Econometric Theory, **0**, 2023, 1–30.
doi:10.1017/S0266466623000348

PERFORMANCE OF EMPIRICAL RISK MINIMIZATION FOR LINEAR REGRESSION WITH DEPENDENT DATA

CHRISTIAN BROWNLEES
Universitat Pompeu Fabra and Barcelona SE

GUÐMUNDUR STEFÁN GUÐMUNDSSON
Aarhus University

This paper establishes bounds on the performance of empirical risk minimization for large-dimensional linear regression. We generalize existing results by allowing the data to be dependent and heavy-tailed. The analysis covers both the cases of identically and heterogeneously distributed observations. Our analysis is nonparametric in the sense that the relationship between the regressand and the regressors is not specified. The main results of this paper show that the empirical risk minimizer achieves the optimal performance (up to a logarithmic factor) in a dependent data setting.

1. INTRODUCTION

Let $\mathcal{D} = \{(Y_t, \boldsymbol{X}_t')'\}_{t=1}^T$ be a sequence of dependent random vectors taking values in $\mathcal{Y} \times \mathcal{X}$ with $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^p$. The p-dimensional vector $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{pt})'$ is used to predict the variable Y_t through the class of linear forecasts given by

$$f_{\boldsymbol{\theta} t} = \theta_1 X_{1t} + \ldots + \theta_p X_{pt}, \tag{1}$$

where $(\theta_1, \ldots, \theta_p)' = \boldsymbol{\theta} \in \mathbb{R}^p$. As is customary in learning theory, the relation between the regressand Y_t and the regressors $X_{1t}, \ldots, X_{pt}$ is not specified, and (1) should be interpreted as a class of prediction rules indexed by $\boldsymbol{\theta} \in \mathbb{R}^p$.

We have benefited from discussions with Liudas Giraitis, Emmanuel Guerre, Petra Laketa, Gabor Lugosi, Stanislav Nagy, Jordi Llorens-Terrazas, Yaping Wang, and Geert Mesters as well as seminar participants at the Granger Center, Nottingham University and School of Economics and Finance, Queen Mary University of London. We would also like to thank the Co-Editor Liangjun Su and two anonymous referees for their useful comments. Christian Brownlees acknowledges support from the Spanish Ministry of Science and Technology (Grant No. MTM2012-37195); the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S) funded by MCIN/AEI/10.13039/501100011033; the Ayudas Fundación BBVA Proyectos de Investigación Cientìfica en Matemáticas 2021. Guðmundur Stefán Guðmundsson acknowledges financial support from Danish National Research Foundation (DNRF Chair Grant No. DNRF154). Address correspondence to Christian Brownlees, Department of Economics and Business, Universitat Pompeu Fabra and Barcelona SE, Barcelona, Spain; e-mail: christian.brownlees@upf.edu.

A prediction rule is to be chosen from the data. The precision of a prediction rule is measured by its average risk defined as

$$R(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(Y_t - f_{\boldsymbol{\theta}t})^2\right].$$

Thus, a natural strategy for choosing a prediction rule from the data consists in minimizing the empirical risk. The empirical risk minimizer (ERM) is defined as

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} R_T(\boldsymbol{\theta}), \text{ where } R_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(Y_t - f_{\boldsymbol{\theta}t})^2. \tag{2}$$

If more than one prediction rule achieves the minimum, we may pick one arbitrarily. Clearly, the ERM in (2) corresponds to the classic least squares estimator. We sometimes denote $\hat{\boldsymbol{\theta}}$ as $\hat{\boldsymbol{\theta}}(\mathcal{D})$ to emphasize that the ERM is a function of the data $\mathcal{D}$. The problem we have described so far is known as linear regression in statistics and econometrics, whereas in learning theory, it is known as linear aggregation (Emery, Nemirovski, and Voiculescu, 2000).

The accuracy of the ERM is measured by its conditional average risk defined as

$$R(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{f}_t)^2 \,\middle|\, \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathcal{D}')\right], \tag{3}$$

where $\hat{f}_t = \hat{\theta}_1 X_{1t} + \cdots + \hat{\theta}_p X_{pt}$ and $\mathcal{D}'$ denotes an independent copy of the data $\mathcal{D}$. The performance measure in (3) can be interpreted as the risk of the ERM obtained from the "training data" $\mathcal{D}'$ over the "validation data" $\mathcal{D}$. This performance measure allows us to keep our analysis close to the bulk of contributions in the learning theory literature (which typically focus on the analysis of i.i.d. data) and facilitates comparisons. We also consider as an alternative accuracy measure the conditional out-of-sample average risk of the ERM, which is more attractive for time series applications. The alternative measure leads a to similar result at the expense of introducing additional notation.

The main objective of this paper is to obtain a bound on the performance of the ERM relative to the optimal risk that can be achieved within the given class of prediction rules. We aim to establish a bound $B_T(p)$ such that $B_T(p) \to 0$ as $T \to \infty$ for which

$$R(\hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} R(\boldsymbol{\theta}) + B_T(p) \tag{4}$$

holds, with high probability, for all (sufficiently large) T. The inequality in (4) is commonly referred to as an *oracle inequality*. Oracle inequalities such as (4) provide non-asymptotic guarantees on the performance of the ERM. The inequality in (4) implies that empirical risk minimization achieves asymptotically the best

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

performance that is possible to attain in the class. We emphasize that in this paper, we study the performance of the ERM for large-dimensional linear regression, meaning that in our analysis, we assume that the number of predictors p is not negligible relative to T (in a sense to be spelled out precisely below). Establishing bounds on the performance of the ERM is a classic problem in learning theory. There is a fairly extensive literature that has studied this problem in the i.i.d. setting (Audibert and Catoni, 2011). The literature conveys that the best possible rate for $B_T(p)$ is of the order p/T, which is referred to as the optimal rate of linear aggregation (Tsybakov, 2003).

The main contribution of this paper consists in establishing oracle inequalities for the ERM when the data are dependent and heavy-tailed. Our analysis covers both the cases of identically and heterogeneously distributed observations (using the jargon of White, 2001). In particular, our main results establish that the ERM achieves the optimal rate of linear aggregation (up to a $\log(T)$ factor) in a dependent data setting. Our analysis highlights a trade-off between the dependence and moment properties of the data on the one hand, and the number of predictors on the other. In particular, we show that the higher the dependence and the lower the number of moments of the data, the lower the maximum rate of growth allowed for the number of predictors. We emphasize that our analysis is nonparametric, in the sense, that the relationship between the regressand and the regressors is assumed to be unknown. Lastly, we remark that the performance bound we recover depends transparently on constants that are straightforward to interpret.

Four remarks are in order before we proceed. First, this work establishes prediction performance guarantees for empirical risk minimization/least squares estimation with dependent data in a large-dimensional setting. These results allow us to determine under which conditions least squares estimation is a reliable estimation strategy in a large-dimensional setup and to appraise more precisely the gains of estimation methodologies specifically designed for such a setup. It is important to acknowledge that estimation methodologies designed for large-dimensional settings (for instance, LASSO) typically achieve substantially better performance guarantees than the ones obtained here. However, these gains come at the expense of additional assumptions. In fact, the performance guarantees obtained here are optimal (up to a logarithmic factor; Tsybakov, 2003).

Second, this paper has a number of connections with the nonparametric literature and, in particular, with nonparametric series methods (Stone, 1985; Andrews, 1991; Newey, 1997; Chen and Shen, 1998; Chen, 2006; Tsybakov, 2014; Belloni et al., 2015). Among these papers, we remark that Chen and Shen (1998) is the only one that considers a non i.i.d. data setup. Let $\{(Y_t, \boldsymbol{W}_t')'\}_{t=1}^T$ be a strictly stationary sequence of random vectors in $\mathcal{Y} \times \mathcal{W} \subset \mathbb{R} \times \mathbb{R}^d$. Then our framework subsumes the problem of estimating the conditional mean of Y_t given $\boldsymbol{W}_t$ on the basis of the approximation given by

$$\mathbb{E}(Y_t|\boldsymbol{W}_t) \approx \theta_1 f_1(\boldsymbol{W}_t) + \cdots + \theta_p f_p(\boldsymbol{W}_t),$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

where $\{f_i\}$ with $f_i : \mathcal{W} \to \mathbb{R}$ is a collection of functions (e.g., B-splines) called a dictionary. We emphasize that, in some sense, our framework is more general since our focus lies on the estimation of the optimal linear prediction rule rather than the conditional mean.

Third, the literature on empirical risk minimization and oracle inequalities for dependent data has been rapidly developing in recent years. Notable contributions in this area include the works of Jiang and Tanner (2010), Fan, Liao, and Mincheva (2011), Caner and Knight (2013), Liao and Phillips (2015), and Miao, Phillips, and Su (2023). We remark that one of the challenges of this literature is that it is not straightforward to apply the theoretical machinery used in learning theory in a dependent data setting. In fact, as forcefully argued in Mendelson (2015), several of the standard results on empirical risk minimization used in learning theory assume i.i.d. bounded data and cannot be extended beyond this setup. In this work, we rely on a proof strategy based on the so-called *small-ball* method developed by Shahar Mendelson and Guillaume Lecué (Mendelson, 2015; Lecué and Mendelson, 2016). The small-ball method allows us to establish sharp bounds on the performance of the ERM under fairly weak moment and dependence assumptions.

Fourth, our analysis aims to provide large-dimensional analogs of some of the classic results of White (2001) for fixed-dimensional linear regression with dependent data. We shall point out the differences between those results and the ones established here.

This paper is related to various strands of the literature. First, it is related to the literature on empirical risk minimization for linear aggregation, which includes Birge and Massart (1998), Bunea, Tsybakov, and Wegkamp (2007), Audibert and Catoni (2011), and Lecué and Mendelson (2016). Second, it is related to the literature on empirical risk minimization for heavy-tailed data, which includes Audibert and Catoni (2011) and Brownlees, Joly, and Lugosi (2015). Third, it is related to the literature on empirical risk minimization for dependent data. In particular, this contribution is close to Jiang and Tanner (2010). Fourth, this paper is related to the vast literature on nonparametric estimation and nonparametric series methods, which includes Chen (2006) and Belloni et al. (2015). Li and Racine (2006) contains a number of important results and references to this literature. Fifth, it is related to the literature on the small-ball method, which includes Mendelson (2018), Lecué and Mendelson (2017), and Lecué and Mendelson (2018). Sixth, it is related to the vast literature on machine learning and large-dimensional modeling, which includes (in econometrics) Kock and Callot (2015), Medeiros and Mendes (2016), Garcia, Medeiros, and Vasconcelos (2017) and Babii, Ghysels, and Striaukas (2023). Hastie, Tibshirani, and Friedman (2001), and Wainwright (2019) contain a number of important results and references to this literature.

The rest of the paper is structured as follows: Section 2 contains preliminaries, additional notation, and assumptions. Section 3 contains an oracle inequality for linear regression with heterogeneously distributed observations. Section 4 contains

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

an analogous result for identically distributed observations. Section 5 contains extensions of the baseline results. Concluding remarks follow in Section 6. All proofs are in the Appendix.

2. NOTATION, PRELIMINARIES, AND ASSUMPTIONS

We introduce the notation used in the remainder of the paper. For a generic vector $\boldsymbol{x} \in \mathbb{R}^d$, we define $\|\boldsymbol{x}\|_r$ as $[\sum_{i=1}^d |x_i|^r]^{1/r}$ for $1 \leq r < \infty$ and $\max_{i=1,\ldots,d} |x_i|$ for $r = \infty$. For a generic random variable $X \in \mathbb{R}$, we define $\|X\|_{L_r}$ as $[\mathbb{E}(|X|^r)]^{1/r}$ for $1 \leq r < \infty$ and $\inf\{a : \mathbb{P}(|X| > a) = 0\}$ for $r = \infty$. For a positive semi-definite matrix $\mathbf{M}$, we use $\mathbf{M}^{\frac{1}{2}}$ to denote the positive semi-definite square root matrix of $\mathbf{M}$ and $\mathbf{M}^{-\frac{1}{2}}$ to denote the generalized-inverse of $\mathbf{M}^{\frac{1}{2}}$.

In this section, we establish a preliminary result and introduce the main assumptions required in our analysis. All results and assumptions are stated for the case of heterogeneously distributed observations. Clearly, these simplify in a straightforward manner if the observations are identically distributed.

We begin by establishing the existence of the optimal prediction rule, that is the *oracle*. Lemma 1 states that there exists an optimal $\boldsymbol{\theta}^*$ that satisfies a Pythagorean-type identity. We remark that the assumptions of Lemma 1 are fairly weak and, in particular, weaker than what we require for the analysis of the ERM.

LEMMA 1. *Let* $\{Y_t\}_{t=1}^T$ *satisfy* $\sup_{1\leq t\leq T} ||Y_t||_{L_2} < \infty$ *and* $\sup_{1\leq i\leq p} \sup_{1\leq t\leq T} ||X_{it}||_{L_2} < \infty$.
Then:

(i) *there exists a* $\boldsymbol{\theta}^* \in \mathbb{R}^p$ *such that*

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} R(\boldsymbol{\theta});$$

(ii) $\boldsymbol{\theta}^*$ *is such that for any* $\boldsymbol{\theta} \in \mathbb{R}^p$ *it holds that*

$$\frac{1}{T}\sum_{t=1}^T \|Y_t - f_t^*\|_{L_2}^2 + \frac{1}{T}\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta}\,t}\|_{L_2}^2 = \frac{1}{T}\sum_{t=1}^T \|Y_t - f_{\boldsymbol{\theta}\,t}\|_{L_2}^2,$$

where $f_t^* = f_{\boldsymbol{\theta}^*\,t}$;

(iii) *if* $\sum_{t=1}^T \mathbb{E}\boldsymbol{X}_t\boldsymbol{X}_t'$ *is positive definite then* $\boldsymbol{\theta}^*$ *is unique.*

Next, we lay out the assumptions we require to establish the properties of the ERM.

A.1 (Moments). *The sequences* $\{Y_t\}_{t=1}^T$, $\{\boldsymbol{X}_t\}_{t=1}^T$, $\{f_t^*\}_{t=1}^T$ *satisfy* $\sup_{1\leq t\leq T} \|Y_t\|_{L_{r_m}} \leq K_m$, $\sup_{1\leq i\leq p} \sup_{1\leq t\leq T} \|X_{it}\|_{L_{r_m}} \leq K_m$ *and* $\sup_{1\leq i\leq p} \sup_{1\leq t\leq T} \|(Y_t - f_t^*)X_{it}\|_{L_{r_m}} \leq K_m$, *for some* $K_m \geq 1$ *and* $r_m > 2$.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Assumption A.1 states that the regressand, predictors, and the product of the predictors and the forecast error of the optimal prediction rule have a number of moments strictly larger than two. The assumption also states that the r_mth moments are bounded by a constant K_m uniformly in t. A few comments are in order. First, this moment assumption is formulated as in White (2001, Chap. 3) in the analysis of linear regression with heterogeneous data. Alternatively, we may state this assumption for the forecast error of the optimal prediction rule and the predictors separately and require at least four moments to exist and to be uniformly bounded. Second, we assume $K_m \geq 1$ to obtain simpler expressions of some of the constants that appear in our analysis. Note that this is without loss of generality. Lastly, we emphasize that this assumption is weaker than what is assumed in a number of contributions on oracle inequalities for dependent data for large-dimensional models such as Jiang and Tanner (2010), Fan et al. (2011), and Kock and Callot (2015) which assume that all moments exist. We remark that assuming that the moments are uniformly bounded is fairly standard in the analysis of regression models with heterogeneous dependent data and that requiring more than two moments to exist is also required to establish consistency of the least squares estimator for fixed-dimensional linear regression (White, 2001, Chap. 3).

