# INFORMATION THEORETIC APPROACH TO HIGH DIMENSIONAL MULTIPLICATIVE MODELS: STOCHASTIC DISCOUNT FACTOR AND TREATMENT EFFECT

CHEN QIU AND TAISUKE OTSU

ABSTRACT. This paper is concerned with estimation of functionals of a latent weight function that satisfies possibly high dimensional multiplicative moment conditions. Main examples are missing data problems, treatment effects, and functionals of the stochastic discount factor in asset pricing. 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 derive asymptotic properties of the proposed estimator, and illustrate the proposed method by a theoretical example on treatment effect analysis and empirical example on the stochastic discount factor.

## 1. INTRODUCTION

1.1. **Motivation.** In empirical analysis, economic information and other statistical information are 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 literature (see Hall, 2004, for a review). In this paper, we consider the following moment condition models taking a multiplicative form:

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

where $X$ is a vector of observables, $\omega : \mathcal{X} \to (0, \infty)$ 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. We are interested in the situation where the observables $X$ and/or functions $g$ are high dimensional (possibly larger than the sample size). Under this condition, our object of interest is set as a functional of the unknown weight function $\omega$:

$$\theta = \mathbb{E}[\omega(X)h(X, Y)], \tag{2}$$

where $Y$ is another vector of observables and $h$ is a vector of known functions of $(X, Y)$. In this paper, we develop a general estimation and inference method for the parameter $\theta$.

Interestingly, this setup can be motivated by somewhat distant economic problems: inference on stochastic discount factors (SDF) 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.

Financial support from the ERC Consolidator Grant (SNP 615882) is gratefully acknowledged (Otsu).

**Example 1** (Stochastic discount factor)**.** Consider a discrete time economy, where all uncertainty is driven by an unobservable state vector $S_t$. Under the assumption of no arbitrage, there exists a strictly positive SDF $m(S_t, S_{t+1})$ such that

$$\mathbb{E}[m(S_t, S_{t+1})R_j(S_t, S_{t+1})] = 1, \tag{3}$$

where $R_j(S_t, S_{t+1})$ is the short term return of asset $j$ between time $t$ and $t+1$. This equation says that any asset $j$ in the market would share the same expected return when discounted by the SDF $m$ (see Cochrane, 2009, for a review). Let $R_{f,t}$ be the risk free return between $t$ and $t+1$ and $X_t = \{R_j(S_t, S_{t+1}) - R_{f,t}\}_{j=1}^K$ be a $K$-vector of observable excess returns. Suppose the SDF $m(S_t, S_{t+1})$ is specified by an unknown function $\omega(X_t)$. Then by subtracting the condition $\mathbb{E}[m(S_t, S_{t+1})R_{f,t}] = 1$ for the risk free asset from (3), we obtain the moment condition

$$\mathbb{E}[\omega(X_t)X_t] = 0,$$

which can be considered as a special case of (1) with $g(X) = X$ and $r = 0$.

Inference on the SDF $\omega$ is one of the central topics in financial economics. For example, Christensen (2016) investigated extraction of permanent and transitory components of the SDF process, which can be formulated as the eigenfunction problem for the matrix

$$\mathbb{E}[s(X_t)s(X_{t+1})']^{-1}\mathbb{E}[\omega(X_t)s(X_t)s(X_{t+1})'],$$

and $s$ is a vector of known basis functions for sieve estimation. Although $\mathbb{E}[s(X_t)s(X_{t+1})']$ can be estimated by the empirical moments, estimation of $\theta = \mathbb{E}[\omega(X_t)s(X_t)s(X_{t+1})']$ poses a substantial challenge. Christensen (2016) considered two cases: (i) $\omega(X_t)$ is directly observable, and (ii) $\omega(X_t)$ is specified by a parametric model, where a preliminary estimator can be plugged-in. Our information theoretic approach will provide an alternative estimation strategy for $\omega$ and $\theta$. □

**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 $i = 1, \ldots, N$, 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 $Y_i^*$ 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 $\theta = \mathbb{E}[\omega(X)YD]$, where $\omega(X) = 1/\mathbb{P}\{D = 1|X\}$. In this setup, many estimation and inference methods for $\theta$ and their generalizations have been proposed (e.g. Tsiatis, 2006), including the inverse probability weighted estimator $n^{-1}\sum_{i=1}^n \tilde{\omega}(X_i)Y_iD_i$, 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 (1) may

be given by

$$\mathbb{E}[\omega(X)g(X)D] = \mathbb{E}[g(X)], \tag{4}$$

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, so-called the balancing covariates approach explores the moment conditions in (4) to find adjusting weights used for estimation of $\theta$ (Zubizarreta, 2015, and Chan, Yam and Zhang, 2016). This paper proposes an alternative estimation method that may be considered as an extension of these papers toward high dimensional environments. □

1.2. **Methodology.** In this paper, we propose an information theoretic approach to estimate the parameters $\theta$ in (2), where the latent weight function $\omega$ is characterized by the moment conditions in (1). Our method allows dimension of the observables and/or moment functions to be high dimensional (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.

A key issue for estimation of $\theta$ is how to evaluate the latent weight function $\omega$ satisfying (1). In this paper, we address this issue by an information theoretic approach. 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 $\mathbb{Q}$ is a tilted model-based measure. Then the moment condition (1) is written as $\mathbb{E}_{\mathbb{Q}}[g(X)] = r$. To approximate the tilted measure $\mathbb{Q}$, we apply the information projection (e.g., Csiszár, 1975, and Kitamura and Stutzer, 1997). In particular, we consider the minimization problem using the Kullback-Leibler divergence

$$\min_{\mathbb{Q}} \int \log \frac{d\mathbb{Q}}{d\mathbb{P}} d\mathbb{Q}, \quad \text{s.t. } \mathbb{E}_{\mathbb{Q}}[g(X)] = r.$$

Under mild regularity conditions, the solution $\mathbb{Q}_*$ is obtained by the Radon-Nikodym derivative

$$\frac{d\mathbb{Q}_*}{d\mathbb{P}} = \exp(\lambda_*' g(x)), \tag{5}$$

where $\lambda_*$ solves $\min_{\lambda} \mathbb{E}_{\mathbb{P}}[\exp(\lambda' g(X)) - \lambda' r]$. Based on this derivative, the pseudo parameter $\theta_*$ can be written as

$$\theta_* = \mathbb{E}_{\mathbb{Q}_*}[h(X, Y)] = \mathbb{E}_{\mathbb{P}}[\exp(\lambda_*' g(X)) h(X, Y)].$$

If the vector of functions $g$ is rich enough to approximate $\omega(x)$ by $\exp(\lambda' g(x))$ with certain $\lambda$, then the parameter of interest $\theta$ can be estimated by using the sample counterpart of $\theta_*$.

Let $\mathbb{E}_n[\cdot]$ be the sample mean and $\|\cdot\|_1$ be the $\ell_1$-norm for a vector. By estimating the population moments $\mathbb{E}_{\mathbb{P}}[\cdot]$ by the sample moments $\mathbb{E}_n[\cdot]$, our information theoretic estimator of $\theta$ is defined as

$$\hat{\theta} = \mathbb{E}_n[\exp(\hat{\lambda}' g(X))h(X,Y)],$$

where

$$\hat{\lambda} = \begin{cases} \arg\min_{\lambda} \mathbb{E}_n[\exp(\lambda' g(X)) - \lambda' r] & \text{(low dimensional case)} \\ \arg\min_{\lambda} \mathbb{E}_n[\exp(\lambda' g(X)) - \lambda' r] + \alpha_n \|\lambda\|_1 & \text{(high dimensional case)} \end{cases}, \tag{6}$$

and $\alpha_n$ is a penalty level chosen by the researcher. If $r$ contains population moments of the observables, we estimate them by the corresponding empirical moments. The $\ell_1$ penalty term for the high dimensional case is introduced to regularize behaviors of the estimator $\hat{\lambda}$. 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 also applied.

We emphasize that although the construction of the estimator $\hat{\lambda}$ in (6) is analogous to the exponential tinting estimator for overidentified moment condition models (Kitamura and Stutzer, 1997, and Imbens, Spady and Johnson, 1998), our setup and properties of the estimator are significantly different from theirs due to three reasons. First, our moment conditions (1) contain the latent weight function $\omega$, and the information projection is applied to estimate $\omega$. Second, the interpretation and property of $\hat{\lambda}$ are different from theirs. In the conventional exponential tilting 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. On the other hand, in our approach, $\hat{\lambda}$ is an estimator for the pseudo parameter $\lambda_*$ and typically does not converge to zero. 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 (possibly larger than the sample size). In such case, the estimator $\hat{\lambda}$ has to be regularized as in (6).[1]

Hereafter the paper is organized as follows. After a brief review of related literature (Section 1.3), we present theoretical properties of our estimator $\hat{\theta}$ for the low dimensional case (Section 2) and high dimensional case (Section 3). The proposed method is illustrated by a theoretical example on treatment effects (Section 4) and empirical example on the SDF (Section 5). All proofs, tables, and figures are contained in Appendix.

1.3. **Related literature.** The construction of our estimator is related to the literature of exponential tilting, empirical likelihood, and its variants (see, Owen, 2001, and Kitamura,

[1]This paper focuses on the information projection by using the Kullback-Leibler divergence to simplify the presentation. It is relatively straightforward to extend our estimation approach to other divergence measures or generalized empirical likelihood criteria (Newey and Smith, 2004).

2006, for surveys). In spite of similarity of the construction of the estimator, however, our setup and property of the Lagrange multiplier $\hat{\lambda}$ are quite different from this literature as discussed in Section 1.2. Indeed our treatment on the Lagrange multiplier 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. 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 $\lambda_*$ in our setup, not zero).

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 overview of these topics.

In the context of missing data and treatment effect analyses, the proposed method, illustrated in Section 4, is closely related to the recent literature on balancing weights (Zubizarreta, 2015, Chan, Yam and Zhang, 2016, and Athey and Imbens, 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.

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, the proposed method can be regarded as an extension to high dimensional environments (especially for a large number of assets). Also, as mentioned in Example 1, this paper can provide an alternative method to extract permanent and transitory components of the SDF process (Christensen, 2016).

**Notation.** Hereafter, we work with triangular array data $\{X_{i,n}, Y_{i,n}\}_{i=1}^n$, which is considered as the first $n$ elements from the infinite sequence $\{X_{i,n}, Y_{i,n}\}_{i=1}^\infty$ generated from a probability measure $\mathbb{P}_n$. Our asymptotic analysis is based on the array asymptotics. To simplify the notation, we suppress the subscripts and denote by $\{X_i, Y_i\}_{i=1}^n$ and $\mathbb{P}$. 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 $A$, $|B| = \sqrt{\text{trace}(B'B)}$ be the Euclidean norm for a

scalar, vector, or matrix $B$, and $a \vee b = \max\{a, b\}$. Finally, let $\lambda_{\max}(C)$ and $\lambda_{\min}(C)$ be the maximum and minimum eigenvalues of a matrix $C$, respectively.

## 2. Low dimensional case

In this section, we present asymptotic properties of our information theoretic estimator $\hat{\theta}$ for the low dimensional case, where the dimension $K$ of the moment conditions $g$ in (1) grows slowly compared to the sample size $n$. In this case, computation of $\hat{\lambda}$ in (6) does not involve the $\ell_1$-penalization. Throughout this section, we impose the following conditions.

**Condition D.** $\{X_i, Y_i\}$ is an independently and identically distributed triangular array. The support $\mathcal{X} \subset \mathbb{R}^p$ of $X$ is a Cartesian product of $p$ compact intervals. $\omega : \mathcal{X} \to (0, \infty)$ is a continuous function bounded from above and away from zero with $\mathbb{E}[\omega(X)^2] < \infty$, and $h$ is a scalar-valued continuous function with $\mathbb{E}[h(X, Y)^2] < \infty$.

**Condition S.** $g$ is a $K$-dimensional vector of continuous functions such that the first element is 1, $\lambda_{\min}(\mathbb{E}[g(X)g(X)']) > c$ for some $c > 0$, and

$$\sup_{x \in \mathcal{X}} |\log \omega(x) - \lambda' g(x)| = O(K^{-\eta}), \tag{7}$$

for some $\lambda \in \mathbb{R}^K$ and $\eta > 0$.

Condition D contains standard assumptions on the data and functions $\omega$ and $h$. Although we focus on independent data, we expect that analogous results can be established under weakly dependent data by employing suitable limit theorems and probabilistic inequalities. Also it is interesting to extend our approach to introduce some blocking scheme for efficiency gain as in Kitamura and Stutzer (1997). To simplify the presentation, we focus on the case where $h$ (and thus $\theta$) is scalar. An extension to the case of vector $\theta$ is straightforward. It is also possible to extend our method to the case where $\theta$ is implicitly defined as a solution of moment conditions $\mathbb{E}[\omega(X)h(X, Y; \theta)] = 0$. Condition S is on the basis functions $g$. The requirement that $g$ contains the constant is for convenience to obtain a simple representation of the density $d\mathbb{Q}^*/d\mathbb{P}$ in (5) (otherwise, it needs to be normalized by $\mathbb{E}_{\mathbb{P}}[\exp(\lambda_*' g(X))]$). The order of the approximation error in (7) depends on the choice of the basis functions $g$ and can be verified by using the results from functional analysis literature (e.g., Lorentz, 1986, and Schumaker, 1981). For example, if we choose polynomials or splines, then the approximation error is of order $O(K^{-s/p})$, where $s$ is the number of continuous derivatives of $\omega$ and $p$ is dimension of $X$. With this respect, $\eta$ can be regarded as a smoothness parameter of the function $\log \omega(x)$.

Let $\hat{\omega}(x) = \exp(\hat{\lambda}' g(x))$ and $\zeta_K = \sup_{x \in \mathcal{X}} |\mathbb{E}[g(X)g(X)']^{-1/2} g(x)|$ be the largest normalized length of the basis functions. Based on the above conditions, the convergence rate of $\hat{\omega}(x)$ and consistency of $\hat{\theta}$ are obtained as follows.

