Journal of Econometrics 235 (2023) 166–179

Contents lists available at ScienceDirect

Journal of Econometrics

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

Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to COVID-19 pandemic

Xu Guo [a], Runze Li [b], Jingyuan Liu [c,*], Mudong Zeng [b]

[a] School of Statistics, Beijing Normal University, Beijing, 100875, China
[b] Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA
[c] MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Lab of Statistics, Xiamen University, Xiamen, 361000, China

ARTICLE INFO

Article history:
Received 2 June 2021
Received in revised form 31 January 2022
Accepted 4 March 2022
Available online 8 April 2022

JEL classification:
C12
C13

Keywords:
Mediation analysis
Penalized least squares
Sparsity
Wald test

ABSTRACT

Mediation analysis draws increasing attention in many research areas such as economics, finance and social sciences. In this paper, we propose new statistical inference procedures for high dimensional mediation models, in which both the outcome model and the mediator model are linear with high dimensional mediators. Traditional procedures for mediation analysis cannot be used to make statistical inference for high dimensional linear mediation models due to high-dimensionality of the mediators. We propose an estimation procedure for the indirect effects of the models via a partially penalized least squares method, and further establish its theoretical properties. We further develop a partially penalized Wald test on the indirect effects, and prove that the proposed test has a χ^2 limiting null distribution. We also propose an F-type test for direct effects and show that the proposed test asymptotically follows a χ^2-distribution under null hypothesis and a noncentral χ^2-distribution under local alternatives. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed tests and compare their performance with existing ones. We further apply the newly proposed statistical inference procedures to study stock reaction to COVID-19 pandemic via an empirical analysis of studying the mediation effects of financial metrics that bridge company's sector and stock return.

1. Introduction

Since the seminal work of Baron and Kenny (1986), mediation analysis has been used in various scientific research, such as economics and finance (Conti et al., 2016; Chernozhukov et al., 2021). It is designed to investigate the mechanisms how exposure variables affect an outcome through intermediate variables, which are termed as mediators. Numerous inference procedures for such mediation models have been studied in both statistic and econometric fields. For instance, in economic policy evaluation, while there certainly is no shortage of techniques assessing effects of policies or other treatments on an outcome (Imbens, 2004; Donald and Hsu, 2014; Abadie and Cattaneo, 2018), mediation analyses move a step further to disentangle such effect into indirect effects through mediators, such as certain economic indices, and direct effects (Celli, 2022). Heckman and Pinto (2015) conducted an econometric mediation analysis with unmeasured and mismeasured

* Corresponding author.
E-mail addresses: xustat12@bnu.edu.cn (X. Guo), rzli@psu.edu (R. Li), jingyuan@xmu.edu.cn (J. Liu), muz149@psu.edu (M. Zeng).

https://doi.org/10.1016/j.jeconom.2022.03.001
0304-4076/

exposure variables. Huber and Frölich (2017) discussed the nonparametric identification of causal direct and indirect effects of a binary treatment based on instrumental variables. See Huber (2020) and Celli (2022) for a comprehensive overview of mediation analysis in economics and econometrics.

On account of modern data-collecting technology, mediation analysis extends its territory to quantitative finance, genomics, internet analysis, biomedical research, among other data-intensive fields. This brings in high-dimensional mediators and requires attention on high-dimensional mediation model (HDMM), where the number of potential mediators is much larger than the sample size. This work is motivated by such a high-dimensional mediation structure when studying the effects of company's belonging sector on stock return via influencing various financial metrics during the COVID-19 period. Direct effects of sectors, as well as financial statements, on stock performance have been extensively studied in literature. See, for instance, Fama and French (1993, 2015), Graham et al. (2002) and Khan et al. (2015). Yet as to be evidently shown by the empirical analysis in Section 3.2, the companies' belonging sectors also significantly affect stock returns indirectly through certain financial metrics in the statements. In our analysis, 550 financial indexes are involved, based on only 490 companies, resulting in high dimensional mediators.

The high-dimensionality, on every account, poses both computational and statistical challenges for carrying out efficient mediation analysis. For instance, the traditional structural equation modeling fails due to the rank-deficiency of the observed covariance matrix. However, notwithstanding the high dimensional mediation structure, the number of truly active mediators is typically assumed small and less than the sample size. This is referred to as the sparsity assumption in the literature, although the sparsity pattern is unknown and thus to be recovered. See, for example, Fan et al. (2020b) and references therein.

Recently, debiased Lasso has been advocated to deal with bias correction and make valid inference for high dimensional data (Zhang and Zhang, 2014; Van de Geer et al., 2014). Cattaneo et al. (2018) developed inference methods for high dimensional linear regression models with heteroscedasticity and the number of included covariates growing as fast as the sample size. In addition, there are other strands of literature focusing on linear regressions with increasing dimensions (Cattaneo et al., 2019; Galbraith and Zinde-Walsh, 2020; Fan et al., 2020a,c). Belloni et al. (2014, 2017), Farrell (2015) and others investigated the inference problem about the average treatment effect in high dimensions. Chernozhukov et al. (2015) provided a general approach based on the idea of orthogonalization. To apply the debiased Lasso to mediation analysis, Zhou et al. (2020) introduced debiased penalized estimators for the direct and indirect effects, with theoretical guarantees of the related tests. However, their method involves estimating high dimensional matrices, leading to potentially unstable estimates and expensive computation. Imposing penalization on all parameters reduces the efficiency of estimators, and hence tests. Wang et al. (2020) systematically discussed the efficiency loss of the debiased methods and presented a thorough comparison among different inference methods.

In this paper, we propose new statistical inference procedures for HDMM. However, there are much less work on statistical inference for HDMM. To our best knowledge, Zhou et al. (2020) is the only one on testing hypothesis on indirect effect with rigorous theoretical analysis. Our inference procedure on indirect effect is distinguished from Zhou et al. (2020) in that we observe the indirect effect in HDMM indeed is a low dimensional parameter and is the difference between the total effect and the direct effect in the HDMM. This motivates us to estimate the total effect via least squares method and the direct effect by partially penalized least squares method, and then estimate the indirect effect by the difference between the estimates of the total effect and the direct effect. We establish the asymptotical normality of the indirect effect estimate and further develop a Wald test for the indirect effect.

We estimate the direct effect in the HDMM by partially penalized least squares method, and propose an F-type test for it. The statistical inference on the direct effect essentially is the same as statistical inference on low dimensional coefficients in high-dimensional linear models. This topic has been studied under the setting in which the covariate vector in the high-dimensional linear models is fixed design (Zhang and Zhang, 2014; Van de Geer et al., 2014; Shi et al., 2019). Due to the nature of HDMM, the design matrix in HDMM must be random rather than fixed since mediators are random. Thus, the statistical setting studied in this paper is different from the one in Shi et al. (2019), in which the covariate vector is assumed to be fixed design. We study the asymptotical property of the proposed estimator in the random-design setting. The random design imposes challenges in deriving the rate of convergence and asymptotical normality of the partially penalized least squares estimates. Under mild regularity conditions, we prove the sparsity and establish the rate of convergence of the partially penalized least squares estimate. We further establish an asymptotical representation of the estimate. Based on the asymptotical representation, we can further derive the asymptotical normality of the estimate and derive the asymptotical distributions of the proposed test for the direct effect under null hypothesis and under local alternative.

We show that the proposed estimate of indirect effect is asymptotically more efficient than the one proposed in Zhou et al. (2020). This is because the debias step of the debiased Lasso inflates the asymptotical variance of the resulting estimate. We conduct Monte Carlo simulation studies to assess the finite sample performance of the proposed estimate in terms of bias and variance and to examine Type I error and power of the proposed test. We also conduct numerical comparisons among the proposed estimate, the oracle estimate and the estimate proposed in Zhou et al. (2020). Our numerical comparison indicates that the proposed estimate performs as well as the oracle one, and outperforms the estimate proposed by Zhou et al. (2020).

We utilize the proposed method to study the mediator role of financial metrics that bridge company's sector and stock return. Our proposed procedure selects six financial metrics out of all the 550 that indeed mediate the pathways linking

company sector and stock return, with interestingly and informatively financial interpretations. We also compare the metrics selected using our data during the COVID-19 period and those classical findings in existing works, including Fama and French (2015) and Edirisinghe and Zhang (2008), among others. We indeed discover some unique patterns and features due to the pandemic. Moreover, according to the proposed tests for effects of sector, both its direct effect and indirect effect via financial metrics are statistically significant. Therefore, evaluating the selected financial metrics, as well as the sector information, might help investors to make wiser investment decisions and choose stocks especially during the pandemic.

The rest of this paper is organized as follows. In Section 2, we develop statistical inference procedures for the indirect and direct effects and establish its theoretical properties. Section 3 presents numerical studies and a real data example. Conclusion and discussion are given in Section 4. All proofs are presented in the supplement of this paper.

2. Tests of hypotheses on indirect and direct effects

Consider the mediation model

$$y = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_1^T \boldsymbol{x} + \varepsilon_1, \tag{2.1}$$

$$\boldsymbol{m} = \Gamma^T \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where y is the outcome, $\boldsymbol{m}$ is the p-dimensional mediator, $\boldsymbol{x}$ is the q-dimensional exposure variable, and a^T denotes the transpose of a. We in this paper assume p is high dimensional, while q is fixed and finite. Correspondingly, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are p- and q-dimensional regression coefficient vectors, and Γ is a $q \times p$ coefficient matrix. For instance, in quantitative finance, the sectors of different companies result in different financial metrics, and subsequently affect stock returns. Therefore, one could adopt model (2.1) and (2.2) to study the mediation effects of financial metrics that bridge company's sector and stock return, where the outcome y is stock return, the exposure variable $\boldsymbol{x}$ are the indicators of companies' sectors, and the mediators $\boldsymbol{m}$ are financial metrics. For more examples, see Zhou et al. (2020), Huber (2020) and Celli (2022).

Following the literature on high-dimensional mediation model (Zhou et al., 2020), we impose a sparsity assumption that only a small proportion of entries in $\boldsymbol{\alpha}_0$ are nonzero. This implies that the corresponding variables in $\boldsymbol{m}$ are actually relevant to y. Notably, from Eq. (2.2), $\boldsymbol{m}$ must be random. We further assume that ε_1 and $\boldsymbol{\varepsilon}$ are independent random errors with $\text{var}(\varepsilon_1) = \sigma_1^2$ and $\text{cov}(\boldsymbol{\varepsilon}) = \Sigma^*$; ε_1 is independent of $\boldsymbol{m}$, $\boldsymbol{x}$, and $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{x}$.

