r=1}^{P} \hat{\omega}_r||\boldsymbol{\beta}_r||_2 + \sum_{a=1}^{P-1}\sum_{b>a}^{P} \hat{\omega}_{ab}||\boldsymbol{\beta}_{ab}||_2 \right),$$

Let $0 \times \infty = 0$, so groups deleted by LASSO are not selected by adaptive group LASSO for sure. $\tilde{\lambda}_n > 0$ is a data driven tuning parameter.

After the selection by step 2, all non-zero coefficients of linear approximation are represented as $\hat{\boldsymbol{\beta}}_{\mathcal{L}}$; non-zero coefficients of the approximate of nonlinear effects are shown as $\hat{\boldsymbol{\beta}}_{\mathcal{R}}$; non-zero coefficients of tensor products are written as $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$. At the same time, all the irrelevant variables or bases are smoothed out since their coefficients are zeros. Additionally, the non-zero βs of Step 2 is a vector $\hat{\boldsymbol{\beta}}_{P_1}$,

$$\hat{\boldsymbol{\beta}}_{P_1} = (\hat{\boldsymbol{\beta}}_{\mathcal{L}}^{\intercal}, \hat{\boldsymbol{\beta}}_{\mathcal{R}}^{\intercal}, \hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\intercal})^{\intercal},$$

where $\hat{\boldsymbol{\beta}}_{\mathcal{L}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{\hat{L}})^{\intercal}$, $\hat{\boldsymbol{\beta}}_{\mathcal{R}} = (\hat{\boldsymbol{\beta}}_1^{\intercal}, \ldots, \hat{\boldsymbol{\beta}}_{\hat{R}}^{\intercal})^{\intercal}$, and $\hat{\boldsymbol{\beta}}_{\mathcal{P}} = (\hat{\boldsymbol{\beta}}_{11'}^{\intercal}, \ldots, \hat{\boldsymbol{\beta}}_{\hat{S}\hat{S}'}^{\intercal})^{\intercal}$.

The model specification we obtained is:

$$h(\boldsymbol{Q}) = \sum_{\boldsymbol{X}_l \in \hat{\mathcal{L}}} \beta_l \boldsymbol{X}_l + \sum_{\boldsymbol{X}_r \in \hat{\mathcal{R}}} m_r(\boldsymbol{X}_r) + \sum_{\boldsymbol{X}_s, \boldsymbol{X}_{s'} \in \hat{\mathcal{S}}} m_{ss'}(\boldsymbol{X}_s, \boldsymbol{X}_{s'}),$$

where,

$$\hat{\mathcal{L}} = \{\boldsymbol{X}_l \in \boldsymbol{Q} : |\hat{\beta}_l| > 0\},$$

$$\hat{\mathcal{R}} = \{\boldsymbol{X}_r \in \boldsymbol{Q} : ||\hat{\boldsymbol{\beta}}_r||_2 > 0\},$$

$$\hat{\mathcal{S}} = \{\boldsymbol{X}_s, \boldsymbol{X}_{s'} \in \boldsymbol{Q} : ||\hat{\boldsymbol{\beta}}_{ss'}||_2 > 0\},$$

Accordingly, $|\hat{\mathcal{L}}| = \hat{L}$, $|\hat{\mathcal{R}}| = \hat{R}$ and $|\hat{\mathcal{S}}| = \hat{S}$. In practice, we include covariates that are selected by both linear and nonlinear parts in the nonlinear set only since this can simplify the model further. The classification above is for theoretical proof purposes.

Next, nonlinear and interactive components are approximated by:

$$\hat{m}_r(\boldsymbol{X}_r) = \boldsymbol{\Phi}(\boldsymbol{X}_r)\hat{\boldsymbol{\beta}}_r, 1 \leqslant r \leqslant \hat{R}$$

$$\hat{m}_{ss'}(\boldsymbol{X}_s, \boldsymbol{X}_{s'}) = \boldsymbol{\Phi}(\boldsymbol{X}_s) \otimes \boldsymbol{\mu}(\boldsymbol{X}_{s'})\hat{\boldsymbol{\beta}}_{ss'}, \quad 1 \leqslant s < s' \leqslant \hat{S}.$$

Meanwhile, we define the matrix of irrelevant components, which are smoothed out by S-LASSO as:

$$\boldsymbol{X}_{\mathcal{L}^C} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_l, \ldots, \boldsymbol{X}_{L^C}), \quad \boldsymbol{X}_l \in \boldsymbol{P} \text{ but } \boldsymbol{X}_l \notin \mathcal{L}.$$

$$\boldsymbol{N}(\boldsymbol{X}_{\mathcal{R}^C}) = (\boldsymbol{\Phi}(\boldsymbol{X}_1), \ldots, \boldsymbol{\Phi}(\boldsymbol{X}_r), \ldots, \boldsymbol{\Phi}(\boldsymbol{X}_{R^C})), \quad \boldsymbol{X}_r \in \boldsymbol{P} \text{ but } \boldsymbol{X}_r \notin \mathcal{R}.$$

$$\boldsymbol{I}(\boldsymbol{X}_{\boldsymbol{\mathcal{S}^C}}) = (\boldsymbol{\Phi}(\boldsymbol{X}_1) \otimes \boldsymbol{\mu}(\boldsymbol{X}_{1'}), \ldots, \boldsymbol{\Phi}(\boldsymbol{X}_{S^C}) \otimes \boldsymbol{\mu}(\boldsymbol{X}_{S^{C'}})), \quad \boldsymbol{X}_s \text{ and } \boldsymbol{X}_{s'} \in \boldsymbol{P} \text{ but } \boldsymbol{X}_s \text{ and } \boldsymbol{X}_{s'} \notin \mathcal{S}.$$

Let $n \times P_1$ matrix $\boldsymbol{Z}_1 = (\boldsymbol{X}_{\mathcal{L}}, \boldsymbol{N}(\boldsymbol{X}_{\mathcal{R}}), \boldsymbol{I}(\boldsymbol{X}_{\mathcal{S}}))$ represent all the relevant components and let $\boldsymbol{\beta}_{P_1}$ be the $P_1 \times 1$ coefficient vector of matrix $\boldsymbol{Z}_1$. Meanwhile, let $n \times P_2$ matrix $\boldsymbol{Z}_2 = (\boldsymbol{X}_{\mathcal{L}^C}, \boldsymbol{N}(\boldsymbol{X}_{\mathcal{R}^C}), \boldsymbol{I}(\boldsymbol{X}_{\mathcal{S}^C}))$, denotes all the irrelevant components. Similarly, let $\boldsymbol{\beta}_{P_2} = (\boldsymbol{\beta}_{\mathcal{L}^C}^\intercal, \boldsymbol{\beta}_{\mathcal{R}^C}^\intercal, \boldsymbol{\beta}_{\mathcal{S}^C}^\intercal)^\intercal$ be the $P_2 \times 1$ coefficient vector of matrix $\boldsymbol{Z_2}$.