**Theorem 1.** Suppose that Conditions D and S hold true. Furthermore, assume $K \to \infty$, $\zeta_K \sqrt{K/n} \to 0$, and $\zeta_K K^{-\eta/2} \to 0$ as $n \to \infty$. Then

$$\sup_{x \in \mathcal{X}} |\hat{\omega}(x) - \omega(x)| = O_p(\zeta_K \{\sqrt{K/n} + K^{-\eta/2}\}),$$

and $\hat{\theta} \xrightarrow{p} \theta$.

Consistency of $\hat{\theta}$ is achieved by first showing that the estimator $\hat{\omega}$ for $\omega$ is also consistent under the sup-norm. Interestingly, although our setup is different from standard nonparametric series estimation, we achieve a similar convergence rate with conventional series estimators for regression models (e.g., Newey, 1994). Indeed, our proof is in line with series estimation methods, where the estimation error of $\omega$ can be decomposed into two parts: sampling error and approximation bias. The sampling error can be controlled by Lemma 1 while the approximation error is dealt with Condition S.

The asymptotic distribution of the estimator $\hat{\theta}$ is obtained as follows.

**Theorem 2.** Suppose that Conditions D and S hold true. Furthermore, suppose $K \to \infty$, $\zeta_K^4 K/\sqrt{n} \to 0$, $\sqrt{n}\zeta_K^2 K^{-\eta/2} \to 0$ as $n \to \infty$, and

$$\sup_{x \in \mathcal{X}} |\mathbb{E}[h(X,Y)|X = x] - \lambda' g(x)| = o(1), \tag{8}$$

for some $\lambda \in \mathbb{R}^K$. Then

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

where $\Omega = \mathbb{E}[\omega(X)^2 \{h(X,Y) - \mathbb{E}[h(X,Y)|X]\}\{h(X,Y) - \mathbb{E}[h(X,Y)|X]\}']$.

This theorem says that our estimator $\hat{\theta}$ for the low dimensional case is $\sqrt{n}$-consistent and asymptotically normal. The asymptotic variance $\Omega$ can be estimated by

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}(X_i)^2 \{h(X_i, Y_i) - \hat{h}^X(X_i)\}\{h(X_i, Y_i) - \hat{h}^X(X_i)\}', \tag{9}$$

where $\hat{h}^X(x)$ is some nonparametric estimator of the conditional mean $\mathbb{E}[h(X,Y)|X = x]$.

Compared to Theorem 1, we strengthen the assumptions in two aspects. First, we require that the conditional expectation function $\mathbb{E}[h(X,Y)|X = x]$ is also approximated by the basis functions $g$. This requirement is mild and verified by the results in functional analysis. By this condition we can rely on a projection argument to achieve mean-square differentiability, which is a key condition to establish the $\sqrt{n}$-consistency (cf. Newey and McFadden, 1994, Theorem 8.1). Second, we impose more stringent conditions on $K$. For example, if we choose the B-spline basis with $K \propto n^a$ for $a > 0$, then it holds $\zeta_K = O(\sqrt{K})$ (Newey, 1997) and the above conditions on $K$ are satisfied with $a \in (1/(\eta - 2), 1/6)$. Therefore, for Theorem 2, $K$ should grow at a sufficiently slow rate.

## 3. High dimensional case

In this section, we now consider the high dimensional case, where the dimension $K$ of the moment functions $g$ can be larger and grow faster than the sample size $n$. In this case, $\hat{\lambda}$ in (6) is computed by the $\ell_1$-penalization. High dimensionality of $g$ can 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 $\omega$ admits certain sparse representation, our penalized estimator can consistently estimate $\omega$ and the parameter of interest $\theta$.

To state the assumptions for the high dimensional case, we introduce further notation. For an index set $S \subset \{1, \ldots, K\}$, let $|S|$ be its cardinality (with slight abuse of notation), $\lambda_S = (\lambda_{1,S}, \ldots, \lambda_{K,S})'$ be a $K$ dimensional vector with $\lambda_{j,S} = \lambda_j \mathbb{I}\{j \in S\}$ for the $j$-th component $\lambda_j$ of $\lambda$, and $\lambda_{S^c} = (\lambda_{1,S^c}, \ldots, \lambda_{K,S^c})'$ with $\lambda_{j,S^c} = \lambda_j \mathbb{I}\{j \notin S\}$. So, $\lambda_S$ and $\lambda_{S^c}$ have non-zero elements only in the index set $S$ and its complement $S^c$, respectively. We first introduce the so-called compatibility condition.

**Condition C.** Let $\mathscr{S}$ be a class of index sets. For each $S \in \mathscr{S}$, there exists some constant $\phi_S > 0$ such that for all $\lambda$ satisfying $\|\lambda_{S^c}\|_1 \leq 3\|\lambda_S\|_1$, it holds $\|\lambda_S\|_1 \leq \frac{|\lambda|\sqrt{|S|}}{\phi_S}$.

This is a high level condition that strengthens the Cauchy-Schwarz inequality between the $\ell_1$- and $\ell_2$-norms. Such compatibility condition is commonly employed in the high dimensional statistics literature, such as the restricted eigenvalue condition in Bickel, Ritov and Tsybakov (2009).

To proceed, we introduce further notation. Recall $\lambda_* = \arg\min_\lambda \mathbb{E}[\exp(\lambda' g(X)) - \lambda' r]$, and let

$$\mathscr{E}(\lambda) = \mathbb{E}[\exp(\lambda' g(X)) - \lambda' r] - \mathbb{E}[\exp(\lambda_*' g(X)) - \lambda_*' r],$$

be the excess risk. Given the class $\mathscr{S}$ with the compatibility constants $\{\phi_S : S \in \mathscr{S}\}$ in Condition C, the oracle $\lambda_o$ is defined as

$$\lambda_o = \arg \min_{\lambda : S_\lambda \in \mathscr{S}} 2\mathscr{E}(\lambda) + \frac{16\alpha_n^2 |S_\lambda|}{\phi_{S_\lambda}^2 \varrho}, \tag{10}$$

where $S_\lambda = \{j : \lambda_j \neq 0\}$, $\alpha_n$ is a penalty level in (6), and $\varrho$ is a constant defined in Condition H below. The minimized value of (10) is denoted as $Q_o$. The oracle $\lambda_o$ plays the role of an approximated counterpart of $\lambda_*$ under sparsity. Note that $\mathscr{E}(\lambda_o) \geq \mathscr{E}(\lambda_*) = 0$ and our sparsity assumption is characterized by the convergence rate of $\mathscr{E}(\lambda_o)$ toward zero. Let

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

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 $\sigma_{\varepsilon,n} \leq \frac{\alpha_n}{8}$, $\varrho$, and $A$ such that for $M = \frac{Q_o}{2\sigma_{\varepsilon,n}}$,

**(i):** $\Pr\{\sup_{\left\|\lambda-\lambda_o\right\|_1 \leq M} |\nu_n(\lambda) - \nu_n(\lambda_o)| \leq \sigma_{\varepsilon,n} M\} \geq 1 - \varepsilon$,

**(ii):** $\{\lambda : \left\|\lambda - \lambda_o\right\|_1 \leq M\} \subseteq \{\lambda : \left\|\lambda - \lambda_*\right\|_\infty \leq A,\ \varrho|\lambda - \lambda_*|^2 \leq \mathscr{E}(\lambda)\}$.

Condition H (i), which is analogous to Lemma 1 (i) for the low dimensional case, controls the empirical process $\nu_n(\lambda)$ around the oracle $\lambda_o$. In contrast to Lemma 1 (i), the neighborhood $\left\|\lambda - \lambda_o\right\|_1 \leq M$ is defined based on the $\ell_1$-norm. This is due to the $\ell_1$-penalization in the objective function of $\hat{\lambda}$. Condition H (i) can be verified by using empirical process theory. [2] Condition H (ii) is similar to Lemma 1 (ii) for the low dimensional case except that the neighborhood by the $\ell_1$-norm is around the oracle $\lambda_o$. By this condition, the excess risk $\mathscr{E}(\lambda)$ can be bounded from below by a quadratic function of $\lambda$.

Under these conditions, the convergence rate of $\hat{\omega}$ and consistency of the parameter estimator $\hat{\theta}$ are established as follows. Let $\sup_{x\in\mathcal{X}} \left\|g(x)\right\|_\infty = O(\tilde{\zeta}_K)$, $s = |S_{\lambda_o}|$, and $\kappa_{o,n} = \mathscr{E}(\lambda_o)\sqrt{\frac{n}{\log K}} \vee s\sqrt{\frac{\log K}{n}}$.

**Theorem 3.** Suppose Conditions D, S, C, and H hold true. Furthermore, assume that $\sup_{x\in\mathcal{X}} |\omega_o(x) - \omega(x)| = O\left(s\sqrt{\frac{\log K}{n}}\right)$ and that $|\lambda_o' g(x)|$ is bounded uniformly over all $x \in \mathcal{X}$ and all $n$ large enough. If $K \to \infty$ and $\tilde{\zeta}_K \kappa_{o,n} \to 0$ as $n \to \infty$, then

$$\sup_{x\in\mathcal{X}} |\hat{\omega}(x) - \omega(x)| = O_p(\tilde{\zeta}_K \kappa_{o,n}),$$

and $\hat{\theta} \xrightarrow{p} \theta$.

This theorem is a counterpart of Theorem 1 for the high dimensional case and establishes the uniform convergence rate of $\hat{\omega}$ and consistency of $\hat{\theta}$. The object $\tilde{\zeta}_K$ depends on the choice of basis functions $g$. For example, if $g$ is a vector of polynomials over $\mathcal{X} = [0,1]^p$ or Haar wavelet basis functions, it is of order $\tilde{\zeta}_K = O(1)$. In this case, if $\mathscr{E}(\lambda_o) = O(s)$, the uniform convergence rate of $\hat{\omega}$ is of order $O_p\left(s\sqrt{\frac{\log K}{n}}\right)$, and the dimension $K$ may grow faster than $n$ even at an exponential rate.

For the high dimensional case, the approximation bias for $\omega$ tends to be larger and is controlled by the approximate sparsity assumption that requires sufficiently fast decays of the excess risk $\mathscr{E}(\lambda_o)$ and approximation error $\sup_{x\in\mathcal{X}} |\omega_o(x) - \omega(x)|$. For example, if $\mathbb{E}[g(X)g(X)']$ is invertible, we can verify that the approximation error is of order $O(\zeta_K\sqrt{\mathscr{E}(\lambda_o)} + K^{-\eta/2})$. Also the sampling error of the $\ell_1$-penalized estimator is controlled by Condition C.

[2] Since our objective function is Lipschitz in a neighborhood of $\lambda_o$, probabilistic inequalities, such as Bühlmann and Van de Geer (2011, Lemma 14.20), can be applied.

The asymptotic distribution of $\hat{\theta}$ for the high dimensional case is obtained as follows. Let $\zeta_s = \sup_{x\in\mathcal{X}} |\mathbb{E}[g_s(X)g_s(X)']^{-1/2}g_s(x)|$ and $g_s(X)$ be an $s$ dimensional subvector of $g(X)$ selected by $S_{\lambda_o}$.

**Theorem 4.** In addition to the assumptions for Theorem 3, suppose $c \leq \lambda_{\min}\{\mathbb{E}[g_s(X)g_s(X)']\} \leq \lambda_{\max}\{\mathbb{E}[g_s(X)g_s(X)']\} \leq C$ for some positive constants $c$ and $C$,

$$\sup_{x\in\mathcal{X}} |\mathbb{E}[h(X,Y)|X=x] - \lambda' g_s(x)| = o(1),$$

for some $\lambda \in \mathbb{R}^s$, and

$$\sqrt{n}\zeta_s^4\kappa_{o,n}^2 \vee \sqrt{n}\tilde{\zeta}_K\zeta_s\kappa_{o,n}^2 \vee \sqrt{n}\zeta_s^2 \sup_{x\in\mathcal{X}} |\omega_o(x) - \omega(x)| \to 0, \tag{11}$$

as $n \to 0$. Then

$$\sqrt{n}(\hat{\theta} - \theta + \Upsilon) \xrightarrow{d} N(0, \Omega),$$

where $\Upsilon$ and $\Omega$ are defined in (29) in Appendix and Theorem 3, respectively.

This theorem says that the estimator $\hat{\theta}$ is $\sqrt{n}$-consistent for $\theta + \Upsilon$ and asymptotically normal. The asymptotic variance $\Omega$ is same as the low dimensional case, and can be estimated as in (9). The main difference with the lower dimensional case is the presence of the bias term $\Upsilon$. A reason for this disappointing but expected result is that the $\ell_1$-norm is not differentiable. Therefore, we cannot employ usual stochastic expansion techniques to bound the mean-square differentiability term, and the bias term arises from the subgradient $\hat{\kappa} = (\text{sign}(\hat{\lambda}_1), \ldots, \text{sign}(\hat{\lambda}_K))'$ of the optimization in (6) with the $\ell_1$-penalty.

The additional conditions in (11) restrict the basis functions $g$, growth rates of $K$ and $s$, and approximation error $\sup_{x\in\mathcal{X}} |\omega_o(x) - \omega(x)|$. For example, if $\tilde{\zeta}_K = O(1)$ and $\mathscr{E}(\lambda_o) = O(s)$, then the first condition in (11) requires $\frac{\zeta_s^4 s^2 \log K}{\sqrt{n}} \to 0$ and the dimension $K$ may grow faster than the sample size $n$ even at an exponential rate (as far as $\zeta_s^4 s^2 = O(n^c)$ for some $c < 1/2$). We also note that Theorem 4 can be applied to the low dimensional case. Under analogous assumptions to Theorem 2, we can show that the $\ell_1$-penalized estimator $\hat{\theta}$ achieves asymptotic normality with a simpler bias term.