Plugging (2.2) into (2.1) yields

$$y = (\boldsymbol{\beta} + \boldsymbol{\alpha}_1)^T \boldsymbol{x} + \varepsilon_1 + \varepsilon_2 = \boldsymbol{\gamma}^T \boldsymbol{x} + \varepsilon_3, \tag{2.3}$$

where $\boldsymbol{\beta} = \Gamma \boldsymbol{\alpha}_0$, $\varepsilon_2 = \boldsymbol{\alpha}_0^T \boldsymbol{\varepsilon}$ with $\text{var}(\varepsilon_2) = \sigma_2^2 = \boldsymbol{\alpha}_0^T \Sigma^* \boldsymbol{\alpha}_0$, $\boldsymbol{\gamma} = \boldsymbol{\beta} + \boldsymbol{\alpha}_1$, and $\varepsilon_3 = \varepsilon_1 + \varepsilon_2$ is the total random error. Following the literature (Imai et al., 2010), we refer $\boldsymbol{\beta}$ to the indirect effect of $\boldsymbol{x}$ on y mediated by $\boldsymbol{m}$, $\boldsymbol{\alpha}_1$ to the direct effect, and $\boldsymbol{\gamma} = \boldsymbol{\alpha}_1 + \boldsymbol{\beta}$ to the total effect. A causal interpretation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_1$ is briefly discussed in the Appendix.

2.1. Estimating indirect and direct effects

In practice, of interest is to test whether there exists significant (joint) indirect effect or not. This can be formulated as the following hypothesis testing problem

$$H_0 : \boldsymbol{\beta} = 0 \text{ versus } H_1 : \boldsymbol{\beta} \neq 0. \tag{2.4}$$

When both p and q are finite-dimensional, $\boldsymbol{\beta}$ can be estimated through $\hat{\boldsymbol{\beta}} = \widehat{\Gamma} \hat{\boldsymbol{\alpha}}_0$, where $\widehat{\Gamma}$ and $\hat{\boldsymbol{\alpha}}_0$ are $\sqrt{n}$-consistently estimated from models (2.1) and (2.2). That is, $\widehat{\Gamma} = \Gamma + \boldsymbol{E}_\gamma$ and $\hat{\boldsymbol{\alpha}}_0 = \boldsymbol{\alpha}_0 + \boldsymbol{e}_\alpha$, where $\boldsymbol{E}_\gamma = O_P(1/\sqrt{n})$ and $\boldsymbol{e}_\alpha = O_P(1/\sqrt{n})$ are estimation errors. Then

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \leq \|\Gamma \boldsymbol{e}_\alpha\| + \|\boldsymbol{E}_\gamma \boldsymbol{\alpha}_0\| + \|\boldsymbol{E}_\gamma \boldsymbol{e}_\alpha\| = O_P(1/\sqrt{n}), \tag{2.5}$$

where $\|\cdot\|$ stands for the Euclidean norm.

When p is high-dimensional, however, the right-hand side of (2.5) is no longer $O_P(1/\sqrt{n})$. This results in potentially non-ignorable estimation error of $\hat{\boldsymbol{\beta}}$. Moreover, $\boldsymbol{\beta}$ is challenging to be estimated through $\Gamma \boldsymbol{\alpha}_0$ as it involves estimation of a high-dimensional matrix and a high-dimensional vector, though, interestingly, $\boldsymbol{\beta} = \Gamma \boldsymbol{\alpha}_0$ is q-dimensional, fixed and finite.

As a key observation from (2.3), the indirect effect $\boldsymbol{\beta} = \boldsymbol{\gamma} - \boldsymbol{\alpha}_1$, is the difference between the total effect and direct effect. This motivates us to estimate $\boldsymbol{\beta}$ by separately estimating $\boldsymbol{\gamma}$ via (2.3) and $\boldsymbol{\alpha}_1$ via (2.1), respectively, rather than estimating the high-dimensional Γ and $\boldsymbol{\alpha}_0$.

Suppose that $\{\boldsymbol{m}_i, \boldsymbol{x}_i, y_i\}$, $i = 1, \ldots, n$ is a random sample from (2.1) and (2.2). Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. Then we estimate $\boldsymbol{\gamma}$ by its least squares estimate

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{2.6}$$

While for the estimator of $\boldsymbol{\alpha}_1$, due to the high-dimensionality of $\boldsymbol{\alpha}_0$, we propose the following partially penalized least squares method:

$$(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_0) = \arg\min_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|), \tag{2.7}$$

where $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)^T$ and $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. The regularization is only applied to the high-dimensional yet sparse $\boldsymbol{\alpha}_0$. We opt not penalize $\boldsymbol{\alpha}_1$ to achieve local power on the direct effect $\boldsymbol{\alpha}_1$ and the indirect effect $\boldsymbol{\beta}$ under local alternatives. See Theorem 2 and Corollary 1 for more details. Thus, our proposal is different from Zhou et al. (2020), which is to develop a debiased estimator of $\tilde{\Sigma}_{XM}\boldsymbol{\alpha}_0$ with $\tilde{\Sigma}_{XM} = E[\boldsymbol{x}\boldsymbol{m}^T]$ rather than estimator of $\boldsymbol{\alpha}_0$ or $\boldsymbol{\beta}$. As a result, the proposal of Zhou et al. (2020) may lead to less efficient estimators due to debiasing. This will be discussed in the next subsection.

2.2. Theoretical results

In this section, we investigate statistical properties of the proposed estimators. We first present some notations and assumptions. For the penalty function, it is assumed that $p_\lambda(t_0)$ is increasing and concave in $t_0 \in [0, \infty)$, and has a continuous derivative $p'_\lambda(t_0)$ with $p'_\lambda(0+) > 0$. Denote $\rho(t_0, \lambda) = p_\lambda(t_0)/\lambda$ for $\lambda > 0$. Further, $\rho'(t_0, \lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0+, \lambda)$ does not depend on λ. Define $\bar{\rho}(\boldsymbol{v}, \lambda) = \{\mathrm{sgn}(v_1)\rho'(|v_1|, \lambda), \ldots, \mathrm{sgn}(v_l)\rho'(|v_l|, \lambda)\}^T$ for a vector $\boldsymbol{v} = (v_1, \ldots, v_l)^T$, where $\mathrm{sgn}(\cdot)$ is the sign function. Define the local concavity of $\rho(\cdot)$ at $\boldsymbol{v}$ as

$$\kappa(\rho, \boldsymbol{v}, \lambda) = \lim_{\epsilon \to 0^+} \max_{1 \le j \le l} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} -\frac{\rho'(t_2, \lambda) - \rho'(t_1, \lambda)}{t_2 - t_1}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_0^T)^T$ and $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_1^{\star T}, \boldsymbol{\alpha}_0^{\star T})^T$, the true value of $\boldsymbol{\theta}$. Further let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}_1^T, \hat{\boldsymbol{\alpha}}_0^T)$ be the estimator of $\boldsymbol{\theta}_0$. Denote $\mathcal{A} = \{j : \alpha_{0j}^\star \neq 0\}$, and $s = |\mathcal{A}|$ is the number of elements in $\mathcal{A}$. Moreover, $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_{0,\mathcal{A}}^T)^T$. And $\boldsymbol{\vartheta}_0, \hat{\boldsymbol{\vartheta}}$ are similarly defined. Let $\boldsymbol{M}^j$ denote the jth column of $\boldsymbol{M}$. Let $\boldsymbol{M}_\mathcal{A}$ be the submatrix of $\boldsymbol{M}$ formed by columns in $\mathcal{A}$. $\boldsymbol{m}_{i,\mathcal{A}}$ is the ith column of the matrix $\boldsymbol{M}_\mathcal{A}^T$. Similarly, let $\boldsymbol{\alpha}_{0,\mathcal{A}}^\star$ be the subvector of $\boldsymbol{\alpha}_0^\star$ formed by elements in $\mathcal{A}$. Define $\mathcal{A}^c = [1, \ldots, p] - \mathcal{A}$ as the complement set of $\mathcal{A}$. Define $\mathcal{N}_0 = \{\boldsymbol{\delta} \in R^s : \|\boldsymbol{\delta} - \boldsymbol{\alpha}_{0,\mathcal{A}}^\star\|_2 \le d_n\}$. Let $\Sigma_{MM} = E[\boldsymbol{m}_\mathcal{A}\boldsymbol{m}_\mathcal{A}^T]$, $\Sigma_{MX} = E[\boldsymbol{m}_\mathcal{A}\boldsymbol{x}^T]$, and $\Sigma_{XX} = E[\boldsymbol{x}\boldsymbol{x}^T]$. Denote

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XM} \\ \Sigma_{MX} & \Sigma_{MM} \end{pmatrix}.$$

In this paper, for a vector $\boldsymbol{v} = (v_1, \ldots, v_l)^T$, $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$ and $\|\boldsymbol{v}\|_2 = (\boldsymbol{v}^T\boldsymbol{v})^{1/2}$. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of the matrix A, respectively. $\|A\|_{2,\infty} = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_\infty$. Further $a \gg b$ means $\lim_{n\to\infty} a/b = \infty$. We impose the following conditions:

A1. $\lambda_{\min}(\Sigma) \ge c > 0$, $\lambda_{\max}(\Sigma) = O(1)$, and $\|\boldsymbol{M}_{\mathcal{A}^c}^T(\boldsymbol{X}, \boldsymbol{M}_\mathcal{A})\|_{2,\infty} = O_P(n)$.

A2. Let d_n be the half minimum signal of $\boldsymbol{\alpha}_{0,\mathcal{A}}^\star$, i.e. $d_n = \min_{j\in\mathcal{A}} |\alpha_{0j}^\star|/2$. Assume that $d_n \gg \lambda_n \gg \max\{\sqrt{s/n}, \sqrt{\log p/n}\}$, $p'_{\lambda_n}(d_n) = o((ns)^{-1/2})$, $\lambda_n\kappa_0 = o(1)$ where $\kappa_0 = \max_{\delta\in\mathcal{N}_0} \kappa(\rho, \delta, \lambda_n)$.

A3. For some $\varpi > 2$, there exists a positive sequence K_n such that $E[\|\boldsymbol{m}_{\mathcal{A}^c}\varepsilon_1\|_\infty^\varpi] \le K_n^\varpi$ and $K_n^2 \log p/n^{1-2/\varpi-\varsigma} \to 0$ for some arbitrary small $\varsigma > 0$. Further assume that $\max_{1\le j\le p+q} E(z_j^4) < C < \infty$, here $\boldsymbol{z} = (\boldsymbol{m}, \boldsymbol{x})$, z_j is the jth component of $\boldsymbol{z}$.