A.2 (Dependence). *Let $\mathcal{F}_{-\infty}^{s}$ and $\mathcal{F}_{s+l}^{\infty}$ be the σ-algebras generated by $\{(Y_t, \boldsymbol{X}_t')' : -\infty \leq t \leq s\}$ and $\{(Y_t, \boldsymbol{X}_t')' : s+l \leq t \leq \infty\}$, respectively, and define the α-mixing coefficients*

$$\alpha(l) = \sup_{s} \sup_{A \in \mathcal{F}_{-\infty}^{s}, B \in \mathcal{F}_{s+l}^{\infty}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\,\mathbb{P}(B)|.$$

The α-mixing coefficients satisfy $\alpha(l) \leq \exp(-K_\alpha l^{r_\alpha})$ for some $K_\alpha > 0$ and $r_\alpha > 0$.

Assumption A.2 states that the sequence $\{(Y_t, \boldsymbol{X}_t')'\}_{t=1}^{T}$ is strongly mixing with geometrically decaying mixing coefficients. The definition of the mixing coefficients is as in White (2001, Defn. 3.42) and does not hinge on the data generating process being stationarity. See also Su and White (2010) for the analysis of α-mixing processes that are not required to be stationary. Note that while this is a stronger assumption than what is required by classical results for consistency and asymptotic normality for the (finite-dimensional) linear regression model that rely on polynomial α-mixing (White, 2001, Chap. 3), geometric α-mixing is commonly used in the analysis of large-dimensional time series models (Jiang and Tanner, 2010; Fan et al., 2011; Kock and Callot, 2015). Moreover, geometric α-mixing is satisfied by many commonly encountered processes such as ARMA and GARCH (Meitz and Saikkonen, 2008).

A.3 (Number of Predictors). *The number of predictors satisfies $p = \lfloor K_p T^{r_p} \rfloor$ for some $K_p > 0$ and $0 \leq r_p < \frac{r_\alpha}{r_\alpha + 1} \wedge \frac{r_m - 2}{2}$.*

Assumption A.3 states that the number of predictors is a function of T. This assumption allows the number of predictors to be constant or to grow sublinearly

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

in T. Importantly, the bound on the rate of growth of the number of predictors p depends on the number of moments and the amount of dependence of the data. The more moments and the less dependence, the higher the maximum rate of growth of the number of predictors. If the data have at least four moments, then the number of predictors is only constrained by the amount of dependence in the data.

A.4 (Eigenvalues). *Define* $\Sigma_t = \mathbb{E}(\boldsymbol{X}_t \boldsymbol{X}_t')$ *and let* $\lambda_{\min}(\Sigma_t)$ *and* $\lambda_{\max}(\Sigma_t)$ *be the smallest and largest eigenvalue of* Σ_t*, respectively. Then the sequence* $\{\Sigma_t\}_{t=1}^T$ *satisfies (i)* $\underline{\lambda} \leq \inf_{1\leq t\leq T} \lambda_{\min}(\Sigma_t)$ *for some* $0 < \underline{\lambda}$ *and (ii)* $\sup_{1\leq t\leq T} \lambda_{\max}(\Sigma_t) \leq \overline{\lambda}$ *for some* $0 < \overline{\lambda} < \infty$.

Assumption A.4 states that the eigenvalues of the covariance matrix of the predictors are bounded from above and bounded away from zero uniformly in t. The assumption that the smallest eigenvalue is bounded away from zero is fairly standard (Newey, 1997). Notice that $\boldsymbol{\theta}^*$ is unique when A.4(i) holds, by Lemma 1(iii). Assuming that the largest eigenvalue of the covariance matrix of the predictors is bounded above uniformly in t is more restrictive. We remark that, as we shall see in detail below, under the additional assumption of identically distributed observations these constraints can be relaxed. In what follows we shall also use the constant $K_\Sigma = \overline{\lambda}/\underline{\lambda}$, which is an upper bound on the condition number of the matrices $\{\Sigma_t\}_{t=1}^T$ and measures the maximum degree of collinearity between the predictors.

A.5 (Distribution). *Consider the sequence of random vectors* $\{\boldsymbol{Z}_t\}_{t=1}^T$ *with* $\boldsymbol{Z}_t = \Sigma_t^{-\frac{1}{2}} \boldsymbol{X}_t$. *Then* $\sup_{1\leq t\leq T} \mathbb{P}(\boldsymbol{Z}_t \in E) \leq K_{\boldsymbol{Z}} \mathbb{P}(\boldsymbol{S} \in E)$ *holds for some p-dimensional spherical random vector* $\boldsymbol{S}$*, some positive constant* $K_{\boldsymbol{Z}}$ *and any* $E \in \mathcal{B}(\mathbb{R}^p)$. *The density of* $\boldsymbol{S}$ *exists and the marginal densities of the components of* $\boldsymbol{S}$ *are bounded from above.*

Assumption A.5 is required to establish upper bounds on the probability of a certain event associated with the vector of predictors $\boldsymbol{X}_t$ in one of the intermediate propositions of our analysis. The probability of this event boils down to a multiple integral that can be expressed using n-spherical coordinates. The spherical distribution bound in A.5 makes it easy to compute such an integral after the n-spherical coordinates transformation. We conjecture that the assumption could be relaxed, however, this would be at the expense of more tedious computations. That being said, the family of spherical distributions is fairly large and includes the appropriately standardized versions of the multivariate Gaussian, Student t, Cauchy, and uniform[1] distributions.[2] Moreover, finite mixtures of spherical distributions are also spherical. Assumption A.5 may be interpreted as a generalization of the

[1] To be precise, the multivariate uniform distribution over the sphere.

[2] For more details on the class of spherical distributions, we refer to Fang, Kotz, and Ng (1990).

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

bounded density assumption typically encountered in the nonparametric literature (Newey, 1997; Li and Racine, 2006; Hansen, 2008). Bounded density assumptions are also formulated in Jiang and Tanner (2010) in the analysis of empirical risk minimization for time series data with bounded support. Last, we remark that this assumption allows for weaker moment conditions than what is imposed by Assumption A.1.

A.6 (Identification/Small-ball). *The sequence* $\{\boldsymbol{X}_t\}_{t=1}^T$ *satisfies, for each* $t = 1, \ldots, T$ *and for each* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$,

$$\mathbb{P}\left(|f_{\boldsymbol{\theta}_1 t} - f_{\boldsymbol{\theta}_2 t}| \geq \kappa_1 \|f_{\boldsymbol{\theta}_1 t} - f_{\boldsymbol{\theta}_2 t}\|_{L_2}\right) \geq \kappa_2,$$

for some $\kappa_1 > 0$ *and* $\kappa_2 > 0$.

Assumption A.6 is the so-called small-ball assumption, and it is stated here as it is formulated in Lecué and Mendelson (2016). This assumption can be interpreted as an identification condition. If we define $\boldsymbol{v} = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$, then the condition is equivalent to $\mathbb{P}\left(|\boldsymbol{v}'\boldsymbol{X}| \geq \kappa_1 \|\boldsymbol{v}'\boldsymbol{X}\|_{L_2}\right) \geq \kappa_2$, which can be seen as requiring that the random variable $\boldsymbol{v}'\boldsymbol{X}$ does not have excessive mass in a neighborhood around zero. We remark that the constants κ_1 and κ_2 measure the strength of the identification in the sense that the larger the value of these constants the stronger the identification condition is. In Section 5, we establish alternative identification assumptions that in turn imply Assumption A.6.

3. DEPENDENT HETEROGENEOUSLY DISTRIBUTED OBSERVATIONS

The ERM performance bound that we derive in this section depends on a constant related to the variance of the gradient of the empirical risk evaluated at the optimal prediction rule (after an appropriate rescaling), that is, $\mathrm{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T (Y_t - f_t^*)\boldsymbol{X}_t\right)$. As is well known, in the standard large sample analysis of linear regression the asymptotic variance of the least squares estimator is typically expressed as a function of the limit of this quantity (White, 2001, Chap. 5). In our analysis, the ERM performance depends on an upper bound on the diagonal elements of this quantity that is given by

$$K_{\sigma^2} = K_m^2 \left(1 + 128 \frac{r_m}{r_m - 2} \sum_{l=1}^{\infty} \alpha(l)^{1 - \frac{2}{r_m}}\right).$$

It is possible to make substantially smaller choices of this constant if we make simplifying assumptions on the setup of our analysis. We explore this in more detail in Section 4.

We can now state the main result of this section.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

THEOREM 1. *Suppose Assumptions A.1–A.6 are satisfied. Then, for all T sufficiently large, the ERM defined in (2) satisfies*

$$R(\hat{\boldsymbol{\theta}}) \leq R(\boldsymbol{\theta}^*) + K_{\sigma^2} \frac{K_\Sigma^3}{\underline{\lambda}} \left(\frac{48}{\kappa_1^2 \kappa_2} \right)^2 \frac{p \log(T)}{T}, \tag{5}$$

with probability at least $1 - 3K_p(2K_m)^{r_m}/(K_{\sigma^2}^{\frac{1}{2}} \log(T)) - o(\log(T)^{-1})$.

The theorem establishes that the ERM for large-dimensional linear regression with heterogeneous dependent data achieves the optimal rate of linear aggregation (up to a $\log(T)$ factor). We remark that A.3 implies that $(p\log(T))/T \to 0$ as $T \to \infty$, which makes the inequality in the theorem an oracle inequality. Note that the bound on the performance of the ERM is proportional to quantities that are associated with a larger asymptotic variability of the least squares estimator. We remark that Theorem 1 may be seen as a non-asymptotic version of classic asymptotic results in the series estimation literature, which establish optimality of the nonparametric least squares estimator. In fact, the convergence rate of p/T (up to a $\log(T)$ factor) is the same as the rate obtained (for instance) in Belloni et al. (2015, Thm. 4.1).

It is interesting to compare Theorem 1 with an analogous result for i.i.d. data. The following result in Lecué and Mendelson (2016, Cor. 1.2) is taken as benchmark.

THEOREM. *Consider the linear regression model*

$$Y_t = \boldsymbol{X}_t' \boldsymbol{\theta}^* + \epsilon_t, \qquad t = 1, \ldots, T,$$

where $\{\boldsymbol{X}_t\}$ *and* $\{\epsilon_t\}$ *are sequences of i.i.d. random variables with* $\mathbb{E}(\epsilon_t) = 0$, $\mathrm{Var}(\epsilon_t) = \sigma^2$, *and* ϵ_t *is independent of* $\boldsymbol{X}_t$. *Assume that there are constants* κ_1 *and* κ_2 *such that*

$$\mathbb{P}\left(|f_{\boldsymbol{\theta}_1 t} - f_{\boldsymbol{\theta}_2 t}| \geq \kappa_1 \|f_{\boldsymbol{\theta}_1 t} - f_{\boldsymbol{\theta}_2 t}\|_{L_2} \right) \geq \kappa_2,$$

for all $\boldsymbol{\theta} \in \mathbb{R}^p$. *Then, for all* $T > (400)^2 p/\kappa_2^2$ *and* $x > 0$ *we have that the ERM defined in (2) satisfies*

$$R(\hat{\boldsymbol{\theta}}) \leq R(\boldsymbol{\theta}^*) + \sigma^2 \left(\frac{16}{\kappa_1^2 \kappa_2} \right)^2 \frac{p}{T} x,$$

with probability at least $1 - \exp(-\kappa_2 T/4) - (1/x)$.

As is immediate to see, we recover an analogous bound to what is established in Lecué and Mendelson (2016). The constant that appears in our risk bound in (5) is much larger than the one in this benchmark result. However, we remark that below we obtain a more favorable bound by simplifying the setup of our analysis. Also, we remark that the result above relies on assuming that the "true model" exists. Lecué and Mendelson (2016) also have results that do not depend on such

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

an assumption but rely on stronger assumptions on the prediction errors of the optimal forecast.

We conclude this section with a sketch of the proof. This is an elegant argument based on Lecué and Mendelson (2016). Define the empirical risk differential for $\boldsymbol{\theta} \in \mathbb{R}^p$ as

$$\widehat{\mathcal{L}}_{\boldsymbol{\theta}} = R_T(\boldsymbol{\theta}) - R_T(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{t=1}^{T}(f_t^* - f_{\boldsymbol{\theta}\,t})^2 + \frac{2}{T}\sum_{t=1}^{T}(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}\,t}).$$

The proof is based on showing that if the condition

$$\frac{1}{T}\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}\,t}\|_{L_2} > \frac{48K_{\sigma^2}^{\frac{1}{2}}K_\Sigma}{\underline{\lambda}\kappa_1^2\kappa_2}\sqrt{\frac{p\log(T)}{T}} \tag{6}$$

holds, then we have that

$$\frac{1}{T}\sum_{t=1}^{T}(f_t^* - f_{\boldsymbol{\theta}\,t})^2 > \left|\frac{2}{T}\sum_{t=1}^{T}(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}\,t})\right| \tag{7}$$

with high probability. This, in turn, implies that for any $\boldsymbol{\theta}$ that satisfies (6) we have $\widehat{\mathcal{L}}_{\boldsymbol{\theta}} > 0$. Since the ERM $\hat{\boldsymbol{\theta}}$ must satisfy $\widehat{\mathcal{L}}_{\hat{\boldsymbol{\theta}}} \leq 0$ then, conditional on the same events, we must have that

$$\frac{1}{T}\sum_{t=1}^{T}\|f_t^* - \hat{f}_t\|_{L_2} \leq \frac{48K_{\sigma^2}^{\frac{1}{2}}K_\Sigma}{\underline{\lambda}\kappa_1^2\kappa_2}\sqrt{\frac{p\log(T)}{T}},$$

which, in turn, implies that

$$R(\hat{\boldsymbol{\theta}}) - R(\theta^*) = \frac{1}{T}\sum_{t=1}^{T}\|f_t^* - \hat{f}_t\|_{L_2}^2 \leq K_{\sigma^2}\frac{K_\Sigma^2}{\underline{\lambda}}\left(\frac{48}{\kappa_1^2\kappa_2}\right)^2\frac{p\log(T)}{T},$$

by an application of Lemma 1.

The following two propositions are key in establishing that the inequality in (7) holds with high probability; and thus, to determine the risk bound in Theorem 1.