To achieve $\sqrt{n}$-consistency for the parameter of interest $\theta$, we can correct the bias term by estimating $\Upsilon$, that is

$$\hat{\Upsilon} = \alpha_n \mathbb{E}_n[\exp(\hat{\lambda}' g(X))h(Y,X)g_{\hat{s}}(X)']\mathbb{E}_n[\exp(\hat{\lambda}' g(X))g_{\hat{s}}(X)g_{\hat{s}}(X)']^{-1}\hat{\kappa}_{\hat{s}},$$

where $g_{\hat{s}}(X)$ and $\hat{\kappa}_{\hat{s}}$ are $\hat{s}$ dimensional subvectors of $g(X)$ and $\hat{\kappa}$ selected by $S_{\hat{\lambda}}$, respectively. The bias corrected estimator is given by $\tilde{\theta} = \hat{\theta} + \hat{\Upsilon}$, and additional regularity conditions guarantee consistency of $\tilde{\theta}$. This is essentially the same idea as in Zhang and Zhang (2014) and van de Geer *et al.* (2014) (called the desparsifying estimator). van

de Geer *et al.* (2014) showed that under certain sparsity assumptions such desparsifying methods will achieve $\sqrt{n}$-consistency for $\theta$.

For the penalty level $\alpha_n$, we shall choose $\alpha_n \propto \sqrt{\frac{\log K}{n}}$ to satisfy Condition H. In practice, $\alpha_n$ may be chosen by a data dependent method, such as cross validation.

## 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 $i = 1, \ldots, n$ ($D_i = 1$ and 0 mean treated and not treated, respectively). 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 unconfoundedness and overlap assumptions, $\tau$ can be identified as (Rosenbaum and Rubin, 1983)

$$\tau = \mathbb{E}[\omega_1(X)DY] - \mathbb{E}[\omega_0(X)(1 - D)Y] \equiv \theta_1 - \theta_0,$$

where $\omega_1(x) = \pi(x)^{-1}$, $\omega_0(x) = \{1 - \pi(x)\}^{-1}$, and $\pi(x) = \Pr\{D = 1|X = x\}$ is the propensity score. We treat $\omega_1$ and $\omega_0$ as latent weight functions, and construct moment conditions as in (1) by utilizing the property of the propensity score:

$$\mathbb{E}[D\omega_1(X)g(X)] = \mathbb{E}[(1 - D)\omega_0(X)g(X)] = \mathbb{E}[g(X)], \tag{12}$$

for any $g$. By applying our methodology based on (12), the weight function $\omega_1$ can be estimated by $\hat{\omega}_1(x) = \exp(\hat{\lambda}_1' g(x))$, where

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

and $\theta_1$ is estimated by $\hat{\theta}_1 = \mathbb{E}_n[\hat{\omega}_1(X)DY]$. Similarly we can estimate $\omega_0$ and $\theta_0$ (by replacing $D$ with $(1 - D)$). Then the average treatment effect $\tau$ can be estimated by $\hat{\tau} = \hat{\theta}_1 - \hat{\theta}_0$.

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 $\mathbb{E}[Y^2(0)] < \infty$, $\mathbb{E}[Y^2(1)] < \infty$, and

$$\begin{aligned} \sup_{x \in \mathcal{X}} |\mathbb{E}[Y(1)|X = x] - \lambda_1' g(x)| &= o(1), \\ \sup_{x \in \mathcal{X}} |\mathbb{E}[Y(0)|X = x] - \lambda_0' g(x)| &= o(1), \end{aligned}$$

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

**(i):** [Low dimensional case] Under the assumptions of Theorem 2, it holds

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

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

**(ii):** [High dimensional case] Under the assumptions of Theorem 4, it holds

$$\sqrt{n}(\hat{\tau} - \tau + \Upsilon^1 - \Upsilon^0) \xrightarrow{d} N(0, \Sigma),$$

where $\Upsilon^1$ and $\Upsilon^0$ are defined as in (29) by setting $h(\cdot) = DY$ and $(1 - D)Y$, respectively.

Proofs are similar to those of Theorems 2 and 4. 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 performance of the proposed method, we consider Example 1 in Section 1 and estimate the SDF in an equity market. We compare out-of-sample performance of the proposed method with other leading factors in empirical finance literature. In particular, the approach adopted by Ghosh, Julliard and Taylor (2016) is a special case of ours for the low (and fixed) dimensional case. Our major findings are: (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 the cross sectional errors are lower, and (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 while Ghosh, Julliard and Taylor's (2016) method shows erratic behaviors.

5.1. **Methodology.** Following the convention in empirical finance, we estimate the SDF on a rolling out-of-sample basis. To be precise, in July year $l$, we form a training subsample using portfolio returns data of past 30 years. Based on this training subsample, we estimate the SDF using the proposed method. In particular, the moment condition in (1) is written as

$$\mathbb{E}[\omega(R_{t_1})R_{t_1}] = 0,$$

where $R_{t_1}$ is a vector of portfolio excess returns at time $t_1$ (a month between year $l$ and $l - 30$) and $\omega$ is the SDF of interest. By applying our method, the SDF can be estimated

by $\hat{\omega}(R_{t_1}) = \frac{\exp(\hat{\lambda}' R_{t_1})}{T_1^{-1}\sum_{t_1=1}^{T_1}\exp(\hat{\lambda}' R_{t_1})}$,[3] where

$$\hat{\lambda} = \begin{cases} \arg\min_{\lambda} T_1^{-1}\sum_{t_1=1}^{T_1}\exp(\lambda' R_{t_1}) & \text{(low dimensional portfolios)} \\ \arg\min_{\lambda} T_1^{-1}\sum_{t_1=1}^{T_1}\exp(\lambda' R_{t_1}) + \alpha_n \left\|\lambda\right\|_1 & \text{(high dimensional portfolios)} \end{cases}.$$

Based on this $\hat{\lambda}$, we predict the SDF using a testing subsample one year ahead from year $l$ with total time periods of $T_2$. Then the estimated out-of-sample SDF from July year $l$ to June year $l+1$ would be $\hat{\omega}(R_{t_2}) = \frac{\exp(\hat{\lambda}' R_{t_2})}{T_2^{-1}\sum_{t_2=1}^{T_2}\exp(\hat{\lambda}' R_{t_2})}$, where $t_2$ is a month between year $l$ and $l+1$. We continue to build the estimated SDF time series in this fashion to cover all periods in our sample.

To test the cross-sectional predictability of our estimated out-of-sample SDF, we use the two-pass regression in empirical finance (Fama and MacBeth, 1973, and Cochrane, 2009). In the first step, we run a time series OLS regression of excess returns $R_j$ on our estimated out-of-sample SDF $\hat{\omega}$ for each portfolio $j$. We record its slope coefficient $\hat{\beta}_j$ as its factor loading. Then in the second step, we run a cross sectional OLS regression from $\bar{R}$ on $\hat{\beta}$, where $\bar{R}$ is a vector of average excess returns for all portfolios, and $\hat{\beta}$ is a vector of estimated factor loadings in the first step. We compare the adjusted R-squared as well as the estimated constant in the second regression to other empirical asset pricing models.

5.2. **Data.** All data are taken from Kenneth French's data library. 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), the out-of-sample evaluation covers from July 1963 to December 2010. We use monthly data so each training subsample is of size $T_1 = 360$ and each testing subsample is of size $T_2 = 12$ (except for the last rolling window where $T_2 = 6$). We only consider equity portfolios returns, which are quoted in %.

We compare three methods: Our method without penalty (essentially, Ghosh, Julliard and Taylor, 2016), our method with $\ell_1$-penalization, and Fama-French three factors model. These methods are compared under the following scenarios.

- **(i):** Low dimensional case: the SDF is constructed from 25 size and book-to-market portfolios, 10 momentum portfolios, 25 size, and long term reversal portfolios, respectively.
- **(ii):** Intermediate case: the SDF is constructed from 100 size and book-to-market portfolios, 49 industry portfolios, and 25 long term reversal and size+25 short term reversal and size+25 momentum portfolios, respectively.
- **(iii):** High dimensional case: use all portfolios available from Kenneth French's data library. Since some data are only available from 1960s, the out-of-sample period

[3]Since the vector $R_{t_1}$ does not contain 1, the solution in (5) needs to be normalized by the empirical average.

can only cover months from July 1993 to December 2010. The SDF is constructed from the two sets of portfolios: **(a)** 300 portfolios that include 100 portfolios based on size and book-to-market, 100 portfolios based on size and operating profitability, and 100 portfolios formed on size and investment, and **(b)** 425 portfolios that include 300 portfolios above, 49 industry portfolios, 25 portfolios on long term reversal and size, 25 portfolios on short term reversal and size, and 25 momentum portfolios.

### 5.3. **Empirical result.**

5.3.1. *Low dimensional and intermediate cases.* Table 1 presents the cross sectional regression results for the low dimensional case. The numbers of portfolios are less than 30 in all panels and the training subsample size is 360. Although penalization seems unnecessary, we present the result when the penalty level is 0.05, a relatively small penalty, for comparison. The numbers in parentheses are t-values for the coefficients above. In all panels, the estimated price of risk is highly significant with the correct sign, either with or without penalty. The adjusted R-squared for the no penalty estimate is larger than the one for the Fama-French model in Panels A and B. The adjusted R-squared for the penalized estimate is worse than the one for the no penalty estimate in these two panels. Since the dimension is low, we expect every portfolio is informative and there is no need for penalization. Moreover, we can see that the intercept estimates are all much smaller than the Fama-French estimates. This also indicates that our model is better than Fama-French three factor models. Panel C is interesting, where the no penalty estimate is worse than the penalized estimate. This result indicates usefulness of penalization even for the low dimensional case.

Table 2 summarizes the results for the intermediate case, where the number of portfolios ranges from 50 to 100. The results are similar to the low dimensional case in Table 1. Our method (with or without penalty) outperforms the Fama-French model for most cases in terms the intercept estimates and adjusted R-squared.

5.3.2. *High dimensional case.* This case is of our major interest, where the no penalty estimate (essentially Ghosh, Julliard and Taylor, 2015) is not applicable or performs erratically, and it is crucial to introduce penalization. In this case, the choice of the penalty level becomes more important. We create a grid from 0.1 to 2 with 0.05 increments, estimate the SDF by our method, and implement the cross sectional regression for each penalty level.

The results are summarized in Figure 1. The SDF estimates without penalization perform very badly with the adjusted R-squared close to 0 and relatively large intercept estimates. As the penalty level increases, the performance of our method gets better. When the penalty level is approximately above 0.5, predictability of our method surpasses

Fama-French, and the intercept estimates are much smaller. Then performance of our method gets worse when the penalty level continues to increase above 1.5. This is expected because the number of portfolios selected will be too small for too large penalty levels and the performance would deteriorate. Based on these results, we set the penalty level at 0.9 for 300 portfolios and 0.85 for 425 portfolios, and report the results in Table 3. We can see that the adjusted R-squared by the penalized SDF estimate is much higher than the one of Fama-French and that its intercept estimate is close to 0 and insignificant. Therefore, our method shows excellent performance upon choosing suitable penalty levels.

The number of active portfolios chosen in each year for each penalty level is summarized in Figures 2 (300 portfolios) and 3 (425 portfolios). Without penalization, too many portfolios will be used and cause undesirable performance. As the penalty level increases, the number of selected portfolios drops quickly, and at the levels used for Table 3, the number of portfolios selected is around 5-10.

5.3.3. *Time series property of penalized SDF estimates.* We illustrate time series properties of the penalized SDF estimates for 300 and 425 portfolios at the penalty levels used for Table 3. The plot is displayed in Figure 4 and the gray shaded areas correspond to NBER recessions. Our SDF estimates catch those macro events very well. In Table 4 we run a time series regression of our SDF estimates on other key factors in the market including Fama-French 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-squared is also small. This indicates that our method catches critical information for asset pricing in the market that cannot be explained by Fama-French and momentum factors.

## Appendix A. Proofs for low dimensional case

**Notation.** In, Appendix A, we use the following notation. Recall $\lambda_* = \arg\min_\lambda \mathbb{E}[\exp(\lambda' g(X)) - \lambda' r]$. Let $\omega_*(x) = \exp(\lambda_*' g(x))$, $A = \mathbb{E}[g(X)g(X)']^{-1/2}$, $\mu(x) = Ag(x)$, $\beta_* = A^{-1}\lambda_*$. Note that $\zeta_K = \sup_{x\in\mathcal{X}} |\mu(x)|$. Also define the empirical process and excess risk function as

$$\begin{aligned} \nu_n(\lambda) &= \mathbb{E}_n[\exp(\lambda' g(X)) - \lambda' r] - \mathbb{E}[\exp(\lambda' g(X)) - \lambda' r], \\ \mathscr{E}(\lambda) &= \mathbb{E}[\exp(\lambda' g(X)) - \lambda' r] - \mathbb{E}[\exp(\lambda_*' g(X)) - \lambda_*' r], \end{aligned}$$

respectively.

### A.1. **Proof of Theorem 1.**

**Proof of the first statement.** By an expansion around $\hat\lambda = \lambda_*$,

$$\begin{aligned} \hat\omega(x) - \omega_*(x) &= \exp\left(\lambda_*' g(x) + t_x(\hat\lambda - \lambda_*)' g(x)\right)(\hat\lambda - \lambda_*)' g(x) \\ &= \exp\left(\beta_*' \mu(x) + t_x(\hat\lambda - \lambda_*)' A^{-1}\mu(x)\right)(\hat\lambda - \lambda_*)' A^{-1}\mu(x), \end{aligned} \tag{13}$$

for some $t_x \in [0,1]$. By Lemma 2 (ii),

$$\sup_{x\in\mathcal{X}} |(\hat\lambda - \lambda_*)' A^{-1}\mu(x)| \le |\hat\lambda - \lambda_*| \sqrt{\lambda_{\min}^{-1}(AA')}\zeta_K = O_p(\zeta_K\sqrt{K/n}). \tag{14}$$

Let $\mathcal{E}_n$ be the event that $\beta_*'\mu(x) + t_x(\hat\lambda - \lambda_*)' A^{-1}\mu(x) \in \Gamma$ for all $x \in \mathcal{X}$, where $\Gamma$ is a compact set defined in (23). From (13) and (14), for each $\delta > 0$, there exists $C_\delta > 0$ such that

$$\Pr\left\{\left\{\sup_{x\in\mathcal{X}} |\hat\omega(x) - \omega_*(x)| > C_\delta \zeta_K\sqrt{K/n}\right\} \cap \mathscr{E}_n\right\} \le \delta,$$

for all $n$ large enough. From (22) and (23) in the proof of Lemma 2 (iii), it holds $\beta_*'\mu(x) \in \mathrm{int}(\Gamma)$ for all $x \in \mathcal{X}$ and $n$ large enough. Therefore, (14) implies $\Pr\{\mathcal{E}_n\} \to 1$.