To emphasize the dependence on the sample size, in the above conditions and the Appendix, we use λ_n to denote the tuning parameter. The first condition is mild and commonly assumed. See for instance Fan and Lv (2011). Condition A2 imposes a minimal signal condition on nonzero elements in $\boldsymbol{\alpha}_0$. Recall that our primary interest is to make statistical inference on direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta} = \boldsymbol{\gamma} - \boldsymbol{\alpha}_1$, and $\boldsymbol{\alpha}_0$ may be treated as a nuisance parameter in this model. We do not make any minimal signal condition on $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}$. Thus, Condition A2 is reasonable to a certain extent. If, in any case where the minimal signal condition is in doubt, debiased procedures may be preferred. Condition A3 is imposed for establishing sparsity result. Compared with existing literature, A3 is very mild. In fact, to simplify the proof, it has been assumed that all covariates are uniformly bounded in the literature. See for instance Wang et al. (2012). Under bounded covariates condition, A3 reduces to $E(|\varepsilon_1|^\varpi) \le C$ by taking K_n as a constant. Furthermore, the dimension p is allowed to increase in a rate of an exponential order of the sample size n according to Conditions A2 and A3.

Theorem 1. *Suppose that Conditions (A1)–(A3) hold, and $s = o(n^{1/2})$, then with probability tending to 1, $\hat{\boldsymbol{\alpha}}_0$ must satisfy (i) $\hat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$. (ii) $\|\hat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}_{0,\mathcal{A}}^\star\|_2 = O_P(\sqrt{s/n})$. Let $\epsilon_1 = (\varepsilon_{11}, \ldots, \varepsilon_{n1})^T$. If further $s = o(n^{1/3})$, we obtain that*

$$\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) = \frac{1}{\sqrt{n}}\Sigma^{-1}\begin{pmatrix} \boldsymbol{X}^T\epsilon_1 \\ \boldsymbol{M}_\mathcal{A}^T\epsilon_1 \end{pmatrix} + o_P(1).$$

The proofs of Theorem 1 and its corollary below are given in the supplement of this paper. Theorem 1 establishes the sparsity of $\hat{\boldsymbol{\alpha}}_0$, the convergence rate of $\hat{\boldsymbol{\alpha}}_{0,\mathcal{A}}$ and the asymptotic representation of $\hat{\boldsymbol{\vartheta}}$. Based on the asymptotical representation, we further obtain the following corollary.

Corollary 1. *Suppose that Conditions (A1)–(A3) hold, and $s = o(n^{1/3})$, we have*

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star) \to N(0, \sigma_1^2(\Sigma_{XX}^{-1} + B)), \text{ and } \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \to N(0, \sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2 B),$$

where $B = \Sigma_{XX}^{-1}\Sigma_{XM}(\Sigma_{MM} - \Sigma_{MX}\Sigma_{XX}^{-1}\Sigma_{XM})^{-1}\Sigma_{MX}\Sigma_{XX}^{-1}$, and $\boldsymbol{\beta}^\star$ is the true value of $\boldsymbol{\beta}$.

This corollary establishes the asymptotic normalities of the estimators $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}$. We next make theoretical comparison with the estimators in Zhou et al. (2020). Note that the asymptotic variance matrices of $\hat{\boldsymbol{\alpha}}_1^Z$ and $\hat{\boldsymbol{\beta}}^Z$ in Zhou et al. (2020) are $\sigma_1^2(\Sigma_{XX}^{-1} + \tilde{B})$ and $\sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2\tilde{B}$, respectively, where $\tilde{\Sigma}_{MM} = E[\boldsymbol{m}\boldsymbol{m}^T]$, $\tilde{\Sigma}_{MX} = E[\boldsymbol{m}\boldsymbol{x}^T]$, $\Sigma_{XX} = E[\boldsymbol{x}\boldsymbol{x}^T]$, and $\tilde{B} = \Sigma_{XX}^{-1}\tilde{\Sigma}_{XM}(\tilde{\Sigma}_{MM} - \tilde{\Sigma}_{MX}\Sigma_{XX}^{-1}\tilde{\Sigma}_{XM})^{-1}\tilde{\Sigma}_{MX}\Sigma_{XX}^{-1}$. To show our proposed estimators are more efficient than the ones in Zhou et al. (2020), it suffices to show that $\tilde{B} > B$. Note that $\Sigma_{XX}^{-1} + B = (I_q, 0_{q\times s})\Sigma^{-1}(I_q, 0_{q\times s})^T$, and

$$\begin{aligned}\Sigma_{XX}^{-1} + \tilde{B} &= (I_q, 0_{q\times p})\begin{pmatrix} E[\boldsymbol{x}\boldsymbol{x}^T] & E[\boldsymbol{x}\boldsymbol{m}^T] \\ E[\boldsymbol{m}\boldsymbol{x}^T] & E[\boldsymbol{m}\boldsymbol{m}^T]\end{pmatrix}^{-1}(I_q, 0_{q\times p})^T \\ &= (I_q, 0_{q\times s})(\Sigma - E[\boldsymbol{x}\boldsymbol{m}_{\mathcal{A}^c}^T]E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{m}_{\mathcal{A}^c}^T]^{-1}E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{x}^T])^{-1}(I_q, 0_{q\times s})^T.\end{aligned}$$

Thus, $\tilde{B} > B$ since $(\Sigma - E[\boldsymbol{x}\boldsymbol{m}_{\mathcal{A}^c}^T]E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{m}_{\mathcal{A}^c}^T]^{-1}E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{x}^T])^{-1} > \Sigma^{-1}$. Hence our proposed estimators are more efficient than the proposal of Zhou et al. (2020). This should not be surprising because the debiased Lasso inflates its asymptotical variance in the debiasing step for high-dimensional linear model (Van de Geer et al., 2014). The proposed partially penalized least squares method does not penalize $\boldsymbol{\alpha}_1$, and hence minimal signal condition on $\boldsymbol{\alpha}_1$ is not required, and the debiased step becomes unnecessary. As discussed by Wang et al. (2020), debiased procedures "achieve bias reduction by essentially allowing all the covariates, including the inactive ones, to be used to adjust for bias". Our procedure aims to work with a sparse model while the debiased procedures are about bias-correction based on all the covariates. Although $\boldsymbol{\alpha}_0$ is sparse, the debiased or desparsified Lasso estimate is not sparse. In other words, debiased procedures do not effectively utilize the sparsity information of nuisance parameter $\boldsymbol{\alpha}_0$.

2.3. Test for indirect effect

To form the test statistic for the indirect effect $\boldsymbol{\beta}$, we first study its asymptotic variance matrix. Let $\hat{\mathcal{A}} = \{j : \hat{\alpha}_{0j} \neq 0\}$. With probability tending to 1, we have $\hat{\mathcal{A}} = \mathcal{A}$. Then the variance matrix Σ and σ_1^2 can be estimated by the estimated sample version and the mean squared errors, respectively.

$$\hat{\Sigma} = \frac{1}{n}\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{M}_{\hat{\mathcal{A}}} \\ \boldsymbol{M}_{\hat{\mathcal{A}}}^T\boldsymbol{X} & \boldsymbol{M}_{\hat{\mathcal{A}}}^T\boldsymbol{M}_{\hat{\mathcal{A}}}\end{pmatrix}, \quad \text{and} \quad \hat{\sigma}_1^2 = \frac{1}{n-\hat{s}-q}\|\boldsymbol{y} - \boldsymbol{M}\hat{\boldsymbol{\alpha}}_0 - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_1\|^2,$$

where $\hat{s} = |\hat{\mathcal{A}}|$. As is shown, $\hat{\sigma}_1^2 = \sigma_1^2 + o_P(1)$. In fact, when $s = o(n^{1/2})$, we have $\hat{\sigma}_1^2 = \sigma_1^2 + O_P(n^{-1/2})$.

As to σ_2^2, we first estimate $\sigma^2 = \text{var}(\varepsilon_3) = \sigma_1^2 + \sigma_2^2$ by the classic least squares residual variance estimator $\hat{\sigma}^2$ based on model (2.3). Thus $\hat{\sigma}_2^2 = \hat{\sigma}^2 - \hat{\sigma}_1^2$. In practice, $\hat{\sigma}_1^2$ may sometimes be larger than $\hat{\sigma}^2$, where we would simply set $\hat{\sigma}_2^2 = 0$. This is possible when no mediators are relevant. That is, $\boldsymbol{\alpha}_0 = 0$, and hence σ_2^2 indeed equals zero.

According to Corollary 1, the asymptotic variance matrices of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}$ can be consistently estimated by:

$$\hat{\sigma}_1^2(I_q, 0_{q\times\hat{s}})\hat{\Sigma}^{-1}(I_q, 0_{q\times\hat{s}})^T;\ \hat{\sigma}_2^2\hat{\Sigma}_{XX}^{-1} + \hat{\sigma}_1^2[(I_q, 0_{q\times\hat{s}})\hat{\Sigma}^{-1}(I_q, 0_{q\times\hat{s}})^T - \hat{\Sigma}_{XX}^{-1}], \tag{2.8}$$

where $\hat{\Sigma}_{XX} = \boldsymbol{X}^T\boldsymbol{X}/n$. Then Wald test statistic for the hypotheses in (2.4) can be derived as

$$S_n = n\hat{\boldsymbol{\beta}}^T\left\{\hat{\sigma}_2^2\hat{\Sigma}_{XX}^{-1} + \hat{\sigma}_1^2[(I_q, 0_{q\times\hat{s}})\hat{\Sigma}^{-1}(I_q, 0_{q\times\hat{s}})^T - \hat{\Sigma}_{XX}^{-1}]\right\}^{-1}\hat{\boldsymbol{\beta}}.$$

Clearly, under H_0, $S_n \to \chi_q^2$, a chi-square random variable with q degrees of freedom.

To investigate the local power of S_n, we consider the local alternative hypotheses H_{1n} : $\boldsymbol{\beta} = \boldsymbol{\Delta}/\sqrt{n}$, where $\boldsymbol{\Delta}$ is a constant vector. From Corollary 1, under such local alternative hypotheses, $S_n \to \chi_q^2(\boldsymbol{\Delta}^T(\sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2 B)^{-1}\boldsymbol{\Delta})$, a chi-square random variable with q degrees of freedom and noncentrality parameter $\boldsymbol{\Delta}^T(\sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2 B)^{-1}\boldsymbol{\Delta}$. Thus, S_n can detect local effects that converge to 0 at root-n rate.