PROPOSITION 1. *Suppose Assumptions A.2–A.6 are satisfied. Then, for all T sufficiently large and any $\boldsymbol{\theta} \in \mathbb{R}^p$,*

$$\frac{1}{T}\sum_{t=1}^{T}(f_t^* - f_{\boldsymbol{\theta}\,t})^2 \geq \frac{\kappa_1^2\kappa_2}{2K_\Sigma}\frac{1}{T}\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}\,t}\|_{L_2}^2$$

holds with probability at least $1 - 8T^{-1} - o(T^{-1})$.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

PROPOSITION 2. *Suppose Assumptions A.1–A.4 are satisfied. Then, for all T sufficiently large and any $\boldsymbol{\theta} \in \mathbb{R}^p / \{\boldsymbol{\theta}^*\}$,*

$$\left| \frac{1}{T} \sum_{t=1}^{T} (Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta} t}) \right| \leq 12 \sqrt{\frac{K_{\sigma^2}}{\underline{\lambda}}} \frac{1}{T} \sum_{t=1}^{T} \|f_t^* - f_{\boldsymbol{\theta} t}\|_{L_2} \sqrt{\frac{p \log(T)}{T}}$$

holds with probability at least $1 - 3K_p(2K_m)^{r_m}/(K_{\sigma^2}^{\frac{1}{2}} \log(T)) - o(\log(T)^{-1})$.

Both propositions exploit a Bernstein-type inequality for α-mixing sequences from Liebscher (1996, Thm. 2.1); (based on the famous covariance inequality of Rio 1995). Proposition 1 uses a covering argument similar to the one used in Jiang and Tanner (2010) and Hansen (2008). Proposition 2 relies on the Bernstein-type inequality and a classic truncation trick used in, for instance, Hansen (2008). See also Dendramis, Giraitis, and Kapetanios (2021) and Babii et al. (2023) for recent developments on concentration inequalities for dependent data with applications to large-dimensional estimation problems.

4. DEPENDENT IDENTICALLY DISTRIBUTED OBSERVATIONS

The constant term in the bound of Theorem 1 can be improved by assuming stationarity.

A.7 (Stationarity). *The sequence of random vectors $\{(Y_t, \boldsymbol{X}_t')'\}_{t=1}^T$ is stationary.*

We remark that Assumptions A.2 and A.7 imply that the data are ergodic.

In the stationary case, it is convenient to state the moment assumption differently.

A.1* (Moments). *The sequences $\{Y_t\}_{t=1}^T$, $\{\boldsymbol{X}_t\}_{t=1}^T$, $\{f_t^*\}_{t=1}^T$ and $\{\boldsymbol{Z}_t\}_{t=1}^T$ with $\boldsymbol{Z}_t = \Sigma_t^{-\frac{1}{2}} \boldsymbol{X}_t$ satisfy $\|Y_t\|_{L_{r_m}} \leq K_m$, $\sup_{1 \leq i \leq p} \|X_{it}\|_{L_{r_m}} \leq K_m$ and $\sup_{1 \leq i \leq p} \|(Y_t - f_t^*) Z_{it}\|_{L_{r_m}} \leq K_m$, for some $K_m \geq 1$ and $r_m > 2$.*

The difference between A.1 and A.1* is that the former assumption bounds the r_mth moment of $(Y_t - f_t^*)X_{it}$ whereas the latter bounds the r_mth moment of $(Y_t - f_t^*)Z_{it}$.

In the stationary case, the assumption on the eigenvalues of Σ, Assumption A.4, can be dropped. In fact, as we show in the proof of Theorem 2, A.1* implies that $\lambda_{\max}(\Sigma) \leq K_m^2 p$. This allows the set of predictors to be generated by a factor model (Forni et al., 2000; Bai and Ng, 2002; Stock and Watson, 2002; Onatski, 2012). Additionally, $\lambda_{\min}(\Sigma)$ is allowed to be zero. This allows the set of predictors to contain some predictors that are perfectly correlated.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Before stating the main result of this section, we introduce a new constant

$$K'_{\sigma^2} = K_m^2 \left(1 + 32 \frac{r_m}{r_m - 2} \sum_{l=1}^{\infty} \alpha(l)^{1-\frac{2}{r_m}}\right).$$

This constant plays the same role as K_{σ^2} and can be interpreted as an upper bound on the diagonal elements of $\mathrm{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (Y_t - f_t^*)\boldsymbol{Z}_t\right)$. Note that $K'_{\sigma^2} \leq K_{\sigma^2}$.

We can now state the main result of this section.

THEOREM 2. *Suppose Assumptions A.1*, A.2, A.3, A.5–A.7 are satisfied. Then, for all T sufficiently large, the ERM defined in (2) satisfies*

$$R(\hat{\boldsymbol{\theta}}) \leq R(\boldsymbol{\theta}^*) + K'_{\sigma^2} \left(\frac{48}{\kappa_1^2 \kappa_2}\right)^2 \frac{p \log(T)}{T}$$

with probability at least $1 - 3K_p K_m^{r_m} / ((K'_{\sigma^2})^{\frac{1}{2}} \log(T)) - o(\log(T)^{-1})$.

Note that the bound in Theorem 2 does not depend on the eigenvalues of Σ. Inspection of the proof shows that if $\{(Y_t, \boldsymbol{X}_t')'\}$ is an i.i.d. sequence, $Y_t - f_t^*$ is independent of $\boldsymbol{X}_t$ and $\mathrm{Var}(Y_t - f_t^*) = \sigma^2$ the bound in Theorem 2 becomes

$$R(\boldsymbol{\theta}^*) + \sigma^2 \left(\frac{48}{\kappa_1^2 \kappa_2}\right)^2 \frac{p \log(T)}{T},$$

which is close to the bound established in Lecué and Mendelson (2016, Cor. 1.2).

5. ADDITIONAL RESULTS

5.1. Alternative Risk Definition

It is important to emphasize that the performance measure defined in (3) is the average risk of the prediction rule over the data $\mathcal{D}$ when $\hat{\boldsymbol{\theta}}$ is estimated using an independent copy of the data $\mathcal{D}'$. This measure may have limited appeal for time series applications since a forecaster typically does not have access to an independent copy of the data. Alternative more appropriate risk measures may be introduced to evaluate the performance of the risk minimizer in a time series context.

Assume that we are interested in predicting the out-of-sample observations $\{(Y_t, \boldsymbol{X}_t')'\}_{t=T+1}^{T+H}$ on the basis of the prediction rule estimated from the in-sample observations $\{(Y_t, \boldsymbol{X}_t')'\}_{t=1}^{T}$. For simplicity, here we focus only on the case of identically distributed observations. We define the average out-of-sample risk of $\boldsymbol{\theta}$ as

$$R_{\mathrm{oos}}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{1}{H} \sum_{t=T+1}^{T+H} (Y_t - f_{\boldsymbol{\theta} t})^2\right],$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

and we measure the accuracy of the ERM $\hat{\boldsymbol{\theta}}$ using the conditional out-of-sample average risk defined as

$$R_{\text{oos}}(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[\frac{1}{H}\sum_{t=T+1}^{T+H}(Y_t - f_{\hat{\boldsymbol{\theta}}\,t})^2 \,\middle|\, (Y_T, \boldsymbol{X}_T')', \ldots, (Y_1, \boldsymbol{X}_1')'\right].$$

For the following result, a slightly stronger version of Assumption A.1 is needed.

A.1** (Moments). *The sequences* $\{Y_t\}_{t=1}^T$, $\{f_t^*\}_{t=1}^T$ *and* $\{\boldsymbol{X}_t\}_{t=1}^T$ *satisfy* $\sup_{1\le i\le p}\sup_{1\le t\le T}\|Y_t - f_t^*\|_{L_{r_m}} \le K_m \sup_{1\le t\le T}\|Y_t\|_{L_{r_m}} \le K_m$ *and* $\sup_{1\le i\le p}\sup_{1\le t\le T}\|X_{it}\|_{L_{r_m}} \le K_m$ *for some* $K_m \ge 1$ *and* $r_m > 4$.

If we define $K_H = 24(K_m^2/\underline{\lambda})\sum_{l=1}^{\infty}\alpha(l)^{\frac{1}{2}}$, we can establish the following theorem.

THEOREM 3. *Suppose Assumptions A.1**, A.2, A.3, A.4(i), A.5–A.7 are satisfied. Then, for all T sufficiently large, the ERM defined in (2) satisfies*

$$R_{\text{oos}}(\hat{\boldsymbol{\theta}}) \le R_{\text{oos}}(\boldsymbol{\theta}^*) + K'_{\sigma^2}\left(\frac{48}{\kappa_1^2\kappa_2}\right)^2\frac{p\log(T)}{T} + K_H\frac{p\log(T)}{H}$$

with probability at least $1 - (6K_pK_m^{r_m}+1)/((K'_{\sigma^2})^{\frac{1}{2}}\log(T)) - o(\log(T)^{-1})$.

A key ingredient in the proof of Theorem 3 is Ibragimov's inequality (Ibragimov, 1962), which bounds the expected value of the difference between the conditional and unconditional expectation as a function of the α-mixing coefficients. It is important to remark that the theorem requires $(p\log(T))/H \to 0$ in order to have that $R_{\text{oos}}(\hat{\boldsymbol{\theta}}) - R_{\text{oos}}(\boldsymbol{\theta}^*) \to 0$. In other words, there exists a "wedge" between $R_{\text{oos}}(\hat{\boldsymbol{\theta}})$ and $R_{\text{oos}}(\boldsymbol{\theta}^*)$ that only vanishes as the forecast horizon H grows large. This may be intuitively explained as follows. The ERM $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}^*$, the minimizer of $R(\boldsymbol{\theta})$. However, the minimizers of $R(\boldsymbol{\theta})$ and $R_{oos}(\boldsymbol{\theta})$ are not guaranteed to be same for finite H and the difference between the two only vanishes as the forecast horizon H grows large.

5.2. Small-Ball Assumption

It is possible to introduce alternative assumptions that imply the small-ball condition stated in A.6. For example, as Lecué and Mendelson (2016) remark, the small-ball condition holds when the L_2 and L_4 norms of $f_{\boldsymbol{\theta}\,t}$ are equivalent. More precisely, if for each $\boldsymbol{\theta} \in \mathbb{R}^p$ (and all $t = 1, \ldots, T$) it holds that $\|f_{\boldsymbol{\theta}\,t}\|_{L_4} \le C\|f_{\boldsymbol{\theta}\,t}\|_{L_2}$, for some constant C (that does not depend on $\boldsymbol{\theta}$ or t). We remark that norm equivalence conditions are commonly used in the literature to establish the properties of empirical risk minimization (see, for instance, Audibert and Catoni, 2011).

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

To give a concrete example, below we show that if the distribution of the standardized predictors $\boldsymbol{Z}_t = \Sigma^{-\frac{1}{2}}\boldsymbol{X}_t$ is spherical then the small-ball assumption is satisfied.

A.6* (Spherical Density). *Consider the sequence of random vectors* $\{\boldsymbol{Z}_t\}_{t=1}^T$ *with* $\boldsymbol{Z}_t = \Sigma_t^{-\frac{1}{2}}\boldsymbol{X}_t$. *Then for each* $t = 1, \ldots, T$ *it holds that* $\boldsymbol{Z}_t \sim \boldsymbol{S}$ *where* $\boldsymbol{S}$ *is a p-dimensional spherical random vector that satisfies* $\sup_{1\leq i\leq p} \|S_i\|_{L_4} < \infty$.

Lemma 2. *Suppose Assumption A.6* holds. Then, for each* $t = 1, \ldots, T$ *and for each* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$, $\mathbb{P}\left(|f_{\boldsymbol{\theta}_1 t} - f_{\boldsymbol{\theta}_2 t}| \geq \kappa_1 \|f_{\boldsymbol{\theta}_1 t} - f_{\boldsymbol{\theta}_2 t}\|_{L_2}\right) \geq \kappa_2$ *holds for some* $\kappa_1 > 0$ *and* $\kappa_2 > 0$.

The proof uses the Paley–Zygmund inequality. A.6* can replace A.6 in Theorems 1 and 2.

6. CONCLUSION

This paper establishes oracle inequalities for the prediction risk of the ERM for large-dimensional linear regression. We generalize existing results by allowing the data to be dependent and heavy-tailed. Our main results show that the ERM achieves optimal performance (up to a logarithmic factor). The results have been established using the small-ball method, which is a powerful technique to obtain oracle inequalities. Future research includes extending these results to regularized empirical risk minimization, analogously to Lecué and Mendelson (2017, 2018).

APPENDICES

A. Proofs

Proof of Lemma 1. (*i*) The existence of $\boldsymbol{\theta}^*$ follows from the fact that $R(\boldsymbol{\theta})$ is quadratic. (*ii*) It is equivalent to show that $\boldsymbol{\theta}^*$ satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}\, t})] = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(Y_t - f_t^*)\boldsymbol{X}_t']\right)(\boldsymbol{\theta}^* - \boldsymbol{\theta}) = 0, \tag{A.1}$$

for any $\boldsymbol{\theta} \in \mathbb{R}^p$. We then have that (A.1) is implied by the first-order condition for a minimum for $R(\boldsymbol{\theta})$, as $\boldsymbol{\theta}^*$ is such that $\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[(Y_t - f_t^*)\boldsymbol{X}_t'] = 0$. (*iii*) This follows from the strict convexity of $R(\boldsymbol{\theta})$. □

Proof of Theorem 1. Define the empirical risk differential for an arbitrary $\boldsymbol{\theta} \in \mathbb{R}^p$ as

$$\widehat{\mathcal{L}}_{\boldsymbol{\theta}} = R_T(\boldsymbol{\theta}) - R_T(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{t=1}^{T}(f_t^* - f_{\boldsymbol{\theta}\, t})^2 + \frac{2}{T}\sum_{t=1}^{T}(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}\, t}).$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Assume that it holds that

$$\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2} > 48\sqrt{\frac{K_{\sigma^2}}{\underline{\lambda}}}\frac{K_\Sigma}{\kappa_1^2\kappa_2}\sqrt{\frac{p\log(T)}{T}}. \tag{A.2}$$

Conditioning on the events of Propositions 1 and 2, for all T sufficiently large, at least with probability $1-3K_p(2K_m)^{r_m}/(K_{\sigma^2}^{\frac{1}{2}}\log(T))-o(\log(T)^{-1})$, we have that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(f_t^*-f_{\boldsymbol{\theta}t})^2 &\overset{(a)}{\geq} \frac{\kappa_1^2\kappa_2}{2K_\Sigma}\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}^2 \overset{(b)}{\geq} \frac{\kappa_1^2\kappa_2}{2K_\Sigma}\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}\\ &\overset{(c)}{>} 24\sqrt{\frac{K_{\sigma^2}}{\underline{\lambda}}}\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}\sqrt{\frac{p\log(T)}{T}} \overset{(d)}{\geq} \left|\frac{2}{T}\sum_{t=1}^{T}(Y_t-f_t^*)(f_t^*-f_{\boldsymbol{\theta}t})\right|,\end{aligned}$$

where (a) follows from Proposition 1, (b) follows from Jensen's inequality, (c) follows from condition (A.2), and (d) follows from Proposition 2. Thus, conditional on the events of Propositions 1 and 2 and assuming (A.2) holds we have with high probability that $\widehat{\mathcal{L}}_{\boldsymbol{\theta}} > 0$. Since the ERM $\hat{\boldsymbol{\theta}}$ satisfies $\widehat{\mathcal{L}}_{\hat{\boldsymbol{\theta}}} \leq 0$ then conditional on the same events we have $\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-\hat{f}_t\|_{L_2} \leq \frac{48K_{\sigma^2}^{\frac{1}{2}}K_\Sigma}{\underline{\lambda}^{\frac{1}{2}}\kappa_1^2\kappa_2}\sqrt{\frac{p\log(T)}{T}}$. The claim follows from:

$$R(\hat{\boldsymbol{\theta}})-R(\boldsymbol{\theta}^*)=\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-\hat{f}_t\|_{L_2}^2 \leq \overline{\lambda}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2 \leq K_{\sigma^2}\frac{K_\Sigma^3}{\underline{\lambda}}\left(\frac{48}{\kappa_1^2\kappa_2}\right)^2\frac{p\log(T)}{T}, \tag{A.3}$$

where the first equality follows from Lemma 1 where the L_2 norm is conditional on $\{\hat{\theta}=\hat{\theta}(\mathcal{D}')\}$, the first inequality follows from $\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-\hat{f}_t\|_{L_2}^2 \leq \overline{\lambda}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2$, and the second inequality follows from $\underline{\lambda}^{\frac{1}{2}}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2 \leq \frac{1}{T}\sum_{t=1}^{T}\|f_t^*-\hat{f}_t\|_{L_2}$. □