Combining these results,

$$\sup_{x\in\mathcal{X}} |\hat\omega(x) - \omega_*(x)| = O_p(\zeta_K\sqrt{K/n}),$$

and the conclusion follows by the triangle inequality and Lemma 2 (iii).

**Proof of the second statement.** Observe that

$$\begin{aligned} |\hat\theta - \theta| &\le |\mathbb{E}_n[\hat\omega(X)h(X,Y)] - \mathbb{E}_n[\omega(X)h(X,Y)]| + |\mathbb{E}_n[\omega(X)h(X,Y)] - \mathbb{E}[\omega(X)h(X,Y)]| \\ &\le \sup_{x\in\mathcal{X}} |\hat\omega(x) - \omega(x)| \frac{1}{n}\sum_{i=1}^n |h(X_i,Y_i)| + o_p(1) = O_p(\zeta_K(\sqrt{K/n} + K^{-\eta/2})) + o_p(1), \end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from the law of large numbers under Condition D, and the equality follows from Theorem 1 (i).

A.2. **Proof of Theorem 2.** Let

$$\begin{aligned} h_i &= h(X_i, Y_i), \qquad h_i^X = \mathbb{E}[h_i|X_i], \qquad \omega_i = \omega(X_i), \qquad \mu_i = \mu(X_i), \\ \omega_{*i} &= \exp(\beta_*'\mu_i), \qquad \hat{\omega}_i = \exp(\hat{\beta}'\mu_i), \end{aligned}$$

By an expansion of $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \exp(\hat{\beta}'\mu_i)h_i$ around $\hat{\beta} = \beta_*$, we obtain

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\omega_i h_i - \theta) + T_1 + T_2 + T_3 + T_4,$$

where

$$\begin{aligned} T_1 &= \mathbb{E}[\omega_{*i}h_i\mu_i']\sqrt{n}(\hat{\beta} - \beta_*), \qquad T_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\omega_{*i}h_i\mu_i' - \mathbb{E}[\omega_{*i}h_i\mu_i']\}(\hat{\beta} - \beta_*), \\ T_3 &= \frac{1}{2}(\hat{\beta} - \beta_*)'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\omega}_i h_i\mu_i\mu_i'\right)(\hat{\beta} - \beta_*), \qquad T_4 = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\omega_{*i}h_i - \omega_i h_i) \end{aligned}$$

and $\tilde{\omega}_i = \exp(\tilde{\beta}'\mu_i)$ with some $\tilde{\beta}$ on the line joining $\hat{\beta}$ and $\beta_*$.

First, we consider $T_2$. The definition of $\zeta_K$ and (19) imply $\mathbb{E}[|\omega_{*i}h_i\mu_i|^2] = O(\zeta_K^2)$. Thus, Chebyshev's inequality and Lemma 2 (ii) imply $T_2 = O_p(\zeta_K\sqrt{K/n})$.

Next, we consider $T_3$. The definition of $\zeta_K$, (19), and Lemma 2 (ii) imply $\mathbb{E}[|\tilde{\omega}_i h_i\mu_i\mu_i'|] = O(\zeta_K^2)$ and $\mathbb{E}[|\tilde{\omega}_i h_i\mu_i\mu_i'|^2] = O(\zeta_K^4)$. Thus, Chebyshev's inequality and Lemma 2 (ii) imply $T_3 = O_p(\zeta_K^2 K/\sqrt{n})$.

Third, we consider $T_4$. From Lemma 2 (iii) and law of large numbers, we have $T_4 = O_p(\sqrt{n}\zeta_K K^{-\eta/2})$.

We now consider $T_1$. By expanding the first order condition of $\hat{\beta}$,

$$0 = \frac{1}{n}\sum_{i=1}^n \exp(\hat{\beta}'\mu_i)\mu_i - Ar = \frac{1}{n}\sum_{i=1}^n (\omega_{*i}\mu_i - Ar) + \frac{1}{n}\sum_{i=1}^n \bar{\omega}_i\mu_i\mu_i'(\hat{\beta} - \beta_*), \qquad (15)$$

where $\bar{\omega}_i = \exp(\bar{\beta}'\mu_i)$ with some $\bar{\beta}$ on the line joining $\hat{\beta}$ and $\beta_*$. Let $\psi = \mathbb{E}[\omega_{*i}h_i\mu_i']$, $\Sigma = \mathbb{E}[\omega_{*i}\mu_i\mu_i']$ and $\bar{\Sigma} = \frac{1}{n}\sum_{i=1}^n \bar{\omega}_i\mu_i\mu_i'$. By solving this for $\hat{\beta} - \beta_*$ and inserting to $T_1$, we have

$$T_1 = -\psi\bar{\Sigma}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \{\omega_{*i}\mu_i - \mathbb{E}[\omega_i\mu_i]\}\right) = T_{11} + T_{12} + T_{13},$$

where

$$\begin{aligned} T_{11} &= -\psi(\bar{\Sigma}^{-1} - \Sigma^{-1})\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \{\omega_{*i}\mu_i - \mathbb{E}[\omega_i\mu_i]\}\right), \\ T_{12} &= -\psi\Sigma^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (\omega_{*i} - \omega_i)\mu_i\right), \qquad T_{13} = -\psi\Sigma^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \{\omega_i\mu_i - \mathbb{E}[\omega_i\mu_i]\}\right). \end{aligned}$$

For $T_{12}$, note that

$$|T_{12}| \leq |\psi| \frac{1}{\sqrt{\lambda_{\min}(\Sigma'\Sigma)}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} - \omega_i)\mu_i \right|.$$

It is easy to see $|\psi| = O(\zeta_K)$ due to the definition of $\zeta_K$. Lemma 2 (iii) and law of large numbers imply $\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} - \omega_i)\mu_i \right| = O_p(\sqrt{n}\zeta_K K^{-\eta/2})$. Since $\lambda_{\min}(\Sigma'\Sigma)$ is bounded away from zero, we have $T_{12} = O_p(\sqrt{n}\zeta_K^2 K^{-\eta/2})$.

For $T_{11}$, note that (15) implies

$$T_{11} = \sqrt{n}\psi(\bar{\Sigma}^{-1} - \Sigma^{-1})\bar{\Sigma}(\hat{\beta} - \beta_*) = \sqrt{n}\psi\Sigma^{-1}(\Sigma - \bar{\Sigma})(\hat{\beta} - \beta_*),$$

which can be bounded as

$$|T_{11}| \leq \sqrt{n}|\psi| \frac{1}{\sqrt{\lambda_{\min}(\Sigma'\Sigma)}} |\Sigma - \bar{\Sigma}| \cdot |\hat{\beta} - \beta_*|.$$

By the triangle inequality,

$$|\Sigma - \bar{\Sigma}| \leq |\mathbb{E}_n[(\bar{\omega}_i - \omega_{*i})\mu_i\mu_i']| + |\mathbb{E}_n[\omega_{*i}\mu_i\mu_i'] - \mathbb{E}[\omega_{*i}\mu_i\mu_i']| = O_p(\zeta_K^3\sqrt{K/n}),$$

where the equality follows from an expansion of $\bar{\omega}_i$ and Chebyshev's inequality. Therefore, we obtain $|\Sigma - \bar{\Sigma}| = O_p(\zeta_K^3\sqrt{K/n})$, and $|\psi| = O(\zeta_K)$ and Lemma 2 (ii) imply $|T_{11}| = O_p(\zeta_K^4 K/\sqrt{n})$.

Now consider $T_{13}$. We show that

$$T_{13} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\omega_i h_i^X - \mathbb{E}[\omega_i h_i^X]\} + o_p(1). \tag{16}$$

To this end, observe that

$$\begin{aligned} & \mathbb{E}[\{\omega_i h_i^X - \mathbb{E}[\omega_i h_i^X]\} - \psi'\Sigma^{-1}\{\omega_i\mu_i - \mathbb{E}[\omega_i\mu_i]\}]^2 \\ \leq\ & \mathbb{E}[\omega_i^2(h_i^X - \psi'\Sigma^{-1}\mu_i)^2] \leq \left(\sup_{x\in\mathcal{X}} \frac{\omega^2(x)}{\omega_*(x)}\right) \mathbb{E}[\omega_{*i}(h_i^X - \psi'\Sigma^{-1}\mu_i)^2] = \left(\sup_{x\in\mathcal{X}} \frac{\omega^2(x)}{\omega_*(x)}\right) \mathbb{E}[(\tilde{h}_i - \gamma_p'\tilde{\mu}_i)^2], \end{aligned}$$

where $\tilde{h}_i = \sqrt{\omega_{*i}}h_i^X$, $\tilde{\mu}_i = \sqrt{\omega_{*i}}\mu_i$, and $\gamma_p = \mathbb{E}[\tilde{\mu}_i\tilde{\mu}_i']^{-1}\mathbb{E}[\tilde{\mu}_i\tilde{h}_i]$. Since $\gamma_p$ is the projection coefficient that solves $\min_\gamma \mathbb{E}[(\tilde{h}_i - \gamma'\tilde{\mu}_i)^2]$, the assumption in (8) guarantees $\mathbb{E}[(\tilde{h}_i - \gamma_p'\tilde{\mu}_i)^2] \to 0$. Therefore, due to boundedness of $\omega$ and $\omega_*$ (by Condition D and Lemma 2 (iii)) and Chebyshev's inequality, (16) is satisfied, and the conclusion is obtained as

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i(h_i - h_i^X) + O_p(r_n)$$

where $r_n = \sqrt{n}\zeta_K^2 K^{-\eta/2} + \zeta_K^4 K/\sqrt{n}$. Since $r_n \to 0$ by the assumption, the conclusion follows by the central limit theorem.

### A.3. **Lemmas.**

**Lemma 1.** Suppose Conditions D and S hold true. Then for every $\varepsilon > 0$ small enough, there exist positive constants $\sigma_\varepsilon$, $\varrho$, and $C$ such that for $\delta = \frac{4\sigma_\varepsilon}{\varrho}\sqrt{\frac{K}{n}}$ and all $n$ large enough,

**(i):** $\Pr\{\sup_{|\lambda-\lambda_*|\leq\delta} |\nu_n(\lambda) - \nu_n(\lambda_*)| \leq \sigma_\varepsilon\delta\sqrt{K/n}\} \geq 1-\varepsilon$,
**(ii):** $\{\lambda : |\lambda-\lambda_*| \leq \delta\} \subseteq \{\lambda : \left\|\lambda-\lambda_*\right\|_\infty \leq C,\ \varrho|\lambda-\lambda_*|^2 \leq \mathscr{E}(\lambda)\}$.

**Lemma 2.** Under Conditions D and S, it holds

**(i):** $\Pr\left\{\mathscr{E}(\hat\lambda) \leq \frac{4\sigma_\varepsilon^2}{\varrho}\frac{K}{n}\right\} \geq 1-\varepsilon$,
**(ii):** $|\hat\lambda - \lambda_*| = O_p(\sqrt{K/n})$,
**(iii):** $\sup_{x\in\mathcal{X}} |\omega_*(x) - \omega(x)| = O(\zeta_K K^{-\eta/2})$.

**Proof of Lemma 1 (i).** Define $\kappa_{\max} = \lambda_{\max}\{\mathbb{E}[g(X)g(X)']\}$, $\kappa_{\min} = \lambda_{\min}\{\mathbb{E}[g(X)g(X)']\}$, $L = \max_{\lambda:|\lambda-\lambda_*|\leq 1, x\in\mathcal{X}} \exp(\lambda' g(x))$, $l = \min_{\lambda:|\lambda-\lambda_*|\leq 1, x\in\mathcal{X}} \exp(\lambda' g(x))$. By Conditions D and S, these are all finite. Pick any $\varepsilon > 0$ and set

$$\sigma_\varepsilon = \frac{4\sqrt{\kappa_{\max}}L}{\varepsilon}, \qquad \varrho = \kappa_{\min} l, \qquad C = 1, \qquad \delta = \frac{4\sigma_\varepsilon}{\varrho}\sqrt{\frac{K}{n}}.$$

Then the conclusion is obtained as

$$\begin{aligned} &\Pr\left\{\sup_{|\lambda-\lambda_*|\leq\delta} |\nu_n(\lambda) - \nu_n(\lambda_*)| > \sigma_\varepsilon\delta\sqrt{\frac{K}{n}}\right\} \\ \leq\ & \mathbb{E}\left[\sup_{|A^{-1}(\lambda-\lambda_*)|\leq\sqrt{\kappa_{\max}}\delta} |\nu_n(\lambda) - \nu_n(\lambda_*)|\right]\frac{1}{\sigma_\varepsilon\delta}\sqrt{\frac{n}{K}} \leq \frac{4\sqrt{\kappa_{\max}}L}{\sigma_\varepsilon} = \varepsilon, \end{aligned}$$

where the first inequality follows from the Markov inequality, and the second inequality follows from Bühlmann and van de Geer (2011, Lemma 14.19).

**Proof of Lemma 1 (ii).** Pick any $\lambda$ satisfying $|\lambda-\lambda_*| \leq \delta$. Since $\delta = \frac{4\sigma_\varepsilon}{\varrho}\sqrt{\frac{K}{n}} \to 0$ as $n\to\infty$, we have $\left\|\lambda-\lambda_*\right\|_\infty \leq 1$ for all $n$ large enough. Also, the second-order expansion of $\mathscr{E}(\lambda)$ around $\lambda = \lambda_*$ yields

$$\mathscr{E}(\lambda) = \frac{1}{2}(\lambda-\lambda_*)'\mathbb{E}[\exp(\tilde\lambda' g(X))g(X)g(X)'](\lambda-\lambda_*) \geq \frac{\kappa_{\min} l}{2}|\lambda-\lambda_*|^2,$$

where $\tilde\lambda$ is a point on the line joining $\lambda$ and $\lambda_*$, and the inequality follows from the definitions of $\kappa_{\min}$ and $l$. Then the conclusion follows by the definition of $\varrho$.

