

# Information theoretic approach to high-dimensional multiplicative models: Stochastic discount factor and treatment effect

CHEN QIU

Department of Economics, Cornell University

TAISUKE OTSU

Department of Economics, London School of Economics

This paper is concerned with estimation of functionals of a latent weight function that satisfies possibly high-dimensional multiplicative moment conditions. Main examples are functionals of stochastic discount factors in asset pricing, missing data problems, and treatment effects. We propose to estimate the latent weight function by an information theoretic approach combined with the  $\ell_1$ -penalization technique to deal with high-dimensional moment conditions under sparsity. We study asymptotic properties of the proposed method and illustrate it by a theoretical example on treatment effect analysis and empirical example on estimation of stochastic discount factors.

**KEYWORDS.** Information theoretic approach, high-dimensional model, stochastic discount factor, treatment effect.

**JEL CLASSIFICATION.** C12, C14.

## 1. INTRODUCTION

### 1.1 *Motivation*

In applied research, economic or statistical information is commonly characterized by moment conditions on observables. The generalized method of moments provides a unified framework to analyze the moment condition models, and numerous extensions have been proposed in the econometrics literature. This paper is concerned with the following moment condition model with a multiplicative moment function:

$$\mathbb{E}[\omega(X)g(X)] = r, \tag{1}$$

where  $X$  is a vector of observables,  $\mathbb{E}[\cdot]$  is expectation under the data generating measure of  $X$ ,  $\omega : \mathcal{X} \rightarrow (0, \infty)$  is an *unknown* weight function,  $g$  is a vector of *known* functions of  $X$ , and  $r$  is a vector of known constants or moments of observables (say,  $r = \mathbb{E}[r(X)]$ )

---

Chen Qiu: [cq62@cornell.edu](mailto:cq62@cornell.edu)

Taisuke Otsu: [t.otsu@lse.ac.uk](mailto:t.otsu@lse.ac.uk)

The authors would like to thank referees for helpful comments. Financial support from the ERC Consolidator Grant (SNP 615882) is gratefully acknowledged (Otsu).

© 2022 The Authors. Licensed under the [Creative Commons Attribution-NonCommercial License 4.0](https://creativecommons.org/licenses/by-nc/4.0/). Available at <http://qeconomics.org>. <https://doi.org/10.3982/QE1603>

for some known  $r(\cdot)$ ). We are interested in the situation where the observables  $X$  and/or vector of functions  $g$  are high dimensional (possibly higher than the sample size).

In general, there exists a nontrivial set  $W$  of  $\omega$  that satisfies (1). In this paper, we introduce an information theoretic approach to select a particular element  $\omega_0 \in W$ , and define the object of interest as its linear functional:

$$\theta_0 = \mathbb{E}[\omega_0(X)h(X, Y)], \quad (2)$$

where  $Y$  is another vector of observables and  $h$  is a vector of known functions of  $(X, Y)$ . This paper develops a general estimation and inference method for the parameter  $\theta_0$  under possibly high-dimensional moment conditions (1).

Interestingly, this setup can be motivated by somewhat distant empirical problems: inference on stochastic discount factors (SDFs) and missing data problems including treatment effect analysis. The latent weight  $\omega$  plays the role of the SDF for the former example, and the (reciprocal of) missing probability or propensity score for the latter.

**EXAMPLE 1 (Stochastic discount factor).** In a discrete time economy with no arbitrage, there exists a strictly positive SDF  $m_t$  such that

$$\mathbb{E}[m_t R_{j,t}] = 1, \quad (3)$$

where  $R_{j,t}$  is the short term return of asset  $j \in \{1, \dots, K-1\}$  between time  $t$  and  $t+1$ , and  $\mathbb{E}[\cdot]$  is the objective expectation operator. This equation says that any asset in the market would share the same expected return when discounted by the SDF  $m_t$  (see [Cochrane \(2009\)](#) for a review). Suppose there also exists a risk free asset with return  $R_{f,t}$ , which satisfies

$$\mathbb{E}[m_t R_{f,t}] = 1. \quad (4)$$

Let  $X_t = (1, R_{1,t} - R_{f,t}, \dots, R_{(K-1),t} - R_{f,t})'$  be a  $K$  dimensional vector of the excess returns and a constant. Since  $\mathbb{E}[m_t] \neq 0$ , (3) and (4) imply

$$\mathbb{E}\left[\frac{m_t}{\mathbb{E}[m_t]} X_t\right] = \mathbf{e}_1, \quad (5)$$

where  $\mathbf{e}_1 = (1, 0, \dots, 0)'$ . Unless the market is complete, the SDF  $m_t$  (and thus  $m_t/\mathbb{E}[m_t]$ ) is generally set identified from the moment condition (5). That is, without further restrictions, any positive random variable  $m_t$  satisfying (5) can be a valid SDF.<sup>1</sup>

In this example, we focus on the case where  $m_t/\mathbb{E}[m_t]$  is written as a function of  $X_t$ . However, this is still not enough to pin down the (normalized) SDF, and there is a set  $W$  of functions of  $X_t$  satisfying (5), that is,

$$\mathbb{E}[\omega(X_t)X_t] = \mathbf{e}_1, \quad \text{for all } \omega \in W. \quad (6)$$

<sup>1</sup>Note that the moment condition (3) also holds conditionally on agents' information sets (say,  $\mathbb{E}[m_t R_{j,t} | \mathcal{I}_{t-1}] = 1$  for the information set  $\mathcal{I}_{t-1}$  at  $t-1$ ) whereas this example focuses on the unconditional moment condition in (3). Thus, the identified set for  $m_t$  by the unconditional moment in (3) is a superset of the one by the conditional moment  $\mathbb{E}[m_t R_{j,t} | \mathcal{I}_{t-1}] = 1$ . Furthermore, since any  $m_t$  satisfying (3) also satisfies (5), the identified set for  $m_t$  by (5) is a superset of the one by (3).

This setup can be considered as a special case of (1) with  $g(X) = X$  and  $r = \mathbf{e}_1$ . In Section 1.2, we present how our methodology can be used to estimate some particular elements in  $W$ .<sup>2</sup>

Inference on SDFs is one of the central topics in financial economics. For example, Christensen (2017) investigated extraction of permanent and transitory components of the SDF process, which requires estimation of  $\mathbb{E}[m_t b(S_t) b(S_{t+1})']$  for a vector of known basis functions  $b(\cdot)$  and state vectors  $S_t$  and  $S_{t+1}$ . Christensen (2017) considered two cases: (i)  $m_t$  is directly observable and (ii)  $m_t$  is replaced with a (parametric or nonparametric) preliminary estimator. Our information theoretic approach will provide nonparametric estimators for some particular choices of  $\omega$  and alternative estimators for  $\mathbb{E}[m_t b(S_t) b(S_{t+1})']$  designed for possibly high-dimensional setups.

**EXAMPLE 2 (Missing data).** Consider the problem of estimating a population mean from incomplete outcome data (see Little and Rubin (2002) for a survey). For each unit  $i = 1, \dots, N$ , we observe an indicator variable  $D_i$  ( $D_i = 1$  if unit  $i$  responds and  $D_i = 0$  otherwise), outcome variable  $Y_i = D_i Y_i^*$  ( $Y_i = 0$  means that  $Y_i^*$  is missing), and vector of covariates  $X_i$ . We are interested in the population mean  $\theta = \mathbb{E}[Y_i^*]$ . Under conditional independence of  $Y^*$  and  $D$  given  $X$  and certain overlap assumptions, the parameter of interest is identified as  $\theta = \mathbb{E}[\omega(X) Y D]$ , where  $\omega(X) = 1/\mathbb{P}\{D = 1|X\}$ . In this setup, many estimation and inference methods for  $\theta$  have been proposed (see, e.g., Tsiatis (2006)), including the inverse probability weighted estimator  $n^{-1} \sum_{i=1}^n \tilde{\omega}(X_i) Y_i D_i$ , where  $\tilde{\omega}(x)$  is a nonparametric estimator of  $1/\mathbb{P}\{D = 1|X = x\}$ .

Our information theoretic approach can be applied in this setup to develop an alternative estimator of  $\theta$ . By the law of iterated expectations, the moment conditions in the form of (1) may be given by

$$\mathbb{E}[\omega(X) g(X) D] = \mathbb{E}[g(X)], \quad (7)$$

for any vector of known functions  $g$ . Then the estimation problem of  $\theta$  can be formulated as a special case of ours by replacing the expectations in (1) and (2) with the conditional expectations given  $D = 1$  and setting  $r = \mathbb{E}[g(X)]$  and  $h(X, Y) = Y$ . In the recent literature of missing data analysis and causal inference, the covariate balancing approach explores the moment conditions in (7) to find suitable weights used for estimation of  $\theta$  (see, e.g., Zubizarreta (2015) and Chan, Yam, and Zhang (2016)). This paper proposes an alternative estimation method that may be considered as an extension of those papers toward high-dimensional setups.

## 1.2 Methodology

In this paper, we propose an information theoretic approach to select some element  $\omega_0$  satisfying (1) and to estimate the parameter  $\theta_0$  in (2) based on  $\omega_0$ . Our approach allows

<sup>2</sup>Based on the framework in Hansen (2014) (see also Chen, Hansen, and Hansen (2020)), the SDF and belief distortions cannot be disentangled. In this context, the object  $m_t$  could be interpreted as the belief distortion required to rationalize an SDF that takes the value 1 almost surely.

high-dimensional observables and/or moment functions (possibly higher than the sample size). This feature is particularly desirable for our motivating examples. For Example 1, the number of assets may be very large. For Example 2, the number of covariates tends to be large so that the conditional independence assumption (unconfoundedness or ignorability in causal analysis) is likely to be satisfied.

More precisely, we regard the latent weight function as the Radon–Nikodym derivative  $\omega = d\mathbb{Q}/d\mathbb{P}$ , where  $\mathbb{P}$  is the data generating measure of  $X$  and  $\mathbb{Q} \ll \mathbb{P}$  is a tilted model-based measure. Since the first elements of  $g$  and  $r$  in (1) are assumed to be 1 (Condition D(3) in the next section), we guarantee that  $\mathbb{E}[\omega(X)] = \int d\mathbb{Q} = 1$ . Letting  $\mathbb{E}_{\mathbb{Q}}[\cdot]$  be expectation under  $\mathbb{Q}$ , the moment condition (1) is written as  $\mathbb{E}_{\mathbb{Q}}[g(X)] = r$ .

In general, there are infinitely many possible choices for the tilted measure  $\mathbb{Q}$ . As a rule to select a particular  $\mathbb{Q}$ , we introduce the information projection based on the  $\phi$ -divergence in the Orlicz space (see, e.g., Csiszár (1995) and Komunjer and Ragusa (2016)). Let  $\phi : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  be a convex and lower-semicontinuous divergence function.<sup>3</sup> We consider the following minimization problem:

$$\min_{\mathbb{Q}} \mathbb{E} \left[ \phi \left( \frac{d\mathbb{Q}}{d\mathbb{P}} \right) \right], \quad \text{s.t. } \mathbb{E}_{\mathbb{Q}}[g(X)] = r, \quad \mathbb{E} \left[ \phi \left( 1 + c \left| \frac{d\mathbb{Q}}{d\mathbb{P}} \right| \right) \right] < \infty \quad \text{for some } c > 0. \quad (8)$$

Under Condition D(4) in the next section, Theorem 3 in Komunjer and Ragusa (2016) implies that the solution of (8) exists and is unique, and by Komunjer and Ragusa (2016, Lemma 5), the primal problem (8) has a well-defined dual problem

$$\min_{\lambda} \mathbb{E} [\phi_*(\lambda'g(X)) - \lambda'r], \quad (9)$$

where  $\phi_*(a) = \sup_{b \in \mathbb{R}} \{ab - \phi(b)\}$  is the convex conjugate of  $\phi$ . Furthermore, let  $\lambda_*$  be the solution of (9). Under Conditions D(3) and D(5) in the next section, we can apply Borwein and Lewis (1993, Corollary 3.6 and primal constraint qualification) implying that the solution  $\mathbb{Q}_*$  of (8) can be characterized as

$$\frac{d\mathbb{Q}_*}{d\mathbb{P}}(\cdot) = \phi_*^{(1)}(\lambda_*'g(\cdot)), \quad (10)$$

where  $\phi_*^{(1)}$  is the first derivative of  $\phi_*$ .

We now define the weight function  $\omega_0$  satisfying (1) of our interest. Since the dimension of  $g$ , denoted by  $K$ , grows as the sample size increases, we define  $\omega_0$  as follows: for each  $x$  in the support  $\mathcal{X}$  of  $X$ ,

$$\omega_0(x) = \begin{cases} \omega(x), & \text{if } \omega \text{ is point identified,} \\ \lim_{K \rightarrow \infty} \frac{d\mathbb{Q}_*}{d\mathbb{P}}(x) = \lim_{K \rightarrow \infty} \phi_*^{(1)}(\lambda_*'g(x)), & \text{if } \omega \text{ is set identified.} \end{cases} \quad (11)$$

That is, if the underlying model that implies (1) uniquely identifies  $\omega$  as  $K \rightarrow \infty$  (as in Example 2),  $\omega_0$  is considered as this identified  $\omega$ . If the underlying model that implies (1) partially or set identifies  $\omega$  even when  $K \rightarrow \infty$  (as in Example 1), we define  $\omega_0(x)$  as the

<sup>3</sup>For convenience, we view  $\phi$  as an extended real valued function defined on  $\mathbb{R}$ . This means: for  $\phi$  defined a priori on  $(0, +\infty)$ , we extend it outside its domain by setting  $\phi(u) = +\infty$  for all  $u \in (-\infty, 0)$  and  $\phi(0) = \lim_{u \rightarrow 0^+} \phi(u)$ .

pointwise limit of  $\frac{d\mathbb{Q}_*}{d\mathbb{P}}(x)$  in (10) for each  $x \in \mathcal{X}$ . Based on  $\omega_0$  defined above, our object of interest is defined as

$$\theta_0 = \mathbb{E}[\omega_0(X)h(X, Y)]. \quad (12)$$

Our estimation methods for  $\omega_0(\cdot)$  and  $\theta_0$  are presented as follows. Let  $\mathbb{E}_n[\cdot]$  be the sample mean,  $\|\cdot\|_1$  be the  $\ell_1$ -norm for a vector, and  $\mathbb{I}\{x \in \mathcal{X}_n\}$  be a trimming term for an increasing sequence  $\{\mathcal{X}_n\}$  to the support  $\mathcal{X}$  of  $X$  to deal with technical problems when  $\mathcal{X}$  is unbounded (see [Chen and Christensen \(2015\)](#)). By taking sample counterparts for the trimmed moment functions, our information theoretic estimator of  $\theta_0$  is obtained as

$$\hat{\theta} = \mathbb{E}_n[\phi_*^{(1)}(\hat{\lambda}'g(X)\mathbb{I}\{X \in \mathcal{X}_n\})h(X, Y)], \quad (13)$$

where

$$\hat{\lambda} = \begin{cases} \arg \min_{\lambda} \mathbb{E}_n[\phi_*(\lambda'g(X)\mathbb{I}\{X \in \mathcal{X}_n\}) - \lambda'r(X)\mathbb{I}\{X \in \mathcal{X}_n\}] \\ \text{(low-dimensional case),} \\ \arg \min_{\lambda} \mathbb{E}_n[\phi_*(\lambda'g(X)\mathbb{I}\{X \in \mathcal{X}_n\}) - \lambda'r(X)\mathbb{I}\{X \in \mathcal{X}_n\}] + \alpha_n \|\lambda\|_1 \\ \text{(high-dimensional case),} \end{cases} \quad (14)$$

$\alpha_n$  is a penalty level chosen by the researcher, and  $r(X)$  may be a vector of known constants (as in [Example 1](#)). The  $\ell_1$ -penalty term for the high-dimensional case is introduced to regularize behaviors of  $\hat{\lambda}$ . Although this paper focuses on the  $\ell_1$ -penalization ([Tibshirani \(1996\)](#)), other penalization methods (such as the smoothly clipped absolute deviation by [Fan and Li \(2001\)](#), and minimax concave penalty by [Zhang \(2010\)](#)) may be applied as well.