2.4. F-Type test on direct effect

It is of interest to test the following hypothesis

$$H_{02} : \boldsymbol{\alpha}_1 = 0 \text{ versus } H_{12} : \boldsymbol{\alpha}_1 \neq 0. \tag{2.9}$$

(2.1) and (2.2) are called complete or full mediation models under H_{02}, while incomplete or partial mediation models under H_{12}.

Testing the hypothesis in (2.9) essentially is to test low dimensional regression coefficients in linear regression model (2.1). This has been studied when the covariates in (2.1) are fixed design (Zhang and Zhang, 2014; Van de Geer et al., 2014; Shi et al., 2019). Due to the nature of mediation model, the covariates in (2.1) are random design. The fixed-design assumption on $\boldsymbol{m}$ is inappropriate in mediation models.

We next propose an F-type test for (2.9), and further show that the proposed F-test asymptotically has a chi-square distribution with q degrees of freedom under H_{02}, and a noncentral chi-square distribution with q degrees of freedom under H_{12}. Similar to F-test, we need to calculate the residual sum of squares (RSS) under the null and alternative hypotheses. Under H_{02}, the penalized least squares function for model (2.1) becomes

$$\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{M}\boldsymbol{\alpha}_0\|^2+\sum_{j=1}^{p}p_\lambda(|\alpha_{0j}|). \tag{2.10}$$

Denote by $\tilde{\boldsymbol{\alpha}}_0$ the resulting penalized least squares estimator. Then the RSS under H_{02} is $\mathrm{RSS}_0=\|\boldsymbol{y}-\boldsymbol{M}\tilde{\boldsymbol{\alpha}}_0\|^2$. Under H_{12}, we can estimate $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ by the partially penalized least squares method in (2.7). Then we calculate $\mathrm{RSS}_1=\|\boldsymbol{y}-\boldsymbol{M}\hat{\boldsymbol{\alpha}}_0-\boldsymbol{X}\hat{\boldsymbol{\alpha}}_1\|^2$, the RSS under H_{12}.

The F-type test for hypothesis (2.9) is defined to be

$$T_n=\frac{\mathrm{RSS}_0-\mathrm{RSS}_1}{\mathrm{RSS}_1/(n-q)}. \tag{2.11}$$

Theorem 2 shows that the asymptotical null distribution of T_n is a chi-square distribution with q degrees of freedom. To evaluate the local power of T_n under local alternative hypotheses, we impose the following assumption.

A4. Consider local alternative hypotheses $H_{1n}:\boldsymbol{\alpha}_1=\boldsymbol{h}_n$. Assume that $\|\boldsymbol{h}_n\|_2=O(\sqrt{1/n})$.

Theorem 2. *Suppose that Conditions (A1)-(A4) hold, and $s=o(n^{1/3})$. It follows that*

$$\sup_x|P(T_n\le x)-P(\chi_q^2(n\boldsymbol{h}_n^T\Phi^{-1}\boldsymbol{h}_n/\sigma_1^2)\le x)|\to 0. \tag{2.12}$$

Here $\Phi=(I_q,0_{q\times s})\Sigma^{-1}(I_q,0_{q\times s})^T$ and $\chi_q^2(n\boldsymbol{h}_n^T\Phi^{-1}\boldsymbol{h}_n/\sigma_1^2)$ is a chi square random variable with q degrees of freedom and noncentrality parameter $n\boldsymbol{h}_n^T\Phi^{-1}\boldsymbol{h}_n/\sigma_1^2$.

Theorem 2 implies that under H_{02}, T_n asymptotically follows χ_q^2 distribution, which does not depend on any parameter in the model. This is similar to the Wilks phenomenon for likelihood ratio test in classical statistical setting. In other words, the Wilks phenomenon still holds in this high dimensional mediation model. Theorem 2 also implies that T_n can detect local alternatives that are distinct from the null hypothesis at the rate of $1/\sqrt{n}$.

Remark. Intuitively, one may also construct a Wald test for the direct effect based on its asymptotic normality. To this end, one needs to estimate the asymptotic covariance matrix of $\hat{\boldsymbol{\alpha}}_1$. This may be tricky under the setting of ultrahigh dimensional sparse linear models. On the other hand, the Wald test is preferable than the F-type test for the indirect effect $\boldsymbol{\beta}$ since $\boldsymbol{\beta}=\Gamma\boldsymbol{\alpha}_0$ is defined as the product of a high dimensional matrix Γ and a high dimensional vector $\boldsymbol{\alpha}_0$. When the null hypothesis $H_0:\boldsymbol{\beta}=0$ holds, either Γ or $\boldsymbol{\alpha}_0$ is zero, associated with two regression models (2.1) and (2.2). Thus constructing an F-test for the indirect effect is difficult since it is not easy, if not impossible, to define the residual sum of squares under the null (RSS_0) and alternative hypotheses (RSS_1). If, for instance, we construct RSS_0 by only regressing $\mathbf{y}$ versus $\mathbf{x}$, and RSS_1 by the same procedure as when constructing (2.11), it would result in only testing the hypothesis that the high dimensional $\boldsymbol{\alpha}_0=0$, rather than $\boldsymbol{\beta}=0$; Furthermore, classical theory for the F-test becomes invalid for high dimensional parameter $\boldsymbol{\alpha}_0$, in particular, when regularization methods are used to estimate the high dimensional parameters.

2.5. Algorithm and tuning parameter selection

To compute the partially penalized estimators $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}$, we apply the local linear approximation algorithm (LLA) in Zou and Li (2008) with the SCAD penalty (Fan and Li, 2001),

$$p'_\lambda(t)=\lambda\{I(t\le\lambda)+\frac{(a\lambda-t)_+}{(a-1)\lambda}I(t>\lambda)\},$$

and set $a=3.7$. The tuning parameter λ for our method is chosen based on the high-dimensional BIC (HBIC) method in Wang et al. (2013). For a fixed regularization parameter λ, define

$$(\hat{\boldsymbol{\alpha}}_0^\lambda,\hat{\boldsymbol{\alpha}}_1^\lambda)=\min_{\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1}\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{M}\boldsymbol{\alpha}_0-\boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2+\sum_{j=1}^{p}p_\lambda(|\alpha_{0,j}|).$$

The minimization of the partially penalized least squares method can be carried out as follows.

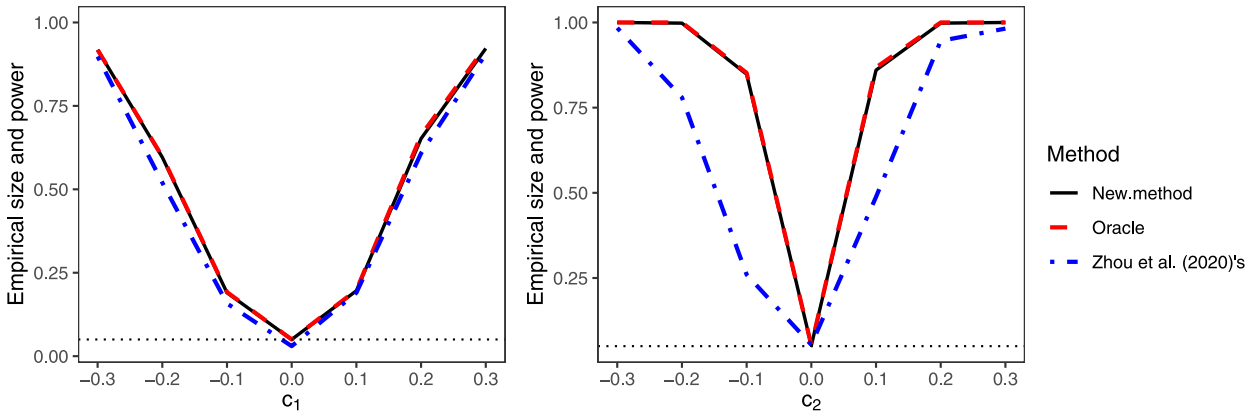

Fig. 1. Left panel is the empirical sizes and powers of S_n, S_n^Z and S_n^O at level $\alpha = 0.05$ over 500 replications for testing indirect effect when $\alpha_1 = 0.5$. Solid line, dotted line and dash-dotted line represent the sizes and powers of S_n, S_n^O, and S_n^Z, respectively. Right panel is empirical sizes and powers of T_n, T_n^Z, and T_n^O at level $\alpha = 0.05$ over 500 replications for testing direct effect when $\beta = 0.7$. The solid line, dotted line, and dash-dotted line represent the sizes and powers of T_n, T_n^O, and T_n^Z, respectively.

1. Get initial values for $\boldsymbol{\alpha}_0^{(0)}, \boldsymbol{\alpha}_1^{(0)}$ by minimizing a partial L_1-penalized least squares: $(\hat{\boldsymbol{\alpha}}_0^{(0)}, \hat{\boldsymbol{\alpha}}_1^{(0)}) = \min_{\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2 + \lambda \sum_{j=1}^p |\alpha_{0,j}|$.
2. Solve $(\hat{\boldsymbol{\alpha}}_0^{(k+1)}, \hat{\boldsymbol{\alpha}}_1^{(k+1)}) = \min_{\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2 + \sum_{j=1}^p p'_\lambda(|\alpha_{0,j}^{(k)}|)|\alpha_{0,j}|$ for $k = 1, 2, \ldots$, until $\{(\hat{\boldsymbol{\alpha}}_0^{(k)}, \hat{\boldsymbol{\alpha}}_1^{(k)})\}$ converges.

In practice, we use a data-driven method to choose the tuning parameter λ. Following Wang et al. (2013), we use the HBIC criterion to choose λ. The HBIC score is defined as $\text{HBIC}(\lambda) = \log(\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2) + \text{df}\log(\log(n))\log(p+q)/n$, where df is the number of variables with nonzero coefficients in $(\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T)^T$. Minimizing $\text{HBIC}(\lambda)$ yields a selection of λ.

3. Numerical studies

In this section, we examine the finite sample performance of the proposed procedures via Monte Carlo simulation studies and illustrate the proposed procedure by a real data example.