3.2 Estimation

OLS can be applied to obtain estimates:

$$\hat{\boldsymbol{\beta}}_{P_1} = (\boldsymbol{Z_1}^\intercal \boldsymbol{Z_1})^{-1} \boldsymbol{Z_1}^\intercal \boldsymbol{Y}.$$

And

$$\hat{\boldsymbol{\beta}}_{P_2} = \boldsymbol{0},$$

$$\hat{\boldsymbol{\beta}}_{P_Z} = (\hat{\boldsymbol{\beta}}_{P_1}^\intercal, \hat{\boldsymbol{\beta}}_{P_2}^\intercal)^\intercal.$$

4 Theoretical results

Firstly, we list some assumptions to facilitate our theoretical analysis.

4.1 Assumption

Assumption 1. *The noise ϵ_i are independent and identically distributed with $E\epsilon_i = 0$ and $Var(\epsilon) = \sigma^2$. Furthermore, it has finite $2k^{th}$ moment with $E(\epsilon_i^{2k}) < \infty$ for $k = 1, 2, \ldots, K$.*

Assumption 2. *Let*

$$\boldsymbol{V} = \frac{1}{n}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^\intercal (\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \begin{Bmatrix} \boldsymbol{V}_{Z_1 Z_1} & \boldsymbol{V}_{Z_1 Z_2} \\ \boldsymbol{V}_{Z_2 Z_1} & \boldsymbol{V}_{Z_2 Z_2} \end{Bmatrix}$$

be the covariance matrix of all the components in step 1. There exist constants c_1, c_2, c_3, *and* c_4 *with* $0 \leqslant c_1 < c_2 \leqslant 1$ *and* $c_3, c_4 > 0$, *such that*

$$P_1 = O(n^{c_1}), \tag{9}$$

$$n^{\frac{1-c_2}{2}} min\{|\beta_l|, ||\boldsymbol{\beta}_r||_2, ||\boldsymbol{\beta}_{ss'}||_2\} \geqslant c_4, \text{ for } \beta_l, \boldsymbol{\beta}_r, \boldsymbol{\beta}_{ss'} \in \boldsymbol{\beta}_{P_1}. \tag{10}$$

$$P_2 = O(n^{(c_2-c_1)k}), \tag{11}$$

$$\lambda_{min}(\boldsymbol{V}_{Z_1 Z_1}) > c_3, \tag{12}$$

Equation 9 and Equation 11 control the maximum dimensions of relevant and irrelevant components respectively. Equation 12 ensures that the minimum eigenvalue of relevant components matrix $\boldsymbol{Z_1}$ is away from 0 to be invertible, where $\lambda_{min}(\boldsymbol{V}_{Z_1 Z_1})$ indicates the smallest eigenvalue of covariance matrix $\boldsymbol{V}_{Z_1 Z_1}$. Finally, Equation 10 limits the decay rate of elements in $\boldsymbol{\beta}_{P_1}$.

Assumption 3. $E(m_r(\boldsymbol{X}_r)) = 0$, $E(m_{ss'}(\boldsymbol{X}_s, \boldsymbol{X}_{s'})) = 0$, *given* $\boldsymbol{X}_j \in \mathcal{R} \cup \mathcal{S}$.

This assumption is for unique identification purpose.

Assumption 4. *0–th, first and second derivatives of* $m_r(\boldsymbol{X}_r)$ *and* $m_{ss'}(\boldsymbol{X}_s, \boldsymbol{X}_{s'})$ *are continuous, for* $X_r \in \mathcal{R}$ *and* $X_s, X_{s'} \in \mathcal{S}$.

This assumption is for approximation accuracy of B-splines bases and their tensor products.

Definition 4.1. Let $\hat{\beta}$ be an estimate of β. Then, $\hat{\beta}$ is **Sign Consistent** with β, shown as $\hat{\beta} =_s \beta$, if and only if

$$sign(\hat{\beta}) = sign(\beta),$$

where $sign(\hat{\beta}) = 1$, if $\hat{\beta} > 0$; $sign(\hat{\beta}) = -1$, if $\hat{\beta} < 0$; and $sign(\hat{\beta}) = 0$, if $\hat{\beta} = 0$. Similarly, Let $\hat{\boldsymbol{\beta}}$ be a vector of estimates of $\boldsymbol{\beta}$. Then $\hat{\boldsymbol{\beta}}$ is **Sign Consistent** with $\boldsymbol{\beta}$, written as $\hat{\boldsymbol{\beta}} =_s \boldsymbol{\beta}$ if and only if each entry is **Sign Consistent**.

Definition 4.2. Let $\hat{\beta}$ be an estimate of β. Then, $\hat{\beta}$ is **Norm Consistent** with β, shown as $\hat{\beta} =_0 \beta$, if and only if

$$sign_0(\hat{\beta}) = sign_0(\beta),$$

where $sign_0(\hat{\beta}) = 1$, if $\hat{\beta} \neq 0$; $sign_0(\hat{\beta}) = 0$, if $\hat{\beta} = 0$. Similarly, Let $\hat{\boldsymbol{\beta}}$ be a vector of estimates of $\boldsymbol{\beta}$. Then $\hat{\boldsymbol{\beta}}$ is **Norm Consistent** with $\boldsymbol{\beta}$, written as $\hat{\boldsymbol{\beta}} =_0 \boldsymbol{\beta}$ if and only if each entry is **Norm Consistent**.

Condition 4.1. Let covariance matrix $\boldsymbol{V}$ satisfies strong irrepresentable condition documented by Zhao and Yu (2006), stating that there exists a positive constant $P_1 \times 1$ vector $\boldsymbol{\eta}$, and

$$|\boldsymbol{V}_{Z_2Z_1}(\boldsymbol{V}_{Z_1Z_1})^{-1} sign(\boldsymbol{\beta}_{\boldsymbol{P_1}})| \leqslant \mathbf{1} - \boldsymbol{\eta},$$

which is true element-wise.

Condition 4.2. Similarly, covariance matrix $\boldsymbol{V}$ satisfies weak irrepresentable condition , if

$$|\boldsymbol{V}_{Z_2Z_1}(\boldsymbol{V}_{Z_1Z_1})^{-1} sign(\boldsymbol{\beta}_{\boldsymbol{P_1}})| < \mathbf{1},$$

which is true element-wise.

