Performance of Empirical Risk Minimization For Principal Component Regression

Christian Brownlees[†,*] Guðmundur Stefán Guðmundsson[‡] Yaping Wang[†]

September 20, 2024

Abstract

This paper establishes bounds on the predictive performance of empirical risk minimization for principal component regression. Our analysis is nonparametric, in the sense that the relation between the prediction target and the predictors is not specified. In particular, we do not rely on the assumption that the prediction target is generated by a factor model. In our analysis we consider the cases in which the largest eigenvalues of the covariance matrix of the predictors grow linearly in the number of predictors (strong signal regime) or sublinearly (weak signal regime). The main result of this paper shows that empirical risk minimization for principal component regression is consistent for prediction and, under appropriate conditions, it achieves near-optimal performance in both the strong and weak signal regimes.

Keywords: empirical risk minimization, principal component regression, time series, oracle inequality

JEL: C13, C14, C22, C55

[†] Department of Economics and Business, Universitat Pompeu Fabra and Barcelona GSE;
e-mail: christian.brownlees@upf.edu, yaping.wang@upf.edu.
[‡] Department of Economics and Business Economics, Aarhus University;
e-mail: gsgudmundsson@econ.au.dk.
[*] Corresponding author.
We have benefited from discussions with Gabor Lugosi, David Rossell and Piotr Zwiernik.
Christian Brownlees acknowledges support from the Spanish Ministry of Science and Technology (Grant MTM2012-37195); the Ayudas Fundación BBVA Proyectos de Investigación Cientìfica en Matemáticas 2021; the Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2011-0075).
Guðmundur Stefán Guðmundsson acknowledges financial support from the Danish National Research Foundation (DNRF Chair grant number DNRF154).

1 Introduction

Principal component regression (PCR) is a regression methodology with a long and well established tradition that can be traced back to at least Hotelling (1957) and Kendall (1957). In a nutshell, PCR consists in forecasting a prediction target of interest on the basis of the principal components extracted from a potentially large set of predictors. PCR is a popular tool for forecasting in macroeconomics where it is documented to perform favourably relative to a number of competing approaches (Stock and Watson, 2012).

In this paper we study the properties of PCR from a learning theory perspective. Our main contribution consists in establishing nonasymptotic prediction performance guarantees for PCR. Our main result may be interpreted as a nonasymptotic analogue of classic asymptotic results on the prediction properties of PCR obtained in Stock and Watson (2002). Our analysis shows that, under appropriate conditions, PCR achieves near-optimal performance. An important feature of our analysis is that we treat PCR as a regularization procedure and we do not assume that the data are generated by a factor model. In particular, as is customary in learning theory, the relation between the prediction target and the predictors is not specified. That being said, as the factor model feinschmecker will recognize, our framework relies on assumptions and proof strategies analogous to the ones used in the factor model literature. In particular we build upon classic contributions such as Bai and Ng (2002), Bai (2003), Fan, Liao, and Mincheva (2011, 2013) and Onatski (2012) among others.

PCR may be described as a two-step procedure. Let $\mathcal{D} = \{(Y_t, \boldsymbol{X}_t')'\}_{t=1}^T$ be a stationary sequence of zero-mean random vectors taking values in $\mathcal{Y} \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}^p$. The goal is to forecast the prediction target Y_t using the p-dimensional vector of predictors $\boldsymbol{X}_t = (X_{1\,t}, \ldots, X_{p\,t})'$. The first step of PCR consists in computing the $T \times K$ principal components matrix $\widehat{\boldsymbol{P}} = (\widehat{\boldsymbol{P}}_1, \ldots, \widehat{\boldsymbol{P}}_T)'$ associated with the $T \times p$ predictor matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)'$, for some appropriate choice of K. This may be defined as the solution of the constrained least squares problem

$$(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{P}}) = \arg \min_{\substack{\boldsymbol{B}\in\mathbb{R}^{p\times K}\\ \boldsymbol{P}\in\mathbb{R}^{T\times K}}} \|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{B}'\|_F^2 \ \text{ s.t. } \frac{1}{T}\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_K, \frac{1}{p}\boldsymbol{B}'\boldsymbol{B} \text{ is diagonal,}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. As is well known, $\widehat{\boldsymbol{P}}$ is given by $\sqrt{T}$ times the first K eigenvectors of the matrix $\boldsymbol{X}\boldsymbol{X}'$. It is useful to remark here that the principal components allow us to express the vector of predictors $\boldsymbol{X}_t$ as a linear combination of the matrix of coefficients $\widehat{\boldsymbol{B}}$ with the K principal components $\widehat{\boldsymbol{P}}_t$ plus a residual vector $\widehat{\boldsymbol{u}}$, that is

$$\boldsymbol{X}_t = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{P}}_t + \widehat{\boldsymbol{u}}_t \ , \tag{1}$$

where $\widehat{\boldsymbol{u}}_t = \boldsymbol{X}_t - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{P}}_t$. The second step of PCR consists in computing the $K \times 1$ least squares coefficients vector $\hat{\boldsymbol{\vartheta}}$ associated with the regression of the $T \times 1$ target variable vector $\boldsymbol{Y} = (Y_1, \ldots, Y_T)'$ on the principal components matrix $\widehat{\boldsymbol{P}}$. This is the solution to the least squares problem

$$\hat{\boldsymbol{\vartheta}} = \arg \min_{\boldsymbol{\vartheta}\in\mathbb{R}^K} \|\boldsymbol{Y} - \widehat{\boldsymbol{P}}\boldsymbol{\vartheta}\|_2^2 \ ,$$

where $\|\cdot\|_2$ denotes the Euclidean norm. It is straightforward to check that $\hat{\boldsymbol{\vartheta}} = \widehat{\boldsymbol{P}}'\boldsymbol{Y}/T$.

PCR may be interpreted as regularized empirical risk minimization.[1] Consider the class of prediction rules indexed by $\boldsymbol{\theta} \in \mathbb{R}^p$ given by $f_{\boldsymbol{\theta}\,t} = \boldsymbol{\theta}'\boldsymbol{X}_t$. Then PCR can be cast as the regularized empirical risk minimization problem given by

$$\hat{\boldsymbol{\theta}}_{PCR} \in \arg \min_{\boldsymbol{\theta}\in\mathbb{R}^p} R_T(\boldsymbol{\theta}) \quad \text{s.t.} \quad \widehat{\boldsymbol{V}}_R'\boldsymbol{\theta} = \boldsymbol{0} \ , \tag{2}$$

where

$$R_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(Y_t - f_{\boldsymbol{\theta}\,t})^2$$

is the empirical risk and

$$\widehat{\boldsymbol{V}}_R = (\widehat{\boldsymbol{v}}_{K+1}, \ldots, \widehat{\boldsymbol{v}}_p) \ ,$$

[1] We remark that in this paper we use the expression "empirical risk minimization for principal component regression" only for simplicity. A more appropriate name for the procedure we study would be "regularized empirical risk minimization based on principal component analysis".

where $\widehat{\boldsymbol{v}}_i$ is the eigenvector associated with the i-th largest eigenvalue of the sample covariance matrix of the predictors $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{X}/T$. The vector $\hat{\boldsymbol{\theta}}_{PCR}$ defined in (2) is the solution to a least squares problem subject to a set of linear constraints. It is straightforward to verify that $\hat{f}_t^{PCR} = \hat{\boldsymbol{\theta}}'_{PCR}\boldsymbol{X}_t = \hat{\boldsymbol{\vartheta}}'\widehat{\boldsymbol{P}}_t$, which implies that the forecasts produced by PCR may be equivalently expressed as linear forecasts based on constrained least squares.

One of the main objectives of learning theory is to obtain a bound on the predictive performance of the ERM relative to the optimal risk that can be achieved within the given class of prediction rules. We define the risk of a prediction rule as $R(\boldsymbol{\theta}) = \mathbb{E}\left[(Y_t - f_{\boldsymbol{\theta}\, t})^2\right]$. The optimal risk is defined as the risk of a prediction rule associated with a $\boldsymbol{\theta}^*$ such that

$$R(\boldsymbol{\theta}^*) = \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathbb{E}\left[(Y_t - f_{\boldsymbol{\theta}\, t})^2\right] \ .$$

The conditional risk of PCR is used to measure predictive performance. This is given by

$$R(\hat{\boldsymbol{\theta}}_{PCR}) = \mathbb{E}\left[(Y_t - \hat{f}_t^{PCR})^2 \middle| \hat{\boldsymbol{\theta}}_{PCR} = \hat{\boldsymbol{\theta}}_{PCR}(\mathcal{D})\right] \ , \tag{3}$$

where $(Y_t, \boldsymbol{X}_t)'$ in (3) denotes an element of the process assumed to be drawn independently of $\mathcal{D}$. The performance measure in (3) can be interpreted as the risk of the ERM obtained from the "training" sample $\mathcal{D}$ over the "validation" observation $(Y_t, \boldsymbol{X}_t)'$. Then our aim is to find a pair $(B_T(p,K), \delta_T)$ such that $B_T(p,K) \to 0$ and $\delta_T \to 0$ as $T \to \infty$ for which

$$R(\hat{\boldsymbol{\theta}}_{PCR}) \leq R(\boldsymbol{\theta}^*) + B_T(p,K) \tag{4}$$

holds with probability at least $1 - \delta_T$ for any T sufficiently large. 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 and imply that the ERM asymptotically performs as well as the best linear predictor. We remark that the performance measure in (3) allows us to keep our analysis close to the bulk of the contributions in the learning theory literature (which typically focus on the analysis of i.i.d. data) and facilitates comparisons. We remark that Brownlees and Guðmundsson (2025) and Brownlees and Llorens-Terrazas (2025) consider alternative performance measures such as the conditional out-of-sample average risk of the ERM, which has a more attractive interpretation for time series applications. It turns out that these alternative measures lead to essentially the same theoretical analysis at the expense of introducing additional notation. We therefore focus on the performance measure defined in (4) for clarity.

This paper is related to various strands of the literature. First, the vast literature on approximate factor models, principal component analysis and spiked covariance models, which includes Wang and Fan (2017), Donoho, Gavish, and Johnstone (2018), Forni, Hallin, Lippi, and Reichlin (2000, 2005), Bai and Ng (2006, 2019), Bai and Li (2012, 2016), Lam, Yao, and Bathia (2011), Fan, Liao, and Wang (2016), Gonçalves and Perron (2020), Barigozzi and Cho (2020), Su and Wang (2017) and Fan, Masini, and Medeiros (2024). In particular, this work is close to the important contribution of Fan *et al.* (2024) that studies the properties of a large class of high-dimensional models, which includes factor models, and establishes results on the predictive performance of such a class. Second, it is related to the literature on the small-ball method, which includes Lecué and Mendelson (2016), Lecué and Mendelson (2017), Mendelson (2018) and Lecué and Mendelson (2018). Third, the literature on empirical risk minimization for linear regression, which includes Birge and Massart (1998) and Tsybakov (2003) among others. In particular, our contribution is close to the subset of the literature that deals with dependent data, as in Jiang and Tanner (2010); Brownlees and Guðmundsson (2025).

The rest of the paper is structured as follows. Section 2 introduces additional basic notation, assumptions and preliminary results. Section 3 contains the main result of the paper and its proof is outlined in Section 4. Concluding remarks follow in Section 5. All the remaining proofs are given in the appendix.

2 Notation, Preliminaries and Assumptions

In this section we lay out the assumptions required for our analysis. Our assumptions may be interpreted as the union of standard conditions employed in the approximate factor model literature (Bai and Ng, 2002; Fan *et al.*, 2011) and conditions employed in the learning literature based on the small-ball method (Mendelson, 2018; Lecué and Mendelson, 2018; Brownlees and Guðmundsson, 2025).

We introduce some basic notation. 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 $\mathbf{M}^{1/2}$ to denote the positive semi-definite square root matrix of $\mathbf{M}$ and $\mathbf{M}^{-1/2}$ to denote the generalized inverse of $\mathbf{M}^{1/2}$. For a generic matrix $\mathbf{M}$ we use $\|\mathbf{M}\|_2$ to denote the spectral norm of $\mathbf{M}$.

We begin with a preliminary lemma that establishes the population analogue of the principal components decomposition of the prediction vector $\boldsymbol{X}_t$ introduced in (1).

Lemma 1. *Let $\boldsymbol{X}_t$ be a zero-mean p-dimensional random vector with $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t')$. For any $K \in \{1, \ldots p\}$, define $\boldsymbol{\Lambda}_K = \text{diag}(\lambda_1, \ldots, \lambda_K)$, $\boldsymbol{\Lambda}_R = \text{diag}(\lambda_{K+1}, \ldots, \lambda_p)$, $\boldsymbol{V}_K = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)$ and $\boldsymbol{V}_R = (\boldsymbol{v}_{K+1}, \ldots, \boldsymbol{v}_p)$ where $\lambda_1, \ldots, \lambda_p$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ denote the sequence of eigenvalues of $\boldsymbol{\Sigma}$ in a non-increasing order and the corresponding sequence of eigenvectors.*

Then, (i) it holds that

$$\boldsymbol{X}_t = \mathbf{B}\boldsymbol{P}_t + \boldsymbol{u}_t \ ,$$

where $\mathbf{B} = \boldsymbol{V}_K\boldsymbol{\Lambda}_K^{1/2}$, $\boldsymbol{P}_t = \boldsymbol{\Lambda}_K^{-1/2}\boldsymbol{V}_K'\boldsymbol{X}_t$ and $\boldsymbol{u}_t = \boldsymbol{V}_R(\boldsymbol{V}_R'\boldsymbol{V}_R)^{-1}\boldsymbol{V}_R'\boldsymbol{X}_t$ with $\mathbf{B}'\mathbf{B}$ diagonal, $\mathbb{E}(\boldsymbol{P}_t\boldsymbol{P}_t') = \boldsymbol{I}_K$, $\mathbb{E}(\boldsymbol{u}_t\boldsymbol{u}_t') = \boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'$ and $\mathbb{E}(\boldsymbol{P}_t\boldsymbol{u}_t') = \mathbf{0}_{K\times p}$. (ii) Let $\boldsymbol{\theta}^* \in \mathbb{R}^p$ be the vector of coefficients of the best linear predictor of Y_t based on $\boldsymbol{X}_t$. Then, it holds that

$$\boldsymbol{X}_t'\boldsymbol{\theta}^* = \boldsymbol{P}_t'\boldsymbol{\vartheta}^* + \boldsymbol{u}_t'\boldsymbol{\gamma}^* \ ,$$

where $\boldsymbol{\vartheta}^* = \boldsymbol{\Lambda}_K^{1/2}\boldsymbol{V}_K'\boldsymbol{\theta}^*$ and $\boldsymbol{\gamma}^* = \boldsymbol{V}_R(\boldsymbol{V}_R'\boldsymbol{V}_R)^{-1}\boldsymbol{V}_R'\boldsymbol{\theta}^*$.

