Fixed-order PCA: Theory for Overestimated Factor Models

Yuan Liao
Department of Economics
University of Iowa

Xin Tong
Department of Mathematics
National University of Singapore

Wanjie Wang
Department of Statistics
National University of Singapore

Dacheng Xiu
Booth School of Business
University of Chicago and NBER

May 18, 2026

Abstract

We develop asymptotic theory for principal component analysis (PCA) of a high-dimensional factor model in which the working dimension R is fixed and only required to satisfy $R \geq r$, where r is the true number of factors. Building on anisotropic local laws from random matrix theory, we show that the "extra" empirical eigencomponents beyond the r-th are asymptotically noise-governed, incoherent, and nearly orthogonal to the factor loadings. We introduce two rotations, an expanded $r \times R$ map H' and a compressed $R \times r$ map H^+, and establish consistency of the estimated factors under both. As an application, we analyze a factor-augmented regression for treatment-effect inference and prove $\sqrt{T}$-asymptotic normality for every fixed $R \geq r$. These results provide a theoretical underpinning for the common empirical practice of adopting a conservative upper bound on the number of factors, and shift the analytical burden from consistent dimension selection to the milder requirement of bounding r from above.

Key words: factor model, fixed-R PCA, treatment effect, overestimated rank

1 Introduction

Factor models provide a parsimonious description of high-dimensional data by decomposing an observation matrix $X \in \mathbb{R}^{N\times T}$ as

$$X = BF' + U, \tag{1.1}$$

where $B \in \mathbb{R}^{N\times r}$ is the loading matrix, $F \in \mathbb{R}^{T\times r}$ collects the latent factors, and $U \in \mathbb{R}^{N\times T}$ collect the idiosyncratic errors. Such models have become central in economics, finance, genomics, and signal processing, and principal component analysis (PCA) is by far the most widely used estimator. The asymptotic theory developed by Connor and Korajczyk (1986); Bai (2003) and others typically rests on a critical prerequisite: the investigator must first consistently estimate the true number of factors r. Under strong-factor assumptions and clear eigenvalue gaps, a variety of methods, such as the information-criteria-based procedure (Bai and Ng, 2002) and the eigenvalue-ratio test (Ahn and Horenstein, 2013), can consistently recover r. In practice, however, the signal is rarely so clean. In such settings, these methods can be unstable and may fail to detect weak factors while also selecting spurious ones. As a result, downstream inference may be adversely affected, particularly when the number of factors is underestimated.

Motivated by this fragility, empirical researchers often proceed by selecting a working number of factors R that is intended to exceed the true r, and then base estimation and inference on this potentially over-specified model. Variants of this approach appear widely in the applied literature, with R typically chosen by rule of thumb or guided by domain knowledge. Despite its prevalence, the statistical properties of this practice remain largely uncharacterized. This paper provides a theoretical foundation for such over-specification.

Contributions

We develop a unified spectral theory for *fixed-order* PCA, wherein the working dimension R is an arbitrary, user-specified integer satisfying $R \geq r$. Our contributions are threefold.

(1) Spectrum and eigenvectors beyond the true factor dimension. Building on anisotropic local laws from random matrix theory (Knowles and Yin, 2017), we characterize the empirical singular values $\lambda_{r+1}(X), \ldots, \lambda_R(X)$ and their associated singular vectors. We show that the "extra" singular values track those of the noise matrix U up to a sharp remainder, and that the overestimated singular vectors are incoherent (also known as localized) with respect to any direction independent of U and are near-orthogonal to the factor space. Our random-matrix-based analysis achieves a strictly sharper alignment rate between the overestimated principal-component directions and the factor loadings than the standard Davis–Kahan / $\sin\Theta$ benchmark. These facts provide the technical basis for the rest of the paper.

