Computational Statistics and Data Analysis 103 (2016) 250–262

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A relative error-based approach for variable selection

Meiling Hao [a], Yunyuan Lin [b,*], Xingqiu Zhao [a]

[a] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

[b] Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 7 October 2015
Received in revised form 6 February 2016
Accepted 23 May 2016
Available online 28 May 2016

Keywords:
Adaptive lasso
ADMM algorithm
Diverging number
Oracle property
Relative error

ABSTRACT

The accelerated failure time model or the multiplicative regression model is well-suited to analyze data with positive responses. For the multiplicative regression model, the authors investigate an adaptive variable selection method via a relative error-based criterion and Lasso-type penalty with desired theoretical properties and computational convenience. With fixed or diverging number of variables in regression model, the resultant estimator achieves the oracle property. An alternating direction method of multipliers algorithm is proposed for computing the regularization paths effectively. A data-driven procedure based on the Bayesian information criterion is used to choose the tuning parameter. The finite-sample performance of the proposed method is examined via simulation studies. An application is illustrated with an analysis of one period of stock returns in Hong Kong Stock Exchange.

1. Introduction

We consider the multiplicative regression model

$$Y_i = \exp(X_i^\top \boldsymbol{\beta})\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where Y_i and X_i are pairs of response and p-vector of predictors, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is the p-vector of regression coefficient and ε is the positive unobservable random error. A multiplicative regression model or multiplicative error model is useful in analyzing data with positive response and heteroscedastic data, which are particularly common in economic, finance, reliability control, biomedical studies, epidemiological and social studies, etc. The estimation of model (1) is usually carried out by transforming the multiplicative models into linear models. However, a linear relationship in the transformed model is not linear in the original model. The analysis results based on linear models need to be transformed back to the original multiplicative measurement scale. A major aspect in regression analysis is variable selection or model selection. A variety of remarkable advancements have been developed for model selection; see for example, Chen and Donoho (1994), Tibshirani (1996), Fan and Li (2001), Shen and Ye (2002), Zou (2006), Wang and Leng (2007), Wang et al. (2007a,b), Wu et al. (2007), Huang et al. (2008a,b), Fan et al. (2009), Pötscher and Schneider (2009), Zhang and Lu (2007), Xu and Ying (2010), Huang et al. (2011), among many others.

The aforementioned estimation and model selection approaches are mostly based on criteria concerning the magnitude of absolute errors, for example, the least squares (LS) and least absolute deviation (LAD). However, in many practical applications, particularly in the analysis of heteroscedastic data, the LS and LAD methods are not adequate as they assign

[*] Corresponding author.
E-mail address: ylin@sta.cuhk.edu.hk (Y. Lin).

http://dx.doi.org/10.1016/j.csda.2016.05.013
0167-9473/© 2016 Elsevier B.V. All rights reserved.

equal weights to the variables. For instance, in the analysis of a number of stocks, comparison of share prices of different stocks is generally meaningless, especially when there is possible share split or reverse split. In lifetime data analysis, longer life time requires more accuracy in terms of absolute error for prediction. In categorical data analysis, more accuracy for prediction in terms of absolute error may be required for a category with larger percentage of observations. There are a number of studies regarding relative errors in the literature; see Narula and Wellington (1977), Makridakis et al. (1984), Khoshgoftaar et al. (1992), Makridakis (1993), Park and Stefanski (1998), Chen et al. (2010), Gneiting (2011), Kolassa and Martin (2011), Zhang and Wang (2013), Tofallis (2014), Demongeot et al. (2015) and Chen et al. (2016), etc. In particular, Chen et al. (2010) proposed the least absolute relative error estimation for model (1) by taking two types of relative errors: one is the absolute error relative to the actual and the other is the absolute error relative to the predicted value of the target, into account in the parameter estimation simultaneously, which enjoys certain dimensionless/unitless and robust properties. Recently, in order to pursue a smooth and convex objective function incorporating relative errors, Chen et al. (2016) introduced a superior criterion called the least product relative error estimation (LPRE) to estimate $\boldsymbol{\beta}$.

As pointed out by Kolassa and Martin (2011) and Tofallis (2014), the most widely used measure for assessing prediction in business and organizations, the mean absolute percentage error or mean magnitude of relative error: the absolute error relative to the target, tends to select models whose prediction error is low. Similar consideration can be found in Demongeot et al. (2015) for a functional framework. A model selection approach, that would have advantages over existing methods in terms of interpretability and prediction accuracy, is much desired. To tackle the problem, in the present paper, we consider to borrow the ideas from Chen et al. (2016) and propose a statistical procedure based on product relative errors and Lasso-type penalties for variable selection and parameter estimation for multiplicative error models. First, the proposed procedure is based on two types of relative errors, which is symmetric in the actual and its predictor and therefore is a balanced and superior criterion compared with the commonly-used mean absolute percentage error. Second, the resultant estimator is dimensionless or scale-free, which retained the original measurement scale. Third, the smooth and convex nature of the LPRE allows numerical simplicity and ensures uniqueness of the solution. With certain proper choice of tuning parameters, the resulting estimator is proved to achieve the oracle property in both settings of fixed and diverging number of variables. The variance estimation can be carried out directly by a plug-in rule. An alternating direction method of multipliers (ADMM) algorithm is proposed for computing the regularization paths effectively. Furthermore, we adopt a BIC-type criterion to select the tuning parameter adaptively.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed procedure along with a large sample theory including model selection consistency and the oracle property with fixed or diverging number of variables. Section 3 presents the ADMM algorithm to compute the resulting estimator. Section 4 reports some supportive simulation results and an application to a real dataset is given in Section 5. Some concluding remarks are given in Section 6 and all technical proofs are deferred to Appendix.

2. Methodology and main results

Let $(X_1^\top, Y_1), \ldots, (X_n^\top, Y_n)$ be n independent and identically distributed (i.i.d.) copies of $(X^\top, Y)$, where $X = (x_1, x_2, \ldots, x_p)^\top$ is the p-vector explanatory variable. Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$. Some notations are needed for ease of presentation. For two vectors $\boldsymbol{a} = (a_1, \ldots, a_d)^\top$ and $\boldsymbol{b} = (b_1, \ldots, b_d)^\top$, we define $\boldsymbol{a}\cdot\boldsymbol{b} = (a_1b_1, a_2b_2, \ldots, a_db_d)^\top$, $\boldsymbol{a}\cdot/\boldsymbol{b} = (a_1/b_1, \ldots, a_d/b_d)^\top$. Throughout the paper, the norm $\|\boldsymbol{a}\|_1 = \sum_{j=1}^d |a_j|$ and $\|\cdot\|$ is the Euclidean norm. For the multiplicative regression model, the least product relative error estimation proposed by Chen et al. (2016) is defined as the minimizer of

$$LPRE_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{ \left| \frac{Y_i - \exp(X_i^\top \boldsymbol{\beta})}{Y_i} \right| \times \left| \frac{Y_i - \exp(X_i^\top \boldsymbol{\beta})}{\exp(X_i^\top \boldsymbol{\beta})} \right| \right\}.$$

Similar to the LARE in Chen et al. (2010), the LPRE accounts for two types of relative errors simultaneously, hence it is symmetric in the target and its predictor. The LPRE can also be regarded as the product of two weighted forms of absolute deviations. In the present paper, we propose a variable selection approach with the product relative errors loss and Lasso-type penalties. As pointed out by Fan and Li (2001) and Zou (2006), the Lasso does not achieve the oracle property in the sense that it cannot simultaneously set all unnecessary regression coefficients to zero correctly with probability tending to one as n increases while having the optimal rate of convergence. To obtain the oracle property, we consider to use the adaptive Lasso penalty in this paper; see Zou (2006), Wang et al. (2007a,b), Zhang and Lu (2007), among many others. A straightforward algebraic calculation of the LPRE criterion function yields

$$LPRE_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{ Y_i \exp(-X_i^\top \boldsymbol{\beta}) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}) - 2 \right\}.$$

To be specific, we define the penalized LPRE estimator $\hat{\boldsymbol{\beta}}_n^*$ as the minimizer of

$$Z_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{ Y_i \exp(-X_i^\top \boldsymbol{\beta}) + \exp(X_i^\top \boldsymbol{\beta}) Y_i^{-1} \right\} + \lambda_n \|\boldsymbol{\beta}\cdot\boldsymbol{\omega}\|_1, \tag{2}$$

where λ_n is the regularization parameter, $\boldsymbol{\omega} = 1\cdot/|\widetilde{\boldsymbol{\beta}}_n|^\gamma = (1/|\widetilde{\beta}_{n1}|^\gamma, 1/|\widetilde{\beta}_{n2}|^\gamma, \ldots, 1/|\widetilde{\beta}_{np}|^\gamma)^\top$, γ is some specified positive number, and $\widetilde{\boldsymbol{\beta}}_n$ is certain consistent estimator of $\boldsymbol{\beta}$. In doing so, we gain several advantages. First, as evidenced in our simulation studies in Section 4, the shrinkage and oracle property of the adaptive Lasso continue to hold in the relative error-based regression. Second, under certain regularity conditions, the smoothness and convexity nature of the LPRE criterion function ensures the uniqueness of the global minimizer of (2), and hence makes its penalized counterpart computationally straightforward and easy to implement. Third, the resulting estimator enjoys the dimensionless or unit-free property.

To study large sample properties of the penalized estimator for fixed dimensionality, we now introduce more notations. Let $\mathcal{A} = \{j_1, \ldots, j_{s^*}\}$ be an arbitrary candidate model index set, $X_{\mathcal{A}} = \{x_{j_1}, \ldots, x_{j_{s^*}}\}$ be the associated covariate set, and $\boldsymbol{\beta}_{\mathcal{A}} = (\beta_{j_1}, \ldots, \beta_{j_{s^*}})^\top$ be the corresponding parameter. Moreover, let $\mathcal{A}_{\mathcal{T}} = \{j : \beta_{j0} \neq 0\}$ be the set of all the true non-zero parameter index and $\mathcal{A}_n^* = \{j : \hat{\beta}_{nj}^* \neq 0\}$ be the set of non-zero estimated coefficient index. Thereby, $\hat{\boldsymbol{\beta}}_{n\mathcal{A}_{\mathcal{T}}}^*$ is the estimator corresponding to the true active set while $\boldsymbol{\beta}_{0\mathcal{A}_{\mathcal{T}}}$ is the true active parameters. We note that $LPRE_n(\boldsymbol{\beta}_0) = (1/n) \sum_{i=1}^n (\varepsilon_i + \varepsilon_i^{-1}) - 2$. Define $V = E(1/\varepsilon - \varepsilon)^2 XX^\top$ and $D = E(\varepsilon + 1/\varepsilon)XX^\top$. Let $V_{\mathcal{A}_{\mathcal{T}}}$ and $D_{\mathcal{A}_{\mathcal{T}}}$ be the submatrices of D and V associated with $\mathcal{A}_{\mathcal{T}}$, respectively. The regularity conditions (C1)–(C6) are deferred to Appendix. Since a root-n consistent $\widetilde{\boldsymbol{\beta}}_n$ is usually readily available when p is fixed, we would use it in the proposed procedure in practice. The following theorem also presents the oracle property of the proposed estimator with a root-n consistent $\widetilde{\boldsymbol{\beta}}_n$.