Parts (i) of Lemma 1 states that $\boldsymbol{X}_t$ can decomposed into a linear combination of the matrix of coefficients $\mathbf{B}$ with the random vector $\boldsymbol{P}_t$ plus the residual random vector $\boldsymbol{u}_t$, where $\boldsymbol{P}_t$ and $\boldsymbol{u}_t$ are orthogonal. We call $\boldsymbol{P}_t$ the population principal components and $\boldsymbol{u}_t$

the idiosyncratic component. Part (*ii*) implies that the best linear predictor for Y_t can be alternatively represented as a function of the population principal components $\boldsymbol{P}_t$ and the idiosyncratic component $\boldsymbol{u}_t$ with the coefficient vectors $\boldsymbol{\vartheta}^*$ and $\boldsymbol{\gamma}^*$. Note that since $\boldsymbol{P}_t$ and $\boldsymbol{u}_t$ are orthogonal, $\boldsymbol{\vartheta}^*$ may be interpreted as the best linear predictor based on the population principal components $\boldsymbol{P}_t$. We remark that, clearly, in the definitions of $\boldsymbol{u}_t$ and $\boldsymbol{\gamma}^*$ the matrix $\boldsymbol{V}_R(\boldsymbol{V}_R'\boldsymbol{V}_R)^{-1}\boldsymbol{V}_R'$ may be simplified to $\boldsymbol{V}_R\boldsymbol{V}_R'$ but we prefer to express it in this way to emphasize that this is a projection matrix. Last note that the decomposition established in part (*ii*) holds for any vector in $\mathbb{R}^p$ but for our purposes it suffices to focus on the vector of coefficients of the best linear predictor. Also notice that the vector of coefficients of the best linear predictor may not be unique, but this raises no issues in our setup.

We lay out the assumptions of our analysis. We say that the d-dimensional random vector $\boldsymbol{U}$ is sub-Gaussian with parameters $C_m > 0$ if, for any $\varepsilon > 0$, it holds that

$$\mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} |\boldsymbol{v}'\boldsymbol{U}| > \varepsilon \right) \leq \exp(-C_m\varepsilon^2) \ .$$

For a univariate random variable U this is equivalent to $\mathbb{P}(|U| > \varepsilon) \leq \exp(-C_m\varepsilon^2)$.

A.1 (Distribution). *(i) There exists a positive constant C_m such that Y_t, $Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*$ and $\boldsymbol{Z}_t = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_t$ with $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t')$ are sub-Gaussian with parameter C_m. (ii) There exists a constant C_Z and a p-dimensional spherical random vector $\boldsymbol{S}$ such that for any $B \in \mathcal{B}(\mathbb{R}^p)$ it holds that $\mathbb{P}(\boldsymbol{Z}_t \in B) \leq C_Z\mathbb{P}(\boldsymbol{S} \in B)$, where the spherical random vector $\boldsymbol{S}$ is such that its density exists and the marginal densities of its components are bounded from above.*

Part (*i*) states that the tails of the data decay exponentially. More precisely it assumes that the prediction target Y_t, the prediction error of the best linear predictor based on the principal components $\boldsymbol{P}_t$ given by $Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*$ and the standardized predictors $\boldsymbol{Z}_t$ have sub-Gaussian tails. We remark that such a condition is fairly standard in the analysis of large-dimensional factor models (Fan *et al.*, 2011). We also remark that the sub-Gaussian condition may be replaced by a sub-Weibull condition at the expense of longer proofs

(Wong, Li, and Tewari, 2020). Part (*ii*) is a regularity condition on the density of the standardized predictors $\boldsymbol{Z}_t$ that 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 same condition is assumed in Brownlees and Guðmundsson (2025).

Let $\mathcal{F}_{-\infty}^{t}$ and $\mathcal{F}_{t+l}^{\infty}$ be the σ-algebras generated by $\{(Y_s, \boldsymbol{X}_s')' : -\infty \leq s \leq t\}$ and $\{(Y_s, \boldsymbol{X}_s')' : t+l \leq s \leq \infty\}$ respectively for some $t \in \mathbb{Z}$ and define the α-mixing coefficients

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

A.2 (Dependence). *There exist constants $C_\alpha > 0$ and $r_\alpha > 0$ such that the α-mixing coefficients satisfy $\alpha(l) \leq \exp(-C_\alpha l^{r_\alpha})$.*

A.2 states that the process $\{(Y_t, \boldsymbol{X}_t')'\}$ has geometrically decaying strong mixing coefficients, which is a fairly standard assumption in the analysis of large dimensional time series models (Jiang and Tanner, 2010; Fan *et al.*, 2011; Kock and Callot, 2015).

A.3 (Eigenvalues). *There is an integer $K \in \{1, \ldots, p\}$, a constant $\alpha \in (1/2, 1]$ and a sequence of non-increasing nonnegative constants $c_1, \ldots, c_p$ with $c_K > 0$ such that, $\lambda_i = c_i p^\alpha$ for $i = 1, \ldots, K$, $\lambda_i = c_i$ for $i = K+1, \ldots, p$.*

A.3 states that the first K eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ diverge as the cross-sectional dimension p becomes large. The rate of divergence is determined by α. We distinguish between two regimes that depend on the value of this constant. When $\alpha = 1$ we say that we are in the strong signal regime, which is analogous to (classic) factor models with pervasive factors (Stock and Watson, 2002; Bai and Ng, 2002; Bai, 2003; Fan *et al.*, 2013). When $\alpha \in (1/2, 1)$ we say that we are in the weak signal regime, which is analogous to weak factor models (Onatski, 2012; Bai and Ng, 2023). We remark that this assumption is weaker than Fan *et al.* (2013), which only allows for strong signals. We also point out that K here is assumed to be known and that there is a large literature devoted to the estimation of this quantity (Bai and Ng, 2002; Amengual and Watson, 2007; Onatski, 2010; Lam and Yao, 2012; Ahn and Horenstein, 2013; Yu, He, and Zhang, 2019). Last, it is important to emphasize that the assumption allows for the non-diverging eigenvalues of $\boldsymbol{\Sigma}$ to be zero, so our framework allows $\boldsymbol{\Sigma}$ to be singular.

A.4 (Number of Predictors and Principal Components). *(i) There are constants $C_p > 0$ and $r_p \in (0, \overline{r}_p)$ such that $p = \lfloor C_p T^{r_p} \rfloor$ where $\overline{r}_p = r_\alpha \wedge \frac{1}{\frac{r_\alpha+1}{r_\alpha} - \alpha}$. (ii) There are constants $C_K > 0$ and $r_K \in [0, \overline{r}_K)$ such that $K = \lfloor C_K T^{r_K} \rfloor$ where $\overline{r}_K = 1 - r_p(1-\alpha) \wedge \frac{1}{3+\frac{2}{r_\alpha}}$.*

A.4 states that the number of predictors and the number of principal components are allowed to increase as a function of the sample size T. The rate of growth of these quantities depends on the rate of decay of the strong mixing coefficients (as measured by r_α) and the strength of the signal (as measured by α). The less dependence and the stronger the signal, the larger the numbers of allowed predictors and principal components. The assumption allows the number of predictors to be larger than the sample size T whereas the number of principal components is at most $T^{1/3}$ (up to a proportionality constant). It is important to emphasize that when the signal is weaker, the maximum rate of growth of the number of predictors is smaller. The condition on the maximum rate of growth of the number of predictors $\overline{r}_p$ is analogous to condition (9) in Uematsu and Yamagata (2022), who study the properties of factor models in the weak signal regime.

A.5 (Identification/Small-ball). *There exist positive constants κ_1 and κ_2 such that, for each $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$, and for each $t = 1, \ldots, T$*

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

A.5 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 $\boldsymbol{\delta} = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$ then the condition is equivalent to $\mathbb{P}\left(|\boldsymbol{\delta}'\boldsymbol{X}_t| \geq \kappa_1 \|\boldsymbol{\delta}'\boldsymbol{X}_t\|_{L_2}\right) \geq \kappa_2$, which can be seen as requiring the random variable $\boldsymbol{\delta}'\boldsymbol{X}_t$ to not have excessive mass in a neighbourhood 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. We also remark that Brownlees and Guðmundsson (2025) discuss conditions that imply A.5.

3 Performance of Empirical Risk Minimization

We now state our main result on the performance of empirical risk minimization.

Theorem 1. *Suppose A.1–A.5 are satisfied.*

Then for any $\eta > 0$ there exists a constant $C > 0$ such that, for any T sufficiently large,

$$R(\hat{\boldsymbol{\theta}}_{PCR}) \leq R(\boldsymbol{\theta}^*) + 2(\boldsymbol{\theta}^*)'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{\theta}^* + C\left[\frac{1}{p^{2\alpha-1}} + \left(\frac{p}{Tp^{\alpha}}\right)^2 p^{\frac{2}{r_\alpha}} + \frac{K}{T}\right]\log(T),$$

holds with probability at least $1 - T^{-\eta}$.

The theorem establishes a regret bound on the excess risk of PCR relative to the best linear predictor that can be obtained on the basis of the predictors $\boldsymbol{X}_t$. The gap is made up of two terms. The first can interpreted as the approximation error of PCR and the second as the estimation error. The approximation error measures the gap between the performance of the best linear predictor based on the population principal components $\boldsymbol{P}_t$ and the best linear predictor based on the predictors $\boldsymbol{X}_t$. The estimation error measures the gap between the performance of PCR relative to the best linear predictor based on the population principal components $\boldsymbol{P}_t$.

A number of additional remarks on Theorem 1 are in order. First, it is insightful to provide an alternative representation of the approximation error of PCR. This may be equivalently expressed as

$$(\boldsymbol{\theta}^*)'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{\theta}^* = \|\boldsymbol{\Lambda}_R^{1/2}\boldsymbol{V}_R'\boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\Lambda}_R^{1/2}\boldsymbol{V}_R'\boldsymbol{V}_R(\boldsymbol{V}_R'\boldsymbol{V}_R)^{-1}\boldsymbol{V}_R'\boldsymbol{\theta}^*\|_2^2 = (\boldsymbol{\gamma}^*)'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{\gamma}^* \ .$$

This highlights that the approximation error of PCR is small when the projection of the best linear predictor $\boldsymbol{\theta}^*$ on the subspace spanned by the population eigenvectors $\boldsymbol{v}_{K+1}, \ldots, \boldsymbol{v}_p$ is small. Differently put, if the contribution of the idiosyncratic component vector $\boldsymbol{u}_t$ is negligible then PCR has a negligible approximation error. Clearly, when $\boldsymbol{\gamma}^* = \mathbf{0}$ the best linear predictor based on $\boldsymbol{X}_t$ coincides with the best linear predictor based on $\boldsymbol{P}_t$ and the approximation error is zero.

Second, it is interesting to provide some comments on the behaviour of the estimation error. Under the rate conditions of A.4 the estimation error is asymptotically negligible as $T \to \infty$. We remark that the condition $p/(p^\alpha T) \to 0$ as $T \to \infty$ is also required by Bai and Ng (2023) in the analysis of approximate weak factor models. The estimation error is also influenced by the degree of persistence in the data, as measured by r_α, and in particular the more dependent the data are the slower the convergence of the estimation error to zero. We remark that the results of Bai and Ng (2023), among others, do not depend on the degree of persistence of the data. This is due to the fact that their analysis relies on higher level conditions that imply sharp rates of convergence of certain key estimators in the analysis. It is also interesting to note that the estimation error is made up of three terms that can be easily associated with the different estimation problems embedded in PCR. The first two terms capture the estimation error of the principal components whereas the third one can be interpreted as the estimation error the principal component regression if the population principal components were observed. Last, an important question concerning the estimation error is whether the rate obtained for it in the theorem is optimal. We provide insights on this question below in Section 3.1.

3.1 Optimal Performance

A natural question that arises upon inspection of Theorem 1 is whether the learning rate for the estimation error is, in some appropriate sense, optimal. We compare the learning rate established by the theorem with the optimal learning rate that could be achieved if the population principal components were observed. In this case it is well known that the optimal rate is of the order K/T (Tsybakov, 2003), which is achieved by the least squares estimator based on the population principal components. In this section we provide conditions under which the optimal rate is achieved up to a logarithmic factor. We call this the near-optimal rate. For ease of exposition we assume that the approximation error is zero throughout this section.

For a given choice of the signal strength α, degree of dependence r_α and the rate of growth of the number of principal components r_K, it is straightforward to verify that the near-optimal rate can be recovered in Theorem 1 provided that

$$\frac{1 - r_K}{2\alpha - 1} < \frac{1 + r_K}{2 - 2\alpha + 2/r_\alpha} \ ,$$

and that the rate of growth of the number of parameters allowed is such a scenario is

$$r_p \in \left[\frac{1 - r_K}{2\alpha - 1}, \frac{1 + r_K}{2 - 2\alpha + 2/r_\alpha} \right] \ .$$

In particular we note that the near-optimal rate can be achieved in the both the strong signal ($\alpha = 1$), and the weak signal ($\alpha < 1$) cases, provided that $\alpha > 2/3$. The larger the values of α, r_α and r_K, the larger the range of admissible growth rates for the number of predictors r_p.

It is interesting to provide concrete examples of these conditions. In the presence of a strong signal ($\alpha = 1$), a fixed number of principal components ($r_K = 0$) and independent data ($r_\alpha = \infty$)[2] we obtain that

$$R(\hat{\boldsymbol{\theta}}) - R(\boldsymbol{\theta}^*) \leq C \left(\frac{1}{p} + \frac{1}{T^2} + \frac{K}{T} \right) \log(T) = O\left(\frac{K}{T} \log(T) \right) ,$$

when $r_p \in [1, \infty)$. In the presence of a weak signal ($\alpha < 1$), a fixed number of principal components ($r_K = 0$) and independent data ($r_\alpha = \infty$) we obtain that

$$R(\hat{\boldsymbol{\theta}}_{PCR}) - R(\boldsymbol{\theta}^*) = O\left(\frac{1}{p^{2\alpha - 1}} + \frac{p^{2-2\alpha}}{T^2} + \frac{K}{T} \right) \log(T) = O\left(\frac{K}{T} \log(T) \right) ,$$

[2]The proofs of this manuscript assume that r_α is finite.

when $r_p \in [1/(2\alpha - 1), 1/(2-2\alpha)]$. In particular when $\alpha = 3/4$ the rate of growth of the number of parameters required to achieve near-optimal performance is $r_p = 2$.

4 Proof of Main Result

We introduce some additional notation. In this section and in the appendices we use the subscript s to denote the index of an observation in the training sample $\mathcal{D}$ and the subscript t to denote the index of a validation observation. Also, for some random variable X we use $\|X\|_{L_2}$ to denote $\sqrt{\mathbb{E}(X^2|\mathcal{D})}$.

The proof of Theorem 1 begins with a decomposition of an upper bound of the excess risk of the empirical risk minimizer. Define the approximate rotation matrix $\boldsymbol{H} = \widehat{\boldsymbol{\Lambda}}_K^{-1/2} \widehat{\boldsymbol{V}}_K' \boldsymbol{V}_K \boldsymbol{\Lambda}_K^{1/2}$ (Bai and Ng, 2002) and the infeasible PCR estimator $\tilde{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^K} \frac{1}{T} \sum_{t=1}^T (Y_t - \boldsymbol{P}_t' \boldsymbol{\vartheta})^2$. The following lemma provides a useful bound for the excess risk. We remark that the lemma only requires stationarity and finite variance.