(2) Estimation of the full factor space. Because the fitted factor matrix $\widehat{F}$ has R columns rather than r, the usual notion of a rotation matrix must be extended. We introduce two rotations, an *expanded* $r \times R$ rotation H and a *compressed* $R \times r$ generalized inverse H^+, and establish convergence rates for $\widehat{F} - FH'$ and $\widehat{F}H^+ - F$. The former is faster because it only requires F to lie in the column space of $\widehat{F}$, whereas the latter must further extract each of the r factor directions from the R-dimensional fitted space.

(3) Robust factor-augmented inference. As an illustrative application, we study treatment-effect estimation in a factor-augmented regression, where the number of latent confounders is unknown and consistent selection of r is not invoked. We show that $\sqrt{T}$-asymptotic normality of the resulting estimator persists for every fixed $R \geq r$ when $r \geq 1$. Adding extra principal components may increase residual variation but does not affect the signal at first order, because the extra components are asymptotically orthogonal to the factor space and, since $R - r$ is fixed, their inclusion inflates the variance by an asymptotically negligible factor of $1 + O(1/T)$.

These results imply that, within the asymptotic regime considered here, overestimating r by a bounded amount is asymptotically harmless, whereas underestimation can have first-order consequences. Fixed-order PCA therefore provides a conservative and transparent procedure.

Related Literature

Robustness to overspecification of the factor number. The consequences of overspecifying factor dimension have been examined in several settings. Moon and Weidner (2015) showed that inference in interactive-effects models remains valid when the number of factors is overspecified, using perturbation theory for linear operators (Kato, 1995). They established conditions under which the inclusion of overestimated eigenvectors does not affect the first order behavior of the parameter of interest. Their operator-perturbation route is more flexible across estimator families, while the random-matrix-based analysis that we adopt in this paper specializes to PCA but yields sharper rates when applicable (e.g., Theorem 2.1(iii)). In addition, we allow the factors to be much weaker in the sense that the top singular values of X can grow much slower than NT, which is a strictly broader regime than the strong-factor scaling employed in Moon and Weidner (2015). Barigozzi and Cho (2020) established consistency of the low-rank component under overestimation through a trimmed PCA estimator that enforces incoherence by post-processing. Fan and Liao (2022) and Choi et al. (2025) propose diversified-projection estimators that are robust to overspecification; their approach requires prespecified weights that are independent of the idiosyncratic noise, typically obtained from external characteristics or a held-out sub-sample. Our analysis shows that, under the setting of (1.1) together with the incoherence conditions of Knowles and Yin (2017), vanilla PCA is itself robust to overspecification, with no trimming, weighting, or external information

required; the incoherence that those papers manufacture by post-processing or by external weights is here a generic property of noise-side singular vectors, supplied by the local law.

Entrywise eigenvector perturbation. A parallel methodological line (Fan et al., 2018; Cape et al., 2019; Abbe et al., 2020; Fan et al., 2021) develops entrywise (sup-norm) perturbation bounds for the leading singular subspace of low-rank, incoherent signal matrices, sharpening the Davis–Kahan $\sin\Theta$ rate by exploiting incoherence of the signal. Importantly, their results only apply to the first r eigenvectors. In contrast, the setting $R > r$ requires a different apparatus: the empirical eigencomponents that govern overspecification lie inside the noise bulk, where no spectral gap is available, so the deterministic perturbation arguments developed in this literature for the leading subspace do not directly apply. We draw instead on the anisotropic local law of Knowles and Yin (2017), which supplies the analogous entrywise control of noise singular vectors in this no-gap regime; this random-matrix-based control is the natural counterpart of the entrywise perturbation bounds derived in the cited works, and is what enables our finer analysis of the extra components.