Proof of Proposition 1. For any $\boldsymbol{\theta}\in\mathbb{R}^p\setminus\{\boldsymbol{\theta}^*\}$ (notice that A.4 implies that $\boldsymbol{\theta}^*$ is unique), define the standardized parameter vector $\boldsymbol{v}=(\boldsymbol{\theta}_*-\boldsymbol{\theta})/\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}^2}$ and note

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(f_t^*-f_{\boldsymbol{\theta}t})^2 &= \frac{\frac{1}{T}\sum_{t=1}^{T}(f_t^*-f_{\boldsymbol{\theta}t})^2}{\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}^2}\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}^2\\ &= \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{X}_t'\boldsymbol{v})^2\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}^2 \geq \frac{\kappa_1^2}{K_\Sigma}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\left\{|\boldsymbol{X}_t'\boldsymbol{v}|\geq\kappa_1K_\Sigma^{-1/2}\right\}}\frac{1}{T}\sum_{t=1}^{T}\|f_t^*-f_{\boldsymbol{\theta}t}\|_{L_2}^2.\end{aligned}$$

Let $g_{\boldsymbol{v}t}=\mathbb{1}_{\{|\boldsymbol{X}_t'\boldsymbol{v}|\geq\kappa_1K_\Sigma^{-1/2}\}}$, define $V=\{\boldsymbol{v}\in\mathbb{R}^p:\frac{1}{T}\mathbb{E}[\sum_{t=1}^{T}(\boldsymbol{X}_t'\boldsymbol{v})^2]=1\}$ and note that

$$\frac{1}{T}\sum_{t=1}^{T}g_{\boldsymbol{v}t}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}g_{\boldsymbol{v}t}+g_{\boldsymbol{v}t}-\mathbb{E}g_{\boldsymbol{v}t} \geq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}g_{\boldsymbol{v}t}-\sup_{\boldsymbol{v}\in V}\left|\frac{1}{T}\sum_{t=1}^{T}g_{\boldsymbol{v}t}-\mathbb{E}g_{\boldsymbol{v}t}\right|,$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

since the standardized parameter vector $\boldsymbol{\nu}$ belongs to V. Let $V_i = \{\boldsymbol{\nu} \in \mathbb{R}^p : \|\boldsymbol{\nu} - \boldsymbol{\nu}_i\|_2 \leq \delta\}$ with $\boldsymbol{\nu}_i \in V$ for $i = 1, \ldots, N_\delta$ denote a δ-covering of V. Then, we have that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\nu}\in V}\left|\frac{1}{T}\sum_{t=1}^{T} g_{\boldsymbol{\nu}t} - \mathbb{E}g_{\boldsymbol{\nu}t}\right| > \varepsilon\right) \leq \sum_{i=1}^{N_\delta}\mathbb{P}\left(\sup_{\boldsymbol{\nu}\in V_i}\left|\frac{1}{T}\sum_{t=1}^{T} g_{\boldsymbol{\nu}t} - \mathbb{E}g_{\boldsymbol{\nu}t}\right| > \varepsilon\right) \\
&\quad \leq \sum_{i=1}^{N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} g_{it} - \mathbb{E}g_{it}\right| > \frac{\varepsilon}{2}\right) \\
&\quad + \sum_{i=1}^{N_\delta}\mathbb{P}\left(\sup_{\boldsymbol{\nu}\in V_i}\left|\frac{1}{T}\sum_{t=1}^{T} g_{\boldsymbol{\nu}t} - \mathbb{E}g_{\boldsymbol{\nu}t} - \left(\frac{1}{T}\sum_{t=1}^{T} g_{it} - \mathbb{E}g_{it}\right)\right| > \frac{\varepsilon}{2}\right),
\end{aligned}$$

where $g_{it} = g_{\boldsymbol{\nu}_i t}$. Proposition B.1 establishes that (i) for each $\boldsymbol{\nu} \in V_i$, we have $|g_{\boldsymbol{\nu}t} - g_{it}| \leq \bar{g}_{it}$, where $\bar{g}_{it}$ is defined in that proposition and (ii) there exists a positive constant K_1 (that does not depend on i, t, and p) such that for all $\delta < K_\Sigma^{-1/2}/(2\bar{\lambda}^{1/2})$ we have that $\mathbb{E}\bar{g}_{it} \leq K_1 p^{1/2}\delta$. Set $\delta = \varepsilon/(8K_1 p^{1/2})$ and note that for all $\varepsilon < 4K_\Sigma^{-1/2}K_1 p^{1/2}/\bar{\lambda}^{1/2}$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\nu}\in V_i}\left|\frac{1}{T}\sum_{t=1}^{T} g_{\boldsymbol{\nu}t} - \mathbb{E}g_{\boldsymbol{\nu}t} - \left(\frac{1}{T}\sum_{t=1}^{T} g_{it} - \mathbb{E}g_{it}\right)\right| > \frac{\varepsilon}{2}\right) \\
&\quad = \mathbb{P}\left(\sup_{\boldsymbol{\nu}\in V_i}\left|\frac{1}{T}\sum_{t=1}^{T}(g_{\boldsymbol{\nu}t} - g_{it}) - (\mathbb{E}g_{\boldsymbol{\nu}t} - \mathbb{E}g_{it})\right| > \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\bar{g}_{it} + \mathbb{E}\bar{g}_{it}) > \frac{\varepsilon}{2}\right) \\
&\quad = \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\bar{g}_{it} - \mathbb{E}\bar{g}_{it}) > \frac{\varepsilon}{2} - \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\bar{g}_{it}\right) \overset{(a)}{\leq} \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\bar{g}_{it} - \mathbb{E}\bar{g}_{it}) > \frac{\varepsilon}{4}\right),
\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}\bar{g}_{it} \leq \varepsilon/8$. Finally, we have that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\nu}\in V}\left|\frac{1}{T}\sum_{t=1}^{T} g_{\boldsymbol{\nu}t} - \mathbb{E}g_{\boldsymbol{\nu}t}\right| > \varepsilon\right) \\
&\quad \leq \sum_{i=1}^{N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} g_{it} - \mathbb{E}g_{it}\right| > \frac{\varepsilon}{2}\right) + \sum_{i=1}^{N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} \bar{g}_{it} - \mathbb{E}\bar{g}_{it}\right| > \frac{\varepsilon}{4}\right) \\
&\quad \leq N_\delta \max_{1\leq i\leq N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z'_{it}\right| > \frac{\varepsilon}{2}\right) + N_\delta \max_{1\leq i\leq N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z''_{it}\right| > \frac{\varepsilon}{4}\right),
\end{aligned}$$

where $Z'_{it} = g_{it} - \mathbb{E}g_{it}$ and $Z''_{it} = \bar{g}_{it} - \mathbb{E}\bar{g}_{it}$. We have that

$$N_\delta \max_{1\leq i\leq N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z'_{it}\right| > \frac{\varepsilon}{2}\right) \overset{(a)}{\leq} \left(1 + \frac{16K_1 p^{1/2}}{\underline{\lambda}^{1/2}\varepsilon}\right)^p \max_{1\leq i\leq N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z'_{it}\right| > \frac{\varepsilon}{2}\right),$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

where (a) follows from the fact that the δ-covering number N_δ of an euclidean sphere of radius C in $\mathbb{R}^p$ satisfies $N_\delta \leq (1+(2C)/\delta)^p$ (Vershynin, 2018, Cor. 4.2.13), and that the covering number of V is smaller than the covering number of $\{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}\|_2 \leq \underline{\lambda}^{-1/2}\}$ since $V \subset \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}\|_2 \leq \underline{\lambda}^{-1/2}\}$. Note that Z'_{it} inherits the mixing properties of $(Y_t, \boldsymbol{X}'_t)'$ and satisfies $\|Z'_{it}\|_{L_\infty} \leq 1$. It follows from Proposition B.3 that for all T sufficiently large and for the choice of ε'_T spelled out in that proposition that $\varepsilon'_T \leq 4K_\Sigma^{-1/2}K_1 p^{1/2}/\bar{\lambda}^{1/2} \wedge \kappa_2/2$ and

$$\left(1+\frac{16K_1p^{1/2}}{\underline{\lambda}^{1/2}\varepsilon'_T}\right)^p \max_{1\leq i\leq N_\delta} \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T Z'_{it}\right| > \frac{\varepsilon'_T}{2}\right) \leq \frac{4}{T} + o\left(\frac{1}{T}\right). \tag{A.4}$$

Using analogous arguments, we have that for all T sufficiently large and for the choice of ε''_T spelled out in Proposition B.3 that $\varepsilon''_T \leq 4K_\Sigma^{-1/2}K_1 p^{1/2}/\bar{\lambda}^{1/2} \wedge \kappa_2/2$ and

$$\begin{aligned} &N_\delta \max_{1\leq i\leq N_\delta} \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T Z''_{it}\right| > \frac{\varepsilon''_T}{4}\right) \\ &\leq \left(1+\frac{16K_1p^{1/2}}{\underline{\lambda}^{1/2}\varepsilon''_T}\right)^p \max_{1\leq i\leq N_\delta} \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T Z''_{it}\right| > \frac{\varepsilon''_T}{4}\right) \leq \frac{4}{T} + o\left(\frac{1}{T}\right). \end{aligned} \tag{A.5}$$

The inequalities in (A.4) and (A.5) imply that for all T sufficiently large we can pick $\varepsilon_T = \varepsilon'_T \wedge \varepsilon''_T$ to obtain

$$\sup_{\boldsymbol{v}\in V}\left|\frac{1}{T}\sum_{t=1}^T g_{\boldsymbol{v}t} - \mathbb{E}g_{\boldsymbol{v}t}\right| \leq \frac{\kappa_2}{2}$$

with probability at least $1-8T^{-1}-o(T^{-1})$. The claim of the proposition follows after noting that with probability at least $1-8T^{-1}-o(T^{-1})$, we have

$$\begin{aligned} &\frac{\kappa_1^2}{K_\Sigma}\frac{1}{T}\sum_{t=1}^T \mathbb{1}_{\{|\boldsymbol{X}'_t\boldsymbol{v}|\geq \kappa_1 K_\Sigma^{-1/2}\}} \frac{1}{T}\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}^2 \\ &\geq \frac{\kappa_1^2}{K_\Sigma}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{P}(|\boldsymbol{X}'_t\boldsymbol{v}|\geq \kappa_1 K_\Sigma^{-1/2}) - \sup_{\boldsymbol{v}\in V}\left|\frac{1}{T}\sum_{t=1}^T g_{\boldsymbol{v}t} - \mathbb{E}g_{\boldsymbol{v}t}\right|\right)\frac{1}{T}\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}^2 \\ &\overset{(a)}{\geq} \frac{\kappa_1^2}{K_\Sigma}\left(\kappa_2 - \frac{\kappa_2}{2}\right)\frac{1}{T}\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}^2 \geq \frac{\kappa_1^2\kappa_2}{2K_\Sigma}\frac{1}{T}\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}^2, \end{aligned}$$

where (a) follows from the fact that $\mathbb{P}(|\boldsymbol{X}'_t\boldsymbol{v}| \geq \kappa_1 K_\Sigma^{-1/2}) \geq \mathbb{P}(|\boldsymbol{X}'_t\boldsymbol{v}| \geq \kappa_1 \|\boldsymbol{X}'_t\boldsymbol{v}\|_{L_2}) \geq \kappa_2$ since $\|\boldsymbol{X}'_t\boldsymbol{v}\|_{L_2} = \|\boldsymbol{X}'_t(\boldsymbol{\theta}^* - \boldsymbol{\theta})\|_{L_2}/\sqrt{\frac{1}{T}\sum_{t=1}^T \|\boldsymbol{X}'_t(\boldsymbol{\theta}^* - \boldsymbol{\theta})\|_{L_2}^2} \geq K_\Sigma^{-1/2}$. □

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Proof of Proposition 2. Define $\nu_t = \mathbb{E}\left[(Y_t - f_t^*)\boldsymbol{X}_t\right]$ and note that Lemma 1 implies

$$\sum_{t=1}^{T} \nu_t'(\boldsymbol{\theta}^* - \boldsymbol{\theta}) = \mathbb{E}\left(\sum_{t=1}^{T}(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}t})\right) = 0 \text{ for any } \boldsymbol{\theta} \in \mathbb{R}^p.$$

For any $\boldsymbol{\theta} \in \mathbb{R}^p \setminus \{\boldsymbol{\theta}^*\}$, we have that (notice that A.4 implies that $\boldsymbol{\theta}^*$ is unique)

$$\mathbb{P}\left(\frac{\left|\sum_{t=1}^{T}(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}t})\right|}{\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}} > \varepsilon\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{R}^p\setminus\{\boldsymbol{\theta}^*\}} \frac{\left|\sum_{t=1}^{T}(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}t})\right|}{\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}} > \varepsilon\right).$$

Define $\boldsymbol{v} = (\boldsymbol{\theta}^* - \boldsymbol{\theta})/(\frac{1}{T}\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2})$ for any $\boldsymbol{\theta} \in \mathbb{R}^p \setminus \{\boldsymbol{\theta}^*\}$ and note that

$$\|\boldsymbol{v}\|_2 = \frac{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}{\frac{1}{T}\sum_{t=1}^{T}\sqrt{(\boldsymbol{\theta}^* - \boldsymbol{\theta})'\mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t')(\boldsymbol{\theta}^* - \boldsymbol{\theta})}} \leq \underline{\lambda}^{-1/2}.$$

Then we have that

$$\sum_{t=1}^{T}\frac{\left|(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}t})\right|}{\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}} = \sum_{t=1}^{T}\frac{\left|(Y_t - f_t^*)\boldsymbol{X}_t'(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right|}{\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}} = \left|\frac{1}{T}\sum_{t=1}^{T}(Y_t - f_t^*)\boldsymbol{X}_t'\boldsymbol{v}\right|$$
$$= \left|\frac{1}{T}\sum_{t=1}^{T}[(Y_t - f_t^*)\boldsymbol{X}_t' - \nu_t']\boldsymbol{v}\right| = \left|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{U}_t'\boldsymbol{v}\right|,$$

where $\boldsymbol{U}_t = (Y_t - f_t^*)\boldsymbol{X}_t - \nu_t$. Next, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{R}^p\setminus\{\boldsymbol{\theta}^*\}}\frac{\left|\sum_{t=1}^{T}(Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta}t})\right|}{\sum_{t=1}^{T}\|f_t^* - f_{\boldsymbol{\theta}t}\|_{L_2}} > \varepsilon\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\leq\underline{\lambda}^{-1/2}}\left|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{U}_t'\boldsymbol{v}\right| > \varepsilon\right)$$
$$\leq \mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\leq\underline{\lambda}^{-1/2}}\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{U}_t\right\|_2\|\boldsymbol{v}\|_2 > \varepsilon\right) \leq \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{U}_t\right\|_2 > \underline{\lambda}^{1/2}\varepsilon\right).$$