Lemma 2. *Suppose that $\{(Y_t, \boldsymbol{X}_t')'\}$ is a stationary sequence taking values in $\mathcal{Y} \times \mathcal{X} \in \mathbb{R} \times \mathbb{R}^p$ such that $\mathbb{E}(Y_t^2) < \infty$ and $\mathbb{E}(X_{it}^2) < \infty$ for all $i = 1, \ldots, p$.*

Then it holds that

$$R(\hat{\boldsymbol{\theta}}_{PCR}) - R(\boldsymbol{\theta}^*) \leq 2 \max_{1 \leq s \leq T} \{Y_s^2\} \mathbb{E}(\|\widehat{\boldsymbol{P}}_t - \boldsymbol{H}\boldsymbol{P}_t\|_2^2 | \mathcal{D}) + 4\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2 + 4\|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2^2 + 2\|\boldsymbol{u}_t' \boldsymbol{\gamma}^*\|_{L_2}^2 . \tag{5}$$

Lemma 2 implies that in order to control the excess risk of PCR it suffices to provide appropriate bounds on the four terms on the right hand side of (5). The following four propositions establish such bounds.

The first two propositions control the terms $\mathbb{E}(\|\widehat{\boldsymbol{P}}_t - \boldsymbol{H}\boldsymbol{P}_t\|_2^2 | \mathcal{D})$ and $\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2$. We remark that these two terms capture the risk of PCR that is due to the estimation of the principal components. The proof of the propositions is based on standard arguments from the approximate factor model literature (Bai and Ng, 2002; Fan *et al.*, 2011).

Proposition 1. *Suppose A.1–A.4 are satisfied.*

Then for any $\eta > 0$ and any T sufficiently large,

$$\mathbb{E}(\|\widehat{\boldsymbol{P}}_t - \boldsymbol{H}\boldsymbol{P}_t\|_2^2|\mathcal{D}) \leq \frac{c_{K+1}}{2c_K^2}\frac{1}{p^{2\alpha-1}},$$

holds with probability at least $1 - T^{-\eta}$.

Proposition 2. *Suppose A.1–A.4 are satisfied.*

Then for any $\eta > 0$ there is $C > 0$ such that, for any T sufficiently large,

$$\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2 \leq C\left\{\frac{K}{T} + \frac{\log(T)}{T} + \left[\frac{(p + \log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{p^\alpha T}\right]^2 + \frac{1}{p^{2\alpha}}\right\}\log(T),$$

holds with probability at least $1 - T^{-\eta}$.

The next proposition controls the term $\|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2^2$. We remark that this term may be interpreted as the risk of the least squares estimator of PCR based on the population principal components. The proof of the proposition is based on the so-called small-ball method, which is the same proof strategy used in Brownlees and Guðmundsson (2025).

Proposition 3. *Suppose A.1–A.5 are satisfied.*

Then for any $\eta > 0$ there is a $C > 0$ such that, for any T sufficiently large,

$$\|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2^2 \leq C\frac{K\log(T)}{T},$$

holds with probability at least $1 - T^{-\eta}$.

The following proposition controls the error $\|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2$, which can be interpreted as the approximation error of PCR.

Proposition 4. *Suppose A.1 and A.3 are satisfied.*

Then it holds that

$$\|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 = (\boldsymbol{\theta}^*)'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{\theta}^* \ .$$

The claim of Theorem 1 follows from Propositions 1 to 4 together with the implication rule, the union bound and the fact that the sub-Gaussian assumption on Y_t (A.1) implies that for each $\eta > 0$ there is a $C > 0$ such that

$$\mathbb{P}\left(\max_{1\le s\le T}\{Y_s^2\} \le C\log(T)\right) \le 1 - T^{-\eta} \ .$$

5 Conclusions

This paper establishes predictive performance guarantees for principal component regression. Our analysis has a number of highlights. First, the analysis we carry out is nonparametric, in the sense that the relation between the prediction target and the predictors is not specified and, in particular, we do not assume that the prediction target is generated by a factor model. Second, our framework considers both the cases in which the largest eigenvalues of the covariance matrix of the predictors diverge linearly in the number of predictors (strong signal regime) or sublinearly (weak signal regime). A highlight of our results is that we show that, under appropriate conditions, PCR achieves optimal performance (up to a logarithmic factor) in both the strong signal and weak signal regimes.

A Proofs

We recall that $\mathbf{\Sigma} = (\boldsymbol{V}_K\mathbf{\Lambda}_K\boldsymbol{V}_K'+\boldsymbol{V}_R\mathbf{\Lambda}_R\boldsymbol{V}_R')$ where $\boldsymbol{V}_K = (\boldsymbol{v}_1,\ldots,\boldsymbol{v}_K)$, $\boldsymbol{V}_R = (\boldsymbol{v}_{K+1},\ldots,\boldsymbol{v}_p)$, $\mathbf{\Lambda}_K = \operatorname{diag}(\lambda_1,\ldots,\lambda_K)$ and $\mathbf{\Lambda}_R = \operatorname{diag}(\lambda_{K+1},\ldots,\lambda_p)$. Analogously, we have that $\widehat{\mathbf{\Sigma}} = (\widehat{\boldsymbol{V}}_K\widehat{\mathbf{\Lambda}}_K\widehat{\boldsymbol{V}}_K'+\widehat{\boldsymbol{V}}_R\widehat{\mathbf{\Lambda}}_R\widehat{\boldsymbol{V}}_R')$ where $\widehat{\boldsymbol{V}}_K = (\hat{\boldsymbol{v}}_1,\ldots,\hat{\boldsymbol{v}}_K)$, $\widehat{\boldsymbol{V}}_R = (\hat{\boldsymbol{v}}_{K+1},\ldots,\hat{\boldsymbol{v}}_p)$, $\widehat{\mathbf{\Lambda}}_K = \operatorname{diag}(\hat\lambda_1,\ldots,\hat\lambda_K)$ and $\widehat{\mathbf{\Lambda}}_R = \operatorname{diag}(\hat\lambda_{K+1},\ldots,\hat\lambda_p)$.

Proof of Lemma 1. (i) We have that $\mathbf{B}\boldsymbol{P}_t + \boldsymbol{u}_t = \boldsymbol{V}_K\boldsymbol{V}_K'\boldsymbol{X}_t + (\boldsymbol{I}_p - \boldsymbol{V}_K\boldsymbol{V}_K')\boldsymbol{X}_t = \boldsymbol{X}_t$. We have that $\mathbf{B}'\mathbf{B} = \mathbf{\Lambda}_K^{1/2}\boldsymbol{V}_K'\boldsymbol{V}_K\mathbf{\Lambda}_K^{1/2} = \mathbf{\Lambda}_K$, that $\mathbb{E}(\boldsymbol{P}_t\boldsymbol{P}_t') = \mathbf{\Lambda}_K^{-1/2}\boldsymbol{V}_K'(\boldsymbol{V}_K\mathbf{\Lambda}_K\boldsymbol{V}_K' + \boldsymbol{V}_R\mathbf{\Lambda}_R\boldsymbol{V}_R')\boldsymbol{V}_K\mathbf{\Lambda}_K^{-1/2} = \boldsymbol{I}_K$, $\mathbb{E}(\boldsymbol{u}_t\boldsymbol{u}_t') = \boldsymbol{V}_R\boldsymbol{V}_R'(\boldsymbol{V}_K\mathbf{\Lambda}_K\boldsymbol{V}_K' + \boldsymbol{V}_R\mathbf{\Lambda}_R\boldsymbol{V}_R')\boldsymbol{V}_R\boldsymbol{V}_R' = \boldsymbol{V}_R\mathbf{\Lambda}_R\boldsymbol{V}_R'$, and that $\mathbb{E}(\boldsymbol{P}_t\boldsymbol{u}_t') = \mathbf{\Lambda}_K^{-1/2}\boldsymbol{V}_K'\left(\boldsymbol{V}_K\mathbf{\Lambda}_K\boldsymbol{V}_K' + \boldsymbol{V}_R\mathbf{\Lambda}_R\boldsymbol{V}_R'\right)\boldsymbol{V}_R\boldsymbol{V}_R' = \mathbf{0}_{K\times p}$. (ii) Note that

$\boldsymbol{X}_t'\boldsymbol{\theta}^* = \boldsymbol{P}_t'\mathrm{B}'\boldsymbol{\theta}^* + \boldsymbol{u}_t'\boldsymbol{\theta}^*$ and that $\boldsymbol{P}_t'\mathrm{B}'\boldsymbol{\theta}^* = \boldsymbol{P}_t'\boldsymbol{\Lambda}_K^{1/2}\boldsymbol{V}_K'\boldsymbol{\theta}^* = \boldsymbol{P}_t'\boldsymbol{\vartheta}^*$ and $\boldsymbol{u}_t'\boldsymbol{\theta}^* = \boldsymbol{u}_t'(\boldsymbol{I}_p - \boldsymbol{V}_K(\boldsymbol{V}_K'\boldsymbol{V}_K)^{-1}\boldsymbol{V}_K')\boldsymbol{\theta}^* = \boldsymbol{u}_t'\boldsymbol{\gamma}^*$ since $\boldsymbol{I}_p - \boldsymbol{V}_K(\boldsymbol{V}_K'\boldsymbol{V}_K)^{-1}\boldsymbol{V}_K'$ is idempotent. □

Proof of Lemma 2. Consider the excess risk decomposition given by

$$
\begin{aligned}
R(\hat{\boldsymbol{\theta}}_{PCR}) - R(\boldsymbol{\theta}^*) &= \|Y_t - \widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 \\
&\quad + \|Y_t - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}}\|_{L_2}^2 \\
&\quad + \|Y_t - \boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 \\
&\quad + \|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 - \|Y_t - \boldsymbol{X}_t'\boldsymbol{\theta}^*\|_{L_2}^2 \ .
\end{aligned}
$$

The projection theorem, the fact $(a+b)^2 \leq 2a^2 + 2b^2$, the fact $\mathbb{E}(\boldsymbol{P}_t\boldsymbol{u}_t') = \mathbf{0}$ and the Cauchy-Schwarz inequality imply

$$
\begin{aligned}
&\|Y_t - \widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 \\
&\quad = \|Y_t - \boldsymbol{X}_t'\boldsymbol{\theta}^* + \boldsymbol{X}_t'\boldsymbol{\theta}^* - \widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{X}_t'\boldsymbol{\theta}^* + \boldsymbol{X}_t'\boldsymbol{\theta}^* - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 \\
&\quad = \|\boldsymbol{X}_t'\boldsymbol{\theta}^* - \widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|\boldsymbol{X}_t'\boldsymbol{\theta}^* - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 \\
&\quad \leq 2\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}} + \boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 + 2\|\boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 \\
&\quad = \|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 + 2\|\boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 + \|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 \\
&\quad \leq 2\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}}\|_{L_2}^2 + 2\|\boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}} - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 + 2\|\boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 + \|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 \\
&\quad \leq 2\|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2^2 + 2\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2 + 2\|\hat{\boldsymbol{\vartheta}}\|_2^2\mathbb{E}(\|\widehat{\boldsymbol{P}}_t - \boldsymbol{H}\boldsymbol{P}_t\|_2^2|\mathcal{D}) + \|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 \ . \qquad (6)
\end{aligned}
$$

To see how the projection theorem applies to the second equality note that $\widehat{\boldsymbol{P}}_t'\hat{\boldsymbol{\vartheta}} = \boldsymbol{X}_t'\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\hat{\boldsymbol{\vartheta}}$ and $\boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}} = \boldsymbol{X}_t'\boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-1/2}\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}$. Furthermore, we have that

$$
\|\hat{\boldsymbol{\vartheta}}\|_2 = \left\|\frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{Y}\right\|_2 \leq \left\|\sqrt{\frac{1}{T}}\widehat{\boldsymbol{P}}'\right\|_2 \left\|\sqrt{\frac{1}{T}}\boldsymbol{Y}\right\|_2 = \sqrt{\frac{1}{T}}\,\|\boldsymbol{Y}\|_2 \leq \max_{1\leq s\leq T}|Y_s| \ . \qquad (7)
$$

Next, the projection theorem, the Cauchy-Schwarz inequality and the inequality $2ab \leq$

$a^2 + b^2$ imply that

$$
\begin{aligned}
&\|Y_t - \boldsymbol{P}_t'\boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}}\|_{L_2}^2 \\
&\quad = \|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^* + \boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}})\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^* + \boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}})\|_{L_2}^2 \\
&\quad = \|\boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}})\|_{L_2}^2 - \|\boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}})\|_{L_2}^2 \\
&\quad = \mathbb{E}\left[[\boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}) - \boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}})][\boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}) + \boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}})] \middle| \mathcal{D} \right] \\
&\quad = \mathbb{E}\left[[\boldsymbol{P}_t'(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}})][\boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}} + \boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}})] \middle| \mathcal{D} \right] \\
&\quad = \mathbb{E}\left[[\boldsymbol{P}_t'(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}})][2\boldsymbol{P}_t'(\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}) + \boldsymbol{P}_t'(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}})] \middle| \mathcal{D} \right] \\
&\quad = \|\boldsymbol{P}_t'(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}})\|_{L_2}^2 + 2\mathbb{E}\left[(\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}})'\boldsymbol{P}_t\boldsymbol{P}_t'(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}) \middle| \mathcal{D} \right] \\
&\quad = \|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2 + 2(\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}})'(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}) \\
&\quad \leq \|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2 + 2\|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2 \leq 2\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2 + \|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2^2 \; . \qquad (8)
\end{aligned}
$$

The projection theorem implies

$$
\|Y_t - \boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}}\|_{L_2}^2 - \|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 = \|\boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 = \|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*\|_2^2, \qquad (9)
$$

$$
\|Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 - \|Y_t - \boldsymbol{X}_t'\boldsymbol{\theta}^*\|_{L_2}^2 = \|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 \; . \qquad (10)
$$

The claim then follows from (6), (7), (8), (9) and (10). □

A.1 Proof of Proposition 1

Proof of Proposition 1. We begin by noting that

$$
\begin{aligned}
\widehat{\boldsymbol{P}}_t - \boldsymbol{H}\boldsymbol{P}_t &= \widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'\boldsymbol{X}_t - \widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'\boldsymbol{V}_K\boldsymbol{\Lambda}_K^{1/2}\boldsymbol{\Lambda}_K^{-1/2}\boldsymbol{V}_K'\boldsymbol{X}_t = \widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'(\boldsymbol{I}_p - \boldsymbol{V}_K\boldsymbol{V}_K')\boldsymbol{X}_t \\
&= \widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t = \widehat{\boldsymbol{\Lambda}}_K^{-1}\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K\widehat{\boldsymbol{V}}_K'\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t \\
&= \widehat{\boldsymbol{\Lambda}}_K^{-1}\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'(\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K\widehat{\boldsymbol{V}}_K' + \widehat{\boldsymbol{V}}_R\widehat{\boldsymbol{\Lambda}}_R\widehat{\boldsymbol{V}}_R')\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t \\
&= \widehat{\boldsymbol{\Lambda}}_K^{-1}\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'\frac{1}{T}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t = \widehat{\boldsymbol{\Lambda}}_K^{-1}\frac{1}{\sqrt{T}}\widehat{\boldsymbol{P}}'\frac{1}{\sqrt{T}}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t \; .
\end{aligned}
$$