Popular choices of the divergence  $\phi$  that will satisfy our regularity conditions are: (i) Kullback–Leibler (KL) divergence (or relative entropy)

$$\phi(x) = \begin{cases} x \log x - x + 1, & x > 0, \\ 1, & x = 0, \\ +\infty, & x < 0, \end{cases}$$

with  $\phi_*(y) = e^y - 1$ , (ii) Pearson's  $\chi^2$  divergence without truncation at zero (PSN1)

$$\phi(x) = \frac{1}{2}x^2 - x + \frac{1}{2},$$

with  $\phi_*(y) = \frac{1}{2}y^2 + y$ , (iii) Pearson's  $\chi^2$  divergence with truncation at zero (PSN2)

$$\phi(x) = \begin{cases} \frac{1}{2}x^2 - x + \frac{1}{2} & \text{for } x \geq 0, \\ +\infty & \text{for } x < 0, \end{cases}$$

with  $\phi_*(y) = \frac{1}{2}(\max\{y, -1\})^2 + \max\{y, -1\}$ .<sup>4</sup>

<sup>4</sup>In our empirical illustration in [Section 5](#), we present results using both versions of Pearson's  $\chi^2$  divergence, and find PSN1 performs slightly better in finite samples. A drawback of PSN1 is that the resulting

We emphasize that although the construction of  $\hat{\lambda}$  in (14) is reminiscent of the generalized empirical likelihood estimator for overidentified moment condition models (Newey and Smith (2004)), our setup and properties of the estimator are significantly different for three reasons. First, our moment conditions in (1) involve the latent weight function  $\omega$ , and the information projection is applied to pin down  $\omega_0$ . Second, the interpretation and property of  $\hat{\lambda}$  are different from theirs. In the conventional generalized empirical likelihood estimator,  $\hat{\lambda}$  plays the role of the Lagrange multiplier or shadow price for the moment conditions, and converges to zero as the sample size increases if the model is correctly specified. On the other hand, in our approach,  $\hat{\lambda}$  is an estimator for the dual parameter  $\lambda_*$  and typically does not converge to zero (even though the moment conditions (1) are correctly specified). With this respect, our method is more in line with the sieve estimation methodology. Finally, we allow the moment conditions (1) to be high dimensional, where  $\hat{\lambda}$  has to be regularized as in (14).

### 1.3 Choice of divergence

To implement our information theoretic estimator  $\hat{\theta}$  in (13), we need to choose the divergence  $\phi$ . When  $\omega$  is point identified by the underlying model implying (1) (as in Example 2), any choice of  $\phi$  satisfying the regularity conditions in the next section yields a consistent and asymptotically normal estimator for  $\theta_0$ .

If  $\omega$  is set identified by (1) (as in Example 1), different choices of  $\phi$  typically select different elements in the identified set  $W$  for  $\omega$ . In this paper, we do not advocate any particular choice of  $\phi$  since its choice usually differs by motivations of researchers.

For instance, in Example 1, choosing a quadratic divergence (e.g.,  $\phi(x) = \frac{1}{2}x^2$ ) picks off the best linear approximation of the projected SDF  $\omega_p(\cdot) = \mathbb{E}[m_t|X_t = \cdot]/\mathbb{E}[m_t]$ . On the other hand, the use of the KL divergence has been motivated by several papers in the literature (Stutzer (1995) and Ghosh, Julliard, and Taylor (2016)): it has a quasi-maximum likelihood interpretation, is consistent with the maximum entropy principle in Bayesian methods, and adds minimum amount of information for the moment conditions to hold. The KL divergence offers a closed-form solution that automatically integrates to 1 and is nonnegative. Moreover, in Example 1, the SDF estimated by the KL divergence is particularly attractive since it is adapted to the popular log-linear modeling of the SDF (e.g., Vasicek (1977)), and consistent with the optimal portfolio choice with an expected utility maximizing investor who has constant absolute risk aversion utility. See Backus, Chernov, and Zin (2014) and Hansen (2014) for further details.

Although a formal discussion of the optimal choice of  $\phi$  is beyond the scope of this paper, we note that the KL divergence requires more stringent regularity conditions (such as existence of higher moments of  $g$  in Condition D(4) below), so it may not be suitable for heavy-tailed data. Therefore, as a general rule of thumb, if some higher moments of  $g$  do not exist, then divergences that impose less stringent conditions for moments (such as the Pearson's  $\chi^2$  divergence) would be more appropriate. In

---

estimate for  $\omega_0$  may take negative values. Christensen and Connault (2019) developed a hybrid divergence that smoothly pastes together KL divergence with a quadratic function. Their hybrid divergence also satisfies our regularity conditions, ensures that the estimate for  $\omega_0$  will always take positive values, and requires weaker moment conditions (see Condition D(4)) compared to the KL divergence.

Section 5, we apply the Pearson's  $\chi^2$  and KL divergences to estimate the SDF and compare their cross-sectional predictability. Both of these estimated SDFs show better predictability than Fama–French's three factors, but exhibit rather different shapes. In the low-dimensional scenario, the SDF estimated by the KL divergence is highly positively skewed and leptokurtic. On the other hand, the SDF estimated by the Pearson's  $\chi^2$  divergence is more symmetric and has low kurtosis. We also find that the performance of the KL divergence is better than the Pearson's  $\chi^2$  divergence in the low-dimensional case in terms of out-of-sample cross-sectional predictability.<sup>5</sup> On the other hand, in high-dimensional scenarios, we need to penalize more aggressively for the KL, and Pearson's  $\chi^2$  divergence performs slightly better than the KL after penalization in terms of out-of-sample cross-sectional predictability. Thus, in our empirical example of estimating out-of-sample SDFs, if higher moments of returns do exist, then divergences that are more sensitive to deviations from one probability measure to another (such as the KL divergence) are more preferable, since they can capture skewness and other higher moment characteristics that might be important in asset markets. Given these theoretical and empirical results, we recommend to use the KL divergence in low-dimensional scenarios and Pearson's  $\chi^2$  divergence in high-dimensional scenarios for our empirical example in Section 5.

#### 1.4 Related literature

The construction of our estimator is related to the method of generalized empirical likelihood (Newey and Smith (2004)). In spite of similarity of the construction of the estimator, however, our setup and properties of  $\hat{\lambda}$  are quite different from this literature. Indeed, our treatment on  $\hat{\lambda}$  shares more similarities with coefficients for basis functions in series or sieve estimation (see Chen (2007) for a review).

In order to deal with high-dimensional moment conditions, we adapt the general theory of the lasso with convex loss functions by van de Geer (2008) and Bühlmann and van de Geer (2011) to our setup. For inference, the debiasing method adopted in Section 3 is similar to Zhang and Zhang (2014) and van de Geer, Bühlmann, Ritov, and Dezeure (2014). Note that the results in Section 3 complement the literature on high-dimensional semiparametric inference with locally/doubly robust moment conditions (e.g., Farrell (2015), Belloni et al. (2017), and Chernozhukov et al. (2018)). Our method can also be compared to high-dimensional versions of empirical likelihood methods, such as Hjort, McKeague, and Van Keilegom (2009), Tang and Leng (2010), and Lahiri and Mukhopadhyay (2012). Again, however, our setup and treatment on  $\hat{\lambda}$  are intrinsically different from this literature (typically  $\hat{\lambda}$  converges to nonzero  $\lambda_*$  in our setup).

The main applications of our method are inference on missing data models, treatment effects, and stochastic discount factors. Here, we only mention closely related papers to clarify our contributions in these fields. See Imbens and Rubin (2015) and Cochrane (2009) for surveys of these topics.

<sup>5</sup>This result may be interpreted as an indication of importance of modeling skewness in financial market (e.g., Kraus and Litzenberger (1976)).



In the realm of asset pricing, our paper is closely related to information theoretic approaches for semi-nonparametric analysis on the SDF (e.g., Kitamura and Stutzer (2002) and Ghosh, Julliard, and Taylor (2016, 2017)). In this context, we make three contributions. First, our method can be regarded as an extension of some existing methods, such as the ones by Ghosh, Julliard, and Taylor (2016, 2017), to high-dimensional setups (especially for a large number of assets). Second, our theoretical analysis for the low-dimensional case in Section 2 provides a theoretical background for the analyses in Ghosh, Julliard, and Taylor (2016, 2017). Third, as mentioned in Example 1, this paper can provide an alternative method to extract permanent and transitory components of SDF processes (Christensen (2017)). Our paper has also been influenced by Hansen (2014) who formulates the problem of estimating SDFs as recovering distorted beliefs (see also Chen, Hansen, and Hansen (2020)). In this context, the estimated SDF in this paper could be interpreted as estimates for the belief distortion required to rationalize an SDF that takes the value 1 almost surely.

In the context of missing data and treatment effect analysis, the proposed method, illustrated in Section 4, is closely related to the literature on balancing weights (Zubizarreta (2015), Chan, Yam, and Zhang (2016), and Athey, Imbens, and Wager (2016)). Compared to Zubizarreta (2015) and Chan, Yam, and Zhang (2016), this paper is considered as an extension toward a high-dimensional setup. Compared to Athey, Imbens, and Wager (2016), this paper proposes an alternative estimation method for treatment effects under high-dimensional covariates by utilizing an information theoretic approach.

### 1.5 Organization

The paper is organized as follows. We first present theoretical properties of our estimator  $\hat{\theta}$  for the low-dimensional case (Section 2) and high-dimensional case (Section 3). Then the proposed method is illustrated by a theoretical example on treatment effects (Section 4) and empirical example on the SDF (Section 5). Proofs and additional tables are contained in the Appendix of the Online Supplementary Material (Qiu and Otsu (2022)).

#### Notation

Hereafter, we work with triangular array data  $\{X_i^{(n)}, Y_i^{(n)}\}_{i=1}^n$ , which are considered as the first  $n$  elements of the infinite sequence  $\{X_i^{(n)}, Y_i^{(n)}\}_{i=1}^\infty$  generated from a probability measure  $\mathbb{P}^{(n)}$ . To simplify the notation, we suppress the upper scripts and denote by  $\{X_i, Y_i\}_{i=1}^n$  and  $\mathbb{P}$ . Our asymptotic analysis is based on the array asymptotics, and the convergence “ $\rightarrow$ ” is understood as the one for  $n \rightarrow \infty$ . Also, let  $\mathbb{E}[\cdot] = \mathbb{E}_{\mathbb{P}}[\cdot]$  be expectation under  $\mathbb{P}$ ,  $\mathbb{E}_n[\cdot]$  be the empirical average,  $\mathbb{I}\{A\}$  be the indicator function for an event  $A$ ,  $|B| = \sqrt{\lambda_{\max}(B'B)}$  be the  $\ell_2$ -norm for a scalar, vector, or matrix  $B$ , and  $a \vee b = \max\{a, b\}$ . For a matrix  $C = [c_{ij}]$ , let  $\lambda_{\max}(C)$  and  $\lambda_{\min}(C)$  be its maximum and minimum eigenvalues, respectively, and denote  $\|C\|_\infty = \max_{1 \leq i, j \leq n} |c_{ij}|$  and  $\|C\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |c_{ij}|$ . Let  $f^{(k)}$  be the  $k$ th derivative of function  $f$ . Finally, “ $A \lesssim B$ ” means there exists some positive constant  $C$  that does not depend on  $n$  and satisfies  $A \leq BC$  for all  $n$  large enough.



## 2. LOW-DIMENSIONAL CASE

In this section, we present asymptotic properties of our information theoretic estimator  $\hat{\theta}$  in (13) for the low-dimensional case, where  $K = \dim(g)$  in (1) grows slowly compared to the sample size  $n$ . In this case, computation of  $\hat{\lambda}$  in (14) does not involve the  $\ell_1$ -penalization. We first impose the following conditions.

CONDITION D.

- (1)  $\{X_i, Y_i\}_{i=1}^n$  is a strictly stationary and ergodic triangular array, and  $\{X_i\}_{i=1}^n$  is  $\alpha$ -mixing with mixing coefficients  $\{\alpha_{X,m}\}$  satisfying  $\sum_{m=1}^n \alpha_{X,m}^{1/2-1/q} \lesssim 1$  for some  $q > 2$ .
- (2) The support  $\mathcal{X} \subseteq \mathbb{R}^p$  of  $X$  is a Cartesian product of  $p$  intervals with nonempty interiors.  $\{\mathcal{X}_n\}$  is an increasing sequence of compact, convex, and nonempty subsets of  $\mathcal{X}$ , and satisfies  $\mathbb{P}\{X \notin \mathcal{X}_n\} = o(n^{-1})$ .
- (3) The first element of  $g$  is 1 and the first element of  $r$  is 1.  $\omega_0$  defined in (11) exists and is a continuous function bounded from above and away from zero with  $\mathbb{E}[\omega_0(X)^2] < \infty$ .  $h$  is a scalar-valued continuous function with  $\mathbb{E}[h(X, Y)^2] < \infty$ .
- (4)  $\phi$  is strictly convex and twice continuously differentiable on  $(0, +\infty)$ , and satisfies  $\phi(1) = \phi^{(1)}(1) = 0$ ,  $\lim_{u \rightarrow 0^+} \phi^{(1)}(u) < 0$ ,  $\lim_{u \rightarrow +\infty} \phi^{(1)}(u) > 0$ ,  $\lim_{u \rightarrow +\infty} \frac{\phi(u)}{u} = +\infty$ , and  $\lim_{u \rightarrow \infty} \frac{u\phi^{(1)}(u)}{\phi(u)} < \infty$ , where  $\phi^{(1)}$  is the first derivative of  $\phi$ .  $\mathbb{E}[\phi_*(a|g_j(X))|] < \infty$  for each  $j = 1, \dots, K$  and  $a > 0$ . There exists some probability measure  $\mathbb{Q}_1$  such that  $\mathbb{E}[\phi(\frac{d\mathbb{Q}_1}{d\mathbb{P}}(X))] < \infty$ .
- (5) There exists some probability measure  $\mathbb{Q}_2$  such that  $\frac{d\mathbb{Q}_2}{d\mathbb{P}}(x)$  is strictly positive and is in the quasi-relative interior of the domain of  $\phi$  for each  $x \in \mathcal{X}$ ,  $\mathbb{E}[\phi(1 + c|\frac{d\mathbb{Q}_2}{d\mathbb{P}}(X))|] < +\infty$  for some  $c > 0$ , and  $\mathbb{E}[g(X)\frac{d\mathbb{Q}_2}{d\mathbb{P}}(X)] = r$ .