Note that $\{\boldsymbol{U}_t\}$ is mean zero, satisfies $\|U_{it}\|_{L_2} \leq K_m$ and

$$\|U_{it}\|_{L_{r_m}} \leq \|(Y_t - f_t^*)X_{it}\|_{L_{r_m}} + \|\nu_{it}\|_{L_{r_m}} = \|(Y_t - f_t^*)X_{it}\|_{L_{r_m}} + \|(Y_t - f_t^*)X_{it}\|_{L_1} \leq 2K_m$$

because of A.1, and inherits the mixing properties of $\{(Y_t, \boldsymbol{X}_t)'\}$ spelled out in A.2. Proposition B.4 then implies that, for all T sufficiently large, we have

$$\sup_{\boldsymbol{\theta}\in\mathbb{R}^p\setminus\{\boldsymbol{\theta}^*\}}\frac{\left|\sum_{t=1}^{T}(Y_t - f_t^*)(f_{\boldsymbol{\theta}t} - f_t^*)\right|}{\sum_{t=1}^{T}\|f_{\boldsymbol{\theta}t} - f_t^*\|_{L_2}} \leq 12\sqrt{\frac{K_{\sigma^2}}{\underline{\lambda}}}\sqrt{\frac{p\log(T)}{T}}$$

with probability at least $1 - 3K_p(2K_m)^{r_m}/(K_{\sigma^2}^{\frac{1}{2}}\log(T)) - o(\log(T)^{-1})$, where K_{σ^2} is the constant σ^2 defined in that proposition. The claim of the proposition then follows. □

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Proof of Theorem 2. We begin by showing that when A.1 is satisfied, we have $\lambda_{\max}(\Sigma) \leq K_m^2 p$. Let $\Sigma = \mathbb{E}(\boldsymbol{X}_t \boldsymbol{X}_t')$, and let $\Sigma_{i\bullet}$ denote the ith row of Σ. Then,

$$\lambda_{\max}(\Sigma) = \sup_{\mathbf{x}\in\mathbb{R}^p:\|\mathbf{x}\|_2=1} \|\Sigma\mathbf{x}\|_2 = \sup_{\mathbf{x}\in\mathbb{R}^p:\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^p (\Sigma_{i\bullet}\mathbf{x})^2} \leq \sup_{\mathbf{x}\in\mathbb{R}^p:\|\mathbf{x}\|_2=1} \sqrt{\sum_{i=1}^p \|\Sigma_{i\bullet}\|_2^2 \|\mathbf{x}\|_2^2}$$
$$= \sqrt{\sum_{i=1}^p \|\Sigma_{i\bullet}\|_2^2} \leq \sqrt{\sum_{i=1}^p \|K_m^2 \mathbf{1}_p\|_2^2} = K_m^2 \sqrt{\sum_{i=1}^p p} = K_m^2 p,$$

where $\mathbf{1}_p$ is p-dimensional vector with entries equal to one.

The proof is similar to the one of Theorem 1 and we only highlight the main differences. In the proof of Proposition 1, define $\boldsymbol{Z}_t = \Sigma^{-\frac{1}{2}} \boldsymbol{X}_t$ and $\boldsymbol{v} = \Sigma^{\frac{1}{2}}(\boldsymbol{\theta}_* - \boldsymbol{\theta})/\|f_t^* - f_{\boldsymbol{\theta} t}\|_{L_2}$. Then A.6 and A.7 imply $\mathbb{P}(|\boldsymbol{Z}_t'\boldsymbol{v}| \geq \kappa_1 \|\boldsymbol{Z}_t'\boldsymbol{v}\|_{L_2}) \geq \kappa_2$. Thus, in that proposition the function $g_{\boldsymbol{v}t}$ can be defined as $\mathbb{1}_{\{|\boldsymbol{Z}_t'\boldsymbol{v}| \geq \kappa_1 \|\boldsymbol{Z}_t'\boldsymbol{v}\|_{L_2}\}}$. Proposition B.1 can then be modified and it is straightforward to see that there exists a $\bar{g}_{it}$ function such that for all $\delta < 1/2$ we have $\mathbb{E}\bar{g}_{it} \leq K_1 p^{\frac{1}{2}} \delta$ for some positive constant K_1. If we set $\delta = \varepsilon/(8K_1 p^{\frac{1}{2}})$ for all $\varepsilon < 4K_1 p^{\frac{1}{2}}$ we get, following the same steps as in Proposition 1 and noting that $\|\boldsymbol{v}\|_2 = 1$, that

$$\mathbb{P}\left(\sup_{\boldsymbol{v}\in V} \left|\frac{1}{T}\sum_{t=1}^T g_{\boldsymbol{v}t} - \mathbb{E}g_{\boldsymbol{v}t}\right| > \varepsilon\right)$$
$$\leq \left(1 + \frac{16K_1 p}{\varepsilon}\right)^p \max_{i=1,\ldots,N_\delta} \left[\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T Z_{it}'\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T Z_{it}''\right| > \frac{\varepsilon}{4}\right)\right].$$

Finally, Proposition B.3 implies that for all T sufficiently large and any $\boldsymbol{\theta} \in \mathbb{R}^p$,

$$\frac{1}{T}\sum_{t=1}^T (f_t^* - f_{\boldsymbol{\theta} t})^2 \geq \frac{\kappa_1^2 \kappa_2}{2} \frac{1}{T}\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta} t}\|_{L_2}^2$$

holds with probability at least $1 - 8T^{-1} - o(T^{-1})$. In the proof of Proposition 2, define $\boldsymbol{v} = \Sigma^{\frac{1}{2}}(\boldsymbol{\theta}^* - \boldsymbol{\theta})/\|f_t^* - f_{\boldsymbol{\theta} t}\|_{L_2}$ and $\boldsymbol{U}_t = (Y_t - f_t^*)\boldsymbol{Z}_t$. Following the steps of the proof of Proposition 2, we have that for any $\boldsymbol{\theta} \in \mathbb{R}^p \setminus \{\boldsymbol{\theta}^*\}$

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{R}^p\setminus\{\boldsymbol{\theta}^*\}} \frac{\left|\sum_{t=1}^T (Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta} t})\right|}{\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta} t}\|_{L_2}} > \varepsilon\right) \leq \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{U}_t\right\|_2 > \varepsilon\right),$$

where we have used the fact that $\|\boldsymbol{v}\|_2 = 1$. Note that $\{\boldsymbol{U}_t\}$ is mean zero, satisfies $\|U_{it}\|_{L_m} \leq K_m$ for each $i = 1, \ldots, p$ because of A.1* and inherits the mixing properties of $\{(Y_t, \boldsymbol{X}_t)'\}$ spelled out in A.2. Applying Proposition B.4, we have that for all T sufficiently large and any $\boldsymbol{\theta} \in \mathbb{R}^p \setminus \{\boldsymbol{\theta}^*\}$,

$$\frac{1}{T}\sum_{t=1}^T (Y_t - f_t^*)(f_t^* - f_{\boldsymbol{\theta} t}) \leq 12\sqrt{K_{\sigma^2}'} \frac{1}{T}\sum_{t=1}^T \|f_t^* - f_{\boldsymbol{\theta} t}\|_{L_2} \sqrt{\frac{p\log(T)}{T}}$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

holds with probability at least $1 - 3K_p K_m^{r_m}/((K'_{\sigma^2})^{\frac{1}{2}} \log(T)) - o(\log(T)^{-1})$ with $K'_{\sigma^2} = K_m^2 \left(1 + 32 \frac{r_m}{r_m - 2} \sum_{l=1}^{\infty} \alpha(l)^{1 - \frac{2}{r_m}}\right)$. Finally, in the proof of Theorem 1, we can replace condition (A.2) with

$$\|f_t^* - f_{\boldsymbol{\theta} t}\|_{L_2} > \frac{48(K'_{\sigma^2})^{\frac{1}{2}}}{\kappa_1^2 \kappa_2} \sqrt{\frac{p \log(T)}{T}}.$$

Following the same steps as in the proof there, we obtain the claim. □

Proof of Theorem 3. We begin by introducing the out-of-sample risk for the "ghost" out-of-sample observations. Let $\{(Y_t^G, (\boldsymbol{X}_t^G)')'\}_{t=T+1}^{T+H}$ denote a sequence of observations from the $\{(Y_t, \boldsymbol{X}_t')'\}$ process that is independent of $\{(Y_t, \boldsymbol{X}_t')'\}_{t=1}^{T}$. Then define

$$R_{\mathsf{oos}}^G(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{1}{H} \sum_{t=T+1}^{T+H} (Y_t^G - f_{\boldsymbol{\theta} t}^G)^2\right]$$

$$R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[\frac{1}{H} \sum_{t=T+1}^{T+H} (Y_t^G - \hat{f}_t^G)^2 \,\middle|\, (Y_T, \boldsymbol{X}_T')', \ldots, (Y_1, \boldsymbol{X}_1')'\right],$$

where $f_{\boldsymbol{\theta} t}^G = \boldsymbol{\theta}' \boldsymbol{X}_t^G$ and $\hat{f}_t^G = \hat{\boldsymbol{\theta}}' \boldsymbol{X}_t^G$ with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\{(Y_1, \boldsymbol{X}_1')', \ldots, (Y_T, \boldsymbol{X}_T')'\})$. Notice that clearly $R_{\mathsf{oos}}^G(\boldsymbol{\theta}) = R_{\mathsf{oos}}(\boldsymbol{\theta})$. We may then note that

$$\begin{aligned} R_{\mathsf{oos}}(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}(\boldsymbol{\theta}^*) &\leq |R_{\mathsf{oos}}(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}})| + |R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}(\boldsymbol{\theta}^*)| \\ &= |R_{\mathsf{oos}}(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}})| + |R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\boldsymbol{\theta}^*)|. \end{aligned}$$

The claim of the theorem follows from the fact that if for some $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, $\delta_1 \in (0,1)$ and $\delta_2 \in (0,1)$ we have that

$$\mathbb{P}(|R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\boldsymbol{\theta}^*)| \geq \varepsilon_1) \leq \delta_1 \tag{A.6}$$

$$\mathbb{P}\left(|R_{\mathsf{oos}}(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}})| \geq \varepsilon_2 \,\middle|\, |R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\boldsymbol{\theta}^*)| \leq \varepsilon_1\right) \leq \delta_2, \tag{A.7}$$

then it follows from the union bound and the total probability theorem that $R_{\mathsf{oos}}(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}(\boldsymbol{\theta}^*) \leq \varepsilon_1 + \varepsilon_2$ with probability at least $1 - 2\delta_1 - \delta_2$. Theorem 2 implies that for all T sufficiently large (A.6) holds for the choice of ε_1 and δ_1 implied by the theorem. Thus, this proof focuses on establishing that (A.7) holds. Denote by $\mathcal{E} = \{R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\boldsymbol{\theta}^*) \leq 1\}$ and note that conditional on $\mathcal{E}$ we have that $1 \geq R_{\mathsf{oos}}^G(\hat{\boldsymbol{\theta}}) - R_{\mathsf{oos}}^G(\boldsymbol{\theta}^*) = \|f_{\boldsymbol{\theta}^* t}^G - f_{\hat{\boldsymbol{\theta}} t}^G\|_{L_2}^2 > \underline{\lambda} \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2^2$. Let $\mathbb{E}_T(\cdot) = \mathbb{E}(\cdot|\mathcal{I}_T)$ be the expectation conditional on information up to time T, with $\mathcal{I}_T$ the information set at time T. This implies that for $r = (1/\underline{\lambda})^{\frac{1}{2}}$ we have

$$\begin{aligned} &|\mathbb{E}_T(Y_{T+h} - f_{\hat{\boldsymbol{\theta}} T+h})^2 - \mathbb{E}_T(Y_{T+h}^G - f_{\hat{\boldsymbol{\theta}} T+h}^G)^2| \\ &\quad \leq \sup_{\boldsymbol{\theta} \in B_2(\boldsymbol{\theta}^*, r)} |\mathbb{E}_T(Y_{T+h} - f_{\boldsymbol{\theta} T+h})^2 - \mathbb{E}_T(Y_{T+h}^G - f_{\boldsymbol{\theta} T+h}^G)^2| \\ &\quad = \sup_{\boldsymbol{\theta} \in B_2(\boldsymbol{\theta}^*, r)} |\mathbb{E}_T(Y_{T+h} - f_{\boldsymbol{\theta} T+h})^2 - \mathbb{E}(Y_{T+h} - f_{\boldsymbol{\theta} T+h})^2| \end{aligned}$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

$$
\begin{aligned}
&\le |\mathbb{E}_T(Y_{T+h} - f_{\boldsymbol{\theta}^* T+h})^2 - \mathbb{E}(Y_{T+h} - f_{\boldsymbol{\theta}^* T+h})^2| \\
&+ \sup_{\boldsymbol{v} \in B_2(\mathbf{0}, r)} \boldsymbol{v}' [\mathbb{E}_T(\boldsymbol{X}_{T+h} \boldsymbol{X}'_{T+h}) - \mathbb{E}(\boldsymbol{X}_{T+h} \boldsymbol{X}'_{T+h})] \boldsymbol{v} \\
&+ 2 \sup_{\boldsymbol{v} \in B_2(\mathbf{0}, r)} |[\mathbb{E}_T((Y_{T+h} - f_{\boldsymbol{\theta}^* T+h}) \boldsymbol{X}_{T+h}) - \mathbb{E}((Y_{T+h} - f_{\boldsymbol{\theta}^* T+h}) \boldsymbol{X}_{T+h})]' \boldsymbol{v}|.
\end{aligned}
$$