Spectral methods with weak or mixed signals. More broadly, our results complement a growing literature on the spectrum of factor and spiked-covariance models with weak or mixed signal strengths (Onatski, 2010, 2012; Wang and Fan, 2017; Freyaldenhoven, 2022; Uematsu and Yamagata, 2023; Bai and Ng, 2023; Giglio et al., 2023). Our theoretical results are developed under a weak-factor regime, in which the signal strength is required only to diverge with the sample size and is not constrained to the strong-factor scaling of Bai (2003); this places our analysis squarely within the asymptotic setting considered by these works and makes the resulting spectral characterizations directly relevant to the inferential and prediction problems they study. Relative to that literature, we provide a single framework that simultaneously handles signal and noise eigencomponents in an overestimated spectrum, with explicit rates that readily plug into other factor-model inference problems. Our approach takes a different route than the factor-number selection Bai and Ng (2002); Onatski (2010); Ahn et al. (2013): rather than asking when these procedures recover r exactly, we ask what inference can be performed given an arbitrary upper bound, which is the practically relevant object once the analyst recognizes that a sharp $\widehat{r}$ is unavailable.

Connection to double machine learning. The factor-augmented inferential application we develop in Section 3 sits within the partialling-out / double machine learning (DML) tradition (Belloni et al., 2014; Chernozhukov et al., 2016; Hansen and Liao, 2018), in which a low-dimensional treatment effect is estimated from a Neyman-orthogonal moment after residualizing both the outcome and the treatment with respect to high-dimensional nuisance components. Two features distinguish our setting. First, the nuisance is the *latent* factor space spanned by F, recovered by PCA from the panel X rather than

constructed by selecting or fitting a function of observed controls; the two nuisances of canonical DML — the conditional expectations of the outcome and the treatment given the controls — share a common latent factor space in our setting, so a single PCA step suffices to residualize both equations, and the "double" structure of DML survives in the residualization rather than in the ML estimation step; the standard DML "product of nuisance rates" condition is replaced by a factor-strength condition stated explicitly in Assumption 3.1(ii). Second, in the factor-augmented regression of Section 3, a structural independence assumption between the regression errors and the factor–noise pair (F, U) supplies the conditional independence that cross-fitting is engineered to deliver in the i.i.d. setting; consequently, neither sample splitting nor data-driven tuning of the working dimension R is required, and any fixed $R \geq r$ with $R - r$ bounded yields valid $\sqrt{T}$ inference.

Organization and Notation

The remainder of the paper is organized as follows. Section 2 introduces the model, fixed-order PCA, and our main spectral results. Section 3 applies these results to inference on a treatment coefficient in a factor-augmented regression. Section 4 reports simulation evidence. Section 6 concludes. All proofs are deferred to the supplementary material.

Throughout the paper, $\lambda_k(A)$ denotes the k-th largest singular value of a matrix A, $\|A\|$ its spectral norm, and $\|A\|_{\mathrm{F}}$ its Frobenius norm. For a full-column-rank matrix A, $P_A = A(A'A)^{-1}A'$; for an arbitrary matrix, A^+ denotes its Moore–Penrose pseudoinverse and $P_A = AA^+$. We write $a_n \asymp b_n$ if $c\, b_n \leq a_n \leq C\, b_n$ for some $0 < c \leq C < \infty$, and $a_n \ll b_n$ (equivalently, $a_n = o(b_n)$) if $a_n / b_n \to 0$.

We write $\Phi(T) \prec \phi(T)$ (*stochastic domination*) if for every $K > 0$ and $\delta > 0$ there exists $C > 0$ such that $\mathrm{P}\{\Phi(T) > T^{\delta}\, C\phi(T)\} \leq T^{-K}$ for all sufficiently large T. This is stronger than the standard O_P notation, which we reserve for ordinary boundedness in probability. We use c and C for generic positive constants whose values may change across displays, and ε for an arbitrarily small positive constant.

2 Fixed-Order PCA: Model and Theory

2.1 Factor Model Setup

We consider the static approximate factor model in which, for each time period $t = 1, \ldots, T$, we observe an $N \times 1$ vector x_t satisfying

$$x_t = B f_t + u_t.$$

Here B is an $N \times r$ matrix of factor loadings, f_t is an $r \times 1$ vector of latent common factors, and u_t is an $N \times 1$ vector of idiosyncratic components. We assume throughout that u_t is independent of f_t and that both have mean zero.