Condition D contains standard assumptions on the data  $\{X_i, Y_i\}_{i=1}^n$ , divergence  $\phi$ , and functions appearing in (1) and (2). Condition D(1) allows the data to be weakly dependent, and covers independent and identically distributed (iid) data as a special case. Condition D(2) is on the support  $\mathcal{X}$  of  $X$  and the trimming set  $\mathcal{X}_n$ . For example, the condition  $\mathbb{P}\{X \notin \mathcal{X}_n\} = o(n^{-1})$  is satisfied with  $\mathcal{X}_n = \{x \in \mathbb{R}^p : |x| \leq n^{1/a}\}$  for  $a \in (0, a_1)$  with  $\mathbb{E}[|X|^{a_1}] < \infty$ .<sup>6</sup> Condition D(3) is on the functions appearing in (1) and (2). The first requirement in Condition D(3) guarantees that  $\mathbb{Q}_*$  in (10) integrates to 1. By [Kojunjer and Ragusa \(2016, Theorem 3 and Lemma 5\)](#), Condition D(4) guarantees that the solution of (8) exists and is unique, and that the primal problem in (8) has the well-defined dual problem in (9). Note that this condition allows unbounded  $g$  as long as  $\mathbb{E}[\phi_*(a|g_j(X))|] < \infty$  for each  $j = 1, \dots, K$  and  $a > 0$ . Condition D(5) combined with the first requirement in Condition D(3) provides a constraint qualification to guarantee the

<sup>6</sup>In this paper, we apply trimming on the support  $\mathcal{X}$  instead of the moment functions  $g$ . The main reason is that the trimming on  $\mathcal{X}$  makes it easier to verify the approximation condition in Condition S(2) (equation (16)) below. We also note that trimming on  $\mathcal{X}$  is adopted by [Chen and Christensen \(2015\)](#).

strong duality between (8) and (9), that is, the unique solution of (8) coincides with the one of (9) by applying [Borwein and Lewis \(1993, Corollary 3.6 and primal constraint qualification\)](#).

To simplify the presentation, we focus on the case where  $h$  (and thus  $\theta_0$ ) is scalar-valued. An extension to the case of vector  $\theta_0$  is straightforward. It is also possible to extend our method to the case where  $\theta_0$  is implicitly defined as a solution of moment conditions  $\mathbb{E}[h(Z, \theta_0, \omega_0(X))] = 0$  for  $Z = (Y, X)'$  and a linear map  $h$  (in  $\omega_0$ ).

Let  $g_n(X) = \mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]^{-1/2}g(X)\mathbb{I}\{X \in \mathcal{X}_n\}$  be the orthonormalized version of  $g$  after trimming. We impose the following assumptions.

**CONDITION S.**

- (1) All eigenvalues of  $\mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]$  are strictly positive for each  $n$ , and  $|\mathbb{E}_n[g_n(X)g_n(X)'] - I| = o_p(1)$ .
- (2) There exists some  $\lambda_b \in \mathbb{R}^K$  such that

$$\sup_{x \in \mathcal{X}_n} |[\phi_*^{(1)}]^{-1}(\omega_0(x)) - \lambda_b' g_n(x)| \lesssim \eta_{K,n}, \quad (15)$$

$$\sqrt{\mathbb{E}[\{\omega_0(X) - \phi_*^{(1)}(\lambda_b' g_n(X))\}^2]} \lesssim s_{K,n}, \quad (16)$$

for some  $\eta_{K,n} \rightarrow 0$  and  $s_{K,n} \rightarrow 0$ .

Condition S lists requirements for the functions  $g$  and  $g_n$ . Condition S(1) contains eigenvalue conditions on  $\mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]$  to guarantee existence of  $g_n$ , and the convergence of the matrix  $\mathbb{E}_n[g_n(X)g_n(X)']$ . This convergence is satisfied if  $\{X_i\}_{i=1}^n$  is iid and  $\zeta_{K,n}^2 \log K \rightarrow 0$ , where  $\zeta_{K,n} = \sup_{x \in \mathcal{X}} |g_n(x)|$  (see, Lemma 3(i) in the Appendix of the Online Supplementary Material). This convergence can be satisfied for dependent data as well. For example, by [Chen and Christensen \(2015, Lemma 2.2\)](#), if  $\{X_i\}_{i=1}^n$  is stationary and  $\beta$ -mixing with mixing coefficients  $\{\beta_m\}$  such that  $\beta_m n/m \rightarrow 0$  for some integer  $m \leq n/2$ , then  $|\mathbb{E}_n[g_n(X)g_n(X)'] - I| = O_p(\sqrt{m\zeta_{K,n}^2 \log K/n})$  provided  $m\zeta_{K,n}^2 \log K/n \rightarrow 0$ . Condition S(2) imposes assumptions on series approximations by  $g_n$  for  $[\phi_*^{(1)}]^{-1}(\omega_0)$ . The orders of the approximation errors  $\eta_{K,n}$  and  $s_{K,n}$  depend on the choices of the basis functions  $g$ , trimming set  $\mathcal{X}_n$ , and smoothness of  $[\phi_*^{(1)}]^{-1}(\omega_0(\cdot))$ . It can be verified by using results from functional analysis literature (e.g., [Lorentz \(1986\)](#) and [Schumaker \(1981\)](#)).

Let  $r_n(X) = \mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]^{-1/2}r(X)\mathbb{I}\{X \in \mathcal{X}_n\}$  and

$$M_{K,n} = \max_{1 \leq j \leq K} \{\mathbb{E}[|g_{nj}(X)|^q]\}^{1/q} \vee \{\mathbb{E}[|r_{nj}(X)|^q]\}^{1/q} \quad \text{for } q \text{ in Condition D(1),}$$

$$\tilde{s}_{K,n} = \sqrt{\frac{1}{n} \left( s_{K,n}^2 + s_{K,n}^{1+2/q} \sum_{m=1}^n \alpha_{X,m}^{1/2-1/q} \right)},$$

$$B_{K,n} = s_{K,n} + \sqrt{\tilde{s}_{K,n}}, \quad \mu_{K,n} = 1 + M_{K,n} \sum_{m=1}^n \alpha_{X,m}^{1/2-1/q}.$$

As in [Komunjer and Ragusa \(2016\)](#), we define  $\phi^{(1)}(0) = \lim_{u \rightarrow 0^+} \phi^{(1)}(u)$ , and  $\phi^{(1)}(+\infty) = \lim_{u \rightarrow +\infty} \phi^{(1)}(u)$ . We impose the following assumptions for the convex conjugate function  $\phi_*$  of the divergence  $\phi$ .

**CONDITION I.**  $\phi_* : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  is strictly convex and three times continuously differentiable on  $(\phi^{(1)}(0), \phi^{(1)}(+\infty))$ . Also,  $\zeta_{K,n}(\sqrt{K\mu_{K,n}/n} + B_{K,n}) \rightarrow 0$ .<sup>7</sup>

Let  $\hat{\omega}(x) = \phi_*^{(1)}(\hat{\lambda}'g(x)\mathbb{I}\{x \in \mathcal{X}_n\})$ . Based on the above conditions, the convergence rates of  $\hat{\omega}(\cdot)$  and consistency of the estimator  $\hat{\theta}$  in (13) are obtained as follows.

**THEOREM 1.** *Suppose that Conditions D, S, and I hold true. Then*

$$\sqrt{\mathbb{E}_n[\{\hat{\omega}(X) - \omega_0(X)\}^2]} = O_p(\sqrt{K\mu_{K,n}/n} + B_{K,n}), \quad (17)$$

$\hat{\theta} \xrightarrow{P} \theta_0$ , and

$$\sup_{x \in \mathcal{X}_n} |\hat{\omega}(x) - \omega_0(x)| = O_p(\zeta_{K,n}\sqrt{K\mu_{K,n}/n} + \zeta_{K,n}B_{K,n} + \eta_{K,n}). \quad (18)$$

The consistency of  $\hat{\theta}$  is established by showing that of  $\hat{\omega}$  under the empirical  $L_2$ -norm in (17). As a byproduct of the proof of (17), we can obtain (18), an upper bound of the uniform convergence rate of  $\hat{\omega}$  over the trimming set  $\mathcal{X}_n$ .<sup>8</sup> Interestingly, although our setup is different from standard nonparametric series estimation and  $\omega_0$  is not a conditional expectation function, we achieve similar convergence rates with conventional series estimators for regression models. Indeed, our proof is in line with series estimation methods, where the estimation error of  $\hat{\omega}$  can be decomposed into two parts: approximation bias (corresponding to  $B_{K,n}$ ) and sampling error (corresponding to  $\sqrt{K\mu_{K,n}/n}$ ). The approximation error is dealt with Lemma 2 while the sampling error is controlled by Lemma 3. In particular,  $\mu_{K,n}$  characterizes a slowdown of the convergence rate for the sampling error due to weak dependence of the data. For iid data, we have  $\mu_{K,n} = 1$  and the sampling error is of order  $\sqrt{K/n}$ . On the other hand,  $\tilde{s}_{K,n}$  is an additional term due to weak dependence in the approximation bias  $B_{K,n}$ . For iid data with  $\sqrt{n}s_{K,n} \rightarrow \infty$ , the bias term becomes a familiar expression  $B_{K,n} = s_{K,n}$ .

We next consider the limiting distribution of our estimator  $\hat{\theta}$ . To this end, we add the following conditions.

<sup>7</sup>If  $\phi_*$  is strictly convex and three times continuously differentiable on  $\mathbb{R}$  (such as the KL and PSN1 divergences), the requirement  $\zeta_{K,n}(\sqrt{K\mu_{K,n}/n} + B_{K,n}) \rightarrow 0$  can be weakened to (i) the second derivative  $\phi_*^{(2)}$  is bounded from above and away from zero, or (ii)  $\zeta_{K,n}(\sqrt{K\mu_{K,n}/n} + B_{K,n}) \lesssim 1$ .

<sup>8</sup>Although this uniform convergence rate is admittedly not optimal, it is sufficient to establish asymptotic normality of our estimator  $\hat{\theta}$  below. It is also an open question whether we can improve the convergence rate in (17) to establish the optimal rate as in [Belloni et al. \(2015\)](#) and [Chen and Christensen \(2015\)](#). Since our estimator  $\hat{\omega}$  and target  $\omega_0$  are more complicated than the least squares estimator for the conditional mean studied in those papers, such analysis will be technically more involving.

CONDITION N.

- (1) There exists a function  $r^h : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X)\mathbb{E}[h(X, Y)|X]]$  and

$$\mathbb{E}[\beta' \{ \omega_0(X)g_n(X) - r_n(X) \} - \{ \omega_0(X)\mathbb{E}[h(X, Y)|X] - r^h(X) \}]^2 = o(n^{-1}), \quad (19)$$

where  $\beta = \mathbb{E}[\phi_*^{(2)}(\lambda'_b g_n(X))g_n(X)g'_n(X)]^{-1}\mathbb{E}[\phi_*^{(2)}(\lambda'_b g_n(X))g_n(X)h(X, Y)]$ .

- (2)  $|\mathbb{E}_n[\phi_*^{(2)}(\lambda'_b g_n(X))g_n(X)g_n(X)'] - \mathbb{E}[\phi_*^{(2)}(\lambda'_b g_n(X))g_n(X)g_n(X)']| = O_p(\Gamma_{K,n})$  for some  $\Gamma_{K,n} \rightarrow 0$ .
- (3)  $\mathbb{E}[h(X, Y)^2|X = \cdot]$  is bounded from above,  $\mathbb{E}[|h(X, Y)|^{q_1/(1-q_1/q)}] < \infty$  for some  $q_1 \in (2, q]$  and  $\mathbb{E}[|r^h(X)|^q] < \infty$ , where  $q > 2$  is defined in Condition D(1).
- (4)  $\{Y_i, X_i\}_{i=1}^n$  is  $\alpha$ -mixing with mixing coefficients  $\{\alpha_{XY,m}\}_{m \in \mathbb{N}}$  satisfying

$$\sum_{m=1}^n \alpha_{XY,m}^{(a/(2+a)) \vee (1/2-1/q_1)} \lesssim 1,$$

for some  $a > 0$  and  $\mathbb{E}[|\Phi|^{2+a}] < \infty$ , where

$$\Phi = \omega_0(X)h(X, Y) - \theta_0 - \{ \omega_0(X)\mathbb{E}[h(X, Y)|X] - r^h(X) \}. \quad (20)$$

Condition N(1) is considered as the mean square continuity condition (cf. Assumption 5.3 in Newey (1994)) in our setup, which guarantees the  $\sqrt{n}$ -consistency of  $\hat{\theta}$  even though  $\hat{\omega}$  converges at a slower rate. Intuitively, (19) requires that  $\mathbb{E}[h(X, Y)|X = \cdot]$  is well approximated by the basis functions  $g_n(\cdot)$ . This requirement is typically verified by the results in functional analysis. The function  $r^h$  should be specified for each application. If  $r(X)$  is a vector of known constants (as in Example 1), we can simply set as  $r^h(X) = \theta_0$ . For Example 2, we can set as  $r^h(X) = \mathbb{E}[Y^*|X]$ . Proposition 2 below gives two examples where (19) is satisfied. Condition N(2) is analogous to Condition S(1). The convergence rate  $\Gamma_{K,n}$  will be  $\sqrt{\zeta_{K,n}^2 \log K/n}$  for the iid case (by Lemma 3(i)), and  $\sqrt{m\zeta_{K,n}^2 \log K/n}$  for the  $\beta$ -mixing case (by adapting Lemma 2.2 in Chen and Christensen (2015)). Condition N(3) contains mild assumptions on  $h$  and  $r^h$ . Condition N(4) requires  $\alpha$ -mixing for  $\{X_i, Y_i\}_{i=1}^n$  to apply a central limit theorem to  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi_i$ , where  $\Phi_i$  is the influence function for  $\hat{\theta}$ .

By imposing Condition N, the limiting distribution of the estimator  $\hat{\theta}$  is obtained as follows.