It follows from Ibragimov's inequality that

$$
\begin{aligned}
&\|\mathbb{E}_T(Y_{T+h} - f_{\boldsymbol{\theta}^* T+h})^2 - \mathbb{E}(Y_{T+h} - f_{\boldsymbol{\theta}^* T+h})^2\|_{L_1} \\
&\quad \le 6\alpha(h)^{\frac{1}{2}} \|(Y_{T+h} - f_{\boldsymbol{\theta}^* T+h})^2\|_{L_2} \le 6\alpha(h)^{\frac{1}{2}} K_m^2, \\
&\| \sup_{\boldsymbol{v} \in B_2(\mathbf{0}, r)} \boldsymbol{v}' [\mathbb{E}_T(\boldsymbol{X}_{T+h} \boldsymbol{X}'_{T+h}) - \mathbb{E}(\boldsymbol{X}_{T+h} \boldsymbol{X}'_{T+h})] \boldsymbol{v} \|_{L_1} \\
&\quad \le \| \max_{ij} |\left[\mathbb{E}_T(\boldsymbol{X}_{T+h} \boldsymbol{X}'_{T+h}) - \mathbb{E}(\boldsymbol{X}_{T+h} \boldsymbol{X}'_{T+h})\right]_{ij}| \|_{L_1} \sup_{\boldsymbol{v} \in B_2(\mathbf{0}, r)} \|\boldsymbol{v}\|_1^2 \le 6\alpha(h)^{\frac{1}{2}} \frac{K_m^2}{\underline{\lambda}} p, \\
&\|2 \sup_{\boldsymbol{v} \in B_2(\mathbf{0}, r)} |[\mathbb{E}_T((Y_{T+h} - f_{\boldsymbol{\theta}^* T+h}) \boldsymbol{X}_{T+h}) - \mathbb{E}((Y_{T+h} - f_{\boldsymbol{\theta}^* T+h}) \boldsymbol{X}_{T+h})]' \boldsymbol{v}| \|_{L_1} \\
&\quad \le 12\alpha(h)^{\frac{1}{2}} \|(Y_{T+h} - f_{\boldsymbol{\theta}^* T+h}) X_{i,T+h}\|_{L_2} \sup_{\boldsymbol{v} \in B_2(\mathbf{0}, r)} \|\boldsymbol{v}\|_1 \le 12\alpha(h)^{\frac{1}{2}} \frac{K_m^2}{\sqrt{\underline{\lambda}}} \sqrt{p}.
\end{aligned}
$$

Thus, conditional on $\mathcal{E}$ and for T sufficiently large we have

$$
\begin{aligned}
&\|\mathbb{E}_T(Y_{T+h} - f_{\hat{\boldsymbol{\theta}} T+h})^2 - \mathbb{E}_T(Y^G_{T+h} - f^G_{\hat{\boldsymbol{\theta}} T+h})^2\|_{L_1} \\
&\quad \le 6\alpha(h)^{\frac{1}{2}} K_m^2 \left(1 + \frac{p}{\underline{\lambda}} + 2\sqrt{\frac{p}{\underline{\lambda}}}\right) \le 24\alpha(h)^{\frac{1}{2}} \frac{K_m^2}{\underline{\lambda}} p.
\end{aligned}
$$

The conditional version of Markov's inequality implies that

$$
\begin{aligned}
\mathbb{P}(|R_{\mathsf{oos}}(\hat{\boldsymbol{\theta}}) - R^G_{\mathsf{oos}}(\hat{\boldsymbol{\theta}})| \ge \varepsilon_2 | \mathcal{E}) &\le \frac{1}{\varepsilon_2} \frac{1}{H} \sum_{h=1}^{H} \|\mathbb{E}_T(Y_{T+h} - f_{\hat{\boldsymbol{\theta}} T+h})^2 - \mathbb{E}(Y_{T+h} - f_{\hat{\boldsymbol{\theta}} T+h})^2\|_{L_1} \\
&\le \frac{24}{\varepsilon_2} \frac{K_m^2}{\underline{\lambda}} \sum_{l=1}^{\infty} \alpha(l)^{\frac{1}{2}} \frac{p}{H},
\end{aligned}
$$

which implies the claim of the theorem. □

Proof of Lemma 2. Let $\boldsymbol{v} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ and note that the Paley–Zygmund inequality implies that for any $\vartheta \in [0, 1]$, we have

$$
\mathbb{P}(|\boldsymbol{v}' \boldsymbol{X}_t| > \vartheta^{\frac{1}{2}} \|\boldsymbol{v}' \boldsymbol{X}_t\|_{L_2}) \ge (1 - \vartheta)^2 \frac{\mathbb{E}(|\boldsymbol{v}' \boldsymbol{X}_t|^2)^2}{\mathbb{E}(|\boldsymbol{v}' \boldsymbol{X}_t|^4)}. \tag{A.8}
$$

Note that A.6* implies that $\boldsymbol{X}_t$ is elliptical. Then $\boldsymbol{v}' \boldsymbol{X}_t = \sigma_t U$ holds where $\sigma_t^2 = \boldsymbol{v}' \Sigma_t \boldsymbol{v}$ and U is an elliptical random variable with zero mean and unit variance (whose distribution does not depend on $\boldsymbol{v}$ nor Σ_t). Thus, we have that the probability in (A.8) is lower bounded by $[(1 - \vartheta)^2 \mathbb{E}(|U|^2)^2] / \mathbb{E}(|U|^4)$, which implies the claim of the lemma. □

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

B. Auxiliary Results

PROPOSITION B.1. *Consider the same setup as in Proposition 1. Let* $V_i = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v} - \boldsymbol{v}_i\|_2 \leq \delta\}$ *with* $\boldsymbol{v}_i \in V$ *for* $i = 1, \ldots, N_\delta$ *denote a* δ*-covering of the set* V *for some* $\delta < K_\Sigma^{-1/2}/(2\bar{\lambda}^{1/2})$. *Define the function* $g_{\boldsymbol{v}t} = \mathbb{1}_{\{|\boldsymbol{X}_t'\boldsymbol{v}| \geq \kappa_1 K_\Sigma^{-1/2}\}}$ *and let* $g_{it} = g_{\boldsymbol{v}_i t}$.

Then (*i*) *for all* $\boldsymbol{v} \in V_i$ *we have that* $|g_{\boldsymbol{v}t} - g_{it}| \leq \bar{g}_{it} = \mathbb{1}_{\{\boldsymbol{X}_t \in S_i\}}$, *where* $S_i = \bigcup_{\boldsymbol{v} \in V_i} \{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}'\boldsymbol{v}| = \kappa_1 K_\Sigma^{-1/2}\}$ *and* (*ii*) *there exists a positive constant* K_1 *that depends on* $K_{\boldsymbol{Z}}$, $\underline{\lambda}$, $\bar{\lambda}$ *and* κ_1 *(and it does not depend on t, i or p) such that* $\mathbb{E}\bar{g}_{it} \leq K_1 p^{1/2}\delta$.

Proof. (*i*) We show that $S_i = \bigcup_{\boldsymbol{v} \in V_i} \{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}'\boldsymbol{v}| = \kappa_1 K_\Sigma^{-1/2}\}$ is the set containing all the vectors $\boldsymbol{x}$ such that the indicator functions $\mathbb{1}_{\{|\boldsymbol{x}'\boldsymbol{v}| \geq \kappa_1 K_\Sigma^{-1/2}\}}$ and $\mathbb{1}_{\{|\boldsymbol{x}'\boldsymbol{v}_i| \geq \kappa_1 K_\Sigma^{-1/2}\}}$ are different. We do so by showing that the complement of S_i is a set of vectors $\boldsymbol{x}$ where the indicator functions are equal. We establish this by contradiction. Assume $\boldsymbol{x}$ is not in S_i and that the indicator functions $\mathbb{1}_{\{|\boldsymbol{x}'\boldsymbol{v}| \geq \kappa_1 K_\Sigma^{-1/2}\}}$ and $\mathbb{1}_{\{|\boldsymbol{x}'\boldsymbol{v}_i| \geq \kappa_1 K_\Sigma^{-1/2}\}}$ are different. Since V_i is convex there must be an intermediate $\dot{\boldsymbol{v}} \in V_i$ such that $|\boldsymbol{x}'\dot{\boldsymbol{v}}| = \kappa_1 K_\Sigma^{-1/2}$ implying that $\boldsymbol{x}$ is in S_i, which leads to a contradiction. (*ii*) Note that

$$S_i = \bigcup_{\boldsymbol{v} \in V_i} \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}'\boldsymbol{v} = \kappa_1 K_\Sigma^{-1/2}\} \cup \bigcup_{\boldsymbol{v} \in V_i} \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}'\boldsymbol{v} = -\kappa_1 K_\Sigma^{-1/2}\} = S_{i+} \cup S_{i-}.$$

In what follows, we bound the probability of the event $\{\boldsymbol{X}_t \in S_{i+}\}$ only as the event $\{\boldsymbol{X}_t \in S_{i-}\}$ can be treated analogously. We divide the proof into four steps. **1**. We work with an appropriately rotated version of $\boldsymbol{X}_t$ denote by $\boldsymbol{Z}$. Let ϑ be the angle between the vector $\Sigma_t^{1/2}\boldsymbol{v}_i$ and $(1, 0, \ldots, 0)'$, and let $\mathbf{R} \in \mathbb{R}^{p \times p}$ be the rotation matrix associated with ϑ. Recall that: (*i*)$\mathbf{R}'\mathbf{R} = \mathbf{I}_p$; (*ii*)$\mathbf{R}\Sigma_t^{1/2}\boldsymbol{v}_i = \|\Sigma_t^{1/2}\boldsymbol{v}_i\|_2(1, 0, \ldots, 0)'$; (*iii*) if we define $W_1 = \{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq 1\}$ and $W_2 = \{\mathbf{w} \in \mathbb{R}^p : \mathbf{w} = \mathbf{R}\mathbf{w}^\star \text{ for some } \mathbf{w}^\star \in W_1\}$, then we have that $W_1 = W_2$. Define $\boldsymbol{Z} = \mathbf{R}\Sigma_t^{-1/2}\boldsymbol{X}_t$ and note that

$$\begin{aligned}
\mathbb{P}(\{\boldsymbol{X}_t \in S_{i+}\}) &= \mathbb{P}\left(\left\{\boldsymbol{X}_t \in \bigcup_{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq 1} \{\mathbf{x} \in \mathbb{R}^p : \boldsymbol{v}_i'\mathbf{x} + \delta\mathbf{w}'\mathbf{x} = \kappa_1 K_\Sigma^{-1/2}\}\right\}\right) \\
&= \mathbb{P}\left(\left\{\mathbf{Z} \in \bigcup_{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq 1} \{\mathbf{z} \in \mathbb{R}^p : \boldsymbol{v}_i'\Sigma_t^{1/2}\mathbf{R}'\mathbf{z} + \delta\mathbf{w}'\Sigma_t^{1/2}\mathbf{R}'\mathbf{z} = \kappa_1 K_\Sigma^{-1/2}\}\right\}\right).
\end{aligned}$$

Define $c_{it} = \|\Sigma_t^{1/2}\boldsymbol{v}_i\|_2$ and note that the set in the last equation is such that

$$\begin{aligned}
&\bigcup_{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq 1} \{\mathbf{z} \in \mathbb{R}^p : \|\Sigma_t^{1/2}\boldsymbol{v}_i\|_2 z_1 + \delta(\mathbf{R}\Sigma_t^{1/2}\mathbf{w})'\mathbf{z} = \kappa_1 K_\Sigma^{-1/2}\} \\
&\quad \subset \bigcup_{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq 1} \{\mathbf{z} \in \mathbb{R}^p : c_{it} z_1 + \bar{\lambda}^{1/2}\delta\mathbf{w}'\mathbf{z} = \kappa_1 K_\Sigma^{-1/2}\} \\
&\quad = \bigcup_{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq 1} \{\mathbf{z} \in \mathbb{R}^p : (c_{it} + \bar{\lambda}^{1/2}\delta w_1) z_1 + \bar{\lambda}^{1/2}\delta\mathbf{w}_{-1}'\mathbf{z}_{-1} = \kappa_1 K_\Sigma^{-1/2}\} = S_{it+}'.
\end{aligned}$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Lastly, we note that for any $i = 1, \ldots, N_\delta$ and $t = 1, \ldots, T$, we have

$$c_{it} = \frac{\|\Sigma_t^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_i)\|_2}{\sqrt{\frac{1}{T}\sum_{t=1}^T \|X_t'(\boldsymbol{\theta}^* - \boldsymbol{\theta}_i)\|_{L_2}^2}} = \sqrt{\frac{(\boldsymbol{\theta}^* - \boldsymbol{\theta}_i)'\Sigma_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_i)}{\frac{1}{T}\sum_{t=1}^T (\boldsymbol{\theta}^* - \boldsymbol{\theta}_i)'\Sigma_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_i)}} > K_\Sigma^{-1/2}$$

and $c_{it} - \bar{\lambda}^{1/2}\delta > K_\Sigma^{-1/2}/2 > 0$. **2**. We construct two sets $S_{it+}'^1$ and $S_{it+}'^2$ such that $S_{it+}' \subset S_{it+}'^1 \cap S_{it+}'^2$. Define

$$S_{it+}'^1 = \left\{ \mathbf{z} \in \mathbb{R}^p : z_1 \leq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} - \bar{\lambda}^{1/2}\delta} + \frac{\bar{\lambda}^{1/2}\delta}{c_{it} - \bar{\lambda}^{1/2}\delta}\sqrt{z_2^2 + \cdots + z_p^2} \right\}, \tag{B.1}$$

that is the set of points "underneath" a hyper-cone. Let $\mathbf{z}$ be in S_{it+}', define $\dot{\mathbf{z}} = \|\mathbf{z}_{-1}\|_2^{-1}(z_2, \ldots, z_p)'$ and note that $\|\dot{\mathbf{z}}\|_2 = 1$. Then for some $\mathbf{w}$ such that $\|\mathbf{w}\|_2 \leq 1$ we have that

$$\begin{aligned} z_1 &= \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1} - \frac{\bar{\lambda}^{1/2}\delta \mathbf{w}_{-1}'\mathbf{z}_{-1}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1} = \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1} - \frac{\bar{\lambda}^{1/2}\delta \mathbf{w}_{-1}'\dot{\mathbf{z}}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1}\|\mathbf{z}_{-1}\|_2 \\ &\leq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} - \bar{\lambda}^{1/2}\delta} + \frac{\bar{\lambda}^{1/2}\delta \|\mathbf{w}_{-1}\|_2\|\dot{\mathbf{z}}\|_2}{K_\Sigma^{-1/2} - \bar{\lambda}^{1/2}\delta}\|\mathbf{z}_{-1}\|_2 \leq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} - \bar{\lambda}^{1/2}\delta} + \frac{\bar{\lambda}^{1/2}\delta}{c_{it} - \bar{\lambda}^{1/2}\delta}\sqrt{z_2^2 + \cdots + z_p^2}, \end{aligned}$$

which implies that $\mathbf{z}$ is also in $S_{i+}'^1$. Define

$$S_{it+}'^2 = \left\{ \mathbf{z} \in \mathbb{R}^p : z_1 \geq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta} - \frac{\bar{\lambda}^{1/2}\delta}{c_{it} - \bar{\lambda}^{1/2}\delta}\sqrt{z_2^2 + \cdots + z_p^2} \right\}, \tag{B.2}$$

that is the set of points "above" a hyper-cone. Let $\mathbf{z}$ in S_{it+}' and define $\dot{\mathbf{z}}$ as above. Then for some $\mathbf{w}$ such that $\|\mathbf{w}\|_2 \leq 1$ we have that

$$\begin{aligned} z_1 &= \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1} - \frac{\bar{\lambda}^{1/2}\delta \mathbf{w}_{-1}'\mathbf{z}_{-1}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1} = \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1} - \frac{\bar{\lambda}^{1/2}\delta \mathbf{w}_{-1}'\dot{\mathbf{z}}}{c_{it} + \bar{\lambda}^{1/2}\delta w_1}\|\mathbf{z}_{-1}\|_2 \\ &\geq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta} - \frac{\bar{\lambda}^{1/2}\delta \|\mathbf{w}_{-1}\|_2\|\dot{\mathbf{z}}\|_2}{K_\Sigma^{-1/2} - \bar{\lambda}^{1/2}\delta}\|\mathbf{z}_{-1}\|_2 \geq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta} - \frac{\bar{\lambda}^{1/2}\delta}{c_{it} - \bar{\lambda}^{1/2}\delta}\sqrt{z_2^2 + \cdots + z_p^2}, \end{aligned}$$

which implies that $\mathbf{z}$ is also in $S_i^{+,2}$.