3.1. Simulation studies

We first examine finite sample performances of the proposed partial-penalization based test statistics, along with comparisons with the oracle test statistics which know the true set $\mathcal{A} = \{j : \alpha_{0j}^\star \neq 0\}$, denoted as S_n^O and T_n^O as a benchmark, and the debiased test statistics S_n^Z and T_n^Z in Zhou et al. (2020), denoted by Zhou et al.'s method in the tables and figures in this section. Note that Zhou et al. (2020) focus on the test of indirect effects. One can derive a valid Wald test for direct effects based on the asymptotical normality established in their paper.

Example 1. In this example, we set $n = 300$, $q = 1$, and $p = 500$. $\boldsymbol{x} \sim N(0, 1)$ and $\boldsymbol{m} = \Gamma^T\boldsymbol{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(0, \Sigma^*)$ with Σ^* being an AR correlation structure. That is, the (i, j)-element of Σ^* equals $\rho^{|i-j|}$ and ρ is set to be 0.5. Take $\Gamma = c_1(\tau_1, \ldots, \tau_p)^T$, where $\tau_k = 0.2k$ for $k = 1, \ldots, 5$, and when $k > 5$, τ_k's are independently generated from $N(0, 0.1^2)$. Set $c_1 = 0$ to examine Type I error rate and $c_1 = \pm 0.1, \pm 0.2, \ldots, \pm 1$ for power when testing the indirect effects.

We generate the response y from model $y = \boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x} + \varepsilon_1$, where $\varepsilon_1 \sim N(0, 0.5^2)$, $\boldsymbol{\alpha}_0 = [1, 0.8, 0.6, 0.4, 0.2, 0, \ldots, 0]^T$ and $\boldsymbol{\alpha}_1 = c_2$ is set in the same fashion as c_1. The simulation results are based on 500 replications. The significance level is set to be 0.05.

We first compare the performances of S_n, S_n^O and S_n^Z for testing the indirect effect $\boldsymbol{\beta}$. We set $c_2 = 0.5$ and $\boldsymbol{\beta} = \Gamma\boldsymbol{\alpha}_0 = 1.4c_1$. The left panel of Fig. 1 depicts power functions of the three tests versus the values of c_1 over $[-0.3, 0.3]$. All the three tests gain larger powers as $|c_1|$ increases. S_n performs as well as the oracle S_n^O, and is generally more powerful than S_n^Z. For instance, when $c_1 = -0.2$, the empirical power of S_n^Z is 0.516, while the empirical powers of S_n and S_n^O are 0.596. These observations are in consistent with the theoretical results in Section 2.

Next, we turn to test the direct effect. Set $c_1 = 0.5$. And c_2 is taken from $0, \pm 0.1, \pm 0.2, \ldots, \pm 1$, where $c_2 = 0$ corresponds to the null hypothesis. The right panel of Fig. 1 depicts the power function of the three tests versus the values of c_2 over $[-0.3, 0.3]$. The proposed test T_n performs almost the same as the oracle one, and is obviously more powerful than the test T_n^Z proposed in Zhou et al. (2020), whose power curve is asymmetric. In fact, when $c_2 = -0.2$, the empirical powers of our test statistic T_n and the oracle test T_n^O are about 1, while that of T_n^Z is only about 0.780.

Table 1
Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

c_1	c_2	New method		Oracle		Zhou et al.'s method	
		$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\alpha}}_1^O$	$\hat{\boldsymbol{\beta}}^O$	$\hat{\boldsymbol{\alpha}}_1^Z$	$\hat{\boldsymbol{\beta}}^Z$
−0.8	0.5	$-0.23_{(4.15)}$	$-0.22_{(13.73)}$	$-0.11_{(4.11)}$	$-0.35_{(13.70)}$	$-11.77_{(6.56)}$	$11.31_{(14.05)}$
−0.4	0.5	$0.18_{(3.13)}$	$-0.33_{(11.98)}$	$0.25_{(3.08)}$	$-0.40_{(11.95)}$	$-3.49_{(5.10)}$	$3.37_{(12.20)}$
0	0.5	$-0.02_{(2.99)}$	$0.39_{(12.61)}$	$-0.00_{(2.99)}$	$0.37_{(12.63)}$	$-0.13_{(8.65)}$	$0.47_{(15.00)}$
0.4	0.5	$0.02_{(3.15)}$	$0.08_{(11.83)}$	$-0.02_{(3.11)}$	$0.12_{(11.81)}$	$-0.60_{(5.31)}$	$0.77_{(12.66)}$
0.8	0.5	$0.31_{(3.79)}$	$0.26_{(12.69)}$	$0.16_{(3.72)}$	$0.42_{(12.63)}$	$-1.57_{(8.57)}$	$2.19_{(15.05)}$
0.5	−0.8	$0.16_{(3.38)}$	$0.79_{(11.62)}$	$0.11_{(3.37)}$	$0.85_{(11.64)}$	$16.37_{(5.61)}$	$-7.63_{(13.13)}$
0.5	−0.4	$-0.01_{(3.43)}$	$0.16_{(12.58)}$	$-0.09_{(3.36)}$	$0.26_{(12.57)}$	$16.05_{(4.00)}$	$-8.08_{(13.64)}$
0.5	0	$0.10_{(3.35)}$	$-0.15_{(12.52)}$	$0.01_{(3.33)}$	$-0.06_{(12.52)}$	$0.66_{(6.56)}$	$-0.71_{(13.82)}$
0.5	0.4	$0.35_{(3.39)}$	$0.01_{(12.26)}$	$0.32_{(3.37)}$	$0.04_{(12.26)}$	$-0.96_{(5.69)}$	$1.30_{(13.10)}$
0.5	0.8	$0.13_{(3.29)}$	$0.24_{(12.10)}$	$0.05_{(3.26)}$	$0.32_{(12.17)}$	$-0.53_{(5.58)}$	$0.84_{(12.86)}$

Table 2
Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

c_1	c_2	Direct effect ($\hat{\alpha}_1$)				Indirect Effect ($\hat{\beta}$)					
		New method		Oracle		New method		Oracle		Zhou et al.'s method	
		std	se(std)	std	se(std)	std	se(std)	std	se(std)	std	se(std)
−0.8	0.5	4.15	$3.88_{(0.23)}$	4.11	$3.89_{(0.23)}$	13.73	$12.56_{(0.72)}$	13.70	$12.56_{(0.72)}$	14.05	$13.43_{(1.03)}$
−0.4	0.5	3.13	$3.16_{(0.18)}$	3.08	$3.17_{(0.18)}$	11.98	$12.38_{(0.73)}$	11.95	$12.38_{(0.73)}$	12.20	$13.14_{(0.85)}$
0	0.5	2.99	$2.90_{(0.17)}$	2.99	$2.91_{(0.17)}$	12.61	$12.26_{(0.66)}$	12.63	$12.26_{(0.66)}$	15.00	$13.12_{(2.62)}$
0.4	0.5	3.15	$3.18_{(0.18)}$	3.11	$3.19_{(0.18)}$	11.83	$12.35_{(0.71)}$	11.81	$12.35_{(0.71)}$	12.66	$13.09_{(0.82)}$
0.8	0.5	3.79	$3.88_{(0.24)}$	3.72	$3.88_{(0.23)}$	12.69	$12.47_{(0.73)}$	12.63	$12.47_{(0.73)}$	15.05	$13.37_{(1.79)}$
0.5	−0.8	3.38	$3.31_{(0.19)}$	3.37	$3.32_{(0.19)}$	11.62	$12.43_{(0.71)}$	11.64	$12.42_{(0.71)}$	13.13	$14.30_{(0.76)}$
0.5	−0.4	3.43	$3.30_{(0.19)}$	3.36	$3.31_{(0.20)}$	12.58	$12.30_{(0.70)}$	12.57	$12.30_{(0.70)}$	13.64	$13.19_{(0.71)}$
0.5	0	3.35	$3.32_{(0.18)}$	3.33	$3.33_{(0.18)}$	12.52	$12.35_{(0.75)}$	12.52	$12.34_{(0.75)}$	13.82	$13.78_{(3.73)}$
0.5	0.4	3.39	$3.32_{(0.19)}$	3.37	$3.33_{(0.19)}$	12.26	$12.39_{(0.71)}$	12.26	$12.39_{(0.71)}$	13.10	$13.14_{(0.75)}$
0.5	0.8	3.29	$3.33_{(0.20)}$	3.26	$3.34_{(0.20)}$	12.10	$12.37_{(0.74)}$	12.17	$12.37_{(0.74)}$	12.86	$13.27_{(1.31)}$

Furthermore, T_n^Z performs unstably according to our simulation studies. To gain insight of this, we explore more on $\hat{\boldsymbol{\alpha}}_1^Z, \hat{\boldsymbol{\beta}}^Z$. The estimates $\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}_1^O, \hat{\boldsymbol{\beta}}^O$ are reported in Table 1 from which it can be seen that the biases of $\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}_1^O, \hat{\boldsymbol{\beta}}^O$ are very small, while $\hat{\boldsymbol{\alpha}}_1^Z$ has a large bias. This may be due to that the direct effect $\boldsymbol{\alpha}_1$ is also penalized in Zhou et al. (2020)'s estimation procedure based on scaled lasso. This makes sense only if the direct effect is expected to be zero. As seen in Table 1, the bias of $\hat{\boldsymbol{\alpha}}_1^Z$ is very small when $c_2 = 0$, yet inversely when $c_2 \neq 0$.

Table 1 also reports standard errors of corresponding estimates. Both the proposed method and oracle outperform (Zhou et al., 2020), especially when estimating $\boldsymbol{\alpha}_1$.

To assess the accuracy of variance estimation of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}$, Table 2 reports their estimated standard errors in two ways. As to each method — new, oracle and Zhou et al.'s method, the first column lists the empirical standard deviations of point estimates $\hat{\boldsymbol{\alpha}}_1$ or $\hat{\boldsymbol{\beta}}$ over 500 replications (they are also recorded in parentheses of Table 1); for the second column, we estimate standard errors of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}$ using formula (2.8) in each simulation run, and reports the average together with standard deviations (in parentheses) over the 500 runs. Note that the R package "freebird" (Zhou et al., 2020) does not provide the estimated standard error of $\hat{\boldsymbol{\alpha}}_1$. From Table 2, for the new method and oracle, the standard errors estimated by Monte Carlo simulations are close to those calculated from formulas; while the two versions of Zhou et al. (2020) depart more.