Theorem 4.1. *Under Assumptions 1-4 and Condition 4.2, and let* $P_Z = P_1 + P_2$, $\boldsymbol{\beta}_{P_Z} = (\boldsymbol{\beta}_{P_1}^\intercal, \boldsymbol{\beta}_{P_2}^\intercal)^\intercal$, *for* $\forall \lambda_n$ *satisfying* $\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2-c_1}{2}})$ *and* $\frac{1}{P_Z}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$ *for* $k = 1, 2, 3 \ldots$, *then the first step of S-LASSO is sign consistent with:*

$$P(\hat{\boldsymbol{\beta}}_{\boldsymbol{P_Z}} =_s \boldsymbol{\beta}_{\boldsymbol{P_Z}}) \geqslant 1 - O(\frac{P_Z n^k}{\lambda_n^{2k}}) \to 1, \text{ as } n \to \infty.$$

Theorem 4.2. *Given the well-chosen number of internal knots* $k_n = [n^v]$ *and under Assumptions 1-4 and Theorem 4.1, S-LASSO is consistent on selection relevant covariates and specification of the correct model:*

$$P(\lim_{n\to\infty} \hat{\boldsymbol{\beta}}_{\mathcal{L}} =_0 \boldsymbol{\beta}_{\mathcal{L}}) \to 0,$$

$$P(\lim_{n\to\infty} \hat{\boldsymbol{\beta}}_{\mathcal{R}} =_0 \boldsymbol{\beta}_{\mathcal{R}}) \to 0,$$

$$P(\lim_{n\to\infty} \hat{\boldsymbol{\beta}}_{\mathcal{S}} =_0 \boldsymbol{\beta}_{\mathcal{S}}) \to 0.$$

5 Simulation study

We generate our model as:

$$y_i = \beta x_{1i} + m_1(x_{2i}) + m_2(x_{3i}, x_{4i}) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\beta x_{1i} = x_{1i}$, $m_1(x_{2i}) = x_{2i}^2$, $m_2(x_{3i}, x_{4i}) = \sin(x_{3i} + x_{4i})$. All three above functions are rescaled to be zero mean and unit variance. Furthermore, we generate P candidate variables x_{pi}. We have all independent variables generated from $Uniform[-2, 2]$ and $\epsilon \sim N(0, \sigma^2)$, where there are no correlations between all potential variables. Two different $\mathbf{P}$ dimensions and three different sample sizes are tested, namely, $P = 30, 50$ with $n = 100, 300, 500$.

In Table 1, we compare the results of S-LASSO and the methods of selecting interactive effects between stock characteristics in Freyberger et al. (2020) (named FNW).

We choose four evenly distributed knots to construct B-splines approximation of nonlinear effects while choosing two evenly distributed knots for each covariate to construct tensor products to keep the group size comparable. Meanwhile, for the FNW methods, we choose all the knots sequences for x_j and $x_j \times x_{j'}$ to be 4, which are also evenly distributed, to approximate both nonlinear and interaction effects among potential covariates. The tuning parameter λ_ns are chosen through BIC for both steps. Here, we define BIC as:

$$BIC = n * \log(MSE) + df * log(n),$$

where n is the number of observation and df represents the degree of freedom in LASSO procedures discussed in Leng et al. (2006).

Furthermore, we define the signal to noise ratio as as $R_\sigma = sd(m(\cdot))/sd(\epsilon)$ to illustrate the robustness of S-LASSO under different noise level.

Table 1: Simulation Example of S-LASSO

			σ=0.25			σ=0.333			σ=0.5		
			INC	CS	MSE	INC	CS	MSE	INC	CS	MSE
P=30	n=100	S-LASSO	52.8	49.4	0.77	48.2	44.2	0.83	32.8	28.4	0.94
			(0.5)	(0.5)	(0.28)	(0.5)	(0.5)	(0.29)	(0.47)	(0.45)	(0.32)
		FNW	0	0	1.13	0	0	1.2	0.2	0.2	1.42
			(0)	(0)	(0.4)	(0)	(0)	(0.41)	(0.04)	(0.04)	(0.49)
	n=300	S-LASSO	95.6	95	0.59	94.8	94	0.66	95	95	0.81
			(0.21)	(0.22)	(0.17)	(0.22)	(0.24)	(0.2)	(0.218)	(0.218)	(0.2)
		FNW	0	0	0.97	0	0	1.02	0	0	1.16
			(0)	(0)	(0.18)	(0)	(0)	(0.17)	(0)	(0)	(0.18)
	n=500	S-LASSO	99.8	99.6	0.53	98.2	97.2	0.59	98.4	98.4	0.75
			(0.04)	(0.06)	(0.08)	(0.133)	(0.165)	(0.11)	(0.126)	(0.126)	(0.138)
		FNW	0	0	0.97	0	0	0.96	0	0	1.15
			(0)	(0)	(0.14)	(0)	(0)	(0.14)	(0)	(0)	(0.15)
P=50	n=100	S-LASSO	34.8	31.6	0.88	34.8	32	0.89	24.8	22	1.01
			(0.48)	(0.47)	(0.31)	(0.48)	(0.47)	(0.32)	(0.43)	(0.41)	0.35
		FNW	0	0	1.24	0	0	1.31	0	0	1.5
			(0)	(0)	(0.48)	(0)	(0)	(0.45)	(0)	(0)	(0.41)
	n=300	S-LASSO	92.8	92.4	0.66	88	88	0.75	85.8	85.8	0.92
			(0.26)	(0.27)	(0.22)	(0.33)	(0.33)	(0.26)	(0.35)	(0.35)	(0.27)
		FNW	0	0	1.01	0	0	1.06	0	0	1.22
			(0)	(0)	(0.2)	(0)	(0)	(0.19)	(0)	(0)	(0.19)
	n=500	S-LASSO	98.6	98.4	0.57	96.8	96.8	0.64	96	96	0.81
			(0.12)	(0.13)	(0.12)	(0.18)	(0.18)	(0.17)	(0.2)	(0.2)	(0.19)
		FNW	0	0	0.98	0	0	1.02	0	0	1.18
			(0)	(0)	(0.14)	(0)	(0)	(0.16)	(0)	(0)	(0.16)

Note: This table compares the performance of S-LASSO and the method used in FNW (2020) under different sample size, n=100, 300, 500; different number of irrelevant variables, P=30, 50; and different levels of noise, $R_\sigma = 4, 3, 2$. INC represents the percentage that all the relevant covariates are correctly included in the model. CS shows the percentage of the whole model that is correctly specified, which means the model not only selects all relevant variables but also gives them a precise specification. MSE indicates the average mean squared error of all repetitions under each method. Simulations are repeated 500 times for each setting. Standard deviations are given in the parentheses.