This implies that for any $\varepsilon > 0$

$$
\begin{aligned}
&\mathbb{P}\left(\mathbb{E}\left[\|\widehat{\boldsymbol{P}}_t - \boldsymbol{H}\boldsymbol{P}_t\|_2^2 \middle| \mathcal{D}\right] \geq \varepsilon\right) \leq \mathbb{P}\left(\mathbb{E}\left[\left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2^2 \left\|\frac{1}{\sqrt{T}}\widehat{\boldsymbol{P}}'\right\|_2^2 \left\|\frac{1}{\sqrt{T}}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t\right\|_2^2 \middle| \mathcal{D}\right] \geq \varepsilon\right) \\
&\quad = \mathbb{P}\left(\left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2^2 \mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t\right\|_2^2 \middle| \mathcal{D}\right] \geq \varepsilon\right) \\
&\quad \leq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2^2 \geq \varepsilon_1\right) + \mathbb{P}\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t\right\|_2^2 \middle| \mathcal{D}\right] \geq \varepsilon_2\right) \ , \qquad (11)
\end{aligned}
$$

for some $\varepsilon_1, \varepsilon_2 > 0$ such that $\varepsilon = \varepsilon_1\varepsilon_2$. We proceed by establishing bounds for the two terms in (11). First, we note that it follows from Proposition A.1 that for any $\eta > 0$, for all T sufficiently large, it holds that

$$
\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2^2 \geq \frac{4}{c_K^2}\frac{1}{p^{2\alpha}}\right) = O\left(\frac{1}{T^\eta}\right) \ . \qquad (12)
$$

Second, we establish a bound for the second term of equation (11). We have that

$$
\begin{aligned}
&\mathbb{P}\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t\right\|_2^2 \middle| \mathcal{D}\right] \geq \varepsilon_2\right) = \mathbb{P}\left(\frac{1}{T}\mathbb{E}\left[\sum_{s=1}^T (\boldsymbol{X}_s'\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t)^2 \middle| \mathcal{D}\right] \geq \varepsilon_2\right) \\
&\quad = \mathbb{P}\left(\frac{1}{T}\mathbb{E}\left[\sum_{s=1}^T (\boldsymbol{X}_s'\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t\boldsymbol{X}_t'\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_s) \middle| \mathcal{D}\right] \geq \varepsilon_2\right) = \mathbb{P}\left(\frac{1}{T}\sum_{s=1}^T \boldsymbol{X}_s'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{X}_s \geq \varepsilon_2\right) \\
&\quad = \mathbb{P}\left(p\left(\frac{1}{T}\sum_{s=1}^T \frac{\boldsymbol{X}_s'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{X}_s}{p} - \frac{\mathrm{tr}(\boldsymbol{\Lambda}_R^2)}{p}\right) + \mathrm{tr}(\boldsymbol{\Lambda}_R^2) \geq \varepsilon_2\right) \\
&\quad \leq \mathbb{P}\left(p\left(\frac{1}{T}\sum_{s=1}^T \frac{\boldsymbol{X}_s'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{X}_s}{p} - \frac{\mathrm{tr}(\boldsymbol{\Lambda}_R^2)}{p}\right) + c_{K+1}^2 p \geq \varepsilon_2\right)
\end{aligned}
$$

where we have used the fact that $\mathbb{E}(\boldsymbol{X}_s'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{X}_s) = \mathrm{tr}(\boldsymbol{\Lambda}_R^{1/2}\boldsymbol{V}_R'\mathbb{E}(\boldsymbol{X}_s\boldsymbol{X}_s')\boldsymbol{V}_R\boldsymbol{\Lambda}_R^{1/2}) = \mathrm{tr}(\boldsymbol{\Lambda}_R^2) < c_{K+1}^2 p$. It is straightforward to verify that the sequence $\{\boldsymbol{X}_s'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{X}_s/p - \mathrm{tr}(\boldsymbol{\Lambda}_R^2)/p\}$ satisfies the conditions of Lemma B.2, implying that for each $\eta > 0$ there exists a constant C such that, for all T sufficiently large, it holds that

$$
\mathbb{P}\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t\right\|_2^2 \middle| \mathcal{D}\right] \geq c_{K+1}^2 p + Cp\sqrt{\frac{\log(T)}{T}}\right) = O\left(\frac{1}{T^\eta}\right) \ . \qquad (13)
$$

The claim of the proposition then follows from inequalities (12) and (13). $\square$

Proposition A.1. *Suppose A.1–A.4 are satisfied.*

Then for any $\eta > 0$, for any T sufficiently large, we have that $\mathbb{P}(\hat{\lambda}_K \geq c_K p^\alpha/2) \geq 1 - T^{-\eta}$.

Proof. A.3 implies $\hat{\lambda}_K \geq \lambda_K - |\lambda_K - \hat{\lambda}_K| \geq c_K p^\alpha - \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$ where the last inequality follows from Weyl's inequality. To establish the claim, it suffices to show that for all T sufficiently large,

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2 \leq \frac{c_K}{2} p^\alpha \tag{14}$$

holds with probability at least $1 - T^{-\eta}$. Using the representation of Lemma 1 we get

$$\begin{aligned}
\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} &= \frac{1}{T}\sum_{t=1}^{T}(\mathbf{B}\boldsymbol{P}_t + \boldsymbol{u}_t)(\mathbf{B}\boldsymbol{P}_t + \boldsymbol{u}_t)' - \mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}_t') \\
&= \frac{1}{T}\mathbf{B}\sum_{t=1}^{T}(\boldsymbol{P}_t\boldsymbol{P}_t' - \boldsymbol{I}_K)\mathbf{B}' + \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{u}_t\boldsymbol{u}_t' - \mathbb{E}\left(\boldsymbol{u}_t\boldsymbol{u}_t'\right)) \\
&+ \frac{1}{T}\sum_{t=1}^{T}(\mathbf{B}\boldsymbol{P}_t\boldsymbol{u}_t') + \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{u}_t\boldsymbol{P}_t'\mathbf{B}') = D_1 + D_2 + D_3 + D_4 \ .
\end{aligned}$$

This, in turn, implies that

$$\mathbb{P}\left(\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2 > \varepsilon\right) \leq \sum_{i=1}^{4}\mathbb{P}\left(\|D_i\|_2 > \varepsilon_i\right) \ , \tag{15}$$

for some $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 > 0$ such that $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$. Proposition A.3 and A.3 imply that for any $\eta > 0$ there is a positive constant C_1 such that

$$\|D_1\|_2 > \left\|\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{P}_t\boldsymbol{P}_t' - \boldsymbol{I}_K)\right\|_2 \|\mathbf{B}\mathbf{B}'\|_2 \geq p^\alpha C_1\left(\sqrt{\frac{K}{T}} + \sqrt{\frac{\log T}{T}}\right) = \varepsilon_1 \tag{16}$$

holds with probability at most $T^{-\eta}$. Proposition A.2 implies that for any $\eta > 0$ there is a positive constant C_2 such that

$$\|D_2\|_2 > C_2\left[\frac{(p + \log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{T} + \sqrt{\frac{(p + \log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{T}}\right] = \varepsilon_2 \tag{17}$$

holds with probability at most T^{η}. Proposition A.4 and A.3 imply that for any $\eta > 0$ there is a positive constant C_3 such that

$$\|D_3\|_2 > \left\| \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{P}_t \boldsymbol{u}_t' \right\|_2 \|\mathbf{B}\|_2 \geq p^{\alpha/2} C_3 \sqrt{\frac{pK \log T}{T}} = \varepsilon_3 \tag{18}$$

holds with probability at most $T^{-\eta}$. Last, note that $\|D_3\|_2 = \|D_4\|_2$. Combining the inequalities in (16), (17), (18) we get

$$\mathbb{P}\left(\left\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\|_2 < \varepsilon \right) \leq 1 - \frac{4}{T^{\eta}} ,$$

where ε may be defined as

$$\varepsilon = p^{\alpha} C_1 \left(\sqrt{\frac{K}{T}} + \sqrt{\frac{\log T}{T}} \right) + p^{\alpha} C_2 \left[\frac{1}{p^{\alpha}} \frac{(p + \log(T))^{\frac{r_\alpha + 1}{r_\alpha}}}{T} + \frac{1}{p^{\alpha}} \sqrt{\frac{(p + \log(T))^{\frac{r_\alpha + 1}{r_\alpha}}}{T}} \right]$$
$$+ 2p^{\alpha} C_3 \sqrt{\frac{p^{1-\alpha} K \log T}{T}}.$$

The claim follows after noting that A.4 implies that, for all T sufficiently large, $\varepsilon \leq \frac{c_K}{2} p^{\alpha}$. □

Proposition A.2. *Suppose A.1–A.4 are satisfied. Then for any $\eta > 0$ there exists a positive constant C such that, for any T sufficiently large,*

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}_t \boldsymbol{u}_t' - \mathbb{E}\left(\boldsymbol{u}_t \boldsymbol{u}_t' \right) \right\|_2 \geq C \left[\frac{(p + \log(T))^{\frac{r_\alpha + 1}{r_\alpha}}}{T} + \sqrt{\frac{(p + \log(T))^{\frac{r_\alpha + 1}{r_\alpha}}}{T}} \right]$$

holds with probability at most $T^{-\eta}$.

Proof. We assume that c_{K+1} is positive (when c_{K+1} is zero the claim is trivial). Consider the isotropic random vectors $\boldsymbol{\Sigma}_u^{-1/2} \boldsymbol{u}_t$ where $\boldsymbol{\Sigma}_u = \mathbb{E}\left(\boldsymbol{u}_t \boldsymbol{u}_t' \right) = \boldsymbol{V}_R \boldsymbol{V}_R' \boldsymbol{\Sigma} \boldsymbol{V}_R \boldsymbol{V}_R'$ and note that Vershynin (2012, Lemma 5.4) implies that if $\mathcal{N}$ is a $\frac{1}{4}$-net of $\mathcal{S}^{p-1}$ then

$$\begin{aligned}
&\left\| \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{u}_t\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2} - \boldsymbol{I}_p \right\|_2 = \max_{\boldsymbol{x}\in\mathcal{S}^{p-1}} \left| \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x})^2 - 1 \right| \\
&\leq 2\max_{\boldsymbol{x}\in\mathcal{N}} \left| \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x})^2 - 1 \right| = 2\max_{\boldsymbol{x}\in\mathcal{N}} \left| \frac{1}{T}\sum_{t=1}^{T} W_{\boldsymbol{x}\,t} \right| , \qquad (19)
\end{aligned}$$

where $W_{\boldsymbol{x}\,t} = (\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x})^2 - 1$. To establish the claim we first show that for all T sufficiently large (19) can be bounded with high probability. We begin by establishing a bound for a fixed vector $\boldsymbol{x} \in \mathcal{N}$. We enumerate three properties of the sequence $\{W_{\boldsymbol{x}\,t}\}$. First, $\mathbb{E}(W_{\boldsymbol{x}\,t}) = 0$ since $\mathbb{E}[(\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x})^2] = \|\boldsymbol{x}\|_2^2 = 1$. Second, $W_{\boldsymbol{x}\,t}$ is the de-meaned square of a sub-Gaussian random variable with parameter C_m (which is independent of $\boldsymbol{x}$). To see this note that $\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x} = \boldsymbol{Z}_t'\boldsymbol{V}_R\boldsymbol{\Lambda}_R^{1/2}\boldsymbol{V}_R'\boldsymbol{V}_R\boldsymbol{\Lambda}_R^{-1/2}\boldsymbol{V}_R'\boldsymbol{x} = \boldsymbol{Z}_t'\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{x}$ where $\boldsymbol{Z}_t = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_t$. Define $\rho = \|\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{x}\|_2$ and notice that $\rho \in [0,1]$. If $\rho \in (0,1]$ then for any $\varepsilon > 0$ it holds

$$\mathbb{P}(|\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x}| \geq \varepsilon) = \mathbb{P}\left(\left| \frac{\boldsymbol{Z}_t'\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{x}}{\rho} \right| \geq \frac{\varepsilon}{\rho} \right) \leq \exp(-C_m(\varepsilon/\rho)^2) \leq \exp(-C_m\varepsilon^2)$$

where $C_m > 0$ is defined in A.1. If $\rho = 0$ then for any $\varepsilon > 0$ it holds $\mathbb{P}(|\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x}| \geq \varepsilon) = 0 \leq \exp(-C_m\varepsilon^2)$. Therefore $\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{x}$ is sub-Gaussian with parameter C_m. Third, the sequence $\{W_{\boldsymbol{x}\,t}\}_{t=1}^T$ inherits the mixing properties of the sequence $\{(Y_t, \boldsymbol{X}_t')'\}_{t=1}^T$ spelled out in A.2. These three facts imply that $\{W_{\boldsymbol{x}\,t}\}$ satisfies the conditions of Bosq (1998, Theorem 1.4) (the C_r inequality and Boucheron, Lugosi, and Massart (2013, Theorem 2.1) imply that the condition spelled out in equation (1.33) of Bosq (1998) is satisfied). Define $\varepsilon_T = \frac{C_1^*(p+\log(T))^{\frac{r\alpha+1}{r\alpha}}}{T} + \left[\frac{C_1^*(p+\log(T))^{\frac{r\alpha+1}{r\alpha}}}{T}\right]^{1/2}$ and $q_T = \left\lceil \frac{T}{C_2^*(p+\log(T))^{\frac{1}{r\alpha}}} \right\rceil - 1$ for positive constants C_1^* and C_2^* to be chosen below. Notice that A.4 implies that for all T sufficiently large it holds that $q_T \in [1, \frac{T}{2}]$, as required by the theorem. Then for any $r \geq 3$ there exist a positive constant C_1 that depends on C_m and r such that, for all T

sufficiently large,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} W_{\boldsymbol{x}\,t}\right| > \varepsilon_T\right) \leq a_{1\,T}\exp\left(-\frac{q_T\varepsilon_T^2}{C_1 + C_1\varepsilon_T}\right) + a_{2\,T}\alpha\left(\left\lfloor\frac{T}{q_T+1}\right\rfloor\right)^{\frac{2r}{2r+1}} \tag{20}$$

holds, where $a_{1\,T} = 2T/q_T + 2[1+\varepsilon_T^2/(C_1+C_1\varepsilon_T)]$, and $a_{2\,T} = 11T(1+C_1/\varepsilon_T)$. We proceed by bounding the r.h.s. of (20). First, for all T sufficiently large, we have

$$\begin{aligned} a_{1\,T}\exp\left(-\frac{q_T\varepsilon_T^2}{C_1+C_1\varepsilon_T}\right) &\leq \left(2T+2+2\frac{\varepsilon_T}{C_1}\right)\exp\left(-\frac{\min(\varepsilon_T,\varepsilon_T^2)}{2C_1}q_T\right) \\ &\leq \exp\left(\log\left(3T+2\frac{\varepsilon_T}{C_1}\right) - \frac{C_1^*(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{2TC_1}\left(\frac{T}{C_2^*(p+\log(T))^{\frac{1}{r_\alpha}}}-1\right)\right) \\ &\leq \exp\left(-\left(\frac{C_1^*}{2C_1C_2^*}-1\right)(p+\log T)\right)\ , \end{aligned} \tag{21}$$

where in the second inequality we use the fact that $\min(\varepsilon_T,\varepsilon_T^2) \geq C_1^*(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}/T$ and the last from the condition $r_p < r_\alpha$. Second, for all T sufficiently large, we have

$$\begin{aligned} a_{2\,T}\alpha\left(\left\lfloor\frac{T}{q_T+1}\right\rfloor\right)^{\frac{2r}{2r+1}} &\leq \exp\left(2\log(T) - C_\alpha\frac{2r}{2r+1}\left(\frac{T}{q_T+1}-1\right)^{r_\alpha}\right) \\ &\leq \exp\left(2\log(T) - \frac{1}{2}C_\alpha(C_2^*)^{r_\alpha}\frac{2r}{2r+1}(p+\log(T))\right) \\ &\leq \exp\left(-\left(C_\alpha(C_2^*)^{r_\alpha}\frac{r}{2r+1}-1\right)(p+\log(T))\right)\ . \end{aligned} \tag{22}$$