Furthermore, Fig. 2 visually compares the standard deviations of $\hat{\boldsymbol{\beta}}$ over 500 point estimates using the new method (x-axis) with those using oracle or Zhou et al.'s method (y-axis), respectively. Each blue diamond or red dot in the figure corresponds to each of the 21 different simulation settings — when holding $c_2 = 0.5$, vary $c_1 = 0, \pm 0.1, \ldots, \pm 1$ in (a) and holding $c_1 = 0.5$, vary $c_2 = 0, \pm 0.1, \ldots, \pm 1$ in (b). The figures imply that the estimated standard errors of the new method are close to oracle, and are generally smaller than those of Zhou et al.'s method. This in turn intuitively illustrates the precision of proposed estimators.

Lastly, Table 3 reports the computing times, where the new method is nearly 1000 times faster than Zhou et al.'s method. The proposed method is very fast and stable because initialized by LASSO estimator, LLA algorithm converges in one step.

Example 2. In this example, we examine the finite sample performances of proposed method when heavy-tail errors are encountered. Specifically, assume now $\varepsilon_1 \sim t_6/\sqrt{6}$. The multiplier $\sqrt{6}$ ensures the equality of variance of ε_1 to that when

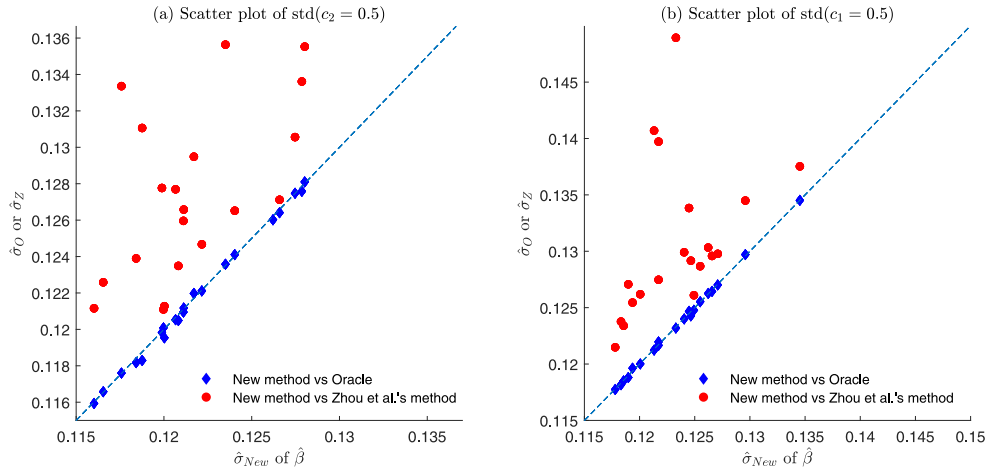

Fig. 2. Scatter plot of standard deviations of $\hat{\boldsymbol{\beta}}$ over 500 point estimates by the new method (x-axis) and by oracle or Zhou et al.'s method (y-axis). Each dot (blue and red) corresponds each of the 21 different simulation settings — when holding $c_2 = 0.5$, vary $c_1 = 0, \pm 0.1, \ldots, \pm 1$ in (a) and holding $c_1 = 0.5$, vary $c_2 = 0, \pm 0.1, \ldots, \pm 1$ in (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison results of the average computing time (in seconds) over 500 replications.

c_1	c_2	New method	Zhou et al.'s method
−0.8	0.5	1.38	1,207.88
−0.4	0.5	1.47	1,327.82
0	0.5	1.31	1,197.66
0.4	0.5	1.52	1,614.84
0.8	0.5	1.22	1,332.24
0.5	−0.8	1.35	1,192.32
0.5	−0.4	1.33	1,329.48
0.5	0	1.48	1,544.23
0.5	0.4	1.50	1,790.34

Table 4
Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2 when $\varepsilon_1 \sim t_6/\sqrt{6}$. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

c_1	c_2	New method		Oracle		Zhou et al.'s method	
		$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\alpha}}_1^O$	$\hat{\boldsymbol{\beta}}^O$	$\hat{\boldsymbol{\alpha}}_1^Z$	$\hat{\boldsymbol{\beta}}^Z$
−0.8	0.5	$0.14_{(4.06)}$	$-0.30_{(12.46)}$	$0.22_{(3.93)}$	$-0.38_{(12.43)}$	$-13.93_{(6.09)}$	$13.50_{(12.84)}$
−0.4	0.5	$0.01_{(1.93)}$	$-0.14_{(6.24)}$	$0.06_{(1.89)}$	$-0.19_{(6.23)}$	$-3.34_{(2.81)}$	$3.23_{(6.43)}$
0	0.5	$0.16_{(3.03)}$	$-0.36_{(12.21)}$	$0.14_{(3.01)}$	$-0.34_{(12.21)}$	$-1.13_{(4.68)}$	$0.86_{(12.74)}$
0.4	0.5	$0.16_{(3.29)}$	$-0.36_{(12.30)}$	$0.09_{(3.26)}$	$-0.28_{(12.29)}$	$-0.77_{(5.19)}$	$0.52_{(13.01)}$
0.8	0.5	$0.28_{(3.07)}$	$-0.26_{(6.67)}$	$0.21_{(3.02)}$	$-0.18_{(6.63)}$	$0.75_{(4.06)}$	$-0.70_{(7.15)}$
0.5	−0.8	$0.19_{(3.44)}$	$-0.37_{(12.34)}$	$0.10_{(3.40)}$	$-0.28_{(12.33)}$	$6.50_{(5.61)}$	$-6.73_{(12.89)}$
0.5	−0.4	$0.16_{(3.45)}$	$-0.32_{(12.32)}$	$0.09_{(3.41)}$	$-0.25_{(12.30)}$	$5.92_{(12.67)}$	$-6.16_{(16.26)}$
0.5	0	$0.19_{(3.42)}$	$-0.34_{(12.34)}$	$0.09_{(3.41)}$	$-0.25_{(12.30)}$	$0.70_{(4.56)}$	$-0.95_{(12.95)}$
0.5	0.4	$0.20_{(3.44)}$	$-0.39_{(12.39)}$	$0.09_{(3.41)}$	$-0.28_{(12.33)}$	$-1.20_{(5.30)}$	$0.93_{(12.98)}$
0.5	0.8	$0.18_{(3.44)}$	$-0.34_{(12.32)}$	$0.09_{(3.41)}$	$-0.25_{(12.30)}$	$-1.17_{(5.29)}$	$0.96_{(13.07)}$

$\varepsilon_1 \sim N(0, 0.5^2)$. All other settings are identical to those in Example 1. We first investigate the performances of S_n, S_n^O and S_n^Z for testing indirect effect $\boldsymbol{\beta}$ via the left panel of Fig. 3. The proposed test S_n performs as well as the oracle one S_n^O in terms of controlling Type-I error rate ($c_1 = 0$) and possessing much larger power than S_n^Z (when $c_1 \neq 0$), especially when $c_1 < 0$. Similar phenomenons are observed in the right penal of Fig. 3 when examining T_n, T_n^O and T_n^Z. The proposed test T_n performs as well as the oracle one, and is more powerful than the test T_n^Z. In fact, when $c_2 = -0.2$, the empirical powers of our test statistic T_n and the oracle test T_n^O are about 1, while that of T_n^Z is only about 0.756. In addition, we also evaluate the accuracy and precision of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}$ through Tables 4 and 5. The overall pattern in these two tables with $\varepsilon_1 \sim t_6/\sqrt{6}$ is very similar to that for $\varepsilon_1 \sim N(0, 0.5^2)$. In sum, the proposed method retains its validity for heavy-tailed error distributions.

Table 5
Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) of different methods with different c_1 and c_2 when $\varepsilon_1 \sim t_6/\sqrt{6}$. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

c_1	c_2	Direct effect ($\hat{\alpha}_1$)				Indirect Effect ($\hat{\beta}$)					
		New method		Oracle		New method		Oracle		Zhou et al.'s method	
		std	se(std)	std	se(std)	std	se(std)	std	se(std)	std	se(std)
−0.8	0.5	4.06	$3.87_{(0.28)}$	3.93	$3.88_{(0.27)}$	12.46	$12.55_{(0.70)}$	12.43	$12.55_{(0.70)}$	12.84	$13.12_{(0.90)}$
−0.4	0.5	1.93	$1.94_{(0.14)}$	1.89	$1.95_{(0.14)}$	6.24	$6.29_{(0.33)}$	6.23	$6.29_{(0.33)}$	6.43	$6.47_{(0.36)}$
0	0.5	3.03	$2.91_{(0.21)}$	3.01	$2.92_{(0.21)}$	12.21	$12.29_{(0.72)}$	12.21	$12.29_{(0.72)}$	12.74	$12.80_{(0.77)}$
0.4	0.5	3.29	$3.17_{(0.23)}$	3.26	$3.18_{(0.23)}$	12.30	$12.35_{(0.72)}$	12.29	$12.35_{(0.72)}$	13.01	$12.95_{(0.82)}$
0.8	0.5	3.07	$2.92_{(0.22)}$	3.02	$2.93_{(0.22)}$	6.67	$6.66_{(0.30)}$	6.63	$6.66_{(0.30)}$	7.15	$6.57_{(0.64)}$
0.5	−0.8	3.44	$3.31_{(0.24)}$	3.40	$3.32_{(0.24)}$	12.34	$12.39_{(0.71)}$	12.33	$12.39_{(0.71)}$	12.89	$12.98_{(0.74)}$
0.5	−0.4	3.45	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.32	$12.39_{(0.71)}$	12.30	$12.39_{(0.71)}$	16.26	$13.01_{(0.87)}$
0.5	0	3.42	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.34	$12.39_{(0.71)}$	12.30	$12.39_{(0.71)}$	12.95	$12.96_{(0.74)}$
0.5	0.4	3.44	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.39	$12.39_{(0.71)}$	12.33	$12.39_{(0.71)}$	12.98	$12.98_{(0.81)}$
0.5	0.8	3.44	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.32	$12.39_{(0.71)}$	12.30	$12.39_{(0.71)}$	13.07	$12.99_{(0.86)}$

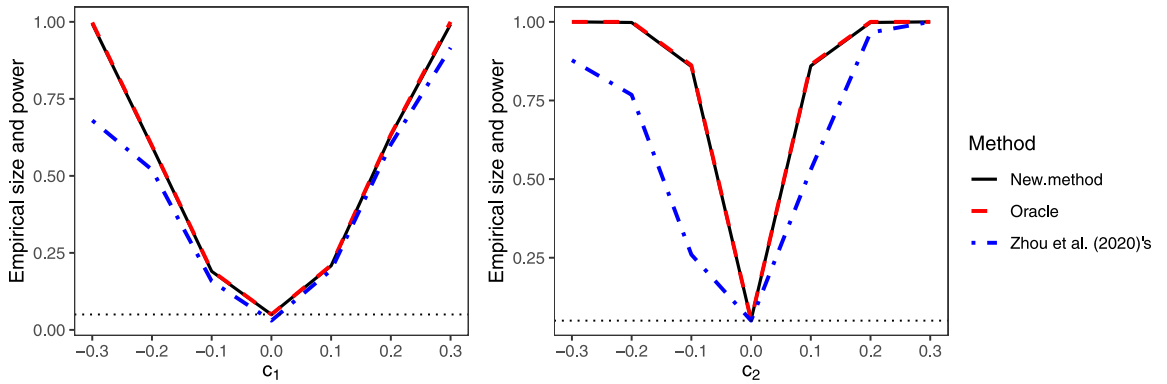

Fig. 3. Left panel is empirical sizes and powers of S_n, S_n^Z and S_n^O when $\varepsilon_1 \sim t_6/\sqrt{6}$ at level $\alpha = 0.05$ over 500 replications for testing indirect effect when $\alpha_1 = 0.5$. Dotted line, solid line, and dash-dotted line represent the sizes and powers of S_n, S_n^O and S_n^Z, respectively. Right panel is empirical sizes and powers of T_n, T_n^Z and T_n^O for testing direct effect when $\beta = 0.7$. The dotted line, solid line, and dash-dotted line represent the sizes and powers of T_n, T_n^O and T_n^Z, respectively.