Theorem 1 (*Oracle Property*). *Suppose Conditions* (C1)–(C5) *hold. If* $\lambda_n\sqrt{n} \to 0$ *and* $\lambda_n n^{(\gamma+1)/2} \to \infty$ *as* $n \to \infty$, *the proposed estimator enjoys the following properties:*

(a) *Variable Selection Consistency:* $\lim_{n\to\infty} P(\mathcal{A}_n^* = \mathcal{A}_{\mathcal{T}}) = 1$.

(b) *Asymptotic Normality:* $\sqrt{n}\{\hat{\boldsymbol{\beta}}_{n\mathcal{A}_{\mathcal{T}}}^* - \boldsymbol{\beta}_{0\mathcal{A}_{\mathcal{T}}}\} \to N(\mathbf{0}, \Sigma)$ *in distribution, where* $\Sigma = D_{\mathcal{A}_{\mathcal{T}}}^{-1} V_{\mathcal{A}_{\mathcal{T}}} D_{\mathcal{A}_{\mathcal{T}}}^{-1}$.

Remark. The tuning parameter λ_n plays a key role of a balance between estimation of $\boldsymbol{\beta}$ and variable selection here. By Theorem 1, one can see that the requirement on the order of λ_n is similar to that of Zou (2006). Theoretically, the order of λ_n varies with the choice of $\widetilde{\boldsymbol{\beta}}_n$. The asymptotic variance Σ is of sandwich formula and can be estimated directly via the plug-in rule. To be specific, for $\hat{\beta}_{nj}^* = 0$, the estimated variance is zero; for $\{\hat{\boldsymbol{\beta}}_{n\mathcal{A}_n^*}^* - \boldsymbol{\beta}_{0\mathcal{A}_n^*}\}$, the estimated variance is $\hat{\Sigma} = (1/n)\hat{D}_{n\mathcal{A}_n^*}^{-1} \hat{V}_{n\mathcal{A}_n^*} \hat{D}_{n\mathcal{A}_n^*}^{-1}$, where

$$\hat{D}_{n\mathcal{A}_n^*} = \frac{1}{n}\sum_{i=1}^n \{Y_i \exp(-X_{i\mathcal{A}_n^*}^\top \hat{\boldsymbol{\beta}}_{n\mathcal{A}_n^*}^*) + Y_i^{-1}\exp(X_{i\mathcal{A}_n^*}^\top \hat{\boldsymbol{\beta}}_{n\mathcal{A}_n^*}^*)\} X_{i\mathcal{A}_n^*} X_{i\mathcal{A}_n^*}^\top,$$

$$\hat{V}_{n\mathcal{A}_n^*} = \frac{1}{n}\sum_{i=1}^n \{-Y_i \exp(-X_{i\mathcal{A}_n^*}^\top \hat{\boldsymbol{\beta}}_{n\mathcal{A}_n^*}^*) + Y_i^{-1}\exp(X_{i\mathcal{A}_n^*}^\top \hat{\boldsymbol{\beta}}_{n\mathcal{A}_n^*}^*)\}^2 X_{i\mathcal{A}_n^*} X_{i\mathcal{A}_n^*}^\top.$$

With diverging number of explanatory variables, we rewrite model (1) as

$$Y_i = \exp(X_{ni}^\top \boldsymbol{\beta}_n)\varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where X_{ni} is the p_n-vector of explanatory variables, $\boldsymbol{\beta}_n$ is the corresponding p_n-vector of regression parameters. Let $\boldsymbol{\beta}_{n0}$ be the true value of $\boldsymbol{\beta}_n$, $\mathcal{A}_{\mathcal{T}} = \{j : \beta_{nj0} \neq 0\}$ be the true active set. And $s_n = \sharp\{j, j \in \mathcal{A}_{\mathcal{T}}\}$ is the number of non-vanishing parameters which may increase with n. For simplicity, we still write X_{ni} as X_i for short. Similarly, we propose to obtain an estimate of $\boldsymbol{\beta}_n$, denoted by $\hat{\boldsymbol{\beta}}_n^\star$, by minimizing $Z_n(\boldsymbol{\beta}_n)$ in (2) over $\boldsymbol{\beta}_n$, where $\widetilde{\boldsymbol{\beta}}_n$ is certain r_n-consistent estimator of $\boldsymbol{\beta}_n$ and $r_n \to \infty$ as $n \to \infty$. Define $d_n = \inf_{j\in\mathcal{A}_{\mathcal{T}}} |\beta_{n0j}|$. Under certain proper conditions, the following lemma characterizes the concentration of $\hat{\boldsymbol{\beta}}_n^\star$.

Lemma 1. *Suppose that Conditions* (C1)–(C3) *and Conditions* (C5)–(C6) *hold. If* $\lambda_n = O(\sqrt{p_n/(s_n n)}d_n^\gamma)$ *and* $p_n^4/n \to 0$ *as* $n \to \infty$, *we have* $\|\hat{\boldsymbol{\beta}}_n^\star - \boldsymbol{\beta}_{n0}\| = O_p(\sqrt{p_n/n})$.

Lemma 1 tells that $\hat{\boldsymbol{\beta}}_n^\star$ is asymptotically unbiased at the same rate as that of Fan and Peng (2004), which implies that the estimator of Chen et al. (2016) is a consistent estimator when p_n is diverging with n. By Lemma 1, we can establish the following main theorem.

Theorem 2 (*Oracle Property*). *Assume Conditions* (C1)–(C3) *and Conditions* (C5)–(C6) *hold. If* $\lambda_n n^{1/2}/d^\gamma \to 0$, $\lambda_n n^{1/2} r_n^\gamma / p_n^{1/2} \to \infty$, $p_n^4/n \to 0$ *as* $n \to \infty$ *and* $s_n = O(n^{1/6})$, $\hat{\boldsymbol{\beta}}_n^\star$ *must satisfy the following properties with probability tending to one:*

(a) *Sparsity:* $\hat{\boldsymbol{\beta}}_{n\mathcal{A}_{\mathcal{T}}{}^c}^\star = \mathbf{0}$, *where* $\mathcal{A}_{\mathcal{T}}{}^c = \{j : \beta_{nj0} = 0\}$.

(b) *Asymptotic Normality:* $\sqrt{n}A_n\widetilde{\Sigma}_n^{-1/2}\{\hat{\boldsymbol{\beta}}^{\star}_{n\mathcal{A}_{\mathcal{T}}} - \boldsymbol{\beta}_{n0\mathcal{A}_{\mathcal{T}}}\} \to N(\mathbf{0}, G)$ *in distribution. Here,* A_n *is a* $q \times s_n$ *deterministic matrix such that* $A_nA_n^T \to G$ *as* $n \to \infty$, *where* G *is a* $q \times q$ *positive definite matrix,* $\widetilde{\Sigma}_n$ *is the same as in* Theorem 1 *and* $\widetilde{\Sigma}_n^{-1/2} = V_{n\mathcal{A}_{\mathcal{T}}}^{-1/2}D_{n\mathcal{A}_{\mathcal{T}}}$.

By Theorem 2, with increasing number of covariates, we estimate the zero component of $\hat{\boldsymbol{\beta}}^{\star}_n$ to be exactly zero and estimate the non-zero component as efficient as when the true model was known. The asymptotic variance can be estimated by the plug-in method.

3. Computation

3.1. The ADMM algorithm

We propose an alternating direction method of multipliers (ADMM) algorithm to solve the penalized LPRE efficiently. The alternating direction method of multipliers (ADMM) was developed in the 1970s, with roots in the 1950s, is well suited to large scale convex optimization; see Boyd et al. (2011), Annergren et al. (2012), Wahlberg et al. (2012) and Shi et al. (2014), and etc. By blending the advantage of the dual ascent with the method of multipliers, the ADMM algorithm enjoys the decomposability and superior convergence. Here, the ADMM decouples the LPRE with L_1 to the smooth term and the ordinary least squares with L_1 term. For illustration, We define $x_{ij}^{**} = x_{ij}/\omega_j, j = 1, 2, \ldots, p$. Then, $X_i^{**} = X_i\boldsymbol{\cdot}/\boldsymbol{\omega}$. The main step is to minimize the following Lasso-type problem:

$$\hat{\boldsymbol{\beta}}_n^{**} \equiv \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i\exp(-X_i^{**\top}\boldsymbol{\beta}) + Y_i^{-1}\exp(X_i^{**\top}\boldsymbol{\beta})\right\} + \lambda_n\|\boldsymbol{\beta}\|_1.$$

Finally, we output $\hat{\boldsymbol{\beta}}_n^{*} = \hat{\boldsymbol{\beta}}_n^{**}\boldsymbol{\cdot}/\boldsymbol{\omega}$. Our proposed algorithm is to solve the above Lasso-type problem. Write

$$Z_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i\exp(-X_i^{**\top}\boldsymbol{\beta}) + \exp(X_i^{**\top}\boldsymbol{\beta})Y_i^{-1}\right\} + \lambda_n\|\boldsymbol{\beta}\|_1.$$

Step 1: Find the minimizer of $(1/n)\sum_{i=1}^{n}\left\{Y_i\exp(-X_i^{**\top}\boldsymbol{\beta}) + \exp(X_i^{**\top}\boldsymbol{\beta})Y_i^{-1}\right\}$, denoted by $\boldsymbol{\beta}_I$, which was shown to be consistent by Chen et al. (2016).

Step 2: The following quadratic approximation is applied to reduce computational cost:

$$\begin{aligned} LPRE_n(\boldsymbol{\beta}) &\approx LPRE_n(\boldsymbol{\beta}_I) + n^{-1/2}W_n^{**\top}\{\boldsymbol{\beta} - \boldsymbol{\beta}_I\} + 1/2\{\boldsymbol{\beta} - \boldsymbol{\beta}_I\}^{\top}D_n^{**}\{\boldsymbol{\beta} - \boldsymbol{\beta}_I\} \\ &\equiv \widehat{LPRE}_n(\boldsymbol{\beta}), \end{aligned}$$

where

$$W_n^{**} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{-Y_i\exp(-X_i^{**\top}\boldsymbol{\beta}_I) + Y_i^{-1}\exp(X_i^{**\top}\boldsymbol{\beta}_I)\}X_i^{**},$$

$$D_n^{**} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i\exp(-X_i^{**\top}\boldsymbol{\beta}_I) + Y_i^{-1}\exp(X_i^{**\top}\boldsymbol{\beta}_I)\}X_i^{**}X_i^{**\top}.$$

It follows from the definition of $\boldsymbol{\beta}_I$ that $W_n^{**} = 0$. Thereby,

$$\widehat{LPRE}_n(\boldsymbol{\beta}) = LPRE_n(\boldsymbol{\beta}_I) + 1/2\{\boldsymbol{\beta} - \boldsymbol{\beta}_I\}^{\top}D_n^{**}\{\boldsymbol{\beta} - \boldsymbol{\beta}_I\}.$$

We compute the minimizer of $\widehat{LPRE}_n(\boldsymbol{\beta}) + \lambda_n\|\boldsymbol{\beta}\|_1$, which approximates the minimizer of $Z_n^*(\boldsymbol{\beta})$ through the following ADMM algorithm.