**THEOREM 2.** *Suppose that the conditions of Theorem 1 and Condition N hold true. In addition,  $\zeta_{K,n}^4 K \mu_{K,n} / \sqrt{n} \rightarrow 0$ ,  $\sqrt{n} \zeta_{K,n} B_{K,n} \rightarrow 0$ , and  $\sqrt{K \mu_{K,n}} \zeta_{K,n} \Gamma_{K,n} \rightarrow 0$ . Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),$$

where  $V = \lim_{n \rightarrow \infty} \text{Var}(\sqrt{n}\mathbb{E}_n[\Phi])$ .

This theorem says that our information theoretic estimator  $\hat{\theta}$  is  $\sqrt{n}$ -consistent and asymptotically normal. For iid data, the variance  $V$  becomes  $\mathbb{E}[\Phi^2]$ , which can be shown to be the semiparametric efficiency bound (see Section 4 for the case of the average treatment effect). Compared to Theorem 1, Theorem 2 requires more stringent conditions on  $K$ . However, we note that the condition  $\zeta_{K,n}^4 K \mu_{K,n} / \sqrt{n} \rightarrow 0$  can be weakened for some choices of  $\phi$ , such as the Pearson's  $\chi^2$  divergence.

The asymptotic variance  $V$  can be estimated by some heteroskedasticity autocorrelation consistent estimator. For example, based on Newey and West (1987),  $V$  can be estimated by

$$\hat{V} = \hat{\gamma}_0 + 2 \sum_{l=1}^{M_n} \left( \frac{M_n - l}{M_n} \right) \hat{\gamma}_l,$$

where  $\hat{\gamma}_l = (n - l)^{-1} \sum_{i=l+1}^n (\hat{\Phi}_i - n^{-1} \sum_{i=1}^n \hat{\Phi}_i)(\hat{\Phi}_{i-l} - n^{-1} \sum_{i=1}^n \hat{\Phi}_i)$  is the sample autocovariance of

$$\hat{\Phi}_i = \mathbb{I}\{X_i \in \mathcal{X}_n\} [\hat{\omega}(X_i)h(X_i, Y_i) - \hat{\theta} - \{\hat{\omega}(X_i)\hat{h}^X(X_i) - \hat{r}^h(X_i)\}],$$

$\hat{h}^X$  and  $\hat{r}^h$  are some nonparametric estimators of  $\mathbb{E}[h(X, Y)|X = \cdot]$  and  $r^h$ , respectively, and  $M_n$  is a tuning parameter. By adapting the proof of Newey and West (1987, Theorem 2) to the present context, the consistency of  $\hat{V}$  is obtained as follows.

**PROPOSITION 1.** *Suppose that the conditions of Theorem 2 hold true. Additionally, assume that  $E[|\Phi_i|^{4q_2+\delta}] < \infty$  for some  $q_2 > 1$  and  $\delta > 0$ ,  $\sum_{m=1}^n \alpha_{XY,m}^{1-1/(2q_2)} \lesssim 1$ ,  $\sup_{x \in \mathcal{X}_n} |\hat{h}^X(x) - \mathbb{E}[h(X, Y)|X = x]| = O_p(R_n)$  and  $\sup_{x \in \mathcal{X}_n} |\hat{r}^h(x) - r^h(x)| = O_p(R_n)$  for  $R_n = \zeta_{K,n} \sqrt{K \mu_{K,n} / n} + \zeta_{K,n} B_{K,n} + \eta_{K,n}$ ,  $M_n \rightarrow \infty$ , and  $M_n R_n \rightarrow 0$ . Then  $\hat{V} \xrightarrow{p} V$ .*

We close this section by providing some specific examples that satisfy (19) in Condition N(1).

**PROPOSITION 2.** *Suppose the assumptions in Theorem 2 except for (19) hold true.*

- (i) *Suppose  $r(X)$  is a vector of known constants,  $\mathbb{P}\{X \notin \mathcal{X}_n\} = o((Kn)^{-1})$  and*

$$\mathbb{E}[\{\mathbb{E}[h(X, Y)|X] - \lambda' g_n(X)\}^2] = o(n^{-1}), \tag{21}$$

*for some  $\lambda \in \mathbb{R}^K$ . Then (19) is satisfied with  $r^h(X) = \theta_0$ .*

- (ii) *In Example 2 on missing data, suppose*

$$\mathbb{E}[\{\mathbb{E}[Y^*|X] - \lambda' g_n(X)\}^2] = o(1),$$

*for some  $\lambda \in \mathbb{R}^K$ . Then (19) is satisfied with  $r^h(X) = \mathbb{E}[Y^*|X]$ .*

Based on Proposition 2, if  $r(X)$  is a vector of known constants, the influence function  $\Phi$  simplifies to  $\Phi = \omega_0(X)\{h(X, Y) - \mathbb{E}[h(X, Y)|X]\}$ .

### 3. HIGH-DIMENSIONAL CASE

In this section, we consider the high-dimensional case, where  $K = \dim(g)$  can be larger and grow faster than the sample size  $n$ . In this case,  $\hat{\lambda}$  in (14) is computed by the  $\ell_1$ -penalization. High dimensionality of  $g$  may be caused by either high dimensionality of the original data  $X$  or many transformations (or basis functions) based on low-dimensional  $X$ . In either case, as far as the latent weight function  $\omega_0$  in (11) admits certain sparse representation, our penalized estimator can consistently estimate  $\omega_0$  and the parameter of interest  $\theta_0$ . In Section 3.1, we study asymptotic properties of  $\hat{\omega}$  to estimate  $\omega_0$ . Then we consider three estimation approaches for  $\theta_0$ , debiasing (Section 3.2), post selection (Section 3.3), and targeted debiasing (Section 3.4), and present conditions to achieve  $\sqrt{n}$ -consistency and asymptotic normality for these estimators for  $\theta_0$ .

#### 3.1 Estimation of $\omega_0$

We first present asymptotic properties of  $\hat{\omega}$ . For the high-dimensional case, we impose the following assumptions.

**CONDITION D'.**  $\{X_i, Y_i\}_{i=1}^n$  is an iid triangular array. The support  $\mathcal{X} \subseteq \mathbb{R}^p$  of  $X$  is a Cartesian product of  $p$  intervals with nonempty interiors. Conditions D(3), (4), and (5) hold true.

For the high-dimensional case, we focus on the case of iid data. An extension to dependent data requires development of empirical process theory for dependent data in our setting, which is beyond the scope of this paper. We also do not use trimming for  $\mathcal{X}$ . Impacts from possible unbounded support are dealt implicitly by the growth rate of  $\sup_{x \in \mathcal{X}} \|g(x)\|_\infty$  and a uniform approximation assumption over  $\mathcal{X}$  (see the statement in Theorem 3).

To state additional conditions for the high-dimensional case, we introduce further notation. For an index subset  $S \subset \{1, \dots, K\}$ , let  $|S|$  be its cardinality,  $\lambda_S = (\lambda_{1,S}, \dots, \lambda_{K,S})'$  be a  $K$  dimensional vector with  $\lambda_{j,S} = \lambda_j \mathbb{1}\{j \in S\}$  for the  $j$ th component  $\lambda_j$  of  $\lambda$ , and  $\lambda_{S^c} = (\lambda_{1,S^c}, \dots, \lambda_{K,S^c})'$  with  $\lambda_{j,S^c} = \lambda_j \mathbb{1}\{j \notin S\}$ . So,  $\lambda_S$  and  $\lambda_{S^c}$  have nonzero elements only in the index set  $S$  and its complement  $S^c$ , respectively. Furthermore, let  $\mathcal{S}$  be a class of index sets.<sup>9</sup> We introduce the so-called compatibility condition.

**CONDITION C.** For each  $S \in \mathcal{S}$ , there exists some constant  $\phi_S > 0$  such that for all  $\lambda$  satisfying  $\|\lambda_{S^c}\|_1 \leq 3\|\lambda_S\|_1$ , it holds  $\|\lambda_S\|_1 \leq \phi_S^{-1} \sqrt{\lambda' \mathbb{E}[g(X)g(X)'] \lambda} \sqrt{|S|}$ .

This is a high level condition that bounds  $\|\lambda_S\|_1$  by the  $L_2$ -norm of its corresponding function  $\lambda'g(\cdot)$ . Such a compatibility condition is commonly employed in the high-dimensional statistics literature, such as the restricted eigenvalue condition in Bickel, Ritov, and Tsybakov (2009). Let

$$\mathfrak{E}(\lambda) = \mathbb{E}[\phi_*(\lambda'g(X)) - \lambda'r(X)] - \mathbb{E}[\phi_*(\lambda'_*g(X)) - \lambda'_*r(X)],$$

<sup>9</sup>Knowledge of  $\mathcal{S}$  can reflect researcher's prior on what might be important sets of covariates. In the worst case of no prior knowledge,  $\mathcal{S}$  should contain all possible index sets for covariates.

be the excess risk. Given  $\mathcal{S}$  with associated compatibility constants  $\{\phi_S : S \in \mathcal{S}\}$  in Condition C, the oracle  $\lambda_{\mathbf{o}}$  is defined as

$$\lambda_{\mathbf{o}} = \arg \min_{\lambda: S_{\lambda} \in \mathcal{S}} 2\mathcal{E}(\lambda) + \frac{8\alpha_n^2}{\phi_{S_{\lambda}}^2 \varrho} |S_{\lambda}|, \quad (22)$$

where  $S_{\lambda} = \{j : \lambda_j \neq 0\}$ ,  $\alpha_n$  is the penalty level in (14), and  $\varrho$  is a constant defined in Condition H below. Let  $Q_{\mathbf{o}}$  be the minimized value of (22) and  $\omega_{\mathbf{o}}(x) = \phi_{S_{\lambda_{\mathbf{o}}}}^{(1)}(\lambda'_{\mathbf{o}} g(x))$ . Note that  $\mathcal{E}(\lambda_{\mathbf{o}}) \geq \mathcal{E}(\lambda_*) = 0$  and a part of our sparsity assumption is characterized by the convergence rate of  $\mathcal{E}(\lambda_{\mathbf{o}})$  toward zero. Let

$$\nu_n(\lambda) = \mathbb{E}_n[\phi_*(\lambda' g(X)) - \lambda' r(X)] - \mathbb{E}[\phi_*(\lambda' g(X)) - \lambda' r(X)],$$

be an empirical process. We impose the following assumptions.

CONDITION H. For every  $\varepsilon > 0$  small enough and  $n$  large enough, there exist positive constants  $\sigma_{\varepsilon, n}$ ,  $\varrho$ , and  $A$  such that for  $M = \frac{Q_{\mathbf{o}}}{2\sigma_{\varepsilon, n}}$ ,

- (1)  $\mathbb{P}\{\sup_{\|\lambda - \lambda_{\mathbf{o}}\|_1 \leq M} |\nu_n(\lambda) - \nu_n(\lambda_{\mathbf{o}})| \leq \sigma_{\varepsilon, n} M\} \geq 1 - \varepsilon$ ,
- (2) for any  $\lambda$  satisfying  $\|\lambda - \lambda_{\mathbf{o}}\|_1 \leq M$ , it holds

$$\sup_{x \in \mathcal{X}} |(\lambda - \lambda_{\mathbf{o}})' g(x)| \leq A, \varrho(\lambda - \lambda_{\mathbf{o}})' \mathbb{E}[g(X)g(X)'](\lambda - \lambda_{\mathbf{o}}) \leq \varepsilon(\lambda),$$

- (3)  $\sigma_{\varepsilon, n} \leq \alpha_n/8$  and  $\alpha_n \propto \sqrt{\log K/n}$  for all  $n \in \mathbb{N}$ .

Condition H(1) controls the empirical process  $\nu_n(\lambda)$  in a neighborhood of the oracle  $\lambda_{\mathbf{o}}$ . Intuitively, we require that  $\nu_n(\lambda) - \nu_n(\lambda_{\mathbf{o}})$  will be small when  $\lambda$  is close to  $\lambda_{\mathbf{o}}$  in terms of the  $\ell_1$ -norm. The order of  $\sigma_{\varepsilon, n}$ , which is typically  $O(\sqrt{\log K/n})$ , can be derived by empirical process theory.<sup>10</sup> By Condition H(2), the excess risk  $\mathcal{E}(\lambda)$  can be bounded from below by a quadratic function of  $\lambda$  when  $\lambda$  is close to  $\lambda_{\mathbf{o}}$  in terms of the  $\ell_1$ -norm. Condition H(3) is on the penalty coefficient  $\alpha_n$ . First,  $\alpha_n$  should be large enough to offset the effect from  $\sigma_{\varepsilon, n}$ . Second, since  $\sigma_{\varepsilon, n}$  is typically of order  $O(\sqrt{\log K/n})$ , we set  $\alpha_n$  as the same order to achieve the fastest convergence in this typical case.<sup>11</sup>

Under these conditions, the convergence rate of  $\hat{\omega}$  and consistency of the parameter estimator  $\hat{\theta}$  are established as follows. Let  $\tilde{\zeta}_K = \sup_{x \in \mathcal{X}} \|g(x)\|_{\infty}$ ,  $s = |S_{\lambda_{\mathbf{o}}}|$ ,  $\kappa_{\mathbf{o}, n} = \mathcal{E}(\lambda_{\mathbf{o}}) \sqrt{\frac{n}{\log K}} \vee s \sqrt{\frac{\log K}{n}}$ , and  $\{\xi_n\}$  and  $\{\mathbf{s}_{\mathbf{o}, n}\}$  be positive sequences such that  $\|\mathbb{E}_n[g(X)g(X)']\|_{\infty} = O_p(\xi_n)$  and  $\sqrt{\mathbb{E}\{\omega_{\mathbf{o}}(X) - \omega_0(X)\}^2} \lesssim \mathbf{s}_{\mathbf{o}, n}$ , respectively.

<sup>10</sup>Since our objective function is Lipschitz in a neighborhood of  $\lambda_{\mathbf{o}}$ , probabilistic inequalities, such as Bühlmann and van de Geer (2011, Lemma 14.20), can be applied.

<sup>11</sup>Generally, there are two data-driven methods to select  $\alpha_n$ . First,  $\alpha_n$  may be chosen by cross validation although it might lack theoretical justification. Second,  $\alpha_n$  can be chosen as the smallest value such that Condition H holds with large probability. That is, we can set  $\alpha_n = 8\hat{\sigma}_{\varepsilon, n}$ , where  $\hat{\sigma}_{\varepsilon, n}$  is an estimator of  $\sigma_{\varepsilon, n}$ , based on the empirical process and moderate deviation theories. See Belloni et al. (2012) for further details.



