UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 2004

TESTING STOCHASTIC DOMINANCE WITH MANY CONDITIONING VARIABLES

Oliver Linton | Myung Hwan Seo | Yoon-Jae Whang

10 February 2020

We propose a test of the hypothesis of conditional stochastic dominance in the presence of many conditioning variables (whose dimension may grow to infinity as the sample size diverges). Our approach builds on a semiparametric location scale model in the sense that the conditional distribution of the outcome given the covariates is characterized by a nonparametric mean function and a nonparametric skedastic function with an independent innovation whose distribution is unknown. We propose to estimate the nonparametric mean and skedastic regression functions by the `1-penalized nonparametric series estimation with thresholding. Under the sparsity assumption, where the number of truly relevant series terms are relatively small (but their identities are unknown), we develop the estimation error bounds for the regression functions and series coefficients estimates allowing for the time series dependence. We derive the asymptotic distribution of the test statistic, which is not pivotal asymptotically, and introduce the smooth stationary bootstrap to approximate its sample distribution. We investigate the finite sample performance of the bootstrap critical values by a set of Monte Carlo simulations. Finally, our method is illustrated by an application to stochastic dominance among portfolio returns given all the past information.

Testing Stochastic Dominance with Many Conditioning Variables*

Oliver Linton† Myung Hwan Seo‡ Yoon-Jae Whang§

February 10, 2020

Abstract

We propose a test of the hypothesis of conditional stochastic dominance in the presence of many conditioning variables (whose dimension may grow to infinity as the sample size diverges). Our approach builds on a semiparametric location scale model in the sense that the conditional distribution of the outcome given the covariates is characterized by a nonparametric mean function and a nonparametric skedastic function with an independent innovation whose distribution is unknown. We propose to estimate the nonparametric mean and skedastic regression functions by the ℓ_1-penalized nonparametric series estimation with thresholding. Under the sparsity assumption, where the number of truly relevant series terms are relatively small (but their identities are unknown), we develop the estimation error bounds for the regression functions and series coefficients estimates allowing for the time series dependence. We derive the asymptotic distribution of the test statistic, which is not pivotal asymptotically, and introduce the smooth stationary bootstrap to approximate its sample distribution. We investigate the finite sample performance of the bootstrap critical values by a set of Monte Carlo simulations. Finally, our method is illustrated by an application to stochastic dominance among portfolio returns given all the past information.

KEYWORDS: Bootstrap; Empirical process; Home bias; LASSO; Power boosting; Sparsity

JEL: C10, C12,C15, C15

*Thanks to Yookyung Lee, Suyeol Kim, Seok Young Hong, and Zhenkai Ran for help with the numerical work.

†Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD. Email: obl20@cam.ac.uk.Thanks to the Cambridge INET for financial support.

‡Seoul National University. Email: myunghseo@snu.ac.kr. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A01033487).

§Seoul National University. Email: whang@snu.ac.kr. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A5A2A01024030).

1 Introduction

The main purpose of this paper is to provide methodology to test stochastic dominance hypotheses in the presence of conditioning information. The first order dominance hypothesis (FSD) we consider is

$$\mathcal{H}_0 : \Pr\left(y_{tj} \leq y | \mathcal{F}_{t-1}\right) \leq \Pr\left(y_{tl} \leq y | \mathcal{F}_{t-1}\right) \tag{1}$$

almost surely for all $y \in \mathbb{R}$, where y_{tj} and y_{tl} denote the outcomes of interest to compare, and $\mathcal{F}_{t-1}$ is information observed at time $t-1$. If this hypothesis holds, then we may conclude that y_{tj} should be preferred to y_{tl} by a large class of individuals who prefer more to less, Levy (2016). The alternative hypothesis we test against is the negation of $\mathcal{H}_0$.

An additional hypothesis of interest is the second order stochastic dominance (SSD) hypothesis that

$$\int_{-\infty}^{x} \Pr\left(y_{tj} \leq y | \mathcal{F}_{t-1}\right) dy \leq \int_{-\infty}^{x} \Pr\left(y_{tl} \leq y | \mathcal{F}_{t-1}\right) dy \tag{2}$$

almost surely for all $x \in \mathbb{R}$. In this case we may conclude that series y_{tj} should be preferred to y_{tl} by a large class of risk averse individuals. Stochastic dominance is central in a number of decision-making contexts, Levy (2016) and Whang (2019).

Allowing for conditioning information is essential in many applied contexts. Gonzalo and Olmo (2014) developed some tests of conditional stochastic dominance for the low dimensional case. In fact, there is nowadays a plethora of available data for both decision-makers and econometricians. We allow the dimensionality of $\mathcal{F}_{t-1}$ to be large and we allow a general functional form for the conditional c.d.f.'s of the outcome variables. In particular, we propose a location-scale model for the observed outcomes with i.i.d. shocks of unknown distribution. The proposed model is semiparametric in the sense that the location and scale functions can be fully nonparametric functions of a finite number of conditioning covariates or can be linear functions of a large number of conditioning covariates, while the unknown error distribution is also left unspecified. The sparsity assumption plays a fundamental role in high-dimensional data analysis. Under this assumption, the information on the large set of conditioning variables can be effectively represented by a small subset of variables, although their identities are unknown to researchers. This plays a key role in determining how to construct a test of the conditional stochastic dominance hypothesis under the high-dimensional setup. In our application below we consider one series to be the return on the US stock market and the other series to be the return on the Rest of the World or Global stock market. We consider a large class of available information $\mathcal{F}_t$. We use modern data techniques to reduce the dimensionality of $\mathcal{F}_t$ to the most essential components, whose dimensionality, nevertheless, is allowed to grow with sample size.

We estimate the unknown location and scale functions by the regularized least squares and the error distribution by the empirical distribution. Specifically, the unknown location function is estimated

by Tibshirani's (1996) least absolute shrinkage and selection operator (LASSO). It is commonly employed for the sparse high-dimensional regression and its statistical properties have been intensely studied for cross sectional data, see e.g. Bickel et al. (2009), but not much work has been published for time series data, Medeiros and Mendes (2016) being an exception. The scale function is also estimated by the LASSO in the regression of the absolute value of the first step residuals. Unlike in other skedastic regressions, we find that the scale normalization by the modulus is more convenient than by the square of the residuals. Then, the error distribution is estimated by the empirical distribution function of the rescaled residuals. To characterize the sharp weak limit of the empirical distribution, however, we need to regularize the residuals by thresholding so that we can control the random variation arising from the imperfect selection of the smallish coefficients. Therefore, we reestimate the location and scale functions by the ordinary least squares with the selected variables by the thresholding. The precise conditions on the thresholding are presented as well as the additional regularity conditions to validate the thresholding.

We develop a time series extension of an exponential inequality that proves useful to obtain deviation bounds for LASSO estimators, which in turn yields tightness of the residual empirical process. The deviation bounds for the LASSO and variants have been developed for random samples by Bickel et al. (2009), Belloni et al. (2012), among many others. See also reviews by Bühlman and van der Geer (2011) and Belloni and Chernozhukov (2013) for instance. They justify the selection consistency based on thresholding the LASSO estimates, provided that so-called beta-min conditions are met. The weak convergence of an empirical process of residuals from a linear regression with increasing numbers of regressors has been studied by many others, e.g. Mammen (1996) and Chen and Lockhart (2001). More recently by Chatterjee et al. (2015) who considered a high-dimensional penalized regression with homoskedastic errors. Our work extends the literature by allowing for dependent data and for the rescaled residuals from a nonparametric location scale regression model. Building on the aforementioned generalizations, we obtain the weak convergence of estimates of the conditional distribution functions and provide the asymptotic distribution of a supremum statistic for the conditional stochastic dominance test.

Our statistic for FSD is the maximum deviation of one estimated conditional distribution function from the other. The conditional distribution functions in our specification are defined on unbounded dimensions because a conditional distribution function is given by the composite of the distribution function and the regression functions, whose domains may belong to an infinite-dimensional space. This is a different feature of our test from the previous finite-dimensional stochastic dominance tests. We establish the weak convergence of the properly centered and scaled estimated conditional distribution functions. The statistic for SSD is the maximum deviation between the unweighted integrals of the estimated conditional distribution functions. The unweighted integral of an empirical

distribution function does not exhibit weak convergence but the difference can be shown to be tight. Then, the statistic converges by the continuous mapping theorem to the supremum of a Gaussian process, which is not pivotal and cannot be tabulated without knowledge of the underlying data distribution.

We propose a smoothed stationary bootstrap to compute the p-value of the statistic. Several issues need to be taken care of to implement the bootstrap correctly due to the high-dimensional time series data and the use of the empirical distribution function. The stationary bootstrap proposed by Politis and Romano (1994) is suitable to control the serial dependence in the observed variables and errors. We do not need to impose the martingale difference condition for the errors. The smoothing is used to impose the smoothness of the error distribution on the bootstrap distribution as in e.g. Neumeyer (2009). Both methods are combined to generate bootstrap samples. Furthermore, the location and scale functions for the bootstrap sample are estimated by the ordinary least squares conditioning on the selection at the original sample, since bootstrapping the LASSO is known to be inconsistent, Chatterjee and Lahiri (2011). Thus, the computational burden is lowered by not implementing the LASSO at the bootstrap.

We investigate the finite sample performance of the bootstrap critical values in some Monte Carlo simulations. We find that the performance is satisfactory even for quite modest sample sizes. Finally, we illustrate our method by an application to the home bias puzzle, which is why investors are so overweighted on the US market. In particular we test whether a Global return series dominates the return on the US market conditioning on a large set of variables available to investors. We find that generally the US market does not dominate the rest of the world when conditioning on available variables but there are some periods where it does.

The paper is organized as follows. The following section introduces our model and the preliminary estimators. The test statistics for FSD and SSD are defined and their asymptotic distributions are derived in Section 3. The smooth stationary bootstrap is introduced in Section 4 and its finite sample property is explored through Monte Carlo simulation in Section 5. A few proposals to improve the power are discussed in Section 4.2. Section 6 illustrates the new test by an application to the home bias puzzle. All the proofs are delegated to the Appendix, where we establish the weak convergence of the residual empirical processes from the (moderately) high-dimensional time series regression and the (conditional) weak convergence of the aforementioned smooth stationary bootstrap residual empirical process. Each of the results in these sections may be of independent interest and thus the results are presented self-contained. For these two sections, we consider a model of moderately high-dimension in the sense that the number of parameters increases but not as fast as to make the OLS infeasible.

Notation: for a vector $x \in \mathbb{R}^s$, $|x|_q := (\sum_{i=1}^n |x_i|^q)^{1/q}$ denotes the ℓ_q-norm of the vector x and

$|x| := |x|_2$. And for an index set S, $x_S := (x_i : i \in S)$ denotes a subvector of x. Let $S(\beta)$ denote the support of a given vector β, that is, $S(\beta) = \{j : \beta_j \neq 0\}$, and let $|S|$ denote the cardinality of an index set S. All the observations are arrays of variables, whose distributions and dimensions may depend on n, but we suppress the dependence on n and do not introduce the subscript n for notational simplicity, unless it is necessary to evade confusion. In particular, X_t denotes the vector of p basis transformations of q_t, while x_t is a subvector of X_t that collects truly relevant elements among X_t. But we use β to indicate the coefficients of both X_t and x_t to ease notation. Thus, if $\widehat{\beta}$ is defined as the OLS estimate in the regression of y_t on x_t, then $X_t^{\intercal}\widehat{\beta}$ should be understood as $X_{t,S}^{\intercal}\widehat{\beta}$.

2 Model

We suppose that the following location scale model generates each outcome variable:

$$y_{tj} = g^j(q_t) + \sigma^j(q_t)\,\varepsilon_{tj}, \quad t = 1, \ldots, n, \tag{3}$$

where the innovation ε_{tj} is an i.i.d. sequence with a marginal distribution F^j and first moment equal to zero. The observed covariate vector $q_t \in \mathbb{R}^k$ may include lagged outcome variables, and its dimension can be large (increases to infinity as n increases to infinity). The functions $g^j, \sigma^j : \mathbb{R}^k \to \mathbb{R}$ are unknown but we specify below some restrictions on them. We suppose that the skedastic functions σ^j are bounded away from zero and ε_{tj} is non-degenerate so that the mean regression error $e_t = \sigma_t \varepsilon_t$ is nonzero with probability one. We assume that ε_{tj} and q_t are mutually independent in which case the conditional distribution of y_{tj} given q_t is characterized by the distribution of ε_{tj} and the functions g^j, σ^j. From here on, we omit the outcome index j when we discuss common features. We will add the superscripts when it is necessary to avoid confusion. Under our assumptions, for a given y and q, the conditional distribution of the typical outcome variable is

$$F(y|q) := \Pr(y_t \leq y | q_t = q) = F\left(\frac{y - g(q)}{\sigma(q)}\right).$$

The c.d.f.'s F^j are of unknown functional form and so are the functions g^j, σ^j. However, we further suppose that there exists a large dimensional observed vector $X_t := (X_{t1}, \ldots, X_{tp})^{\intercal} :=$
$(X_1(q_t), \ldots, X_p(q_t))^{\intercal} = X(q_t) \in \mathbb{R}^p$ such that

$$g(q_t) = \beta_0^{\intercal} X_t + r_{gt}, \tag{4}$$

where β_0 is sparse (contains many zeros essentially) and the approximation error $r_{gt} \to 0$ at a proper rate as the dimensionality of X_t expands with sample size. We do not need to know *a priori* the exact identity of X_t, it suffices to have a superset of it. One prominent model is that the regression

function is partially linear $g(q_t) = g_1(q_{1t}) + \alpha_g^\intercal q_{2t}$, where q_{2t} is a vector of large dimensions, while q_{1t} is of small, fixed dimensions, and g_1 is of unknown form so that $g_1(q_1) = \sum_{l=1}^{\infty} \theta_l \psi_l(q_1)$, where ψ_l are known basis functions and θ_l are unknown parameters, in which case we can take X_t to contain $q_{2t}, \psi_1(q_{1t}), \ldots, \psi_m(q_{1t})$ for some finite truncation m. The key question in practice is how to assign the elements of q to q_1 or q_2, but our method and theory do not require us to take a position on this: we can include a superset where we expand all continuous variables and the selection will choose which ones essentially require nonlinear treatment.

An analogous assumption is imposed on the skedastic function σ so that

$$\sigma_t = \sigma(q_t) = \gamma_0^\intercal X_t + r_{\sigma t}, \tag{5}$$

where γ_0 is a sparse vector and the approximation error $r_{\sigma t} \to 0$ as the dimension of X_t grows. The vector X_t for the function σ may be different from X_t in the specification of g in (4) but we let them be the same for notational simplicity. It may be viewed as the union of the two if they are different.

We allow σ to determine the scale of the overall error term and impose the scale normalization on the distribution of ε_t, specifically we suppose that $E|\varepsilon_t| = 1$.[1]

2.1 Estimation

The dimension p of the series terms X_t can be potentially huge because the dimension of q_t itself may be high or the number of the basis transformations that are employed may be large. The series can be composed of various interactions, their dummies, polynomials, and B-splines and even interactions of the basic series transformations. Thus, it is generally not feasible to estimate β by unrestricted least squares; we estimate the coefficients β by a regularized least squares. Our procedure takes three steps. We first employ a weighted LASSO procedure, and then select elements of X_t by thresholding, and finally reestimate β by the OLS on the selected variables.

First, we describe the weighted LASSO, i.e., the ℓ_1-penalized least squares with penalty varying for each element in β, which is

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{t=1}^{n} \left| y_t - X_t^\intercal \beta \right|_2^2 + \lambda \left| D\beta \right|_1, \tag{6}$$

where D is a diagonal weighting matrix and λ is a tuning parameter. For instance, in the standard scale normalization D's j-th diagonal element is $(n^{-1}\sum_{t=1}^{n} X_{tj}^2)^{1/2}$. Song and Bickel (2011) proposed a lag-order dependent penalty k or k^2 when $X_{tj} = y_{t-k}$.

[1] Since $E|e_t|^i = E|\sigma_t|^i E|\varepsilon_t|^i$, it does not matter which moment of $|\varepsilon_t|$ we choose to normalize. In the case where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $E|\varepsilon_t| = \sigma_\varepsilon\sqrt{2/\pi}$, i.e., we would normalize $\sigma_\varepsilon = \sqrt{\pi/2}$ instead of 1. Note that the function σ_t is still viewed as the conditional variance of the error e_t given q_t upto a scale adjustment, since $E(e_t^2|\mathcal{F}_{t-1}) = \sigma_t^2 \cdot E\varepsilon_t^2$. While $i = 2$ is more common in the GARCH literature, we let $i = 1$ as it is more convenient for our test procedure later on.

Let $\widehat{\beta}_{lasso}$ denote the resulting LASSO estimate. Second, we threshold this estimator further. Let

$$\widehat{S} = \left\{ j : \left| \widehat{\beta}_{lasso,j} \right| > \lambda_{thr} \right\} \tag{7}$$

where the threshold parameter λ_{thr} is strictly bigger than the LASSO parameter λ.

Third, we re-estimate β by the OLS method on the selected variables defined by $\widehat{S}$. Specifically,

$$\widehat{\beta}_{Tasso} = \arg \min_{\beta : \beta_j = 0, j \notin \widehat{S}} \sum_{t=1}^{n} \left(y_t - X_t^{\intercal} \beta \right)^2, \tag{8}$$

that is, $\widehat{\beta}_{Tasso,\widehat{S}}$ is equivalent to the OLS estimator[2] $\widehat{\beta}$ from the linear regression of y_t on $\widehat{x}_t := X_{t,\widehat{S}}$. Then, we set

$$\widehat{g}_t = \widehat{g}\left(q_t\right) = X_t^{\intercal} \widehat{\beta}_{Tasso} = \widehat{x}_t^{\intercal} \widehat{\beta},$$

and define the residual $\widehat{e}_t = y_t - \widehat{g}_t$ for each t.

To estimate γ, note that under our scale normalization we have $|e_t| = \sigma\left(q_t\right) + \eta_t = \gamma_0^{\intercal} X_t + r_{\sigma t} + \eta_t$, where $E\left[\eta_t | q_t\right] = 0$. We replace the unobserved e_t by the residual $\widehat{e}_t$ and proceed as for the estimation of β. Specifically, let

$$\widehat{\gamma}_{lasso} := \underset{\gamma \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{t=1}^{n} \left(|\widehat{e}_t| - X_t^{\intercal} \gamma \right)^2 + \mu \left| Q \gamma \right|_1, \tag{9}$$

where Q is a weight matrix such as the diagonal matrix whose j-th element is $(n^{-1} \sum_{t=1}^{n} w_{tj}^2)^{1/2}$ and μ is a penalty parameter. Then, apply the thresholding to determine a selection

$$\widehat{S}_{\gamma} = \left\{ j : |\widehat{\gamma}_{lasso,j}| \geq \mu_{thr} \right\},$$

for some μ_{thr}. Next, let $\widehat{w}_t = X_{t,\widehat{S}_{\gamma}}$ and $\widehat{\gamma}$ denote the OLS estimate of $|\widehat{e}_t|$ on $\widehat{w}_t$. We may also define $\widehat{\gamma}_{Tasso}$ as the thresholded LASSO estimate as for $\widehat{\beta}_{Tasso}$. Then, we may set

$$\widehat{\sigma}_t = \widehat{\sigma}\left(q_t\right) = X_t^{\intercal} \widehat{\gamma}_{Tasso} = \widehat{w}_t^{\intercal} \widehat{\gamma},$$

provided that $w_t^{\intercal} \widehat{\gamma}_{Tasso} > 0$. In the event that $w_t^{\intercal} \widehat{\gamma}_{Tasso} \leq 0$, which happens with low probability, we set $\widehat{\sigma}_t = \sum_{t=1}^{n} |\widehat{e}_t| / n$.

3 Test Statistics

This section introduces our test statistics for the FSD and SSD hypotheses and then develops their asymptotic distributions.

We observe the dataset $\{y_{t1}, y_{t2}, q_t\}_{t=1}^{n}$. The testing proceeds as follows.

[2]There is some abuse of notation in defining $\widehat{\beta}_{Tasso}$ and $\widehat{\beta}$ as the two have different dimensions. That is, $\widehat{\beta}_{Tasso}$ denotes the estimate of the coefficient of X_t while $\widehat{\beta}$ corresponds to that of $\widehat{x}_t$, which is a subset of X_t due to the selection. We keep this notation as it simplifies the notation without much confusion.

1. For each $j = 1, 2$, run the regression of y_{tj} on X_t by the thresholded LASSO as described in Section 2.1 to get $\widehat{S}^j$.

2. Let $\widehat{S} = \widehat{S}^1 \cup \widehat{S}^2$ and define $\widehat{x}_t := X_{t,\widehat{S}}$, i.e. the collection of selected elements of X_t in at least one of the two regressions, and $\widehat{x}(q) = X_{\widehat{S}}(q)$.

3. Let $\widehat{\beta}^j$ denote the OLS estimate[3] in the regression of y_{tj} on $\widehat{x}_t$ and define the residual $\widehat{e}_{tj} = y_{tj} - \widehat{x}_t^{\intercal}\widehat{\beta}^j$.

4. Likewise, for each $j = 1, 2$, run the skedastic regression of $|\widehat{e}_{tj}|$ on X_t as explained in Section 2.1 and compute $\widehat{S}_\sigma$, $\widehat{w}_t$, $\widehat{w}(q)$, and $\widehat{\gamma}^j$ analogously to $\widehat{S}$, $\widehat{x}_t$, $\widehat{x}(q)$, and $\widehat{\beta}^j$ in the preceding steps, respectively.

5. For each $j = 1, 2$, construct the scaled residual $\widehat{\varepsilon}_{tj} = (y_{tj} - \widehat{x}_t^{\intercal}\widehat{\beta}^j)/\widehat{\sigma}_t^j$, its empirical distribution function
$$\widehat{F}^j(\tau) = \frac{1}{n}\sum_{t=1}^{n} 1\{\widehat{\varepsilon}_{tj} \leq \tau\},$$
and let $\widehat{\tau}^j(y, q) = (y - \widehat{x}(q)^{\intercal}\widehat{\beta}^j)/\widehat{\sigma}^j(q)$.

6. Construct the test statistic[4] for the FSD hypothesis
$$T_n = \sqrt{n}\sup_{y,q}\left(\widehat{F}^1\left(\widehat{\tau}^1(y,q)\right) - \widehat{F}^2\left(\widehat{\tau}^2(y,q)\right)\right),$$
and, for the SSD hypothesis,
$$\begin{aligned} U_n &= \sqrt{n}\sup_{y,q}\int_{-\infty}^{y}\widehat{F}^1\left(\widehat{\tau}^1(u,q)\right)du - \int_{-\infty}^{y}\widehat{F}^2\left(\widehat{\tau}^2(u,q)\right)du \\ &= \sup_{y,q}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\widehat{\sigma}^1(q)\left(\widehat{\tau}^1(y,q) - \widehat{\varepsilon}_{t1}\right)_+ - \widehat{\sigma}^2(q)\left(\widehat{\tau}^2(y,q) - \widehat{\varepsilon}_{t2}\right)_+\right]. \end{aligned}$$

We may also by analogy construct tests for stochastic maximality in the case with multiple prospects, McFadden (1989).