3.2. Real data analysis

We apply the proposed method to an empirical analysis to examine whether financial statements items and metrics mediate the relationship between company sectors and stock price recovery after COVID-19 pandemic outbreak. While investors and researchers have reached a consensus ages ago that stock returns highly rely on companies' belonging sectors, recent studies more focus on using financial statements or market conditions to predict stock returns. Fama and French (1993)'s pioneering proposal of the three-factor model started this era, which captures patterns of return using market return, firm size and book-to-market ratio factors. Callen and Segal (2004) showed that accruals, cash flow, growth in operating income significantly influence stocks return. Edirisinghe and Zhang (2008) developed a relative financial strength metric based on data envelopment analysis (Farrell, 1957), and found that return on assets and solvency ratio has high correlation with stock price return. To enhance prediction accuracy, deep neural network and data mining techniques were developed, with model inputs as historical financial statements and output as stock price return (Enke and Thawornwong, 2005; Lee et al., 2019). Meanwhile, it is reasonable to hypothesize that companies' sectors affect stock performances via influencing the associated financial metrics. Few existing works, however, study the mediating effects of such financial metrics. Hence our analysis aims to fill in this gap, and use the proposed mediation analysis to select important financial metrics, as well as to test the direct and indirect effects of companies' sectors on returns.

In addition, we in this analysis are specifically interested in the stock performance of S&P 500 component companies during the COVID-19 pandemic period. As is known, the outbreak of the COVID-19 dealt a shock to the U.S. economy with unprecedented speed, and the government had to take a lockdown to stop spread of virus. The lockdown took a toll in the U.S. economy: business were closed, millions of people lost jobs and the price of an oil futures contract fell below zero. The crisis spread to the U.S. stock market, dragging down the major index S&P 500 by 33.92%. To help businesses, households and the economy, the Federal Reserve and the White House launched various rescue programs and take measures to stabilize energy prices from the end of March, 2020. Therefore, all these events and measures led the U.S. stock market to a V-shape pattern, thanks to which, the general financial rules from classical literature may not directly apply any more.

Table 6
The estimated coefficients and associated standard errors of direct and indirect effect in the stock data analysis.

Sectors	Direct effect	std	Indirect effect	std
Intercept	−0.2845	0.0248	−0.0831	0.0162
Basic materials	0.0912	0.0354	0.0709	0.0224
Communication services	0.1471	0.0430	0.0673	0.0269
Consumer cyclical	0.0126	0.0287	0.0662	0.0185
Consumer defensive	0.1563	0.0332	0.0998	0.0221
Financial services	0.0217	0.0293	0.0549	0.0192
Healthcare	0.1621	0.0301	0.1211	0.0196
Industrials	0.0583	0.0289	0.0939	0.0188
Real.Estate	0.0117	0.0348	0.0798	0.0217
Technology	0.1042	0.0290	0.1220	0.0191
Utilities	0.1083	0.0337	0.0592	0.0214

Admittedly, some recent literature studied the economic reaction to COVID-19 pandemic from sector or company level data (Ramelli and Wagner, 2020; Zhang et al., 2020; Baker et al., 2020; Gormsen and Koijen, 2020; De Vito and Gomez, 2020). Thorbecke (2020) analyzed sector-specific and macroeconomic variables as contributing factors to stock return in COVID-19 downturn and found that idiosyncratic factors negatively affected energy and consumer cyclical sectors. Hassan et al. (2020) investigated companies' transcripts of quarterly earnings call from January to September 2020 to investigate senior management's and major market participants' opinions about future prospects. They discovered several important factors related to accounting and business fundamentals, including supply chain, production and operations and financing, that are highly associated with stock market recovery from COVID-19. However, these methods mainly rely on prior financial knowledge to select low dimensional data for modeling, while ignore important company level factors. In addition, these methods only consider the relation of stock return to either sector level or company level while failing to recognize that the company's financial plays a role in mediating stock sector effects to stock price return. Therefore, we use the proposed method to study the financial statement items or metrics that mediate the relationship between firm sectors and stock performance in this special period. This work may then shed light on how to select valuable stocks during a pandemic or any adverse event likewise.

In the mediation models, the response is taken to be the stock return from its highest price before the pandemic in February, 2020 to April 30th, 2020. The closed price is adjusted for both dividends and splits. The potential mediators in $\boldsymbol{m}$ are 550 accounting metrics from financial statements of associated companies, scratched from Yahoo Finance on April 30, 2020. We obtain annual reports of firms from fiscal year 2015Y to 2019Y and the first three quarterly reports in 2019. We use the latest annual report of firms to compute financial metrics and use previous annual reports to compute average growth rate of each financial metrics. In practice, financial analysts use the latest financial statements and the news to gauge future stock price performance. Thus, we focus on the companies that released their latest reports, either quarterly or annual report, during the investigation period from February to April 2020. This results in that 490 companies in the S&P 500 are included this example. Thus the total sample size is 490.

The exposure variables in $\boldsymbol{x}$, are companies' sectors according to Global Industry Classification Standard (GICS) that are coded as dummy variables. GICS classifies companies into eleven sectors: basic materials, communication services, consumer cyclical, consumer defensive, energy, financial services, healthcare, industrials, real estate, technology and utilities. We set energy sector as baseline level. In this empirical analysis, we select λ by the high-dimensional BIC (HBIC, Wang et al. (2013)), in which the authors demonstrate that the HBIC balances model complexity and prediction, and prove that the HBIC selects the optimal tuning parameter which asymptotically identifies the oracle estimator under the high dimensional linear regression model setting. The selected $\lambda = 0.1125$ — this yields six financial metrics as selected mediators. It is worth to noting that six financial metrics are not as few as they seem to be because the financial metrics are correlated, thus other potentially relevant metrics are to some extent represented by the six chosen ones. We also validate model assumption via residual plot. See Section S.4 in the supplement for details.

Table 6 presents the estimated direct and indirect effects of companies' sectors, together with their standard errors. The test statistic $S_n = 57.857$ with P-value $< 10^{-8}$ for testing indirect effect $H_0 : \boldsymbol{\beta} = 0$, and $T_n = 731.47$ with P-value $< 10^{-15}$ for testing direct effect $H_0 : \boldsymbol{\alpha}_1 = 0$, indicating both the direct and indirect effect are significant. Note that this is not a experimental study, like clinical trials, and the $\boldsymbol{x}$ is an exposure rather than treatment variable. The coefficients of $\boldsymbol{x}$ may not be interpreted as causal effects of $\boldsymbol{x}$, but they may have a descriptive interpretation. For example, as shown in Table 6, stocks in sectors such as 'Communication services' and 'Healthcare' are more likely to outperform benchmark than 'Industrials'. Furthermore, sectors influence the stocks performance partly through business operation reflected by selected financial metrics, and the indirect effects are significantly positive.

The selected mediating metrics, their associated estimated coefficients in model (2.1), as well as their brief descriptions, are presented in Table 7. These selected metrics are of their own significance. For instance, the first three chosen metrics in Table 7, namely return on assets, gross margin and annual growth rate of operating income, reflect firms' revenue. Return on assets is an indicator of how well a firm utilizes its assets, by determining how profitable a firm is relative to its total assets. A firm with a higher return-on-assets value is preferred, as the firm squeezes more out of limited resources to make a profit. Gross margin is the portion of sales revenue a firm retains after subtracting costs of producing the goods

Table 7
Selected importance mediators and their coefficients.

Selected mediator	Estimated coefficient (std)	Description
Return on assets	0.0677 (0.0060)	Net income divided by the total assets
Gross margin	0.0134 (0.0062)	The difference between the revenue and cost of goods sold divided by revenue
AGR* Operating Income	0.0169 (0.0055)	Revenues subtract the cost of goods sold and operating expenses
AGR* Quick ratio	0.0190 (0.0054)	Total current assets minus inventory divided by total current liabilities
Debt to assets	−0.0193 (0.0058)	Total debts divided by total assets
Receivables turnover (days)	−0.0151 (0.0055)	Average receivables divided by net credit sales times 360 days

* AGR: average growth rate, calculated as the average of growth rates for the metrics from 2015Y to 2019Y.

it sells and the services it provides. A firm that has higher gross margin is more likely to retain more profit for every dollar of good sold. Annual growth rate of operating income shows the firm's growth of generating operating income compared with previous year. Operating income measures the amount of profit realized from a business's operation, after deducting operating expenses such as wages, depreciation, and cost of goods sold. A firm with high growth of operating income can avoids unnecessary production costs, and improve core business efficiency. In a word, a firm with a higher return on assets, gross margin and growing operating income is considered more profitable, and hence, more likely to attract investors.