From the results in Table 1, S-LASSO overperforms FNW under all scenarios. Because in FNW, they treat interaction term $x_j \times x_{j'}$ as a new variable and construct B-spline space based on this covariate. Therefore, only certain forms of pairwise interactions with input $x_j \times x_{j'}$ can be detected. Hence, for nearly all the simulation settings, FNW can neither include all the relevant covariates nor specify the model correctly, given the interactive function form $\sin(x_{3i} + x_{4i})$. However, S-LASSO employs tensor products of B-splines to approximate potential interactions and has decent accuracy on both including all relevant covariates and choosing the correct model. We use this simulation to show the limitation of FNW and demonstrate that tensor products can accommodate more comprehensive forms of interactions. Additionally, although prediction is not the primary goal of S-LASSO, it has much smaller MSE compared with FNW.

Furthermore, S-LASSO works better for small P large n circumstances, and the highest percentage of selecting the relevant covariates and specifying the correct model can be 99.8% and 99.6% individually. For the most challenging condition under P=50, n=100 and $\sigma = 0.5$, S-LASSO also has acceptable performance with an accuracy of 24.8% and 22% respectively.

S-LASSO is also robust under different levels of noise for all settings. As shown from different rows of Table 1, the accuracy is similar across three different noise levels.

6 Empirical Study

6.1 Introduction

In this section, we revisit the question proposed by Freyberger et al. (2020), where they try to detect the influence of firms' characteristics on stock returns non-parametrically. They specify assets returns as additive non-parametric functions of lagged corresponding assets characteristics such as book-to-market ratio, profitability, etc. Their model is:

$$E(\boldsymbol{Y}_{t+1}|\boldsymbol{W}_t) = \boldsymbol{\theta}_t + \sum_{r=1}^{R} m_r(\boldsymbol{X}_{rt}), \tag{13}$$

where $\boldsymbol{Y}_{t+1}$ is a $n \times 1$ vector of stock excess returns at time t; $\boldsymbol{W_t}$ is a $n \times W$ matrix of W asset-relevant characteristics that are observed at time t. At the right hand side of Equation 13, they select R additive non-parametric uni-variate unknown functions of characteristics that are relevant to predict stock excess returns, and $\boldsymbol{\theta_t}$ is the intercept.

To further investigate the interactive effects between assets sizes with other characteristics, they propose a model to accommodate pairwise interactions:

$$E(\boldsymbol{Y}_{t+1}|\boldsymbol{W}_t) = \boldsymbol{\theta_t} + \sum_{r=1}^{R} m_r(\boldsymbol{X}_{rt}) + \sum_{s}^{S} m_s(\boldsymbol{X}_{st} \cdot \boldsymbol{X}_{size,t}), \tag{14}$$

where they consider the unknown function form taking input as $\boldsymbol{X}_s \cdot \boldsymbol{X}_{size}$. As discussed in Introduction and exemplified in Simulation, $m_s(\boldsymbol{X}_s \cdot \boldsymbol{X}_{size}) \neq m_s(\boldsymbol{X}_s, \boldsymbol{X}_{size})$, this specification of interactions may restrict the form of interactive effects to be multiplicity only. Furthermore, they do not include linear parts, which have both computational simplicity and quicker rate of convergence.

In this section, we apply S-LASSO to short rolling window data to revisit the effects of assets characteristics on stock returns and their interactive effects with firms sizes. We further divide uni-variate effects to be linear or nonlinear. The model is specified as:

$$E(\boldsymbol{Y}_{t+1}|\boldsymbol{W}_t) = \boldsymbol{\theta_t} + \sum_{l=1}^{L} \beta_l \boldsymbol{X}_l + \sum_{r=1}^{R} m_r(\boldsymbol{X}_{rt}) + \sum_{s=1}^{S} m_s(\boldsymbol{X}_{st}, \boldsymbol{X}_{size,t}), \tag{15}$$

where the notations are similar to Equation 14. However, we add a linear term to capture the linear effects of some characteristics, which can increase the rate of convergence and simplify the model and interpretation. Meanwhile, we relax the pairwise interaction between characteristics to a more general form. Similarly, we also assume that both slope parameters and characteristic functions are time-invariant. Therefore, for those nonlinear and interactive characteristics, each characteristic and each pair among them share a certain form of variation.

6.2 Data Description

Monthly stock returns are collected from CRSP (Center for Research in Security Prices) and security-specific characteristics date is from Compustat. In terms of stock returns, we correct all returns of delisted stock as in Hou et al. (2015). Furthermore, we subtract Fama-French's monthly risk-free returns from monthly stock returns to attain $\boldsymbol{Y}$ from July 1967 to June 2017, 600 months in total. As for security-related characteristics matrix $\boldsymbol{W}$, is constructed using the same way of Freyberger et al. (2020). After trading off the number of assets kept and characteristics' availability, we select 33 characteristics, which are documented in the Appendix. We use balance sheet data ending at fiscal year $t-1$ to predict stock excess returns from July $t-1$ to June t. Some characteristics are updated annually, so we take them unchanged during the fiscal year t. Finally, we merge stock returns and security-specific characteristics.

6.3 Variable Selection and Model Specification

We apply non-overlapping rolling window analysis in this empirical study. The purpose is to understand whether there are any time variations in Equation 15. In each rolling block, we use pooled panel data to apply S-LASSO. We omit the heterogeneity to assume that the same characteristic has an identical functional form within each rolling window.

For each characteristic, we choose the number of knots to be 6 to construct B-spline bases, which are used to approximate nonlinear effects and choose the number of knots to be 3 for tensor product bases, which are con-

structed to approximate interactive effects. Next, we substitute all the levels, B-spline bases and tensor products into the S-LASSO algorithm.

There are two steps for S-LASSO to work, and for both steps, similar to simulation studies, we choose λ_n and $\tilde{\lambda}_n$ through BIC.

We summarize selection results in Table 2, Table 3 and Table 4, respectively. Columns of these tables are rolling window time periods while each row presents selection results of each characteristic separately. We use ✓ to show that the corresponding characteristic is selected in a certain rolling block. We omit some rolling blocks due to the non-invertible characteristics matrix. Table 2 documents selection results of characteristics' linear effects on assets excess returns. We do not include characteristics that have both linear and nonlinear effects in Table 2 as the general effects of these characteristics should be concluded as nonlinear. Compared with Table 3, characteristics that only have linear effects on assets returns are uncommon. However, some characteristics experience persistent linear effects on stock returns, such as "C2A" (ratio of cash and short-term investments to total assets), "PCM" (price-to-cost margin), "r_{12_7}" (cumulative past return from 12 to 7 months). Table 2 demonstrates that most uni-variate effects from characteristics are nonlinear, and some of them are long-lasting. "LME" (total market capitalization of the previous month), "A2ME" (assets to market capitalization), "AT" (total assets), "E2P" (earnings to price) and "ROA" (return-on-assets) are selected by all rolling windows. Meanwhile, "Investment", "Q" (Tobin's Q), "ROE" (return-on-equity), "r_{2_1}" (short-term reversal 2 to 1 month) and "S2P" (sales-to-price) are frequently chosen. As for interactive effects with firms' sizes, we use "LME" (total market capitalization of the previous month) as the measure of firms' sizes. Table 4 shows the characteristics that have interactive effects with "LME". The interactive effects are not limited to be multiplicity by our method. "Free_cash" is more influential on stock returns when interacting with firms' sizes. "A2ME", "AT", "Q" and "ROA" also substantially interact with "LME".