Let $X = (x_1, \ldots, x_T)$, $F = (f_1, \ldots, f_T)'$, and $U = (u_1, \ldots, u_T)$. Then X admits the matrix representation

$$X = M + U, \qquad M = BF',$$

where $M \in \mathbb{R}^{N \times T}$ is the low-rank signal matrix and $U \in \mathbb{R}^{N \times T}$ collects idiosyncratic errors.

We model idiosyncratic components as $u_t = \Sigma_e^{1/2} e_t$, where e_t is an $N \times 1$ vector of standardized shocks and Σ_e is an $N \times N$ positive-definite matrix. In matrix form,

$$U = \Sigma_e^{1/2} E, \qquad E = (e_1, \ldots, e_T),$$

and we denote by $\sigma_1, \ldots, \sigma_N$ the eigenvalues of Σ_e. The matrix Σ_e is allowed to be a general positive definite matrix subject to the spectral regularity conditions in Assumption 2.1(ii) below: its eigenvalues are bounded and the associated deformed Marchenko–Pastur law is regular. Together these conditions accommodate cross-sectional heteroskedasticity and weak cross-sectional dependence in the idiosyncratic components, but rule out approximate factor-like structures within u_t that would create outlying spikes in the spectrum of Σ_e; the latter would correspond to additional, undetected factors and is best modelled by enlarging r.

Two features of this setup deserve emphasis. First, B is treated as deterministic and F as random throughout; statements such as "$\max_{i \le N} \|b_i\| = O(1)$" below should be read as deterministic bounds uniform in N. Second, all dependence in U is channeled through the deterministic operator $\Sigma_e^{1/2}$ acting on entries $e_{i,t}$ that are independent across both i and t (Assumption 2.1(i)). This is the standard random-matrix setting in which the anisotropic local law of Knowles and Yin (2017) applies, but it is more restrictive than the classical approximate-factor-model framework of Bai (2003), which permits weak *serial* as well as cross-sectional dependence in u_t. We emphasize this trade-off explicitly: on one hand, both cross-sectional and serial independence are required by the local-law we use to characterize the spectral theory of the extra eigenvectors. On the other hand, a more general HAC-type extension to weakly serially-dependent u_t is the most empirically pressing direction for future work, which we list accordingly in Section 6. The applications for which our setup is best suited are therefore those in which independence across (i, t) is plausible by sampling design — repeated cross-sections, randomized rollouts, large-panel snapshots in survey data. Meanwhile, in this paper the factor process f_t is indeed allowed to be serially dependent.

Because B and F are identified only up to a rotation, we fix a canonical normalization tied to the singular vectors of M. Let $M = \Xi_r L_r V_r'$ denote the SVD of BF', write $S_B = N^{-1} B'B$ and $S_f = T^{-1} F'F$, and let $H_B, H_F \in \mathbb{R}^{r \times r}$ be the rotation matrices that

simultaneously diagonalize $S_B^{1/2} S_f S_B^{1/2}$ and $S_f^{1/2} S_B S_f^{1/2}$, normalized so that

$$\frac{1}{\sqrt{N}} B \;=\; \Xi_r \, H_B^{-1}, \qquad \frac{1}{\sqrt{T}} F \;=\; V_r \, H_F^{-1}, \qquad (H_F' H_B)^{-1} \;=\; (NT)^{-1/2} \, L_r. \tag{2.1}$$

The construction of H_B and H_F from (S_B, S_f) and the verification of (2.1) are entirely algebraic; the explicit formulas and computation are deferred to Section B. The identities in (2.1) are the only consequences used in the body of the paper.

2.2 Fixed-Order PCA

Throughout, the true number of latent factors $r = \dim(f_t) \geq 0$ is assumed to be fixed and bounded. When factors are not strong (the singular values of M grow slower than $\sqrt{NT}$), consistently estimating r typically requires stringent technical conditions and performs poorly in finite samples; information criteria and eigenvalue-ratio methods are well documented to under- or over-estimate r in finite sample, depending on signal strength.