On the other hand, both the average growth rate of quick ratio and debt to assets are indicators of financial leverage of a firm. Quick ratio of a firm is defined as the dollar amount of liquid assets dividing that of current liabilities, where liquid assets are the portion of assets that can be quickly converted into cash with minimal impact on the price received in open market, while current liabilities are a firm's debts or obligations to be paid to creditors within one year. Thus a large quick ratio indicates that the firm is fully equipped with enough assets to be instantly liquidated to pay off its current liabilities. Debt to assets is the total amount of debt relative to assets owned by a firm. It reflects a firm's financial stability. Therefore, a firm with a higher quick ratio or a lower debt to assets might be more likely to survive when it is difficult to finance through borrowing and cover its debts, thus is more favorable to investors during the economy lockdown.

Lastly, receivables turnover quantifies a firm's effectiveness in collecting its receivables or money owed by clients. It shows how well a firm uses and manages the credit it extends to customers and how quickly that short-term debt is paid. Receivables turnover can be negative when net credit sale is negative because the client pre-pay for the product or service. A negative receivables turnover means that the firm are less susceptible to counter-party credit risk because it already receives the cash from its client before delivering the service or shipping out the product. This is especially important during liquidity dry periods when the clients may default or delay payment due to lack of cash. Therefore, a firm that has a negative receivables turnover is preferred.

On all accounts, one might incorporate the analysis results as reference when seeking for a stock portfolio during the financial crisis caused by pandemic. First, the sectors in 'Healthcare', 'Consumer defensive', 'Communication service', 'Utility' and 'Technology' have the top five positive direct effects on stock return. In terms of the financial metrics, we may focus on those reported in Table 7 to filter stocks. For example, we shall select firms that have higher values in AGR operating income, gross margin, quick ratio, and return on assets but lower values in debt to assets and receivable turnover.

Moreover, we compare our findings with those selected in established models. For instance, our method picks profitability factors like return on assets, which is also selected in Fama and French (2015), as profitability is the core of a firm's stock performance. But we do not include metrics representing size of firm, valuation of stock price or investment that were covered by Fama and French (2015). For firm size factor, there is no evidence that small-size firms recovered faster or slower than larger-size ones. For valuation of stock price factor, previous price valuation ratio changed significantly due to stock price change and is no longer reliable to predict future stock return. For investment factor, it is less important for a short-term stock price movement. Compared with Edirisinghe and Zhang (2008), our method also picks profitability (return on assets), liquidity (quick ratio) and solvency (debt to assets) metrics, as in Edirisinghe and Zhang (2008). During the crisis, a firm facing liquidity crunch could not access to credit. Therefore, a firm with sufficient cash and less debt is more easily to survive and less likely to be forced to liquidate valuable assets at unfavorable prices. And its stock would be safer and more attractive to investors. But we did not select metrics of earnings per share or about capital intensity as in Edirisinghe and Zhang (2008). The lockdown dramatically changes a firm's revenue structure and capital allocation, and hence reduces predictive capability of these metrics to short-term recovery.

4. Conclusion

In this paper, we propose statistical inference procedures for the indirect effects in high dimensional mediation model. We introduce a partially penalized least squares method and study its statistical properties under random design, and show that the proposed estimators are more efficient than existing ones. We further develop a partially penalized Wald test to detect the indirect effect, with a χ^2 limiting null distribution, and develop an F-type test for the direct effect

and reveal Wilks phenomenon in the high-dimensional mediation model. The proposed inference procedures are used to analyze the mediation effects of various financial metrics on the relationship between company's sector and the stock return.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to the associate editor and referees for their constructive comments and suggestions that led to significant improvement of an early manuscript. Guo's research was supported by National Natural Science Foundation of China (NNSFC) grant 12071038 and Beijing Natural Science Foundation, China (1212004). Liu's research was supported by grants from NNSFC, China grants 11701034 and 71988101. Li and Zeng's research was supported by National Science Foundation, USA, DMS 1820702, 1953196 and 2015539. All the authors contribute equally to the paper, and are listed in the alphabetical order.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.03.001.

References

Abadie, A., Cattaneo, M.D., 2018. Econometric methods for program evaluation. Annu. Rev. Econ. 10, 465–503.
Baker, S.R., Bloom, N., Davis, S.J., Kost, K., Sammon, M., Viratyosin, T., 2020. The unprecedented stock market reaction to COVID-19. Rev. Asset Pricing Stud. 10 (4), 742–758.
Baron, R.M., Kenny, D.A., 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J. Personal. Soc. Psychol. 51 (6), 1173.
Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85 (1), 233–298.
Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. Rev. Econom. Stud. 81 (2), 608–650.
Callen, J.L., Segal, D., 2004. Do accruals drive firm-level stock returns? A variance decomposition analysis. J. Account. Res. 42 (3), 527–560.
Cattaneo, M.D., Jansson, M., Ma, X., 2019. Two-step estimation and inference with possibly many included covariates. Rev. Econom. Stud. 86 (3), 1095–1122.
Cattaneo, M.D., Jansson, M., Newey, W.K., 2018. Inference in linear regression models with many covariates and heteroscedasticity. J. Amer. Statist. Assoc. 113 (523), 1350–1361.
Celli, V., 2022. Causal mediation analysis in economics: Objectives, assumptions, models. J. Econ. Surv. 36, 214–234.
Chernozhukov, V., Hansen, C., Spindler, M., 2015. Valid post-selection and post-regularization inference: An elementary, general approach. Annu. Rev. Econ. 7 (1), 649–688.
Chernozhukov, V., Kasahara, H., Schrimpf, P., 2021. Causal impact of masks, policies, behavior on early COVID-19 pandemic in the US. J. Econometrics 220 (1), 23–62.
Conti, G., Heckman, J.J., Pinto, R., 2016. The effects of two influential early childhood interventions on health and healthy behaviour. Econ. J. 126 (596), F28–F65.
De Vito, A., Gomez, J.-P., 2020. Estimating the COVID-19 cash crunch: Global evidence and policy. J. Account. Public Policy 39 (2), 106741.
Donald, S.G., Hsu, Y.-C., 2014. Estimation and inference for distribution functions and quantile functions in treatment effect models. J. Econometrics 178, 383–397.
Edirisinghe, N.C.P., Zhang, X., 2008. Portfolio selection under DEA-based relative financial strength indicators: case of US industries. J. Oper. Res. Soc. 59 (6), 842–856.
Enke, D., Thawornwong, S., 2005. The use of data mining and neural networks for forecasting stock market returns. Expert Syst. Appl. 29 (4), 927–940.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.
Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.
Fan, Y., Demirkaya, E., Li, G., Lv, J., 2020a. RANK: Large-scale inference with graphical nonlinear knockoffs. J. Amer. Statist. Assoc. 115 (529), 362–379.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Fan, J., Li, R., Zhang, C.-H., Zou, H., 2020b. Statistical Foundations of Data Science. Chapman and Hall/CRC.
Fan, J., Lv, J., 2011. Nonconcave penalized likelihood with NP-dimensionality. IEEE Trans. Inform. Theory 57 (8), 5467–5484.
Fan, Y., Lv, J., Sharifvaghefi, M., Uematsu, Y., 2020c. IPAD: stable interpretable forecasting with knockoffs inference. J. Amer. Statist. Assoc. 115 (532), 1822–1834.
Farrell, M.J., 1957. The measurement of productive efficiency. J. R. Stat. Soc. Ser. A (General) 120 (3), 253–281.
Farrell, M.H., 2015. Robust inference on average treatment effects with possibly more covariates than observations. J. Econometrics 189 (1), 1–23.
Galbraith, J.W., Zinde-Walsh, V., 2020. Simple and reliable estimators of coefficients of interest in a model with high-dimensional confounding effects. J. Econometrics 218 (2), 609–632.
Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 (3), 1166–1202.
Gormsen, N.J., Koijen, R.S., 2020. Coronavirus: Impact on stock prices and growth expectations. Rev. Asset Pricing Stud. 10 (4), 574–597.
Graham, C.M., Cannice, M.V., Sayre, T.L., 2002. The value-relevance of financial and non-financial information for internet companies. Thunderbird Int. Bus. Rev. 44 (1), 47–70.
Hassan, T.A., Hollander, S., Van Lent, L., Schwedeler, M., Tahoun, A., 2020. Firm-Level Exposure to Epidemic Diseases: COVID-19, SARS, and H1N1. Technical Report, National Bureau of Economic Research.

Heckman, J., Pinto, R., 2015. Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. Econometric Rev. 34 (1), 6–31.
Huber, M., 2020. Mediation analysis. In: Handbook of Labor, Human Resources and Population Economics. Springer, pp. 1–38.
Huber, M., Frölich, M., 2017. Direct and indirect treatment effects: Causal chains and mediation analysis with instrumental variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 79, 1645–1666.
Imai, K., Keele, L., Tingley, D., 2010. A general approach to causal mediation analysis. Psychol. Methods 15 (4), 309.
Imbens, G.W., 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. Rev. Econ. Stat. 86 (1), 4–29.
Khan, M.N., Khokhar, I., et al., 2015. The effect of selected financial ratios on profitability: an empirical analysis of listed firms of cement sector in Saudi Arabia. Q. J. Econom. Res. 1 (1), 1–12.
Lee, T.K., Cho, J.H., Kwon, D.S., Sohn, S.Y., 2019. Global stock market investment strategies based on financial network indicators using machine learning techniques. Expert Syst. Appl. 117, 228–242.
Ramelli, S., Wagner, A.F., 2020. Feverish stock price reactions to COVID-19. Rev. Corp. Finance Stud. 9 (3), 622–655.
Shi, C., Song, R., Chen, Z., Li, R., 2019. Linear hypothesis testing for high dimensional generalized linear models. Ann. Statist. 47 (5), 2671.
Thorbecke, W., 2020. The impact of the COVID-19 pandemic on the US economy: evidence from the stock market. J. Risk Financial Manag. 13 (10), 233.
Wang, J., He, X., Xu, G., 2020. Debiased inference on treatment effect in a high-dimensional model. J. Amer. Statist. Assoc. 115 (529), 442–454.
Wang, L., Kim, Y., Li, R., 2013. Calibrating non-convex penalized regression in ultra-high dimension. Ann. Statist. 41 (5), 2505–2536.
Wang, L., Wu, Y., Li, R., 2012. Quantile regression for analyzing heterogeneity in ultra-high dimension. J. Amer. Statist. Assoc. 107 (497), 214–222.
Zhang, D., Hu, M., Ji, Q., 2020. Financial markets under the global pandemic of COVID-19. Finance Res. Lett. 36, 101528.
Zhang, C.-H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 217–242.
Zhou, R.R., Wang, L., Zhao, S.D., 2020. Estimation and inference for the indirect effect in high-dimensional linear mediation models. Biometrika 107 (3), 573–589.
Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist. 36 (4), 1509.