[3]There is some abuse of notation in defining $\widehat{\beta}_{Tasso}$ and $\widehat{\beta}$ as the two have different dimensions. That is, $\widehat{\beta}_{Tasso}$ denotes the estimate of the coefficient of X_t while $\widehat{\beta}$ corresponds to that of $\widehat{x}_t$, which is a subset of X_t due to the selection. We keep this notation as it simplifies the notation without much confusion.

[4]In practice, the supremums are approximated by the maximums over some grid of (y, q). This grid search becomes computationally challenging as the dimension of q grows.

3.1 Asymptotic Distribution

We next present the large sample distributions of the test statistics. Define the following empirical processes:

$$\bar{T}_n(y,q) = \sqrt{n}\left[\widehat{F}^1\left(\widehat{\tau}^1(y,q)\right) - \widehat{F}^2\left(\widehat{\tau}^2(y,q)\right) - \left(F^1\left(\tau^1(y,q)\right) - F^2\left(\tau^2(y,q)\right)\right)\right],$$

$$\bar{U}_n(y,q) = \sqrt{n}\left[\int_{-\infty}^{y} \widehat{F}^1\left(\widehat{\tau}^1(u,q)\right) - F^1\left(\tau^1(u,q)\right) du - \int_{-\infty}^{y} \widehat{F}^2\left(\widehat{\tau}^2(u,q)\right) - F^2\left(\tau^2(u,q)\right) du\right].$$

Note that $T_n = \sup_{y,q} \bar{T}_n(y,q)$ and $U_n = \sup_{y,q} \bar{U}_n(y,q)$ under the least favorable case of the null hypothesis (i.e., when $F^1 = F^2$).

First, we collect the regularity conditions. Recall that we assume the array structure.

Assumption 1 *Suppose that (3), (4), (5), and the following hold for each n:*

1. *$\{X_t e_t, t = 1, 2, \ldots\}$ is a β-mixing array with coefficient $\beta(m)$ satisfying that for some positive h_1, h_2, b, and c,*

$$\beta(m) \leq \exp\left(-cm^{h_1}\right)$$

$$\Pr\left(|X_{ti}e_t| > \tau\right) \leq H(\tau) := \exp\left(1 - (\tau/b)^{h_2}\right) \quad \text{for all } i \tag{10}$$

$$h := \left(h_1^{-1} + h_2^{-1}\right)^{-1} < 1, \tag{11}$$

$$\log p = O\left(n^{h_1 \wedge (1 - 2h_1(1-h)/h)}\right). \tag{12}$$

2. *The marginal distribution functions F^j of ε_{tj} and the regression functions g^j and σ^j are twice continuously differentiable with uniformly bounded derivatives. Furthermore, the density functions f^j of ε_{tj} are strictly positive throughout $\mathbb{R}$.*

3. *The processes σ_t and σ_t^{-1} are bounded almost surely.*

4. *$E \sup_t (|r_{gt}| + |r_{\sigma t}|) = o\left(n^{-1/2}\right)$.*

5. *The sparsity parameter $s = |S(\beta_0)| + |S(\gamma_0)|$ satisfies: $s^4 \log^3 s = o(n)$, $\lambda_{thr} = o\left(\min\left\{|\beta_{0j}| : \beta_{0j} \neq 0\right\}\right)$ and $\mu_{thr} = o\left(\min\left\{|\gamma_{0j}| : \gamma_{0j} \neq 0\right\}\right)$.*

The definition for the β-mixing coefficient is given in Section A.3, which provides more discussion regarding the mixing rate restriction. The β-mixing condition is known to be more convenient to work with than the strong α-mixing condition as it allows for decoupling as in Berbee's lemma, see e.g. Doukhan et al. (1995). We impose this to establish the weak convergence of the residual

based empirical processes of time series data. Otherwise, the weaker strong mixing condition will be sufficient.

There is a trade-off between the decay rate of the mixing coefficients and that of the tail probability. As $h < 1$ due to (11), we need

$$h_1 (h-1)/h > -\alpha, \tag{13}$$

for some $\alpha < 1/2$ to allow p in (12) to grow. If $h_2 < 1$, this and (11) cannot be met simultaneously. And for $h_2 > 1$, this condition implies that $h_1 > (1-\alpha)h_2/(h_2-1)$.Thus, the smaller h_2, the larger h_1 is needed, that is, a trade-off between h_1, the mixing decay rate, and h_2, the tail decay rate. Refer to Merlevede et al. (2011) for more discussion on the case $h \geq 1$. The tail condition (10) is a weaker form of $E \exp\left(\delta |v_t|^{h_2}\right) < \infty$ with some positive δ due to the Markov inequality.

The condition on the approximation error in Assumption 1.4 appears more stringent than the usual condition of $o_p(s/n)$ but it is required to control the error in approximating F by the distribution function of $\sigma_t^{-1}(e_t + r_{gt})$. In addition, it seems natural to expect that the approximation error gets smaller as the number of terms used in the approximation increases as described in condition 5. This is the so-called *sparsity* assumption. In fact, this condition also imposes the so-called *beta-min* condition that the size of the minimal signal is well-separated from zero so that the selection by thresholding can be perfect asymptotically.

However, the identity of those terms is unknown and not unique. One way to understand the target value β_0 is via the following oracle problem:

$$s = \min \left\{ \underset{s'}{\operatorname{argmin}} \left(\min_{|\beta|_0 = s'} E\left(g(q_t) - X_t^\intercal \beta\right)^2 \right) + \sigma_e^2 \frac{s'}{n} \right\}$$

$$\beta_0 \in \underset{|\beta|_0 \leq s}{\operatorname{argmin}} E\left(g(q_t) - X_t^\intercal \beta\right)^2.$$

See e.g. Belloni and Chernozhukov (2013). Without a proper rank condition on X_t, β_0 is not unique. In fact, we do not need any unique approximation but any approximation that satisfies condition Assumption 1.4. It is in fact more flexible than the conventional series estimation of the regression functions since it does not demand the identity of the more important series terms. Furthermore, it even allows for combining very different bases such as polynomials and B-splines. The same comments apply to the approximation of $\sigma(\cdot)$.

The beta-min condition is a strong assumption that enables weak convergence of our test statistic. It is a sufficient but not necessary condition. It is an interesting future research topic to relax this. It has been relaxed in some special cases such as Chernozukov et al. (2017, 2018). Here, we discuss some challenges that lie in wait for this extension in our testing problem. First, our test concerns not a finite dimensional parameter but whole conditional distribution functions whose number of

conditioning variables diverges. There should be a strict restriction on the number of conditioning variables due to the curse of dimensionality. The beta-min condition imposes this restriction. Second, our test builds on the empirical distribution function of the growing dimensional regression residuals. It is well established by Mammen (1996) and Chen and Lockhart (2001) that the dimension cannot be bigger than the sample size raised to a power less than one half for the tightness of the empirical process. Thus, there might not be a weak limit without the beta-min or similar condition. Third, a finite sample Gaussian approximation may be considered instead of the weak limit of the test statistic. For instance, Chernozhukov et al. (2014) develop a Gaussian approximation to the empirical process which may not be P-Donsker at its supremum. However, the extension from their setting to the residual empirical process from the high dimensional time series regression is challenging, if it is even possible.

Although a full rank condition for X_t is impossible for high-dimensional regression, a reasonable rank condition is required to guarantee a good performance of the LASSO and the OLS after thresholding. Thus, we introduce the following definition.

Definition 1 *A $p \times p$ matrix Σ is compatible for an index set S with a compatibility constant $\phi > 0$, if*

$$|S| \frac{b^{\intercal} \Sigma b}{|b_S|_1^2} \geq \phi^2,$$

for any $b \in \mathbb{R}^p$ such that $|b_{S^c}|_1 \leq 3 |b_S|_1$.

Sufficient conditions for the compatibility condition are extensively discussed in Bühlmann and van der Geer (2011, Section 6) and are given by Basu and Michailidis (2015) for certain dependent data. In view of Bühlmann and van der Geer (2011, Corollary 6.8), the compatibility condition does not have to hold for the Gram matrix $n^{-1} \sum_{t=1}^{n} X_t X_t^{\intercal}$ with a uniform compatibility constant almost surely. It is sufficient to assume the following conditions:

Assumption 2 *The matrix $EX_t X_t^{\intercal}$ is compatible for $S(\beta_0)$ with some ϕ_β.*

Assumption 3 *Assume that $EX_t X_t^{\intercal}$ is compatible for $S(\gamma_0)$ with some ϕ_γ.*

The tuning parameters $\lambda, \lambda_{thr}, \mu$, and μ_{thr} need to diminish as the sample size n diverges. More, precisely, we require that

$$\sqrt{\frac{\log(np)}{n}} = o(\lambda), \tag{14}$$

which is a lower bound for λ. In practice, the cross validation method is commonly used for the selection of λ.

Next, λ_{thr} needs to satisfy the following lower bound

$$\lambda s = o\left(\lambda_{thr}\right). \tag{15}$$

Since the deviation bound for $|\widehat{\beta}_{lasso} - \beta_0|_1$ is $O_p\left(\lambda s\right)$, as shown in Theorem 6 in Section A.4, the proposed threshold makes $\widehat{S}$ collect only the relatively significant variables. Other popular alternatives include the smoothly clipped absolute deviation penalty (SCAD), Fan and Li (2001), and the adaptive LASSO, Zou (2006). We have chosen the thresholded LASSO mainly because it is explicit about its selection and is computationally more efficient since it is a convex optimization. The adaptive LASSO is more difficult to analyze theoretically as it is not explicit about its variable selection. The estimation of the support of a coefficient vector is a challenging problem. See van der Geer et al. (2011).

A notable difference in the LASSO estimation of the feasible skedastic regression is that the LASSO penalty term μ needs to be bigger than λ, the LASSO penalty in the mean regression. This is because the estimation error in the dependent variable $|\widehat{e}_t|$ in the feasible skedastic regression inflates the standard error of the estimator so that we need

$$\lambda\sqrt{s} = o\left(\mu\right) \text{ and } \mu s_\gamma = o\left(\mu_{thr}\right). \tag{16}$$

To characterize the asymptotic distributions, recall that $x_t = X_{t,S}$ and $w_t = X_{t,S_\gamma}$ and define

$$\mathcal{D}\left(y,q\right) = \mathbb{D}\left(y,q\right) + f\left(\tau\left(y,q\right)\right) D_1 - \tau\left(y,q\right) f\left(\tau\left(y,q\right)\right) D_2,$$

where $\tau\left(y,q\right) = \left(y - g\left(q\right)\right)/\sigma\left(q\right)$, and $\mathbb{D}$ and $D = \left(D_1, D_2\right)^\intercal$ are centered Gaussians with covariance kernels specified as follows:

$$E\mathbb{D}\left(y_1,q_1\right)\mathbb{D}\left(y_2,q_2\right) = \operatorname{cov}\left(1\left\{\varepsilon_{t1} \le \tau_1\right\} - 1\left\{\varepsilon_{t2} \le \tau_1\right\}, 1\left\{\varepsilon_{t1} \le \tau_2\right\} - 1\left\{\varepsilon_{t2} \le \tau_2\right\}\right)$$

with $\tau_i = \left(y_i - g\left(q_i\right)\right)/\sigma\left(q_i\right)$, $i = 1, 2$, and

$$EDD^\intercal = \lim_{n\to\infty} E \begin{bmatrix} \widetilde{x}_t^2\left(\varepsilon_{t1} - \varepsilon_{t2}\right)^2, & \widetilde{x}_t\left(\varepsilon_{t1} - \varepsilon_{t2}\right)\widetilde{w}_t\left(|\varepsilon_{t1}| - |\varepsilon_{t2}|\right) \\ . & \widetilde{w}_t^2\left(|\varepsilon_{t1}| - |\varepsilon_{t2}|\right)^2 \end{bmatrix}$$

$$E\mathbb{D}\left(y_1,q_1\right)D = \lim_{n\to\infty} E\left(1\left\{\varepsilon_{t1} \le \tau_1\right\} - 1\left\{\varepsilon_{t2} \le \tau_1\right\}\right) \begin{pmatrix} \widetilde{x}_t\left(\varepsilon_{t1} - \varepsilon_{t2}\right) \\ \widetilde{w}_t\left(|\varepsilon_{t1}| - |\varepsilon_{t2}|\right) \end{pmatrix},$$

with $\widetilde{x}_t = \mu_x^\intercal\left(Ex_tx_t^\intercal\right)^{-1}x_t$ and $\widetilde{w}_t = \mu_w^\intercal\left(Ew_tw_t^\intercal\right)^{-1}w_t\sigma_t$.

We denote by "$\Longrightarrow$" the weak convergence in $\ell^\infty(\mathbb{R}\times\mathcal{Q})$ that is *uniform* over the distributions of conditioning variables and errors in the sense of Theorem 2.8.2 of van der Vaart and Wellner (1996). Then, the following theorem establishes the uniform P-Donsker property of our test statistics.

Theorem 1 *Suppose that Assumptions 1 - 3 and conditions (14), (15), and (16) hold. Then, as* $n \to \infty$

$$\bar{T}_n(\cdot,\cdot) \Longrightarrow \mathcal{D}(\cdot,\cdot), \quad \bar{U}_n(\cdot,\cdot) \Longrightarrow \int_{-\infty}^{\cdot} \mathcal{D}(u,\cdot)\, du, \quad .$$

The theorem involves several nontrivial extensions of the existing statistical convergence results concerning the weak convergence of the empirical distribution functions of high-dimensional regression residuals, the deviation bounds for the weighted lasso estimator for the time series regression and for the skedastic regression, and the weak convergence of functions defined on an unbounded domains. These results may be of independent interest and thus are presented in the Appendix in a more self-contained manner. Building on these results, the convergence of T_n follows from the continuous mapping theorem.

The weak convergence of the SSD statistic cannot result from applying the continuous mapping theorem to the residual empirical process unless the support of the integral is bounded. It is well-known that the sample analogue higher order stochastic dominance test statistics needs proper weighting functions to control the tail behavior. In the case of the SSD hypothesis, Horvath et al. (2006) illustrate that the weak convergence can be achieved without a weighting function in a simpler case where one can observe ε_{1t}'s and ε_{2t}'s directly, while the previous literature like Linton et al. (2005) has assumed a bounded support, which is a special case of a weighting function of an indicator for a bounded set. Later, Linton et al. (2010) reintroduced a weighting function to study the bootstrap of the stochastic dominance test using residuals.

3.2 Local Power

To analyze the power of the test, we study the centering terms in $\bar{T}_n$ and $\bar{U}_n$, that is, $\sqrt{n}\left(F^1(y|q) - F^2(y|q)\right)$ and its integrated version. Let:

$$g^2(q) = g^1(q) + \delta_{1n}(q) \tag{17}$$

$$\sigma^2(q) = \sigma^1(q) + \delta_{2n}(q) \tag{18}$$

$$F^2(\tau) = F^1(\tau) + \delta_{3n}(\tau), \tag{19}$$

where recall that $\int \tau dF^2(\tau) = \int \tau dF^1(\tau) = 0$ and $\int |\tau|\, dF^2(\tau) = \int |\tau|\, dF^1(\tau) = 1$. We allow $F^j(y|q)$ to change with the sample size n, but have suppressed the dependence of $F^j(\cdot), g^j(\cdot)$, and $\sigma^j(\cdot)$ on the sample size n for the sake of notational simplicity. By the mean value theorem, for any given y and q,

$$\begin{aligned} F^1(y|q) - F^2(y|q) &= -F^2\left(\tau^2\right) + F^2\left(\tau^1\right) - F^2\left(\tau^1\right) + F^1\left(\tau^1\right) \\ &= \frac{\partial F^2(\bar{\tau})}{\partial \tau}\left(\frac{1}{\bar{\sigma}(q)}\right)\left(\delta_{1n}(q) + \bar{\tau}\delta_{2n}(q)\right) - \delta_{3n}\left(\tau^1\right), \end{aligned}$$

where $\bar{\tau} = (y - \bar{g}(q))/\bar{\sigma}(q)$ and $(\bar{g}(q), \bar{\sigma}(q))$ is a mean value between $(g^1(q), \sigma^1(q))$ and $(g^2(q), \sigma^2(q))$. If $\delta_{in} = 0$, for all $i = 1, 2, 3$,then it corresponds to the least favorable case of the null hypothesis.

We derive the asymptotic distribution of the test statistic T_n under the drifting sequence of models

$$(\delta_{1n}(q), \delta_{2n}(q), \delta_{3n}(\tau)) = \frac{1}{\sqrt{n}}(\delta_1(q), \delta_2(q), \delta_3(\tau)), \tag{20}$$

for all n, where δ_i is continuous and bounded for all i, q, and τ. Here, $\delta_3(\tau)$ stands for the deviation of the distribution functions of ε_{tj}s, which satisfy $E\varepsilon_{tj} = 0$ and $E|\varepsilon_{tj}| = 1$. Thus, it should satisfy some regularity conditions. First, the continuity of $F^j(\tau)$ means that $\delta_3(\tau) \to 0$ as $\tau \to \pm\infty$. As $E\varepsilon_{tj} = 0$, $\int_{-\infty}^{0} F^j(y)\,dy = \int_0^{\infty}(1 - F^j(y))\,dy$ by the integral-by-parts, yielding $\int_{-\infty}^{0} F(y) + \delta_{3n}(y)\,dy = \int_0^{\infty}(1 - F(y) - \delta_{3n}(y))\,dy$ and thus $\int_{-\infty}^{0}\delta_3(y)\,dy = -\int_0^{\infty}\delta_3(y)\,dy$ and $\int_{-\infty}^{\infty}\delta_3(y)\,dy = 0$. Similarly, applying the integral-by-parts to the restriction that $E|\varepsilon_{tj}| = 1$, we further restrict δ_3 to satisfy $\int_{-\infty}^{\infty}\delta_{3n}(y)\,dy = 2\int_0^{\infty}(F^2(y) - F^1(y))\,dy = 2\int_0^{\infty}\delta_{3n}(y)\,dy$. Unless $\delta_3(\tau) = 0$ for all τ, $\delta_3(\cdot)$ should take both positive and negative values. Our discussion is summarized in the following theorem.

Theorem 2 *Suppose that Assumptions 1 - 3 and conditions (14), (15), and (16) hold. Then, under (20) as* $n \to \infty$

$$T_n \Longrightarrow \sup_{y,q}[\mathcal{D}(y,q) + \mathcal{B}(y,q)],$$
$$U_n \Longrightarrow \sup_{y,q}\int_{-\infty}^{y}[\mathcal{D}(u,q) + \mathcal{B}(u,q)]\,du.$$

This theorem derives the asymptotic distribution under a sequence of alternative hypotheses. When $\mathcal{B}(y,q) \leq 0$,the sequence obeys the null hypothesis. Thus, the deterministic non-centrality function $\mathcal{B}(y,q)$ determines the local power of the test. For a given critical value c_α of significance level α from the distribution of $\sup_{y,q}\mathcal{D}(y,q)$ or $\sup_{y,q}\int_{-\infty}^{y}\mathcal{D}(u,q)\,du$ a non-trivial test demands that the probability of $\sup_{y,q}[\mathcal{D}(y,q) + \mathcal{B}(y,q)]$ or $\sup_{y,q}\int_{-\infty}^{y}[\mathcal{D}(u,q) + \mathcal{B}(u,q)]\,du$ greater than c_α exceeds α. Equivalently, $\mathcal{B}(\widehat{y},\widehat{q})$ and $\int_{-\infty}^{\widehat{y}}\mathcal{B}(u,\widehat{q})\,du$ are greater than 0 with a probability bigger than α, where $(\widehat{y},\widehat{q})$ denotes the maximizers of the stochastic processes, $[\mathcal{D}(y,q) + \mathcal{B}(y,q)]$ or its integrated process, respectively.

We discuss some sufficient conditions for this. To begin with, suppose $\delta_1(\cdot) > 0$ while[5] $\delta_2(\cdot) = 0$, and $\delta_3(\cdot) = 0$. Then, as $\frac{\partial F(\tau)}{\partial\tau}\frac{1}{\sigma(q)} > 0$ for any y and q, $\mathcal{B}(y,q) > 0$ for all y and q thus implying a non trivial power. Similarly, if $\delta_3(\cdot) < 0$, the tests have non trivial powers. When $\delta_2(q) \neq 0$, we note that $\tau = (y - g(q))/\sigma(q)$ can take both positive and negative values for any q and so can $\tau\delta_2(q)$. Thus, $\mathcal{B}(y,q) < 0$ for some y and q and $\mathcal{B}(y,q) > 0$ for others, with δ_1 and δ_3 held fixed at

[5]In fact, it is well-known that $F_X(\tau) \leq F_Z(\tau)$ implies that $EX \geq EZ$ by the integral by parts formula. Thus, the conditional mean relation $g^1(q) < g^2(q)$ implies that $F^1(y|q) > F^2(y|q)$ for some y.

zero. The local power in this case depends on the joint distribution of $\mathcal{B}(\widehat{y}, \widehat{q})$ and $(\widehat{y}, \widehat{q})$ in a rather complicated manner.

4 Bootstrap for Inference

This section presents a bootstrap algorithm to approximate the p-values of our test statistics. Our procedure, we name the smooth stationary bootstrap, combines two separate methods in the literature to take care of the complexity of our test statistics due to the temporal dependence and the highly nonlinear nature of the statistics. In the Appendix we review the stationary bootstrap algorithm and the smooth bootstrap for a generic sequence of variable $\mathcal{Z}_n$. Here, we combine them to approximate the distribution of our residual based processes.

4.1 Bootstrap Test Statistic

The asymptotic distributions of T_n and U_n are not pivotal hence the critical values cannot be tabulated once and for all. Thus, we introduce a bootstrap procedure that is based on the post-selection dataset $\{\widehat{x}_t, \widehat{w}_t, y_t,\, t = 1, \ldots, n\}$. The procedure is as follows

1. Fix constants a_n and π_n within the interval $(0, 1)$ and a smooth distribution function G and generate $\{i_t^*, \eta_t\}$ as follows

   (a) Let d_t and i_t, $t = 1, \ldots, n$, be random draws from Bernoulli(π_n) and Uniform$\{1, \ldots, n\}$ distributions, respectively.

   (b) Let $i_1^* = i_1$. For $t = 2, \ldots, n$, let

   $$i_t^* = \left(i_{t-1}^* + 1\right)(1 - d_t) + i_t d_t$$

   with the convention that $i_{t-1}^* + 1 = 1$ if $i_{t-1}^* = n$.

   (c) Let η_t be i.i.d.