Step 3: Initiate $k = 0, \rho = 1, \alpha = 1, \boldsymbol{u}_k = \mathbf{0}, \boldsymbol{z}_k = \mathbf{0}$;

Step 4: $\boldsymbol{\beta}_{k+1}^* = (D_n^{**} + \rho I)^{-1}\boldsymbol{q}$, $\boldsymbol{\beta}_{k+1} = \alpha\boldsymbol{\beta}_{k+1}^* + (1-\alpha)\boldsymbol{z}_k$, $\boldsymbol{q} = D_n^{**}\boldsymbol{\beta}_I + \rho(\boldsymbol{z}_k - \boldsymbol{u}_k)$, $\boldsymbol{z}_{k+1} = S_{\lambda_n/\rho}(\boldsymbol{\beta}_{k+1} + \boldsymbol{u}_k)$, $\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + (\boldsymbol{\beta}_{k+1} - \boldsymbol{z}_{k+1})$. Besides, the soft thresholding operator $S_\kappa(s)$ is defined as:

$$S_\kappa(s) = \begin{cases} s - \kappa, & s > \kappa; \\ 0, & |s| < \kappa; \\ s + \kappa, & s < -\kappa. \end{cases}$$

According to Boyd et al. (2011), only when both $\|\boldsymbol{\beta}_{k+1} - \boldsymbol{z}_{k+1}\|$ and $\|\rho(\boldsymbol{z}_{k+1} - \boldsymbol{z}_k)\|$ are smaller than ϵ_{k+1}^{pri} and ϵ_{k+1}^{dual} respectively, or $k+1$ is equal to a pre-set K, one can stop the iteration. Otherwise, we set $k = k+1$ and the algorithm goes back to *step* 4. Here,

$$\epsilon_{k+1}^{pri} = \sqrt{p}\epsilon^{abs} + \epsilon^{rel}\max(\|\boldsymbol{\beta}_{k+1}\|, \|\boldsymbol{z}_{k+1}\|);$$
$$\epsilon_{k+1}^{dual} = \sqrt{p}\epsilon^{abs} + \epsilon^{rel}\|\rho\boldsymbol{u}_{k+1}\|,$$

where $\epsilon^{abs} > 0$ is an absolute tolerance and $\epsilon^{rel} > 0$ is a relative tolerance, and both of them are very small.

Step 5: Let m represent the number of iterations. Then $\hat{\boldsymbol{\beta}}_n^{**} = \boldsymbol{z}_m$.

Moreover, similar to that of Boyd et al. (2011), the superfluous parameter ρ in the ADMM algorithm can vary in different steps to ease the computational cost. Besides, α is called the relaxation operator. For $\alpha \in [1.5, 1.8]$, it is called over-relaxation. The performance of convergence is usually better for over-relaxation. In our algorithm, we set $\alpha = 1.8$, $\rho = 1$, $\epsilon^{abs} = 10^{-4}$ and $\epsilon^{rel} = 10^{-2}$.

3.2. Choice of tuning parameter

The choice of tuning parameter is of key importance for variable selection. There are many well-known methods for choosing such regularization parameters; see Hastie et al. (2011). For practical application, we consider to minimize the Bayesian information criterion (BIC), a consistent data-driven variable selection procedure to select λ adaptively; see Wang et al. (2007a,b), Wang et al. (2009). For some specified γ, for example, $\gamma = 1$, the following BIC type principle is used to select λ:

$$\mathrm{BIC}_1(\lambda) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \exp(-X_i^\top \hat{\boldsymbol{\beta}}_{\lambda,\gamma}) + \exp(X_i^\top \hat{\boldsymbol{\beta}}_{\lambda,\gamma})Y_i^{-1} - 2\right\}\right] + C_n df_{\lambda,\gamma}\frac{\log(n)}{n}, \tag{3}$$

where $df_{\lambda,\gamma} = \sharp\{j, \hat{\beta}_{\lambda,\gamma} \neq 0\}$. We let $C_n = 1$ for fixed p. On the other hand, when p_n is diverging with the sample size, C_n has to satisfy $C_n s_n \log(n)/n \to 0$ as $n \to \infty$ theoretically. Alternatively, one may consider to minimize the following BIC-type criterion to choose λ:

$$\mathrm{BIC}_2(\lambda) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log(Y_i) - X_i^\top \hat{\boldsymbol{\beta}}_{\lambda,\gamma}\right\}^2\right] + C_n df_{\lambda,\gamma}\frac{\log(n)}{n}. \tag{4}$$

4. Simulation studies

Simulation studies are conducted to examine the finite-sample performance of the proposed method compared with least squares. We simulate the data from two settings of the multiplicative regression model $Y = \exp(X^\top \boldsymbol{\beta}_0)\varepsilon$. In the first setting, we consider fixed dimensionality with $p = 8$ and $\boldsymbol{\beta}_{\mathcal{A}_\mathcal{T}} = (1, 0.5, 0.25)^\top$ or $\boldsymbol{\beta}_{\mathcal{A}_\mathcal{T}} = (1.5, 2, 3)^\top$. The covariate X is generated from the multivariate normal distribution with covariance matrix $V = (\rho_{ij})$, where $\rho_{ij} = \sigma^{|j-i|}$ with $\sigma = \pm 0.5, \pm 0.8$, respectively. In the second setting, we consider a diverging number of explanatory variables with $p_n = [4n^{1/4}] - 4$ where $[x]$ is the integer part of a real number x, and $\boldsymbol{\beta}_{\mathcal{A}_\mathcal{T}} = (4, 3, 2, 1.5, 1, 2.5)^\top$ or $\boldsymbol{\beta}_{\mathcal{A}_\mathcal{T}} = (4, 3, 2, 1, 0.5, 0.25)^\top$. The covariate X is generated from the multivariate normal distribution with identity covariance matrix. We consider two distributions for $\log(\varepsilon)$: $N(0, 1)$ and Uniform$[-2, 2]$. Throughout the simulations, we randomly put the nonzero components in the p-vector regression coefficient with the rest being zero. The simulations are based on 1000 replications and the sample size $n = 250, 500, 1000$. To avoid a lengthy paper, we only present the results with moderate sample size $n = 250$ for the first setting.

We refer our method with adaptive Lasso penalty as PR-aLasso. For comparison, we also compute the LS and LAD with adaptive Lasso penalty, denoted by LS-aLasso and LAD-aLasso, respectively. Moreover, as it is suggested by one reviewer, we also compare the proposed method with a straightforward application of the BIC principle when p is fixed, that is the best subset selection for the LPRE estimate via criteria (3) and (4), denoted by BIC$_1$ and BIC$_2$, respectively. For the selection of λ, we set $\gamma = 1$ for the two BIC criteria in (3) and (4), while we set $C_n = \sqrt{\log(p_n)}$ in (3) and $C_n = \log\log(p_n)$ in (4) for diverging p_n. The results are summarized in Tables 1–6. Tables 1–4 present the selection performance of different approaches in terms of: the rate of over-fitted models (OF) meaning the portion of $\mathcal{A}_n^* \supset \mathcal{A}_T$, the rate of the correctly fitted models (CF) meaning the portion of $\mathcal{A}_n^* = \mathcal{A}_T$ exactly, the false positive rate (FP) meaning the portion of occasions on which the model selected contains some zero components, and the false negative rate (FN) meaning the portion of occasions on which the model selected excludes some nonzero components. Tables 5–6 present the estimation performance in terms of bias (BIAS), the empirical standard error (SE), the estimated standard error (SEE) and the empirical coverage probability (CP) of the 95% confidence interval based on the normal approximation.

It can be seen from the tables that the proposed method performs the best in model selection and parameter estimation compared with other penalized methods in all cases when $\log(\varepsilon)$ follows Uniform$[-2, 2]$. And the proposed method is generally comparable to the penalized LS when $\log(\varepsilon)$ follows $N(0, 1)$ and it is comparable to BIC$_1$ and BIC$_2$ in the first setting. In particular, Tables 1–4 show that the PR-aLasso(BIC$_1$) almost has the smallest FN in the first setting and the largest CF in the second setting.

5. Application

The dataset to be analyzed is obtained by the Bloomberg. It consists of records for 437 companies in the Hong kong Stock Exchange, including the monthly closed price in January and February 2012 and the values of 12 factors in January 2012 related the company: the open-price (OPX), asset turnover (ATO), profit margin (PM), degree of financial leverage

Table 1
Variable selection results with $p = 8$, $\beta_{\mathcal{A}_\mathcal{T}} = (\beta_{\mathcal{A}_\mathcal{T}1}, \beta_{\mathcal{A}_\mathcal{T}2}, \beta_{\mathcal{A}_\mathcal{T}3})^\top = (1, 0.5, 0.25)^\top$ and $n = 250$.

σ	Method	$\log(\varepsilon) \sim N(0, 1)$				$\log(\varepsilon) \sim U(-2, 2)$			
		OF	CF	FP	FN	OF	CF	FP	FN
−0.5	BIC$_1$	0.170	0.744	0.226	0.086	0.078	0.825	0.126	0.097
	BIC$_2$	0.064	0.839	0.104	0.097	0.066	0.720	0.130	0.214
	PR-aLasso(BIC$_1$)	0.262	0.669	0.287	0.069	0.184	0.713	0.222	0.103
	PR-aLasso(BIC$_2$)	0.166	0.744	0.191	0.090	0.140	0.702	0.175	0.158
	LS-aLasso	0.170	0.740	0.189	0.090	0.163	0.658	0.209	0.179
	LAD-aLasso	0.060	0.661	0.075	0.279	0.067	0.396	0.130	0.537
0.5	BIC$_1$	0.178	0.674	0.228	0.148	0.090	0.828	0.123	0.082
	BIC$_2$	0.061	0.754	0.093	0.185	0.062	0.725	0.126	0.213
	PR-aLasso(BIC$_1$)	0.207	0.642	0.232	0.151	0.185	0.697	0.224	0.118
	PR-aLasso(BIC$_2$)	0.106	0.683	0.122	0.211	0.118	0.684	0.150	0.198
	LS-aLasso	0.108	0.690	0.124	0.202	0.144	0.636	0.184	0.220
	LAD-aLasso	0.023	0.468	0.033	0.509	0.037	0.353	0.081	0.610
−0.8	BIC$_1$	0.107	0.529	0.342	0.364	0.049	0.610	0.227	0.341
	BIC$_2$	0.031	0.552	0.107	0.417	0.033	0.378	0.255	0.589
	PR-aLasso(BIC$_1$)	0.244	0.471	0.426	0.285	0.209	0.483	0.408	0.308
	PR-aLasso(BIC$_2$)	0.167	0.507	0.346	0.326	0.159	0.468	0.346	0.373
	LS-aLasso	0.176	0.513	0.360	0.311	0.182	0.412	0.384	0.406
	LAD-aLasso	0.060	0.446	0.237	0.494	0.051	0.146	0.208	0.803
0.8	BIC$_1$	0.113	0.518	0.331	0.369	0.045	0.583	0.208	0.372
	BIC$_2$	0.045	0.534	0.231	0.421	0.030	0.361	0.221	0.609
	PR-aLasso(BIC$_1$)	0.192	0.500	0.352	0.308	0.126	0.477	0.315	0.397
	PR-aLasso(BIC$_2$)	0.106	0.528	0.257	0.366	0.079	0.419	0.244	0.502
	LS-aLasso	0.106	0.547	0.252	0.347	0.082	0.378	0.256	0.540
	LAD-aLasso	0.027	0.394	0.136	0.579	0.021	0.162	0.131	0.817

Table 2
Variable selection results with $p = 8$, $\beta_{\mathcal{A}_\mathcal{T}} = (\beta_{\mathcal{A}_\mathcal{T}1}, \beta_{\mathcal{A}_\mathcal{T}2}, \beta_{\mathcal{A}_\mathcal{T}3})^\top = (1.5, 2, 3)^\top$ and $n = 250$.