3. We establish an upper bound on the probability of the event $\{\mathbf{Z} \in S_{it+}'\}$. Note that $S_{it+}' \subset S_{i+}'^1 \cap S_{i+}'^2 = A_i \cup B_i \cup C_i$ where

$$\begin{aligned} A_{it} &= S_{it+}'^1 \cap \left\{ \mathbf{z} \in \mathbb{R}^p : z_1 \geq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} - \bar{\lambda}^{-1/2}\delta} \right\} \\ B_{it} &= \left\{ \mathbf{z} \in \mathbb{R}^p : \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta} \leq z_1 \leq \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} - \bar{\lambda}^{1/2}\delta} \right\} \\ C_{it} &= S_{it+}'^2 \cap \left\{ \mathbf{z} \in \mathbb{R}^p : z_1 < \frac{\kappa_1 K_\Sigma^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta} \right\}. \end{aligned}$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Then we have that $\mathbb{P}(\boldsymbol{X}_t \in S_{i+}) < \mathbb{P}(\mathbf{Z} \in A_{it}) + \mathbb{P}(\mathbf{Z} \in B_{it}) + \mathbb{P}(\mathbf{Z} \in C_{it})$. Using Proposition B.2 and A.5, we have that

$$\mathbb{P}(\mathbf{Z} \in A_{it}) \leq K_{\mathbf{Z}} K_{\Sigma}^{1/2} \bar{\lambda}^{1/2} \sqrt{\frac{\pi}{2}} p^{1/2} \delta$$

$$\mathbb{P}(\mathbf{Z} \in B_{it}) = \mathbb{P}\left(\frac{\kappa_1 K_{\Sigma}^{-1/2}}{c_{it} + \bar{\lambda}^{1/2}\delta} \leq Z_1 \leq \frac{\kappa_1 K_{\Sigma}^{-1/2}}{c_{it} - \bar{\lambda}^{1/2}\delta} \right)$$

$$\leq K_{\mathbf{Z}} \sup_s f_{S_1}(s) \kappa_1 K_{\Sigma}^{-1/2} \frac{2\bar{\lambda}^{1/2}\delta}{(c_{it} - \bar{\lambda}^{1/2}\delta)(c_{it} + \bar{\lambda}^{1/2}\delta)} \leq 8 K_{\mathbf{Z}} \kappa_1 K_{\Sigma}^{1/2} \sup_s f_{S_1}(s) \bar{\lambda}^{1/2} \delta$$

$$\mathbb{P}(\mathbf{Z} \in C_{it}) \leq K_{\mathbf{Z}} K_{\Sigma}^{1/2} \bar{\lambda}^{1/2} \sqrt{\frac{\pi}{2}} p^{1/2} \delta.$$

4. It follows from the inequalities above, and by using analogous steps to bound the probability of the event $\mathbb{P}(\boldsymbol{X}_t \in S_{i-})$, that there exists a positive constant K_1 that depends on $K_{\mathbf{Z}}, \boldsymbol{S}, \underline{\lambda}, \bar{\lambda}$, and κ_1, but does not depend on i and t or p, such that $\mathbb{P}(\boldsymbol{X}_t \in S_i) \leq K_1 p^{1/2} \delta$. □

PROPOSITION B.2. *Let $\mathbf{Z}$ be a p-dimensional random vector. Suppose $\mathbb{P}(\mathbf{Z} \in E) \leq K_{\mathbf{Z}} \mathbb{P}(\boldsymbol{S} \in E)$ holds for some p-dimensional spherical random vector $\boldsymbol{S}$ whose density is assumed to exist, some positive constant $K_{\mathbf{Z}}$ and any $E \in \mathcal{B}(\mathbb{R}^p)$. Define the set $S = \{\mathbf{z} \in \mathbb{R}^p : a \leq z_1 \leq a + b\sqrt{z_2^2 + \cdots + z_p^2}\}$ for some $a, b > 0$.*

Then, there is a positive constant C such that $\mathbb{P}(\mathbf{Z} \in S) \leq C p^{\frac{1}{2}} b$.

Proof. For convenience, we show this result for $p > 2$ and for $a = 0$. We have

$$\mathbb{P}(\mathbf{Z} \in S) = \mathbb{P}\left(0 \leq Z_1 \leq b\sqrt{Z_2^2 + \ldots + Z_p^2}\right) = \mathbb{P}\left(0 \leq Z_1^2 \leq b^2 (Z_2^2 + \cdots + Z_p^2)\right)$$

$$= \mathbb{P}\left(0 \leq \frac{Z_1^2}{\|\mathbf{Z}\|_2^2} \leq b^2 \frac{\|\mathbf{Z}\|_2^2 - Z_1^2}{\|\mathbf{Z}\|_2^2}\right) = \mathbb{P}\left(0 \leq \frac{Z_1}{\|\mathbf{Z}\|_2} \leq \frac{b}{\sqrt{1+b^2}}\right).$$

Consider the p-spherical transformation of $\mathbf{Z}$ (Fang and Zhang, 1990, Exam. 1.6.8)

$$(Z_1, \ldots, Z_i, \ldots, Z_p)' = r \left(\cos\theta_1, \ldots, \prod_{k=1}^{i-1} \sin\theta_k \cos\theta_i, \ldots, \prod_{k=1}^{p-2} \sin\theta_k \sin\theta_{p-1} \right)',$$

where $r \in [0, \infty)$, $\theta_i \in [0, \pi]$ for $1 \leq i \leq p-2$ and $\theta_{p-1} \in [0, 2\pi]$. We remark that r denotes $\|\mathbf{Z}\|_2$ and that the angles $\theta_1, \ldots, \theta_{p-1}$ are set according to the following scheme: θ_1 is the angle between the z_1 axis and the vector $\mathbf{Z}$; θ_2 is the angle between the projection of the $\mathbf{Z}$ vector on the span generated by $z_2, \ldots, z_p$, which we denote by $\mathbf{Z}^{(1)}$, and the z_2 axis; θ_3 is the angle between the projection of $\mathbf{Z}^{(1)}$ on the span generated by $z_3, \ldots, z_p$, which we denote by $\mathbf{Z}^{(2)}$, and the z_3 axis; …; θ_{p-1} is the angle between the projection of $\mathbf{Z}^{(p-2)}$ on the span generated by Z_{p-1}, Z_p and the z_{p-1} axis. If we let ϑ denote the angle such that $\cos(\vartheta) = b/\sqrt{1+b^2}$, then we have

$$\mathbb{P}\left(0 \leq \frac{Z_1}{\|\mathbf{Z}\|_2} \leq \frac{b}{\sqrt{1+b^2}}\right) = \mathbb{P}\left(0 \leq \cos\theta_1 \leq \frac{b}{\sqrt{1+b^2}}\right) = \mathbb{P}\left(\vartheta \leq \theta_1 \leq \frac{\pi}{2}\right). \quad \textbf{(B.3)}$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

We use $K_{\mathbf{Z}}$ and the distribution of $\boldsymbol{S}$ to bound the probability in (B.3). Fang et al. (1990, Thm. 2.11) establish that the density of θ_1 implied by $\boldsymbol{S}$ is given $f_{\theta_1}(t) = \frac{\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p-1}{2}\right)} \sin^{p-2} t$. Then we have that (B.3) is upper bounded by

$$K_{\mathbf{Z}} \int_{\vartheta}^{\pi/2} f_{\theta_1}(t)dt \overset{(a)}{\leq} \frac{K_{\mathbf{Z}}}{\sqrt{\pi}} \frac{\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p-1}{2}\right)} \int_{\vartheta}^{\pi/2} 1dt \overset{(b)}{\leq} \frac{K_{\mathbf{Z}}}{\sqrt{2\pi}} p^{1/2} \left(\frac{\pi}{2} - \vartheta\right) \overset{(c)}{\leq} \frac{K_{\mathbf{Z}}}{2} \sqrt{\frac{\pi}{2}} p^{1/2} \frac{b}{\sqrt{1+b^2}},$$

where (a) follows from the fact that for any θ it holds that $\sin^{p-2}\theta \leq 1$, (b) follows from the fact that for $x > 0$ and $s \in (0,1)$ it holds that $\Gamma(x+1)/\Gamma(x+s) < (x+1)^{1-s}$ (Gautschi's inequality), and (c) follows from the fact that $\frac{\pi}{2} - \vartheta = \frac{\pi}{2} - \arccos(\cos(\vartheta)) \leq \frac{\pi}{2}\cos(\vartheta)$. The claim then follows since, for any $b > 0$, we have that $b/\sqrt{1+b^2} < b$. □

PROPOSITION B.3. *Let $\{Z_t\}_{t=1}^T$ be a sequence of centered Bernoulli random variables. Suppose that the α-mixing coefficients of the sequence satisfy $\alpha(l) < \exp(-K_\alpha l^{r_\alpha})$ for some $K_\alpha > 0$ and $r_\alpha > 0$.*

Define $p = \lfloor K_p T^{r_p} \rfloor$ for some $K_p > 0$ and $r_p \in [0, r_\alpha/(r_\alpha+1))$ and define

$$\varepsilon_T = \sqrt{\frac{K_1 K_2 p \log(T)}{T^{\frac{r_\alpha}{r_\alpha+1}}}} + \sqrt{\frac{K_2 \log(T)}{T^{\frac{r_\alpha}{r_\alpha+1}}}},$$

where $K_1 = 3/2$ and $K_2 = 64\sigma^2$ with $\sigma^2 = (\frac{1}{4} + 8\sum_{l=1}^{\infty}\alpha(l))$.

Then, for any $K_3 > 0$ and all T sufficiently large, it holds that

$$\left(1 + \frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z_t\right| > \varepsilon_T\right) \leq \frac{4}{T} + o\left(\frac{1}{T}\right).$$

Proof. We begin by noting that Z_t is zero mean and that $\sup_{1\leq t\leq T}\|Z_t\|_{L_\infty} \leq 1$, thus it satisfies the mixing and moment conditions of Theorem 2.1 of Liebscher (1996). Define $M_T = \lfloor T^{\frac{1}{r_\alpha+1}} \rfloor$ and note that for all $T \geq 2$ we have that $M_T \in [1,T]$ and $4M_T < T\varepsilon_T$, as required by the theorem. Then, we have

$$\mathbb{P}\left(\left|\sum_{t=1}^{T} Z_t\right| > T\varepsilon_T\right) \leq 4\exp\left(-\frac{T\varepsilon_T^2}{64D(T,M_T)/M_T + \frac{8}{3}M_T\varepsilon_T}\right) + 4\frac{T}{M_T}e^{-K_\alpha M_T^{r_\alpha}}$$

with $D(T,M_T) = \sup_{0\leq j\leq T-1}\mathbb{E}\left[\left(\sum_{t=j+1}^{j+M_T\wedge T} Z_t\right)^2\right]$. Define $\gamma(l) = \sup_{1\leq t\leq T-l}|\mathrm{Cov}(Z_t, Z_{t+l})|$ for $l = 0,\ldots,T-1$. Note that $\gamma(0) \leq 1/4$ and, by Billingsley's inequality (Bosq, 1998, Cor. 1.1), that $\gamma(l) \leq 4\alpha(l)$ for $l = 1,\ldots,T-1$. Thus, it holds that $D(T,M_T) \leq M_T\gamma(0) + 2M_T\sum_{l=1}^{M_T-1}\gamma(l) \leq M_T(\frac{1}{4} + 8\sum_{l=1}^{\infty}\alpha(l)) = M_T\sigma^2$. We have

$$\left(1 + \frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z_t\right| > \varepsilon_T\right)$$
$$\leq 4\left(1 + \frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p \exp\left(-\frac{T\varepsilon_T^2}{64\sigma^2 + \frac{8}{3}M_T\varepsilon_T}\right) + 4\left(1 + \frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p \frac{T}{M_T}e^{-K_\alpha M_T^{r_\alpha}}$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

$$\leq 4\left(1+\frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p \exp\left(-\frac{T^{\frac{r_\alpha}{r_\alpha+1}}\varepsilon_T^2}{64\sigma^2+\frac{8}{3}\varepsilon_T}\right)+8\left(1+\frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p T^{\frac{r_\alpha}{r_\alpha+1}} \exp\left(-\frac{K_\alpha}{2^{r_\alpha}} T^{\frac{r_\alpha}{r_\alpha+1}}\right)$$

$$\overset{(a)}{\leq} 4\left(1+\frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p \exp\left(-\frac{T^{\frac{r_\alpha}{r_\alpha+1}}}{64\sigma^2}\varepsilon_T^2\right)+8\left(1+\frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p T^{\frac{r_\alpha}{r_\alpha+1}} \exp\left(-\frac{K_\alpha}{2^{r_\alpha}} T^{\frac{r_\alpha}{r_\alpha+1}}\right)$$

$$= A_T + B_T,$$

where (a) follows from the fact that $x^2/(64s^2+8/3x) > x^2/(64s^2+8/3)$ for any x in $(0,1]$ and $\sigma^2 > 1/4$. Note that for all T sufficiently large, we have

$$\begin{aligned}
\log\left(1+\frac{K_3 p^{\frac{1}{2}}}{\varepsilon_T}\right)^p &= p\log\left(\varepsilon_T + K_3 p^{\frac{1}{2}}\right) - p\log(\varepsilon_T) \\
&= \frac{1}{2}p\log(p) + p\log\left(\frac{\varepsilon_T}{p^{\frac{1}{2}}}+K_3\right) - p\log(\varepsilon_T) \\
&\leq \frac{1}{2}p\log(K_p) + \frac{r_p}{2}p\log(T) + p\log(1+K_3) - p\log(\varepsilon_T) \\
&\leq \left(\frac{1}{2}+\frac{r_p}{2}+1+\frac{r_\alpha}{2(r_\alpha+1)}\right)p\log(T) < K_1 p\log(T),
\end{aligned}$$

where the last inequality follows from the fact that $r_p < 1$ and $r_\alpha/(2(r_\alpha+1)) < 1$. Finally, the claim follows after noting that for all T sufficiently large, we have

$$A_T \leq 4\exp\left(K_1 p\log(T) - \frac{T^{\frac{r_\alpha}{r_\alpha+1}}}{K_2}\varepsilon_T^2\right)$$

$$\overset{(a)}{\leq} 4\exp\left(K_1 p\log(T) - K_1 p\log(T) - \log(T)\right) \leq 4\exp\left(-\log(T)\right) = \frac{4}{T},$$

where (a) follows from the fact that $(x+y)^2 \geq x^2+y^2$ for $x, y \geq 0$, and that

$$B_T \leq 8\left(1+\frac{K_3 p^{1/2}}{\varepsilon_T}\right)^p T^{\frac{r_\alpha}{r_\alpha+1}} \exp\left(-\frac{K_\alpha}{2^{r_\alpha}} T^{\frac{r_\alpha}{r_\alpha+1}}\right)$$

$$\leq 8\exp\left(\frac{r_\alpha}{r_\alpha+1}\log(T) + K_1 K_p T^{r_p}\log(T) - \frac{K_\alpha}{2^{r_\alpha}} T^{\frac{r_\alpha}{r_\alpha+1}}\right) = o\left(\frac{1}{T}\right). \qquad \square$$