2. For each $j = 1, 2$, construct the bootstrap sample $\left\{x_t^* = \widehat{x}_{i_t^*}, w_t^* = \widehat{w}_{i_t^*}\right\}$ and $\{\varepsilon_{tj}^* = \widehat{\varepsilon}_{i_t^*, j} + a_n \eta_t\}$, respectively, and then compute

   $$y_{tj}^* = x_t^{*\intercal} \widehat{\beta}^j + w_t^{*\intercal} \widehat{\gamma}^j \cdot \varepsilon_{tj}^*, \quad t = 1, \ldots, n.$$

3. For each $j = 1, 2$, obtain the OLS estimates $\widehat{\beta}^{j*}$ with the bootstrap sample $\{x_t^*, y_{tj}^*\}$, i.e.,

   $$\widehat{\beta}^{j*} = \left(\sum_{t=1}^{n} x_t^* x_t^{*\intercal}\right)^{-1} \sum_{t=1}^{n} x_t^* y_{tj}^*,$$

and compute the bootstrap OLS residuals $\widehat{e}^*_{tj} = y^*_t - x^{*\intercal}_t \widehat{\beta}^{j*}$, $\quad t = 1, \ldots, n$. Then, compute:

$$\widehat{\gamma}^{j*} = \left(\sum_{t=1}^{n} w^*_t w^{*\intercal}_t \right)^{-1} \sum_{t=1}^{n} w^*_t \left| \widehat{e}^*_{tj} \right|,$$

$$\widehat{\varepsilon}^*_{tj} = \widehat{e}^*_{tj} \left(w^{*\intercal}_t \widehat{\gamma}^{j*} \right)^{-1}, \text{ and } \widehat{\tau}^{j*}(y,q) = \left(\widehat{w}(q)^{\intercal} \widehat{\gamma}^{j*} \right)^{-1} \left(y - \widehat{x}(q)^{\intercal} \widehat{\beta}^{j*} \right).$$

4. Define the empirical distribution functions:

$$\widehat{F}^{j*}(\tau) = \frac{1}{n} \sum_{t=1}^{n} 1\left(\widehat{\varepsilon}^*_{tj} \leq \tau \right) \qquad ; \qquad F^{j*}(\tau) = \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{\tau - \widehat{\varepsilon}_{tj}}{a_n} \right).$$

Then construct the bootstrap statistic, for the FSD test

$$T^*_n = \sqrt{n} \sup_{y,q} \left[\widehat{F}^{1*}\left(\widehat{\tau}^{1*}(y,q) \right) - \widehat{F}^{2*}\left(\widehat{\tau}^{2*}(y,q) \right) - \left(F^{1*}\left(\widehat{\tau}^{1}(y,q) \right) - F^{2*}\left(\widehat{\tau}^{2}(y,q) \right) \right) \right],$$

and for the SSD test,

$$U^*_n = \sqrt{n} \sup_{y,q} \int_{-\infty}^{y} \left[\widehat{F}^{1*}\left(\widehat{\tau}^{1*}(u,q) \right) - F^{1*}\left(\widehat{\tau}^{1}(u,q) \right) - \left(\widehat{F}^{2*}\left(\widehat{\tau}^{2*}(u,q) \right) - F^{2*}\left(\widehat{\tau}^{2}(u,q) \right) \right) \right] du.$$

The bootstrap critical value c^*_α for a prespecified significance level α is then computed from the distribution of T^*_n, which can be approximated from the empirical distribution of the simulated bootstrap statistics by repeating steps 1-4. Also, the bootstrap p-value, that is, the conditional probability that $T^*_n > T_n$, can be approximated as the proportion of the generated $\{T^*_n\}$ that are greater than equal to T_n. The same applies for the SSD test.

We next make some remarks on our bootstrap procedure. First, note that the centering term of the bootstrap residual empirical process is not the empirical distribution function $\widehat{F}_n$ of the original sample. The proper centering reflects the smoothing by η_t and thus the c.d.f. $F^*(\cdot)$ is continuous unlike $\widehat{F}_n(\cdot)$. Second, our bootstrap scheme mimics the OLS steps only using the selected regressors. An alternative bootstrap scheme is to resample both untransformed regressors q_t and errors $\widetilde{\varepsilon}_t$ and perform the threshold LASSO for each bootstrap sample. This is computationally much more demanding since each bootstrap iteration now involves LASSO estimation with high dimensional variables. Thus, we do not pursue this route in this paper.

Next we establish the asymptotic validity of our bootstrap test.

Assumption 4 *Let the i-th derivative of a function g be denoted by $g^{(i)}$ and assume the following:*

1. *The function G is κ-times differentiable with $\kappa \geq 3$ and the first derivative $G^{(1)}$ of G is a (symmetric) probability density function such that $\int G^{(1)}(z) z^2 dz < \infty$ and $G^{(v)}$ is bounded for all $v \leq \kappa$ and $\int G^{(v)}(z)^2 dz < \infty$ for $v < \kappa$.*

2. The bandwidth a_n satisfies that $a_n^4 n \to 0$ as $n \to \infty$.

Theorem 3 *Suppose that Assumptions 1 - 4 and conditions (14), (15), and (16) hold. Let c_α^* denote the bootstrap critical value of level α for T_n. Then, under $\mathcal{H}_0$, we have*

$$\limsup_{n\to\infty} \Pr\left\{T_n > c_\alpha^*\right\} \leq \alpha$$

for any $0 < \alpha < 1$, while under $\mathcal{H}_1$,

$$\Pr\left\{T_n > c_\alpha^*\right\} \to 1$$

for any $0 < \alpha < 1$. The same holds true for U_n.

This result shows the size control and power property of our test.

4.2 Boosting Power

Before concluding the description of our test, we note that Linton et al. (2010) proposed the so-called contact set approach to improve the power of the *unconditional* stochastic dominance test. It works well with the unconditional dominance testing because the contact set is estimated on the real line. However, it is less practical in our setting since we need to estimate the set for each value of conditioning variables whose dimension grows. To mitigate this problem, we propose to apply a *screening principle*, which is to test certain implications of the null hypothesis with a *higher criticism*. This approach is advocated by e.g. Fan et al. (2015).

One implication of the first- and second-order stochastic dominance of y_t^1 over y_t^2 (conditional on $q_t = q$) is the dominance of the conditional means, i.e.,

$$E\left(Y_t^1 | q_t = q\right) \geq E\left(Y_t^2 | q_t = q\right). \tag{21}$$

The negation of this implication implies the negation of the null hypothesis. Using the conditional mean function $\widehat{g}\left(q_t\right) = X_t^\top \widehat{\beta}_{Tasso}$, which is estimated to construct our main test statistic T_n in Section 2.1, we can screen this implication for a sequence of values of $q_t \in \{q_1, ..., q_J\}$ or $X_t \in \{x_1 = x\left(q_1\right), \ldots, x_J = x\left(q_J\right)\}$ by statistics

$$t_k = 1\left\{\frac{x_k^\intercal\left(\widehat{\beta}^2 - \widehat{\beta}^1\right)}{\widehat{\sigma}_k} > c^*\right\},\ k = 1, \ldots, J,$$

for some scaling $\widehat{\sigma}_k$ and a critical value c^*. If $t_k = 1$ for any k, we can stop and conclude that the null is rejected. Otherwise, we resort to our test statistic T_n. To justify this initial screening, the value c^* needs to satisfy the high criticism property that

$$\lim_{n\to\infty} \Pr\left\{\max_{k\in\{1,2,\ldots,J\}} \frac{x_k^\intercal\left(\widehat{\beta}^2 - \widehat{\beta}^1\right)}{\widehat{\sigma}_k} \leq c^*\right\} = 1$$

under the null hypothesis.

For sieve nonparametric regression with dependent data, Lemma 2.4 in Chen and Christensen (2015) provides a uniform deviation bound $||\widehat{h} - h_0||_\infty \leq O_p(\zeta_n \lambda_n \sqrt{n^{-1} \log n}) +$ bias, where $\widehat{h}$ and h_0 stand for the estimated and true regression functions, respectively, $\lambda_n = \lambda_{\min}^{-1/2} (EX(q_t) X(q_t)^\intercal)$, and ζ_n is the size of the regressor vector $\sup_q \|X(q)\|$. When the support of q_t is bounded, $\zeta_{K,n} \lambda_{K,n} = O(\sqrt{K})$ for commonly used linear sieves, see Chen and Christensen (2015) for more detailed discussion. When testing (21), the biases will cancel out or negative. Thus, we may set $c^* = \zeta_n \lambda_n \sqrt{n^{-1} \log n} \log \log n$.

As for x_k, good candidates are those that promote the sparse alternatives. However, we do not consider the k-th unit vector $\iota_k = (0, \ldots, 0, 1, 0, \ldots, 0)^\intercal$, under which t_k would concern the significance of each elementwise difference in $\widehat{\beta}^j$, because $X_t = X(q_t)$. Rather, we consider x_k of the form $X(q_k)$ for a grid of $\{q_k\}$.

Choosing a proper scaling is a challenging issue in high-dimensional inference. We suggest to set $\widehat{\sigma}_k^2 = n \sum_{i=1}^J x_{ki}^2 \left(\sum_{t=1}^n X_{ti}^2\right)^{-2} \sum_{t=1}^n X_{ti}^2 \left(x_k^\top \widehat{\gamma}\right)^2 \pi/2$, which corresponds to the case where there is no correlation among the X_{ti}'s. These estimates are uniformly bounded.

5 Monte Carlo Simulation

In this section, we provide Monte Carlo simulation results that evaluate the finite sample performance of our statistic. We generated the data for $j = 1, 2$ as follows:

$$y_t^j = \beta^j q_{1,t} + c_v \cdot (|q_{1,t}| + 1) \cdot \varepsilon_t,$$

where $c_v = 0.3, \varepsilon_t$ is an i.i.d. normal error term with mean 0 and $\mathbb{E}[|\varepsilon_t|] = 1$. On the other hand, the explanatory variables $q_{i,t}$ are generated by following time series process:

$$q_{i,t} = a + b q_{i,t-1} + e_{i,t}$$

where $i = 1, 2$, $a = 0, b = 0.5$, and $t = 1, 2, \ldots, n$. Here t starts from -99 and thus the first 100 observations are discarded. We estimate the model based on $X_t = X(q_{1,t}, q_{2,t})$, which are transformations of $q_{1,t}, q_{2,t}$ and are common for $j = 1, 2$. They are powers and interaction terms of $q_{1,t}, q_{2,t}$ up to polynomial order of 10. Hence, X_t has 65 variables excluding the intercept. We also add additional variables for X_t so that the high-dimensional setting $n << p$ holds. The additional variables may vary according to subsections.

Through this section, the parameters for LASSO is $\lambda = c_{v'} \cdot \sqrt{\log p / n}$ where n is the sample size, p is the number of covariates, and $c_{v'} \propto c_v$. Any variable is selected by LASSO if its estimated coefficient is larger than $2.0 \cdot \lambda$. Then we again run OLS with selected variables.

Table 1: Rejection probability with higher polynomial orders

order	10	15	20	25	30	35	40
$n \setminus p$	65	135	230	350	495	665	860
100	0.077	0.062	0.065	0.073	0.068	0.082	0.084
200	0.075	0.048	0.060	0.055	0.048	0.060	0.047
300	0.055	0.043	0.048	0.053	0.058	0.045	0.051
400	0.052	0.056	0.055	0.041	0.047	0.051	0.044
500	0.048	0.051	0.047	0.045	0.051	0.039	0.046

Table 2: Rejection probability with additional q pairs

New Pairs	1	3	5	7	10	13	15
$n \setminus p$	130	260	390	520	715	910	1040
100	0.071	0.070	0.073	0.085	0.058	0.063	0.068
200	0.061	0.062	0.064	0.057	0.062	0.056	0.061
300	0.058	0.070	0.062	0.044	0.072	0.054	0.055
400	0.048	0.067	0.062	0.057	0.069	0.075	0.052
500	0.048	0.054	0.058	0.062	0.067	0.065	0.054

5.1 Size Simulation

We increase p by adding more terms to X_t with three different ways. First, we grow the polynomial order of $q_{1,t}, q_{2,t}$ constructing X_t. Second, we can generate more $q_{3,t}, q_{4,t}, \ldots$ pairs and add them to X_t. Finally, we add lagged $q_{1,t}, q_{2,t}$ terms and its powers (up to 10). We draw 10^5 random grid points from the uniform distribution of grid support. Then to obtain the supremum of our objective function, we evaluate the object at each point and take maximum.

We report the rejection rate at the significance level of $\alpha = 0.05$ with the true parameter value of $\beta^1 = \beta^2 = 1$ out of 1000 simulation iterations. Tables 1, 2, and 3 report the rejection frequencies for each case.

We also experiment with different values of $b = 0.3, 0.4, \ldots, 0.9$ to examine the effect of higher serial correlation in q_t. See Table 4 for the results, which reports minor over rejection tendency for bigger values of b.

5.2 Power Simulation

In this section, we fix Max Lag = 30 so that $p = 665$. All other settings are same with the size simulation. Then we evaluate the power performance of our test in three ways. First, by changing

Table 3: Rejection probability with lagged q terms

Max lag	5	10	15	20	25	30	35	40
$n \setminus p$	165	265	365	465	565	665	765	865
100	0.072	0.072	0.082	0.095	0.088	0.070	0.079	0.078
200	0.064	0.075	0.071	0.070	0.067	0.045	0.067	0.070
300	0.048	0.074	0.082	0.091	0.077	0.060	0.084	0.070
400	0.070	0.073	0.068	0.068	0.071	0.072	0.062	0.070
500	0.063	0.071	0.078	0.070	0.086	0.087	0.075	0.067

Table 4: Rejection probability with different AR coefficients b

$n \setminus b$	0.3	0.4	0.5	0.6	0.7	0.8	0.9
100	0.082	0.089	0.087	0.093	0.089	0.088	0.132
200	0.068	0.072	0.080	0.073	0.090	0.088	0.089
300	0.076	0.077	0.082	0.068	0.078	0.070	0.092
400	0.082	0.093	0.057	0.086	0.079	0.081	0.083
500	0.088	0.085	0.072	0.066	0.079	0.060	0.094

$\beta^2 = 1.0, 1.1, \ldots, 2.0$. Second, we shift y^2 by adding $\alpha = 0.1, \ldots, 1.0$. We only report the simulated rejection probability up to the case $\beta^2 = 1.5$ or $\alpha = 0.5$ since beyond the value the rejection probabilities are almost identical to 1. Tables 5 and 6 report the results for each case. As expected the rejection frequencies grow as the sample size increases and as the alternative models move away from the null model.

Third, we change the error distribution by letting ε_t^2 follow $(Z^2 - 1)/0.9680$, i.e. chi-square with one degrees of freedom normalized to mean 0 and the first absolute moment 1. The following Figure 1 compares it with normal distribution with mean 0 and the first absolute moment 1.

Table 7 shows that the power improves as the sample size increases.

Table 5: Rejection probability with β^2 being $1.0, 1.1, \cdots, 1.5$

$n \setminus \beta^2$	1.0	1.1	1.2	1.3	1.4	1.5
100	0.090	0.124	0.284	0.433	0.636	0.801
200	0.065	0.135	0.357	0.625	0.816	0.935
300	0.079	0.181	0.465	0.764	0.936	0.976
400	0.091	0.192	0.557	0.866	0.969	0.981
500	0.082	0.238	0.686	0.933	0.977	0.987

Table 6: Rejection probability after shifting y^2 by α

$n \setminus \alpha$	0.0	0.1	0.2	0.3	0.4	0.5
100	0.070	0.251	0.488	0.741	0.918	0.988
200	0.086	0.324	0.764	0.970	0.987	0.994
300	0.079	0.452	0.860	0.983	0.994	0.986
400	0.082	0.530	0.933	0.981	0.989	0.994
500	0.081	0.581	0.968	0.983	0.992	0.998

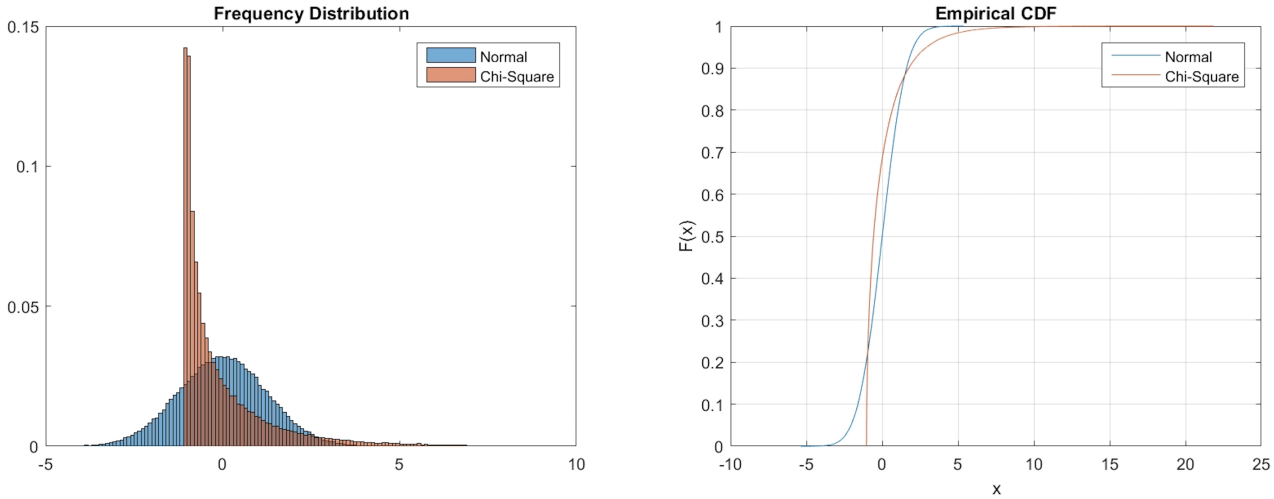


Figure 1: Distributions

Table 7: Rejection probability of normal vs. chi-square error distribution

n	Rejection Prob.
100	0.043
200	0.094
300	0.182
400	0.327
500	0.417

Table 8: Sample Statistics

	Mean	Std	Skew	Kurt	$\rho(1)$	$\rho(2)$
US	0.0314	1.134	-0.126	10.979	-0.0285	-0.0431
Global	0.0191	0.916	-0.216	10.306	0.1519	-0.0376

6 Application

We apply our method to the comparison of US and Global equity returns. The home bias puzzle has been investigated by a number of authors including Chan, Covrig, and Ng (2005), French and Poterba (1991), and Lewis (1999). Levy and Levy (2014) argue that despite a significant reduction in implicit and explicit transaction costs around the world, the US home bias in stock and bond returns has not disappeared, that is, domestic investors invest a higher fraction of their wealth in domestic stocks than is warranted by the global mean variance trade-off. We shed some light on this by comparing the conditional return distributions of US stocks and international stocks to see whether such domestic biased strategies could be justified when accounting for many conditioning variables in a general way and when adopting stochastic dominance rather than mean variance as the criterion.

The dataset we use is the Fama-French US and Global risk premium daily series from 7 August 1992 to 30 June 2016 obtained from Kenneth French's Data Library (6020 sample data in total, with extra 20 observations in use to accommodate lagged returns). The two return series have contemporaneous correlation of around 0.842. The sample statistics are reported in Table 8.

We next test the conditional dominance of the US series over the global series with 674 conditioning variables detailed below in Table 9.[6] We provide more detail on the variables in the appendix.

We conduct a non-overlapping rolling window analysis of size 500, roughly two years, to allow for nonstationarity. We plot the series of p-values reported from 12 windows below in Figure 2. Throughout, λ is set to be $\sqrt{\log(np)/n} = 0.1595$, the lower bound given in (14), and the LASSO threshold constant is set to be 0.1. The result reveals that the null hypothesis suggesting the conditional dominance of the US series over the global series is rejected at the 5% level of significance, except for the periods of 1992-1994, 2004-2006, and 2008-2012 where we do not have sufficient statistical evidence to so conclude. It appears that those years have been somewhat different relative to the rest of the sample.

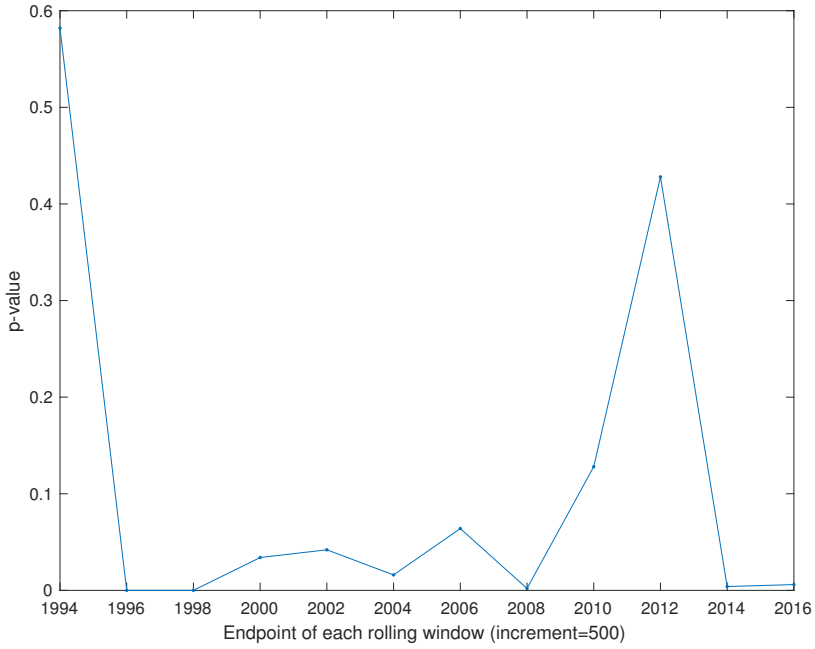
We next investigate the selection process, i.e., which covariates survived. For the period from 07/08/2000 to 06/08/2002, we calculate the sample correlations between the conditioning variables

[6]We carried out Linton, Maasoumi and Whang's (2005) LMW test of stochastic dominance using subsampling based critical values. We cannot reject the SSD hypothesis unconditionally.