**Proof of Lemma 2 (i).** Pick any $\varepsilon > 0$ small enough and $n \in \mathbb{N}$ large enough. Then take $\delta = \frac{4\sigma_\varepsilon}{\varrho}\sqrt{\frac{K}{n}}$ to satisfy the statements in Lemma 1. Let $t = \frac{\delta}{\delta+|\hat\lambda-\lambda_*|}$ and $\bar\lambda = t\hat\lambda + (1-t)\lambda_*$. Due to the convexity of the exponential function,

$$\begin{aligned} \mathbb{E}_n[\exp(\bar\lambda' g(X)) - \bar\lambda' r] &\leq t\mathbb{E}_n[\exp(\hat\lambda' g(X)) - \hat\lambda' r] + (1-t)\mathbb{E}_n[\exp(\lambda_*' g(X)) - \lambda_*' r] \\ &\leq \mathbb{E}_n[\exp(\lambda_*' g(X)) - \lambda_*' r], \end{aligned}$$

where the second inequality follows from the definition of $\hat{\lambda}$. Thus, we have

$$\mathscr{E}(\bar{\lambda}) \leq \nu_n(\lambda_*) - \nu_n(\bar{\lambda}),$$

and it is enough to bound the incremental empirical process part. Since

$$|\bar{\lambda} - \lambda_*| = \frac{\delta|\hat{\lambda} - \lambda_*|}{\delta + |\hat{\lambda} - \lambda_*|} \leq \delta, \tag{17}$$

Lemma 1 (i) and the above inequality imply

$$\Pr\{\mathscr{E}(\bar{\lambda}) \leq \sigma_\varepsilon \delta \sqrt{K/n}\} \geq 1 - \varepsilon.$$

Note that $xy \leq x^2 + \frac{y^2}{4}$ for any $x, y \in \mathbb{R}$. Thus by setting $x = \delta\sqrt{\varrho}$ and $y = 8\sigma_\varepsilon\sqrt{\frac{K}{\varrho n}}$, we have $\sigma_\varepsilon \delta \sqrt{K/n} \leq \frac{4\sigma_\varepsilon^2 K}{\varrho n}$ and

$$\Pr\left\{\mathscr{E}(\bar{\lambda}) \leq \frac{4\sigma_\varepsilon^2}{\varrho}\frac{K}{n}\right\} \geq 1 - \varepsilon.$$

Now, by (17) and Lemma 1 (ii), it holds

$$|\bar{\lambda} - \lambda_*| \leq \sqrt{\frac{\mathscr{E}(\bar{\lambda})}{\varrho}} \leq \frac{1}{2}\delta,$$

where the second inequality follows from Lemma 2 (i). This inequality and $|\bar{\lambda} - \lambda_*| = \frac{\delta|\hat{\lambda}-\lambda_*|}{\delta+|\hat{\lambda}-\lambda_*|}$ imply $|\hat{\lambda} - \lambda_*| \leq \delta$. Thus, Lemma 1 (i) and the above inequality imply

$$\Pr\{\mathscr{E}(\hat{\lambda}) \leq \sigma_\varepsilon \delta \sqrt{K/n}\} \geq 1 - \varepsilon.$$

Due to $\sigma_\varepsilon \delta \sqrt{K/n} \leq \frac{4\sigma_\varepsilon^2 K}{\varrho n}$, the conclusion follows.

**Proof of Lemma 2 (ii).** In the proof of Part (i) of this lemma, we obtained $|\hat{\lambda} - \lambda_*| \leq \delta = \frac{4\sigma_\varepsilon}{\varrho}\sqrt{\frac{K}{n}}$, which implies the conclusion.

**Proof of Lemma 2 (iii).** Recall $\mu(x) = Ag(x)$. By Condition S, there exists $b_* \in \mathbb{R}^K$ such that

$$\sup_{x\in\mathcal{X}} |\log(\omega(x)) - b_*'\mu(x)| = O(K^{-\eta}). \tag{18}$$

By Condition D, both $\underline{\gamma} = \inf_{x\in\mathcal{X}} \log(\omega(x))$ and $\overline{\gamma} = \sup_{x\in\mathcal{X}} \log(\omega(x))$ are finite. Thus, there exists $C_1 > 0$ such that

$$b_*'\mu(x) \in [\underline{\gamma} - C_1K^{-\eta}, \overline{\gamma} + C_1K^{-\eta}], \tag{19}$$

for all $x \in \mathcal{X}$ and $n$ large enough. Also, (18) guarantees

$$\omega(x) - \exp(b_*'\mu(x)) \in [\exp(b_*'\mu(x) - C_1K^{-\eta}) - \exp(b_*'\mu(x)), \exp(b_*'\mu(x) + C_1K^{-\eta}) - \exp(b_*'\mu(x))],$$

for all $x \in \mathcal{X}$ and $n$ large enough. By applying the mean value theorem to the upper and lower bounds, there exist $c_1, c_2 > 0$ such that

$$\begin{aligned}\exp(b_*'\mu(x) + C_1K^{-\eta}) - \exp(b_*'\mu(x)) &\leq c_1C_1K^{-\eta},\\ \exp(b_*'\mu(x) - C_1K^{-\eta}) - \exp(b_*'\mu(x)) &\geq -c_2C_1K^{-\eta},\end{aligned}$$

for all $x \in \mathcal{X}$ and $n$ large enough. Thus, we have

$$\sup_{x\in\mathcal{X}} |\omega(x) - \exp(b_*'\mu(x))| = O(K^{-\eta}). \tag{20}$$

Now define

$$Q(b) = \mathbb{E}[\exp(b'\mu(X)) - b'Ar], \qquad Q_*(b) = \mathbb{E}\left[\frac{\omega(X)\exp(b'\mu(X))}{\exp(b_*'\mu(X))} - b'Ar\right].$$

Note that $\beta_* = A^{-1}\lambda_*$ solves $\min_b Q(b)$, and $b_*$ solves $\min_b Q_*(b)$ because

$$\frac{\partial Q_*(b_*)}{\partial b} = \mathbb{E}[\omega(X)\mu(X) - Ar] = 0. \tag{21}$$

Define the set $\mathit{\Pi}_K = \{b : |b - b_*| \leq K^{-\eta/2}\}$ and its boundary $\partial\mathit{\Pi}_K = \{b : |b - b_*| = K^{-\eta/2}\}$. We show the claim:

$$\text{if } Q(b_*) - Q(b) < 0 \text{ for all } b \in \partial\mathit{\Pi}_K, \text{ then } \beta_* \in \mathit{\Pi}_K. \tag{22}$$

We prove (22) by its contrapositive. Suppose $\beta_* \notin \mathit{\Pi}_K$. Since $Q(b)$ is strictly convex in $b$, $\beta_*$ is a unique global minimizer of $Q(b)$. Thus, if $\beta_* \notin \mathit{\Pi}_K$, then there exists $\tilde{b} \in \partial\mathit{\Pi}_K$ such that $Q(\tilde{b}) < Q(b_*)$. This completes the proof of (22).

Furthermore, for any $b \in \mathit{\Pi}_K$, it holds $|b'\mu(x) - b_*'\mu(x)| \leq K^{-\eta/2}\zeta_K$, i.e.,

$$b'\mu(x) \in [b_*'\mu(x) \pm K^{-\eta/2}\zeta_K] \subset [\underline{\gamma} - C_1K^{-\eta} - K^{-\eta/2}\zeta_K, \overline{\gamma} + C_1K^{-\eta} + K^{-\eta/2}\zeta_K],$$

for all $x \in \mathcal{X}$ and $n$ large enough, where the set inclusion follows from (19). Since $K \to \infty$, there exists $\epsilon > 0$ such that

$$b'\mu(x) \in \Gamma = [\underline{\gamma} - \epsilon, \overline{\gamma} + \epsilon], \tag{23}$$

for all $b \in \mathit{\Pi}_K$, $x \in \mathcal{X}$, and $n$ large enough.

From (20) and (23), there exists $C_2 > 0$ such that

$$|Q_*(b) - Q(b)| = \left|\mathbb{E}\left[\frac{\exp(b'\mu(X))}{\exp(b_*'\mu(X))}\{\omega(X) - \exp(b_*'\mu(X))\}\right]\right| \leq C_2K^{-\eta},$$

for all $b \in \mathit{\Pi}_K$, $x \in \mathcal{X}$, and $n$ large enough. Thus, for all $b \in \partial\mathit{\Pi}_K$,

$$Q(b_*) - Q(b) \leq Q_*(b_*) - Q_*(b) + 2C_2K^{-\eta}. \tag{24}$$

By an expansion around $b = b_*$ using (21), there exist $\tilde{b} \in (b, b_*)$ and $c_3 > 0$ such that

$$\begin{aligned} Q_*(b) - Q_*(b_*) &= \frac{1}{2}(b - b_*)' \mathbb{E}\left[ \frac{\omega(X)\exp(\tilde{b}'\mu(X))}{\exp(b_*'\mu(X))} \mu(X)\mu(X)' \right] (b - b_*) \\ &> c_3 |b - b_*|^2 = c_3 K^{-\eta}, \end{aligned} \tag{25}$$

for all $b \in \partial \Pi_K$, $x \in \mathcal{X}$, and $n$ large enough, where the inequality follows from (23) and $\mathbb{E}[\mu(X)\mu(X)'] = I_K$. By (24) and (25), we can take large enough $C_2$ and small enough $c_3$ to satisfy the condition of (22), which implies

$$|b_* - \beta_*| \leq K^{-\eta/2}, \tag{26}$$

for all $n$ large enough. Therefore, by an expansion with (23), there exists $C_3 > 0$ such that

$$|\exp(b_*'\mu(x)) - \exp(\beta_*'\mu(x))| \leq C_3 |b_* - \beta_*| \cdot |\mu(x)| \leq C_3 \zeta_K K^{-\eta/2},$$

for all $x \in \mathcal{X}$ and $n$ large enough. By using this and (18), the conclusion is obtained as

$$\sup_{x \in \mathcal{X}} |\omega_*(x) - \omega(x)| = O(\zeta_K K^{-\eta/2} + K^{-\eta}) = O(\zeta_K K^{-\eta/2}).$$

## Appendix B. Proofs for high dimensional case

**Notation.** In addition to the notation in Appendix A, let $\omega_o(x) = \exp(\lambda_o' g(x))$.

B.1. **Proof of Theorem 3.** By the mean value theorem, there exists $t_x \in [0, 1]$ such that

$$\hat{\omega}(x) - \omega_o(x) = \exp\left( \lambda_o' g(x) + t_x (\hat{\lambda} - \lambda_o)' g(x) \right) (\hat{\lambda} - \lambda_o)' g(x), \tag{27}$$

for each $x \in \mathcal{X}$. By the Hölder inequality and Lemma 3 (ii),

$$\sup_{x \in \mathcal{X}} |t_x (\hat{\lambda} - \lambda_o)' g(x)| \leq \left\| \hat{\lambda} - \lambda_o \right\|_1 \tilde{\zeta}_K = O_p(\tilde{\zeta}_K \kappa_{o,n}). \tag{28}$$

Since $|\lambda_o' g(x)|$ is bounded uniformly over all $x \in \mathcal{X}$ and all $n$ large enough, the result in (28) guarantees $\Pr\{\mathcal{E}_n\} \to 1$, where $\mathcal{E}_n$ is the event that $\exp\left( \lambda_o' g(x) + t_x (\hat{\lambda} - \lambda_o)' g(x) \right)$ lies in a fixed bounded set for all $x \in \mathcal{X}$.

On the event $\mathcal{E}_n$, (27) and (28) imply

$$\sup_{x \in \mathcal{X}} |\hat{\omega}(x) - \omega_o(x)| \leq C_2 \left\| \hat{\lambda} - \lambda_o \right\|_1 \tilde{\zeta}_K = O_p(\tilde{\zeta}_K \kappa_{o,n}),$$

for some $C_2 > 0$. Therefore, from $\Pr\{\mathcal{E}_n\} \to 1$ and the assumption on $\sup_{x \in \mathcal{X}} |\omega_o(x) - \omega(x)|$, the conclusion is obtained as

$$\sup_{x \in \mathcal{X}} |\hat{\omega}(x) - \omega(x)| \leq \sup_{x \in \mathcal{X}} |\hat{\omega}(x) - \omega_o(x)| + \sup_{x \in \mathcal{X}} |\omega_o(x) - \omega(x)| = O_p(\tilde{\zeta}_K \kappa_{o,n}).$$

The proof of consistency of $\hat{\theta}$ is similar to that of Theorem 1.