**THEOREM 3.** *Suppose Conditions D', C, and H hold true.  $\phi_* : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  is strictly convex and three times continuously differentiable, and either (i) the second derivative  $\phi_*^{(2)}$  is bounded from above and away from zero, or (ii)  $\tilde{\zeta}_K \kappa_{\mathbf{o},n} \lesssim 1$ . Furthermore, assume that  $\mathbf{s}_{\mathbf{o},n} \rightarrow 0$ ,  $\kappa_{\mathbf{o},n} \xi_n^{1/2} \rightarrow 0$ , and  $\sup_{x \in \mathcal{X}} |\omega_{\mathbf{o}}(x) - \omega_0(x)| \lesssim 1$ . Then*

$$\sqrt{\mathbb{E}_n[\{\hat{\omega}(X) - \omega_0(X)\}^2]} = O_p(\kappa_{\mathbf{o},n} \sqrt{\xi_n} + \mathbf{s}_{\mathbf{o},n}), \quad (23)$$

and  $\hat{\theta} \xrightarrow{p} \theta_0$ . If we additionally assume  $\tilde{\zeta}_K \kappa_{\mathbf{o},n} \rightarrow 0$  and  $\sup_{x \in \mathcal{X}} |\omega_{\mathbf{o}}(x) - \omega_0(x)| \rightarrow 0$ , then

$$\sup_{x \in \mathcal{X}} |\hat{\omega}(x) - \omega_0(x)| \xrightarrow{p} 0. \quad (24)$$

This theorem, a counterpart of Theorem 1 for the high-dimensional case, establishes the empirical  $L_2$  convergence rate of  $\hat{\omega}$ , which is used to derive the consistency of  $\hat{\theta}$ . Note that we only require boundedness for the uniform approximation error  $\sup_{x \in \mathcal{X}} |\omega_{\mathbf{o}}(x) - \omega_0(x)|$  by the oracle. The object  $\tilde{\zeta}_K$  depends on the choice of basis functions  $g$  and  $\mathcal{X}$ . For example, if  $g$  is a vector of polynomials over  $\mathcal{X} = [0, 1]^p$ , it holds  $\tilde{\zeta}_K = O(1)$ . The object  $\xi_n$  measures the growth rate of the sup-norm of  $\mathbb{E}_n[g(X)g(X)']$ . It can be controlled by Hoeffding's inequality, and is typically of order  $O(\|\mathbb{E}[g(X)g(X)']\|_\infty)$  (or  $O(1)$  for certain basis functions). In this case, if we further assume  $\mathcal{E}(\lambda_{\mathbf{o}}) = O(s \log K/n)$  and  $\mathbf{s}_{\mathbf{o},n} = O(s \sqrt{\log K/n})$ , then the empirical  $L_2$  convergence rate of  $\hat{\omega}$  is of order  $O_p(s \sqrt{\log K/n})$  and the dimension  $K$  may grow faster than  $n$  even at an exponential rate. For the high-dimensional case, the approximation bias for  $\omega_0$  tends to be larger and is controlled by the approximate sparsity assumption that requires sufficiently fast decay rates of the excess risk  $\mathcal{E}(\lambda_{\mathbf{o}})$  and approximation error  $\mathbf{s}_{\mathbf{o},n}$ . A byproduct of this theorem is the uniform consistency in (24) under additional assumptions.

Although the estimator  $\hat{\theta}$  is consistent for  $\theta_0$ , it does not achieve the  $\sqrt{n}$ -consistency and asymptotic normality in general. In the following subsections, we present three approaches to modify the estimator for  $\theta_0$  to achieve the  $\sqrt{n}$ -consistency and asymptotic normality.

### 3.2 Debiased estimator for $\theta_0$

In this subsection, we consider a debiased estimation method for  $\theta_0$  in the high-dimensional setup. It is well known that plug-in methods to estimate finite dimensional objects, where the first step is implemented by the lasso, typically cannot achieve the  $\sqrt{n}$ -consistency. In statistics literature, several procedures are proposed to debias the lasso estimators to achieve the  $\sqrt{n}$ -consistency and asymptotic normality for finite dimensional objects of interest (see, e.g., Zhang and Zhang (2014) and van de Geer et al. (2014)). It is natural to ask whether such debiasing procedures may be applied to our setup. However, in our setting, it seems the debiasing procedure achieves  $\sqrt{n}$ -consistency and asymptotic normality for  $\theta_0$  only under certain stringent conditions.

To illustrate this point, suppose  $\phi_*^{(2)}(\cdot) = c_* > 0$  for some known constant  $c_*$  (e.g., by choosing  $\phi(x) = \frac{1}{2}x^2$ ). Let  $\hat{\kappa} = (\text{sign}(\hat{\lambda}_1), \dots, \text{sign}(\hat{\lambda}_K))'$  and  $\hat{\Theta}$  be an approximation of

the ‘inverse’ of  $\mathbb{E}_n[g(X)g(X)']$  (which may not exist in the high-dimensional case). Here, we consider the debiased estimator

$$\hat{\theta}_{DB} = \mathbb{E}_n[\{\phi_*^{(1)}(\hat{\lambda}'g(X)) + \alpha_n g(X)' \hat{\Theta} \hat{\kappa}\} h(X, Y)], \quad (25)$$

where the additional term  $\alpha_n g(\cdot)' \hat{\Theta} \hat{\kappa}$  corrects the first-order bias from the plug-in estimation by  $\hat{\lambda}$ . We note that this additional term will be different if we drop the requirement  $\phi_*^{(2)}(\cdot) = c_* > 0$ . To establish the  $\sqrt{n}$ -consistency and asymptotic normality of  $\hat{\theta}_{DB}$ , we impose the following assumptions. Let  $\hat{\beta}_{DB} = \hat{\Theta}' \mathbb{E}_n[g(X)h(X, Y)]$ .

CONDITION DB.

- (1) There exist functions  $r^h, \tilde{r}^h, \tilde{h}^X : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X)\mathbb{E}[h(X, Y)|X]]$ ,  $\mathbb{E}[\tilde{r}^h(X)] = \mathbb{E}[\omega_0(X)\tilde{h}^X(X)]$ , and

$$\begin{aligned} \mathbb{E}_n[\hat{\beta}'_{DB}\{\omega_0(X)g(X) - r(X)\} - \{\omega_0(X)\tilde{h}^X(X) - \tilde{r}^h(X)\}]^2 &= o_p(n^{-1}), \\ (\mathfrak{s}_{\mathbf{0},n}^2 + \mathfrak{s}_{\mathbf{0},n}n^{-1/2})\mathbb{E}_n[\tilde{h}^X(X) - \hat{\beta}'_{DB}g(X)]^2 &= o_p(n^{-1}). \end{aligned}$$

- (2)  $\sqrt{n}\kappa_{\mathbf{0},n}\|\mathbb{E}_n[h(X, Y)g(X)]\|_\infty\|I - \mathbb{E}_n[g(X)g(X)']\hat{\Theta}\|_1 = o_p(1)$ .

Condition DB highlights two key requirements for achieving the  $\sqrt{n}$ -consistency and asymptotic normality of the debiased estimator  $\hat{\theta}_{DB}$ . Condition DB(1) is a natural extension of Condition N(1) under the high-dimensional case. It requires that  $\hat{\beta}'_{DB}g(\cdot)$  should converge fast enough to some function  $\tilde{h}^X(\cdot)$ . Intuitively,  $\tilde{h}^X(\cdot)$  can be understood as an approximation of  $\mathbb{E}[h(X, Y)|X = \cdot]$ . This is a key condition to correct the bias from the second step to compute  $\hat{\theta}_{DB}$ . On the other hand, Condition DB(2) controls the  $\ell_1$ -regularization bias. It says the matrix  $\hat{\Theta}$  should be selected to guarantee  $\|I - \mathbb{E}_n[g(X)g(X)']\hat{\Theta}\|_1$  to be sufficiently small.

The  $\sqrt{n}$ -normality of the debiased estimator  $\hat{\theta}_{DB}$  is obtained as follows. Let  $\{\tau_n\}$  be a positive sequence such that  $\sqrt{\mathbb{E}[\{\mathbb{E}[h(X, Y)|X] - \tilde{h}^X(X)\}^2]} \vee \sqrt{\mathbb{E}[\{r^h(X) - \tilde{r}^h(X)\}^2]} \lesssim \tau_n$ .

**THEOREM 4.** *Suppose Conditions D', C, H, and DB hold true and  $\phi_*^{(2)}(\cdot) = c_* > 0$  for some known constant  $c_*$ . If  $\sup_{x \in \mathcal{X}} \mathbb{E}[h(X, Y)^2|X = x] \lesssim 1$ ,  $\mathfrak{s}_{\mathbf{0},n} \rightarrow 0$ ,  $\tau_n \rightarrow 0$ , and  $\sqrt{n}\mathfrak{s}_{\mathbf{0},n}\tau_n \rightarrow 0$ , then*

$$\sqrt{n}(\hat{\theta}_{DB} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}[\Phi^2]).$$

Theorem 4 gives conditions under which the debiased estimator  $\hat{\theta}_{DB}$  can achieve the  $\sqrt{n}$ -normality. It seems the requirements on  $\hat{\Theta}$  listed in Condition DB are difficult to avoid. In fact, our debiasing procedure may be considered as an intermediate procedure between the parametric debiasing of Zhang and Zhang (2014) and van de Geer et al. (2014), and the complete debiasing of Farrell (2015) and Belloni et al. (2012). It is beyond the scope of this paper to study a practical way of finding the matrix  $\hat{\Theta}$  (e.g., by adapting the lasso with nodewise regression in van de Geer et al. (2014)), and we leave this for future research.

3.3 Post selection estimator for  $\theta_0$ 

Given that the debiasing procedure in the last subsection requires relatively strong conditions, we propose the following post selection method to obtain a  $\sqrt{n}$ -consistent estimator for  $\theta_0$ :

- (1) Compute  $\hat{\lambda}$  in (14) for the high-dimensional case. Let  $\mathbf{s} = |\hat{S}|$  be the cardinality of the selected set  $\hat{S} = \{j : \hat{\lambda}_j \neq 0\}$ .
- (2) Let  $g_{\mathbf{s}}$  and  $r_{\mathbf{s}}$  be the  $\mathbf{s}$ -dimensional functions corresponding to the selected set  $\hat{S}$ . Implement (14) for the low-dimensional case (i.e., without the  $\ell_1$ -penalty) based on  $g_{\mathbf{s}}$  and  $r_{\mathbf{s}}$ . Denote the solution of this step as

$$\hat{\Lambda} = \arg \min_{\Lambda \in \mathbb{R}^{\mathbf{s}}} \mathbb{E}_n[\phi_*(\Lambda' g_{\mathbf{s}}(X)) - \Lambda' r_{\mathbf{s}}(X)]. \quad (26)$$

- (3) Construct the post selection estimator as

$$\tilde{\theta} = \mathbb{E}_n[\phi_*^{(1)}(\hat{\Lambda}' g_{\mathbf{s}}(X))h(X, Y)]. \quad (27)$$

To study asymptotic properties of the post selection estimator  $\tilde{\theta}$ , we introduce some notation. Let  $\Lambda_* = \arg \min_{\Lambda \in \mathbb{R}^{\mathbf{s}}} \mathbb{E}[\phi_*(\Lambda' g_{\mathbf{s}}(X)) - \Lambda' r_{\mathbf{s}}(X)]$  be the population counterpart of (26), and  $\omega_*(x) = \phi_*^{(1)}(\Lambda_*' g_{\mathbf{s}}(x))$ , which is an approximation of  $\omega_0$  using the selected vector  $g_{\mathbf{s}}$ . Note that  $\omega_*$  could be different from  $\omega_{\mathbf{o}}$  selected by the oracle  $\lambda_{\mathbf{o}}$ . Also, define

$$\beta_{\mathbf{s}} = \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}}(X))g_{\mathbf{s}}(X)g_{\mathbf{s}}(X)']^{-1} \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}}(X))g_{\mathbf{s}}(X)\mathbb{E}[h(X, Y)|X]],$$

and  $\tilde{h}^X(x) = \beta_{\mathbf{s}}' g_{\mathbf{s}}(x)$ . We impose the following conditions.

**CONDITION N'**. There exist functions  $r^h, \tilde{r}^h : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X) \times \mathbb{E}[h(X, Y)|X]]$ ,  $\mathbb{E}[\tilde{r}^h(X)] = \mathbb{E}[\omega_0(X)\tilde{h}^X(X)]$ , and

$$\mathbb{E}[\beta_{\mathbf{s}}'\{\omega_0(X)g_{\mathbf{s}}(X) - r_{\mathbf{s}}(X)\} - \{\omega_0(X)\tilde{h}^X(X) - \tilde{r}^h(X)\}]^2 \rightarrow 0. \quad (28)$$

Condition N' can be viewed as an extension of the mean square continuity (as in Assumption 5.3 of Newey (1994)) for imperfect model selection, where  $\tilde{h}^X(\cdot) = \beta_{\mathbf{s}}' g_{\mathbf{s}}(\cdot)$  is understood as an approximation of  $\mathbb{E}[h(X, Y)|X = \cdot]$  based on the selected basis functions  $g_{\mathbf{s}}$ . In the case of imperfect model selection (i.e.,  $\hat{S} \neq S_{\lambda_{\mathbf{o}}}$ ),  $\omega_*$  and  $\tilde{h}^X$  may not approximate  $\omega_0$  and  $h^X$  well enough, respectively. We impose the following conditions for those approximation errors.

**CONDITION S'**. For each  $n$ , all eigenvalues of  $\mathbb{E}[g_{\mathbf{s}}(X)g_{\mathbf{s}}(X)']$  are bounded from above and away from zero, conditional on the selected set  $\hat{S}$ . Also, for some positive sequences  $\{\mathbf{s}_{\mathbf{s},n}\}$  and  $\{\tau_{\mathbf{s},n}\}$ ,

$$\sqrt{\mathbb{E}[\{\omega_0(X) - \omega_*(X)\}^2]} \lesssim \mathbf{s}_{\mathbf{s},n}, \quad (29)$$

$$\sqrt{\mathbb{E}[\{\mathbb{E}[h(X, Y)|X] - \tilde{h}^X(X)\}^2]} \lesssim \tau_{\mathbf{s},n}. \quad (30)$$

Because of the imperfect model selection,  $\varsigma_{\mathbf{s},n}$  and  $\tau_{\mathbf{s},n}$  may not vanish sufficiently fast as in Theorem 2. Instead, we only require  $\varsigma_{\mathbf{s},n}$  and  $\tau_{\mathbf{s},n}$  to be  $O(1)$ . Let  $\zeta_{\mathbf{s}} = \sup_{x \in \mathcal{X}} |g_{\mathbf{s}}(x)|$ .