σ	Method	$\log(\varepsilon) \sim N(0, 1)$				$\log(\varepsilon) \sim U(-2, 2)$			
		OF	CF	FP	FN	OF	CF	FP	FN
−0.5	BIC$_1$	0.198	0.802	0.198	0	0.099	0.901	0.099	0
	BIC$_2$	0.082	0.918	0.082	0	0.083	0.917	0.083	0
	PR-aLasso(BIC$_1$)	0.141	0.859	0.141	0	0.068	0.932	0.068	0
	PR-aLasso(BIC$_2$)	0.070	0.930	0.070	0	0.051	0.949	0.051	0
	LS-aLasso	0.060	0.940	0.060	0	0.068	0.932	0.068	0
	LAD-aLasso	0.031	0.969	0.031	0	0.065	0.935	0.065	0
0.5	BIC$_1$	0.199	0.801	0.199	0	0.097	0.903	0.097	0
	BIC$_2$	0.064	0.936	0.064	0	0.087	0.913	0.087	0
	PR-aLasso(BIC$_1$)	0.132	0.868	0.132	0	0.069	0.931	0.069	0
	PR-aLasso(BIC$_2$)	0.043	0.957	0.043	0	0.051	0.949	0.051	0
	LS-aLasso	0.045	0.955	0.045	0	0.064	0.936	0.064	0
	LAD-aLasso	0.018	0.982	0.018	0	0.064	0.936	0.064	0
−0.8	BIC$_1$	0.196	0.804	0.196	0	0.081	0.919	0.081	0
	BIC$_2$	0.078	0.922	0.078	0	0.083	0.917	0.083	0
	PR-aLasso(BIC$_1$)	0.156	0.844	0.156	0	0.069	0.931	0.069	0
	PR-aLasso(BIC$_2$)	0.082	0.918	0.082	0	0.054	0.946	0.054	0
	LS-aLasso	0.081	0.919	0.081	0	0.072	0.928	0.072	0
	LAD-aLasso	0.021	0.979	0.021	0	0.072	0.928	0.072	0
0.8	BIC$_1$	0.185	0.815	0.185	0	0.079	0.921	0.079	0
	BIC$_2$	0.071	0.929	0.071	0	0.076	0.924	0.076	0
	PR-aLasso(BIC$_1$)	0.139	0.861	0.139	0	0.061	0.939	0.061	0
	PR-aLasso(BIC$_2$)	0.051	0.949	0.051	0	0.047	0.953	0.047	0
	LS-aLasso	0.051	0.949	0.051	0	0.061	0.939	0.061	0
	LAD-aLasso	0.015	0.985	0.015	0	0.064	0.936	0.064	0

(DFL), sales-growth rate (SGR), revenue-sequential-growth (RGR), accounts receivable turnover (ACCT), inventory-growth to sales-growth (INVGR), price-to-book ratio (PB), the logarithm of total assets (ASSET), the logarithm of other assets (OASS), and the price-to-earning ratio (PE). If other assets are 0, we set OASS to be 0. As pointed out by Wang et al. (2009), these variables are among the most important explanatory factors in the prediction of future earnings. In particular, open-price (OPX) is the security first traded on the first day in one month; asset turnover (ATO) is a financial ratio measuring the efficiency of a company's use of its assets; profit margin (PM) is a measure of the company's profitability; degree of financial leverage (DEL) summarizes the affect of a particular amount of financial leverage has on a company's earnings; sales growth rate (SGR), revenue-sequential-growth (RGR) and inventory-growth to sales-growth (INVGR) measure the actual previous period growth; the account receivable turnover (ACCT) represents the average net sales in a period; price-to-book ratio is

Table 3
Variable selection results with $p = [4n^{1/4}] - 4$ and $\beta_{\mathcal{A}_\mathcal{T}} = (4, 3, 2, 1, 0.5, 0.25)^\top$.

n	Method	$\log(\varepsilon) \sim N(0, 1)$				$\log(\varepsilon) \sim U(-2, 2)$			
		OF	CF	FP	FN	OF	CF	FP	FN
250	PR-aLasso(BIC$_1$)	0.098	0.796	0.101	0.106	0.059	0.791	0.063	0.150
	PR-aLasso(BIC$_2$)	0.174	0.769	0.181	0.057	0.145	0.734	0.150	0.121
	LS-aLasso	0.182	0.762	0.187	0.056	0.171	0.684	0.179	0.145
	LAD-aLasso	0.081	0.711	0.085	0.208	0.092	0.395	0.118	0.513
500	PR-aLasso(BIC$_1$)	0.091	0.904	0.091	0.005	0.036	0.952	0.036	0.012
	PR-aLasso(BIC$_2$)	0.128	0.871	0.128	0.001	0.099	0.889	0.099	0.012
	LS-aLasso	0.127	0.872	0.127	0.001	0.141	0.843	0.141	0.016
	LAD-aLasso	0.059	0.912	0.059	0.029	0.088	0.625	0.092	0.287
1000	PR-aLasso(BIC$_1$)	0.044	0.956	0.044	0	0.017	0.983	0.017	0
	PR-aLasso(BIC$_2$)	0.070	0.930	0.070	0	0.051	0.949	0.051	0
	LS-aLasso	0.054	0.946	0.054	0	0.078	0.922	0.078	0
	LAD-aLasso	0.018	0.982	0.018	0	0.062	0.850	0.062	0.088

Table 4
Variable selection results with $p = [4n^{1/4}] - 4$ and $\beta_{\mathcal{A}_\mathcal{T}} = (4, 3, 2, 1.5, 1, 2.5)^\top$.

n	Method	$\log(\varepsilon) \sim N(0, 1)$				$\log(\varepsilon) \sim U(-2, 2)$			
		OF	CF	FP	FN	OF	CF	FP	FN
250	PR-aLasso(BIC$_1$)	0.046	0.954	0.046	0	0.017	0.983	0.017	0
	PR-aLasso(BIC$_2$)	0.103	0.897	0.103	0	0.073	0.927	0.073	0
	LS-aLasso	0.100	0.900	0.100	0	0.096	0.904	0.096	0
	LAD-aLasso	0.046	0.954	0.046	0	0.102	0.898	0.102	0
500	PR-aLasso(BIC$_1$)	0.040	0.960	0.040	0	0.010	0.990	0.010	0
	PR-aLasso(BIC$_2$)	0.076	0.924	0.076	0	0.050	0.950	0.050	0
	LS-aLasso	0.065	0.935	0.065	0	0.063	0.937	0.063	0
	LAD-aLasso	0.036	0.964	0.036	0	0.063	0.937	0.063	0
1000	PR-aLasso(BIC$_1$)	0.030	0.970	0.030	0	0.010	0.990	0.010	0
	PR-aLasso(BIC$_2$)	0.046	0.954	0.046	0	0.036	0.964	0.036	0
	LS-aLasso	0.033	0.967	0.033	0	0.043	0.957	0.043	0
	LAD-aLasso	0.005	0.995	0.005	0	0.029	0.971	0.029	0

Table 5
Estimation results with $p = 8, n = 250, \log(\varepsilon) \sim N(0, 1)$ and $\beta_{\mathcal{A}_\mathcal{T}} = (\beta_{\mathcal{A}_\mathcal{T}1}, \beta_{\mathcal{A}_\mathcal{T}2}, \beta_{\mathcal{A}_\mathcal{T}3})^\top = (1.5, 2, 3)^\top$.

σ	Method	$\beta_{\mathcal{A}_\mathcal{T}1}$				$\beta_{\mathcal{A}_\mathcal{T}2}$				$\beta_{\mathcal{A}_\mathcal{T}3}$			
		BIAS	SE	SEE	CP (%)	BIAS	SE	SEE	CP (%)	BIAS	SE	SEE	CP (%)
−0.5	BIC$_1$	0.001	0.084	0.075	92.1	−0.001	0.076	0.066	91.9	0.000	0.076	0.074	94.7
	BIC$_2$	0.000	0.082	0.075	92.6	−0.001	0.072	0.066	93.3	0.000	0.077	0.075	94.9
	PR-aLasso(BIC$_1$)	−0.011	0.083	0.075	92.4	−0.006	0.072	0.066	92.5	−0.008	0.077	0.074	93.6
	PR-aLasso(BIC$_2$)	−0.017	0.081	0.076	92.9	−0.009	0.070	0.066	92.6	−0.012	0.076	0.075	93.6
	LS-aLasso	−0.012	0.079	0.074	93.6	−0.006	0.067	0.064	93.8	−0.009	0.073	0.073	95.1
	LAD-aLasso	−0.013	0.097	0.096	93.6	−0.005	0.081	0.083	95.1	−0.008	0.096	0.095	94.3
0.5	BIC$_1$	0.003	0.074	0.067	92.1	0.001	0.082	0.074	92.4	0.001	0.082	0.076	93.0
	BIC$_2$	0.002	0.070	0.067	93.2	0.001	0.081	0.075	92.8	0.0010	0.081	0.077	93.1
	PR-aLasso(BIC$_1$)	−0.003	0.071	0.067	92.8	0.002	0.080	0.075	92.9	−0.004	0.082	0.076	92.8
	PR-aLasso(BIC$_2$)	−0.009	0.068	0.067	93.2	0.001	0.080	0.075	92.6	−0.007	0.081	0.077	93.1
	LS-aLasso	−0.005	0.066	0.065	94.4	0.001	0.075	0.073	93.5	−0.005	0.076	0.075	94.7
	LAD-aLasso	−0.006	0.085	0.085	94.9	−0.001	0.093	0.095	94.8	−0.005	0.097	0.098	95.2
−0.8	BIC$_1$	0.002	0.119	0.088	94.9	−0.004	0.132	0.120	93.2	−0.004	0.115	0.106	92.6
	BIC$_2$	0.002	0.102	0.086	90.2	−0.004	0.128	0.120	93.6	−0.003	0.115	0.107	92.6
	PR-aLasso(BIC$_1$)	−0.006	0.107	0.088	90.8	−0.024	0.117	0.120	92.9	−0.025	0.131	0.107	91.8
	PR-aLasso(BIC$_2$)	−0.008	0.099	0.086	91.7	−0.030	0.128	0.120	93.2	−0.033	0.114	0.107	91.8
	LS-aLasso	−0.005	0.090	0.084	93.6	−0.027	0.122	0.118	93.5	−0.028	0.111	0.105	92.6
	LAD-aLasso	0.001	0.107	0.108	95.5	−0.027	0.155	0.154	95.1	−0.025	0.136	0.138	94.3
0.8	BIC$_1$	0.002	0.106	0.087	91.8	0.000	0.123	0.108	91.4	0.001	0.132	0.121	93.1
	BIC$_2$	0.001	0.097	0.085	92.8	0.001	0.118	0.108	91.8	0.002	0.126	0.120	93.8
	PR-aLasso(BIC$_1$)	−0.006	0.099	0.087	92.4	−0.005	0.120	0.108	92.8	0.008	0.130	0.121	93.1
	PR-aLasso(BIC$_2$)	−0.013	0.093	0.085	92.3	−0.011	0.117	0.108	93.1	0.014	0.127	0.120	93.4
	LS-aLasso	−0.009	0.089	0.083	93.7	−0.009	0.110	0.106	93.5	0.010	0.122	0.118	94.3
	LAD-aLasso	−0.010	0.108	0.108	94.8	−0.008	0.136	0.138	94.8	0.020	0.152	0.154	95.2