Table 9: Description of the conditioning variables

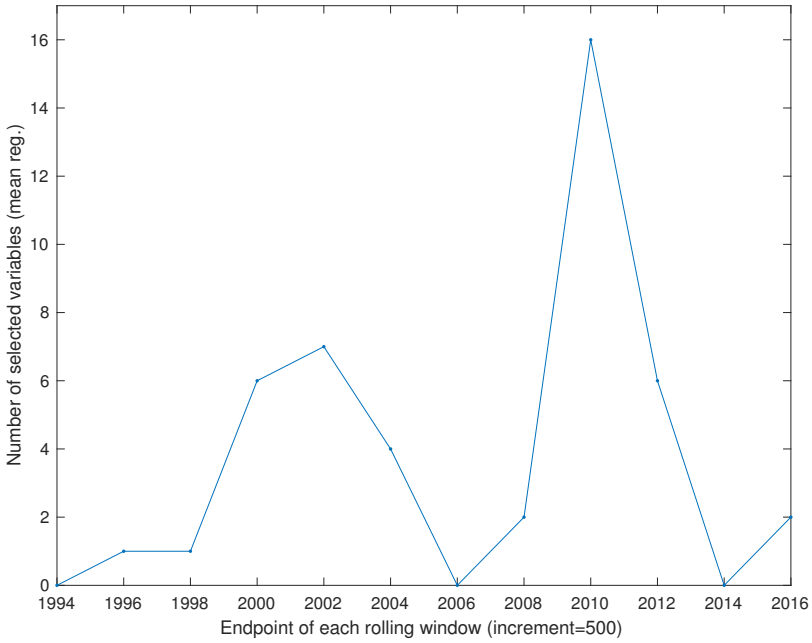
Index	Description
# 1-40	Lagged returns (max. lag = 20)
# 41-200	Powers of lags
# 201-600	Interactions
# 601-638	Momentum measures
# 639-657	Changes in trading volume
# 658-665	Relative strength Indices
# 666-669	Moving average oscillators
# 670-674	Day of the week dummy variables

Figure 2: Tests



p-values from the rolling window analysis

Figure 3: Selection



Number of selected variables from the mean regression

Table 10: Correlations of the selected variables, an example

Variable Index	Rank	Correlation (abs.)	Sign of the correlation
# 256	1	0.169272187	+
# 234	2	0.151388803	+
# 424	3	0.145195343	−
# 253	4	0.143038571	−
# 351	6	0.137275434	+
# 650	9	0.129063574	+
# 1	291	0.042979460	+

and the US series, and rank them in descending order of the absolute value of the correlations. Table 10 reports the correlations of 7 "selected variables" from the mean regression (cf. Figure 3); the result suggests that the selected variables tends to be those with high correlations in general, with 6 out of 7 variables listed on top 9 out of 674.

7 Conclusion

The concept of stochastic dominance has been playing an indispensable role in various economic analyses. This paper extends the econometric literature to the data rich environment by considering an abundant set of conditioning information that a rational investor may take account of. We achieve this by working with a location and scale semiparametric model, which allows for very general specification of the mean and variance in terms of a large number of conditioners, while using conventional nonparametric methods to estimate the error distribution. Although we focus on the stochastic dominance in every possible scenario by considering the supremum statistic over all realizations of conditioning variables, the results readily carry over to the supremum statistics over a particular event of interest. There are a number of other extensions that we are interested in pursuing. Firstly, the case where the outcome is the same random variable but the conditioning information sets may be different or nested is of interest. This poses some unique problems when selection is employed to choose relevancy of predictors. Secondly, we may consider the case where there are many outcome variables and the hypothesis of interest is stochastic maximality, McFadden (1989).

References

[1] Akritas, M.G., Van Keilegom, I., 2001. Non-parametric estimation of the residual distribution. Scandinavian Journal of Statistics 28, 549-567

[2] Bai, J., 1994. Weak convergence of the sequential empirical processes of residuals in ARMA models. The Annals of Statistics, 2051-2061

[3] Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics 43, 1535-1567

[4] Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C, 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369-2429.

[5] Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19, 521-547

[6] Belloni, A., Chernozhukov, V., & Hansen, C. 2014. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2), 608-650.

[7] Bickel, P.J., Ritov, Y.a., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics 37, 1705-1732

[8] Boldin, M.V., 1983. Estimation of the distribution of noise in an autoregression scheme. Theory of Probability & Its Applications 27, 866-871

[9] Bühlmann, P., Van De Geer, S., 2011. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media

[10] Chatterjee, A., Gupta, S., Lahiri, S.N., 2015. On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property. Journal of Econometrics 186, 317-324

[11] Chen, G., Lockhart, R.A., 2001. Weak convergence of the empirical process of residuals in linear models with many parameters. Annals of statistics, 748-762

[12] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. 2017. Double/debiased/neyman machine learning of treatment effects. American Economic Review, 107(5), 261-65.

[13] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins. 2018 Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, 21(1), C1-C68.

[14] Chernozhukov, V., Chetverikov, D., & Kato, K. 2014. Gaussian approximation of suprema of empirical processes. The Annals of Statistics, 42(4), 1564-1597.

[15] Davidson, J., 1994. Stochastic limit theory: An introduction for econometricians. OUP Oxford.

[16] Doukhan, P., Massart, P., Rio, E., 1995. Invariance principles for absolutely regular empirical processes. In: Annales de l'IHP Probabilites et statistiques, pp. 393-427. Gauthier-Villars

[17] Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96, 1348-1360

[18] French, K.R., Poterba, J.M., 1991. Investor diversification and international equity markets. National Bureau of Economic Research

[19] Goncalves, S., Politis, D., 2011. Discussion: Bootstrap methods for dependent data: A review. Journal of the Korean Statistical Society 40, 383-386

[20] Gonzalo, J., and J. Olmo (2014). Conditional stochastic dominance tests in dynamic settings. International review 55(3) 819-838.

[21] Koul, H.L., 1970. Some convergence theorems for ranks and weighted empirical cumulatives. The Annals of Mathematical Statistics, 1768-1773

[22] Koul, H.L., Levental, S., 1989. Weak convergence of the residual empirical process in explosive autoregression. The Annals of Statistics 17, 1784-1794

[23] Levy, H., Levy, M., 2014. The home bias is here to stay. Journal of Banking & Finance 47, 29-40

[24] Lewis, K.K., 1999. Trying to explain home bias in equities and consumption. Journal of economic literature 37, 571-608

[25] Ling, S., 1998. Weak convergence of the sequential empirical processes of residuals in nonstationary autoregressive models. The Annals of Statistics 26, 741-754

[26] Linton, O., Maasoumi, E., & Whang, Y. J. (2005). Consistent testing for stochastic dominance under general sampling schemes. The Review of Economic Studies, 72(3), 735-765.

[27] Loynes, R.M., 1980. The empirical distribution function of residuals from generalised regression. The Annals of Statistics, 285-298

[28] McFadden, D. (1989), "Testing for stochastic dominance," in Part II of T. Fomby and T.K. Seo (eds.) *Studies in the Economics of Uncertainty* (in honor of J. Hadar), Springer-Verlag.

[29] Mammen, E., 1996. Empirical process of residuals for high-dimensional linear models. The annals of statistics 24, 307-335

[30] Medeiros, M.C., Mendes, E.F., 2016. ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. Journal of Econometrics 191, 255-271

[31] Muller, U.U., Schick, A., Wefelmeyer, W., 2004. Estimating linear functionals of the error distribution in nonparametric regression. Journal of Statistical Planning and Inference 119, 75-93

[32] Neumeyer, N., 2009. Smooth Residual Bootstrap for Empirical Processes of Nonparametric Regression Residuals. Scandinavian Journal of Statistics 36, 204-228

[33] Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. Journal of the American Statistical association 89, 1303-1313

[34] Seo, M.H. and Otsu, T., 2018. Local M-estimation with discontinuous criterion for dependent and limited observations. The Annals of Statistics, 46(1), 344-369.

[35] Schick, A., Wefelmeyer, W., 2002. Estimating the innovation distribution in nonlinear autoregressive models. Annals of the Institute of Statistical Mathematics 54, 245-260

[36] Song, S., Bickel, P.J., 2011. Large vector auto regressions. arXiv preprint arXiv:1106.3915

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

[38] van de Geer, S., Buhlmann, P., Zhou, S., 2011. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). Electronic Journal of Statistics 5, 688-749

[39] Whang, Y. J. (2019). Econometric Analysis of Stochastic Dominance: Concepts, Methods, Tools, and Applications. Cambridge University Press.

[40] Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American statistical association 101, 1418-1429

Appendix

The first two sections of this appendix derive the weak convergence of the residual-based empirical distribution functions from high-dimensional time series regression and deviation bounds for the weighted lasso estimates when the regression under consideration involves time series data and/or it is the feasible skedastic regression. These results can be of independent interests. Then, we turn to the proof of main theorems.

A Appendix

A.1 Data series used in application

Suppose that Y_t is the daily return on the benchmark and R_t is the daily return on the alternative. We consider the following price based predictors: Lagged returns Y_{t-j}, R_{t-j}, $j = 1, \ldots, 10$; Powers of lags Y_{t-j}^k, R_{t-j}^k; $j = 1, \ldots, 10$, $k = 2, 3, 4, 5$; Interactions $Y_{t-j}R_{t-k}$, $j, k = 1, \ldots, 10$; Momentum measures $\sum_{j=1}^k Y_{t-j}$, $\sum_{j=1}^k R_{t-j}$ for $k = 2, 3, \ldots, 11$; Local trends of window periods of 5, 10, 15, 20 days, respectively. For example, fit the linear regression $\log P_{Yi} = \alpha + \beta \cdot i + \varepsilon_i$,using the data $i = t-1, t-2, \ldots, t-k$. Then include the return forecast $\widehat{\alpha} + \widehat{\beta} t - \log P_{t-1}$. Relative strength indices which are the percentages of the previous 5, 10, 15, 20 days that returns were positive, respectively. Moving average oscillators, each of which is the difference between an average of the closing prices over the previous q_1 days and that over the previous q_2 days, where $q_1 < q_2$, $q_1 = 1, 5, 10, 15$ and $q_2 = 5, 10, 15, 20$. Nonparametric regressions $E(Y_t|Z_{t-j}), E(R_t|Z_{t-j})$, where Z is an observed state variable, for example $Z_{t-j} = Y_{t-j}$. We also included additional nonprice based variables such as: Trading Volume Changes $\log V_{t-j} - \log V_{t-j-1}$, $j = 1, \ldots, 10$; Day of the week dummies; Term spread; Junk spread; Industrial production and inflation (at best these are monthly observed); Rescaled time t/T. For changes in trading volume and the moving average oscillators, we take the S&P 500 daily trading volume and closing prices data from Yahoo Finance, respectively.

We plot in Figure 4 the estimated conditional means of the two series across the rolling window period, which shows how closely the conditional means move.

A.2 Smooth Stationary Bootstrap

The stationary bootstrap of a sequence of variable $\mathcal{Z}_n = \{v_1, \ldots, v_n\}$, Politis and Romano (1994), may be understood as a way to generate random sequences of indexes $\{i_t^*\}$, with parameter $0 < \pi_n < 1$,

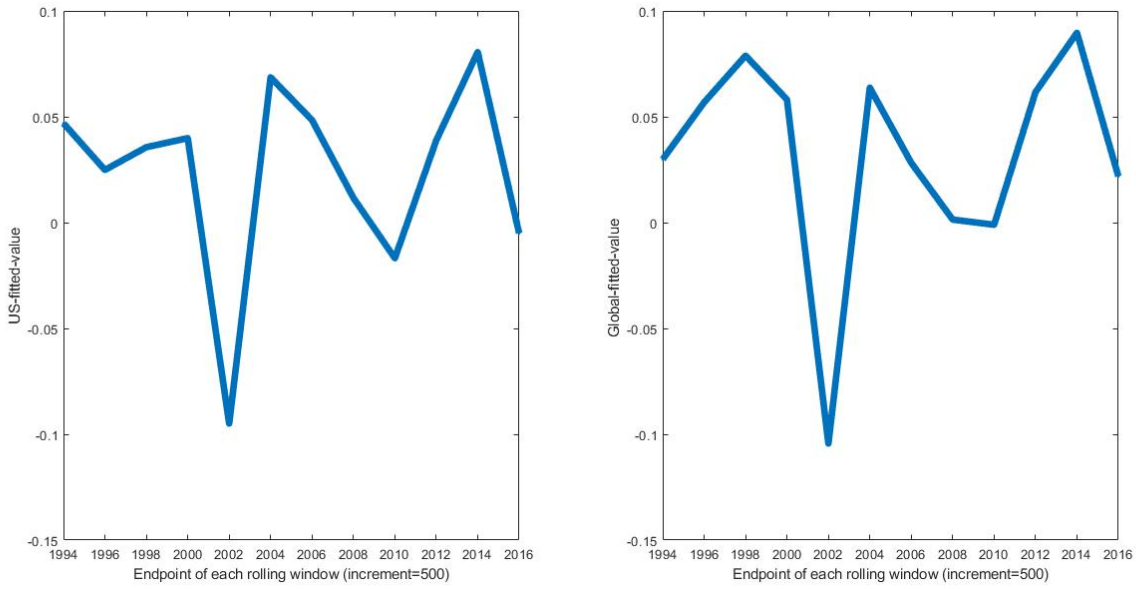


Figure 4: Estimated Conditional Means

to guarantee the stationarity of the resulting sequence $\{v_{i_t^*}\}$. Specifically,[7]

1. Let d_t and i_t, $t = 1, \ldots, n$, be random draws from Bernoulli(π_n) and Uniform$\{1, \ldots, n\}$ distributions, respectively.

2. Let $i_1^* = i_1$.

3. For $t = 2, \ldots, n$, let
$$i_t^* = \left(i_{t-1}^* + 1\right)(1 - d_t) + i_t d_t$$
with the convention that $i_{t-1}^* + 1 = 1$ if $i_{t-1}^* = n$.

The smooth bootstrap by Neumeyer (2009) adds a continuous variable to the original nonparametric bootstrap sample to make the resulting bootstrap variable continuous. The smoothing is introduced when the continuity of the distribution is a key condition to characterize the asymptotic distribution of the statistic of interest, as in for example the residual empirical process. Let η_t be generated from a continuous distribution $G(\cdot)$, which is independent of both $\{i_t^*\}$ and $\mathcal{Z}_n$. The idea

[7]Strictly speaking, we can set $\pi_n = 1$ since the innovations in our model come from an independently distributed sequence and the serial correlations in the conditioning variables do not affect the limit distribution. This added generality, however, may mitigate the effect of such serial correlations coming from the approximation errors in finite samples.

of the smooth stationary bootstrap is to construct the bootstrap sample as follows: for $t = 1, , \ldots, n$ let

$$v_t^* = v_{i_t^*} + a_n \eta_t,$$

where $a_n \to 0$ as $n \to \infty$ is a smoothing parameter. Neumeyer's work is for i.i.d. data and we extend it to dependent samples by combining it with the stationary bootstrap.

Remarks on smooth stationary bootstrap:

1. Another way to describe the smooth stationary bootstrap scheme is that, when $v_{t-1}^* = v_{i_{t-1}^*}$, v_t^* is determined as $a_n \eta_t$ plus a random draw from $\mathcal{Z}_n$ with probability π_n or the next observation $v_{i_{t-1}^*+1}$ of v_t^* in $\mathcal{Z}_n$ with probability $1 - \pi_n$. That is,

$$v_t^* = \begin{cases} v_{i_t} + a_n \eta_t & \text{with probability } \pi_n \\ v_{i_{t-1}^*+1} + a_n \eta_t & \text{with probability } 1 - \pi_n. \end{cases}$$

2. Note that the (stationary) marginal distribution of i_{t-1}^* is Uniform$\{1, \ldots, n\}$ and that $\left\{i_{t-1}^*\right\}$ is serially dependent.

3. Suppose that v_t is univariate. According to Politis and Romano (1994) and Goncalves and Politis (2011), this bootstrap sample's distribution, conditional on the original sample, is also a strictly stationary Markov chain. The conditional distribution $G^*(s) := \Pr\left\{v_t^* \leq s | \mathcal{Z}_n\right\}$ of v_t^*, conditioning on the original observation $\mathcal{Z}_n$, is given by

$$G^*(s) = \frac{1}{n} \sum_{j=1}^{n} G\left(\frac{s - v_j}{a_n}\right).$$

4. The choice of smoothing parameter π_n is a challenging issue. Politis and Romano (1994) showed that the optimal rate in terms of the mean squared error of the sample mean is $n^{-1/3}$, but suggested $\pi_n = 1/b$ where b is the block length in the moving block bootstrap.

5. The moment of $1\left\{v_t^* \leq s\right\}$ and the inner products of $1\left\{v_{t_1}^* \leq s_1\right\}$ and $1\left\{v_{t_2}^* \leq s_2\right\}$ converges to the corresponding moment and inner products of $1\left\{v_t \leq s\right\}$ and $1\left\{v_{t_1} \leq s_1\right\}$ and $1\left\{v_{t_2} \leq s_2\right\}$, respectively, as $a_n \to 0$.

6. The weighting constant a_n is similar to the smoothing parameter in the kernel density estimation. Neumeyer (2009) employed $a_n = n^{-1/4}$ and $0.5n^{-1/4}$.

A.3 Residual Empirical process from high-dimensional time series regression

This section can be read independently and extends Mammen (1996) and Chen and Lockhart's (2001) residual empirical process with many parameters results to the setup of time series regression with conditionally heteroskedastic errors. The cases for residuals from parametric linear or nonlinear models have been considered by Koul (1970) and Loynes (1980). The extension to the time series case was made by Boldin (1983), Bai (1994), Ling (1998) and Schick and Wefelmeyer (2002). For the empirical distribution of non-parametrically or semi-parametrically estimated residuals from cross sectional regression models, see Akritas and Van Keilegom (2001) for instance. The challenge in our case where we have dependent observations and a growing numbers of parameters lies in the limited availability of a proper maximal inequality.

Specifically, we consider

$$y_t = x_t^\intercal \beta_0 + r_{gt} + e_t, \quad e_t = \sigma_t \varepsilon_t \tag{22}$$
$$\sigma_t = w_t^\intercal \gamma_0 + r_{\sigma t}$$

where $\{x_t, w_t\}$ and $\{\varepsilon_t\}$ can be serially correlated but mutually independent of each other and r_{gt} and $r_{\sigma t}$ are approximation errors. Compared to our main model characterized in (3), (4), and (5), this model imposes the sparsity and the regression is performed only for those relevant variables. Recall the notation that $x_t = X_{t,S}$ and $w_t = X_{t,S_\gamma}$.

Let the OLS residual be denoted by $\widehat{e}_t = y_t - x_t^\intercal \widehat{\beta}$, where $\widehat{\beta}$ is the OLS estimate. Next, the unknown parameter γ_0 can be estimated by regressing $|\widehat{e}_t|$ on w_t (the skedastic regression). Then, let $\widehat{\sigma}_t = w_t^\intercal \widehat{\gamma}$ and introduce the scaled residual $\widehat{\varepsilon}_t = \widehat{\sigma}_t^{-1} \widehat{e}_t$, and the (scaled) residual empirical process

$$\widehat{\mathbb{Z}}_n(\tau) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(1\left\{ \widehat{\varepsilon}_t \leq \tau \right\} - F(\tau) \right). \tag{23}$$

This process has some distinct features from previous works in that it allows for the conditional heteroskedasticity of unknown form and the time series dependence.

We define the infeasible empirical process of the unobservable error term ε_t,

$$\mathbb{Z}_n(\tau) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(1\left\{ \varepsilon_t \leq \tau \right\} - F(\tau) \right).$$

We also define a correction term that is due to the estimation error in the first step:

$$\widetilde{B}_n(\tau) = f(\tau) \mu_x^\intercal \sqrt{n} \left(\widehat{\beta} - \beta \right) - \tau f(\tau) \mu_w^\intercal \sqrt{n} \left(\widehat{\gamma} - \gamma \right),$$

where $\left(\mu_x^\intercal, \mu_w^\intercal\right)^\intercal = E\sigma_t^{-1}\left(x_t^\intercal, w_t^\intercal\right)^\intercal$. It can be shown that (see e.g. Theorem 2.8.3 in van der Vaart and Wellner 1996)

$$\left(\begin{array}{c} \mathbb{Z}_n(\tau) \\ \widetilde{B}_n(\tau) \end{array}\right) \Longrightarrow \left(\begin{array}{c} \mathbb{Z}(\tau) \\ f(\tau) Z_1 - \tau f(\tau) Z_2 \end{array}\right),$$

where $\Rightarrow$ signifies the weak convergence as introduced ahead of Theorem 2 and $\mathbb{Z}(\tau)$ and $Z = (Z_1, Z_2)^\intercal$ are a centered Gaussian process and a centered bivariate Gaussian vector, respectively, whose covariances are characterized by

$$\begin{aligned} E\mathbb{Z}(\tau_1)\mathbb{Z}(\tau_2) &= \operatorname{cov}\left(1\{\varepsilon_t \le \tau_1\}, 1\{\varepsilon_t \le \tau_2\}\right) \\ E\mathbb{Z}(\tau)Z^\intercal &= \lim_{n\to\infty} E\left[\left(\widetilde{x}_t\varepsilon_t, \widetilde{w}_t\left(|\varepsilon_t| - 1\right)\right) 1\{\varepsilon_t \le \tau\}\right] \\ EZZ^\intercal &= \lim_{n\to\infty} E\left[\begin{array}{cc} \widetilde{x}_t^2\varepsilon_t^2, & \widetilde{x}_t\varepsilon_t\widetilde{w}_t\left(|\varepsilon_t| - 1\right) \\ \cdot & \widetilde{w}_t^2\left(|\varepsilon_t| - 1\right)^2 \end{array}\right], \end{aligned}$$

where $\widetilde{x}_t = \mu_x^\intercal\left(Ex_tx_t^\intercal\right)^{-1} x_t$ and $\widetilde{w}_t = \mu_w^\intercal\left(Ew_tw_t^\intercal\right)^{-1} w_t\sigma_t$.

Theorem 4 *Suppose Assumption 1 holds. Then,*

$$\sup_{\tau\in\mathbb{R}}\left|\widehat{\mathbb{Z}}_n(\tau) - \mathbb{Z}_n(\tau) - \widetilde{B}_n(\tau)\right| = o_p(1) \tag{24}$$

$$\sup_{\tau\in\mathbb{R}}\left|\int_{-\infty}^{\tau}\widehat{\mathbb{Z}}_n - \int_{-\infty}^{\tau}\mathbb{Z}_n - \int_{-\infty}^{\tau}\widetilde{B}_n\right| = o_p(1). \tag{25}$$

Furthermore,

$$\widehat{\mathbb{Z}}_n(\tau) \Longrightarrow \mathbb{Z}(\tau) + f(\tau) Z_1 - \tau f(\tau) Z_2. \tag{26}$$

The conditions in Assumption 1.1 can be weakened so that the random vector q_t is strictly stationary β-mixing with mixing coefficient of order $\beta_m = O(\rho^m)$ for some $0 < \rho < 1$ and has bounded fourth moments. The rather strong tail condition is to control the behavior of the lasso estimator for the high-dimensional regression, which can be weakened when we do not need selection using the lasso. Note that it is not clear if there exists a weak limit of $\int_{-\infty}^{\tau}\mathbb{Z}_n$ as the natural semi-metric of L_2-norm does not make the functions space of this empirical process totally bounded due to the integration. And this difficulty can be mitigated later when we consider the differences for the second order stochastic dominance.