To avoid this instability, we adopt *fixed-order PCA*. Rather than attempting consistent recovery of r, the statistician specifies an integer R (the *working number of factors*) and imposes only

$$R \geq r.$$

The estimator is then constructed from the first R empirical singular components of X, regardless of whether R matches the true factor dimension.

The requirement $R \geq r$ is still substantive: one needs a credible upper bound on the factor dimension, supplied for example by a conservative screen from information criteria, or by scientific constraints on the number of latent channels. When no such upper bound is available, fixed-order PCA should be viewed as a sensitivity analysis over a small grid of plausible R's rather than as a replacement for factor-number learning.

From a machine-learning perspective, the working dimension R is the tuning parameter of fixed-order PCA. The main results of this section can therefore be read as a robustness-to-tuning-parameter statement: valid inference holds for any fixed R within a range bounded below by the unknown r and bounded above by a constant, without data-driven optimization of R.

Write the singular value decomposition of X as

$$X = \widehat{\Xi}_R \widehat{L}_R \widehat{V}_R' + \widehat{\Xi}_{-R} \widehat{L}_{-R} \widehat{V}_{-R}',$$

where $\widehat{\Xi}_R \in \mathbb{R}^{N \times R}$, $\widehat{V}_R \in \mathbb{R}^{T \times R}$, and $\widehat{L}_R \in \mathbb{R}^{R \times R}$ collect the top-R left singular vectors, right singular vectors, and singular values, and $(\widehat{\Xi}_{-R}, \widehat{L}_{-R}, \widehat{V}_{-R})$ collect the remaining components. The PCA estimators of loadings and factors are

$$\widehat{B} = \sqrt{N} \, \widehat{\Xi}_R, \qquad \widehat{F} = \frac{1}{\sqrt{N}} X' \widehat{\Xi}_R = \frac{1}{\sqrt{N}} \widehat{V}_R \widehat{L}_R,$$

and the low-rank component is estimated by singular-value thresholding,

$$\widehat{M} = \widehat{B}\widehat{F}' = \widehat{\Xi}_R \widehat{L}_R \widehat{V}_R' = X\widehat{V}_R\widehat{V}_R'.$$

In particular, $\widehat{S}_f := T^{-1}\widehat{F}'\widehat{F} = (TN)^{-1}\widehat{L}_R^2$; since the top R singular values of X are positive almost surely under Assumption 2.1, $\widehat{S}_f$ is invertible (almost surely).

A central contribution of this paper is to show that fixed-order PCA still consistently recovers the factor space spanned by B even when R strictly exceeds r. The additional empirical eigencomponents (those beyond the true factor dimension) do not contaminate the recovery of the true factor space; instead, they converge to well-characterized noise-governed directions. The leading r components of fixed-order PCA asymptotically reconstruct the true factor space, while the remaining $R - r$ components behave in a controlled and predictable manner.

2.3 Asymptotic Behavior of the Extra Eigencomponents

The main objective of this section is to establish the consistency of PCA when $R > r$. We partition the empirical eigencomponents as

$$\widehat{\Xi}_R = (\widehat{\Xi}_r, \widehat{\Xi}_{-r}), \qquad \widehat{V}_R = (\widehat{V}_r, \widehat{V}_{-r}),$$

where $\widehat{\Xi}_r$ denotes the leading r left singular vectors, the usual *spiked* eigenvectors, and $\widehat{\Xi}_{-r}$ collects the remaining $R-r$ eigenvectors, which we refer to as the *extra* (or overestimated) eigenvectors. When $R = r$, we set $\widehat{\Xi}_{-r} = \emptyset$. The analogous decomposition applies to $\widehat{V}_R$.

The behavior of the extra eigenvectors is the key obstacle: $\widehat{\Xi}_{-r}$ carries no factor signal, and its asymptotic effect depends delicately on the noise U. For PCA-based inference to remain valid under overspecification, $\widehat{\Xi}_{-r}$ must be incoherent, in the sense that its entries spread approximately uniformly across coordinates rather than concentrating on a few positions; coherent eigenvectors (also known as localized eigenvectors), in contrast, would interact spuriously with deterministic directions of interest. We establish this incoherency for $\widehat{\Xi}_{-r}$ as Theorem 2.1(ii) below.