used to compare a company's current market price to its book value while price-to-earning ratio measures the price paid for a share relative to the annual income or profit per share earned by the firm.

It is known that the stocks of different companies have different units that are not well defined. Comparison of the absolute estimation errors of the share prices of different stocks is not of practical referential value to practitioners. A

Table 6
Estimation results with $p = 8, n = 250, \log(\varepsilon) \sim U(-2, 2), \beta_{\mathcal{A}_T} = (\beta_{\mathcal{A}_T 1}, \beta_{\mathcal{A}_T 2}, \beta_{\mathcal{A}_T 3})^\top = (1.5, 2, 3)^\top$.

σ	Method	$\beta_{\mathcal{A}_T 1}$				$\beta_{\mathcal{A}_T 2}$				$\beta_{\mathcal{A}_T 3}$			
		BIAS	SE	SEE	CP (%)	BIAS	SE	SEE	CP (%)	BIAS	SE	SEE	CP (%)
−0.5	BIC$_1$	0.001	0.066	0.062	94.0	−0.001	0.072	0.069	94.5	−0.001	0.074	0.071	95.4
	BIC$_2$	0.001	0.066	0.062	94.2	−0.001	0.071	0.071	94.5	−0.001	0.074	0.069	95.4
	PR-aLasso(BIC$_1$)	−0.006	0.065	0.062	94.1	−0.012	0.071	0.069	94.4	−0.008	0.073	0.071	95.1
	PR-aLasso(BIC$_2$)	−0.008	0.066	0.062	93.4	−0.015	0.073	0.069	93.3	−0.010	0.074	0.071	94.2
	LS-aLasso	−0.008	0.079	0.076	93.6	−0.014	0.086	0.085	94.4	−0.010	0.088	0.087	94.6
	LAD-aLasso	−0.013	0.136	0.127	92.9	−0.022	0.148	0.142	92.4	−0.015	0.152	0.145	93.0
0.5	BIC$_1$	0.001	0.064	0.062	93.8	−0.004	0.075	0.071	93.4	−0.000	0.072	0.069	93.7
	BIC$_2$	0.001	0.063	0.062	93.9	−0.003	0.075	0.071	93.2	−0.000	0.072	0.069	93.7
	PR-aLasso(BIC$_1$)	−0.006	0.063	0.063	94.0	−0.007	0.074	0.071	93.8	−0.001	0.072	0.069	93.7
	PR-aLasso(BIC$_2$)	−0.009	0.065	0.062	93.2	−0.009	0.075	0.071	93.4	−0.002	0.073	0.069	93.3
	LS-aLasso	−0.008	0.076	0.076	94.8	−0.009	0.089	0.087	93.6	−0.002	0.088	0.084	93.6
	LAD-aLasso	−0.014	0.132	0.127	92.5	−0.009	0.156	0.145	92.3	−0.007	0.148	0.141	93.6
−0.8	BIC$_1$	−0.002	0.094	0.079	91.0	−0.002	0.116	0.111	94.7	−0.001	0.104	0.099	93.9
	BIC$_2$	−0.003	0.094	0.079	90.9	−0.002	0.117	0.111	94.6	−0.001	0.104	0.099	93.9
	PR-aLasso(BIC$_1$)	−0.010	0.090	0.079	91.8	−0.025	0.117	0.111	94.0	−0.027	0.105	0.099	92.3
	PR-aLasso(BIC$_2$)	−0.011	0.090	0.078	91.9	−0.028	0.119	0.111	93.5	−0.031	0.109	0.099	91.0
	LS-aLasso	−0.012	0.108	0.097	93.0	−0.031	0.143	0.136	93.0	−0.034	0.128	0.122	92.6
	LAD-aLasso	−0.011	0.182	0.162	92.6	−0.050	0.246	0.229	92.1	−0.054	0.217	0.204	92.3
0.8	BIC$_1$	0.002	0.086	0.079	93.9	−0.005	0.120	0.111	93.0	−0.001	0.105	0.099	93.8
	BIC$_2$	0.003	0.086	0.079	93.7	−0.004	0.120	0.111	93.1	−0.001	0.104	0.099	93.9
	PR-aLasso(BIC$_1$)	−0.007	0.083	0.079	94.4	−0.009	0.120	0.111	93.1	0.002	0.106	0.099	93.5
	PR-aLasso(BIC$_2$)	−0.010	0.084	0.078	93.6	−0.010	0.121	0.111	92.7	0.004	0.107	0.099	93.2
	LS-aLasso	−0.008	0.100	0.096	94.7	−0.010	0.145	0.137	93.4	0.002	0.128	0.122	93.9
	LAD-aLasso	−0.011	0.176	0.162	93.0	−0.006	0.249	0.229	93.1	−0.001	0.217	0.204	93.4

Table 7
Analysis of Hong Kong stock return data with different methods.

Variables	Method					
	PR-aLasso(BIC$_1$)	PR-aLasso(BIC$_2$)	LS-aLasso	LAD-aLasso	BIC$_1$	BIC$_2$
INT	−0.23144	−0.23144	−0.15647		−0.32834	−0.32834
OPX						
ATO						
PM						
DFL						
SGR	−0.00004	−0.00004			−0.00011	−0.00011
RGR	−0.00002	−0.00002				
ACCT	0.00038	0.00038	0.00024		0.00056	0.00056
INVGR						
PB						
ASSET	0.03782	0.03782	0.02991	0.01296	0.04824	0.04824
OASS						
PE						

Note: — indicates the insignificant covariates.

relative error-based model selection and estimation approach that is unit-free makes more sense here. We note that the relative error is approximately equal to the absolute error only when the size of the relative error is small. We consider to fit model (1) with an intercept term (INI) that relates the ratio of the closed-price of February 2012 to that of January 2012 to the 12 predictors for regression analysis. Similar to the simulation studies, we select the variables and estimate the coefficients with different approaches. The tuning parameter is selected by (3) with $C_n = \sqrt{\log(\log(p_n))}$ and (4) with $C_n = \log(\log(p_n))$. For the penalized LAD, we set $C_n = \log(\log(p_n))$. The results are reported in Table 7. It can be seen that the proposed method is able to select small signals, which is in accordance with the simulation results. Moreover, except LAD, all the methods select ACCT and ASSET, whose coefficients are positive. It reveals that the closed-price ratio is positively associated with ACCT and ASSET via our analysis with this particular dataset, which partly suggests that the share price of a company with high accounts receivable turnover (ACCT) tends to go up, while a firm with larger total assets (ASSET) has a higher profitability. Besides, the negative coefficients of the SGR and RGR may due to the falling of Euro, which arises the market slump from late 2011 to early 2012.

As it is suggested by the referees, we perform prediction for the monthly closed prices for the following three months, respectively, to evaluate the prediction power of different methods. The prediction accuracy are measured by three criteria: the mean of absolute prediction error $\{|Y_{te} - \hat{Y}|\}$ (MPE); the mean of squared prediction errors $\{|Y_{te} - \hat{Y}|^2\}$ (MSPE); the mean of absolute relative errors $\{|Y_{te} - \hat{Y}|^2/(\hat{Y}Y_{te})\}$ (MPPE). The results were summarized in Table 8. It can be concluded that the proposed method is most powerful in terms of prediction power in this real example.

Table 8
Comparisons of prediction power with different methods for the real example.

Months	Method						
	Errors	PR-aLasso(BIC$_1$)	PR-aLasso(BIC$_2$)	LS-aLasso	LAD-aLasso	BIC$_1$	BIC$_2$
Feb.	MPE	0.2783	0.2783	0.2807	0.2847	0.2808	0.2808
	MSPE	0.1372	0.1372	0.1388	0.1443	0.1383	0.1383
	MPPE	0.0985	0.0985	0.1008	0.1045	0.0982	0.0982
Mar.	MPE	0.2693	0.2693	0.2701	0.2731	0.2743	0.2743
	MSPE	0.11717	0.1171	0.1178	0.1199	0.1205	0.1205
	MPPE	0.1206	0.1206	0.1230	0.1255	0.1208	0.1208
Apr.	MPE	0.2702	0.2702	0.2709	0.2730	0.2742	0.2742
	MSPE	0.1168	0.1168	0.1162	0.1172	0.1217	0.1217
	MPPE	0.1104	0.1104	0.1110	0.1124	0.1124	0.1124

6. Concluding remarks

This paper complements to the literature with a relative error-based model selection and estimation approach, that is possibly superior to the existing methods in terms of prediction power and interpretability, and therefore may have wider applications in financial/economic data analysis and survival data analysis, as shown in the simulation studies and the real data analysis of this paper. Such consideration by taking both types of relative errors into account for parameter estimation and model selection may be extended to various parametric and semiparametric models. For other parametric or semiparametric models, the complication for an relative error-based approach to work theoretically is that additional constraints on the error distribution and the nonparametric component need to be imposed for model identifiability. It remains unclear how to make such a constraint under minimum conditions. We shall work along this direction and find reliable computation procedures for possible extensions to other semiparametric models.