Combining (21) and (22) we get that for a given $\boldsymbol{x}$, for all T sufficiently large, it holds

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} W_{\boldsymbol{x}\,t}\right| > \varepsilon_T\right) &\leq \exp\left(-\left(\frac{C_1^*}{2C_1C_2^*}-1\right)(p+\log T)\right) \\ &+ \exp\left(-\left(C_\alpha(C_2^*)^{r_\alpha}\frac{r}{2r+1}-1\right)(p+\log(T))\right)\ . \end{aligned} \tag{23}$$

The next step consist in taking the union bound over all the vectors $\boldsymbol{x} \in \mathcal{N}$. It follows from Vershynin (2012, Lemma 5.2) that the cardinality of a $\frac{1}{4}$-net $\mathcal{N}$ of the unit sphere $\mathcal{S}^{p-1}$ is bounded by 9^p. Setting $C_2^* = \left(\frac{(\eta+\log(9)+1)(2r+1)}{rC_\alpha}\right)^{\frac{1}{r_\alpha}}$ and $C_1^* = 2C_1C_2^*(\eta+\log(9)+1)$

in (23) we then obtain that, for all T sufficiently large,

$$\mathbb{P}\left(\max_{\boldsymbol{x}\in\mathcal{N}}\left|\frac{1}{T}\sum_{t=1}^{T}W_{\boldsymbol{x}\,t}\right|>\varepsilon_T\right)\le 9^p\max_{\boldsymbol{x}\in\mathcal{N}}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}W_{\boldsymbol{x}\,t}\right|>\varepsilon_T\right)$$
$$\le 2\exp\left(\log(9)p-(\eta+\log(9))\left(p+\log(T)\right)\right)<2\exp(-\eta\log(T))=\frac{2}{T^\eta}\ .$$

The claim follows after noting that, for all T sufficiently large, it holds

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t'-\mathbb{E}\left(\boldsymbol{u}_t\boldsymbol{u}_t'\right)\right\|_2>2\|\boldsymbol{\Sigma}_u\|_2\varepsilon_T\right)=\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Sigma}_u^{-1/2}\boldsymbol{u}_t\boldsymbol{u}_t'\boldsymbol{\Sigma}_u^{-1/2}-\boldsymbol{I}_p\right\|_2>2\varepsilon_T\right)$$
$$\le\mathbb{P}\left(\max_{\boldsymbol{x}\in\mathcal{N}}\left|\frac{1}{T}\sum_{t=1}^{T}W_{\boldsymbol{x}\,t}\right|>\varepsilon_T\right)=O\left(\frac{1}{T^\eta}\right),$$

and noting that A.3 implies that $\|\boldsymbol{\Sigma}_u\|_2=c_{K+1}$. □

Let $\lambda_{\min}(\boldsymbol{M})$ denote the smallest eigenvalue of the square matrix $\boldsymbol{M}$.

Proposition A.3. *Suppose A.1, A.2 and A.4 are satisfied. Then for any $\eta>0$ there exists a positive constant C such that, for any T sufficiently large,*

$$(i)\ \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\mathbb{E}\left(\boldsymbol{P}_t\boldsymbol{P}_t'\right)\right\|_2\ge C\left(\sqrt{\frac{K}{T}}+\sqrt{\frac{\log T}{T}}\right)\right)=\frac{1}{T^\eta}\ , \tag{24}$$

$$(ii)\ \mathbb{P}\left(\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)\le\frac{1}{2}\right)=\frac{1}{T^\eta}\ , \tag{25}$$

$$(iii)\ \mathbb{P}\left(\left\|\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)^{-1}-\left(\mathbb{E}\left(\boldsymbol{P}_t\boldsymbol{P}_t'\right)\right)^{-1}\right\|_2\ge C\left(\sqrt{\frac{K}{T}}+\sqrt{\frac{\log T}{T}}\right)\right)=\frac{1}{T^\eta}\ . \tag{26}$$

Proof. (i) Following the same arguments of Proposition A.2 and denoting by $\mathcal{N}$ the $\frac{1}{4}$-net of the unit sphere $\mathcal{S}^{K-1}$ we get that

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\mathbb{E}\left(\boldsymbol{P}_t\boldsymbol{P}_t'\right)\right\|_2\le 2\max_{\boldsymbol{x}\in\mathcal{N}}\left|\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{P}_t'\boldsymbol{x})^2-1\right|=2\max_{\boldsymbol{x}\in\mathcal{N}}\left|\frac{1}{T}\sum_{t=1}^{T}W_{\boldsymbol{x}\,t}\right|\ , \tag{27}$$

where $W_{\boldsymbol{x}\,t}=(\boldsymbol{P}_t'\boldsymbol{x})^2-1$. To establish the claim, we first show that for all T sufficiently large (27) can be bounded with high probability. We note two properties of the sequence

$\{W_{\boldsymbol{x}\,t}\}$. First, $\mathbb{E}(W_{\boldsymbol{x}\,t}) = 0$ since $\mathbb{E}[(\boldsymbol{P}_t'\boldsymbol{x})^2] = \|\boldsymbol{x}\|_2^2 = 1$. Second, $W_{\boldsymbol{x}\,t}$ is sub-exponential with parameter C_m' (which is independent of $\boldsymbol{x}$ and depending on C_m). To see this note that $|\boldsymbol{P}_t'\boldsymbol{x}| = |\boldsymbol{Z}_t'\boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}'\boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-1/2}\boldsymbol{x}| = |\boldsymbol{Z}_t'\boldsymbol{V}_K\boldsymbol{x}|$ where $\boldsymbol{Z}_t = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_t$ and that $\|\boldsymbol{V}_K\boldsymbol{x}\|_2 \in [0,1]$. Consider the decomposition $\sum_{t=1}^T W_{\boldsymbol{x}\,t} = \sum_{t=1}^T W_{\boldsymbol{x}\,t}' + \sum_{t=1}^T W_{\boldsymbol{x}\,t}''$ where $W_{\boldsymbol{x}\,t}' = W_{\boldsymbol{x}\,t}\mathbb{1}\left(|W_{\boldsymbol{x}\,t}| \leq b_T\right) - \mathbb{E}\left(W_{\boldsymbol{x}\,t}\mathbb{1}\left(|W_{\boldsymbol{x}\,t}| \leq b_T\right)\right)$ and $W_{\boldsymbol{x}\,t}'' = W_{\boldsymbol{x}\,t}\mathbb{1}\left(|W_{\boldsymbol{x}\,t}| > b_T\right) - \mathbb{E}\left(W_{\boldsymbol{x}\,t}\mathbb{1}\left(|W_{\boldsymbol{x}\,t}| > b_T\right)\right)$. Then for any $\varepsilon > 0$ we have that

$$\mathbb{P}\left(\left|\sum_{t=1}^T W_{\boldsymbol{x}\,t}\right| > \varepsilon\right) \leq \mathbb{P}\left(\left|\sum_{t=1}^T W_{\boldsymbol{x}\,t}'\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\left|\sum_{t=1}^T W_{\boldsymbol{x}\,t}''\right| > \frac{\varepsilon}{2}\right) .$$

The sequence $\{W_{\boldsymbol{x}\,t}'\}_{t=1}^T$ is such that $\|W_{\boldsymbol{x}\,t}'\|_\infty < 2b_T$ and has the same mixing properties as $\{(Y_t, \boldsymbol{X}_t)\}_{t=1}^T$ spelled out in A.2. These two facts imply that $\{W_{\boldsymbol{x}\,t}'\}$ satisfies the conditions of Liebscher (1996, Theorem 2.1). Define $\varepsilon_T = C^*(\sqrt{KT} + \sqrt{T\log T})$, $b_T = \frac{2}{C_m'}[\log(2\|W_{\boldsymbol{x}\,t}\|_{L_2}\sqrt{T}/C^*) + C^*(K + \log T)]$ and $M_T = \lfloor b_T^{-1}T^{\frac{1}{2}}/(\sqrt{K} + \sqrt{\log T})\rfloor$ for positive constant C^* to be chosen below. For all T sufficiently large, we have that $M_T \in [1, T]$ (note A.4 implies $r_k < 1/3$) and $4(2b_T)M_T < \varepsilon_T$, as required by the theorem. Then, for all T sufficiently large,

$$\mathbb{P}\left(\left|\sum_{t=1}^T W_{\boldsymbol{x}\,t}'\right| > \varepsilon_T\right) < 4\exp\left(-\frac{\varepsilon_T^2}{64\frac{T}{M_T}D(T, M_T) + \frac{16}{3}b_T M_T \varepsilon_T}\right) + 4\frac{T}{M_T}\exp\left(-C_\alpha M_T^{r_\alpha}\right)$$

holds with $D(T, M_T) = \mathbb{E}\left[\left(\sum_{t=1}^{M_T} W_{\boldsymbol{x}\,t}'\right)^2\right]$. Define $\gamma(l) = |\operatorname{Cov}(W_{\boldsymbol{x}\,t}', W_{\boldsymbol{x}\,t+l}')|$ for $l = 0, \ldots, T-1$ and note that $D(T, M_T) \leq M_T \sum_{l=-T+1}^{T-1} \gamma(l)$. Define $C_{m\,4} = \|W_{\boldsymbol{x}\,t}\|_{L_4}^2$ (which is a constant depending only on C_m). Then, for $l = 0, \ldots$ it holds that

$$\gamma(l) \leq 12\alpha(l)^{\frac{1}{2}}\|W_{\boldsymbol{x}\,t}'\|_{L_4}^2 \leq 48\alpha(l)^{\frac{1}{2}}\|W_{\boldsymbol{x}\,t}\|_{L_4}^2 \leq 48\alpha(l)^{\frac{1}{2}}C_{m,4} ,$$

where the first inequality follows from Davydov's inequality (Bosq, 1998, Corollary 1.1) and the second follows from the fact that $\|W_{\boldsymbol{x}\,t}'\|_{L_4} \leq 2\|W_{\boldsymbol{x}\,t}\mathbb{1}\left(|W_{\boldsymbol{x}\,t}| \leq b_T\right)\|_{L_4} \leq 2\|W_t\|_{L_4}$. This implies that $D(T, M_T) < C_{\sigma^2}M_T$ where $C_{\sigma^2} = 96C_{m,4}\sum_{l=0}^\infty \alpha(l)^{\frac{1}{2}} \vee 1$. We then use the inequality $a^2 + b^2 \leq (a+b)^2$ for $a, b > 0$ to get that

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{t=1}^{T} W'_{\boldsymbol{x}\,t}\right| > \varepsilon_T\right) &\leq 4\exp\left(-\frac{C^{*2}(\sqrt{KT} + \sqrt{T\log(T)})^2}{64TC_{\sigma^2} + \frac{16}{3}b_T M_T \varepsilon_T}\right) + 4\frac{T}{M_T}\exp\left(-C_\alpha M_T^{r_\alpha}\right) \\
&\leq 4\exp\left(-\frac{C^{*2}KT + C^{*2}T\log(T)}{64TC_{\sigma^2} + \frac{16}{3}C^*T}\right) + 4\,T\exp\left(-C_\alpha M_T^{r_\alpha}\right) \\
&\leq 4\exp\left(-\frac{C^{*2}}{64C_{\sigma^2} + \frac{16}{3}C^*}\left(K + \log(T)\right)\right) + 4\exp\left(\log(T) - C_\alpha M_T^{r_\alpha}\right) \ .
\end{aligned}$$

Note that for all T sufficiently large, we have $M_T^{r_\alpha} > K + \log(T)$ since $r_K < \frac{1}{3+\frac{2}{r_\alpha}}$ as implied by A.4. We may write that

$$\mathbb{P}\left(\left|\sum_{t=1}^{T} W'_{\boldsymbol{x}\,t}\right| > \varepsilon_T\right) \leq 5\exp\left(-\frac{C^{*2}}{64C_{\sigma^2} + \frac{16}{3}C^*}\left(K + \log(T)\right)\right) \ . \tag{28}$$

The sequence $\{W''_{\boldsymbol{x}\,t}\}_{t=1}^{T}$ is such that, for all T large enough,

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{t=1}^{T} W''_{\boldsymbol{x}\,t}\right| > \varepsilon_T\right) &\leq \frac{1}{\varepsilon_T}\mathbb{E}\left|\sum_{t=1}^{T} W''_{\boldsymbol{x}\,t}\right| \leq \frac{T}{\varepsilon_T}\mathbb{E}|W''_{\boldsymbol{x}\,t}| \leq \frac{2T}{\varepsilon_T}\mathbb{E}\left|W_{\boldsymbol{x}\,t}\mathbb{1}\left(|W_{\boldsymbol{x}\,t}| > b_T\right)\right| \\
&\leq \frac{2T}{\varepsilon_T}\|W_{\boldsymbol{x}\,t}\|_{L_2}\|\mathbb{1}\left(|W_{\boldsymbol{x}\,t}| > b_T\right)\|_{L_2} = \frac{2T}{\varepsilon_T}\|W_{\boldsymbol{x}\,t}\|_{L_2}\mathbb{P}(|W_{\boldsymbol{x}\,t}| > b_T)^{\frac{1}{2}} \leq \frac{2T\|W_{\boldsymbol{x}\,t}\|_{L_2}}{\varepsilon_T}\exp\left(-\frac{C'_m b_T}{2}\right) \\
&= \frac{2T\|W_{\boldsymbol{x}\,t}\|_{L_2}}{C^*(\sqrt{KT} + \sqrt{T\log(T)})}\exp\left(-\frac{C'_m b_T}{2}\right) \leq \frac{2\|W_{\boldsymbol{x}\,t}\|_{L_2}\sqrt{T}}{C^*}\exp\left(-\frac{C'_m b_T}{2}\right) \\
&= \exp\left(\log\left(\frac{2\|W_{\boldsymbol{x}\,t}\|_{L_2}\sqrt{T}}{C^*}\right) - \frac{C'_m b_T}{2}\right) = \exp\left(-C^*(K + \log(T))\right) \ ,
\end{aligned} \tag{29}$$

where the first inequality follows from Markov's inequality. Combining (28) and (29) we have that, for all T sufficiently large,

$$\max_{x\in\mathcal{N}}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} W_{\boldsymbol{x}\,t}\right| > \frac{\varepsilon_T}{T}\right) \leq \exp\left(-C^*(K + \log(T))\right) + 5\exp\left(-\frac{C^{*2}}{64C_{\sigma^2} + \frac{16}{3}C^*}\left(K + \log(T)\right)\right) \ ,$$

for the fixed vector $\boldsymbol{x}$. It remains to take the union bound over all the vectors $\boldsymbol{x} \in \mathcal{N}$.