A.3.1 Auxiliary Lemmas

We require certain maximal inequalities to show Theorem 4. Specifically, we employ the maximal inequality developed by Doukhan, Massart and Rio (1995). It builds on some high-level assumptions and indeterminate quantities that need to be verified and computed carefully. To

state the theorem, it is useful to have some definitions. Let $\mathcal{F}_a^b$ denote the sigma field generated by a sequence of given random variables $\{\zeta_a, \ldots, \zeta_b\}$. Define the mixing coefficient $\beta_m = 2^{-1} \sup \sum_{(i,j)\in I\times J} |P\{A_i \cap B_j\} - P\{A_i\} P\{B_j\}|$, where the supremum is taken over all finite partitions $\{A_i, i \in I\}$ that is $\mathcal{F}_{-\infty}^0$ measurable and $\{B_j, j \in J\}$ that is $\mathcal{F}_m^\infty$ measurable. Introduce a norm of a random function $g(\zeta_t)$

$$\|g\|_{2,\beta} = \sqrt{\int_0^1 \beta^{-1}(u)\mathfrak{Q}_g(u)^2 du},$$

where $\beta^{-1}(u)$ is the cadlag inverse of the β-mixing coefficients and $\mathfrak{Q}_g(u)$ is the inverse of the tail probability function $z \mapsto P\{|g| > z\}$. The function $\mathfrak{Q}_g(u)$, called the quantile function in Doukhan, Massart and Rio (1995), is different from the familiar quantile function $u \mapsto \inf\{x : u \leq P\{|g(\zeta_t)| \leq x\}\}$. Also let

$$\mathcal{G}_{n,\delta}^\beta = \left\{g : \|g\|_{2,\beta} < \delta\right\},$$

and $G_{n,\delta}$ denote an envelope of $\mathcal{G}_{n,\delta}^\beta$. In comparison, we use $\|\cdot\|_p$ to denote the standard L_p-norm for random variables. Now, we reiterate their Theorem 3 in the following.

Theorem 5 *Let $\{\zeta_t\}$ be a strictly stationary and absolutely regular process with β-mixing coefficient $\beta_m = O(\rho^m)$ for some $0 < \rho < 1$. Then, there exists a positive constant C, which depends only on $\int_0^1 \beta^{-1}(u)\,du$, such that*

$$E \sup_{g\in\mathcal{G}_{n,\delta}^\beta} |\frac{1}{\sqrt{n}} \sum_{t=1}^n (g(\zeta_t) - Eg(\zeta_t))| \leq C[1 + \delta^{-1}\mathfrak{q}_{G_{n,\delta}}(\min\{1, v_n(\delta)\})]\varphi_n(\delta), \tag{27}$$

where

$$\mathfrak{q}_{G_{n,\delta}}(v) = \sup_{u\leq v} \mathfrak{Q}_{G_{n,\delta}}(u) \sqrt{\int_0^u \beta^{-1}(\widetilde{u})d\widetilde{u}}$$

with the envelope function $G_{n,\delta}$ of $\mathcal{G}_{n,\delta}^\beta$, and $v_n(\delta)$ is the unique solution of

$$\frac{v_n(\delta)^2}{\int_0^{v_n(\delta)} \beta^{-1}(\widetilde{u})d\widetilde{u}} = \frac{\varphi_n(\delta)^2}{n\delta^2}, \tag{28}$$

and where

$$\varphi_n(\delta) = \int_0^\delta \sqrt{\log N_{[]}(\nu, \mathcal{G}_{n,\delta}^\beta, \|\cdot\|_{2,\beta})}d\nu.$$

Note that $u^{-1}\int_0^u \beta^{-1}(\widetilde{u})d\widetilde{u}$ is bounded away from zero and bounded above by the condition that $\beta_m = O(\rho^m)$ and $\epsilon < \rho < 1 - \epsilon$.

Next, we derive a lemma that verifies the high-level conditions in the above theorem and compute the explicit formula for the bound on the right of the inequality in (27).

Lemma 1 *Consider functions* $g_b(\cdot,\cdot):\mathbb{R}\times\mathbb{R}^s\to\mathbb{R}$, *indexed by* $b=\left(b_1,b_2^\intercal\right)^\intercal$, *such that* $g_b(\varepsilon,\zeta)=1\left\{\varepsilon-b_1-b_2^\intercal\zeta\le 0\right\}-1\left\{\varepsilon\le b_1\right\}$ *and a collection of such functions*

$$\mathcal{G}_{n,\delta}=\left\{g_b(\cdot,\cdot):b_1\in\mathbb{R},|b_2|\le\delta\right\}$$

for some finite C *and* $e^{-\sqrt{s}}<\delta<s^{-1}\epsilon$. *Let* $\{\varepsilon_t,\zeta_t\}$ *satisfy the mixing condition of being a strictly stationary and absolutely regular process with* β*-mixing coefficient* $\beta_m=O(\rho^m)$ *for some* $\epsilon\le\rho\le 1-\epsilon$ *and the moment conditions of* $\inf_{\theta:|\theta|=1}E\left|\zeta_t^\intercal\theta\right|>0$ *and* $\sup_j E\zeta_{jt}^2<\infty$. *Then,*

$$E\sup_{g\in\mathcal{G}_{n,\delta}}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^n\left(g(\varepsilon_t,\zeta_t)-Eg(\varepsilon_t,\zeta_t)\right)\right|\le C\left(1+\frac{\sqrt{s\log s}-\log\delta}{s^{1/4}\sqrt{\delta n}}\right)s^{1/4}\sqrt{\delta}\left(\sqrt{s\log s}-\log\delta\right).$$

Proof. We apply Theorem 5. Since the supremum in the theorem is over $\mathcal{G}_{n,\delta}^\beta$ while it is over $\mathcal{G}_{n,\delta}$ in this lemma, we first need to establish the relation between different norms. To this end, it was shown by Doukhan et al. (1995) and Seo and Otsu (2018) that for bounded random variables obeying the mixing condition of the lemma,

$$\|\cdot\|_2\le\|\cdot\|_{2,\beta}\le C\|\cdot\|_2\sup_{u\in[0,1]}u^{-1}\int_0^u\beta^{-1}(\widetilde{u})d\widetilde{u} \tag{29}$$

where $C=\sup_{u\in[0,1]}u^{-1}\int_0^u\beta^{-1}(\widetilde{u})d\widetilde{u}$, which is bounded for all $\rho\in[\epsilon,1-\epsilon]$. Hereafter, C, C_1,c, etc denote generic positive finite constants. Note that for any $g\in\mathcal{G}_n$,

$$\begin{aligned}\|g\|_{2,\beta}^2\le C\|g\|_2^2&=CE\left|1\left\{\varepsilon_t\le b_1+\zeta_t^\intercal b_2\right\}-1\left\{\varepsilon_t\le b_1\right\}\right|^2\\&=CE\left|F\left(b_1+\zeta_t^\intercal b_2\right)-F(b_1)\right|\le C_1E\left|\zeta_t^\intercal b_2\right|\le C_2E|\zeta_t||b_2|\le c\sqrt{s}|b_2|,\end{aligned}$$

where the constants depend only on the distribution of the sample and the inequalities follow from the boundedness of f, Cauchy Schwarz inequality, and the fact that $E|\zeta_t|\le\sqrt{E|\zeta_t|^2}\le\sqrt{s\max_j E\zeta_{jt}^2}$, respectively. This implies that

$$\mathcal{G}_{n,\delta}\subset\mathcal{G}_{n,Cs^{1/4}\delta^{1/2}}^\beta \tag{30}$$

because $\|g\|_{2,\beta}\le cs^{1/4}\delta^{1/2}$ for any b such that $|b|\le\delta$.

Now the bound in (27) involves the bracketing numbers in terms of $\|\cdot\|_{2,\beta}$-norm but the norm $\|\cdot\|_{2,\beta}$ is not convenient to compute the bracketing numbers. However, based on the norm relation we have just derived in (29), we may proceed it with $\|\cdot\|_2$ .It is useful to note that for any $\bar{b}$, and

$\eta > 0$, $\widetilde{\zeta}_t = \left(1, \zeta_t^\intercal\right)^\intercal$,

$$
\begin{aligned}
& E \sup_{b:\left|b_2-\bar{b}_2\right|<\eta,\left|F(b_1)-F\left(\bar{b}_1\right)\right|<\eta} \left|1\left\{\varepsilon_t \leq \widetilde{\zeta}_t^\intercal b\right\} - 1\left\{\varepsilon_t \leq \widetilde{\zeta}_t^\intercal \bar{b}\right\}\right|^2 \\
& \leq E \sup_{\left|b_2-\bar{b}_2\right|<\eta} 1\left\{F^{-1}\left(-\eta + F\left(\bar{b}_1\right)\right) < \varepsilon_t - \zeta_t^\intercal b_2 \leq F^{-1}\left(\eta + F\left(\bar{b}_1\right)\right)\right\} \\
& + E1\left\{\widetilde{\zeta}_t^\intercal \bar{b} - \left|\zeta_t\right|\eta < \varepsilon_t \leq \widetilde{\zeta}_t^\intercal \bar{b} + \left|\zeta_t\right|\eta\right\} \\
& \leq 2E\left(F\left(\zeta_t^\intercal \bar{b}_2 + \left|\widetilde{\zeta}_t\right|\eta + F^{-1}\left(\eta + F\left(\bar{b}_1\right)\right)\right) - F\left(\zeta_t^\intercal \bar{b}_2 - \left|\widetilde{\zeta}_t\right|\eta + F^{-1}\left(-\eta + F\left(\bar{b}_1\right)\right)\right)\right) \\
& \leq 2C\left(E\left|\widetilde{\zeta}_t\right| + 1\right)\eta, \qquad (31)
\end{aligned}
$$

where the last inequality is due to the boundedness of the derivative of F and an expansion that

$$
F^{-1}\left(\eta + F\left(\bar{b}_1\right)\right) - F^{-1}\left(-\eta + F\left(\bar{b}_1\right)\right) \leq 2\eta f\left(\bar{b}_1 + \eta\right)^{-1}
$$

and that

$$
\begin{aligned}
& \left(F\left(\zeta_t^\intercal \bar{b}_2 + \left|\widetilde{\zeta}_t\right|\eta + F^{-1}\left(\eta + F\left(\bar{b}_1\right)\right)\right) - F\left(\zeta_t^\intercal \bar{b}_2 - \left|\widetilde{\zeta}_t\right|\eta + F^{-1}\left(-\eta + F\left(\bar{b}_1\right)\right)\right)\right) \\
& \leq f\left(\zeta_t^\intercal \bar{b}_2 + \left|\widetilde{\zeta}_t\right|\eta + F^{-1}\left(\eta + F\left(\bar{b}_1\right)\right)\right)\left(2\left|\widetilde{\zeta}_t\right|\eta + 2\eta f\left(\bar{b}_1 + \eta\right)^{-1}\right)
\end{aligned}
$$

when $\left(\zeta_t^\intercal b_2 + \left|\widetilde{\zeta}_t\right|\eta + F^{-1}\left(\eta + F\left(\bar{b}_1\right)\right)\right) > \left(\zeta_t^\intercal \bar{b}_2 - \left|\widetilde{\zeta}_t\right|\eta + F^{-1}\left(-\eta + F\left(\bar{b}_1\right)\right)\right)$, while the other case is handled similarly. Recall that $E\left|\zeta_t\right| = O\left(\sqrt{s}\right)$. Then, let $d\left(b, b_j\right) = \left|F\left(b_1\right) - F\left(b_{1j}\right)\right| + \left|b_2 - b_{2j}\right|$ and consider the brackets with upper and lower bounds $1\left\{\varepsilon_t \leq \widetilde{\zeta}_t^\intercal b_j\right\} \pm B_j\left(\varepsilon_t, \widetilde{\zeta}_t\right)$, where

$$
B_j\left(\varepsilon_t, \widetilde{\zeta}_t\right) = \sup_{b:d(b,b_j)<\nu^2/C\sqrt{s}} \left|1\left\{\varepsilon_t \leq \widetilde{\zeta}_t^\intercal b\right\} - 1\left\{\varepsilon_t \leq \widetilde{\zeta}_t^\intercal b_j\right\}\right|
$$

for a sequence of b_js and some finite C_2. Then, L_2-norms of these brackets are bounded by ν due to (31). In particular, we choose b_js to be the centers of the set of hypercubes of equal side length $\eta_1 = 2\nu^2/C_2 s$ with circumradius $\nu^2/C_2\sqrt{s}$ that partition $[0,1] \times [-C_1\delta, C_1\delta]^s$. Here recall that the circumradius of the hypercube of side lengths $\left(a_1, \ldots, a_{s+1}\right)^\intercal$ is $\sqrt{\sum_{j=1}^{s+1} a_j^2}/2$. The number of such hypercubes is $\left(2/\eta_1\right)^{s+1} C\left(C_1\delta\right)^s = \left(C_2 s/\nu^2\right)^{s+1} C\left(C_1\delta\right)^s$. Thus, a direct algebra using the indefinite integral formula $\int \log x dx = \text{const.} + x(\log x - 1)$ yields that

$$
\begin{aligned}
\varphi_n(\delta) &\leq \int_0^\delta \sqrt{\log\left(\left(C_2 s/\nu^2\right)^{s+1} C\left(C_1\delta\right)^s\right)} d\nu \\
&\leq \delta\left(\sqrt{(s+1)\left(C_3 + \log s\right)} - 2\log\delta\right),
\end{aligned}
$$

for some $C_3 < \infty$.

Given the mixing condition, v_n that solves (28) satisfies

$$v_n(\delta) \leq C\frac{\varphi_n(\delta)^2}{n\delta^2} \leq C\frac{\left(C+\sqrt{(s+1)\log s}-2\log\delta\right)^2}{n} \leq 1.$$

Then, noting that $G_{n,\delta}$ is bounded and $\sqrt{\int_0^u \beta^{-1}(\widetilde{u})d\widetilde{u}} \leq Cu$ uniformly, we conclude that

$$\mathfrak{q}_{G_{n,\delta}}(v_n(\delta)) \leq C\sqrt{v_n(\delta)} \leq \frac{\sqrt{s\log s}-2\log\delta}{\sqrt{n}}.$$

Putting all these together yields the bound in (27) as

$$C\left(1+\frac{\sqrt{s\log s}-2\log\delta}{\delta\sqrt{n}}\right)\delta\left(\sqrt{s\log s}-2\log\delta\right). \tag{32}$$

Finally, revoking the relation in (30), we replace δ in (32) with $cs^{1/4}\delta^{1/2}$. ■

Next, we give a maximal inequality for the SSD statistic. The bound here are not equivalent to the preceding lemma due to the unboundedness of the class of functions and a different entropy number.

Lemma 2 *Consider a collection of functions*

$$\mathcal{G}_{n,\delta} = \left\{g_b(\varepsilon,\zeta) = \left(\varepsilon - b_1 - b_2^{\intercal}\zeta \leq 0\right)_+ - \left(\varepsilon \leq b_1\right)_+ : b_1 \in \mathbb{R}, |b_2| \leq \delta\right\},$$

for some $e^{-\sqrt{s}} < \delta < s^{-1}\epsilon$. Let $\{\varepsilon_t, \zeta_t\}$ satisfy the mixing condition of being a strictly stationary and absolutely regular process with β-mixing coefficient $\beta_m = O(\rho^m)$ for some $\epsilon \leq \rho \leq 1-\epsilon$ and the moment conditions of $\inf_{\theta:|\theta|=1} E\left|\zeta_t^{\intercal}\theta\right| > 0$ and $\sup_j E\zeta_{jt}^2 < \infty$. Then,

$$E\sup_{g\in\mathcal{G}_{n,\delta}}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(g(\varepsilon_t,\zeta_t) - Eg(\varepsilon_t,\zeta_t)\right)\right| \leq C\left(1+\frac{\sqrt{s\log s}-2\log s^{1/2}\delta}{s^{1/2}\delta\sqrt{n}}\right)s^{1/2}\delta\left(\sqrt{s\log s}-2\log s^{1/2}\delta\right).$$

Proof. This proof highlights the difference from that of the preceding Lemma 1. First, in this case, our class of functions are unbounded and thus (29) and (30) need to be modified. Due to Lemmas 1 and 2 of Doukhan et al (1995), we have

$$\|g\|_2 \leq \|g\|_{2,\beta} \leq \|g\|_{\phi,2}\sqrt{1+\int_0^1 \phi^*\left(\beta^{-1}(u)\right)du}, \tag{33}$$

where $\phi^*(y) = \sup_{x>0}\left[xy - \phi(x)\right]$ for a given function ϕ, $\|g\|_{\phi,2} = \inf\left\{c>0 : E\phi\left(|g/c|^2\right) \leq 1\right\}$. This is the Orlicz norm and is equivalent to $\|g\|_{2p}$ when $\phi(x) = x^p$ and $p > 1$. Furthermore, when $\phi(x) = x^p$, we have $\int_0^1 \phi^*\left(\beta^{-1}(u)\right)du = C_p\sum_{n=1}^{\infty} n^{\frac{1}{p-1}}\beta_n$, where C_p is a constant dependent only on p, see Doukhan et al (1995 p. 404). Also

$$E\left|\left(\varepsilon_t - \widetilde{\zeta}_t^{\intercal}b\right)_+ - \left(\varepsilon_t - b_1\right)_+\right|^{2p} \leq E\left|\zeta_t^{\top}b_2\right|^{2p} = O\left(s^p\right)|b_2|^{2p},$$

where $E\left|\widetilde{\zeta}_t\right|^{2p} = O\left(s^p\right)$ by Jensen's inequality. This yields that

$$\mathcal{G}_{n,\delta} \subset \mathcal{G}^{\beta}_{C\delta s^{1/2}}. \tag{34}$$

Next, to compute the entropy with L_{2p}-norm due to (33), note that for any $\bar{b}$, and $\eta > 0$, $\widetilde{\zeta}_t = \left(1, \zeta_t^{\intercal}\right)^{\intercal}$,

$$\begin{aligned}
&2^{1-2p} E \sup_{b:\left|b_2-\bar{b}_2\right|<\eta,\left|F(b_1)-F\left(\bar{b}_1\right)\right|<\eta} \left|\left(\varepsilon_t - \widetilde{\zeta}_t^{\intercal} b\right)_+ - \left(\varepsilon_t - b_1\right)_+ - \left(\left(\varepsilon_t - \widetilde{\zeta}_t^{\intercal} \bar{b}\right)_+ - \left(\varepsilon_t - \bar{b}_1\right)_+\right)\right|^{2p} \\
&\leq E \sup_{b:\left|b_2-\bar{b}_2\right|<\eta,\left|F(b_1)-F\left(\bar{b}_1\right)\right|<\eta} \left|\left(\varepsilon_t - \widetilde{\zeta}_t^{\intercal} b\right)_+ - \left(\varepsilon_t - b_1\right)_+ - \left(\left(\varepsilon_t - \zeta_t^{\intercal} b_2 - \bar{b}_1\right)_+ - \left(\varepsilon_t - \bar{b}_1\right)_+\right)\right|^{2p} \\
&+ E \sup_{b:\left|b_2-\bar{b}_2\right|<\eta,\left|F(b_1)-F\left(\bar{b}_1\right)\right|<\eta} \left|\left(\left(\varepsilon_t - \zeta_t^{\intercal} b_2 - \bar{b}_1\right)_+ - \left(\varepsilon_t - \bar{b}_1\right)_+\right) - \left(\left(\varepsilon_t - \widetilde{\zeta}_t^{\intercal} \bar{b}\right)_+ - \left(\varepsilon_t - \bar{b}_1\right)_+\right)\right|^{2p} \\
&\leq E\left|\zeta_t\right|^{2p} \left(\left|\bar{b}_2\right| + \eta\right)^{2p} 1\left\{F^{-1}\left(-\eta + F\left(\bar{b}_1\right)\right) < \varepsilon_t \leq F^{-1}\left(\eta + F\left(\bar{b}_1\right)\right)\right\} \\
&+ E\left(\left|\zeta_t\right| \eta\right)^{2p} \\
&\leq E\left|\widetilde{\zeta}_t\right|^{2p} \left(\eta^{2p} + \eta\left|\bar{b}_2\right|^{2p}\right).
\end{aligned}$$

Recall that $E\left|\widetilde{\zeta}_t\right|^{2p} = O\left(s^p\right)$. Then, η_1 in the proof of Lemma 1 can be set as either $\eta_1 = \left(2\nu/\left(C_2\delta\sqrt{s}\right)\right)^{2p}$ or $\eta_1 = \nu/\sqrt{s}$. And proceed along the proof with $\eta_1 = \left(2\nu/\left(C_2\delta\sqrt{s}\right)\right)^{2p}$ to note that

$$\begin{aligned}
\varphi_n(\delta) &\leq \int_0^{\delta} \sqrt{\log\left(\left(\nu^2/s\right)^{-s/p} \delta^{2s}\right)} d\nu \\
&\leq \delta\left(\sqrt{\left(s/p\right)\left(C_3 + \log s\right)} - 2\log\delta\right),
\end{aligned}$$

which is equivalent to the bound for $\varphi_n(\delta)$ in the proof of Lemma 1, meaning that we get the bound corresponding to (32), which is

$$C\left(1 + \frac{\sqrt{s\log s} - 2\log\delta}{\delta\sqrt{n}}\right)\delta\left(\sqrt{s\log s} - 2\log\delta\right).$$

Finally, revoking the relation in (34), we replace δ in (33) with $cs^{1/2}\delta$ to achieve the following bound

$$C\left(1 + \frac{\sqrt{s\log s} - 2\log s^{1/2}\delta}{s^{1/2}\delta\sqrt{n}}\right)s^{1/2}\delta\left(\sqrt{s\log s} - 2\log s^{1/2}\delta\right).$$

The case with $\eta_1 = \nu/\sqrt{s}$ is analogous and the details are omitted. ■

Next, we state some standard results for the sake of later reference.

Lemma 3

$$\left|\widehat{\beta} - \beta_0\right|_1 = O_p\left(s\sqrt{\frac{\log s}{n}}\right) \quad \left|\widehat{\beta} - \beta_0\right|_2 = O_p\left(\sqrt{\frac{s\log s}{n}}\right).$$

Proof. Note that the Hölder inequality yields that

$$\left|\widehat{\beta}-\beta_0\right|_1 \leq \left|\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t^\top\right)^{-1}\right|_\infty \left|\frac{1}{n}\sum_{t=1}^{n} x_t e_t\right|_1$$

and $\left|\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t^\top\right)^{-1}\right|_\infty = O_p(1)$ as the minimum eigenvalue of $\frac{1}{n}\sum_{t=1}^{n} x_t x_t^\top$ is bounded away from zero with probability approaching to one due to Lemma 8, while $\left|\frac{1}{n}\sum_{t=1}^{n} x_t e_t\right|_\infty \leq \left|\frac{1}{n}\sum_{t=1}^{n} x_t r_{gt}\right|_\infty + \left|\frac{1}{n}\sum_{t=1}^{n} x_t \sigma_t \varepsilon_t\right|_\infty = O_p(\log s/\sqrt{n})$ due to Assumption 1 and Lemma 6. Then, $\left|\frac{1}{n}\sum_{t=1}^{n} x_t e_t\right|_1 \leq s\left|\frac{1}{n}\sum_{t=1}^{n} x_t e_t\right|_\infty$. We may proceed similarly for $\left|\widehat{\beta}-\beta_0\right|_2$ using the Cauchy Schwarz inequality and bounding $\left|\frac{1}{n}\sum_{t=1}^{n} x_t e_t\right|_2 \leq \sqrt{s\left|\frac{1}{n}\sum_{t=1}^{n} x_t e_t\right|_\infty^2} = O_p\left(\sqrt{s\log s/n}\right)$ as above. ■

Lemma 4 *For $\tau \in \mathbb{R}$,*

$$\sqrt{n}\left(\mathbb{Z}_n(\tau), \mu\left(\widehat{\beta}-\beta_0\right), \mu_x^\intercal\left(\widehat{\gamma}-\gamma_0\right)\right)^\top \Longrightarrow \left(\mathbb{Z}(\tau), Z^\top\right)^\top,$$

which is a centered normal variate with

$$\begin{aligned} E\mathbb{Z}(\tau_1)\mathbb{Z}(\tau_2) &= \operatorname{cov}\left(1\left\{\varepsilon_t \leq \tau_1\right\}, 1\left\{\varepsilon_t \leq \tau_2\right\}\right) \\ E\mathbb{Z}(\tau)Z^\intercal &= \lim_{n\to\infty} E\left[\left(\widetilde{x}_1\varepsilon_1, \widetilde{w}_1\left(|\varepsilon_1|-1\right)\right)1\left\{\varepsilon_1 \leq \tau\right\}\right]. \end{aligned}$$

and

$$EZZ^\intercal = \lim_{n\to\infty} E\begin{bmatrix} \widetilde{x}_t^2\varepsilon_t^2, & \widetilde{x}_t\varepsilon_t\widetilde{w}_t\left(|\varepsilon_t|-1\right) \\ \cdot & \widetilde{w}_t^2\left(|\varepsilon_t|-1\right)^2 \end{bmatrix},$$

where $\widetilde{x}_t = \mu_x^\intercal\left(Ex_t x_t^\intercal\right)^{-1} x_t$ and $\widetilde{w}_t = \mu_w^\intercal\left(Ew_t w_t^\intercal\right)^{-1} w_t\sigma_t$.