Acknowledgments

The authors are grateful to the Editor, the AE and two referees for their careful review and constructive comments that lead to substantial improvements in the paper. Lin's work is supported by the Hong Kong Research Grants Council (Grant No. 509413) and a Direct Grant for Research at the Chinese University of Hong Kong (Grant No. 4053136). Zhao's research is partly supported by the Research Grant Council of Hong Kong (Grant No. 503513), the Natural Science Foundation of China (No. 11371299), and The Hong Kong Polytechnic University.

Appendix

The following regularity conditions are needed to study the asymptotic properties of the proposed estimator.

(C1) The error term satisfies $E(\varepsilon - 1/\varepsilon|X) = 0$.
(C2) There exists $\delta > 0$ such that $E\{(\varepsilon + 1/\varepsilon)\exp(\delta\|X\|)\} < \infty$.
(C3) The matrix $D = E\{XX^\top(\varepsilon + 1/\varepsilon)\}$ is positive definite.
(C4) There exists $\delta > 0$, such that $E\{(\varepsilon + 1/\varepsilon)\|X\|^3\exp(\delta\|X\|)\} < \infty$.
(C5) Define $V \equiv E\{XX^\top(-\varepsilon + 1/\varepsilon)^2\}$. There exist constants c_1, c_2, such that $0 < c_1 < \lambda_{\min}(V) \le \lambda_{\max}(V) < c_2 < \infty$. Besides, $E(x_jx_k)^2(-\varepsilon + 1/\varepsilon)^4 < c_3 < \infty$, $E(x_jx_k)^2(-\varepsilon + 1/\varepsilon)^2 < c_4 < \infty$.
(C6) There exists $\delta > 0$ such that $E\{(-\varepsilon + 1/\varepsilon)^2(x_jx_kx_l)^2\exp(\delta\|X\|)\} < c_5 < \infty$.

Proof of Theorem 1. First, we wish to prove $\|\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$. Let $\mathscr{B} = \{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{u}n^{-1/2}, \|\boldsymbol{u}\| \le C\}$ for some constant C. It suffices to show that for any $\varepsilon > 0$, there exists a sufficiently large constant C, such that

$$P\{\sup_{\|\boldsymbol{u}\|=C} Z_n(\boldsymbol{\beta}_0 + \boldsymbol{u}n^{-1/2}) > Z_n(\boldsymbol{\beta}_0)\} > 1 - \varepsilon$$

for n large enough. Then, there is a local minimizer of $Z_n(\boldsymbol{\beta})$ in $\mathscr{B}$. Under conditions (C2)–(C3), the local minimizer is the global minimizer of $Z_n(\boldsymbol{\beta})$. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n}$. We define

$$\begin{aligned}\psi_n(\boldsymbol{u}) &= n\left\{Z_n\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - Z_n(\boldsymbol{\beta}_0)\right\}\\ &= \sum_{i=1}^n\left\{Y_i\exp\left(-X_i^\top\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\right)\right) + Y_i^{-1}\exp\left(X_i^\top\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\right)\right)\right\} - \sum_{i=1}^n\Big\{Y_i\exp(-X_i^\top\boldsymbol{\beta}_0)\\ &\quad + Y_i^{-1}\exp(X_i^\top\boldsymbol{\beta}_0)\Big\} + n\lambda_n\sum_{j=1}^p\left\{\left|\omega_j\left(\beta_{j0} + \frac{u_j}{\sqrt{n}}\right)\right| - |\omega_j\beta_{j0}|\right\}\end{aligned}$$

$$
\begin{aligned}
&= W_n^\top \boldsymbol{u} + \boldsymbol{u}^\top D_n \boldsymbol{u}/2 + \frac{1}{6n^{3/2}} \sum_{i=1}^{n} \sum_{j,k,l}^{p} \Big\{ -Y_i \exp(-X_i^\top \boldsymbol{\beta}^*) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}^*) \Big\} x_{ij} x_{ik} x_{il} \\
&\quad \times u_l u_j u_k + n\lambda_n \sum_{j=1}^{p} \left\{ \left| \omega_j \left(\beta_{j0} + \frac{u_j}{\sqrt{n}} \right) \right| - \left| \omega_j \beta_{j0} \right| \right\} \\
&\equiv I_1 + I_2 + I_3 + I_4,
\end{aligned}
$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n}$, and

$$
W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -Y_i \exp(-X_i^\top \boldsymbol{\beta}_0) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}_0) \right\} X_i,
$$

$$
D_n = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \exp(-X_i^\top \boldsymbol{\beta}_0) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}_0) \right\} X_i X_i^\top .
$$

It follows directly from conditions (C1) and (C5) that $W_n \to W$ in distribution, where $W \sim N(0, V)$. Moreover, under condition (C3), $D_n \to D$ in probability. Hence,

$$
\|I_1\| = O_p(1)\|\boldsymbol{u}\|
$$

and

$$
\|I_2\| = \boldsymbol{u}^\top D\boldsymbol{u}/2 + o_p(1)\|\boldsymbol{u}\|^2.
$$

Next, under condition (C4), it can be shown that

$$
\|I_3\| = o_p(1)\|\boldsymbol{u}\|^2.
$$

As $\lambda_n\sqrt{n} \to \lambda_0$ and $\sqrt{n}\{|(\beta_{0j} + u_j/\sqrt{n})| - |\beta_{0j}|\} \to \{u_j(\beta_{0j})I(\beta_{0j} \neq 0) + |u_j|I(\beta_{0j} = 0)\}$, we have

$$
I_4 > \lambda_n\sqrt{n} \sum_{j=1}^{s} u_j \mathrm{sgn}(\beta_{j0})\omega_j = O(1)\|\boldsymbol{u}\|.
$$

Thereby, $\psi_n(\boldsymbol{u})$ is positive for sufficiently large C. As a result, we have shown $\|\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

Given that $\lambda_n\sqrt{n}\omega_j \to 0$ in probability $\forall j \in \mathcal{A}_\mathcal{T}$ and $\lambda_n\sqrt{n}\omega_j = \lambda_n\sqrt{n}n^\gamma(\sqrt{n}\widetilde{\beta}_{nj})^{-\gamma} \to \infty$ for any $j \in \mathcal{A}_\mathcal{T}{}^c$, we have $\psi_n(u) \xrightarrow{d} \psi(u)$, where

$$
\psi(\boldsymbol{u}) = \begin{cases} W_{\mathcal{A}_\mathcal{T}}^\top \boldsymbol{u}_{\mathcal{A}_\mathcal{T}} + \boldsymbol{u}_{\mathcal{A}_\mathcal{T}}^\top D_{\mathcal{A}_\mathcal{T}} \boldsymbol{u}_{\mathcal{A}_\mathcal{T}}/2 & \text{for} \quad u_j = 0 \quad \forall j \in \mathcal{A}_\mathcal{T}; \\ \infty & \text{otherwise} \end{cases}
$$

by the Slutsky's theorem. Together with condition (C3) and Geyer (1994), it can be shown that $\sqrt{n}\{\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_0\} \to \arg\min(\psi(\boldsymbol{u}))$ in distribution. Denote $\boldsymbol{u}^* \equiv \arg\min(\psi(\boldsymbol{u}))$. Note that $D_{\mathcal{A}_\mathcal{T}}^{-1} W_{\mathcal{A}_\mathcal{T}} \sim N(0, D_{\mathcal{A}_\mathcal{T}}^{-1} V_{\mathcal{A}_\mathcal{T}} D_{\mathcal{A}_\mathcal{T}}^{-1})$. Hence, $\boldsymbol{u}_{\mathcal{A}_\mathcal{T}}^* \to N(0, D_{\mathcal{A}_\mathcal{T}}^{-1} V_{\mathcal{A}_\mathcal{T}} D_{\mathcal{A}_\mathcal{T}}^{-1})$ in distribution and $\boldsymbol{u}_{\mathcal{A}_\mathcal{T}{}^c} \to 0$ in probability as $n \to \infty$.

Now we are in a position to show the model selection consistency. For any $j \in \mathcal{A}_\mathcal{T}$, it follows from the asymptotic normality that $\hat{\beta}_{nj}^* \to \beta_{0j}$ in probability, which implies that $P(j \in \mathcal{A}_n^*) \to 1$, namely $P(\mathcal{A}_\mathcal{T} \subseteqq \mathcal{A}_n^*) \to 1$. We then only need to prove $P(\mathcal{A}_\mathcal{T}{}^c \subseteqq \mathcal{A}_n^{*c}) \to 1$. Namely, $P(j' \in \mathcal{A}_n^*) \to 0$ for any $j' \in \mathcal{A}_\mathcal{T}{}^c$. For any $j' \in \mathcal{A}_n^*$, by the KKT condition, we have

$$
\begin{aligned}
&W_{nj'} + (D_n\boldsymbol{u})_{j'} + \lambda_n\sqrt{n}\omega_{j'}\mathrm{sgn}(u_{j'}) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -Y_i \exp(-X_i^\top \boldsymbol{\beta}_0) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}_0) \right\} x_{ij'} \\
&\qquad + \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \exp(-X_i^\top \boldsymbol{\beta}_0) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}_0) \right\} x_{ij'} X_i^\top \boldsymbol{u} + \lambda_n\sqrt{n}\omega_{j'}\mathrm{sgn}(u_{j'}) \\
&\quad = 0.
\end{aligned}
$$

By the Central Limit Theorem, the law of large numbers and the Slutsky's theorem, we have

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{-Y_i \exp(-X_i^\top \boldsymbol{\beta}_0) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}_0)\} x_{ij'} + \frac{1}{n} \sum_{i=1}^{n} \{Y_i \exp(-X_i^\top \boldsymbol{\beta}_0) + Y_i^{-1} \exp(X_i^\top \boldsymbol{\beta}_0)\} x_{ij'} X_i^\top \boldsymbol{u} \to Z,
$$

in distribution, where Z is a normal random variable. It follows from $\lambda_n\sqrt{n}\omega_{j'}\mathrm{sgn}(u_{j'}) \to \pm\infty$ in probability that $P(j' \in \mathcal{A}_n^*) = P(Z = \pm\infty) \to 0$. The proof of Theorem 1 is complete.