Using the same arguments of Proposition A.2 we have that, for all T sufficiently large,

$$\mathbb{P}\left(\max_{x\in\mathcal{N}}\left|\frac{1}{T}\sum_{t=1}^{T}W_{\boldsymbol{x}\,t}\right|>\frac{\varepsilon_T}{T}\right)\leq 9^K\max_{x\in\mathcal{N}}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}W_{\boldsymbol{x}\,t}\right|>\frac{\varepsilon_T}{T}\right)$$
$$\leq 9^K\exp\left(-C^*(K+\log(T))\right)+9^K\cdot 5\exp\left(-\frac{C^{*2}}{64C_{\sigma^2}+\frac{64}{3}C^*}\left(K+\log(T)\right)\right)\leq\frac{1}{T^\eta}$$

where the last inequality follows for a sufficiently large choice of the constant C^*. The claim follows after noting that, for all T sufficiently large, it holds

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\mathbb{E}\left(\boldsymbol{P}_t\boldsymbol{P}_t'\right)\right\|_2>2\left(\sqrt{\frac{K}{T}}+\sqrt{\frac{\log(T)}{T}}\right)\right)\leq\mathbb{P}\left(\max_{\boldsymbol{x}\in\mathcal{N}}\left|\frac{1}{T}\sum_{t=1}^{T}W_{\boldsymbol{x}\,t}\right|>\frac{\varepsilon_T}{T}\right)\leq\frac{1}{T^\eta}\,.$$

(ii) We begin by noting that

$$\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)=\min_{\boldsymbol{x}\in\mathcal{S}^{K-1}}\boldsymbol{x}'\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)\boldsymbol{x}$$
$$=\min_{\boldsymbol{x}\in\mathcal{S}^{K-1}}\boldsymbol{x}'\boldsymbol{I}_K\boldsymbol{x}-\boldsymbol{x}'\left(\boldsymbol{I}_K-\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)\boldsymbol{x}\geq 1-\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\boldsymbol{I}_K\right\|_2\,.$$

It then follows from part (i) that for any $\eta>0$, for all T sufficiently large, $1-\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\boldsymbol{I}_K\right\|_2\geq\frac{1}{2}$ holds with at least probability $1-T^{-\eta}$, which implies the claim.

(iii) First, we condition on the event of part (ii), which implies that for any $\eta>0$, for all sufficiently large T, $\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)^{-1}$ exists. Then we note that conditional on the same event for any $\eta>0$ there exists a positive constant C such that, for all sufficiently large

T, it holds that

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)^{-1} - \mathbb{E}\left(\boldsymbol{P}_t\boldsymbol{P}_t'\right)^{-1}\right\|_2 \geq \frac{C}{2}\left(\sqrt{\frac{K}{T}}+\sqrt{\frac{\log(T)}{T}}\right)\right) \\
&= \mathbb{P}\left(\left\|\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)^{-1}\left(\boldsymbol{I}_K - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)\right\|_2 \geq \frac{C}{2}\left(\sqrt{\frac{K}{T}}+\sqrt{\frac{\log(T)}{T}}\right)\right) \\
&\leq \mathbb{P}\left(\left\|\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'\right)^{-1}\right\|_2 \geq \frac{1}{2}\right) + \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t' - \mathbb{E}\left(\boldsymbol{P}_t\boldsymbol{P}_t'\right)\right\|_2 \geq C\left(\sqrt{\frac{K}{T}}+\sqrt{\frac{\log(T)}{T}}\right)\right) \\
&= \frac{2}{T^{\eta}} ,
\end{aligned}
$$

where we have used the fact that $\mathbb{E}(\boldsymbol{P}_t\boldsymbol{P}_t')^{-1} = \boldsymbol{I}_K$. The unconditional probability of this event being realized is $(2T^{-\eta})(1 - T^{-\eta}) = T^{-\eta}$. □

Proposition A.4. *Suppose A.1, A.2, A.3 and A.4 are satisfied. Then for any $\eta > 0$ there exits a positive constant C such that, for any T sufficiently large, it holds that*

$$
\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{u}_t'\right\|_2 \geq C\sqrt{\frac{pK\log(T)}{T}}\right) = \frac{1}{T^{\eta}} .
$$

Proof. We assume that c_{K+1} is positive (when c_{K+1} is zero the claim is trivial). Let $V_{ij,t}$ denote $P_{it}u_{jt}$ and $\boldsymbol{V}_{j,t}$ denote the j-th column of $\boldsymbol{P}_t\boldsymbol{u}_t'$ where $1 \leq i \leq K$ and $1 \leq j \leq p$. We begin by showing that, for each j, the sequence of K-dimensional random vectors $\{\boldsymbol{V}_{j,t}\}_{t=1}^{T}$ satisfies the conditions required in Lemma B.2. Lemma B.1.(ii) and (iii) establish that $\boldsymbol{P}_t$ and $\boldsymbol{u}_t$ are sub-Gaussian random vectors. Standard results on sub-Gaussian random variables imply that

$$
\mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}|\boldsymbol{V}_{j\,t}'\boldsymbol{v}| > \varepsilon\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}|\boldsymbol{P}_t'\boldsymbol{v}||u_{jt}| > \varepsilon\right) \leq \exp(-C_m'\varepsilon)
$$

for some $C_m' > 0$. Lemma 1 implies that $\mathbb{E}(\boldsymbol{V}_{j,t}) = \mathbb{E}(\boldsymbol{P}_t u_{jt}) = \boldsymbol{0}$, hence $\boldsymbol{V}_{j,t}$ is a zero-mean random vector. Moreover, standard properties of strong mixing processes imply that $\{\boldsymbol{V}_{j,t}\}_{t=1}^{T}$ inherits the mixing properties of the sequence $\{(Y_t, \boldsymbol{X}_t)'\}_{t=1}^{T}$ spelled out in A.2. Lastly, A.4 implies that $K = \lfloor C_K T^{r_K} \rfloor$ where $r_K < 1$. Then the conditions of

Lemma B.2 applied to the sequence $\{\boldsymbol{V}_{j,t}\}_{t=1}^T$ are satisfied and picking $\eta' = \eta + r_p$ we get that there exists a C such that, for all T sufficiently large,

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{P}_t\boldsymbol{u}_t'\right\|_2 \geq C\sqrt{\frac{pK\log(T)}{T}}\right) \leq \mathbb{P}\left(\sqrt{p}\max_{1\leq j\leq p}\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{V}_{j,t}\right\|_2 \geq C\sqrt{\frac{pK\log(T)}{T}}\right)$$
$$\leq\ p\max_{1\leq j\leq p}\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{Z}_{j,t}\right\|_2 \geq C\sqrt{\frac{K\log(T)}{T}}\right) = \frac{p}{T^{\eta'}} = C\frac{T^{r_p}}{T^{\eta+r_p}} = \frac{C}{T^{\eta}}\ .$$

□

A.2 Proof of Proposition 2

Proof of Proposition 2. Recall that $\tilde{\boldsymbol{\vartheta}} = \left(\frac{1}{T}\boldsymbol{P}'\boldsymbol{P}\right)^{-1}\frac{1}{T}\boldsymbol{P}'\boldsymbol{Y}$ and $\hat{\boldsymbol{\vartheta}} = \frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{Y}$ and note that

$$\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}} = \tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}^{-1}\boldsymbol{H}\boldsymbol{H}'\hat{\boldsymbol{\vartheta}} = \left(\frac{1}{T}\boldsymbol{P}'\boldsymbol{P}\right)^{-1}\frac{1}{T}\boldsymbol{P}'\boldsymbol{Y} - \boldsymbol{H}^{-1}\boldsymbol{H}\boldsymbol{H}'\frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{Y}$$
$$= \left[\left(\frac{1}{T}\boldsymbol{P}'\boldsymbol{P}\right)^{-1}\frac{1}{T}\boldsymbol{P}'\boldsymbol{Y} - \frac{1}{T}\boldsymbol{P}'\boldsymbol{Y}\right] - \boldsymbol{H}^{-1}\left(\boldsymbol{H}\boldsymbol{H}'\frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{Y} - \frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{Y} + \frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{Y} - \frac{1}{T}\boldsymbol{H}\boldsymbol{P}'\boldsymbol{Y}\right)$$
$$= \left[\left(\frac{1}{T}\boldsymbol{P}'\boldsymbol{P}\right)^{-1} - \boldsymbol{I}_K\right]\frac{1}{T}\boldsymbol{P}'\boldsymbol{Y} - \boldsymbol{H}^{-1}\left[(\boldsymbol{H}\boldsymbol{H}' - \boldsymbol{I}_K)\sqrt{\frac{1}{T}}\widehat{\boldsymbol{P}}' + \sqrt{\frac{1}{T}}\left(\widehat{\boldsymbol{P}}' - \boldsymbol{H}\boldsymbol{P}'\right)\right]\sqrt{\frac{1}{T}}\boldsymbol{Y}\ .$$

Then the triangle inequality and the implication rule imply that for any $\varepsilon > 0$ and $\varepsilon_i > 0$ for $i = 1, \ldots, 5$ such that $\varepsilon = \varepsilon_1\varepsilon_2 + \varepsilon_3\varepsilon_5 + \varepsilon_4\varepsilon_5$ it holds that

$$\mathbb{P}(\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2 > \varepsilon) \leq \mathbb{P}\left(\left\|\left(\frac{1}{T}\boldsymbol{P}'\boldsymbol{P}\right)^{-1} - \boldsymbol{I}_K\right\|_2 > \varepsilon_1\right) + \mathbb{P}\left(\left\|\frac{1}{\sqrt{T}}\boldsymbol{P}'\right\|_2\left\|\frac{1}{\sqrt{T}}\boldsymbol{Y}\right\|_2 > \varepsilon_2\right)$$
$$+\,\mathbb{P}\left(\|\boldsymbol{H}\boldsymbol{H}' - \boldsymbol{I}_K\|_2 > \varepsilon_3\right) + \mathbb{P}\left(\frac{1}{\sqrt{T}}\left\|\widehat{\boldsymbol{P}}' - \boldsymbol{H}\boldsymbol{P}'\right\|_2 > \varepsilon_4\right) + \mathbb{P}\left(\|\boldsymbol{H}^{-1}\|_2\left\|\frac{1}{\sqrt{T}}\boldsymbol{Y}\right\|_2 > \varepsilon_5\right)\ . \tag{30}$$

The claim then follows from Proposition A.7, Proposition A.3, Proposition A.5 and Proposition A.6 by setting $\varepsilon_1 = C_1(\sqrt{K/T} + \sqrt{\log(T)/T})$, $\varepsilon_2 = \sqrt{6\log(T)/C_m}$, $\varepsilon_3 = C_2[\sqrt{K/T} + \sqrt{\log(T)/T} + (p + \log(T))^{\frac{r\alpha+1}{r\alpha}}/(p^\alpha T) + 1/p^\alpha]$, $\varepsilon_4 = C_3[(p + \log(T))^{\frac{r\alpha+1}{r\alpha}}/(p^\alpha T) + 1/p^\alpha]$ and $\varepsilon_5 = \sqrt{3\log(T)/(2C_m)}$. The same propositions imply that the r.h.s. of (30) can be bounded by $O(T^{-\eta})$. □

Proposition A.5. *Suppose A.1–A.4 are satisfied.*

Then for all $\eta > 0$ there exists a $C > 0$ such that, for any T sufficiently large,

$$\sqrt{\frac{1}{T}} \left\| \widehat{\boldsymbol{P}} - \boldsymbol{P}\boldsymbol{H}' \right\|_2 \leq C \left[\frac{(p + \log(T))^{\frac{r_\alpha + 1}{r_\alpha}}}{p^\alpha T} + \frac{1}{p^\alpha} \right]$$

holds with probability at least $1 - T^{-\eta}$.

Proof of Proposition A.5. Let $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-1/2}$ and $\widehat{\boldsymbol{P}} = \boldsymbol{X}\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2}$ where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)'$. We note that

$$\begin{aligned}
\widehat{\boldsymbol{P}} - \boldsymbol{P}\boldsymbol{H}' &= \boldsymbol{X}\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2} - \boldsymbol{X}\boldsymbol{V}_K\boldsymbol{\Lambda}_K^{1/2}\boldsymbol{\Lambda}_K^{-1/2}\boldsymbol{V}_K'\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2} = \boldsymbol{X}(\boldsymbol{I}_p - \boldsymbol{V}_K\boldsymbol{V}_K')\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2} \\
&= \boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2} = \boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K\widehat{\boldsymbol{V}}_K'\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{\Lambda}}_K^{-1} \\
&= \boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'(\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K\widehat{\boldsymbol{V}}_K' + \widehat{\boldsymbol{V}}_R\widehat{\boldsymbol{\Lambda}}_R\widehat{\boldsymbol{V}}_R')\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{\Lambda}}_K^{-1} = \frac{1}{T}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{V}}_K\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{\Lambda}}_K^{-1} \\
&= \frac{1}{\sqrt{T}}\boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}'\frac{1}{\sqrt{T}}\widehat{\boldsymbol{P}}\widehat{\boldsymbol{\Lambda}}_K^{-1} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t'\frac{1}{\sqrt{T}}\widehat{\boldsymbol{P}}\widehat{\boldsymbol{\Lambda}}_K^{-1} \ .
\end{aligned}$$

This implies that for any $\varepsilon, \varepsilon_1, \varepsilon_2 > 0$ such that $\varepsilon = \varepsilon_1\varepsilon_2$ it holds

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{\frac{1}{T}}\left\|\widehat{\boldsymbol{P}} - \boldsymbol{P}\boldsymbol{H}'\right\|_2 > \varepsilon\right) \leq \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t'\right\|_2 \left\|\frac{1}{\sqrt{T}}\widehat{\boldsymbol{P}}\right\|_2 \left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2 > \varepsilon\right) \\
&\quad = \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t'\right\|_2 \left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2 > \varepsilon\right) \leq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2 > \varepsilon_1\right) + \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t'\right\|_2 > \varepsilon_2\right) \ .
\end{aligned}$$

We proceed by bounding the two terms on the r.h.s. of the last inequality. First we note that Proposition A.1 implies that, for all T sufficiently large

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Lambda}}_K^{-1}\right\|_2 > \frac{2}{c_K p^\alpha}\right) = 1 - \mathbb{P}\left(\frac{1}{\widehat{\lambda}_K} \leq \frac{2}{c_K p^\alpha}\right) = O\left(\frac{1}{T^\eta}\right) \ . \tag{31}$$

Second, we note that

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t'\right\|_2 \leq \left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t' - \mathbb{E}(\boldsymbol{u}_t\boldsymbol{u}_t')\right\|_2 + \|\mathbb{E}(\boldsymbol{u}_t\boldsymbol{u}_t')\|_2 = \left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t' - \mathbb{E}\left(\boldsymbol{u}_t\boldsymbol{u}_t'\right)\right\|_2 + c_{K+1} \ .$$

Proposition A.2 then implies that, for all T sufficiently large, we have

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t'\right\|_2 > C\left(\frac{(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{T}+\sqrt{\frac{(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{T}}\right)+c_{K+1}\right)=O\left(\frac{1}{T^\eta}\right). \tag{32}$$

The claim of the proposition then follows from inequalities (31) and (32) after noting that

$$C\left(\frac{(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{T}+\sqrt{\frac{(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{T}}\right)+c_{K+1}\leq C'\left(\frac{(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{T}+c_{K+1}\right).$$

□

Proposition A.6. *Suppose A.1–A.4 are satisfied.*

Then for any $\eta > 0$ there exist a positive constant C such that, for any T sufficiently large,

$$\|\boldsymbol{H}\boldsymbol{H}'-\boldsymbol{I}_K\|_2\leq C\left(\sqrt{\frac{K}{T}}+\sqrt{\frac{\log(T)}{T}}+\frac{(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}}{p^\alpha T}+\frac{1}{p^\alpha}\right),$$

holds with probability at least $1-T^{-\eta}$.