B.2. **Proof of Theorem 4.** Let $a_s$ and $a_{s^c}$ mean $s$- and $(K-s)$ dimensional subvectors of $a \in \mathbb{R}^K$, respectively, and define

$$\begin{aligned} g_{si} &= g_s(X_i), \qquad A_s = \mathbb{E}[g_{si}g_{si}']^{-1/2}, \qquad \mu_i = A_s g_{si}, \\ h_i &= h(X_i, Y_i), \qquad h_i^X = \mathbb{E}[h_i|X_i], \qquad \omega_{oi} = \omega_o(X_i), \\ \psi_s &= \mathbb{E}[\omega_{oi}h_i g_{si}'], \qquad \Sigma_s = \mathbb{E}[\omega_{oi}g_{si}g_{si}']. \end{aligned}$$

Based on this notation, the bias term is defined as

$$\Upsilon = \Upsilon_1 + \Upsilon_2 + \Upsilon_3, \tag{29}$$

where

$$\begin{aligned} \Upsilon_1 &= \psi_s \Sigma_s^{-1} \mathbb{E}_n[\omega_{oi}\{\exp(\hat{\lambda}_{s^c}' g_{s^c i}) - 1\} g_{si}], \\ \Upsilon_2 &= \alpha_n \psi_s \Sigma_s^{-1} \hat{\kappa}_s, \qquad \Upsilon_3 = \mathbb{E}_n[\omega_{oi}h_i\{1 - \exp(\hat{\lambda}_{s^c}' g_{s^c i})\}], \end{aligned}$$

and $\hat{\kappa}$ is a $K$-vector with the $j$-th element $\hat{\kappa}_j = \text{sign}(\hat{\lambda}_j)$. Note that

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \exp(\hat{\lambda}' g_i) h_i = \frac{1}{n}\sum_{i=1}^n \exp(\hat{\lambda}_s' g_{si}) \exp(\hat{\lambda}_{s^c}' g_{s^c i}) h_i.$$

By an expansion of $\exp(\hat{\lambda}_s' g_{si})$ around $\hat{\lambda}_s = \lambda_{os}$ combined with $\lambda_{os}' g_{si} = \lambda_o' g_i$, we obtain

$$\sqrt{n}(\hat{\theta} - \theta + \Upsilon) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\omega_i h_i - \theta) + T_1 + T_2 + T_3 + T_4 + T_5 + T_6,$$

where

$$\begin{aligned} T_1 &= \mathbb{E}[\omega_{oi}h_i g_{si}']\sqrt{n}(\hat{\lambda}_s - \lambda_{os}) + \sqrt{n}(\Upsilon_1 + \Upsilon_2), \qquad T_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\omega_{oi}h_i g_{si}' - \mathbb{E}[\omega_{oi}h_i g_{si}']\}(\hat{\lambda}_s - \lambda_{os}), \\ T_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\omega_{oi}\exp(\hat{\lambda}_{s^c}' g_{s^c i}) h_i g_{si}' - \omega_{oi}h_i \mu_{si}']\}(\hat{\lambda}_s - \lambda_{os}), \\ T_4 &= \frac{1}{2}(\hat{\lambda}_s - \lambda_{os})' \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\omega}_i h_i \mu_{si}\mu_{si}'\right)(\hat{\lambda}_s - \lambda_{os}), \qquad T_5 = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\omega_{oi}h_i - \omega_i h_i), \\ T_6 &= \sqrt{n}\Upsilon_3 + \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\omega_{oi}h_i \exp(\hat{\lambda}_{s^c}' g_{s^c i}) - \omega_{oi}h_i\}, \end{aligned}$$

and $\tilde{\omega}_i = \exp(\tilde{\lambda}_s' g_{si}) \exp(\hat{\lambda}_{s^c}' g_{s^c i})$ with some $\tilde{\lambda}_s$ on the line joining $\hat{\lambda}_s$ and $\lambda_{os}$. Note that $T_6 = 0$.

First, consider $T_2$. Similar to the proof of Theorem 3, and using the definition of $\zeta_s$, we have $\mathbb{E}[|\omega_{oi}h_i g_{si}|^2] = O(\zeta_s^2)$. Thus, Chebyshev's inequality and Lemma 3 (ii) imply $T_2 = O_p(\zeta_s \kappa_{o,n})$.

Second, consider $T_3$. Observe that

$$\begin{aligned} T_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\omega_{oi}\exp(\hat{\lambda}'_{s^c}g_{s^ci})h_ig'_{si} - \omega_{oi}h_ig'_{si}\}(\hat{\lambda}_s - \lambda_{os}) \\ &\leq \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\|\{\exp(\hat{\lambda}'_{s^c}g_{s^ci}) - 1\}\omega_{oi}h_ig'_{si}\right\|_\infty\right)\left\|\hat{\lambda}_s - \lambda_{os}\right\|_1 \\ &\leq \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\exp(\hat{\lambda}'_{s^c}g_{s^ci}) - 1|\right)O_p(\zeta_s\kappa_{o,n}), \end{aligned}$$

where the last inequality is due to the definition of $\zeta_s$ and Lemma 3 (ii). By an expansion around $\hat{\lambda}_{s^c} = \lambda_{s^c}$ (note: $\hat{\lambda}_{s^c} = 0$),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\exp(\hat{\lambda}'_{s^c}g_{s^ci}) - 1| = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\|\exp(\tilde{\lambda}'_{s^c}g_{s^ci})g'_{s^ci}\right\|_\infty\left\|\hat{\lambda}_{s^c}\right\|_1 = O_p(\sqrt{n}\tilde{\zeta}_K\kappa_{o,n}),$$

where the first inequality follows from Hölder's inequality, and the second equality follows from the definition of $\tilde{\zeta}_K$ and Lemma 3 (ii). Therefore, $T_3 = O_p(\sqrt{n}\tilde{\zeta}_K\zeta_s\kappa^2_{o,n})$.

Third, we consider $T_4$ and $T_5$. Lemma 3 (ii) implies $T_4 = O_p(\sqrt{n}\zeta_s^2\kappa^2_{o,n})$. Also, the law of large numbers implies $T_5 = O_p(\sqrt{n}\sup_{x\in\mathcal{X}}|\omega_o(x) - \omega(x)|)$.

Finally, consider $T_1$. Denote $Q_n(\lambda) = \mathbb{E}_n[\exp(\lambda' g_i) - \lambda' r]$. Then the estimator is written as $\hat{\lambda} = \arg\min_\lambda Q_n(\lambda) + \alpha_n\left\|\lambda\right\|_1$ and satisfies the first order condition

$$0 = \frac{\partial Q_n(\hat{\lambda})}{\partial\lambda} + \alpha_n\hat{\kappa}.$$

By focusing on the subvector selected by $S_o$,

$$\begin{aligned} \left(\frac{\partial Q_n(\hat{\lambda})}{\partial\lambda}\right)_s &= \frac{1}{n}\sum_{i=1}^{n}\exp(\hat{\lambda}' g_i)g_{si} - r_s \\ &= \frac{1}{n}\sum_{i=1}^{n}\exp(\lambda'_{os}g_{si})\exp(\hat{\lambda}'_{s^c}g_{s^ci})g_{si} - r_s + \frac{1}{n}\sum_{i=1}^{n}\bar{\omega}_i g_{si}g'_{si}(\hat{\lambda}_s - \lambda_{os}), \end{aligned}$$

where $\bar{\omega}_i = \exp(\bar{\lambda}'_s g_{si})\exp(\hat{\lambda}'_{s^c}g_{s^ci})$ and $\bar{\lambda}_s$ is a point on the line joining $\hat{\lambda}_s$ and $\lambda_{os}$. Therefore, by inserting to the first order condition,

$$\hat{\lambda}_s - \lambda_{os} = -\bar{\Sigma}_s^{-1}(\Xi_s + \alpha_n\hat{\kappa}_s), \tag{30}$$

where $\Xi_s = \mathbb{E}_n[\omega_{oi}\exp(\hat{\lambda}'_{s^c}g_{s^ci})g_{si} - r_s]$ and $\bar{\Sigma}_s = \frac{1}{n}\sum_{i=1}^{n}\bar{\omega}_i g_{si}g'_{si}$. Let $\bar{\Upsilon}_2 = \alpha_n\psi_s\bar{\Sigma}_s^{-1}\hat{\kappa}_s$. By using (30), $T_1$ is decomposed as

$$T_1 = -\psi_s\bar{\Sigma}_s^{-1}\sqrt{n}\Xi_s - \sqrt{n}\bar{\Upsilon}_2 + \sqrt{n}(\Upsilon_1 + \Upsilon_2) = T_{11} + T_{12} + T_{13} + T_{14},$$

where

$$
\begin{aligned}
T_{11} &= -\psi_s \Sigma_s^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{oi} \exp(\hat{\lambda}'_{s^c} g_{s^c i}) - \omega_{oi}) g_{si} \right) + \sqrt{n} \Upsilon_1 \\
T_{12} &= -\psi_s \{\bar{\Sigma}_s^{-1} - \Sigma_s^{-1}\} \sqrt{n} \Xi_s - \sqrt{n}(\bar{\Upsilon}_2 + \Upsilon_2), \qquad T_{13} = -\psi_s \Sigma_s^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{oi} - \omega_i) g_{si} \right). \\
T_{14} &= -\psi_s \Sigma_s^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\omega_i g_{si} - \mathbb{E}[\omega_i g_{si}]\} \right)
\end{aligned}
$$

Note that $T_{11} = 0$. By (30), $T_{12}$ is written as

$$
\begin{aligned}
T_{12} &= -\sqrt{n} \psi_s \{\bar{\Sigma}_s^{-1} - \Sigma_s^{-1}\} [-\bar{\Sigma}_s (\hat{\lambda}_s - \lambda_{os}) - \alpha_n \hat{\kappa}_s] - \sqrt{n}(\bar{\Upsilon}_2 + \Upsilon_2) \\
&= \sqrt{n} \psi_s \{\bar{\Sigma}_s^{-1} - \Sigma_s^{-1}\} \bar{\Sigma}_s (\hat{\lambda}_s - \lambda_{os}).
\end{aligned}
$$

Now, since $T_{12}$, $T_{13}$, and $T_{14}$ are all low dimensional, they can be analyzed in the same manner as in the proof of Theorem 2. Therefore, we have

$$
\begin{aligned}
T_{12} &= O_p(\sqrt{n} \zeta_s^4 \kappa_{o,n}^2), \qquad T_{13} = O_p \left( \sqrt{n} \zeta_s^2 \sup_{x \in \mathcal{X}} |\omega_o(x) - \omega(x)| \right), \\
T_{14} &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\omega_i h_i^X - \mathbb{E}[\omega_i h_i^X]\} + o_p(1).
\end{aligned}
$$

Combining these results,

$$
\sqrt{n}(\hat{\theta} - \theta + \Upsilon) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \{\omega_i h_i - \theta\} - \{\omega_i h_i^X - \mathbb{E}[\omega_i h_i^X]\} \right] + O_p(r_n),
$$

where $r_n = \sqrt{n} \zeta_s^4 \kappa_{o,n}^2 + \sqrt{n} \tilde{\zeta}_K \zeta_s \kappa_{o,n}^2 + \sqrt{n} \zeta_s^2 \sup_{x \in \mathcal{X}} |\omega_o(x) - \omega(x)|$. Since $r_n \to 0$ by assumption, the central limit theorem yields the conclusion.

## B.3. **Lemmas.**

**Lemma 3.** Under Conditions D, C, and H, it holds

**(i):** $\Pr \left\{ \frac{1}{2} \mathscr{E}^{\varrho}(\hat{\lambda}) + \alpha_n \left\| \hat{\lambda} - \lambda_o \right\|_1 \leq 4 \mathscr{E}^{\varrho}(\lambda_o) + \frac{32 \alpha_n^2 s}{\phi^2_{S_{\lambda_o}} \varrho} \right\} \geq 1 - \varepsilon,$

**(ii):** $\mathscr{E}^{\varrho}(\hat{\lambda}) = O_p \left( \kappa_{o,n} \sqrt{\frac{\log K}{n}} \right)$ and $\left\| \hat{\lambda} - \lambda_o \right\|_1 = O_p(\kappa_{o,n})$.

**Proof of Lemma 3 (i).** Pick any $\varepsilon > 0$ small enough and $n \in \mathbb{N}$ large enough to satisfy Condition H. Then set $M = \frac{Q_o}{2\sigma_{\varepsilon,n}}$ and take $\bar{\lambda} = t\hat{\lambda} + (1-t)\lambda_o$ with $t = \frac{M}{M + \left\| \hat{\lambda} - \lambda_o \right\|_1}$. Due to the definition of $\hat{\lambda}$ in (6) and convexity of its objective function, we have

$$
\mathbb{E}_n[\exp(\bar{\lambda}' g(X)) - \bar{\lambda}' r] + \alpha_n \left\| \bar{\lambda} \right\|_1 \leq \mathbb{E}_n[\exp(\lambda_o' g(X)) - \lambda_o' r] + \alpha_n \left\| \lambda_o \right\|_1,
$$

and thus

$$\begin{aligned}\mathscr{E}(\bar{\lambda})+\alpha_n\big\|\bar{\lambda}\big\|_1 &\leq -\{\nu_n(\bar{\lambda})-\nu_n(\lambda_o)\}+\mathscr{E}(\lambda_o)+\alpha_n\big\|\lambda_o\big\|_1\\ &\leq \mathscr{E}(\lambda_o)+\alpha_n\big\|\lambda_o\big\|_1+\frac{Q_o}{2},\end{aligned}\tag{31}$$

with probability at least $1-\varepsilon$, where the second inequality follows from Condition H (i) combined with $\big\|\bar{\lambda}-\lambda_o\big\|_1=\frac{M\big\|\hat{\lambda}-\lambda_o\big\|_1}{M+\big\|\hat{\lambda}-\lambda_o\big\|_1}\leq M$. Hereafter all inequalities involving $\bar{\lambda}$ hold true with probability at least $1-\varepsilon$.

Note that $\lambda=\lambda_{S_{\lambda_o}}+\lambda_{S^c_{\lambda_o}}$, and particularly $\lambda_{o,S_{\lambda_o}}=\lambda_o$ and $\lambda_{o,S^c_{\lambda_o}}=0$. Thus, (31) and the triangle inequality imply

$$\begin{aligned}\mathscr{E}(\bar{\lambda})+\alpha_n\big\|\bar{\lambda}_{S^c_{\lambda_o}}\big\|_1 &\leq \mathscr{E}(\lambda_o)+\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1+\frac{Q_o}{2}\\ &\leq Q_o+\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1,\end{aligned}\tag{32}$$

where the second inequality follows from $\mathscr{E}(\lambda_o)\leq\frac{Q_o}{2}$ (due to the definition of $Q_o$). Thus, the triangle inequality yields

$$\mathscr{E}(\bar{\lambda})+\alpha_n\big\|\bar{\lambda}-\lambda_o\big\|_1\leq Q_o+2\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1.\tag{33}$$

In order to bound the right hand side of (33), we consider two cases: (I) $2\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1<Q_o$, and (II) $2\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1\geq Q_o$.

**Case (I)** $2\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1<Q_o$**.**

In this case, (33) and Condition H (iii) imply

$$\mathscr{E}(\bar{\lambda})+\alpha_n\big\|\bar{\lambda}-\lambda_o\big\|_1<2Q_o\leq\frac{\alpha_n M}{2},\tag{34}$$

and thus $\big\|\bar{\lambda}-\lambda_o\big\|_1\leq\frac{M}{2}$.