Proof of Lemma 1. Denote $\mathscr{B} \equiv \{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_{n0} + \boldsymbol{u}\alpha_n, \|\boldsymbol{u}\| \leq C\}$ where $\alpha_n = p_n^{1/2}n^{-1/2}$. To show $\|\hat{\boldsymbol{\beta}}_n^\star - \boldsymbol{\beta}_{n0}\| = O_p(\sqrt{p_n/n})$, it suffices to show that for any $\varepsilon > 0$, there exists a sufficiently large constant C such that

$$P\left\{\sup_{\|\boldsymbol{u}\|=C} Z_n(\boldsymbol{\beta}_{n0} + \boldsymbol{u}\alpha_n) > AQ_n(\boldsymbol{\beta}_{n0})\right\} > 1 - \varepsilon$$

for n large enough. Similar to the arguments in the proof of Theorem 1, the local minimizer of $Z_n(\boldsymbol{\beta}_n)$ in $\mathscr{B}$ is the global minimizer of $Z_n(\boldsymbol{\beta}_n)$. Denote $\boldsymbol{\beta} = \boldsymbol{\beta}_{n0} + \boldsymbol{u}\alpha_n$. Write

$$\begin{aligned}
\tau_n(u) &= n\{Z_n(\boldsymbol{\beta}_{n0} + \boldsymbol{u}\alpha_n) - Z_n(\boldsymbol{\beta}_{n0})\} \\
&= \sum_{i=1}^{n}\{Y_i \exp(-X_i^\top(\boldsymbol{\beta}_{n0} + \boldsymbol{u}\alpha_n)) + Y_i^{-1}\exp(X_i^\top(\boldsymbol{\beta}_{n0} + \boldsymbol{u}\alpha_n))\} \\
&\quad - \{Y_i \exp(-X_i^\top \boldsymbol{\beta}_{n0}) + Y_i^{-1}\exp(X_i^\top \boldsymbol{\beta}_{n0})\} + n\lambda_n \sum_{j=1}^{p_n}\left\{\left|\omega_j(\beta_{nj0} + u_j\alpha_n)\right| - \left|\omega_j\beta_{nj0}\right|\right\} \\
&\geq \sqrt{n}\alpha_n W_n^\top \boldsymbol{u} + n\alpha_n^2\boldsymbol{u}^\top D_n\boldsymbol{u}/2 + \frac{\alpha_n^3}{6}\sum_{i=1}^{n}\sum_{j,k,l}^{p_n}\{-Y_i\exp(-X_i^\top\boldsymbol{\beta}^*) + Y_i^{-1}\exp(X_i^\top\boldsymbol{\beta}^*)\} \\
&\quad \times x_{ij}x_{ik}x_{il}u_ju_lu_k + n\lambda_n\sum_{j=1}^{s_n}\left\{\left|\omega_j(\beta_{nj0} + u_j\alpha_n)\right| - \left|\omega_j\beta_{nj0}\right|\right\} \\
&\equiv I_1 + I_2 + I_3 + I_4,
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_{n0} + \boldsymbol{u}\alpha_n$ and $\boldsymbol{\beta}_{n0}$. On the other hand, by the Cauchy–Schwarz inequality, we have

$$|I_1| = \sqrt{n}\alpha_n|W_n^\top\boldsymbol{u}| \leq \sqrt{n}\alpha_n\|\boldsymbol{u}\|\,\|W_n\|.$$

Under conditions (C1) and (C5), $\|W_n\| = O_p(\sqrt{p_n})$, which implies $I_1 = O_p(\sqrt{np_n})\alpha_n\|u\|$. Next,

$$\begin{aligned}
|I_2| &= \frac{n}{2}\alpha_n^2\boldsymbol{u}^\top D_n\boldsymbol{u} \\
&= \frac{n}{2}\alpha_n^2\boldsymbol{u}^\top D\boldsymbol{u} + \frac{n}{2}\alpha_n^2\boldsymbol{u}^\top(D_n - D)\boldsymbol{u} \\
&\equiv I_{21} + I_{22}.
\end{aligned}$$

Given $p_n^4/n \to 0$ as $n \to \infty$ and condition (C5), we have

$$\begin{aligned}
P\left(\|D_n - D\| > \frac{\varepsilon}{p_n}\right) &\leq \frac{p_n^2}{\varepsilon^2}E\|D_n - D\|^2 \\
&\leq \frac{p_n^2}{n^2\varepsilon^2}E\sum_{i=1}^{n}\sum_{j=1}^{p_n}\sum_{k=1}^{p_n}\left[x_{ij}x_{ik}(\varepsilon_i + \varepsilon_i^{-1}) - E\{x_{ij}x_{ik}(\varepsilon_i + \varepsilon_i^{-1})\}\right]^2 \\
&= \frac{p_n^4}{n\varepsilon^2} = o(1).
\end{aligned}$$

Thereby, $I_{22} = o_p(1/p_n)n\alpha_n^2\|\boldsymbol{u}\|$. In view of $I_{21} = n\alpha_n^2\boldsymbol{u}^\top D\boldsymbol{u}/2$, we obtain $I_2 = n\alpha_n^2\boldsymbol{u}^\top D\boldsymbol{u}/2\{1 + o_p(1)\}$. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
|I_3| &= \left|\frac{\alpha_n^3}{6}\sum_{i=1}^{n}\sum_{j,k,l}^{p_n}\{-Y_i\exp(-X_i^\top\boldsymbol{\beta}^*) + Y_i^{-1}\exp(X_i^\top\boldsymbol{\beta}^*)\}x_{ij}x_{ik}x_{il}u_lu_ju_k\right| \\
&\leq \frac{\alpha_n^3}{6}\left(\sum_{j,k,l}^{p_n}\left[\sum_{i=1}^{n}x_{ij}x_{ik}x_{il}\left\{-Y_i\exp(-X_i^\top\boldsymbol{\beta}_0^*) + Y_i^{-1}\exp(X_i^\top\boldsymbol{\beta}_0^*)\right\}\right]^2\right)^{1/2}\|\boldsymbol{u}\|^3.
\end{aligned}$$

Together with condition (C6), we have $|I_3| = O_p(p_n^{3/2}\alpha_n)n\alpha_n^2\|\boldsymbol{u}\|^2$. Lastly, by the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
|I_4| &= n\lambda_n\sum_{j=1}^{s_n}\left\{\left|\omega_j(\beta_{nj0} + u_j\alpha_n)\right| - \left|\omega_j(\beta_{nj0})\right|\right\} \\
&= n\lambda_n\alpha_n\sum_{j=1}^{s_n}u_j\mathrm{sgn}(\beta_{n0j})\omega_{nj}
\end{aligned}$$

$$\begin{aligned}&\le n\lambda_n\alpha_n\|\boldsymbol{u}\|\,\|\boldsymbol{\omega}_{ns_n}\|\\ &= O(n\lambda_n s_n^{1/2}\alpha_n\|\boldsymbol{u}\|d^{-\gamma}).\end{aligned}$$

Therefore, when C is sufficiently large, $\tau_n(\boldsymbol{u}) > 0$. The proof of Lemma 1 is complete.

Proof of Theorem 2. We first prove part (1) in Theorem 2. Without loss of generality, we assume the active set $\mathcal{A}_{\mathcal{T}} = \{1, 2, \ldots, s_n\}$. Then we can write $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{n_1}^\top, \boldsymbol{\beta}_{n_2}^\top)^\top$ $\boldsymbol{\beta}_{n0} = (\boldsymbol{\beta}_{n_10}^\top, \boldsymbol{\beta}_{n_20}^\top)^\top$. In order to prove the sparsity, it suffices to show $Z_n((\boldsymbol{\beta}_{n_1}^\top, \mathbf{0})^\top) = \min_{\|\boldsymbol{\beta}_{n_2}\|\le\varepsilon_n} Z_n((\boldsymbol{\beta}_{n_1}^\top, \boldsymbol{\beta}_{n_2}^\top)^\top)$, in which $\|\boldsymbol{\beta}_{n_1} - \boldsymbol{\beta}_{n_10}\| = O_p(\sqrt{p_n/n})$, $\varepsilon_n = C\sqrt{p_n/n}$ and C is a positive constant. To this end, we need to show for any $j \in \mathcal{A}_{\mathcal{T}}{}^c$,

$$\begin{aligned}&\frac{\partial Z_n(\boldsymbol{\beta}_n)}{\partial\beta_{n_2j}} > 0 \quad \text{for } -\varepsilon_n < \beta_{n_2j} < 0;\\ &\frac{\partial Z_n(\boldsymbol{\beta}_n)}{\partial\beta_{n_2j}} < 0 \quad \text{for } 0 < \beta_{n_2j} < \varepsilon_n.\end{aligned}\tag{5}$$

By a Taylor expansion,

$$\begin{aligned}n\frac{\partial Z_n(\boldsymbol{\beta}_n)}{\partial\beta_{n_2j}} &= \sum_{i=1}^n x_{ij}(-\varepsilon_i + \varepsilon_i^{-1}) + \sum_{l=1}^{p_n}\sum_{i=1}^n x_{ij}x_{il}(\varepsilon_i + \varepsilon_i^{-1})(\beta_{nl} - \beta_{n0l}) + \frac{1}{2}\sum_{l,k}^{p_n}\sum_{i=1}^n x_{ij}x_{il}x_{ik}\\ &\quad\times\left\{-Y_i\exp(-X_i^\top\boldsymbol{\beta}_n^*) + Y_i^{-1}\exp(X_i^\top\boldsymbol{\beta}_n^*)\right\}(\beta_{nl} - \beta_{n0l})(\beta_{nk} - \beta_{n0k}) + n\lambda_n\omega_j\text{sgn}(\beta_{n_2j}).\end{aligned}$$

It can be shown along similar lines of Lemma 1 that

$$\begin{aligned}n\frac{\partial Z_n(\boldsymbol{\beta}_n)}{\partial\beta_{n_2j}} &= O_p(\sqrt{np_n}) + n\lambda_n\omega_j\text{sgn}(\beta_{n_2j})\\ &= n\lambda_n\omega_j\{\sqrt{np_n}/(n\lambda_n w_j) + \text{sgn}(\beta_{n_2j})\}.\end{aligned}$$

Under the assumptions that $\lambda_n n^{1/2}r_n^\gamma/p_n^{1/2} \to \infty$ as $n \to \infty$, $\omega_j = O_p(r_n^\gamma)$ and $\sqrt{np_n}/(n\lambda_n w_j) = o(1)$, we have $\partial Z_n(\boldsymbol{\beta}_n)/\partial\beta_{n_2j} = n\lambda_n\omega_j\{o(1) + \text{sgn}(\beta_{n_2j})\}$. Hence, part (1) is proved.