PROPOSITION B.4. *Let $\{\mathbf{Z}_t\}_{t=1}^T$ be a sequence of p-dimensional zero-mean random vectors. Suppose that (i) $\sup_{1\leq i\leq p}\sup_{1\leq t\leq T}\|Z_{it}\|_{L_2} \leq K_m$ and $\sup_{1\leq i\leq p}\sup_{1\leq t\leq T}\|Z_{it}\|_{L_{r_m}} \leq 2K_m$ for some $K_m \geq 1$ and $r_m > 2$; (ii) the α-mixing coefficients of the sequence satisfy $\alpha(l) < \exp(-K_\alpha l^{r_\alpha})$ for some $K_\alpha > 0$ and $r_\alpha > 0$; and (iii) $p = \lfloor K_p T^{r_p} \rfloor$ for some $K_p > 0$ and $r_p \in [0, (r_m-2)/2 \wedge 1)$.*

Then, for all T sufficiently large, it holds that

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^T \mathbf{Z}_t\right\|_2 > 12\sigma\sqrt{\frac{p\log(T)}{T}}\right) \leq \frac{3K_p(2K_m)^{r_m}}{\sigma\log(T)} + o\left(\frac{1}{\log(T)}\right),$$

where $\sigma^2 = K_m^2(1+128\frac{r_m}{r_m-2}\sum_{l=1}^\infty \alpha(l)^{1-\frac{2}{r_m}})$.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Proof. For any positive constant K, we have that

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}_t\right\|_2 > K\sqrt{\frac{p\log(T)}{T}}\right) \le \mathbb{P}\left(\max_{1\le i\le p}\left|\frac{1}{T}\sum_{t=1}^{T}Z_{it}\right| > K\sqrt{\frac{\log(T)}{T}}\right)$$
$$\le p\max_{1\le i\le p}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z_{it}\right| > K\sqrt{T\log(T)}\right).$$

Let $\sum_{t=1}^{T}Z_{it}=\sum_{t=1}^{T}Z'_{it}+\sum_{t=1}^{T}Z''_{it}$, where $Z'_{it}=Z_{it}\mathbb{1}(|Z_{it}|\le b_T)-\mathbb{E}(Z_{it}\mathbb{1}(|Z_{it}|\le b_T))$ and $Z''_{it}=Z_{it}\mathbb{1}(|Z_{it}|> b_T)-\mathbb{E}(Z_{it}\mathbb{1}(|Z_{it}|> b_T))$. For any $\lambda\in(0,1)$, we have

$$p\max_{1\le i\le p}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z_{it}\right| > K\sqrt{T\log(T)}\right)$$
$$\le p\max_{1\le i\le p}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z'_{it}\right| > \lambda K\sqrt{T\log(T)}\right)+p\max_{1\le i\le p}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z''_{it}\right| > (1-\lambda)K\sqrt{T\log(T)}\right).$$

The sequence $\{Z'_{it}\}_{t=1}^{T}$ has the same mixing properties as $\{\mathbf{Z}_t\}_{t=1}^{T}$ and $\sup_{1\le i\le p}\sup_{1\le t\le T}\|Z'_{it}\|_\infty < 2b_T$. Define $\varepsilon'_T=\lambda KT^{\frac{1}{2}}\sqrt{\log(T)}$, $b_T=(T^{\frac{1+2r_p}{2}}\sqrt{\log(T)})^{\frac{1}{r_m-1}}$ and $M_T=\lfloor b_T^{-1}T^{\frac{1}{2}}/\sqrt{\log(T)}\rfloor$. For all T sufficiently large, the conditions of Theorem 2.1 of Liebscher (1996) are satisfied, since for all T sufficiently large, we have that $M_T\in[1,T]$ and $4(2b_T)M_T<\varepsilon'_T$ and we have

$$p\max_{1\le i\le p}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z'_{it}\right| > \varepsilon'_T\right)$$
$$\le 4p\exp\left(-\frac{(\varepsilon'_T)^2}{64\frac{T}{M_T}D(T,M_T)+\frac{16}{3}b_TM_T\varepsilon'_T}\right)+4\frac{pT}{M_T}\exp\left(-K_\alpha M_T^{r_\alpha}\right)$$

with $D(T,M_T)=\sup_{0\le j\le T-1}\mathbb{E}\left[\left(\sum_{t=j+1}^{j+M_T\wedge T}Z'_{it}\right)^2\right]$. Define $\gamma(l)=\sup_{1\le i\le p}\sup_{1\le t\le T-l}|\mathrm{Cov}(Z'_{it},Z'_{it+l})|$ for $l=0,\dots,T-1$ and note that $D(T,M_T)\le M_T\sum_{l=-T+1}^{T-1}\gamma(l)$. Next, we note that $\gamma(0)\le K_m^2$ since

$$\mathrm{Var}(Z'_{it})=\|Z_{it}\mathbb{1}(|Z_{it}|\le b_T)\|_{L_2}^2-[\mathbb{E}(Z_{it}\mathbb{1}(|Z_{it}|\le b_T))]^2\le\|Z_{it}\|_{L_2}^2\le K_m^2.$$

Davydov's inequality (Bosq, 1998, Cor. 1.1) implies that

$$\gamma(l)\le 4\frac{r_m}{r_m-2}\alpha(l)^{1-\frac{2}{r_m}}\|Z'_{it}\|_{L_{r_m}}\|Z'_{it+l}\|_{L_{r_m}}\le 64K_m^2\frac{r_m}{r_m-2}\alpha(l)^{1-\frac{2}{r_m}},$$

for $l=1,\dots,T-1$, where we have used the fact that

$$\|Z'_{it}\|_{L_{r_m}}\le\|Z_{it}\mathbb{1}(|Z_{it}|\le b_T)\|_{L_{r_m}}+\|\mathbb{E}(Z_{it}\mathbb{1}(|Z_{it}|\le b_T))\|_{L_{r_m}}$$
$$\le\|Z_{it}\mathbb{1}(|Z_{it}|\le b_T)\|_{L_{r_m}}+\|Z_{it}\mathbb{1}(|Z_{it}|\le b_T)\|_{L_1}$$
$$\le 2\|Z_{it}\mathbb{1}(|Z_{it}|\le b_T)\|_{L_{r_m}}\le 2\|Z_{it}\|_{L_{r_m}}\le 4K_m.$$

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

These together imply that $D(T,M_T) \leq M_T K_m^2(1+128\frac{r_m}{r_m-2}\sum_{l=1}^{\infty}\alpha(l)^{1-\frac{2}{r_m}}) = M_T\sigma^2$. For any K that satisfies

$$K > \frac{1}{\lambda}\left(8\sqrt{\sigma^2+\frac{1}{9}}+\frac{8}{3}\right), \tag{B.4}$$

we have $1-\lambda^2K^2/(64\sigma^2+\frac{16}{3}\lambda K) < 0$. Notice that the condition is satisfied, for instance, by $K = \lambda^{-1}8\sqrt{2\sigma^2}$ since $\sigma^2 \geq 1$. Thus, for any K that satisfies this, we have

$$\begin{aligned} p\max_{1\leq i\leq p}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z'_{it}\right| > \varepsilon'_T\right) &\leq 4K_p\exp\left(r_p\log(T) - \frac{\lambda^2K^2T\log(T)}{64\sigma^2T+\frac{16}{3}\lambda KT}\right) + 4K_pT^{1+r_p}\exp\left(-K_\alpha M_T^{r_\alpha}\right) \\ &\leq 4K_p\exp\left(\left[r_p - \frac{\lambda^2K^2}{64\sigma^2+\frac{16}{3}\lambda K}\right]\log(T)\right) + 4K_pT^{1+r_p}\exp\left(-K_\alpha M_T^{r_\alpha}\right) \\ &\leq o\left(\frac{1}{\log(T)}\right). \end{aligned}$$

Let $\varepsilon''_T = (1-\lambda)KT^{\frac{1}{2}}\sqrt{\log(T)}$ and note that

$$\begin{aligned} p\max_{1\leq i\leq p}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z''_{it}\right| > \varepsilon''_T\right) &\overset{(a)}{\leq} \frac{p}{\varepsilon''_T}\max_{1\leq i\leq p}\mathbb{E}\left|\sum_{t=1}^{T}Z''_{it}\right| \leq \frac{p}{\varepsilon''_T}\sum_{t=1}^{T}\max_{1\leq i\leq p}\mathbb{E}|Z''_{it}| \\ &\leq \frac{2p}{\varepsilon''_T}\sum_{t=1}^{T}\max_{1\leq i\leq p}\mathbb{E}\,|Z_{it}\mathbb{1}(|Z_{it}| > b_T)| \\ &\overset{(b)}{\leq} \frac{2p}{\varepsilon''_T}\sum_{t=1}^{T}\frac{\max_{1\leq i\leq p}\mathbb{E}|Z_{it}|^{r_m}}{b_T^{r_m-1}} \leq \frac{2pT(2K_m)^{r_m}}{\varepsilon''_Tb_T^{r_m-1}} = \frac{2K_p(2K_m)^{r_m}}{(1-\lambda)K\log(T)}, \end{aligned}$$

where (a) follows from Markov's inequality and (b) from the inequality $\mathbb{E}(|Z\mathbb{1}(|Z| > b)|) \leq \mathbb{E}(|Z|^r)/b^{r-1}$ for a random variable Z with finite rth moment and positive constant b. The claim follows after picking $\lambda = 8\sqrt{2}/12$ and noticing that $K = 12\sigma$ satisfies (B.4). □

REFERENCES

Andrews, D.W.K. (1991) Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59(2), 307–345.

Audibert, J.-Y. & O. Catoni (2011) Robust linear least squares regression. *Annals of Statistics* 39, 2766–2794.

Babii, A., E. Ghysels, & J. Striaukas (2023) High-dimensional Granger causality tests with an application to VIX and news. *Journal of Financial Econometrics*, forthcoming.

Bai, J. & S. Ng (2002) Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.

Belloni, A., V. Chernozhukov, D. Chetverikov, & K. Kato (2015) Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2), 345–366.

Birge, L. & P. Massart (1998) Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* 4, 329–375.

Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*, 2nd Edition. Springer.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Brownlees, C., E. Joly, & G. Lugosi (2015) Empirical risk minimization for heavy-tailed losses. *Annals of Statistics* 43, 2507–2536.

Bunea, F., A.B. Tsybakov, & M.H. Wegkamp (2007) Aggregation for Gaussian regression. *Annals of Statistics* 35, 1673–1697.

Caner, M. & K. Knight (2013) An alternative to unit root tests: Bridge estimators differentiate between nonstationary versus stationary models and select optimal lag. *Journal of Statistical Planning and Inference* 143, 691–715.

Chen, X. (2006). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. E. Leamer (eds), *Handbook of Econometrics*, pp. 5549–5632. North-Holland.

Chen, X. & X. Shen (1998) Sieve extremum estimates for weakly dependent data. *Econometrica* 66(2), 289–314.

Dendramis, Y., L. Giraitis, & G. Kapetanios (2021) Estimation of time-varying covariance matrices for large datasets. *Econometric Theory* 37(6), 1100–1134.

Emery, M., A. Nemirovski, & D. Voiculescu (2000) Ecole d'Ete de Probabilites de Saint-Flour XXVIII-1998. In P. Bernard (ed), *Lecture Notes in Mathematics*, Vol. 1738, pp. 87–285.

Fan, J., Y. Liao, & M. Mincheva (2011) High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* 39, 3320–3356.

Fang, K.-T., S. Kotz, & K.W. Ng (1990) *Symmetric Multivariate and Related Distributions*. Chapman and Hall).

Fang, K.-T. & Y.-T. Zhang (1990) *Generalized Multivariate Analysis*. Springer.

Forni, M., M. Hallin, M. Lippi, & L. Reichlin (2000) The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* 82, 540–554.

Garcia, M.G., M.C. Medeiros, & G.F. Vasconcelos (2017) Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting* 33(3), 679–693.

Hansen, B.E. (2008) Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 24, 726–748.

Hastie, T., R. Tibshirani, & J. Friedman (2001) *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

Ibragimov, I.A. (1962) Some limit theorems for stationary processes. *Theory of Probability and its Applications* 7, 349–382.

Jiang, W. & M.A. Tanner (2010) Risk minimization for time series binary choice with variable selection. *Econometric Theory* 26, 1437–1452.

Kock, A.B. & L. Callot (2015) Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* 186, 325–344.

Lecué, G. & S. Mendelson (2016) Performance of empirical risk minimization in linear aggregation. *Bernoulli* 22, 1520–1534.

Lecué, G. & S. Mendelson (2017) Regularization and the small-ball method ii: Complexity dependent error rates. *Journal of Machine Learning Research* 18, 1–48.

Lecué, G. & S. Mendelson (2018) Regularization and the small-ball method i: Sparse recovery. *Annals of Statistics* 46, 611–641.

Li, Q. & J. Racine (2006) *Nonparamtric Econometrics: Theory and Practice*. Princeton University Press.

Liao, Z. & P.C.B. Phillips (2015) Automated estimation of vector error correction models. *Econometric Theory* 31(3), 581–646.

Liebscher, E. (1996) Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications*, 65, 69–80.

Medeiros, M.C. & E.F. Mendes (2016) ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics* 191, 255–271.

Meitz, M. & P. Saikkonen (2008) Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models. *Econometric Theory* 24(5), 1291–1320.

Mendelson, S. (2015) Learning without concentration. *Journal of the ACM* 62(3), 1–25.

Mendelson, S. (2018) Learning without concentration for general loss functions. *Probability Theory and Related Fields* 171(1), 459–502.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press

Miao, K., P.C.B. Phillips, & L. Su (2023) High-dimensional VARs with common factors. *Journal of Econometrics* 233(1), 155–183.

Newey, W.K. (1997) Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79(1), 147–168.

Onatski, A. (2012) Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* 168(2), 244–258.

Rio, E. (1995) The functional law of the iterated logarithm for stationary strongly mixing sequences. *Annals of Probability* 23, 1188–1203.

Stock, J.H. & M.W. Watson (2002) Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–1179.

Stone, C. (1985) Additive regression and other nonparametric models. *Annals of Statistics* 13(2), 689–705.

Su, L. & H. White (2010) Testing structural change in partially linear models. *Econometric Theory* 26(6), 1761–1806.

Tsybakov, A.B. (2003) Optimal rates of aggregation. In B. Schölkopf and M.K. Warmuth (eds), *Learning Theory and Kernel Machines*. Lecture Notes in Computer Science, Vol. 2777. Springer.

Tsybakov, A.B. (2014) Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians (Seoul, August 2014)*, pp. 225–246.

Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M.J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

White, H. (2001) *Asymptotic Theory for Econometricians*, *Revised Edition*. Academic Press.

https://doi.org/10.1017/S0266466623000348 Published online by Cambridge University Press