**Case (II)** $2\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1\geq Q_o$**.**

In this case, (32) and $\lambda_{o,S^c_{\lambda_o}}=0$ guarantees

$$\big\|\bar{\lambda}_{S^c_{\lambda_o}}-\lambda_{o,S^c_{\lambda_o}}\big\|_1=\big\|\bar{\lambda}_{S^c_{\lambda_o}}\big\|_1\leq 3\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_{o,S_{\lambda_o}}\big\|_1\leq\frac{3\sqrt{s}}{\phi_{S_{\lambda_o}}}|\bar{\lambda}-\lambda_o|,\tag{35}$$

where the last inequality follows from Condition C. Observe that

$$\begin{aligned}\mathscr{E}(\bar{\lambda})+\alpha_n\big\|\bar{\lambda}-\lambda_o\big\|_1 &\leq 4\alpha_n\big\|\bar{\lambda}_{S_{\lambda_o}}-\lambda_o\big\|_1\leq\frac{4\alpha_n\sqrt{s}}{\phi_{S_{\lambda_o}}}|\bar{\lambda}-\lambda_o|\\ &\leq \frac{4\alpha_n\sqrt{s}}{\phi_{S_{\lambda_o}}}(|\bar{\lambda}-\lambda_*|+|\lambda_*-\lambda_o|).\end{aligned}$$

where the first inequality follows from (33) and the condition of Case (II), the second inequality follows from (35), and the third inequality follows from triangle inequality.

Now by using $xy \leq x^2 + \frac{y^2}{4}$ for any $x, y \in \mathbb{R}$, we obtain

$$\frac{4\alpha_n\sqrt{s}}{\phi_{S_{\lambda_o}}}|\bar{\lambda} - \lambda_*| \leq \frac{1}{2}\left(\varrho|\bar{\lambda} - \lambda_*|^2 + \frac{16\alpha_n s}{\phi^2_{S_{\lambda_o}}\varrho}\right) \leq \frac{1}{2}\left(\mathscr{E}(\bar{\lambda}) + \frac{16\alpha_n s}{\phi^2_{S_{\lambda_o}}\varrho}\right),$$

where the second inequity follows from Condition H (ii). Similarly,

$$\frac{4\alpha_n\sqrt{s}}{\phi_{S_{\lambda_o}}}|\lambda_o - \lambda_*| \leq \frac{1}{2}\mathscr{E}(\lambda_o) + \frac{8\alpha_n^2 s}{\phi^2_{S_{\lambda_o}}\varrho}.$$

Combining these results with the definition of $Q_o$,

$$\mathscr{E}(\bar{\lambda}) + \alpha_n\left\|\bar{\lambda} - \lambda_o\right\|_1 \leq \frac{1}{2}\mathscr{E}(\bar{\lambda}) + \frac{1}{2}\mathscr{E}(\lambda_o) + \frac{16\alpha_n^2 s}{\phi^2_{S_{\lambda_o}}\varrho} \leq \frac{1}{2}\mathscr{E}(\bar{\lambda}) + Q_o, \tag{36}$$

which implies (by Condition H (ii-c)) $\left\|\bar{\lambda} - \lambda_o\right\|_1 \leq \frac{2\sigma_\varepsilon M}{\alpha_n} \leq \frac{M}{4}$.

Therefore, for both cases, it holds $\left\|\bar{\lambda} - \lambda_o\right\|_1 \leq \frac{M}{2}$ and also $\left\|\hat{\lambda} - \lambda_o\right\|_1 \leq M$, i.e., $\hat{\lambda}$ is close enough to $\lambda_o$ to invoke Condition H (i).

Repeat the proof above by replacing $\bar{\lambda}$ with $\hat{\lambda}$. Then we obtain the counterparts of (34) and (36) with replacements of $\bar{\lambda}$ with $\hat{\lambda}$, i.e.,

$$\frac{1}{2}\mathscr{E}(\hat{\lambda}) + \alpha_n\left\|\hat{\lambda} - \lambda_o\right\|_1 \leq 2Q_o,$$

with probability at least $1 - \varepsilon$. Therefore, the conclusion follows.

**Proof of Lemma 3 (ii).** By setting $\alpha_n \propto \sqrt{\frac{\log K}{n}}$, Part (i) of this lemma implies

$$\frac{1}{2}\mathscr{E}(\hat{\lambda}) + \sqrt{\frac{\log K}{n}}\left\|\hat{\lambda} - \lambda_o\right\|_1 = O_p\left(\mathscr{E}(\lambda_o) \vee \frac{s\log K}{n}\right),$$

and the conclusion follows.

## Appendix C. Tables and figures

Table 1. Cross sectional regression in low dimensional case

| | Const. | $\lambda_{SDF}$ | $\lambda_{RM}$ | $\lambda_{SMB}$ | $\lambda_{HML}$ | Adjusted $R^2$ |
|---|---|---|---|---|---|---|
| Panel A: 25 size and book-to-market | | | | | | |
| SDF: No penalty | 0.649 | -0.257 | | | | 0.844 |
| | (13.977) | (-11.438) | | | | |
| SDF: $\alpha = 0.05$ | 0.720 | -0.124 | | | | 0.625 |
| | (10.146) | (-6.400) | | | | |
| 3 Factors | 1.668 | | -0.751 | 0.204 | 0.437 | 0.714 |
| | (4.401) | | (-2.067) | (3.853) | (6.773) | |
| Panel B: 10 momentum | | | | | | |
| SDF: No penalty | 0.752 | -0.168 | | | | 0.918 |
| | (21.715) | (-10.056) | | | | |
| SDF: $\alpha = 0.05$ | 0.716 | -0.129 | | | | 0.908 |
| | (18.714) | (-9.493) | | | | |
| 3 Factors | 2.365 | | -1.198 | -0.068 | -1.485 | 0.815 |
| | (1.576) | | (-0.754) | (-0.057) | (-1.615) | |
| Panel C: 25 long term reversal and size | | | | | | |
| SDF: No penalty | 0.741 | -0.215 | | | | 0.505 |
| | (8.023) | (-5.049) | | | | |
| SDF: $\alpha = 0.05$ | 0.382 | -0.180 | | | | 0.785 |
| | (4.372) | (-9.416) | | | | |
| 3 Factors | 0.702 | | 0.219 | 0.111 | 0.633 | 0.754 |
| | (2.541) | | (0.833) | (1.678) | (5.051) | |

Note: Cross sectional regression results in the low dimensional case. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1963 to December 2010, using portfolios in each corresponding panel. Panel A presents results using 25 size and book-to-market portfolios, Panel B presents results using 20 momentum portfolios, and Panel C is concerned with results using 25 long term reversal and size portfolios. 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. In each panel the first row is about the estimated SDF when no penalty is imposed, the second row is the estimated SDF when penalty level is at 0.05, and the third row is the seminal Fama-French three factor models.

TABLE 2. Cross sectional regression in intermediate case

| | Const. | $\lambda_{SDF}$ | $\lambda_{RM}$ | $\lambda_{SMB}$ | $\lambda_{HML}$ | Adjusted $R^2$ |
|---|---|---|---|---|---|---|
| Panel A: 100 size and book-to-market | | | | | | |
| SDF: No penalty | 1.033 | -0.926 | | | | 0.581 |
| | (52.744) | (-11.532) | | | | |
| SDF: $\alpha = 0.1$ | 0.725 | -0.273 | | | | 0.652 |
| | (20.435) | (-13.367) | | | | |
| 3 Factors | 1.575 | | -0.639 | 0.190 | 0.439 | 0.627 |
| | (8.618) | | (-3.670) | (5.577) | (11.175) | |
| Panel B: 49 industry | | | | | | |
| SDF: No penalty | 0.800 | -0.129 | | | | 0.329 |
| | (16.239) | (-4.852) | | | | |
| SDF: $\alpha = 0.1$ | 0.686 | -0.065 | | | | 0.294 |
| | (0.686) | (-0.065) | | | | |
| 3 Factors | 1.064 | | -0.008 | -0.096 | -0.109 | -0.002 |
| | (6.229) | | (-0.047) | (-0.923) | (-1.151) | |
| Panel C: 25 long term reversal+25 short term reversal+25 momentum | | | | | | |
| SDF: No penalty | 1.083 | -1.919 | | | | 0.605 |
| | (48.960) | (-10.698) | | | | |
| SDF: $\alpha = 0.1$ | 1.130 | -0.484 | | | | 0.441 |
| | (43.162) | (-7.705) | | | | |
| 3 Factors | 1.416 | | -0.432 | 0.293 | 0.012 | 0.153 |
| | (4.489) | | (-1.454) | (3.370) | (0.064) | |

Note: Cross-sectional regression results in the intermediate case. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1963 to December 2010, using portfolios in each corresponding panel. Panel A presents results using 100 size and book-to-market portfolios, Panel B presents results using 49 industry portfolios, and Panel C presents results using 75 portfolios listed in the beginning of the panel. 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. In each panel the first row is about the estimated SDF when no penalty is imposed, the second row is the estimated SDF when penalty level is at 0.1, and the third row is the seminal Fama-French three factor models.

FIGURE 1. Summary of cross sectional regression against different penalty levels in high dimension case ($K = 300$ or 425; $T = 360$)