Proof. This is straightforward by the central limit theorem and law of large numbers for stationary arrays, see e.g. Davidson (1994) while the uniform tightness of $\sqrt{n}\mathbb{Z}_n(\tau)$ is standard, see e.g. Theorem 2.8.3 in van der Vaart and Wellner (1996) as the class of indicator functions of half intervals is a V-C subgraph class of functions. ■

Now, we turn to the proof of the main theorem in this section.

A.3.2 Proof of Theorem 4

To begin with, it is useful to note that

$$\begin{aligned} \left\{\widehat{\varepsilon}_t \leq \tau\right\} &= \left\{\widehat{e}_t - \tau\widehat{\sigma}_t \leq 0\right\} \\ &= \left\{\varepsilon_t + \sigma_t^{-1}r_{gt} - \sigma_t^{-1}x_t^\intercal\left(\widehat{\beta}-\beta\right) - \tau\sigma_t^{-1}\widehat{\sigma}_t \leq 0\right\} \\ &= \left\{\varepsilon_t + \sigma_t^{-1}r_{gt} + \tau\sigma_t^{-1}r_{\sigma t} \leq \tau + \sigma_t^{-1}x_t^\intercal\left(\widehat{\beta}-\beta\right) + \sigma_t^{-1}w_t^\intercal\left(\widehat{\gamma}-\gamma\right)\tau\right\} \end{aligned}$$

and that,

$$\begin{aligned}
&\sup_{\tau,c}\left|\Pr\left\{\varepsilon_t+\sigma_t^{-1}r_{gt}+\tau\sigma_t^{-1}r_{\sigma t}\le c\right\}-F\left(c\right)\right| \\
&=\sup_{\tau,c}E\left|F\left(c-\sigma_t^{-1}r_{gt}-\tau\sigma_t^{-1}r_{\sigma t}\right)-F\left(c\right)\right| \\
&\le \bar{f}E\sup_{\tau}\left|\sigma_t^{-1}r_{gt}+\tau\sigma_t^{-1}r_{\sigma t}\right|=o\left(n^{-1/2}\right), \qquad (35)
\end{aligned}$$

where the first equality follows from the independence of ε_t from x_t, the inequality is due to the boundedness of the density of ε_t and the last equality is due to Assumption 1 and the boundedness of τ and σ_t^{-1}. For the same reason, $\Pr\left\{\sup_\tau\left|\sigma_t^{-1}r_{gt}+\tau\sigma_t^{-1}r_{\sigma t}\right|\le b_n^{-1}\right\}\to 1$ with $b_n=n^{1/2}\left(s\log s\right)^{-1/2}$.

Let $r_t=\sigma_t^{-1}r_{gt}+\tau\sigma_t^{-1}r_{\sigma t}$ and consider the following process

$$\mathbb{Z}_n(b,g,\tau)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(1\left\{\varepsilon_t\le\tau+\sigma_t^{-1}x_t^{\intercal}b+\sigma_t^{-1}w_t^{\intercal}g\tau+r_t\right\}-F(\tau)\right),$$

on $\Theta_n=\left\{|b|,|g|\le Cb_n^{-1},|\tau|\le C\right\}$ for a given $C<\infty$. Here the size of the index set reflects the rate, at which the estimators $\widehat{\beta}$ and $\widehat{\gamma}$ converges to the true value β_0 and γ_0. And note that $\mathbb{Z}_n\left(\widehat{\beta}-\beta_0,\widehat{\gamma}-\gamma_0,\tau\right)=\widehat{\mathbb{Z}}_n\left(\tau\right)$. Then, write

$$\begin{aligned}
\mathbb{Z}_n(b,g,\tau)&=\mathbb{Z}_n(b,g,\tau)-\mathbb{Z}_n(\tau)+\mathbb{Z}_n(\tau) \\
&=\mathbb{M}_n(b,g,\tau)-\mathbb{M}_n(\tau)+E\mathbb{Z}_n(b,g,\tau)+\mathbb{Z}_n(\tau), \qquad (36)
\end{aligned}$$

where $\mathbb{M}_n(b,g,\tau)=\mathbb{Z}_n(b,g,\tau)-E\mathbb{Z}_n(b,g,\tau)$ is an empirical process and $\mathbb{M}_n\left(\tau\right)=\mathbb{M}_n\left(0,0,\tau\right)$. Note that $\mathbb{Z}_n(\tau)=\mathbb{Z}_n(0,0,\tau)$ and $E\mathbb{Z}_n(\tau)=0$. First, we show that

$$E\sup_{b,g,\tau}\left|\mathbb{M}_n(b,g,\tau)-\mathbb{M}_n(0,0,\tau)\right|\to 0, \qquad (37)$$

where the supremum is over Θ_n. We can apply Lemma 1 and a truncation argument to verify (37). Note that for any C_{1n} and C_2,

$$\begin{aligned}
&\left\{1\left\{\varepsilon_t\le\tau+\sigma_t^{-1}x_t^{\intercal}b+\sigma_t^{-1}w_t^{\intercal}g\tau+r_t\right\}-1\left\{\varepsilon_t\le\tau\right\}:|\tau|\le C_{1n},|b|,|g|\le C_2b_n^{-1}\right\} \\
&\subset\mathcal{G}_n=\left\{g_{t\tau\theta}=1\left\{\varepsilon_t-\tau-\zeta_t^{\intercal}\theta\le 0\right\}-1\left\{\varepsilon_t\le\tau\right\}:|\tau|\le C_{1n},|\theta|\le b_n^{-1}C_2C_{1n}\right\},
\end{aligned}$$

where $\zeta_t=\left(\sigma_t^{-1}x_t^{\intercal},\sigma_t^{-1}w_t^{\intercal},r_t\right)^{\intercal}$. Then, apply Lemma 1 by setting δ proportional to $b_n^{-1}C_{1n}$ and $s=o\left(n^{1/2}\right)$. Then, the resulting bound becomes $O\left(n^{-1/4}s\left(\log s\right)^{3/4}\right)$, since $\frac{\sqrt{s\log s}-\log\delta}{s^{1/4}\sqrt{\delta n}}=o\left(1\right)$ for $s/\sqrt{n}=o\left(1\right)$. This allows for s order up to $n^{1/4}$ aside from the logarithmic factor. Furthermore, for $\tau<-C_{1n}$, for any $\varepsilon>0$, the following holds for sufficiently large n

$$\begin{aligned}
E\left(1\left\{\varepsilon_t-\tau-\zeta_t^{\intercal}\theta\tau\le 0\right\}-1\left\{\varepsilon_t\le\tau\right\}\right)&=E\left|F\left(\tau+\zeta_t^{\intercal}\theta\tau\right)-F\left(\tau\right)\right| \\
&\le\int f\left(\tau\left(1+\bar{z}\right)\right)zp\left(\frac{z}{a_n}\right)\frac{1}{a_n}dz \\
&\le f\left(-C_{1n}\left(1-\varepsilon\right)\right)a_n=o\left(n^{-1/2}\right), \qquad (38)
\end{aligned}$$

where $p(\cdot)$ denotes the maximum of densities of $\zeta_t^\top a$ over the unit vectors a and $a_n = b_n^{-1} C_{1n}$, by the change of variables. The argument for the case $\tau > C_{1n}$ is similar and thus omitted.

Next, the standard Taylor series expansion for $E\mathbb{Z}_n(b, g, \tau)$ yields that

$$\begin{aligned} E\mathbb{Z}_n(b, g, \tau) &= \frac{1}{\sqrt{n}} E \sum_{t=1}^{n} \left(F\left(\tau + \sigma_t^{-1} x_t^\intercal b + \sigma_t^{-1} w_t^\intercal g\tau - r_t\right) - F(\tau) \right) \\ &= \frac{1}{\sqrt{n}} E \sum_{t=1}^{n} f(\widehat{\tau}) \left(\sigma_t^{-1} x_t^\intercal b + \sigma_t^{-1} w_t^\intercal \tau g - r_t\right) \end{aligned}$$

for some value $\widehat{\tau}$ between τ and $\tau + \sigma_t^{-1} x_t^\intercal b + \sigma_t^{-1} w_t^\intercal g\tau + r_t$. And the boundedness of the first derivative of f means that

$$\begin{aligned} &\frac{1}{\sqrt{n}} \left| E \sum_{t=1}^{n} f(\widehat{\tau}) \sigma_t^{-1} x_t^\intercal b - f(\tau) E \sigma_t^{-1} x_t^\intercal b \right| \\ &\leq C \frac{1}{\sqrt{n}} E \sum_{t=1}^{n} \left(\left| \sigma_t^{-1} x_t^\intercal b + \sigma_t^{-1} w_t^\intercal g\tau - r_t \right| \left| \sigma_t^{-1} x_t^\intercal b \right| \right) = o(1) \end{aligned}$$

if $\sqrt{n} E r_t^2, n^{-1/2} E \sum_{t=1}^{n} \left(\sigma_t^{-1} w_t^\intercal g\right)^2$, and $n^{-1/2} E \sum_{t=1}^{n} \left(\sigma_t^{-1} x_t^\intercal b\right)^2$ are $o(1)$, by the Cauchy Schwarz inequality. However, $\sqrt{n} E r_t^2 = o(1)$ by assumption and $n^{-1/2} E \sum_{t=1}^{n} \left(\sigma_t^{-1} x_t^\intercal b\right)^2$ is bounded by $C n^{-1/2} E \sum_{t=1}^{n} \left(x_t^\intercal b\right)^2$ by the boundedness of σ_t^{-1}, which is $o(1)$ as $|b| \leq b_n^{-1} = n^{-1/2} (s \log s)^{1/2}$. By proceeding similarly for the term with $w_t^\intercal \tau g$ and due to (35), we can conclude

$$\left| E\mathbb{Z}_n(b, g, \tau) - f(\tau) E \left(\sigma_t^{-1} x_t^\intercal \sqrt{n} b + \sigma_t^{-1} w_t^\intercal \tau \sqrt{n} g\right) \right| = o(1),$$

uniformly in b, g,and τ on Θ_n. By setting $b = \left(\widehat{\beta} - \beta_0\right)$ and $g = (\widehat{\gamma} - \gamma_0)$ and applying Lemma 4 yields the weak limit (26).

The proof of (25) is analogous. We can redefine $\mathbb{Z}_n(b, g, \tau)$ in (36) as

$$\mathbb{Z}_n(b, g, \tau) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(\left(-\varepsilon_t + \tau + \sigma_t^{-1} x_t^\intercal b + \sigma_t^{-1} w_t^\intercal g\tau + r_t\right)_+ - E\left(-\varepsilon_t + \tau\right)_+ \right),$$

and proceed to show (37) by applying Lemma 2 similarly for the proof of (24) using Lemma 1 in there. The truncation argument is applied by showing $\mathbb{M}_n(b, g, \tau) - \mathbb{M}_n(0, 0, \tau) - (\mathbb{M}_n(b, g, C_{1n}) - \mathbb{M}_n(0, 0, C_{1n}))$ is $o_p(1)$ uniformly over $|\tau| > C_{1n}$ for the same reason as (38). ■

A.4 Weighted LASSO Regressions in Time Series

We derive oracle inequalities that show the variable selection property and deviation bounds for the parameter estimates and prediction. Despite the huge literature on the lasso, the results for the time series model are limited. Recently, Medeiros and Mendes (2016) studies the lasso estimate for the

dependent data but the bounds in there are not as sharp as those in this section. And Basu and Michailidis (2015) focuses on the Gaussian vector autoregression. Furthermore, no result is available for the lasso estimator of the skedastic regression.

Let $S(\beta)$ denote the support of a given vector β while $|S|$ denote the cardinality of an index set S. Recall that $\beta_S = (\beta_j : j \in S)$ denotes the $|S|$-dimensional subvector of a p-dimensional vector β. Also, recall that $x_t = X_{t,S}$ and $\widehat{x}_t = X_{t,\widehat{S}}$

Let $\widetilde{\beta} := \widehat{\beta}_{lasso}$ and $\widetilde{\gamma} := \widehat{\gamma}_{lasso}$ throughout this section to ease notation. Also let X, Y and $\mathtt{e}$ denote the matrices stacking $X_t^\intercal$, y_t, and ε_t, respectively. The results in this section holds under more general strong mixing (α-mixing) conditions than the absolutely regular conditions. See the conditions in Proposition 1 below for specifics.

A.4.1 Weighted LASSO with Dependent Data

Theorem 6 *Let Assumption 1, and 2 hold. Then, the followings hold with probability approaching one when λ satisfies the constraint (14):*
Part (i).

$$\frac{2}{n}\sum_{t=1}^{n}\left(X_t^\intercal\widehat{\beta}_{lasso} - g(q_t)\right)^2 + \lambda\left|\widehat{\beta}_{lasso} - \beta_0\right|_1 \leq 6\frac{1}{n}\sum_{t=1}^{n} r_{gt}^2 + \frac{48\lambda^2|S|}{\phi_\beta^2},$$

and if $\lambda s = o(\min\{|\beta_{0j}| : \beta_{0j} \neq 0\})$ in addition, then $\Pr\left(\widehat{\beta}_{lasso,j} \neq 0 : j \in S\right) \to 1$.
Part (ii). Furthermore, assume that $\lambda_{thr} = o(\min\{|\beta_{0j}| : \beta_{0j} \neq 0\})$. Then $\Pr\left(\widehat{S} = S\right) \to 1$ and

$$\left|\widehat{\beta}_{Tasso} - \beta_0\right|_2 = O_p\left(\sqrt{\frac{s_x \log s_x}{n}}\right).$$

Also, for any $a \neq 0$

$$\sqrt{n}a^\intercal\left(\widehat{\beta}_{Tasso} - \beta_0\right) \Longrightarrow \mathcal{N}\left(0, a_S M^{-1}\Omega M^{-1} a_S\right),$$

where $M = E\left(x_t x_t^\intercal\right)$ and $\Omega = E x_t x_t^\intercal e_t^2$.

Proof of Theorem 6 The proof of Part (i) consists of three lemmas, Lemma 5, 7, and 8. More specifically, Lemma 5 gives the deviation bound in the theorem conditional on the other two lemmas' conclusion. Here, we assume x_t's elements are scale normalized, i.e. $D = I$. The claim that $\Pr\left\{\widehat{S} \supset S\right\}$ then follows from the triangle inequality that $\left|\widehat{\beta}_i\right| = \left|\widehat{\beta}_i - \beta_{0i} + \beta_{0i}\right| \geq \left||\beta_{0i}| - \left|\widehat{\beta}_i - \beta_{0i}\right|\right| > 0$ under the uniform beta-min condition $\lambda s = o(\min\{|\beta_{0j}| : \beta_{0j} \neq 0\})$. Similarly in Part (ii) we can derive the perfect variable selection property of the thresholded lasso and under the perfect variable selection the asymptotic normality is standard. In fact, given the result in

Part (i), $\Pr\left\{\widehat{S} \supset S\right\}$, it remains to argue that $\widehat{\beta}_i$ is smaller than λ_{thr} if $\beta_{0i} = 0$ with probability approaching one. But this is obvious due to the deviation bound in Part (i) as $\left|\widehat{\beta} - \beta_0\right|_1 = O_p(\lambda s)$ and $\lambda s = o(\lambda_{thr})$. ■

Now, we present the promised lemmas.

Lemma 5 *Consider a sequence of events*

$$\mathcal{A}_n = \left\{8n^{-1}\left|\mathsf{e}^\intercal X\right|_\infty \leq \lambda\right\}. \tag{39}$$

Conditional on $\mathcal{A}_n$, we have

$$4n^{-1}\left|X\widetilde{\beta} - g\right|_2^2 + 3\lambda\left|\widetilde{\beta}_{S^c}\right|_1 \leq 4n^{-1}\left|r_y\right|_2^2 + 5\lambda\left|\widetilde{\beta}_S - \beta_0\right|_1.$$

Furthermore, if $n^{-1}X^\intercal X$ is compatible for S with compatibility constant $\phi = \phi_\beta/\sqrt{2}$, then

$$2n^{-1}\left|X\widetilde{\beta} - g\right|_2^2 + \lambda\left|\widetilde{\beta} - \beta_0\right|_1 \leq 6n^{-1}\left|r_y\right|_2^2 + 24\lambda^2|S|\,\phi^{-2}.$$

Proof. Since $\widetilde{\beta}$ is the minimizer,

$$\frac{1}{n}\left|Y - X\widetilde{\beta}\right|_2^2 + \lambda\left|\widetilde{\beta}\right|_1 \leq \frac{1}{n}\left|Y - X\beta_0\right|_2^2 + \lambda\left|\beta_0\right|_1.$$

Then, as $Y = X\beta_0 + r_y + \mathsf{e}$,

$$n^{-1}\left|X\widetilde{\beta} - g\right|_2^2 + \lambda\left|\widetilde{\beta}\right|_1 \leq n^{-1}\left|r_y\right|_2^2 + 2n^{-1}\mathsf{e}^\intercal X\left(\widetilde{\beta} - \beta_0\right) + \lambda\left|\beta_0\right|_1. \tag{40}$$

By the Hölder inequality, $n^{-1}\left|\mathsf{e}^\intercal X\left(\widetilde{\beta} - \beta_0\right)\right| \leq n^{-1}\left|\mathsf{e}^\intercal X\right|_\infty\left|\widetilde{\beta} - \beta_0\right|_1$. Then, recall the condition (39) of the lemma and proceed as in Bühlmann and van der Geer's (2011) Lemma 6.3 to obtain the first part of the lemma.

The second part of the lemma is Bühlmann and van der Geer's (2011) Theorem 6.2. ■

Next, we derive a maximal inequality for the strong (α-) mixing array to control the probability of the event $\mathcal{A}_n$ in Lemma 5 (39). Let α_m denote the strong mixing coefficient, which is the supremum of $\left|P\{G \cap H\} - P\{G\}P\{H\}\right|$ over every element $G \in \mathcal{F}_{-\infty}^0$ and $H \in \mathcal{F}_m^\infty$. The sequence α_m is bounded by the β-mixing coefficient β_m by construction.

The following simplified version of Merlevede et al's (2011) Proposition 2 serves as a building block to Lemma 6.

Proposition 1 *Let an array $\{\xi_t\}$ be a strictly stationary α-mixing array satisfying that for some positive h_1, h_2, b, and c,*

$$\begin{aligned}\alpha(m) &\leq \exp\left(-cm^{h_1}\right)\\ \Pr\{|\xi_t| > t\} &\leq H(t) := \exp\left(1 - (t/b)^{h_2}\right)\\ h_1^{-1} + h_2^{-1} &> 1.\end{aligned} \tag{41}$$

Also let $h = \left(h_1^{-1} + h_2^{-1}\right)^{-1}$, $M = b\left(1 + n^{h_1}\right)^{1/h_2}$ *and* $\xi_t^M = \xi_t 1\left\{|\xi_t| \leq M\right\} + M1\left\{|\xi_t| > M\right\}$. *Then, for any positive* $\ell < cn^{h_1(h-1)/h}$, *there exists a* $K < \infty$, *which does not depend on* ℓ, *such that*

$$\log E\left(\exp\left(\ell \sum_{t=1}^{n}\left(\xi_t^M - E\xi_t^M\right)\right)\right) \leq K\frac{n\ell^2}{1 - \ell n^{h_1(1-h)/h}}.$$

Here, K *is universal for a class of distributions with uniformly bounded* h_1, b, *and* c, *and* $h \leq 1 - \epsilon$ *for some positive* ϵ.

In applying this bound, it is worth noting that the permissible ℓ should not be too small and it is linked to the mixing decay rate and the tail probability through h_1 and h_2. Then,

Lemma 6 *Let* $\{\xi_{ti}\}$, $i = 1, \ldots, p$, *be* α*-mixing arrays satisfing the conditions in Proposition 1. Also assume that*

$$\log p = O\left(n^{h_1 \wedge (1 - 2h_1(1-h)/h)}\right). \tag{42}$$

Then,

$$\max_{i \leq p}\left|\sum_{t=1}^{n}\xi_{ti}\right| = O_p\left(\sqrt{n \log p}\right).$$

There are two upper bounds for p, in (42). Both are increasing functions of h_1 but the value of h_2 determines which is lower.

Proof. We begin with a maximal inequality for the sums of truncated variables. For any L, Jensen's inequality yields

$$E\max_{i \leq p}\left|\sum_{t=1}^{n}\left(\xi_{ti}^M - E\xi_{ti}^M\right)\right| \leq L\log\left[E\exp\left(L^{-1}\max_{i \leq p}\left|\sum_{t=1}^{n}\left(\xi_{ti}^M - E\xi_{ti}^M\right)\right|\right)\right].$$

Since $e^{|x|} \leq e^x + e^{-x}$ and the moment bound is uniform in i in Proposition 1 with $\ell = L^{-1}$,

$$\begin{aligned}
L\log\left[E\exp\left(L^{-1}\max_{i \leq p}\left|\sum_{t=1}^{n}\left(\xi_{ti}^M - E\xi_{ti}^M\right)\right|\right)\right] &= L\log\left[E\max_{i \leq p}\exp\left(L^{-1}\left|\sum_{t=1}^{n}\left(\xi_{ti}^M - E\xi_{ti}^M\right)\right|\right)\right] \\
&\leq L\log\left[2pE\exp\left(L^{-1}\sum_{t=1}^{n}\left(\xi_{ti}^M - E\xi_{ti}^M\right)\right)\right] \\
&\leq L\log(2p) + \frac{Kn}{L - n^{h_1(1-h)/h}} \\
&\leq (3 + K)\sqrt{n\log p},
\end{aligned}$$

where the last inequality follows by setting $L = n^{h_1(1-h)/h} + (n/\log 2p)^{1/2}$ due to (42).