CONDITION I'.  $\phi_* : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  is strictly convex and three times continuously differentiable,

$$\sup_{x \in \mathcal{X}} \phi_*^{(2)}(\Lambda'_* g_{\mathbf{s}}(x)) \lesssim 1, \quad \text{and} \quad \sup_{\Lambda \in \mathbb{R}^{\mathbf{s}}: |\Lambda - \Lambda_*| \lesssim \sqrt{\zeta_{\mathbf{s}}^2/n}} \mathbb{E}_n[\phi_*^{(3)}(\Lambda'_* g_{\mathbf{s}}(X))^2] = O_p(1).$$

Condition I' is a counterpart of Condition I, and imposes additional requirements on the conjugate function  $\phi_*$ , which can be trivially satisfied for some divergence, such as  $\phi(x) = \frac{1}{2}x^2$ . This condition can be also satisfied if  $\sup_{x \in \mathcal{X}} |[\phi_*^{(1)}]^{-1}(\omega_0(x)) - \Lambda'_* g_{\mathbf{s}}(x)| \lesssim 1$ , that is, the selected component  $\Lambda'_* g_{\mathbf{s}}(\cdot)$  is not too far from  $[\phi_*^{(1)}]^{-1}(\omega_0(\cdot))$ .

Under these conditions, the  $\sqrt{n}$ -normality of the post selection estimator  $\tilde{\theta}$  is obtained as follows.

THEOREM 5. *Suppose Conditions D', S', I', and N' hold true. In addition,  $\zeta_{\mathbf{s}}^2 \log \mathbf{s}/n \rightarrow 0$ ,  $\zeta_{\mathbf{s}}^6/\sqrt{n} \rightarrow 0$ , and  $\mathbb{E}[(\Phi + v_1 + v_2 + v_3)^2] < \infty$ , where  $\Phi$  is defined in (20). Then*

$$\sqrt{n}(\tilde{\theta} - \theta_0 + b) \xrightarrow{d} N(0, \mathbb{E}[(\Phi + v_1 + v_2 + v_3)^2]), \quad (31)$$

where  $b = \mathbb{E}[(\omega_0(X) - \omega_*(X))(h^X(X) - \tilde{h}^X(X))]$ ,

$$v_1 = (\omega_*(X) - \omega_0(X))(h(X, Y) - h^X(X)),$$

$$v_2 = \omega_0(X)(h^X(X) - \tilde{h}^X(X)) + \tilde{r}^h(X) - r^h(X),$$

$$v_3 = (\omega_*(X) - \omega_0(X))(h^X(X) - \tilde{h}^X(X)) - \mathbb{E}[(\omega_*(X) - \omega_0(X))(h^X(X) - \tilde{h}^X(X))].$$

Furthermore, if  $\varsigma_{\mathbf{s},n} \rightarrow 0$ ,  $\tau_{\mathbf{s},n} \rightarrow 0$ , and  $\sqrt{n}\varsigma_{\mathbf{s},n}\tau_{\mathbf{s},n} \rightarrow 0$ , then

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}[\Phi^2]). \quad (32)$$

This theorem characterizes effects of the imperfect model selection from the first step lasso procedure.  $b$  is an additional bias term, and  $v_1$ ,  $v_2$ , and  $v_3$  are additional variance terms. In particular,  $v_1$  is due to imperfect approximation of  $\omega_0$  by  $\omega_*$ ,  $v_2$  is due to imperfect approximation of  $h^X$  by  $\tilde{h}^X$ , and  $v_3$  is due to slow approximation of both  $h^X$  and  $\omega_0$ . For the case of (32), we can conduct inference on  $\theta_0$  by estimating the asymptotic variance  $\mathbb{E}[\Phi^2]$ . On the other hand, if the imperfect model selection is severe in the sense of  $\varsigma_{\mathbf{s},n} = \tau_{\mathbf{s},n} = O(1)$ , the post selection estimator  $\tilde{\theta}$  will have the asymptotic bias  $b$  and additional terms in the variance as in (31). Valid inference in this general case is left for future research.

3.4 Targeted debiasing estimator for  $\theta_0$ 

In this subsection, we discuss a targeted debiasing procedure, which is between the debiasing procedure for the whole vector  $\hat{\lambda}$  in Section 3.2 and post selection procedure in Section 3.3.

Without loss of generality, we assume the first  $\mathbf{s}$  elements of  $\{1, \dots, K\}$  are selected by  $\hat{\lambda}$ . Suppose that  $\hat{\Theta}_{\mathbf{s}}$  is a good approximation of the inverse of the  $\mathbf{s} \times \mathbf{s}$  matrix  $\mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}}g_{\mathbf{s}}(X))g_{\mathbf{s}}(X)g_{\mathbf{s}}(X)']$ . For example, a practical choice for  $\hat{\Theta}_{\mathbf{s}}$  would be the empirical counterpart  $(\mathbb{E}_n[\phi_*^{(2)}(\hat{\lambda}'_{\mathbf{s}}g_{\mathbf{s}}(X))g_{\mathbf{s}}(X)g_{\mathbf{s}}(X)'])^{-1}$ . Define the targeted debiasing version  $\hat{\lambda}_{\text{TD}}$  of  $\hat{\lambda}$  as

$$\hat{\lambda}_{\text{TD}} = (\hat{\Lambda}'_{\mathbf{s}}, 0'_{K-\mathbf{s}})', \quad \hat{\Lambda}_{\mathbf{s}} = \hat{\lambda}_{\mathbf{s}} + \hat{\Theta}_{\mathbf{s}}\alpha_n\hat{\kappa}_{\mathbf{s}},$$

and  $0_{K-\mathbf{s}}$  is the  $(K - \mathbf{s})$ -dimensional vector of zeros. That is, we only correct the bias for the selected elements by  $\hat{S}$ . Then  $\theta_0$  is estimated by

$$\hat{\theta}_{\text{TD}} = \mathbb{E}_n[\phi_*^{(1)}(\hat{\lambda}'_{\text{TD}}g(X))h(X, Y)]. \quad (33)$$

Let  $\tilde{\gamma}_n = \kappa_{\mathbf{o}, n} \vee \sqrt{\mathbf{s} \log K/n}$ ,  $\omega_{\mathbf{s}}(x) = \phi_*^{(1)}(\lambda'_{\mathbf{os}}g_{\mathbf{s}}(x))$ , and  $\tilde{h}_{\text{TD}}^X(x) = \tilde{\beta}'_{\mathbf{s}}g_{\mathbf{s}}(x)$ , where

$$\tilde{\beta}_{\mathbf{s}} = \mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}}g_{\mathbf{s}}(X))g_{\mathbf{s}}(X)g_{\mathbf{s}}(X)']^{-1}\mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}}g_{\mathbf{s}}(X))g_{\mathbf{s}}(X)\mathbb{E}[h(X, Y)|X]].$$

To derive the limiting distribution of  $\hat{\theta}_{\text{TD}}$ , we add the following assumptions.

CONDITION TD.

- (1) There exist functions  $r^h, \tilde{r}_{\text{TD}}^h : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X)\mathbb{E}[h(X, Y)|X]]$ ,  $\mathbb{E}[\tilde{r}_{\text{TD}}^h(X)] = \mathbb{E}[\omega_0(X)\tilde{h}_{\text{TD}}^X(X)]$ , and

$$\mathbb{E}[\{\tilde{\beta}'_{\mathbf{s}}(\omega_0(X)g_{\mathbf{s}}(X) - r(X)) - (\omega_0(X)\tilde{h}_{\text{TD}}^X(X) - \tilde{r}_{\text{TD}}^h(X))\}^2] \rightarrow 0.$$

- (2)  $|\hat{\Theta} - Q^{(2)}(\lambda_{\mathbf{os}})^{-1}| = O_p(Q_n)$  and  $\sqrt{n}\tilde{\gamma}_n\zeta_{\mathbf{s}}Q_n \rightarrow 0$ .

- (3) Condition I' holds true with  $\Lambda_*$  and  $\sqrt{\zeta_{\mathbf{s}}^2/n}$  replaced by  $\lambda_{\mathbf{os}}$  and  $\tilde{\gamma}_n$ , respectively.

- (4) Condition S' holds true with  $\omega_*$  and  $\tilde{h}^X$  replaced by  $\omega_{\mathbf{s}}$  and  $\tilde{h}_{\text{TD}}^X$ , respectively.

Condition TD(1) is a counterpart of Condition N'(2). The roles of Conditions TD(3)–(4) for the targeted debiasing procedure are same as Conditions I' and S' for the post-selection procedure, respectively. Condition TD(2) is concerned with quality of the targeted debiasing procedure. Under these conditions, the targeted debiasing estimator  $\hat{\theta}_{\text{TD}}$  admits the same asymptotic representation as the post selection estimator.

**THEOREM 6.** *Suppose Conditions D', C, H, and TD hold true. Additionally, assume  $\sqrt{n}\kappa_{\mathbf{o}, n}\zeta_{\mathbf{s}}^4 \rightarrow 0$ ,  $\sqrt{n}\zeta_{\mathbf{s}}^2\tilde{\gamma}_n^2 \rightarrow 0$ , and  $\mathbb{E}[(\Phi + \tilde{v}_1 + \tilde{v}_2 + \tilde{v}_3)^2] < \infty$ . Then*

$$\sqrt{n}(\hat{\theta}_{\text{TD}} - \theta_0 + \tilde{b}) \xrightarrow{d} N(0, \mathbb{E}[(\Phi + \tilde{v}_1 + \tilde{v}_2 + \tilde{v}_3)^2]).$$

where  $\tilde{b}, \tilde{v}_1, \tilde{v}_2$ , and  $\tilde{v}_3$  are same as those in Theorem 5 with replacements of  $\omega_*$ ,  $\tilde{h}^X$ , and  $\tilde{r}^h$  with  $\omega_{\mathbf{s}}$ ,  $\tilde{h}_{\text{TD}}^X$ , and  $\tilde{r}_{\text{TD}}^h$ , respectively.

## 4. THEORETICAL APPLICATION: TREATMENT EFFECT

In this section, we extend Example 2 in Section 1 and consider estimation of the average treatment effect. Let  $D_i$  be the indicator of a treatment for individual  $i = 1, \dots, n$  ( $D_i = 1$  and  $0$  mean treated and not treated, resp.). For each  $i$ , there exist two potential outcomes,  $Y_i(1)$  if treated and  $Y_i(0)$  if not treated. The observable outcome is  $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$ . Also, let  $X_i$  be covariates of individual  $i$ . Based on a random sample  $\{D_i, Y_i, X_i\}_{i=1}^n$ , we wish to estimate the average treatment effect  $\tau = \mathbb{E}[Y(1) - Y(0)]$ . Under the unconfoundedness and overlap assumptions,  $\tau$  can be identified as (Rosenbaum and Rubin (1983))

$$\tau = \mathbb{E}[\omega^t(X)DY] - \mathbb{E}[\omega^u(X)(1-D)Y] \equiv \theta^t - \theta^u,$$

where  $\omega^t(x) = \pi(x)^{-1}$ ,  $\omega^u(x) = \{1 - \pi(x)\}^{-1}$ , and  $\pi(x) = \Pr\{D = 1 | X = x\}$  is the propensity score. We treat  $\omega^t$  and  $\omega^u$  as latent weight functions, and construct moment conditions as in (1) by utilizing the property of the propensity score:

$$\mathbb{E}[D\omega^t(X)g(X)] = \mathbb{E}[(1-D)\omega^u(X)g(X)] = \mathbb{E}[g(X)], \quad (34)$$

for any  $g$ . By applying our methodology based on (34), the weight function  $\omega^t$  can be estimated by

$$\begin{cases} \hat{\omega}^t(x) = \phi_*^{(1)}(\hat{\lambda}'_1 g(x)) & \text{(low-dimensional case),} \\ \tilde{\omega}^t(x) = \phi_*^{(1)}(\hat{\Lambda}'_1 g(x)) & \text{(high-dimensional case),} \end{cases}$$

where

$$\hat{\lambda}_1 = \begin{cases} \arg \min_{\lambda} \mathbb{E}_n[D\phi_*(\lambda'g(X)) - \lambda'g(X)] & \text{(low-dimensional case),} \\ \arg \min_{\lambda} \mathbb{E}_n[D\phi_*(\lambda'g(X)) - \lambda'g(X)] + \alpha_{1n}\|\lambda\|_1 & \text{(high-dimensional case),} \end{cases}$$

$$\hat{\Lambda}_1 = \arg \min_{\Lambda \in \mathbb{R}^{\mathbf{s}_1}} \mathbb{E}_n[D\phi_*(\Lambda'g_{\mathbf{s}_1}(X)) - \Lambda'g_{\mathbf{s}_1}(X)],$$

and  $g_{\mathbf{s}_1}$  is the  $\mathbf{s}_1$ -dimensional functions corresponding to  $\hat{S}_1 = \{j : \hat{\lambda}_{1j} \neq 0\}$ .

Then  $\theta^t$  can be estimated by  $\hat{\theta}^t = \mathbb{E}_n[\hat{\omega}^t(X)DY]$  for the low-dimensional case, or by the post-selection estimator  $\tilde{\theta}^t = \mathbb{E}_n[\tilde{\omega}^t(X)DY]$  for the high-dimensional case. Similarly, we can estimate  $\omega^u$  and  $\theta^u$  (by replacing  $D$  with  $(1 - D)$ ). The average treatment effect  $\tau$  can be estimated by  $\hat{\tau} = \hat{\theta}^t - \hat{\theta}^u$  for the low-dimensional case, or  $\tilde{\tau} = \tilde{\theta}^t - \tilde{\theta}^u$  for the high-dimensional case. By applying the results in the previous sections, we obtain the following corollary.

**COROLLARY 1.** *Consider the setup of this section. Suppose  $D \perp (Y(1), Y(0)) | X$  (unconfoundedness condition), and the propensity score  $\pi$  is bounded away from 0 and 1 over the compact support  $\mathcal{X}$  (overlap condition). Furthermore, assume  $\mathbb{E}[Y^2(0)] < \infty$  and  $\mathbb{E}[Y^2(1)] < \infty$ .*

(i) [Low-dimensional case] Under the assumptions of Theorem 2, in particular, if

$$\begin{aligned} \sup_{x \in \mathcal{X}} |\mathbb{E}[Y(1)|X = x] - \lambda'_1 g(x)| &\rightarrow 0, \\ \sup_{x \in \mathcal{X}} |\mathbb{E}[Y(0)|X = x] - \lambda'_0 g(x)| &\rightarrow 0, \end{aligned}$$

for some  $\lambda_1, \lambda_0 \in \mathbb{R}^K$ , it holds

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \Sigma),$$

$$\text{where } \Sigma = \mathbb{E}[\{\mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0)|X] - \tau\}^2 + \frac{\text{Var}(Y(1)|X)}{\pi(X)} + \frac{\text{Var}(Y(0)|X)}{1-\pi(X)}].$$

(ii) [High-dimensional case] Under the assumptions of Theorem 5, it holds

$$\sqrt{n}(\tilde{\tau} - \tau + b_{ps}) \xrightarrow{d} N(0, \Sigma_{ps}),$$

where the formula of  $b_{ps} \geq 0$  and  $\Sigma_{ps} \geq \Sigma$  can be found accordingly via Theorem 5.