Proof. By repeated application of the triangle inequality we get

$$\begin{aligned}
\|\boldsymbol{H}\boldsymbol{H}'-\boldsymbol{I}_K\|_2 &\leq \left\|\boldsymbol{H}\boldsymbol{H}'-\frac{1}{T}\boldsymbol{H}\boldsymbol{P}'\boldsymbol{P}\boldsymbol{H}'\right\|_2+\left\|\frac{1}{T}\boldsymbol{H}\boldsymbol{P}'\boldsymbol{P}\boldsymbol{H}'-\boldsymbol{I}_K\right\|_2\\
&\leq \|\boldsymbol{H}\|_2^2\left\|\boldsymbol{I}_K-\frac{1}{T}\boldsymbol{P}'\boldsymbol{P}\right\|_2+\left\|\frac{1}{T}\boldsymbol{H}\boldsymbol{P}'\boldsymbol{P}\boldsymbol{H}'-\frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{P}\boldsymbol{H}'+\frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{P}\boldsymbol{H}'-\frac{1}{T}\widehat{\boldsymbol{P}}'\widehat{\boldsymbol{P}}\right\|_2\\
&\leq \|\boldsymbol{H}\|_2^2\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\boldsymbol{I}_K\right\|_2+\frac{1}{T}\left\|(\boldsymbol{P}\boldsymbol{H}'-\widehat{\boldsymbol{P}})'\boldsymbol{P}\boldsymbol{H}'\right\|_2+\frac{1}{T}\left\|\widehat{\boldsymbol{P}}'(\boldsymbol{P}\boldsymbol{H}'-\widehat{\boldsymbol{P}})\right\|_2\\
&\leq \|\boldsymbol{H}\|_2^2\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\boldsymbol{I}_K\right\|_2+\frac{1}{\sqrt{T}}\left\|\widehat{\boldsymbol{P}}-\boldsymbol{P}\boldsymbol{H}'\right\|_2\left(\frac{1}{\sqrt{T}}\|\boldsymbol{P}\|_2\|\boldsymbol{H}\|_2+\frac{1}{\sqrt{T}}\|\widehat{\boldsymbol{P}}\|_2\right).
\end{aligned}$$

For any $\varepsilon > 0$, $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, $\varepsilon_3 > 0$ and $\varepsilon_4 > 0$ such that $\varepsilon = \varepsilon_1\varepsilon_2 + \varepsilon_3\varepsilon_4$ it holds that

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{H}\boldsymbol{H}'-\boldsymbol{I}_K\|_2\geq\varepsilon\right)&\leq\mathbb{P}\left(\|\boldsymbol{H}\|_2^2\geq\varepsilon_1\right)+\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{P}_t\boldsymbol{P}_t'-\boldsymbol{I}_K\right\|_2\geq\varepsilon_2\right)\\
&+\mathbb{P}\left(\sqrt{\frac{1}{T}}\left\|\widehat{\boldsymbol{P}}-\boldsymbol{P}\boldsymbol{H}'\right\|_2\geq\varepsilon_3\right)+\mathbb{P}\left(\frac{1}{\sqrt{T}}\|\boldsymbol{P}\|_2\,\|\boldsymbol{H}\|_2+\frac{1}{\sqrt{T}}\|\widehat{\boldsymbol{P}}\|_2>\varepsilon_4\right).
\end{aligned} \tag{33}$$

The claim then follows from Proposition A.7, Propositions A.3 and Proposition A.5 by setting $\varepsilon_1 = 2c_1/c_K$, $\varepsilon_2 = C_1(\sqrt{K/T}+\sqrt{\log T/T})$, $\varepsilon_3 = C_2[(p+\log(T))^{\frac{r_\alpha+1}{r_\alpha}}/(p^\alpha T)+1/p^\alpha]$ and $\varepsilon_4 = 2\sqrt{c_1/c_K}+1$. The same propositions imply that the r.h.s. of (33) can be bounded by $O(T^{-\eta})$. □

Proposition A.7. *Suppose A.1–A.4 are satisfied. Then for any $\eta > 0$, for any T sufficiently large, it holds that (i) $\mathbb{P}(\sqrt{1/T}\|\boldsymbol{P}\|_2 \geq \sqrt{2}) = O(T^{-\eta})$, (ii) $\mathbb{P}(\|\boldsymbol{H}\|_2 \geq \sqrt{2c_1/c_K}) = O(T^{-\eta})$, (iii) $\mathbb{P}(\|\boldsymbol{H}^{-1}\|_2 \geq 2\sqrt{c_1/c_K}) = O(T^{-\eta})$, (iv) $\mathbb{P}(\sqrt{1/T}\|\boldsymbol{Y}\|_2 > \sqrt{\eta\log(T)/C_m}) = O(T^{-\eta})$.*

Proof. (*i*) By the triangle inequality, the fact that $\mathbb{E}\left(\boldsymbol{P}_t\boldsymbol{P}_t'\right) = \boldsymbol{I}_K$ and Proposition A.3 we have that for any $\eta > 0$ there exists a positive constant C such that, for all T sufficiently large,

$$\left\|\sqrt{\frac{1}{T}}\boldsymbol{P}\right\|_2^2 \leq \left\|\frac{1}{T}\boldsymbol{P}'\boldsymbol{P} - \boldsymbol{I}_K\right\|_2 + \|\boldsymbol{I}_K\|_2 \leq C\left(\sqrt{\frac{K}{T}} + \sqrt{\frac{\log T}{T}}\right) + 1 \leq 2$$

holds with probability at least $1 - O(T^{-\eta})$.

(*ii*) We begin by noting that

$$\|\boldsymbol{H}\|_2 = \|\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\widehat{\boldsymbol{V}}_K'\boldsymbol{V}_K\boldsymbol{\Lambda}_K^{1/2}\|_2 \leq \|\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\|_2\|\widehat{\boldsymbol{V}}_K\|_2\|\boldsymbol{V}_K\|_2\|\boldsymbol{\Lambda}_K^{1/2}\|_2 = \sqrt{c_1p^\alpha}\|\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\|_2 \ ,$$

where we have used the fact that A.3 implies $\|\boldsymbol{\Lambda}_K^{1/2}\|_2 = \sqrt{c_1p^\alpha}$. Now we may write

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{H}\|_2 \geq \sqrt{2\frac{c_1}{c_K}}\right) &\leq \mathbb{P}\left(\sqrt{c_1p^\alpha}\|\widehat{\boldsymbol{\Lambda}}_K^{-1/2}\|_2 \geq \sqrt{2\frac{c_1}{c_K}}\right) = \mathbb{P}\left(c_1p^\alpha\frac{1}{\widehat{\lambda}_K} \geq 2\frac{c_1}{c_K}\right)\\ &= 1 - \mathbb{P}\left(\widehat{\lambda}_K \geq c_K\frac{p^\alpha}{2}\right) \leq O\left(\frac{1}{T^\eta}\right) \ ,\end{aligned}$$

where the last inequality is implied by Proposition A.1.

(iii) We begin by noting that

$$\|\boldsymbol{H}^{-1}\|_2 = \|\boldsymbol{H}'(\boldsymbol{H}')^{-1}\boldsymbol{H}^{-1}\|_2 \leq \|\boldsymbol{H}\|_2\|(\boldsymbol{H}')^{-1}\boldsymbol{H}^{-1}\|_2 = \|\boldsymbol{H}\|_2\|\boldsymbol{I}_K + (\boldsymbol{H}')^{-1}\boldsymbol{H}^{-1} - \boldsymbol{I}_K\|_2$$
$$\leq \|\boldsymbol{H}\|_2 + \|\boldsymbol{H}\|_2\|(\boldsymbol{H}')^{-1}\boldsymbol{H}^{-1} - \boldsymbol{I}_K\|_2 \leq \|\boldsymbol{H}\|_2 + \|\boldsymbol{H}\|_2 \frac{\|\boldsymbol{H}\boldsymbol{H}' - \boldsymbol{I}_K\|_2}{1 - \|\boldsymbol{H}\boldsymbol{H}' - \boldsymbol{I}_K\|_2} ,$$

where we have used the inequality

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}\|_2 \leq \frac{\|\boldsymbol{A}^{-1}\|_2^2\|\boldsymbol{B} - \boldsymbol{A}\|_2}{1 - \|\boldsymbol{A}^{-1}\|_2\|\boldsymbol{B} - \boldsymbol{A}\|_2} ,$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible $n \times n$ matrices. The claim then follows from Proposition A.6 and part (ii) of this proposition.

(iv) By A.1, for any $\varepsilon > 0$, $\mathbb{P}\left(\sqrt{1/T}\|\boldsymbol{Y}\|_2 > \varepsilon\right) \leq \mathbb{P}\left(\max_{t \leq T} |Y_t| > \varepsilon\right) \leq T\mathbb{P}\left(|Y_t| > \varepsilon\right) \leq T\exp\left(-C_m\varepsilon^2\right)$. Thus, by choosing $\varepsilon = \sqrt{\eta \log(T)/C_m}$ we have that $\sqrt{1/T}\|\boldsymbol{Y}\|_2 > \sqrt{(1+\eta)\log(T)/C_m}$ holds at most with probability $O(T^{-\eta})$. □

A.3 Proof of Proposition 3

Proof of Proposition 3. Define the empirical risk differential for an arbitrary $\boldsymbol{\vartheta} \in \mathbb{R}^K$ as

$$\widehat{\mathcal{L}}_{\boldsymbol{\vartheta}} = R_T(\boldsymbol{\vartheta}) - R_T(\boldsymbol{\vartheta}^*) = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)^2 + \frac{2}{T}\sum_{t=1}^{T}(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)(\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*) .$$

Assume that it holds that

$$\|\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2} > \frac{72C_{\sigma^2}^{\frac{1}{2}}}{\kappa_1^2\kappa_2}\sqrt{\frac{K\log(T)}{T}} . \tag{34}$$

Condition on the events of Proposition A.8 and Proposition A.9, for any $\eta > 0$, for any T sufficiently large, with probability at least $1 - O(T^{-\eta})$, we have that

$$\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)^2 \overset{(a)}{\geq} \frac{\kappa_1^2\kappa_2}{2}\|\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 \overset{(b)}{>} 36C_{\sigma^2}^{\frac{1}{2}}\|\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}\sqrt{\frac{K\log(T)}{T}}$$
$$\overset{(c)}{\geq} \left|\frac{2}{T}\sum_{t=1}^{T}(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)(\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)\right| ,$$

where (a) follows from Proposition A.8, (b) follows from condition (34), and (c) follows from Proposition A.9. Thus, conditional on the events of Proposition A.8 and Proposition A.9 and assuming (34) holds we have high probability that $\widehat{\mathcal{L}}_{\boldsymbol{\vartheta}} > 0$. Since the empirical risk minimizer $\hat{\boldsymbol{\vartheta}}$ satisfies $\widehat{\mathcal{L}}_{\hat{\boldsymbol{\vartheta}}} \leq 0$ then conditional on the same events we have $\|\boldsymbol{P}_t'\hat{\boldsymbol{\vartheta}} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2} \leq \frac{72C_{\sigma^2}^{\frac{1}{2}}}{\kappa_1^2\kappa_2}\sqrt{\frac{K\log(T)}{T}}$. The claim then follows after noting that

$$\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^*\|_2 = \|\boldsymbol{P}_t'\tilde{\boldsymbol{\vartheta}} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}^2 \leq C_{\sigma^2}\left(\frac{72}{\kappa_1^4\kappa_2^2}\right)^2\frac{K\log(T)}{T} . \tag{35}$$

It is important to emphasize that the L_2 norm in (35) is the L_2 norm conditional on $\{\tilde{\boldsymbol{\vartheta}} = \tilde{\boldsymbol{\vartheta}}(\mathcal{D})\}$. □

Proposition A.8. *Suppose A.1–A.5 are satisfied.*

Then for any $\eta > 0$, for all T sufficiently large and any $\boldsymbol{\vartheta} \in \mathbb{R}^K$,

$$\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta})^2 \geq \frac{\kappa_1^2\kappa_2}{2}\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta}\|_{L_2}^2 ,$$

holds with probability at least $1 - T^{-\eta}$.

Proof of Proposition A.8. This follows from Brownlees and Guðmundsson (2025, Proposition 1 and Theorem 2). Note that the proposition there establishes an analogous claim for $\eta = 1$, but inspection of the proof shows that it is straightforward to allow for any $\eta > 0$. Note that those results require that $r_K < r_\alpha/(r_\alpha + 1)$, which is implied by A.4. Also, note that A.5 implies that the small-ball condition is also satisfied by $\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta}$. □

Proposition A.9. *Suppose A.1–A.3 are satisfied. Then for any $\eta > 0$ and any $\boldsymbol{\vartheta} \in \mathbb{R}^K$,* *there exists a positive constant C such that, for all T sufficiently large*

$$\left|\frac{1}{T}\sum_{t=1}^{T}(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)(\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta})\right| \leq C\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta}\|_{L_2}\sqrt{\frac{K\log(T)}{T}} \ ,$$

holds with probability at least $1 - T^{-\eta}$.