Empirical results demonstrate the power of S-LASSO to select relevant variables and specify a flexible regression model. We show that asset-related characteristics are relevant to predict stock excess returns. Specifically, the form of each characteristic is different, which includes but is not limited to linear effects, nonlinear effects and interactions with firms' sizes. Although most uni-variate functions of characteristics are nonlinear, however, linear functions, which have both computational and convergence advantages, are still important. S-LASSO can not only specify linear parts but also select more general interactive effects with firms' sizes since it uses tensor products to approximate more complicated bi-variate functions.

6.4 Selection Results

Table 2: Summary of Linear Effects of Characteristics on Assets Excess Returns

Characteristics	65-68	68-71	71-73	73-76	76-79	79-82	85-88	88-91	91-94	94-97	97-00	03-06	06-09	09-12
LME														
A2ME														
AT														
ATO														
BEME														
C2A		✓	✓		✓						✓		✓	✓
C2D														
CTO	✓		✓	✓	✓									
Delceq														
DelGmSale														
Delshrout								✓						
E2P														
EPS				✓	✓									
Free_cash	✓		✓	✓				✓						
Investment														
IPM					✓									
Lev	✓			✓	✓									
LTurnover														
PCM	✓	✓	✓		✓	✓								
PM	✓	✓		✓										
Prof														
Q														
ROA														
ROC														
ROE														
r12_2														
r12_7			✓	✓	✓	✓			✓	✓	✓			
r6_2			✓	✓	✓	✓								
r2_1	✓	✓				✓	✓							
S2C														
S2P												✓		
Sales_g		✓												
SGA2S	✓													

This table shows selection results of characteristics that only have linear effects on predicting assets excess returns through three-year rolling windows from July 1965-June 2012. ✓ represents the characteristic is selected in the corresponding rolling window shown in the column.

Table 3: Summary of nonlinear Effects of Characteristics on Assets Excess Returns

Characteristics	65-68	68-71	71-73	73-76	76-79	79-82	85-88	88-91	91-94	94-97	97-00	03-06	06-09	09-12
LME	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A2ME	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ATO	✓				✓	✓		✓	✓	✓	✓	✓	✓	✓
BEME	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
C2A	✓													
C2D			✓					✓	✓	✓		✓		
CTO							✓	✓	✓	✓	✓	✓	✓	✓
Delceq		✓	✓			✓		✓		✓	✓	✓	✓	✓
DelGmSale	✓		✓					✓	✓	✓		✓	✓	✓
Delshrout			✓		✓		✓		✓	✓	✓	✓	✓	✓
E2P	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EPS	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
Free_cash												✓		
Investment	✓				✓		✓	✓	✓	✓		✓	✓	✓
IPM	✓	✓	✓			✓	✓	✓	✓		✓	✓	✓	✓
Lev			✓				✓	✓	✓	✓	✓	✓	✓	✓
LTurnover									✓	✓		✓	✓	✓
PCM										✓		✓	✓	✓
PM					✓		✓	✓		✓	✓	✓	✓	✓
Prof	✓				✓	✓		✓	✓	✓			✓	✓
Q		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ROA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ROC		✓			✓			✓		✓	✓	✓	✓	✓
ROE						✓	✓	✓	✓	✓	✓	✓	✓	✓
r12_2	✓						✓	✓	✓			✓	✓	✓
r12_7								✓				✓		
r6_2							✓	✓	✓	✓	✓	✓	✓	✓
r2_1			✓	✓				✓	✓	✓	✓	✓	✓	
S2C			✓						✓		✓	✓		
S2P			✓		✓		✓	✓	✓	✓	✓	✓	✓	✓
Sales_g			✓	✓				✓		✓		✓		
SGA2S	✓						✓				✓	✓		

This table shows selection results of characteristics that have nonlinear effects on predicting assets excess returns through three-year rolling windows from July 1965-June 2012. ✓ represents the characteristic is selected in the corresponding rolling window shown in the column.

Table 4: Summary of Interactive Effects of Characteristics with Size on Assets Excess Returns

Characteristics	65-68	68-71	71-73	73-76	76-79	79-82	85-88	88-91	91-94	94-97	97-00	03-06	06-09	09-12
A2ME	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ATO		✓												
BEME	✓		✓		✓		✓					✓		
C2A	✓		✓		✓		✓		✓	✓	✓	✓		
C2D														
CTO	✓		✓					✓		✓		✓	✓	
Delceq														
DelGmSale	✓													
Delshrout							✓			✓				
E2P		✓				✓		✓		✓	✓		✓	
EPS			✓	✓	✓	✓		✓		✓				
Free_cash					✓			✓	✓	✓		✓	✓	
Investment					✓	✓						✓		
IPM														
Lev					✓	✓			✓	✓	✓	✓		
LTurnover		✓							✓			✓		
PCM			✓		✓	✓			✓	✓		✓		
PM								✓						
Prof						✓								
Q	✓		✓			✓	✓	✓	✓	✓	✓	✓		
ROA	✓	✓	✓		✓				✓	✓	✓	✓	✓	
ROC	✓		✓						✓					
ROE														
r12_2												✓		
r12_7														
r6_2													✓	
r2_1									✓	✓				
S2C						✓								
S2P		✓				✓		✓		✓		✓	✓	
Sales_g														
SGA2S			✓					✓	✓			✓		

This table shows selection results of characteristics that have interactive effects with firms' sizes (LME) on predicting assets excess returns through three-year rolling windows from July 1965-June 2012. ✓ represents the characteristic is selected in the corresponding rolling window shown in the column.

7 Conclusion

We propose a more general variable selection and model specification method, called Specification LASSO (S-LASSO). S-LASSO is designed under sparsity, to specify a partially linear additive non-parametric regression model with pairwise interactions among regressors. Firstly, S-LASSO considers all possibilities through levels, B-splines bases and tensor products of all variables. Then, there are two steps for S-LASSO to work. In the first step, we apply LASSO to give preliminary selection. In the second step, an adaptive group LASSO is employed to give the final selection results in a group manner, using estimates in the first step as discriminatory group penalty. We illustrate the satisfactory accuracy of S-LASSO through simulation studies. Empirically, S-LASSO is applied to a characteristics-based asset pricing model. We show that security-specific characteristics have linear, nonlinear and interactive effects with firms' sizes on assets excess returns, which complements current literature.