Next, we wish to show part (2). With a slight abuse of notation, we let $Z_n(\boldsymbol{\beta}_{n_1}) = Z_n((\boldsymbol{\beta}_{n_1}^\top, \mathbf{0}^\top)^\top)$. In view of the fact that $\boldsymbol{\beta}_{n_1}$ should satisfy $\nabla Z_n(\boldsymbol{\beta}_{n_1}) = \mathbf{0}$, by a Taylor expansion of $\nabla Z_n(\boldsymbol{\beta}_{n_1})$ at $\boldsymbol{\beta}_{n_10}$, we have

$$\begin{aligned}n\left\{\frac{\partial Z_n(\boldsymbol{\beta}_{n_1})}{\partial\boldsymbol{\beta}_{n_1}} - \frac{\partial Z_n(\boldsymbol{\beta}_{n_10})}{\partial\boldsymbol{\beta}_{n_10}}\right\} &= \sum_{i=1}^n X_{i\mathcal{A}_{\mathcal{T}}}X_{i\mathcal{A}_{\mathcal{T}}}^\top(\varepsilon_i + \varepsilon_i^{-1})(\boldsymbol{\beta}_{n_1} - \boldsymbol{\beta}_{n_10}) + \frac{1}{2}\sum_{i=1}^n\sum_{k,l}^{p_n} X_{i\mathcal{A}_{\mathcal{T}}}x_{ik}x_{il}\\ &\quad\times\left\{-Y_i\exp(-X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*) + Y_i^{-1}\exp(X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*)\right\}(\beta_{n_1l} - \beta_{n_10l})(\beta_{n_1k} - \beta_{n_10k}),\end{aligned}\tag{6}$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_{n1}$ and $\boldsymbol{\beta}_{n10}$. A simple manipulation of the first term of (6) yields

$$\begin{aligned}&\sum_{i=1}^n X_{i\mathcal{A}_{\mathcal{T}}}X_{i\mathcal{A}_{\mathcal{T}}}^\top(\varepsilon_i + \varepsilon_i^{-1})(\boldsymbol{\beta}_{n_1} - \boldsymbol{\beta}_{n_10})\\ &= -\sum_{i=1}^n X_{i\mathcal{A}_{\mathcal{T}}}(-\varepsilon_i + \varepsilon_i^{-1}) - n\lambda_n\boldsymbol{\omega}_{ns_n}^\top(\text{sgn}(\boldsymbol{\beta}_{n_1}) - \text{sgn}(\boldsymbol{\beta}_{n_10})) - \frac{1}{2}\sum_{i=1}^n\sum_{k,l}^{s_n} X_{i\mathcal{A}_{\mathcal{T}}}x_{ik}x_{il}\\ &\quad\times\left\{-Y_i\exp(-X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*) + Y_i^{-1}\exp(X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*)\right\}(\beta_{n_1l} - \beta_{n_10l})(\beta_{n_1k} - \beta_{n_10k})\\ &\equiv L_1 + L_2 + L_3.\end{aligned}$$

First, it follows from $\lambda_n n^{1/2}/d^\gamma \to 0$ that $L_2 = o_p(n^{1/2})$. Next, by the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}|L_3| &= \left|\frac{1}{2}\sum_{i=1}^n\sum_{k,l}^{s_n} X_{i\mathcal{A}_{\mathcal{T}}}x_{ik}x_{il}\left\{-Y_i\exp(-X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*) + Y_i^{-1}\exp(X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*)\right\}(\beta_{n_1l} - \beta_{n_10l})(\beta_{n_1k} - \beta_{n_10k})\right|\\ &\le \left(\sum_{j,k,l}^{s_n}\left[\sum_{i=1}^n x_{ij}x_{ik}x_{il}\left\{-Y_i\exp(-X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*) + Y_i^{-1}\exp(X_{i\mathcal{A}_{\mathcal{T}}}^\top\boldsymbol{\beta}_n^*)\right\}\right]^2\right)^{1/2}\|\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n10}\|^2\\ &= O_p(s_n^{3/2}p_n).\end{aligned}$$

Under the assumptions that $p_n^4/n = o(1)$ and $s_n = O(n^{1/6})$, it can be shown that $L_3 = o_p(n^{1/2})$. Also it can be shown along the lines of the proofs of $\|D_n - D\| = o_p(1/p_n)$ that $\|V_{n\mathcal{A}_{\mathcal{T}}} - V_{\mathcal{A}_{\mathcal{T}}}\| = o_p(1/p_n)$. Hence, we have

$$n^{1/2}A_n\widetilde{\Sigma}_n^{-1/2}(\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n10}) = A_nV_{\mathcal{A}_{\mathcal{T}}}^{-1/2}W_{n\mathcal{A}_{\mathcal{T}}} + o_p(1).$$

To prove the asymptotic normality, we need to show that $A_n V_{\mathcal{A}_\mathcal{T}}^{-1/2} W_{n\mathcal{A}_\mathcal{T}}$ satisfies the Linderberg–Feller Central Limit theorem. Let $Y_{ni} = A_n V_{\mathcal{A}_\mathcal{T}}^{-1/2} W_{ni\mathcal{A}_\mathcal{T}}$, $i = 1, 2, \ldots, n$, and $I(\cdot)$ be the indicator function. Then, for any $\varepsilon > 0$,

$$\begin{aligned}\sum_{i=1}^{n} E\left[Y_{ni}^2 I\{\|Y_{ni}\| > \varepsilon\}\right] &= nE\|Y_{n1}\|^2 I\{\|Y_{n1}\| > \varepsilon\} \\ &\leq n\left(E\|Y_{n1}\|^4\right)^{1/2} \{P(I(\|Y_{n1}\| > \varepsilon))\}^{1/2}.\end{aligned}$$

Together with condition (C2) and $A_n A_n^\top \to G$, we have

$$P(I(\|Y_{n1}\| > \varepsilon)) = \frac{E\|A_n V_{\mathcal{A}_\mathcal{T}}^{-1/2} W_{n\mathcal{A}_\mathcal{T}1}\|^2}{\varepsilon} = O(n^{-1}).$$

Moreover,

$$\begin{aligned}E\|Y_{n1}\|^4 &= E\|A_n V_{\mathcal{A}_\mathcal{T}}^{-1/2} W_{n\mathcal{A}_\mathcal{T}1}\|^4 \\ &\leq \frac{1}{n^2}\lambda_{\max}(A_n A_n^\top)^2 \lambda_{\min}(V)^{-2} E\|X_{\mathcal{A}_\mathcal{T}}^\top X_{\mathcal{A}_\mathcal{T}}(\varepsilon_1 - \varepsilon_1^{-1})^2\|^2 \\ &= O_p(s_n^4/n^2).\end{aligned}$$

Thus, it follows that

$$\sum_{i=1}^{n} E\left\{Y_{ni}^2 I(\|Y_{ni}\| > \varepsilon)\right\} = O(n(s_n^2 n^{-1})n^{-1/2}) = o(1).$$

Note that $\sum_{i=1}^n \mathrm{cov}(Y_{ni}) \to G$ as $n \to \infty$. Therefore, it follows from the Linderberg–Feller central limit theorem that

$$n^{1/2} A_n \widetilde{\Sigma}_n^{-1/2}(\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n10}) \to N(\mathbf{0}, G).$$

The proof of Theorem 2 is complete.

References

Annergren, M., Hansson, A., Wahlberg, B., 2012. An ADMM algorithm for solving l_1 regularized MPC. In: Proc. IEEE 51st Conf. Decision Control, pp. 4486–4491.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 31–122.

Chen, S., Donoho, D., 1994. Basis pursuit. In: 28th Asilomar Conf. Signals, Systems and Computers, Asilomar.

Chen, K., Guo, S., Lin, Y., Ying, Z., 2010. Least absolute relative error estimation. J. Amer. Statist. Assoc. 105, 1104–1112.

Chen, K., Lin, Y., Wang, Z., Ying, Z., 2016. Least product relative error estimation. J. Multivariate Anal. 144, 91–98.

Demongeot, J., Hamie, A., Laksaci, A., Rachdi, M., 2015. Relative-error prediction in nonparametric functional statistics: Theory and practice. J. Multivariate Anal..

Fan, J., Feng, Y., Wu, Y., 2009. Network exploration via the adaptive lasso and SCAD penalties. Ann. Appl. Stat. 3, 521–541.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928–961.

Geyer, C.J., 1994. On the asymptotics of constrained M-estimation. Ann. Statist. 22, 1993–2010.

Gneiting, T., 2011. Making and evaluating point forecasts. J. Amer. Statist. Assoc. 106, 746–762.

Hastie, T., Tibshirani, R., Friedman, J., 2011. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Huang, J., Horowitz, J.L., Ma, S., 2008a. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist. 587–613.

Huang, J., Ma, S., Li, H., Zhang, C., 2011. The sparse Laplacian shrinkage estimator for high-dimensional regression. Ann. Statist. 39 (4), 2021–2046.

Huang, J., Ma, S., Zhang, C., 2008b. Adaptive lasso for sparse high-dimensional regression models. Statist. Sinica 18, 1603–1618.

Khoshgoftaar, T.M., Bhattacharyya, B.B., Richardson, G.D., 1992. Predicting software errors, during development, using nonlinear regression models: A comparative study. IEEE Trans. Reliab. 41, 390–395.

Kolassa, S., Martin, R., 2011. Percentage errors can ruin your day (and rolling the dice show how). Foresight: Int. J. Appl. Forecast. 23, 21–27.

Makridakis, S., 1993. Accuracy measures: theoretical and practical concerns. Int. J. Forecast. 9, 527–529.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R., 1984. The Forecasting Accuracy of Major Time Series Methods. Wiley, New York.

Narula, S.C., Wellington, J.F., 1977. Prediction, linear regression and the minimum sum of relative errors. Technometrics 19, 185–190.

Park, H., Stefanski, L.A., 1998. Relative-error prediction. Statist. Probab. Lett. 40, 227–236.

Pötscher, B.M., Schneider, U., 2009. On the distribution of the adaptive lasso estimator. J. Statist. Plann. Inference 139, 2775–2790.

Shen, X., Ye, J., 2002. Adaptive model selection. J. Amer. Statist. Assoc. 97, 210–221.

Shi, W., Ling, Q., Yuan, K., Wu, G., Yin, W., 2014. On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans. Signal Process. 62, 1750–1761.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.

Tofallis, C., 2014. A better measure of relative prediction accuracy for model selection and model estimation. J. Oper. Res. Soc. 103, 1–11.

Wahlberg, B., Boyd, S., Annergren, M., Wang, Y., 2012. An ADMM algorithm for a class of total variation regularized estimation problems. In: Proc. 16th IFAC Symp. Syst. Ident., vol. 16, pp. 83–88.

Wang, H., Leng, C., 2007. Unified lasso estimation by least squares approximation. J. Amer. Statist. Assoc. 102, 1418–1429.

Wang, H., Li, G., Jiang, G., 2007a. Robust regression shrinkage and consistent variable selection through the LAD-lasso. J. Bus. Econom. Statist. 25, 347–355.

Wang, H., Li, B., Leng, C., 2009. Shrinkage tuning parameter selection with a diverging number of parameters. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 671–683.

Wang, H., Li, R., Tsai, C.L., 2007b. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553–568.

Wu, Y., Boos, D.D., Stefanski, L.A., 2007. Controlling variable selection by the addition of pseudovariables. J. Amer. Statist. Assoc. 102, 235–243.

Xu, J., Ying, Z., 2010. Simultaneous estimation and variable selection in median regression using Lasso-type penalty. Ann. Inst. Statist. Math. 62, 487–514.

Zhang, H.H., Lu, W., 2007. Adaptive lasso for Cox's proportional hazards model. Biometrika 94, 691–703.

Zhang, Q., Wang, Q., 2013. Local least absolute relative error estimating approach for partially linear multiplicative model. Statist. Sinica 23, 1091–1116.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.