For the remainder term, we show by the union bound for the maximum and Markov inequality that

$$\begin{aligned}\Pr\left\{\max_{i\leq p}\sum_{t=1}^{n}\left|\xi_{ti}-\xi_{ti}^{M}+E\xi_{ti}^{M}\right|>C\right\} &\leq p\Pr\left\{\sum_{t=1}^{n}\left|\xi_{ti}-\xi_{ti}^{M}+E\xi_{ti}^{M}\right|>C\right\}\\ &\leq C^{-1}npE\left|\xi_{t}-\xi_{t}^{M}+E\xi_{t}^{M}\right|\\ &\leq 2C^{-1}np\int_{M}^{\infty}H\left(x\right)dx\\ &=O\left(np\exp\left(-n^{h_1}\right)\right),\end{aligned}$$

where the last equality follows from the standard algebra, that is, the integral is bounded by $MH(M)=\exp\left(-n^{h_1}\right)$ in general and $M^{-1}H\left(M\right)$ if $h_2>1$ since $H\left(x\right)$ decays at an exponential rate.

Finally, note that

$$\begin{aligned}\max_{i\leq p}\left|\sum_{t=1}^{n}\xi_{ti}\right| &\leq \max_{i\leq p}\left|\sum_{t=1}^{n}\xi_{ti}^{M}-E\xi_{ti}^{M}\right|+\max_{i\leq p}\left|\sum_{t=1}^{n}\xi_{ti}-\left(\xi_{ti}^{M}-E\xi_{ti}^{M}\right)\right|\\ &=O_p\left(\sqrt{n\log p}\right)+O_p\left(np\exp\left(-n^{h_1}\right)\right)\end{aligned}$$

and recall that $\log p=O\left(n^{h_1}\right)$ from (42), which implies that $O_p\left(np\exp\left(-n^{h_1}\right)\right)=O_p\left(\sqrt{n\log p}\right)$. ■

The preceding maximal inequality leads to the following bounds on the probability of conditioning event $\mathcal{A}_n$.

Lemma 7 *Suppose* $\frac{\log(n\vee p)}{\lambda\sqrt{n}}=o\left(1\right)$. *Then,*

$$\Pr\left(4n^{-1}\left|\mathsf{e}^{\intercal}X\right|_{\infty}>\lambda\right)=o\left(1\right).$$

Proof. This is a direct consequence of Lemma 6 under Assumption 1. ■

Lemma 8 *Under Assumption 2,* $n^{-1}X^{\intercal}X$ *is compatible for* S *with compatibility constant* $\phi_{\beta}/\sqrt{2}$ *with probability approaching one.*

Proof. Due to Assumption 2 and Bühlmann and van der Geer's (2011) Corollary 6.8 it is sufficient to show that $\left|n^{-1}X^{\intercal}X-EX_tX_t^{\intercal}\right|_{\infty}$ is bounded by $\left(32s\right)^{-1}\phi^2$ with probability approaching one. However, in view of Lemma 6, $\left|n^{-1}X^{\intercal}X-EX_tX_t^{\intercal}\right|_{\infty}=O_p\left(n^{-1/2}\sqrt{\log np}\right)=o\left(\phi^2/s\right)$. ■

A.4.2 Skedastic LASSO Regression

Theorem 7 *Let Assumption 1, and 3 hold. And suppose that*

$$\mu = o(1) \text{ and } \lambda\sqrt{s} = o(\mu).$$

Then, the following holds with probability approaching one :
Part (i)

$$\frac{2}{n}\sum_{t=1}^{n}\left(w_t^{\intercal}\widehat{\gamma}_{lasso} - \sigma(q_t)\right)^2 + \mu\left|\widehat{\gamma}_{lasso} - \gamma_0\right|_1 \leq \frac{6}{n}\sum_{t=1}^{n} r_{\sigma_t}^2 + \frac{48\mu^2|S_\gamma|}{\phi_\gamma^2}$$

Part (ii) if $\mu_{thr} = o(\min\{|\gamma_{0j}| : \gamma_{0j} \neq 0\})$ *in addition, then*

$$\Pr\left(\widehat{S}_\gamma = S_\gamma\right) \to 1$$

and

$$\left|\widehat{\gamma}_{Tasso} - \gamma_0\right|_2 = O_p\left(\sqrt{\frac{|S_\gamma|\log|S_\gamma|}{n}}\right).$$

Also, for any $a \neq 0$

$$\sqrt{n}a^{\intercal}\left(\widehat{\gamma}_{Tasso} - \gamma_0\right) \Longrightarrow \mathcal{N}\left(0, a_0 M_w^{-1}\Omega_w M_w^{-1}a_0\right),$$

where $M_w = E\left(w_{0t}w_{0t}^{\intercal}\right)$ *and* $\Omega = Ew_{0t}w_{0t}^{\intercal}\eta_t^2$ *with* $w_{0t} = (w_{tj} : \gamma_{0j} \neq 0)$ *and* $a_0 = (a_j : \gamma_{0j} \neq 0)$.

Proof of Theorem 7 Recall the definition of $\widehat{e}_t$ and write that

$$|\widehat{e}_t| - w_t^{\intercal}\gamma = \eta_t + \left(\sigma_t - w_t^{\intercal}\gamma\right) + \left(|\widehat{e}_t| - |e_t|\right)$$

and by the triangle inequality

$$\left||\widehat{e}_t| - |e_t|\right| \leq |\widehat{e}_t - e_t| = \left|g(q_t) - \widehat{x}_t^{\intercal}\widehat{\beta}\right|.$$

Then,

$$\begin{aligned}\left(\widehat{e}_t - w_t^{\intercal}\gamma\right)^2 - \left(\widehat{e}_t - w_t^{\intercal}\gamma_0\right)^2 = \left(\sigma_t - w_t^{\intercal}\gamma\right)^2 - r_{\sigma t}^2 \\ - 2\eta_t w_t^{\intercal}(\gamma - \gamma_0) - 2w_t^{\intercal}(\gamma - \gamma_0)\left(|\widehat{e}_t| - |e_t|\right).\end{aligned}$$

By an iterated applications of the Hölder inequality,

$$\begin{aligned}\left|\frac{1}{n}\sum_{t=1}^{n}\left(g_t - \widehat{x}_t^{\intercal}\widehat{\beta}\right)w_t^{\intercal}\left(\widehat{\gamma} - \gamma_0\right)\right| &\leq \left|\frac{1}{n}\sum_{t=1}^{n} w_t\left(g_t - \widehat{x}_t^{\intercal}\widehat{\beta}\right)\right|_\infty \left|\widehat{\gamma} - \gamma_0\right|_1 \\ &\leq \sup_{1\leq i\leq p}\left(\frac{1}{n}\sum_{t=1}^{n} w_{ti}^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=1}^{n}\left(g_t - \widehat{x}_t^{\intercal}\widehat{\beta}\right)^2\right)^{1/2}\left|\widehat{\gamma} - \gamma_0\right|_1\end{aligned}$$

and

$$\left|\frac{1}{n}\sum_{t=1}^{n}\eta_t w_t^{\intercal}\left(\gamma-\gamma_0\right)\right| \leq \left|\frac{1}{n}\sum_{t=1}^{n}\eta_t w_t\right|_{\infty}\left|\gamma-\gamma_0\right|_1.$$

Conditional on

$$\mathcal{E}_n = \left\{\sup_{1\leq i\leq p}\left(\frac{1}{n}\sum_{t=1}^{n}w_{ti}^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=1}^{n}\left(g_t-\widehat{x}_t^{\intercal}\widehat{\beta}\right)^2\right)^{1/2}+\left|\frac{1}{n}\sum_{t=1}^{n}\eta_t w_t\right|_{\infty}\leq\frac{\mu}{8}\right\},$$

the remaining steps of the proof are identical to that of Theorem 6. That is, we have arrived at the equation (40) and the conditional event $\mathcal{E}_n$ corresponds to the event $\mathcal{A}_n$, which bounds $|n^{-1}\mathsf{e}^{\intercal}X|_{\infty}$.

Finally, note that the probability of $\mathcal{E}_n$ converges to 1 due to Theorem 6 and Lemma 6. ■

A.5 Proofs of Main Theorems

A.5.1 Proofs for Section 3

Proof of Theorem 2 Since $\widehat{S}=S$ with probability approaching one due to Theorem 6, we assume $\widehat{S}=S$ without loss of generality and thus we define $\widehat{g}^{\boldsymbol{j}}(q_t)=x_t^{\intercal}\widehat{\beta}^{\boldsymbol{j}}$.

Convergence of $\bar{T}_n$. Recall that

$$\widehat{\mathbb{Z}}_n^{\boldsymbol{j}}\left(\frac{y-\widehat{g}^{\boldsymbol{j}}(q)}{\widehat{\sigma}^{\boldsymbol{j}}(q)}\right)=\sqrt{n}\left(\widehat{F}^{\boldsymbol{j}}(y|q)-F^{\boldsymbol{j}}(y|q)\right)$$

and let $\tau^{\boldsymbol{j}}(y,q)=\frac{y-g^{\boldsymbol{j}}(q)}{\sigma^{\boldsymbol{j}}(q)}$ for each $j=1,2$. Since the sum of uniform P-Donskers is a uniform P-Donsker, to derive the limit distribution of T_n, we first show that for each $j=1,2$ the process $\widehat{\mathbb{Z}}_n^{\boldsymbol{j}}\left(\frac{y-\widehat{g}^{\boldsymbol{j}}(q)}{\widehat{\sigma}^{\boldsymbol{j}}(q)}\right)$ is a uniform P-Donsker, then derive the limit of $\sqrt{n}\left(F^{\boldsymbol{2}}(y|q)-F^{\boldsymbol{1}}(y|q)\right)$ explicitly and the limit covariance kernel of $\widehat{\mathbb{Z}}_n^{\boldsymbol{1}}\left(\frac{y-\widehat{g}^{\boldsymbol{1}}(q)}{\widehat{\sigma}^{\boldsymbol{1}}(q)}\right)-\widehat{\mathbb{Z}}_n^{\boldsymbol{2}}\left(\frac{y-\widehat{g}^{\boldsymbol{2}}(q)}{\widehat{\sigma}^{\boldsymbol{2}}(q)}\right)$, finally apply the uniform continuous mapping theorem in Linton, Song, and Whang (2010) since the supremum is a Lipschitz continuous operator.

However, Lemma 9 below shows the uniform weak convergence of $\widehat{\mathbb{Z}}_n^{\boldsymbol{j}}\left(\frac{y-\widehat{g}^{\boldsymbol{j}}(q)}{\widehat{\sigma}^{\boldsymbol{j}}(q)}\right)$ to the limit Gaussian process $\mathbb{Z}^{\boldsymbol{j}}\left(\frac{y-g(q)}{\sigma(q)}\right)+B^{\boldsymbol{j}}\left(\frac{y-g(q)}{\sigma(q)}\right)$ for each j. Furthermore, a direct computation of the covariance terms as in the proof of Theorem 4 yields the desired form of the limit Gaussian process.

Convergence of $\bar{U}_n$ Since Theorem 4 establishes an asymptotic equivalence, we need to derive (i) the weak convergence of $\widetilde{B}_n(\tau)=E\mathbb{Z}_n^{\boldsymbol{1}}(b^{\boldsymbol{1}},g^{\boldsymbol{1}},\tau)-E\mathbb{Z}_n^{\boldsymbol{2}}(b^{\boldsymbol{2}},g^{\boldsymbol{2}},\tau)$ when $b^{\boldsymbol{j}}=\widehat{\beta}^{\boldsymbol{j}}-\beta_0^{\boldsymbol{j}}$ and $g^{\boldsymbol{j}}=\widehat{\gamma}^{\boldsymbol{j}}-\gamma_0^{\boldsymbol{j}}$ for $j=1,2$ and (ii) the weak convergence of

$$A_n(\tau)=\int_{-\infty}^{\tau}\mathbb{Z}_n^{\boldsymbol{1}}(u)-\mathbb{Z}_n^{\boldsymbol{2}}(u)\,du,\quad \tau\in\mathbb{R}.$$

The former follows from the Taylor series expansion of $E\mathbb{Z}_n(b, g, \tau)$ and the standard CLT as in Lemma 4, analogously as for FSD. Note that $A_n(\tau)$ is an empirical process indexed by functions of $(\tau - \varepsilon_{1t})_+ - (\tau - \varepsilon_{2t})_+$ with $\tau \in \mathbb{R}$. It is easy to see that we can set $F = |\varepsilon_{1t} - \varepsilon_{2t}|$ as an envelope, whose second moment is bounded. Also, note that this class of functions is a V-C subgraph class as the index τ varies the function values monotonically. Thus, we can apply the uniform weak convergence result as in Theorem 2.8.3 in van der Vaart and Wellner (1996).

To deal with the composite process where we plug in $\widehat{\tau}^{\mathbf{1}}(y, q)$ and $\widehat{\tau}^{\mathbf{2}}(y, q)$, we consider a generalized version of $A_n(\tau)$, that is,

$$A_n(\tau, \tau_1) = \int_{-\infty}^{\tau} \mathbb{Z}_n^{\mathbf{1}}(u) - \int_{-\infty}^{\tau+\tau_1} \mathbb{Z}_n^{\mathbf{2}}(u)\, du, \quad \tau \in \mathbb{R} \text{ and } \tau_1 \in (-\delta, \delta)$$

for some $\delta > 0$. As for $A_n(\tau)$, we can see that we can set the envelope $F = |\varepsilon_{1t} - \varepsilon_{2t}| + \delta$ and the argument for V-C subgraph class remains valid as τ_1 is bounded. Then, the equivalence relation

$$\begin{aligned}
&\int_{-\infty}^{\widehat{\tau}^{\mathbf{1}}(y,q)} \widehat{\mathbb{Z}}_n^{\mathbf{1}}(u) - \int_{-\infty}^{\widehat{\tau}^{\mathbf{2}}(y,q)} \widehat{\mathbb{Z}}_n^{\mathbf{2}}(u)\, du \\
&= \int_{-\infty}^{\widehat{\tau}^{\mathbf{1}}(y,q)} \widehat{\mathbb{Z}}_n^{\mathbf{1}}(u) - \int_{-\infty}^{\widehat{\tau}^{\mathbf{2}}(y,q)} \widehat{\mathbb{Z}}_n^{\mathbf{2}}(u)\, du - \int_{-\infty}^{\tau(y,q)} \widehat{\mathbb{Z}}_n^{\mathbf{1}}(u) - \widehat{\mathbb{Z}}_n^{\mathbf{2}}(u)\, du \\
&+ \int_{-\infty}^{\tau(y,q)} \widehat{\mathbb{Z}}_n^{\mathbf{1}}(u) - \widehat{\mathbb{Z}}_n^{\mathbf{2}}(u)\, du \\
&= \int_{-\infty}^{\tau(y,q)} \widehat{\mathbb{Z}}_n^{\mathbf{1}}(u) - \widehat{\mathbb{Z}}_n^{\mathbf{2}}(u)\, du + o_p(1),
\end{aligned}$$

follows similarly as in Lemma 9, which also shows that $\widehat{\tau}^{\boldsymbol{j}}(y, q) \xrightarrow{p} \tau(y, q)$ uniformly in y and q, for both $j = 1, 2$. ■

Now, we present the lemma cited above. Here, we omit the superscript $\boldsymbol{j}$.

Lemma 9 *Under the assumptions of Theorem 2,*

$$\widehat{\mathbb{Z}}_n\left(\frac{y - \widehat{g}(q)}{\widehat{\sigma}(q)}\right) \Rightarrow \mathbb{Z}\left(\frac{y - g(q)}{\sigma(q)}\right) + B\left(\frac{y - g(q)}{\sigma(q)}\right).$$

Proof. Recall from Theorem 4 that

$$\widehat{\mathbb{Z}}_n(\tau) = \mathbb{Z}_n(\tau) + \widetilde{B}_n(\tau) + o_p(1) \tag{43}$$

and

$$\mathbb{Z}_n(\tau) + \widetilde{B}_n(\tau) \Rightarrow \mathbb{Z}(\tau) + B(\tau)$$

over the real line due to due to the standard CLT and uniform weak convergence of empirical distribution functions, see e.g. Theorem 2.8.3 in van der Vaart and Wellner (1996). To extend this

result, introduce semimetrics

$$\begin{aligned}\dot{\rho}(\tau_1,\tau_2) &= \left(E\left(1\left\{\varepsilon_t \leq \tau_1\right\} - 1\left\{\varepsilon_t \leq \tau_2\right\}\right)^2\right)^2 \\ &= \left|F(\tau_1) - F(\tau_2)\right|^{1/2}\end{aligned}$$

and

$$\ddot{\rho}(z_1,z_2) = \left(E\left(1\left\{\varepsilon_t \leq \tau(z_1)\right\} - 1\left\{\varepsilon_t \leq \tau(z_2)\right\}\right)^2\right)^2$$

with $z = (y,q)$. Certainly, the semimetric space $(\mathbb{R},\dot{\rho})$ is totally bounded. Since $\dot{\rho}(\tau_1,\tau_2) = \ddot{\rho}(z_1,z_2)$ for $\tau_i = \tau(z_i)$, $(\mathbb{R}^\infty,\ddot{\rho})$ is totally bounded. Also, the stochastic $\dot{\rho}$-equicontinuity of $\widehat{\mathbb{Z}}_n(\tau)$ is equivalent to the stochastic $\ddot{\rho}$-equicontinuity of $\widehat{\mathbb{Z}}_n(\tau(y,q))$ not to mention the finite-dimensional convergence of $\widehat{\mathbb{Z}}_n(\tau)$ being identical to that of $\widehat{\mathbb{Z}}_n(\tau(y,q))$. Thus, the uniform weak convergence of $\widehat{\mathbb{Z}}_n(\tau(y,q))$ over $\mathbb{R}^\infty$ follows.

Furthermore, note that $\widehat{\tau}(z)$ is uniformly $\dot{\rho}$-consistent since

$$\begin{aligned}\sup_z \left|\dot{\rho}\left(\widehat{\tau}(z),\tau(z)\right)\right|^2 &= \sup_z \left|F\left(\widehat{\tau}(z)\right) - F\left(\tau(z)\right)\right| \\ &\leq \sup_{(y,q)}\left|F\left(\frac{y-\widehat{g}(q)}{\widehat{\sigma}(q)}\right) - F\left(\frac{y-g(q)}{\widehat{\sigma}(q)}\right)\right| + \sup_{(y,q)}\left|F\left(\frac{y-g(q)}{\sigma(q)}\right) - F\left(\frac{y-g(q)}{\widehat{\sigma}(q)}\right)\right| \\ &= O_p\left(\left\|\widehat{g}-g\right\|_\infty\right) + O_p\left(\left\|\widehat{\sigma}-\sigma\right\|_\infty\right) = o_p(1),\end{aligned}$$

where the last equality is due to the uniform convergence result in Chen and Christensen (2015) and Assumption 1-1. This and the uniform stochastic $\dot{\rho}$-equicontinuity of $\widehat{\mathbb{Z}}_n(\tau)$ imply that

$$\sup_{(y,q)}\left|\widehat{\mathbb{Z}}_n\left(\widehat{\tau}(y,q)\right) - \widehat{\mathbb{Z}}_n\left(\tau(y,q)\right)\right| = o_p(1).$$

This establishes the lemma. ■

A.5.2 Proofs for Section ??

In the following, we maintain the convention that the quantities defined with superscript "*" denote the bootstrap counterparts of the original terms in the preceding proofs and thus we do not redefine them. And let "$\Rightarrow$ in P" denote the weak convergence of bootstrap statistics conditional on the original data, and define the stochastic order notation $O_{p^*}(1)$ and $o_{p^*}(1)$ in terms of the conditional distribution given the original data. Specifically, for any $\epsilon > 0$, $\tau_n^* = o_p^*(1)$ if $P^*\left\{|\tau_n^*| > \epsilon\right\} \xrightarrow{p} 0$, and $\tau_n^* = O_p^*(1)$ if there exists $C = O_p(1)$ such that $P^*\left\{|\tau_n^*| > C\right\} < \epsilon$ for all sufficiently large n.

Proof of Theorem 3 We imitate the derivation of the asymptotic null distributions in Theorem 2. Some intermediate steps are given in separate lemmas later in this section.

First, Theorem 8 below shows that the process $\widehat{\mathbb{Z}}_n^{\boldsymbol{j}*}(\tau)$ is a uniform P-Donsker and establishes the weak convergence of $\widehat{\mathbb{Z}}_n^{\boldsymbol{j}*}(\tau) \Longrightarrow^* \mathbb{Z}^{\boldsymbol{j}}(\tau) + B^{\boldsymbol{j}}\left(\frac{y-g(q)}{\sigma(q)}\right)$ in P, for $\boldsymbol{j} = \mathbf{1}, \mathbf{2}$. Next, suppressing the supscript $\boldsymbol{j}$, we consider the composite process $\widehat{\mathbb{Z}}_n^*(\widehat{\tau}^*(y,q))$ and show that $\widehat{\mathbb{Z}}_n^*(\tau(y,q))$ is a uniform P-Donsker and that

$$\widehat{\mathbb{Z}}_n^*(\widehat{\tau}^*(y,q)) - \widehat{\mathbb{Z}}_n^*(\tau(y,q)) = o_p^*(1).$$

This follows for the same reasoning as Lemma 9. That is, since the process $\widehat{\mathbb{Z}}_n^*(\tau(y,q))$ on $\mathbb{R}^\infty$ is a uniform P-Donsker as shown in the first part of the proof of Lemma 9, it remains to show that $\widehat{\tau}^*(y,q)$ is uniformly $\dot{\rho}$-consistent to $\tau(y,q)$. Note that with $z = (y,q)$

$$\begin{aligned}
\sup_z \left|\dot{\rho}\left(\widehat{\tau}^*(z), \tau(z)\right)\right|^2 &= \sup_z \left|F\left(\widehat{\tau}^*(z)\right) - F\left(\tau(z)\right)\right| \\
&\leq \sup_{(y,q)} \left|F\left(\frac{y-\widehat{g}^*(q)}{\widehat{\sigma}^*(q)}\right) - F\left(\frac{y-g(q)}{\widehat{\sigma}^*(q)}\right)\right| + \sup_{(y,q)} \left|F\left(\frac{y-g(q)}{\sigma(q)}\right) - F\left(\frac{y-g(q)}{\widehat{\sigma}^*(q)}\right)\right| \\
&= O_p^*\left(\|\widehat{g}^* - g\|_\infty\right) + O_p^*\left(\|\widehat{\sigma}^* - \sigma\|_\infty\right) = o_p^*(1),
\end{aligned}$$

by the uniform consistency result in Chen and Christensen (2015).

The preceding steps establish the weak convergence of $\widehat{\mathbb{Z}}_n^{j*}(\widehat{\tau}^*(y,q))$ for each $j = 1, 2$ and our statistic is the sup of the difference. In view of the continuous mapping theorem for the sup operator and the permanence property of P-Donsker for the sum of two classes of functions, we only have to verify the covariance kernels converges properly. This step proceeds as in the derivation of (46) in the proof of Theorem 8 and thus details are omitted.