Assumption 2.1. *The idiosyncratic shocks and their cross-sectional covariance satisfy:*

(i) The entries $e_{i,t}$ of e_t are independent random variables with $\mathbb{E}[e_{i,t}] = 0$ and $\mathbb{E}[e_{i,t}^2] = 1$. In addition, $e_{i,t}$ is subexponential, in the sense that $\|e_{i,t}\|_{\psi_1} := \inf\{K > 0 : \mathbb{E}\exp(|e_{i,t}|/K) \le 2\} < \infty$.

(ii) The eigenvalues of Σ_e lie in $[c, C]$ for constants $0 < c \le C < \infty$ that do not depend on N, and their associate deformed Marchenko–Pastur (MP) has regular edges and bulks, as defined in Definition A.1 in Appendix A.

We require the noise be subexponential-tailed, as defined in Vershynin (2018). Condition (ii) allows us to develop a local-law of the spectrum of X for factor models. The

MP law of U can be determined using its Stieltjes transform. Assuming it has regular edges and bulks, we show that the analysis of X can proceed by leveraging the eigenvalue rigidity, edge spacing, and anisotropic incoherence properties of U established in Knowles and Yin (2017). In part (ii) of this assumption, the required regularity on edges and bulks for the eigenvalues of Σ_e are standard, and we defer the detailed definition to Definition A.1. Here are two examples to satisfy this condition. Let the ordered eigenvalues of Σ_e be $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_N$, and their empirical spectral distribution be $\pi = \frac{1}{N}\sum_{i=1}^{N} \delta_{\sigma_i}$, where δ_{σ_i} denotes the Dirac measure at σ_i.

1. **Discrete limit.** The measure π is supported on K fixed values $\{s_k\}_{k=1}^{K}$, and $\pi(s_k)$ converges to a limit as $T \to \infty$.

2. **Continuous limit.** There exists a measure π_∞ supported on $[a, b]$ whose density is bounded in $[\tau, \tau^{-1}]$ for some $\tau > 0$, and π converges weakly to π_∞ as $T \to \infty$.

Therefore, Assumption 2.1(ii) is satisfied in standard examples: the homoskedastic case $\Sigma_e = I\sigma$, sparse perturbations of the identity, and Σ_e with bounded condition number whose limiting spectral measure has positive density.

Assumption 2.2. *The loadings, factors, and signal strength satisfy:*

(i) The loadings B are deterministic with row norms uniformly bounded in N, and the factors F are random with row norms bounded in probability:

$$\max_{i \leq N} \|b_i\| \leq C, \qquad \max_{t \leq T} \|f_t\| = O_P(1).$$

(ii) $N/T \to \phi$ for some $\phi \in (0, \infty)$ with $\phi \neq 1$.

(iii) There exists a sequence $\nu_M \to \infty$ with $\nu_M = O(\sqrt{N})$ such that

$$c\nu_M\sqrt{T} \leq \lambda_r(M) < \cdots < \lambda_1(M) \leq C\nu_M\sqrt{T},$$

and

$$c \leq \lambda_r(S_f) \leq \cdots \leq \lambda_1(S_f) \leq C, \qquad S_f := \frac{1}{T}F'F.$$

The restriction $\phi \neq 1$ is inherited from Knowles and Yin (2017): the anisotropic local law guarantees incoherence of the singular vectors of U only when the aspect ratio is bounded away from the MP edge. We do not claim that the conclusions fail at or near $\phi = 1$; rather, the available local-law input used here does not provide the required singular-vector incoherence there. To our knowledge, the extent to which this is a proof artifact versus a genuine phase transition for incoherence remains open. Recent work on the MP edge in the regime $\phi \to 1$ (see, e.g., the line of research developed by L. Erdős

and collaborators on rigidity at the soft and hard edges) studied the rate at which ϕ converges to one, and suggested that the natural fluctuation scale changes from the bulk $T^{-1/2}$ to an edge scale of order $T^{-2/3}$; whether this scale is compatible with the entrywise bounds we require for Theorem 2.1 is an interesting question for future work, and would be the natural route to closing the $\phi \neq 1$ gap.