Proofs are similar to those of Theorems 2 and 5. This corollary may be considered as an extension of [Chan, Yam, and Zhang \(2016\)](#) to the high-dimensional case by using the  $\ell_1$ -penalized estimator. Note that the asymptotic variance  $\Sigma$  is the semiparametric efficiency bound for  $\tau$  established in [Hahn \(1998\)](#).

## 5. EMPIRICAL APPLICATION: STOCHASTIC DISCOUNT FACTOR

To illustrate the performance of our proposed method, we consider [Example 1](#) and estimate the normalized SDF in an equity market. We compare out-of-sample performances of the proposed method and Fama–French three-factor method.

To make the results comparable with existing literature (e.g., [Fama and French \(1993\)](#), [Lewellen, Nagel, and Shanken \(2010\)](#), and [Ghosh, Julliard, and Taylor \(2016\)](#)), our out-of-sample evaluation covers from July 1963 to December 2010. All returns data are taken from Kenneth French's data library and are quoted in %. We note the approach adopted by [Ghosh, Julliard, and Taylor \(2016\)](#) is a special case of ours for the low-(and fixed) dimensional case using the KL divergence without trimming.

Our major findings are as follows. (i) In the low-dimensional setup where the number of portfolios in the market is small, predictability of our method is at least as good as the Fama–French three-factors model, and our method shows lower cross-sectional errors. (ii) In a relatively high-dimensional setup where the number of portfolios is similar to the number of training periods, upon choosing suitable penalty levels, our method outperforms the Fama–French three factors model. Also [Ghosh, Julliard, and Taylor's \(2016\)](#) method shows erratic behaviors in this case. (iii) Our methods are robust against different choices of  $\phi$  and trimming, but the SDFs extracted by different  $\phi$  have different shapes, especially in terms of skewness and kurtosis. (iv) In a low-dimensional case, the KL divergence performs better than the Pearson's  $\chi^2$  divergence. In high-dimensional case with penalization, the Pearson's  $\chi^2$  divergence performs better than the KL divergence.



### 5.1 Step-by-step implementation

We first give a detailed procedure of implementing our proposed method to estimate out-of-sample SDF and test its cross-sectional predictability.

**5.1.1 Form training and testing samples in rolling windows** Let  $\mathcal{L}$  be a set of indexes of years for which we want to estimate monthly out-of-sample SDFs. Let  $R_t$  be a  $K - 1$  dimensional vector of portfolio excess returns in month  $t$ . Following the convention in empirical finance, in July of each year  $l \in \mathcal{L}$ , we form a training sample  $\{R_t\}_{t \in \mathcal{T}_1(l)}$  of monthly returns in the past 30 years and a testing sample  $\{R_t\}_{t \in \mathcal{T}_2(l)}$  of returns 12 months ahead. That is, in each rolling window, the training sample size is  $|\mathcal{T}_1(l)| = 360$ , and the testing sample size is  $|\mathcal{T}_2(l)| = 12$ .<sup>12</sup> Let  $\{R_t\}_{t \in \tilde{\mathcal{T}}_1(l)}$  be the sample of monthly returns after trimming. The actual training sample size after trimming is  $|\tilde{\mathcal{T}}_1(l)|$ . If there is no trimming, it holds  $\tilde{\mathcal{T}}_1(l) = \mathcal{T}_1(l)$ .

**5.1.2 Out-of-sample prediction** We create a grid of possible values for the penalty  $\alpha_n$ . For each  $\alpha_n$  in the grid points and each  $l \in \mathcal{L}$ , we compute the following:

- (1) If the KL divergence is used, the out-of-sample prediction for the SDF is given by

$$\hat{\omega}_j = \frac{\exp(\hat{\lambda}' R_j)}{|\mathcal{T}_2(l)|^{-1} \sum_{t \in \mathcal{T}_2(l)} \exp(\hat{\lambda}' R_t)}, \quad (35)$$

for each  $j \in \mathcal{T}_2(l)$ , where

$$\hat{\lambda} = \arg \min_{\lambda \in \mathbb{R}^{K-1}} |\tilde{\mathcal{T}}_1(l)|^{-1} \sum_{t \in \tilde{\mathcal{T}}_1(l)} \exp(\lambda' R_t) + \alpha_n \|\lambda\|_1.$$

- (2) For other divergences, the out-of-sample prediction for the SDF is given by

$$\hat{\omega}_j = \phi_*^{(1)}(\hat{\lambda}' X_j), \quad (36)$$

for each  $j \in \mathcal{T}_2(l)$ , where  $X_j = (1, R_j)'$ ,

$$\hat{\lambda} = \arg \min_{\lambda \in \mathbb{R}^K} |\tilde{\mathcal{T}}_1(l)|^{-1} \sum_{t \in \tilde{\mathcal{T}}_1(l)} \phi_*(\lambda' X_t) - \lambda' \mathbf{e}_1 + \alpha_n \|\lambda\|_1,$$

and  $\mathbf{e}_1 = (1, 0, \dots, 0)'$  is a  $K$  dimensional vector.

- (3) Repeat (1) and (2) for each year  $l \in \mathcal{L}$ .

**5.1.3 Testing cross-sectional predictability** Based on the constructed time series of the predicted SDFs  $\{\hat{\omega}_j\}_{j \in \mathcal{T}_2(l), l \in \mathcal{L}}$ , we test its cross-sectional predictability using standard two-pass regression in empirical finance (Fama and MacBeth (1973) and Cochrane (2009)). Empirical performances of extracted out-of-sample SDFs depend on the penalty

<sup>12</sup>This is except for the last rolling window in which the testing sample size is 6.

level  $\alpha_n$  for our method. We recommend to select  $\alpha_n$  in a given grid to lead to the best predictability. There are different measures of predictability in the literature. In this empirical exercise, we set the optimal penalty level as the one that leads to the smallest magnitude of the estimated constant and the largest adjusted  $R^2$ .

## 5.2 Main empirical results

**5.2.1 Low-dimensional case: 25 size and book-to-market portfolios** This is arguably a low-dimensional scenario. We present results using three divergences: KL, PSN1, and PSN2. Table 1 presents some summary statistics of predicted SDFs without penalization. As we can see, the predicted SDFs by KL are positively skewed with high kurtosis compared to the ones by PSN1 and PSN2. By truncating at zero, PSN2 excludes negative values for predicted SDFs and yields positive skewness.

Table 2 presents cross-sectional regression results for the 25 Fama–French size and book-to-market portfolios. Panel A summarizes results without trimming. Although penalization seems unnecessary, we also present predictability results with  $\alpha_n = 0.05$  for comparison. Without penalty, all three choices of divergences work well: (i) the estimated prices of risk are highly significant with the correct sign, (ii) the adjusted  $R^2$ 's are larger than the one for the Fama–French model, and (iii) the intercept estimates are much smaller than the one by the Fama–French model. These results indicate that the proposed method outperforms the Fama–French three-factor model in our empirical example. We note that the KL divergence works better than the PSN1 and PSN2 divergences in terms the adjusted  $R^2$  in this case, and that the performances of PSN1 and PSN2 are very similar. For our method, the estimates with penalization underperform the ones without it. Since the dimension is low, we expect every portfolio is informative and there is no need for penalization.

We also report results after trimming extreme values of returns in Panel B of Table 2. For each training sample formed for year  $l$ , we remove returns that are either too big or

TABLE 1. Summary statistics of predicted SDF using 25 portfolios, no penalty, no trimming.

	KL	PSN1	PSN2
min	0.044	−1.544	0
max	4.965	3.100	3.458
Mean	1	1.08 <sup>13</sup>	0.972
25%	0.654	0.744	0.660
Median	0.924	1.088	0.975
75%	1.199	1.372	1.277
Standard deviation	0.544	0.555	0.483
Skewness	2.229	−0.200	0.430
Kurtosis	13.166	5.117	4.252

<sup>13</sup>For PSN1 and PSN2, the mean of the predicted SDF is not exactly 1 because they are out-of-sample prediction. On the other hand, for KL, the mean of the predicted SDF is always 1 by construction of (35).

TABLE 2. Cross-sectional regression for 25 size and book-to-market portfolios.

	Intercept	$\lambda_{\text{SDF}}$	$\lambda_{\text{RM}}$	$\lambda_{\text{SMB}}$	$\lambda_{\text{HML}}$	Adjusted $R^2$
3 Factors	1.668 (4.401)		-0.751 (-2.067)	0.204 (3.853)	0.437 (6.773)	0.714
Panel A: No trimming						
KL: No penalty	0.649 (13.977)	-0.257 (-11.438)				0.844
KL: $\alpha_n = 0.05$	0.720 (10.146)	-0.124 (-6.400)				0.625
PSN1: No penalty	0.735 (14.628)	-0.194 (-8.938)				0.767
PSN1: $\alpha_n = 0.05$	0.755 (11.555)	-0.116 (-6.478)				0.631
PSN2: No penalty	0.675 (11.126)	-0.167 (-8.270)				0.737
PSN2: $\alpha_n = 0.05$	1.616 (22.725)	-0.077 (-7.312)				0.686
Panel B: With trimming						
KL: trim 1%	0.660 (13.423)	-0.260 (-10.573)				0.822
KL: trim 2.5%	0.624 (10.080)	-0.284 (-8.890)				0.765
PSN1: trim 1%	0.748 (14.257)	-0.195 (-8.284)				0.738
PSN1: trim 2.5%	0.695 (11.339)	-0.203 (-7.853)				0.717
PSN2: trim 1%	0.716 (11.736)	-0.168 (-7.563)				0.701
PSN2: trim 2.5%	0.776 (11.475)	-0.141 (-5.927)				0.587

*Note:* The estimated SDF is derived in a rolling window out-of-sample fashion from July 1963 to December 2010. Panel A presents results without trimming, and Panel B presents results with trimming. The second column is the estimated constant in each model, the last column records the adjusted  $R^2$ , and the other columns summarize estimated price of risk. Numbers in the bracket are the corresponding  $t$ -values.

too small. For each period  $t$ , the vector of returns  $R_t$  is trimmed as

$$R_t \mathbb{I}\{\|R_t\|_\infty \leq Q_{1-a}\},$$

where  $Q_{1-a}$  is the  $(1 - a)$ th empirical quantile of  $\{\|R_t\|_\infty\}$  across all months used for training, that is, from July 1933 to June 2010. We consider  $a = 0.01$  and  $0.025$ . As we can see in Table 2, after trimming, predictability in terms of the adjusted  $R^2$  slightly decreases for all divergences. The KL divergence seems more sensitive to extreme values than the unrestricted PSN1 divergence. Forcing nonnegativity, PSN2 divergence also increases sensitivity of the results to extreme values.

For robustness checks, we also report results using the KL divergence for other low- and intermediate-dimensional portfolios in Tables 5 and 6 in the Appendix of the Online Supplementary Material. An interesting case is in Panel B of Table 5, where the estimate

without penalization is worse than the penalized estimate. This result indicates usefulness of penalization even for the low-dimensional case.<sup>14</sup>

**5.2.2 High-dimensional case: 300 portfolios** In this case, the estimate without penalization (essentially, the one by Ghosh, Julliard, and Taylor (2016)) is not applicable or performs erratically, and it is crucial to introduce some penalization. We focus on two divergences, KL and PSN1. For KL, the grid for the penalty level  $\alpha_n$  ranges from 0 to 2 with 0.05 increments. For PSN1, the grid for  $\alpha_n$  ranges from 0 to 1 with 0.025 increments. We estimate the SDFs by our method and implement the cross-sectional regression for each penalty level.

The results are summarized in Figure 1. Performances of the two divergences are similar. The SDF estimates without penalization perform very badly with the adjusted  $R^2$  close to 0 and relatively large intercept estimates. As the penalty level increases, predictability of our method gets better and outperforms Fama–French. The intercept estimates of our methods are also much smaller compared to Fama–French. However, the performance of our method gets worse when the penalty level continues to increase ( $\alpha_n > 1.5$  for KL and  $\alpha_n > 0.45$  for PSN1). This is expected because the number of selected portfolios will be too small for too large penalty levels and the performance would deteriorate. Based on these results, we set the optimal penalty level at 0.9 for KL and 0.475 for PSN1, and report more detailed results in Table 3. We can see that the adjusted  $R^2$  by the SDF estimates using penalization is much higher than the one of Fama–French, and that its intercept estimate is much closer to 0. Therefore, our method shows excellent performance upon choosing suitable penalty levels. We find that in this high-dimensional scenario, PSN1 works better than KL in terms of the adjusted  $R^2$ . This may be due to nonexistence of higher moments of certain returns in the presence of many portfolios. Moreover, we find that at the optimal penalty level (i.e., 0.9 for KL and 0.475 for PSN1), KL and PSN1 select 5 and 18 portfolios on average, respectively. Out of all rolling windows, 33% of times PSN1 with the optimal penalty level includes all portfolios selected by KL with optimal penalty level, 56% of times PSN1 only misses one or two portfolios selected by KL, and 11% of times PSN1 misses three or four portfolios selected by KL.

**5.2.3 Time series property of penalized SDF estimates** We illustrate time series properties of the SDF estimates with penalization for 300 portfolios. The penalty levels are chosen at 0.9 for KL and 0.475 for PSN1. The time series plot is displayed in Figure 2 and the grey shaded areas correspond to NBER recessions. In Table 4, we run a time series regression of our SDF estimates on other key factors in the market including Fama–French

<sup>14</sup>Underperformance of the estimate without penalization (for both low- and high-dimensional cases) may be due to nonexistence of higher moments. Note that both Ghosh, Julliard, and Taylor (2016) (with no penalization) and our method using the KL divergence (with  $\ell_1$ -penalization) rely upon finite exponential moments  $\mathbb{E}[\exp(\lambda'R_t)]$ , which require infinite order of moments of  $R_t$ . If some higher moments of  $R_t$  do not exist, the estimator without penalization will behave erratically. Although formal analysis is beyond the scope of this paper, we conjecture that our  $\ell_1$ -penalization may effectively remove such problematic components. Also, if nonexistence of higher moments is a significant concern, we can choose a different divergence function, such as the PSN1 or PSN2, which requires less stringent conditions for higher moments.