The verification of convergence of U_n^* follows the same reasoning for that of T_n^* imitating the derivation of the convergence of the original sample statistic.

Finally, the consistency of the test is straightforward since the bootstrap statistic $T_n^* = O_p(1)$ under both the null and alternative hypotheses while $T_n, U_n \to +\infty$ under a fixed alternative that $F^1(y|q) > F^2(y|q)$ for some (y,q). That is, at such a (y,q) we have

$$\sqrt{n}\left(\widehat{F}^{\mathbf{1}}(y|q) - \widehat{F}^{\mathbf{2}}(y|q)\right) = \sqrt{n}\left(F^{\mathbf{1}}(y|q) - F^{\mathbf{2}}(y|q) + O_p(1)\right) \to +\infty$$

and thus the supremum and the supremum of their integrals also diverge. ■

The following lemma, which is a modification of the maximal inequality in Lemma 1 to allow for unbounded functions, is first verified to establish Theorem 8 afterward.

Lemma 10 *Consider functions $g(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^s \to \mathbb{R}$ indexed by τ and b such that $g(e_t, x_t) = \frac{1}{a_n} G^{(1)}\left(\frac{\tau - e_t}{a_n}\right) x_t^\intercal b$ and its collection*

$$\mathcal{G}_{n,\delta} = \{g(\cdot,\cdot) : |\tau| \leq C, |b| \leq \delta\}$$

for some finite C and $\delta > 0$. Let $\{e_t, x_t\}$ be a strictly stationary and absolutely regular process with β-mixing coefficient $\beta_m = O(\rho^m)$ for some $\epsilon < \rho < 1-\epsilon$ and the moment conditions $\inf_{\theta:|\theta|=1} E\left|x_t^{\intercal}\theta\right| > 0$ and $\sup_j Ex_{jt}^{2p} < \infty$ for some $p > 1$. Then, for $\delta^2 < a_n^3 s$ and some finite C'

$$E \sup_{g \in \mathcal{G}_{n,\delta}} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(g(e_t, x_t) - Eg(e_t, x_t) \right) \right| \leq C' \frac{s}{a_n^{3/2}} \delta \log s\delta^{-1}.$$

Proof of Lemma 10 We modify the proof of Lemma 1. We begin with establishing $\|g\|_{2,\beta} \leq C \|g\|_{2p}$ with $p > 1$ and some bounded C, which only depends on the distribution of the sample. For a given function ϕ, let $\phi^*(y) = \sup_{x>0} [xy - \phi(x)]$. Then, due to Lemma 1 and 2 of Doukhan et al (1995),

$$\|g\|_2 \leq \|g\|_{2,\beta} \leq \|g\|_{\phi,2} \sqrt{1 + \int_0^1 \phi^* \left(\beta^{-1}(u)\right) du},$$

where $\|g\|_{\phi,2} = \inf\left\{c > 0 : \mathrm{E}\phi\left(|g/c|^2\right) \leq 1\right\}$, which is the Orlicz norm and is equivalent to $\|g\|_{2p}$ when $\phi(x) = x^p$. Furthermore, when $\phi(x) = x^p$, we have $\int_0^1 \phi^* \left(\beta^{-1}(u)\right) du = C_p \sum_{n=1}^{\infty} n^{\frac{1}{p-1}} \beta_n$, where C_p is a constant dependent only on p, see Doukhan et al (1995 p. 404). Note that given the mixing condition the sum $\sum_{n=1}^{\infty} n^{\frac{1}{p-1}} \beta_n$ is bounded for any $p > 1$. Then,

$$\begin{aligned}
\|g\|_{2,\beta}^{2p} &\leq C \|g\|_{2p}^{2p} = CE \left| \frac{1}{a_n} G^{(1)} \left(\frac{\tau - e_t}{a_n} \right) x_t^{\intercal} b \right|^{2p} \\
&\leq C' a_n^{-2p} E |x_t|^{2p} |b|^{2p} \\
&\leq C' s^p |b|^{2p} a_n^{-2p},
\end{aligned}$$

due to the boundedness of $G^{(1)}$, Cauchy-Schwarz inequality, and the fact that

$$E|x_t|^{2p} = s^p E \left(s^{-1} \sum_{j=1}^{s} x_{tj}^2 \right)^p \leq s^p E \left(s^{-1} \sum_{j=1}^{s} x_{tj}^{2p} \right) \leq s^p \max_j E|x_{tj}|^{2p}. \tag{44}$$

This yields that

$$\mathcal{G}_{n,\delta} \subset \mathcal{G}^{\beta}_{n, Cs^{1/2}\delta a_n^{-1}}.$$

This replaces (30). On the other hand, it follows from the rank condition on x_t and the fact that $G^{(1)}$ is a density that

$$E \left| \frac{1}{a_n} G^{(1)} \left(\frac{\tau - e_t}{a_n} \right) x_t^{\intercal} b \right|^2 \geq C |b|^2,$$

which implies that

$$\mathcal{G}^{\beta}_{n,\delta} \subset \mathcal{G}_{n, C_1 \delta}. \tag{45}$$

The next step to compute the bracketing number depends on the bound in (31), which in the current lemma becomes for any $\bar{\theta}$, where $\theta = \left(b^{\intercal}, \tau\right)^{\intercal}$,

$$
\begin{aligned}
& E \sup_{\theta:\left|\theta-\bar{\theta}\right|\leq v} \frac{1}{a_n^2} \left| \left(G^{(1)}\left(\frac{\tau - e_t}{a_n}\right) x_t^{\intercal} b - G^{(1)}\left(\frac{\bar{\tau} - e_t}{a_n}\right) x_t^{\intercal} \bar{b} \right) \right|^2 \\
& \leq \frac{3}{a_n^2} E \sup_{\theta:\left|\theta-\bar{\theta}\right|\leq v} \left| \left(G^{(1)}\left(\frac{\tau - e_t}{a_n}\right) x_t^{\intercal} \left(b - \bar{b}\right) \right) \right|^2 \\
& + \frac{3}{a_n^2} E \sup_{\theta:\left|\theta-\bar{\theta}\right|\leq v} \left| \left(G^{(1)}\left(\frac{\tau - e_t}{a_n}\right) - G^{(1)}\left(\frac{\bar{\tau} - e_t}{a_n}\right) \right) x_t^{\intercal} \bar{b} \right|^2 \\
& \leq \frac{1}{a_n} E\left|x_t\right|^2 v^2 + \frac{v^2}{a_n^3} E\left|x_t\right|^2 \left|\bar{b}\right|^2 \leq s v^2 \left(\frac{1}{a_n} + \frac{\delta^2}{a_n^3}\right) \leq \frac{2 s v^2}{a_n}.
\end{aligned}
$$

The remaining steps proceeds similarly given these updated bounds and thus details are omitted. ■

In the following, we employ the common notation of the superscript "*" to indicate the bootstrap quantities, "$\Rightarrow$ in P" to denote the weak convergence of bootstrap statistics conditional on the original data, and the stochastic order notation $O_{p^*}(1)$ and $o_{p^*}(1)$ in terms of the conditional distribution given the original data. Specifically, for any $\epsilon > 0$, $\tau_n^* = o_p^*(1)$ if $P^*\left\{|\tau_n^*| > \epsilon\right\} \xrightarrow{p} 0$, and $\tau_n^* = O_p^*(1)$ if there exists $C = O_p(1)$ such that $P^*\left\{|\tau_n^*| > C\right\} < \epsilon$ for all sufficiently large n.

Theorem 8 *Let Assumption 1 hold and x_t^* and ε_t^* satisfy the mixing condition in Assumption 1.1. Furthermore, if $a_n = o\left(s^{-1}/\log s\right)$ and $s\sqrt{\log s} n^{-1/4} = o\left(a_n\right)$, then*

$$
\widehat{\mathbb{Z}}_n^*(\tau) \Longrightarrow \mathbb{Z}(\tau) + f(\tau) Z_1 - \tau f(\tau) Z_2 \text{ in } P
$$

where $\mathbb{Z}$, Z_1 and Z_2 are the same as in Theorem 4.

Proof of Theorem 8 First, as for $\widehat{\mathbb{Z}}_n(\tau)$, we begin with verifying the conditional weak convergence of the bootstrap empirical process

$$
\widehat{\mathbb{Z}}_n^*(\tau) \equiv \sqrt{n}\left(\widehat{F}^*(\tau) - F^*(\tau)\right) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(1\left\{ y_t^* - x_t^{*\intercal} \widehat{\beta}^* \leq \tau \right\} - G\left(\frac{\tau - \widehat{\varepsilon}_t}{a_n}\right) \right),
$$

where the last equality follows from $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{1}{n} \sum_{j=1}^{n} G\left(\frac{\tau - \widehat{\varepsilon}_j}{a_n}\right) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} G\left(\frac{\tau - \widehat{\varepsilon}_t}{a_n}\right)$. Note that we begin with the case where σ_t is fixed at 1.

Since $y_t^* = x_t^{*\intercal} \widehat{\beta} + \varepsilon_t^*$,

$$
\left\{ y_t^* - x_t^{*\intercal} \widehat{\beta}^* \leq \tau \right\} = \left\{ \varepsilon_t^* - x_t^{*\intercal} \left(\widehat{\beta}^* - \widehat{\beta} \right) \leq \tau \right\}.
$$

Then, by the same reasoning as in Lemma 3, $r_n \left|\widehat{\beta}^* - \widehat{\beta}\right|_2 = O_p^*(1)$, with $r_n = \sqrt{n}/\sqrt{s \log s}$. Next, recall that

$$\varepsilon_t^* = \widehat{\varepsilon}_{i_{t-1}^*} + a_n \eta_t = y_{i_{t-1}^*} - x_{i_{t-1}^*}^{\intercal} \widehat{\beta} + a_n \eta_t$$

and define

$$\mathbb{Z}_n^*(b, \tau) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(1\left\{ \varepsilon_t^* - x_t^{*\intercal} b \leq \tau \right\} - G\left(\frac{\tau - \widehat{\varepsilon}_t}{a_n} \right) \right),$$

so that $\widehat{\mathbb{Z}}_n^*(\tau) = \mathbb{Z}_n^*\left(\widehat{\beta}^* - \widehat{\beta}, \tau\right)$. Then, we can imitate the convergence of $\mathbb{Z}_n^*$ conditional on $\mathcal{X}_n$ since the conditional distribution of $\{\varepsilon_t^*\}$ and $\{x_t^*\}$ conditional on $\mathcal{X}_n$ satisfies the stationarity and mixing conditions. In particular,

$$\begin{aligned}
\mathbb{M}_n^*(b, \tau) &= \mathbb{Z}_n^*(b, \tau) - E^* \mathbb{Z}_n^*(b, \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left[1\left\{ \varepsilon_t^* - x_t^{*\intercal} b \leq \tau \right\} - G\left(\frac{\tau - \left(\widehat{\varepsilon}_t - x_t^{\intercal} b\right)}{a_n} \right) \right].
\end{aligned}$$

And by Lemma 1

$$\sup_{\tau} \sup_{|b| \leq r_n^{-1} C} \left| \mathbb{M}_n^*(b, \tau) - \mathbb{M}_n^*(0, \tau) \right| = o_p^*(1),$$

for any $C < \infty$. Then, since $E^* \mathbb{Z}_n^*(0, \tau) = 0$,

$$\begin{aligned}
\mathbb{Z}_n^*(b, \tau) &= \mathbb{M}_n^*(b, \tau) - \mathbb{M}_n^*(0, \tau) + E^* \mathbb{Z}_n^*(b, \tau) + \mathbb{Z}_n^*(0, \tau) \\
&= E^* \mathbb{Z}_n^*(b, \tau) + \mathbb{Z}_n^*(\tau) + o_p^*(1),
\end{aligned}$$

where $\mathbb{Z}_n^*(\tau) := \mathbb{Z}_n^*(0, \tau)$.

We begin with the conditional weak convergence of $\mathbb{Z}_n^*(\tau) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(1\left\{ \varepsilon_t^* \leq \tau \right\} - G\left(\frac{\tau - \widehat{\varepsilon}_t}{a_n} \right) \right)$. The uniform tightness of the process follows from the same argument as that of $\mathbb{Z}_n(\tau)$. For the finite dimensional convergence, first note that the variable $1\left\{ \varepsilon_t^* \leq \tau \right\} - \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{\tau - \widehat{\varepsilon}_t}{a_n} \right)$ is centered, stationary, and mixing (conditional on the original sample) with covariance (conditional on the original sample)

$$\begin{aligned}
&\operatorname{cov}^*\left(1\left\{ \varepsilon_t^* \leq \tau_1 \right\}, 1\left\{ \varepsilon_s^* \leq \tau_2 \right\} \right) \\
&= E^* 1\left\{ \varepsilon_t^* \leq \tau_1 \right\} 1\left\{ \varepsilon_s^* \leq \tau_2 \right\} - \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{\tau_1 - \widehat{\varepsilon}_t}{a_n} \right) \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{\tau_2 - \widehat{\varepsilon}_t}{a_n} \right) \\
&\xrightarrow{p} E 1\left\{ \varepsilon_t \leq \tau_1 \right\} 1\left\{ \varepsilon_s \leq \tau_2 \right\} - F(\tau_1) F(\tau_2). \qquad (46)
\end{aligned}$$

To see this, first suppose $s = t$ and $\tau_1 = \tau_2$ and note that $E^* 1\left\{ \varepsilon_t^* \leq \tau_1 \right\} 1\left\{ \varepsilon_s^* \leq \tau_2 \right\} = E^* 1\left\{ \varepsilon_t^* \leq \tau_1 \right\} = \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{\tau_1 - \widehat{\varepsilon}_t}{a_n} \right)$. The convergence of $\frac{1}{n} \sum_{t=1}^{n} G\left(\frac{\tau_1 - \widehat{\varepsilon}_t}{a_n} \right)$ to $F(\tau_1)$ in P is straightforward from the

uniform law of large numbers for triangular arrays and the mean value expansion of

$$\begin{aligned} & EG\left(\frac{\tau_1-\varepsilon_t+x_t^{\intercal}\left(\widehat{\beta}-\beta_0\right)}{a_n}\right) \\ & = EG\left(\frac{\tau_1-\varepsilon_t}{a_n}\right)-\frac{1}{a_n}EG^{(1)}\left(\frac{\tau_1-\widehat{\varepsilon}_t}{a_n}\right)x_t^{\intercal}\left(\widehat{\beta}-\beta_0\right), \end{aligned}$$

for which we note that

$$EG\left(\frac{\tau_1-\varepsilon_t}{a_n}\right)=\frac{1}{a_n}\int G^{(1)}\left(\frac{\tau_1-s}{a_n}\right)F(s)\,ds\to F(\tau_1),$$

by the integration by parts and then by the change-of-variables and the dominated convergence theorem and that $\frac{1}{a_n}EG^{(1)}\left(\frac{\tau_1-\widehat{\varepsilon}_t}{a_n}\right)x_t^{\intercal}\left(\widehat{\beta}-\beta_0\right)=o(1)$ due to the deviation bound that $E\left(x_t^{\intercal}\left(\widehat{\beta}-\beta_0\right)\right)^2=o(a_n)$ and the fact that $EG\left(\frac{\tau_1-\widehat{\varepsilon}_t}{a_n}\right)=O(a_n)$.

Now, if $s=t$ and $\tau_1\neq\tau_2$, by the same argument as above, $E^*1\{\varepsilon_t^*\leq\tau_1\}1\{\varepsilon_s^*\leq\tau_2\}\xrightarrow{p}F(\tau_1\bigwedge\tau_2)$. And for $t=s+k$ and $k>0$, by construction

$$\operatorname{cov}^*\left(1\{\varepsilon_t^*\leq\tau_1\},1\{\varepsilon_s^*\leq\tau_2\}\right)=(1-\pi_n)^k\operatorname{cov}^*\left(1\left\{\varepsilon_{j_s^*+k}\leq\tau_1\right\},1\left\{\varepsilon_{j_s^*}\leq\tau_2\right\}\right).$$

And

$$E^*1\left\{\varepsilon_{j_s^*+k}\leq\tau_1\right\}1\left\{\varepsilon_{j_s^*}\leq\tau_2\right\}=\frac{1}{n}\sum_{t=1}^{n}G\left(\frac{\tau_1-\widehat{\varepsilon}_{t+k}}{a_n}\right)G\left(\frac{\tau_2-\widehat{\varepsilon}_t}{a_n}\right)$$

with the convention that $\widehat{\varepsilon}_{t+k}=\widehat{\varepsilon}_{t+k-n}$ if $t+k>n$. Following very similar algebra as above, we can show that it converges in probability to $E1\{\varepsilon_t\leq\tau_1\}1\{\varepsilon_s\leq\tau_2\}$. Finally, by Theorem 1 of Politis and Romano (1994), the sum over k also converges.

Turn to the limit of $E^*\mathbb{Z}_n^*(b,\tau)$. For some mean value $\widetilde{b}$, we can expand

$$\begin{aligned} E^*\mathbb{Z}_n^*(b,\tau) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[G\left(\frac{\left(\tau-\left(\widehat{\varepsilon}_t-x_t^{\intercal}b\right)\right)}{a_n}\right)-G\left(\frac{\tau-\widehat{\varepsilon}_t}{a_n}\right)\right] \\ &= \frac{1}{na_n}\sum_{t=1}^{n}G^{(1)}\left(\frac{\tau-\left(\widehat{\varepsilon}_t-x_t^{\intercal}\widetilde{b}\right)}{a_n}\right)x_t^{\intercal}\sqrt{n}b \end{aligned}$$

and show that

$$\sup_{\tau}\sup_{|b|\leq Cr_n^{-1}}\left|\frac{1}{na_n}\sum_{t=1}^{n}G^{(1)}\left(\frac{\tau-\left(\widehat{\varepsilon}_t-x_t^{\intercal}\widetilde{b}\right)}{a_n}\right)x_t^{\intercal}\sqrt{n}b-f(\tau)Ex_t^{\intercal}\sqrt{n}b\right|=o_p(1).$$

Let $\widehat{k}_t := x_t^{\intercal}\left(\widehat{\beta} - \beta_0 + \widetilde{b}\right)$. By the Cauchy-Schwarz inequality and the triangle inequality, the LHS of the proceeding equation is bounded by

$$\begin{aligned}
&\sup_{\tau}\sup_{|b|\leq Cr_n^{-1}}\left|\frac{1}{\sqrt{n}a_n}\sum_{t=1}^{n}\left(G^{(1)}\left(\frac{\tau-\varepsilon_t-\widehat{k}_t}{a_n}\right)x_t^{\intercal}b - E\left[G^{(1)}\left(\frac{\tau-\varepsilon_t-\widehat{k}_t}{a_n}\right)x_t^{\intercal}b\right]\right)\right| \\
&+\sup_{\tau}\sup_{|b|\leq Cr_n^{-1}}\left|\frac{1}{a_n}E\left[\left(G^{(1)}\left(\frac{\tau-\varepsilon_t-\widehat{k}_t}{a_n}\right) - G^{(1)}\left(\frac{\tau-\varepsilon_t}{a_n}\right)\right)x_t^{\intercal}b\sqrt{n}\right]\right| \\
&+\sup_{\tau}\sup_{|b|\leq Cr_n^{-1}}\left|\left[\frac{1}{a_n}EG^{(1)}\left(\frac{\tau-\varepsilon_t}{a_n}\right) - f(\tau)\right]Ex_t^{\intercal}b\sqrt{n}\right|.
\end{aligned}$$

For the first term, we can apply the maximal inequality in Lemma 10. For the second, we note that since $G^{(1)}$ has a bounded derivative, we can bound it by

$$a_n^{-2}E\left|\widehat{k}_t x_t^{\intercal}b\sqrt{n}\right| \leq a_n^{-2}\left|Ex_t x_t^{\intercal}\right|_{\infty}\left(|b|_1 + \left|\widehat{\beta}-\beta_0\right|_1\right)^2\sqrt{n} = o(1),$$

due to the Hölder inequality and the triangle inequality. Finally, the standard argument in the kernel density estimation using change-of-variables formula and dominated convergence theorem yields that $a_n^{-1}E\left[G^{(1)}\left(\frac{\tau-\varepsilon_t}{a_n}\right)\right] - f(\tau) = O(a_n)$ uniformly in τ and $E\left|x_t^{\intercal}b\sqrt{n}\right| = O\left(\sqrt{s\log s}\right)$ uniformly in b. Thus, the last term is $O\left(a_n\sqrt{s\log s}\right) = o(1)$ as well.

The proof for the general case proceeds similarly for the known σ_t case but with much heavier notation. As in the proofs of Theorem 4, we can replace $\widehat{\varepsilon}_t^*$ with $\varepsilon_t^* - \sigma_t^{*-1}x_t^{*\intercal}\left(\widehat{\beta}^* - \widehat{\beta}\right) - \sigma_t^{*-1}w_t^{*\intercal}\left(\widehat{\gamma}^* - \widehat{\gamma}\right)\tau$ when we consider the set $\{\widehat{\varepsilon}_t^* \leq \tau\}$, where $\varepsilon_t^* = \widehat{\varepsilon}_{i_{t-1}^*} + a_n\eta_t$. Compared to the case of known σ_t, we have an additional term $\sigma_t^{*-1}w_t^{*\intercal}\left(\widehat{\gamma}^* - \widehat{\gamma}\right)\tau$. It demands us to derive the limit of $\left(\widehat{\gamma}^* - \widehat{\gamma}\right)$ and introduce $\mathbb{Z}_n^*(b,g,\tau)$ analogously to $\mathbb{Z}_n(b,g,\tau)$ in the proof of Theorem 4. The convergence of $\left(\widehat{\gamma}^* - \widehat{\gamma}\right)$ is standard as $\widehat{\gamma}^*$ is an OLS estimator. Then, the same maximal inequality in Lemma 1 yields that

$$\sup_{\tau}\sup_{|g|,|b|\leq r_n^{-1}C}\left|\mathbb{M}_n^*(b,g,\tau) - \mathbb{M}_n^*(0,0,\tau)\right| = o_p^*(1),$$

where $\mathbb{M}_n^*(b,g,\tau) = \mathbb{Z}_n^*(b,g,\tau) - E^*\mathbb{Z}_n^*(b,g,\tau)$.

Turning to $E^*\mathbb{Z}_n^*(b,g,\tau)$, note that

$$E^*\mathbb{Z}_n^*(b,g,\tau) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[G\left(a_n^{-1}\left(\tau - \left(\widehat{\varepsilon}_t - \frac{x_t^{\intercal}b}{w_t^{\intercal}\widehat{\gamma}} - \tau\frac{w_t^{\intercal}g}{w_t^{\intercal}\widehat{\gamma}}\right)\right)\right) - G\left(a_n^{-1}\left(\tau - \widehat{\varepsilon}_t\right)\right)\right],$$

and its expansion is similar to that of $E^*\mathbb{Z}_n^*(b,\tau)$ and the details are omitted for the sake of space. ■