Part (iii) is the factor-strength condition: equivalently,

$$\lambda_k(B'B) \asymp \nu_M^2, \quad k = 1, \ldots, r.$$

The sequence ν_M indexes the signal strength and governs the sharpness of all subsequent results. At one extreme, $\nu_M \asymp \sqrt{N}$ recovers the strong-factor setting of Bai (2003); at the other, the results remain informative as long as $\nu_M \to \infty$, which is considerably weaker.

The role of ν_M across the main results is as follows. The spectral statements in Theorem 2.1(ii)–(iii) require only $\nu_M \to \infty$ for the first r singular values to be separated from the noise, while the factor-space recovery of Theorem 2.4 adds the mild $T^{\varepsilon} = o(\nu_M)$ requirement, which is satisfied by any polynomial rate for $\nu_M \to \infty$. Furthremore, the inference theorem strengthens this to the product-of-rates condition $\sqrt{T} = o(\nu_M^2)$ in Assumption 3.1(ii), the analogue of the DML rate condition.

The factor-strength part of Assumption 2.2 is imposed only when $r \geq 1$. Meanwhile we allow a special case $r = 0$, that is, there is no factor present so $X = U$. In this case, statements about extra components reduce to the corresponding noise-only local-law statements, which we state in Corollary 2.2 below. This case is interesting in applications where statisticians are concerned about the impact of confounding factors but are not sure whether they are present. Therefore, allowing $r = 0$ as a special case makes the PCA-based inference be also robust to whether confounding factors are present.

The following theorem is the analytical foundation of the paper. It shows that the extra spectrum closely resembles the spectrum of the noise matrix, and that the extra eigenvectors are incoherent and nearly orthogonal to the factor directions. The theorem is stated for $r \geq 1$, the setting of direct interest for factor-model inference, whereas the case $r = 0$ is recorded as Corollary 2.2 below.

In the theorem we use the notion of stochastic domination $\prec$ to describe the rate of convergence, which is slightly stronger than the notion of $O_P(\cdot)$. While the definition is given at the end of Section 1, roughly speaking $X_T \prec a_T$ means $X_T = O_P(T^{\varepsilon} a_T)$ for an arbitrarily small $\varepsilon > 0$.

Theorem 2.1 (Extra spectrum, $R > r$)**.** *Under Assumptions 2.1 and 2.2 with $r \geq 1$, the bounds in (i)–(iii) below all hold uniformly in $k \in \{1, \ldots, R - r\}$ on a single high-probability event.*

(i) Singular values. *For any* $R > r \geq 1$ *and* $k = 1, \ldots, R - r$,

$$\lambda_k(U)^2 \geq \lambda_{r+k}(X)^2, \qquad \lambda_k(U)^2 - \lambda_{r+k}(X)^2 \prec \nu_M^{-2} T.$$

(ii) Incoherence of extra eigenvectors. *For any sequence of unit vectors* $\{\eta_i, \zeta_t : i \leq N, t \leq T\}$, *which are independent of* U, $\|\eta_i\| = 1$ *and* $\|\zeta_t\| = 1$, $i = 1...N$, $t = 1, ..., T$,

$$\max_{i \leq N} \|\eta_i' \widehat{\Xi}_{-r}\| \prec \nu_M^{-1}, \qquad \max_{t \leq T} \|\zeta_t' \widehat{V}_{-r}\| \prec \nu_M^{-1}.$$

(iii) Near-orthogonality with factors.

$$\|B\|_{\mathrm{F}}^{-1} \|B' \widehat{\Xi}_{-r}\| \prec \nu_M^{-2}, \qquad \|F\|_{\mathrm{F}}^{-1} \|F' \widehat{V}_{-r}\| \prec \nu_M^{-2}.$$