Appendices

A Proofs

Let $\boldsymbol{\beta}_{P_Z} = (\boldsymbol{\beta}_{P_1}^\intercal, \boldsymbol{\beta}_{P_2}^\intercal)$, $\boldsymbol{Z} = (\boldsymbol{Z_1}, \boldsymbol{Z_2})$, β_i is the i^{th} element of $\boldsymbol{\beta}$. $\boldsymbol{\beta_j}$ is the j^{th} group of $\boldsymbol{\beta}_{P_Z}$, and $\boldsymbol{X_j}$ is the covariates matrix of $\boldsymbol{Z}$ in the second group.

In the first step, after applying KKT conditions, we obtain Lemma A.1 below.

Lemma A.1.

$$\frac{d||\boldsymbol{Y}-\boldsymbol{\theta}-\boldsymbol{Z\beta}||^2}{d\beta_i} = \lambda_n sign(\hat{\beta}_i) \quad for\ \hat{\beta}_i \neq 0,$$

$$\frac{d||\boldsymbol{Y}-\boldsymbol{\theta}-\boldsymbol{Z\beta}||^2}{d\beta_i} \leqslant \lambda_n sign(\hat{\beta}_i) \quad for\ \hat{\beta}_i = 0.$$

Lemma A.2. *Under Strong Irrepresentable Condition holds and a constant $\eta > 0$, then:*

$$P(\hat{\boldsymbol{\beta}}_{P_Z} =_s \boldsymbol{\beta}_{P_Z}) \geqslant P(E_A \cap E_B),$$

where:

$$E_A = \{\frac{1}{\sqrt{n}}|(\boldsymbol{V}_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal U}| < \sqrt{n}(|\boldsymbol{\beta}_{P_1}| - \frac{\lambda_n}{2n}|(\boldsymbol{V}_{Z_1Z_1})^{-1}sign(\boldsymbol{\beta}_{P_1})|)\}$$

$$E_B = \{\frac{1}{\sqrt{n}}|\boldsymbol{V}_{Z_2Z_1}(\boldsymbol{V}_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal U} - \boldsymbol{Z_2^\intercal U}| \leqslant \frac{\lambda_n}{2\sqrt{n}}\eta\},$$

The above equations hold for each entry.

The Lemma A.2 is borrowed from Proposition 1. of Zhao and Yu (2006). Proofs can be found in their Appendix.

Proof of Theorem 4.1 : We give some notations before the proof. Let $\boldsymbol{\tau} = \frac{1}{\sqrt{n}}(V_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal U}$, and $\boldsymbol{\upsilon} = \frac{1}{\sqrt{n}}(V_{Z_2Z_1}(V_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal U} - \boldsymbol{Z_2^\intercal U})$.

By Lemma A.2 we have:

$$1-P(E_A\cap E_B) \leqslant P(E_A^c)+P(E_B^c) \leqslant \sum_{i=1}^{P_1} P(|\tau_i| \geqslant \sqrt{n}(|\beta_{P_1 i}|-\frac{\lambda_n}{2n}(\boldsymbol{V}_{Z_1Z_1})^{-1}sign(\beta_{P_1 i}))+\sum_{i=1}^{P_2} P(|\upsilon_i| \geqslant \frac{\lambda_n}{2\sqrt{n}}\eta_i).$$

Then we have

$$\boldsymbol{F_\tau} = \frac{1}{\sqrt{n}}(\boldsymbol{V}_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal},$$

therefore,

$$\boldsymbol{F_\tau F_\tau^\intercal} = (\boldsymbol{V}_{Z_1Z_1})^{-1}.$$

Given $\lambda_{min}(\boldsymbol{V}_{Z_1Z_1}) > c_3$, then we have $\boldsymbol{V}_{Z_1Z_1}^{-1} < c_5$ for each entry. Similarly, let

$$\boldsymbol{F_\upsilon} = \frac{1}{\sqrt{n}}(\boldsymbol{V}_{Z_2Z_1}(\boldsymbol{V}_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal} - \boldsymbol{Z_2^\intercal}),$$

and

$$\boldsymbol{F_\upsilon F_\upsilon^\intercal} = \frac{1}{n}\boldsymbol{Z_2^\intercal}(\boldsymbol{I} - \boldsymbol{Z_1^\intercal V}_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal})\boldsymbol{Z_2}.$$

Since $\boldsymbol{I} - \boldsymbol{Z_1^\intercal}(\boldsymbol{V}_{Z_1Z_1})^{-1}\boldsymbol{Z_1^\intercal}$ is idempotent, which only has the eigenvalues of 1 and 0, therefore $\boldsymbol{F_\upsilon F_\upsilon^\intercal} \leqslant c_4$ for each diagonal element.

Furthermore, we have:

$$\frac{\lambda_n}{n}|(\boldsymbol{V}_{Z_1Z_1})^{-1}sign(\boldsymbol{\beta}_{P_1})| \leqslant \frac{c_5\lambda_n}{n}\|\boldsymbol{\beta}_{P_1}\|_2$$

Given $E(\epsilon_i^{2k}) < \infty$, then we have $E(\tau_i^{2k}) < \infty$ and $E(\upsilon_i^{2k}) < \infty$. Therefore, the tail probability of τ_i is bounded by:

$$P(\tau_i > T) = O(T^{-2k}),$$

furthermore, under $\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2-c_1}{2}})$,

$$\sum_{i=1}^{P_1} P(|\tau_i| \geqslant \sqrt{n}(|\beta_{P_1i}| - \frac{\lambda_n}{2n}(\boldsymbol{V}_{Z_1Z_1})^{-1}sign(\beta_{P_1i})) = P_1O(n^{-kc_2}) = o(\frac{P_Zn^k}{\lambda_n^{2k}}). \tag{16}$$

Similarly,

$$\sum_{i=1}^{P_2} P(|\upsilon_i| \geqslant \frac{\lambda_n}{2\sqrt{n}}\eta_i) = P_2O(\frac{n^k}{\lambda_n^{2k}}) = o(P_Z\frac{n^k}{\lambda_n^{2k}}). \tag{17}$$

Then, combining Equation 16 and Equation 17 gives Theorem 4.1. □

After grouping all the coefficients from step 1, we use $\boldsymbol{\beta_j}$ to represent the j^{th} group of $\boldsymbol{\beta}_{P_Z}$.

we apply the KKT condtions again to obtain the Lemma A.3

Lemma A.3.

$$\frac{d||\boldsymbol{Y} - \boldsymbol{\theta} - \boldsymbol{Z\beta}||^2}{d\boldsymbol{\beta_j}} = \hat{\omega}_j\tilde{\lambda}_n\frac{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}}{\|\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}\|_2} \quad \textit{for } \|\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}\|_2 \neq 0,$$

$$\frac{d||\boldsymbol{Y} - \boldsymbol{\theta} - \boldsymbol{Z\beta}||^2}{d\boldsymbol{\beta_j}} \leqslant \hat{\omega}_j\tilde{\lambda}_n \quad \textit{for } \|\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}\|_2 = 0,$$