Proof of Proposition A.9. We begin by noting that for any $\boldsymbol{\vartheta} \in \mathbb{R}^K \setminus \boldsymbol{\vartheta}^*$ we have that

$$\begin{aligned}\left|\sum_{t=1}^{T}\frac{(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)(\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta})}{T\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta}\|_{L_2}}\right| &= \left|\frac{1}{T}\sum_{t=1}^{T}(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)\boldsymbol{P}_t'\boldsymbol{\nu}\right| \\ = \left|\frac{1}{T}\sum_{t=1}^{T}[(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)\boldsymbol{P}_t']\boldsymbol{\nu}\right| &= \left|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{W}_t'\boldsymbol{\nu}\right| \ ,\end{aligned}$$

where $\boldsymbol{W}_t = (W_{1\,t}, \ldots, W_{K\,t})'$ with $W_{i\,t} = (Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)P_{it}$ and $\boldsymbol{\nu} = (\boldsymbol{\vartheta}^* - \boldsymbol{\vartheta})/\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta}\|_{L_2}$. Note that $\|\boldsymbol{\nu}\|_2 = 1$. Then, for any $\boldsymbol{\vartheta} \in \mathbb{R}^K \setminus \boldsymbol{\vartheta}^*$ it holds

$$\begin{aligned}&\mathbb{P}\left(\frac{\left|\sum_{t=1}^{T}(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)(\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta})\right|}{T\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta}\|_{L_2}} > \varepsilon\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{\vartheta}\in\mathbb{R}^K\setminus\boldsymbol{\vartheta}^*}\frac{\left|\sum_{t=1}^{T}(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)(\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta})\right|}{T\|\boldsymbol{P}_t'\boldsymbol{\vartheta}^* - \boldsymbol{P}_t'\boldsymbol{\vartheta}\|_{L_2}} > \varepsilon\right) \\ &= \mathbb{P}\left(\sup_{\boldsymbol{\nu}:\|\boldsymbol{\nu}\|_2=1}\left|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{W}_t'\boldsymbol{\nu}\right| > \varepsilon\right) \leq \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{W}_t\right\|_2 > \varepsilon\right) \ .\end{aligned}$$

A.1 and Lemma B.1 imply that $\boldsymbol{W}_t$ is sub-exponential for some parameter $C_m' > 0$ for each i. Standard properties of strong mixing processes imply that $\{\boldsymbol{W}_t\}$ inherits the mixing properties of $\{(Y_t, \boldsymbol{X}_t')'\}$ spelled in A.2. Assuming that A.4 is satisfied we have that Lemma B.2 holds which implies that, for any $\eta > 0$ there exists a positive constant C such that, for all T sufficiently large,

$$\sup_{\boldsymbol{\vartheta}\in\mathbb{R}^K\setminus\boldsymbol{\vartheta}^*}\frac{\left|\sum_{t=1}^{T}(Y_t - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)(\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*)\right|}{T\|\boldsymbol{P}_t'\boldsymbol{\vartheta} - \boldsymbol{P}_t'\boldsymbol{\vartheta}^*\|_{L_2}} \leq C\sqrt{\frac{K\log(T)}{T}}$$

holds with probability at least $1 - O(T^{-\eta})$. This implies the claim. □

A.4 Proof of Proposition 4

Proof of Proposition 4. The claim follows from the fact that $\|\boldsymbol{u}_t'\boldsymbol{\gamma}^*\|_{L_2}^2 = (\boldsymbol{\gamma}^*)'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{\gamma}^* = (\boldsymbol{\theta}^*)'\boldsymbol{V}_R\boldsymbol{\Lambda}_R\boldsymbol{V}_R'\boldsymbol{\theta}^*$ from Lemma 1. □

B Auxiliary Results

Lemma B.1. *Suppose A.1 and A.3 are satisfied.*

Then we have that (i) there exists some constant $C_1 > 0$ such that $\boldsymbol{P}_t$ is a sub-Gaussian vector with parameter C_1; (ii) if $c_{K+1} > 0$ then there exists some constant $C_2 > 0$ such that $\boldsymbol{u}_t$ is sub-Gaussian vector with parameter C_2 otherwise $\boldsymbol{u}_t$ is degenerate at zero.

Proof. (i) Since $\boldsymbol{P}_t = \boldsymbol{\Lambda}_K^{-1/2}\boldsymbol{V}_K'\boldsymbol{X}_t = \boldsymbol{V}_K'\boldsymbol{Z}_t$, we have that for any $\varepsilon > 0$

$$\mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} |\boldsymbol{v}'\boldsymbol{P}_t| > \varepsilon\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} |\boldsymbol{v}'\boldsymbol{Z}_t| > \varepsilon\right) \leq \exp\left(-C_m\varepsilon^2\right) .$$

Then, we have that $\boldsymbol{P}_t$ is a sub-Gaussian vector with parameter $C_1 = C_m$.
(ii) We assume that $c_{K+1} > 0$. Since $\boldsymbol{u}_t = \boldsymbol{V}_R\boldsymbol{V}_R'\boldsymbol{X}_t = \boldsymbol{V}_R\boldsymbol{\Lambda}_R^{1/2}\boldsymbol{V}_R'\boldsymbol{Z}_t$, we have that for any $\varepsilon > 0$

$$\mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} |\boldsymbol{v}'\boldsymbol{u}_t| > \varepsilon\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} c_{K+1}^{1/2}|\boldsymbol{v}'\boldsymbol{Z}_t| > \varepsilon\right) \leq \exp\left(-\frac{C_m}{c_{K+1}}\varepsilon^2\right) ,$$

where we have used the fact that $\|\boldsymbol{V}_R\boldsymbol{\Lambda}_R^{1/2}\boldsymbol{V}_R'\|_2 \leq \|\boldsymbol{V}_R\|_2\|\boldsymbol{\Lambda}_R^{1/2}\|_2\|\boldsymbol{V}_R'\|_2 = c_{K+1}^{1/2}$. Then we have that that $\boldsymbol{u}_t$ is a sub-Gaussian vector with parameter $C_2 = C_m/c_{K+1}$. □

Lemma B.2. *Let $\{\boldsymbol{Z}_t\}_{t=1}^T$ be a stationary sequence of d-dimensional zero-mean random vectors. Suppose (i) for any $\varepsilon > 0$ it holds $\sup_{1\leq i\leq d}\mathbb{P}(|Z_{i\,t}| > \varepsilon) \leq \exp(-C_m\varepsilon)$ for some $C_m > 0$; (ii) the α-mixing coefficients of the sequence satisfy $\alpha(l) < \exp(-C_\alpha l^{r_\alpha})$ for some $C_\alpha > 0$ and $r_\alpha > 0$; and (iii) $d = \lfloor C_d T^{r_d}\rfloor$ for some $C_d > 0$ and $r_d \in [0, 1]$.*

Then for any $\eta > 0$ there exists a positive constant C_η such that, for any T sufficiently *large, it holds that*

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Z}_t\right\|_2 > C_\eta\sqrt{\frac{d\log(T)}{T}}\right) \leq \frac{1}{T^\eta} \ .$$

Proof. Let C^* denote a positive constant to be chosen below. Note that

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Z}_t\right\|_2 \geq C^*\sqrt{\frac{d\log(T)}{T}}\right) \leq d\max_{1\leq i\leq d}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z_{it}\right| \geq C^*\sqrt{T\log(T)}\right) \ .$$

Let $\sum_{t=1}^{T}Z_{it} = \sum_{t=1}^{T}Z'_{it} + \sum_{t=1}^{T}Z''_{it}$ where $Z'_{it} = Z_{it}\mathbb{1}\left(|Z_{it}| \leq b_T\right) - \mathbb{E}\left(Z_{it}\mathbb{1}\left(|Z_{it}| \leq b_T\right)\right)$ and $Z''_{it} = Z_{it}\mathbb{1}\left(|Z_{it}| > b_T\right) - \mathbb{E}\left(Z_{it}\mathbb{1}\left(|Z_{it}| > b_T\right)\right)$. Then we have

$$\mathbb{P}\left(\left|\sum_{t=1}^{T}Z_{it}\right| > C^*\sqrt{T\log(T)}\right) \leq \mathbb{P}\left(\left|\sum_{t=1}^{T}Z'_{it}\right| > \frac{C^*}{2}\sqrt{T\log(T)}\right) + \mathbb{P}\left(\left|\sum_{t=1}^{T}Z''_{it}\right| > \frac{C^*}{2}\sqrt{T\log(T)}\right) \ .$$

The sequence $\{Z'_{it}\}_{t=1}^{T}$ has the same mixing properties as $\{\boldsymbol{Z}_t\}_{t=1}^{T}$ and $\sup_{1\leq i\leq d}\|Z'_{it}\|_\infty < 2b_T$. Define $\varepsilon_T = (C^*/2)\sqrt{T\log(T)}$, $b_T = 2(r_d + 1/2 + \eta)C_m^{-1}\log(T)$ and $M_T = \lfloor b_T^{-1}\sqrt{T/\log(T)}\rfloor$. For any T sufficiently large $M_T \in [1, T]$ and $4(2b_T)M_T < \varepsilon_T$, implying that the conditions of Theorem 2.1 of Liebscher (1996) are satisfied. Then we have

$$\mathbb{P}\left(\left|\sum_{t=1}^{T}Z'_{it}\right| > \varepsilon_T\right) < 4\exp\left(-\frac{\varepsilon_T^2}{64\frac{T}{M_T}D(T, M_T) + \frac{16}{3}b_T M_T\varepsilon_T}\right) + 4\frac{T}{M_T}\exp\left(-C_\alpha M_T^{r_\alpha}\right) \ ,$$

with $D(T, M_T) = \sup_{1\leq i\leq d}\mathbb{E}\left[\left(\sum_{t=1}^{M_T}Z'_{it}\right)^2\right]$. Define $\gamma(l) = \sup_{1\leq i\leq d}|\operatorname{Cov}(Z'_{it}, Z'_{it+l})|$ for $l = 0, 1, \ldots$ and note that $D(T, M_T) \leq M_T\sum_{l=-M_T+1}^{M_T-1}\gamma(l)$. For $l = 0, \ldots$ it holds that $\gamma(l) \leq 12\alpha(l)^{\frac{1}{2}}\sup_{1\leq i\leq d}\|Z'_{it}\|_{L_4}^2 \leq 48\alpha(l)^{\frac{1}{2}}\sup_{1\leq i\leq d}\|Z_{it}\|_{L_4}^2$ where the first inequality follows from Davydov's inequality (Bosq, 1998, Corollary 1.1) and the second one follows from the fact that $\|Z'_{it}\|_{L_4} \leq 2\|Z_{it}\mathbb{1}\left(|Z_{it}| \leq b_T\right)\|_{L_4} \leq 2\|Z_{it}\|_{L_4}$ This implies that $D(T, M_T) < M_T C_{\sigma^2}$ where $C_{\sigma^2} = (96C_{m,4}\sum_{l=0}^{\infty}\alpha(l)^{\frac{1}{2}}) \vee 1$ where $C_{m,4} = \sup_{1\leq i\leq d}\|Z_{it}\|_{L_4}^2$. We then have that

$$\begin{aligned} d\max_{1\leq i\leq d}\mathbb{P}\left(\left|\sum_{t=1}^{T}Z'_{it}\right| > \varepsilon_T\right) &\leq 4C_d\exp\left(r_d\log(T) - \frac{(C^*)^2\log(T)}{256C_{\sigma^2} + \frac{32}{3}C^*}\right) + 4C_d T^{1+r_d}\exp\left(-C_\alpha M_T^{r_\alpha}\right) \\ &< 4C_d\exp\left(-\eta\log(T)\right) + 4C_d\exp\left(2\log(T) - C_\alpha M_T^{r_\alpha}\right) = \frac{5C_d}{T^\eta} \ , \end{aligned} \tag{36}$$

where the second inequality follows from a sufficiently large choice of the constant C^*. The sequence $\{Z''_{it}\}_{t=1}^T$ is such that

$$\begin{aligned}
&d \max_{1\le i\le d} \mathbb{P}\left(\left|\sum_{t=1}^T Z''_{it}\right| > \varepsilon_T\right) \le \frac{d}{\varepsilon_T} \max_{1\le i\le d} \mathbb{E}\left|\sum_{t=1}^T Z''_{it}\right| \le \frac{d}{\varepsilon_T}\sum_{t=1}^T \max_{1\le i\le d} \mathbb{E}|Z''_{it}| \\
&\le \frac{2dT}{\varepsilon_T} \max_{1\le i\le d} \mathbb{E}\left|Z_{it}\mathbb{1}\left(|Z_{it}| > b_T\right)\right| \le \frac{2dT}{\varepsilon_T} \max_{1\le i\le d} \|Z_{it}\|_{L_2}\|\mathbb{1}\left(|Z_{it}| > b_T\right)\|_{L_2} \\
&= \frac{2dT}{\varepsilon_T} \max_{1\le i\le d} \|Z_{it}\|_{L_2}\mathbb{P}(|Z_{it}| > b_T)^{\frac{1}{2}} < \frac{2dT}{\varepsilon_T}\sigma^2 \exp\left(-\frac{C_m}{2}b_T\right) \\
&\le \frac{4C_d\sigma^2}{C^*\sqrt{\log T}} \exp\left(\left(r_d + \frac{1}{2}\right)\log(T) - \frac{C_m}{2}b_T\right) < \frac{1}{T^\eta}\ , \qquad (37)
\end{aligned}$$

where the first inequality follows from Markov's inequality and $\sigma^2 = \sup_{1\le i\le d} \|Z_{it}\|_{L_2}$. Equations (36) and (37) imply the claim. □

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, **81**(3), 1203–1227.

Amengual, D. and Watson, M. W. (2007). Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business & Economic Statistics*, **25**(1), 91–96.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, **71**(1), 135–171.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, **40**(1).

Bai, J. and Li, K. (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. *Review of Economics and Statistics*, **98**(2), 298–309.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, **70**, 191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, **74**(4), 1133–1150.

Bai, J. and Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics*, **212**(1), 78–96.

Bai, J. and Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*, **235**(2), 1893–1916.

Barigozzi, M. and Cho, H. (2020). Consistent estimation of high-dimensional factor models when the factor number is over-estimated. *Electronic Journal of Statistics*, **14**(2), 2892–2921.

Birge, L. and Massart, P. (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.* Springer, New York, Second edition.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Brownlees, C. and Guðmundsson, G. S. (2025). Performance of empirical risk minimization for linear regression with dependent data. *Econometric Theory*, **forthcoming**.

Brownlees, C. and Llorens-Terrazas, J. (2025). Empirical risk minimization for time series: Nonparametric performance bounds for prediction. *Journal of Econometrics*, **forthcoming**.

Donoho, D., Gavish, M., and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, **46**(4).

Fan, J., Liao, Y., and Mincheva, M. (2011). High Dimensional Covariance Matrix Estimation in Approximate Factor Models. *The Annals of Statistics*, **39**, 3320–3356.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**(4), 603–680.

Fan, J., Liao, Y., and Wang, W. (2016). Projected principal component analysis in factor models. *The Annals of Statistics*, **44**(1).

Fan, J., Masini, R. P., and Medeiros, M. C. (2024). Bridging Factor and Sparse Models. *The Annals of Statistics*, **forthcoming**.

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

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, **100**(471), 830–840.

Gonçalves, S. and Perron, B. (2020). Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, **218**(2), 476–495.

Hotelling, H. (1957). the relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, **10**(2), 69–79.

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

Kendall, M. (1957). *A Course In Multivariate Statistics*. Griffin, London.

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

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, **40**(2), 694 – 726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**(4), 901–918.

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

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

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

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.

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

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, **92**(4), 1004–1016.

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

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

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, **30**(4), 481–493.

Su, L. and Wang, X. (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics*, **198**(1), 84–101.

Tsybakov, A. B. (2003). Optimal Rate of Linear Aggregation. In *Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg.

Uematsu, Y. and Yamagata, T. (2022). Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, **41**(1), 213–227.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, chapter 5, page 210–268. Cambridge University Press, Cambridge, England.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, **45**(3).

Wong, K. C., Li, Z., and Tewari, A. (2020). Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, **48**(2), 1124 – 1142.

Yu, L., He, Y., and Zhang, X. (2019). Robust factor number specification for large-dimensional elliptical factor model. *Journal of Multivariate Analysis*, **174**, 104543.