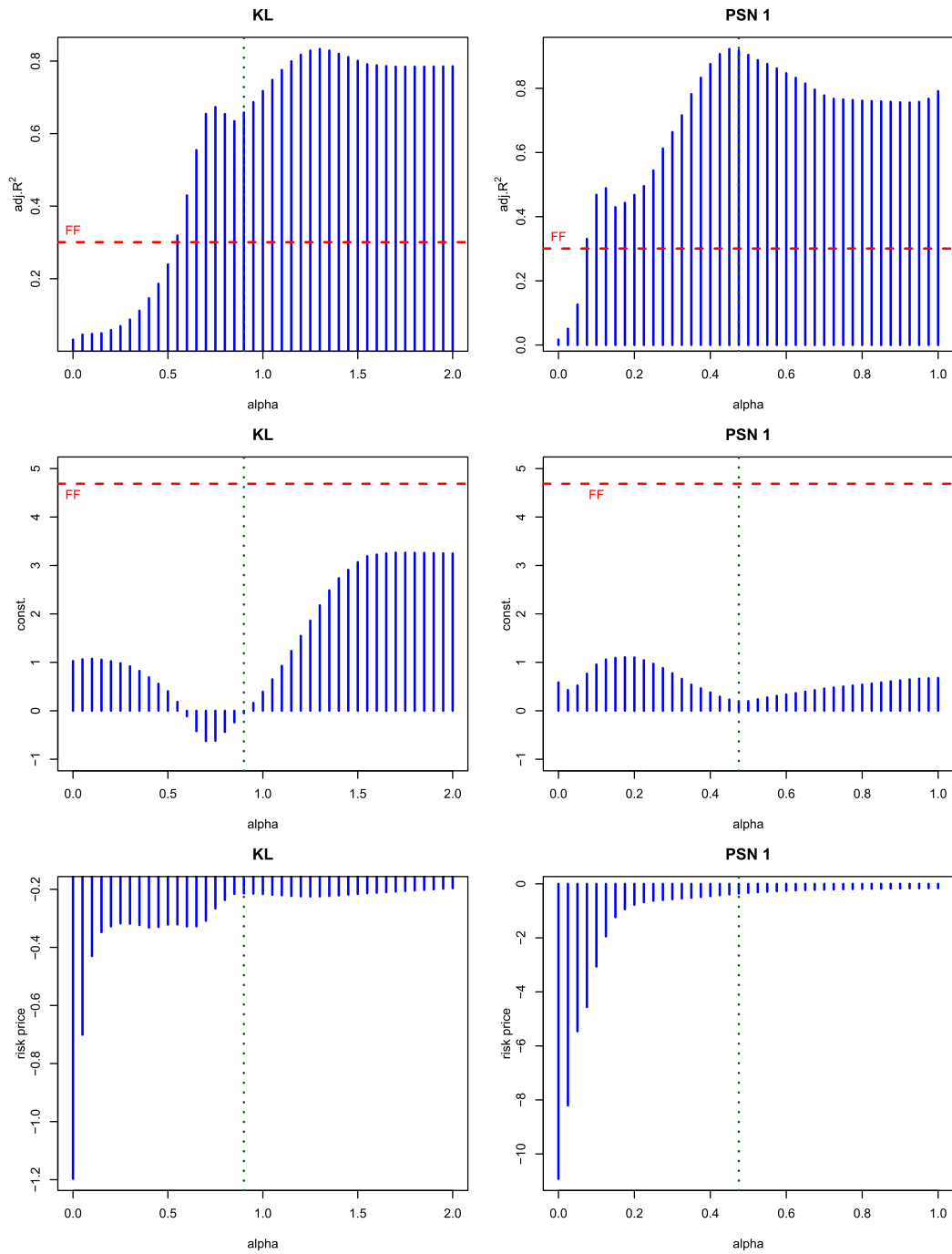


FIGURE 1. Summary of cross-sectional regression against different penalty levels in high-dimensional case ( $K = 300, |\tilde{\mathcal{T}}_1| = 360$ ).

TABLE 3. Cross-sectional regression in high-dimensional case: 300 portfolios.

	Intercept	$\lambda_{\text{SDF}}$	$\lambda_{\text{RM}}$	$\lambda_{\text{SMB}}$	$\lambda_{\text{HML}}$	Adjusted $R^2$
KL: $\alpha_n = 0.1$	1.027 (14.062)	-1.197 (-3.306)				0.032
KL: $\alpha_n = 0.9$	-0.050 (-0.851)	-0.214 (-24.017)				0.658
PSN1: $\alpha_n = 0.05$	0.521 (5.576)	-5.458 (-6.651)				0.126
PSN1: $\alpha_n = 0.475$	0.200 (8.281)	-0.361 (-58.153)				0.919
3 Factors	4.687 (10.986)		-3.891 (-9.998)	0.699 (5.295)	-0.517 (-2.900)	0.301

Note: Cross-sectional regression results with 300 portfolios. The 300 portfolios are composed of: 100 size and book-to-market portfolios, 100 size and operating profitability portfolios, and 100 size and investment portfolios. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1993 to December 2010. The second column is the estimated constant in each model, the last column records the adjusted  $R^2$ , and the other columns summarize estimated price of risk. Numbers in the bracket are the corresponding  $t$ -values.

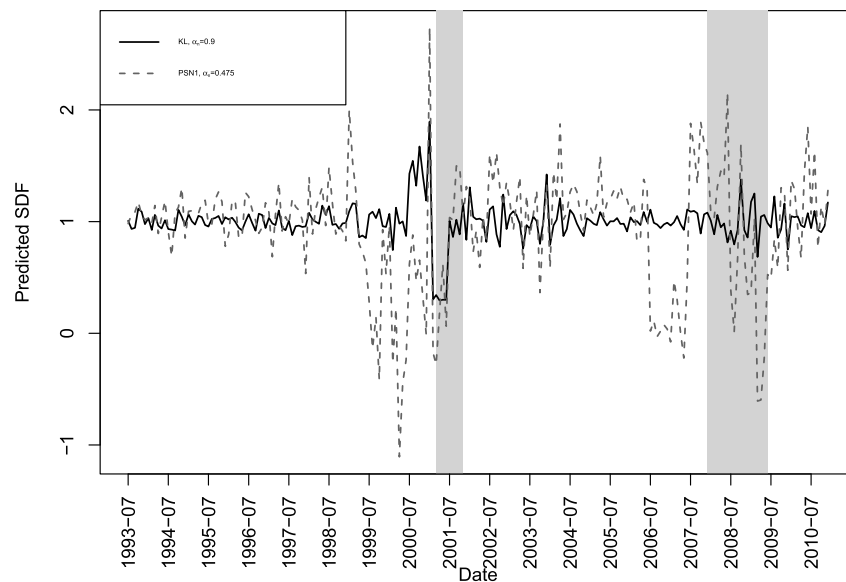


FIGURE 2. Time series plot of estimated SDF in high-dimensional case: July 1993–December 2010 (Grey shaded area represents NBER recessions).

three factors and momentum factors. We can see that correlations of our SDF estimates with those leading factors are very small, and the adjusted  $R^2$  is also small. This indicates that our method may capture critical information for asset pricing in the market that cannot be explained by Fama–French or momentum factors.

TABLE 4. Time series properties of estimated SDF from 300 portfolios.

Intercept	$\beta_{RM}$	$\beta_{SMB}$	$\beta_{HML}$	$\beta_{MOM}$	Adjusted $R^2$
Panel A: KL, $\alpha_n = 0.9$					
1.011 (85.846)	-0.004 (-1.427)	-0.014 (-4.106)	-0.007 (-1.851)	-0.007 (-3.264)	0.118
Panel B: PSN1, $\alpha_n = 0.475$					
0.962 (26.596)	-0.020 (-2.343)	-0.028 (-2.657)	-0.051 (-4.377)	-0.001 (-0.177)	0.096

*Note:* Time series regression of estimated SDF extracted from 300 portfolios against key factors in the market. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1993 to December 2010. Panel A presents results using KL divergence and when penalty level is 0.9, and Panel B presents results using PS1 divergence and when penalty level is 0.475. The first column is the estimated intercept in each regression, the last column records the adjusted  $R^2$ , and the other columns summarize estimated beta for each factor. Numbers in the bracket are the corresponding  $t$ -values.

## REFERENCES

- Athey, S., G. W. Imbens, and S. Wager (2016), "Approximate residual balancing: de-biased inference of average treatment effects in high dimensions." Working paper. [70]
- Backus, D., M. Chernov, and S. Zin (2014), "Sources of entropy in representative agent models." *Journal of Finance*, 69, 51–99. [68]
- Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012), "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica*, 80, 2369–2429. [77, 79]
- Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015), "Some new asymptotic theory for least squares series: Pointwise and uniform results." *Journal of Econometrics*, 186, 345–366. [73]
- Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017), "Program evaluation and causal inference with high-dimensional data." *Econometrica*, 85, 233–298. [69]
- Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009), "Simultaneous analysis of lasso and Dantzig selector." *Annals of Statistics*, 37, 1705–1732. [76]
- Borwein, J. M. and A. S. Lewis (1993), "Partially-finite programming in L1 and the existence of maximum entropy estimates." *SIAM Journal on Optimization*, 3 (2), 248–267. [66, 72]
- Bühlmann, P. and S. van de Geer (2011), *Statistics for High-Dimensional Data*. Springer. [69, 77]
- Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016), "Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting." *Journal of the Royal Statistical Society, B*, 78, 673–700. [65, 70, 84]
- Chen, X. (2007), "Large sample sieve estimation of semi-nonparametric models." In *Handbook of Econometrics*, Vol. 6B (J. Heckman and E. Leamer, eds.). Elsevier. [69]



- Chen, X. and T. Christensen (2015), “Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions.” *Journal of Econometrics*, 188, 447–465. [67, 71, 72, 73, 74]
- Chen, X., L. P. Hansen, and P. G. Hansen (2020), “Robust identification of investor beliefs.” *Proceedings of the National Academy of Sciences*, 117 (52), 33130–33140. [65, 70]
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018), “Double/debiased machine learning for treatment and structural parameters.” *Econometrics Journal*, 21, C1–C68. [69]
- Christensen, T. M. (2017), “Nonparametric stochastic discount factor decomposition.” *Econometrica*, 85, 1501–1536. [65, 70]
- Christensen, T. M. and B. Connault (2019), “Counterfactual sensitivity and robustness.” arXiv:1904.00989. arXiv preprint. [68]
- Cochrane, J. H. (2009), *Asset Pricing*, revised edition. Princeton University Press. [64, 69, 85]
- Csiszár, I. (1995), “Generalized projections for non-negative functions.” *Acta Mathematica Hungarica*, 68, 161–185. [66]
- Fama, E. F. and K. R. French (1993), “Common risk factors in the returns on stocks and bonds.” *Journal of Financial Economics*, 33, 3–56. [84]
- Fama, E. F. and J. D. MacBeth (1973), “Risk, return, and equilibrium: Empirical tests.” *Journal of Political Economy*, 81, 607–636. [85]
- Fan, J. and R. Li (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties.” *Journal of the American Statistical Association*, 96, 1348–1360. [67]
- Farrell, M. H. (2015), “Robust inference on average treatment effects with possibly more covariates than observations.” *Journal of Econometrics*, 189, 1–23. [69, 79]
- Ghosh, A., C. Julliard, and A. P. Taylor (2016), “An information-theoretic asset pricing model.” Working paper. [68, 70, 84, 88]
- Ghosh, A., C. Julliard, and A. P. Taylor (2017), “What is the consumption-CAPM missing? An information-theoretic framework for the analysis of asset pricing models.” *Review of Financial Studies*, 30, 442–504. [70]
- Hahn, J. (1998), “On the role of the propensity score in efficient semiparametric estimation of average treatment effects.” *Econometrica*, 66, 315–331. [84]
- Hansen, L. P. (2014), “Nobel lecture: Uncertainty outside and inside economic models.” *Journal of Political Economy*, 122, 945–987. [65, 68, 70]
- Hjort, N. L., I. W. McKeague, and I. Van Keilegom (2009), “Extending the scope of empirical likelihood.” *Annals of Statistics*, 37, 1079–1111. [69]
- Imbens, G. W. and D. B. Rubin (2015), *Causal Inference*. Cambridge University Press. [69]

- Kitamura, Y. and M. Stutzer (2002), “Connections between entropic and linear projections in asset pricing estimation.” *Journal of Econometrics*, 107, 159–174. [70]
- Komunjer, I. and G. Ragusa (2016), “Existence and characterization of conditional density projections.” *Econometric Theory*, 32, 947–987. [66, 71, 73]
- Kraus, A. and R. H. Litzenberger (1976), “Skewness preference and the valuation of risk assets.” *The Journal of Finance*, 31 (4), 1085–1100. [69]
- Lahiri, S. and S. Mukhopadhyay (2012), “A penalized empirical likelihood method in high dimensions.” *Annals of Statistics*, 40, 2511–2540. [69]
- Lewellen, J., S. Nagel, and J. Shanken (2010), “A skeptical appraisal of asset pricing tests.” *Journal of Financial Economics*, 96, 175–194. [84]
- Little, R. J. A. and D. B. Rubin (2002), *Statistical Analysis With Missing Data*. Wiley. [65]
- Lorentz, G. G. (1986), *Approximations of Functions*. Chelsea. [72]
- Newey, W. K. (1994), “The asymptotic variance of semiparametric estimators.” *Econometrica*, 62, 1349–1382. [74, 80]
- Newey, W. K. and R. J. Smith (2004), “Higher order properties of GMM and generalized empirical likelihood estimators.” *Econometrica*, 72, 219–255. [68, 69]
- Newey, W. K. and K. D. West (1987), “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.” *Econometrica*, 55, 703–708. [75]
- Qiu, C. and T. Otsu (2022), “Supplement to ‘Information theoretic approach to high-dimensional multiplicative models: Stochastic discount factor and treatment effect.’” *Quantitative Economics Supplemental Material*, 13, <https://doi.org/10.3982/QE1603>. [70]
- Rosenbaum, P. R. and D. B. Rubin (1983), “The central role of the propensity score in observational studies for causal effects.” *Biometrika*, 70, 41–55. [83]
- Schumaker, L. L. (1981), *Spline Functions: Basic Theory*. Wiley. [72]
- Stutzer, M. (1995), “A Bayesian approach to diagnosis of asset pricing models.” *Journal of Econometrics*, 68, 367–397. [68]
- Tang, C. Y. and C. Leng (2010), “Penalized high-dimensional empirical likelihood.” *Biometrika*, 97, 905–920. [69]
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society, B*, 58, 267–288. [67]
- Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*. Springer. [65]
- van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014), “On asymptotically optimal confidence regions and tests for high-dimensional models.” *Annals of Statistics*, 42, 1166–1202. [69, 78, 79]
- van de Geer, S. A. (2008), “High-dimensional generalized linear models and the lasso.” *Annals of Statistics*, 36, 614–645. [69]

Vasicek, O. (1977), “An equilibrium characterisation of the term structure.” *Journal of Financial Economics*, 5, 177–188. [68]

Zhang, C.-H. (2010), “Nearly unbiased variable selection under minimax concave penalty.” *Annals of Statistics*, 38, 894–942. [67]

Zhang, C.-H. and S. S. Zhang (2014), “Confidence intervals for low dimensional parameters in high dimensional linear models.” *Journal of the Royal Statistical Society, B*, 76, 217–242. [69, 78, 79]

Zubizarreta, J. R. (2015), “Stable weights that balance covariates for estimation with incomplete outcome data.” *Journal of the American Statistical Association*, 110, 910–922. [65, 70]

---

Co-editor Andres Santos handled this manuscript.

Manuscript received 14 April, 2020; final version accepted 17 May, 2021; available online 5 August, 2021.