Similar to Lemma 5 and Lemma 6 of Huang et al. (2010), we give the following Lemmas:

Lemma A.4. *Under Assumptions 1-4 and Condition 4.1-4.2:*

$$P(\|\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} - \boldsymbol{\beta}_{\boldsymbol{j}}\|_2 \geqslant \|\boldsymbol{\beta}_{\boldsymbol{j}}\|_2, \quad \exists \boldsymbol{X}_{\boldsymbol{j}} \in \mathcal{L} \cup \mathcal{R} \cup \mathcal{S}) \to 0.$$

Lemma A.5. *Under Assumptions 1-4 and Condition 4.1-4.2:*

$$P(\|\boldsymbol{X}_{\boldsymbol{j}}^{\intercal}(\boldsymbol{Y} - \boldsymbol{Z}_{\boldsymbol{1}}\boldsymbol{\beta}_{\boldsymbol{1}})\|_{\boldsymbol{2}} > \tilde{\lambda}_n \hat{\omega}_j/2, \quad \exists \boldsymbol{X}_{\boldsymbol{j}} \notin \mathcal{L} \cup \mathcal{R} \cup \mathcal{S}) \to 0$$

Proofs of Lemma A.4 and Lemma A.5 can be found in the Appendix of Huang et al. (2010).

Proof of Theorem 4.2 : Theorem 4.2 satisfies the Condition 1 of Huang et al. (2010). Under Theorem **??**, and Lemma A.3, we set $\boldsymbol{\zeta} = (\frac{\hat{\omega}_j \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}}{2\|\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}\|})$, for $\boldsymbol{X}_{\boldsymbol{j}} \in \mathcal{L} \cup \mathcal{R} \cup \mathcal{S}$.

Therefore, we have:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{P_1}} = (\boldsymbol{Z}_{\boldsymbol{1}}^{\intercal}\boldsymbol{Z}_{\boldsymbol{1}})^{-\boldsymbol{1}}\boldsymbol{Z}_{\boldsymbol{1}}^{\intercal}(\boldsymbol{Y} - \tilde{\lambda}_n\boldsymbol{\zeta}).$$

To proof Theorem 4.2, equivalently, we need to proof:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\boldsymbol{P_1}} &=_{\boldsymbol{0}} \boldsymbol{\beta}_{\boldsymbol{P_1}} \\ \|\boldsymbol{Z}_j^{\intercal}(\boldsymbol{Y} - \boldsymbol{Z}_1\boldsymbol{\beta}_{P_1})\|_2 &\leqslant \tilde{\lambda}_n\hat{\omega}_j/2 \quad \forall j \notin \mathcal{L} \cup \mathcal{R} \cup \mathcal{S} \end{aligned}.$$

This is equivalently to show:

$$\begin{aligned} \|\boldsymbol{\beta}_j\|_2 - \|\hat{\boldsymbol{\beta}}_j\|_2 &< \|\boldsymbol{\beta}_j\|_2 \qquad \forall j \in \mathcal{L} \cup \mathcal{R} \cup \mathcal{S} \\ \|\boldsymbol{Z}_j^{\intercal}(\boldsymbol{Y} - \boldsymbol{Z}_1\boldsymbol{\beta}_{P_1})\|_2 &\leqslant \tilde{\lambda}_n\hat{\omega}_j/2 \quad \forall j \notin \mathcal{L} \cup \mathcal{R} \cup \mathcal{S} \end{aligned}.$$

Therefore,

$$\begin{aligned} P(\hat{\boldsymbol{\beta}}_{P_Z} \neq_0 \boldsymbol{\beta}_{P_Z}) \leqslant \quad & P(\boldsymbol{\beta}_j\|_2 - \|\hat{\boldsymbol{\beta}}_j\|_2 \geqslant \|\boldsymbol{\beta}_j\|_2 \qquad \exists j \in \mathcal{L} \cup \mathcal{R} \cup \mathcal{S}) \\ & + P(\|\boldsymbol{Z}_j^{\intercal}(\boldsymbol{Y} - \boldsymbol{Z}_1\boldsymbol{\beta}_{P_1})\|_2 > \tilde{\lambda}_n\hat{\omega}_j/2 \quad \exists j \notin \mathcal{L} \cup \mathcal{R} \cup \mathcal{S}) \end{aligned}.$$

Theorem 4.1 shows

$$\hat{\omega}_j \to \infty, \quad \forall j \notin \mathcal{L} \cup \mathcal{R} \cup \mathcal{S},$$

where ω_j is the specific penalty parameter of the j^{th} coefficient group.

Then,

$$P(\|\boldsymbol{Z}_j^{\intercal}(\boldsymbol{Y} - \boldsymbol{Z}_1\boldsymbol{\beta}_{P_1})\|_2 > \tilde{\lambda}_n\hat{\omega}_j/2, \quad \exists j \notin \mathcal{L} \cup \mathcal{R} \cup \mathcal{S}) \to 0$$

Therefore, under Lemma A.4 and Lemma A.5, the Theorem 4.2 follows. □

A.1 Characteristics

Table 5: Characteristic Details

Name	Description	Reference
A2ME	We define assets-market cap as total assets (AT) over market capitalization as of December t-1. Market capitalization is the product of shares outstanding (SHROUT) and price(PRC).	Bhandari (1988)
AT	Total assets (AT)	Gandhi and Lusting (2015)
ATO	Net sales over lagged net operating assets. Net operating assets are the difference between operating assets and operating liabilities. Operating assets are total assets (AT) minus cash and short-term investments (CHE), minus investment and other advances (IVAO). Operating liabilities are total assets (AT), minus debt in current liabilities(DLC),minus long-term debt (DLTT),minus minority interest (MIB), minus preferred stock (PSTK), minus common equity (CEQ).	Soliman(2008)
BEME	Ratio of book value of equity to market value of equity. Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITC), minus preferred stock (PS). SH is shareholder‘s equity (SEQ). If missing, SH is the sum of common equity (CEQ) and preferred stock (PS). If missing, SH is the difference between total assets (AT) and total liabilities (LT). Depending on availability, we use the redemption (item PSTKRV), liquidating (item PSTKL), or par value (item PSTK) for PS. The market value of equity is as of December t-1. The market value of equity is the product of shares outstanding (SHROUT) and price (PRC).	Rosenberg, Reid and Lanstein (1985) Davis, Fama, and French (2000)
C	Ration of cash and short-term investments (CHE) to total assets (AT)	Palazzo