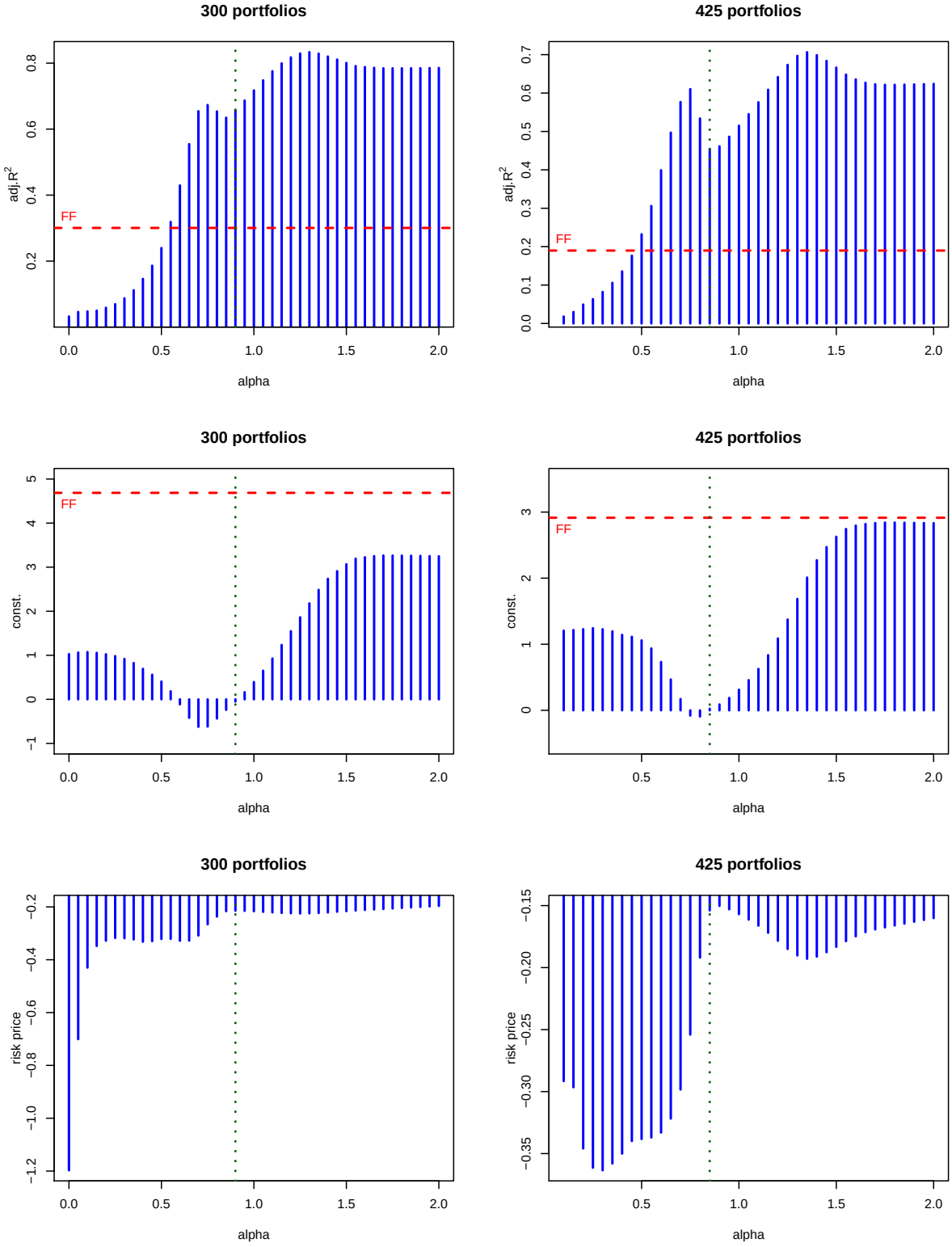

FIGURE 2. Number of active portfolios selected under 300 portfolios case

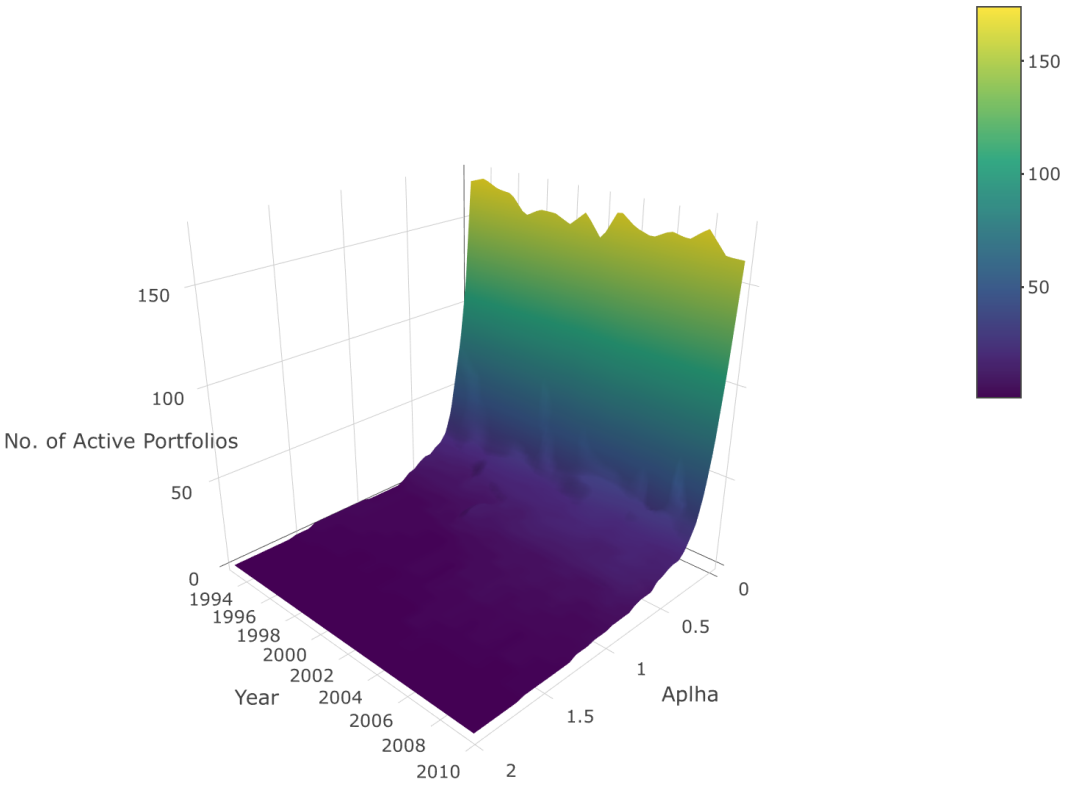

FIGURE 3. Number of active portfolios selected under 425 portfolios case

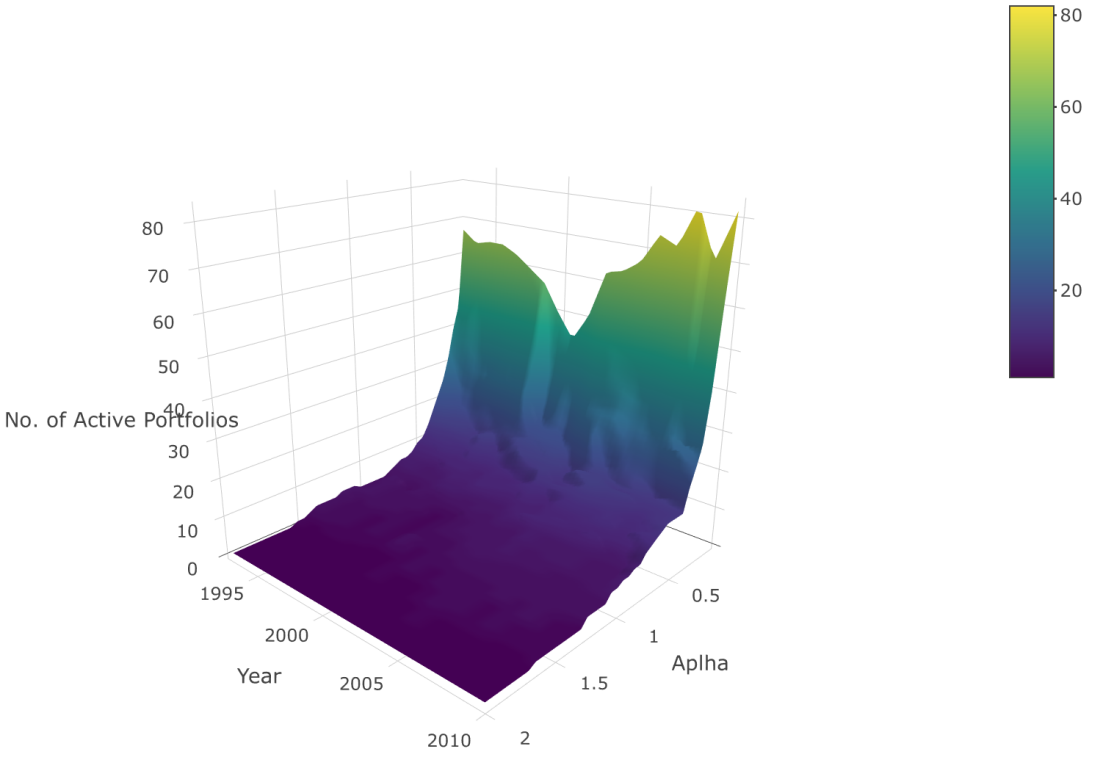

TABLE 3. Cross sectional regression in high dimensional case

| | Const. | $\lambda_{SDF}$ | $\lambda_{RM}$ | $\lambda_{SMB}$ | $\lambda_{HML}$ | Adjusted $R^2$ |
|---|---|---|---|---|---|---|
| Panel A: 300 portfolios<br>100 size & book-to-market+100 size & operating profitability+100 size & investment | | | | | | |
| SDF: $\alpha = 0.1$ | 1.027<br>(14.062) | -1.197<br>(-3.306) | | | | 0.032 |
| SDF: $\alpha = 0.9$ | -0.050<br>(-0.851) | -0.214<br>(-24.017) | | | | 0.658 |
| 3 Factors | 4.687<br>(10.986) | | -3.891<br>(-9.998) | 0.699<br>(5.295) | -0.517<br>(-2.900) | 0.301 |
| Panel B: 425 portfolios<br>300 in Panel A+49 industry+25 long term rev.+25 short term rev.+25 momentum | | | | | | |
| SDF: $\alpha = 0.1$ | 1.206<br>(12.684) | -0.292<br>(-2.967) | | | | 0.018 |
| SDF: $\alpha = 0.85$ | 0.024<br>(0.383) | -0.154<br>(-18.800) | | | | 0.455 |
| 3 Factors | 2.914<br>(10.507) | | -2.121<br>(-8.339) | 0.659<br>(6.331) | -0.305<br>(-2.205) | 0.190 |

Note: Cross-sectional regression results in the high dimensional case. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1993 to December 2010, using portfolios in each corresponding panel. Panel A presents results using 300 portfolios, and Panel B presents results using 425 portfolios. 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. In each panel the first row is about the estimated SDF when the penalty level is set at 0.1, the second row is the estimated SDF when penalty level is at 0.9 and 0.85, respectively, and the third row is the seminal Fama-French three factor models.

FIGURE 4. Time series plot of estimated SDF in high dimensional case: July 1993 - December 2010.
Grey shaded area represents NBER recessions

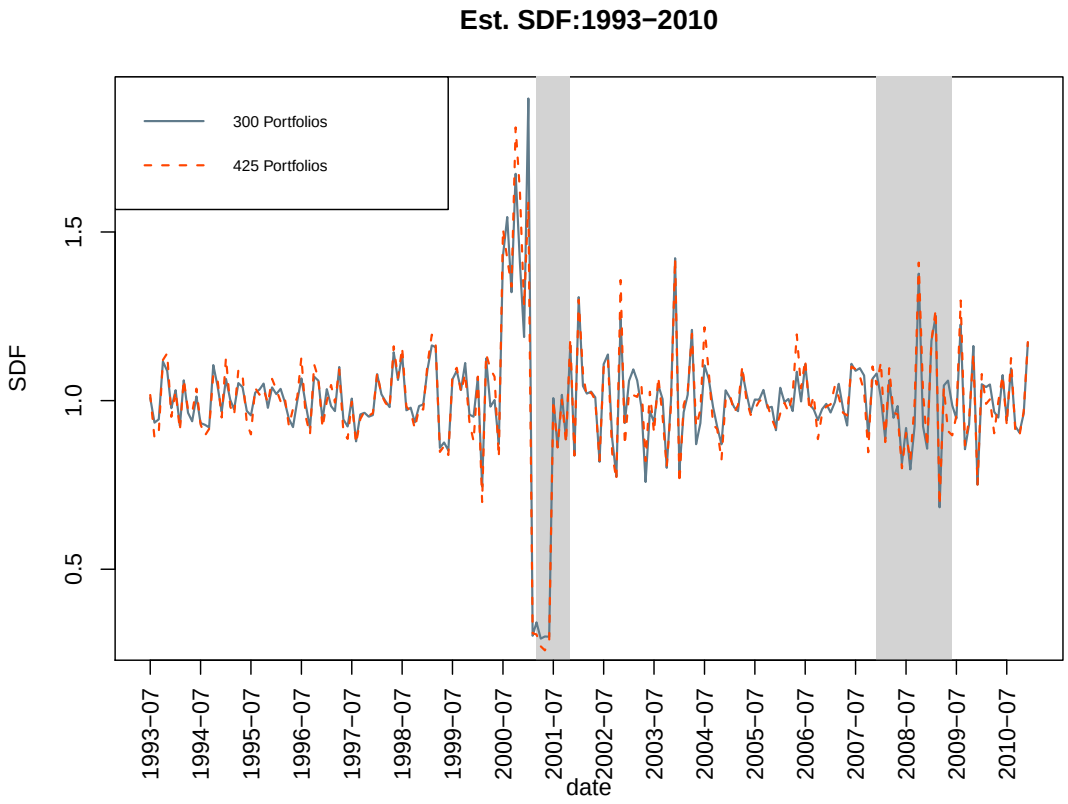

TABLE 4. Time series properties of estimated SDF from high dimensional case

| $\alpha$ | $\beta_{RM}$ | $\beta_{SMB}$ | $\beta_{HML}$ | $\beta_{MOM}$ | Adjusted $R^2$ |
|---|---|---|---|---|---|
| Panel A: 300 portfolios, $\alpha = 0.9$ | | | | | |
| 1.011 | -0.004 | -0.014 | -0.007 | -0.007 | 0.118 |
| (85.846) | (-1.427) | (-4.106) | (-1.851) | (-3.264) | |
| Panel B: 425 Portfolios, $\alpha = 0.85$ | | | | | |
| 1.012 | -0.006 | -0.016 | -0.003 | -0.008 | 0.171 |
| (84.730) | (-2.340) | (-4.712) | (-0.879) | (-3.594) | |

Note: Time series regression of estimated SDF extracted from the high dimensional case 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, using portfolios and penalty level in each corresponding panel. Panel A presents results using 300 portfolios and when penalty level is 0.9, and Panel B presents results using 425 portfolios and when penalty level is set at 0.85. The first column is the estimated constant (or, "alpha") 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


[1] Athey, S., Imbens, G. W. and S. Wager (2016) Approximate residual balancing: de-biased inference of average treatment effects in high dimensions, Working paper.

[2] Bickel, P. J., Ritov, Y. and A. B. Tsybakov (2009) Simultaneous analysis of lasso and dantzig selector, *Annals of Statistics*, 37, 1705-1732.

[3] Bühlmann, P. and S. van de Geer (2011) *Statistics for High-Dimensional Data*, Springer.

[4] Chan, K. C. G., Yam, S. C. P. 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.

[5] Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models, in Heckman, J. and E. Leamer (eds.), *Handbook of Econometrics*, vol. 6B, ch. 76, Elsevier.

[6] Christensen, T. M. (2016) Nonparametric stochastic discount factor decomposition, Working paper.

[7] Cochrane, J. H. (2009) *Asset Pricing*, revised ed., Princeton University Press.

[8] Csiszár, I. (1975) $I$-divergence geometry of probability distributions and minimization problems, *Annals of Probability*, 3, 146-158.

[9] 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.

[10] Fama, E. F. and J. D. MacBeth (1973) Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy*, 81, 607-636.

[11] 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.

[12] Ghosh, A., Julliard, C. and A. P. Taylor (2016) An information-theoretic asset pricing model, Working paper.

[13] Ghosh, A., Julliard, C. 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.

[14] Hall, A. R. (2004) *Generalized Method of Moments*, Oxford University Press.

[15] Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica*, 66, 315-331.

[16] Hjort, N. L., I. W. McKeague and I. van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079-1111.

[17] Imbens, G. W. and D. B. Rubin (2015) *Causal Inference*, Cambridge University Press.

[18] Imbens, G. W., Spady, R. H. and P. Johnson (1998) Information theoretic approaches to inference in moment condition models, *Econometrica*, 66, 333-357.

[19] Kitamura, Y. (2007) Empirical likelihood methods in econometrics : theory and practice, in Blundell, R., Newey, W. K. and T. Persson (eds.), *Advances in Economics and Econometrics*, vol. III, Cambridge University Press.

[20] Kitamura, Y. and M. Stutzer (1997) An information-theoretic alternative to generalized method of moments estimation, *Econometrica*, 65, 861-874.

[21] Kitamura, Y. and M. Stutzer (2002) Connections between entropic and linear projections in asset pricing estimation, *Journal of Econometrics*, 107, 159-174.

[22] Lahiri, S. and S. Mukhopadhyay (2012) A penalized empirical likelihood method in high dimensions, *Annals of Statistics*, 40, 2511-2540.

[23] Lewellen, J., Nagel, S. and J. Shanken (2010) A skeptical appraisal of asset pricing tests, *Journal of Financial Economics*, 96, 175-194.

[24] Little, R. J. A. and D. B. Rubin (2002) *Statistical Analysis with Missing Data*, Wiley.

[25] Lorentz, G. G. (1986) *Approximations of Functions*, Chelsea.

[26] Newey, W. K. (1994) The asymptotic variance of semiparametric estimators, *Econometrica*, 62, 1349-1382.

[27] Newey, W. K. (1997) Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, 79, 147-168.

[28] Newey, W. K. and D. McFadden (1994) Large sample estimation and hypothesis testing, in Engle, R. F. and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, ch. 36, Elsevier.

[29] Newey, W. K. and R. J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica*, 72, 219-255.

[30] Owen, A. B. (2001) *Empirical Likelihood*, CRC press.

[31] 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.

[32] Schumaker, L. L. (1981) *Spline Functions: Basic Theory*, Wiley.

[33] Tang, C. Y. and C. Leng (2010) Penalized high-dimensional empirical likelihood, *Biometrika*, 97, 905-920.

[34] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, B 58, 267-288.

[35] Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*, Springer.

[36] van de Geer, S. A. (2008) High-dimensional generalized linear models and the lasso, *Annals of Statistics*, 36, 614-645.

[37] van de Geer, S., Bühlmann, P., Ritov, Y. and R. Dezeure (2014) On asymptotically optimal confidence regions and tests for high-dimensional models, *Annals of Statistics*, 42, 1166-1202.

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

[39] 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.

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

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

*E-mail address*: c.qiu@lse.ac.uk

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

*E-mail address*: t.otsu@lse.ac.uk