C2D	Cash flow to price is the ratio of income and extraoridinary items (IB) and depreciation and amortization (dp) to total liabilities (LT).	
CTO	We define caoital turnover as ratio of net sales (SALE) to lagged total assets (AT).	Haugen and Baker (1996)
Debt2P	Debt to price is the radio of long-term debt (DLTT) and debt in current liabilities (DLC) to the market capitalization as of December t-1 . Market capitalization is the product of shares outstanding (SHROUT) and price (PRC).	Litzenberger and Ramaswamy (1979)
Δceq	The percentage change in the book value of equity (CEQ).	Richardson et al. (2005)
$\Delta(\Delta Gm - Sales)$	The difference in the percentage change in gross margin and the percentage change in sales (SALE). We define gross margin as the difference in sales (SALE) and costs of goods sold (COGS).	Abarbanell and Bushee (1997)
$\Delta Shrout$	The definition of the percentage change in shares outstanding (SHROUT).	Pontiff and Woodgate (2008)
$\Delta PI2A$	We define the change in property, plants ,and equipment as changes in property,plants,and equipment (PPEGT) and inventory (INVT) over lagged total assets (TA).	Lyandres , Sun, and Zhang (2008)
DTO	We define turnover as ratio of daily volume (VOL) to shares outstanding (SHROUT) minus the daily market turnover and de-trend it by its 180 trading day median. We scale down the volume of NASDAQ securities by 38% after 1997 and by 50% before that to address the issue of double-counting of volume for NASDAQ securities.	Garfinkel (2009); Anderson and Dyl (2005)
E2P	We define earnings to price as the ratio of income before extraordinary items (IB) to the market capitalization as December t-1 Market capitalization is the product of share outstanding (SHROUT) and price (PRC).	Basu (1983)

EPS	We define earnings per share as the ratio of income before extraordinary items (IB) to share outstanding (SHROUT) as of December t-1	Basu (1997)
Investment	We define investment as the percentage year-on-year growth rate in total assets (AT).	Cooper, Gulen and Schill(2008)
IPM	We define pre-tax profit margin as ratio of pre-tax income (PI) to sales (SALE).	
Lev	leverage is the ratio of long-term debt (DLTT) and debt in the current liabilities (DLC) to the sum of long-term debt, debt in current liabilities, and stockholders' equity (SEQ)	Lewenllen (2015)
LME	Size is the total market capitalization of the previous month defined as price (PRC) times shares outstanding (SHROUT)	Fama and French (1992)
Turnover	Turnover is last month's volume (VOL) over shares outstanding (SHROUT).	Datar, Naik and Radcliffe (1998)
OL	Operating leverage is the sum of cost of goods sold (COGS) and selling, general, and administrative expenses (XSGA) over total assets.	Novy-Marx (2011)
PCM	The price-to-cost margin is the difference between net sales (SALE) and costs of goods sold (COGS) divided by net sales (SALE).	Gorodnichenko and Weber (2016) and D'Acunto, Liu, Pflucger and Wcber (2017)
PM	The profit margin is operating income after depreciation (OIADP) over sales (SALE)	Soliman (2008)
Q	Tobin's Q is total assets (AT), the market value of equity (SHROUT times PRC) minus cash and short-term investments (CEQ) minus deferred taxes (TXDB) scaled by total assets (AT).	
ROA	Return-on-assets is income before extraordinary items (IB) to lagged total assets (AT).	Balakrishnan, Bartov and Faurel (2010)
ROC	ROC is the ratio of market value of equity (ME) plus long-term debt (DLTT)minus total assets to Cash and Short-Term Investments (CHE).	Chandrashekar and Rao (2009)

ROE	Return-on-equity is income before extraordinary items (IB) to lagged book-value of equity.	in Haugen and Baker (1996)
r_{12-2}	We define momentum as cumulative return from 12 months before the return prediction to two months before.	Fama and French (1996)
r_{12-7}	We define intermediate momentum as cumulative return from 12 months before the return prediction to seven months before.	Novy-Marx (2012)
r_{6-2}	We definer_{6-2} as cumulative return from 6 months before the return prediction to two months before.	Jegadeesh and Titman (1993)
r_{2-1}	We define short-term reversal as lagged one-month return.	Jegadeesh(1990)
S2C	Sales-to-cash is the ratio of net sales (SALE) to Cash and Short-Term Investments (CHE).	following Ou and Penman (1989)
Sales-G	Sales growth is the percentage growth rate in annual sales (SALE).	Lakonishok, Shleifer , and Vishmy (1994)
SAT	We define asset turnover as the ratio of sales (SALE) to total assets (AT).	Soliman (2008)
SGA2S	SGA to sales is the ratio of selling ,general and administrative expenses (XSGA) to net sales (SALE).	

References

A. Deaton and J. Muellbauer. *Economics and consumer behavior*. Cambridge university press, 1980.

C. Dong, O. B. Linton, and B. Peng. A weighted sieve estimator for nonparametric time series models with nonstationary variables. *SSRN Working paper*, 2019.

Y. Fan and Q. Li. A kernel-based method for estimating additive partially linear models. *Statistica Sinica*, pages 739–762, 2003.

J. Freyberger, A. Neuhierl, and M. Weber. Dissecting characteristics nonparametrically. Technical Report 5, 2020.

T. Hastie and R. Tibshirani. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, pages 1005–1016, 1990.

K. Hou, C. Xue, and L. Zhang. Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705, 2015.

J. Huang, J. L. Horowitz, and F. Wei. Variable selection in non-parametric additive models. *Annals of Statistics*, 38(4):2282–2313, 2010.

C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273, 2006.

Q. Li. Efficient estimation of additive partially linear models. *International Economic Review*, 41(4):1073–1092, 2000.

H. Liang, S. W. Thurston, D. Ruppert, T. Apanasovich, and R. Hauser. Additive partial linear models with measurement errors. *Biometrika*, 95(3):667–678, 2008.

Y. Lin and H. H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34(5):2272–2297, 2006.

O. Linton. Miscellanea efficient estimation of additive nonparametric regression models. *Biometrika*, 84(2): 469–473, 1997.

O. Linton and W. Härdle. Estimation of additive regression models with known links. *Biometrika*, 83(3):529–540, 1996.

O. B. Linton. Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, 16(4):502–523, 2000.

S. Ma and L. Yang. Spline-backfitted kernel smoothing of partially linear additive model. *Journal of Statistical Planning and Inference*, 141(1):204–219, 2011.

L. Schumaker. Spline functions: basic theory. *John Wiley&Sons, New York*, 1981.

S. Sperlich, D. Tjøstheim, and L. Yang. Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, 18(2):197–251, 2002.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

H. Wang and C. Leng. A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12): 5277–5286, 2008.

L. Wang, L. Yang, et al. Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *The Annals of Statistics*, 35(6):2474–2503, 2007.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov): 2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476): 1418–1429, 2006.