The theorem admits the following geometric interpretation. The signal matrix M contributes r singular values of order $\nu_M \sqrt{T}$, while the noise matrix U has Marchenko–Pastur spectrum on the order of $\sqrt{T}$. When $\nu_M \gg 1$, the empirical singular spectrum of X exhibits the so-called "canonical outlier-plus-bulk pattern": r spike outliers at scale $\nu_M \sqrt{T}$, well separated from a Marchenko–Pastur bulk concentrated at scale $\sqrt{T}$. They contain essentially all of the factor information in X. In addition, the next $R-r$ empirical singular vectors $\widehat{\Xi}_{-r}$ of X lie in the orthogonal complement of this signal subspace and are therefore noise-driven; their geometry is governed by the local law for U. The contribution of the signal to these extra subspaces is a second-order effect, captured by the ν_M^{-2} rate in Theorem 2.1(iii).

Part (ii) of Theorem 2.1 implies $\|\eta' \widehat{\Xi}_{-r}\| \xrightarrow{P} 0$ for every unit vector η independent of U, at rate ν_M^{-1}. This is precisely the incoherence condition: were $\widehat{\Xi}_{-r}$ sparsely concentrated, say at the first coordinate, then taking $\eta = (1, 0, \ldots, 0)'$ would produce $\eta' \widehat{\Xi}_{-r} = 1$, contradicting the conclusion. In other words, even though the extra eigenvectors carry no signal, they are diffuse rather than coherent.

This theorem leads to three statistical insights. First, result (ii) imply the ℓ_∞ perturbation bounds for $R > r$, by setting η and ζ as canonical basis vectors:

$$\|\widehat{\Xi}_{-r}\|_\infty = \max_k \|\eta_k' \widehat{\Xi}_{-r}\| = O_P(\nu_M^{-1})$$

where $\eta_k' = (0, .., 0, 1, 0...)$, taking value one on its k th element (similarly we have bounds for $\|\widehat{V}_{-r}\|_\infty$). The related results on deterministic ℓ_∞ perturbation arguments of Fan et al. (2018) and Abbe et al. (2020) only apply to the first r eigenvectors, but not for $\widehat{\Xi}_{-r}$ or $\widehat{V}_{-r}$. The technical challenge in extending $R = r$ to $R > r$ is that the extra singular vectors correspond to singular values without clear separation from neighboring singular values, so the standard eigen-gap arguments do not apply to them. Our Theorem 2.1(ii) leverages the random-matrix local law to analyze the entrywise structure of the extra

singular vectors.

Secondly, the improvement from ν_M^{-1} in (ii) to ν_M^{-2} in (iii) is genuine: the extra eigenvectors are driven by U and thus nearly orthogonal to the factor space. The gap between the signal and noise singular values is what drives the fast rate of convergence ν_M^{-2}. The rate is also strictly sharper than what a Davis–Kahan / $\sin\Theta$ argument would deliver. The standard $\sin\Theta$ bound for a perturbed singular subspace yields only $O_P(\nu_M^{-1})$, with no improvement for the eigenvectors orthogonal to factors. By contrast, an arbitrary η in (ii) carries no a priori alignment with the signal block, so only the leading-order incoherence is available, which explains why the rate in (ii) is slower.

Lastly, each extra empirical singular vector, denoted by $\widehat{\xi}_k$ as the kth column of $\widehat{\Xi}_R$ for $k > r$, admits the orthogonal decomposition

$$\widehat{\xi}_k = \Xi_r x_k + \Xi_c y_k$$

where the signal-aligned coefficient $x_k \in \mathbb{R}^r$ shrinks at the strictly faster rate $\|x_k\| = O_P(\nu_M^{-2})$ while the noise-aligned coefficient $y_k \in \mathbb{R}^{N-r}$ inherits the entrywise incoherency of the underlying noise singular vectors. The difference in the rates, ν_M^{-1} for arbitrary deterministic directions in (ii) versus ν_M^{-2} for the factor block in (iii), is the geometric reason why over-estimation in PCA is asymptotically benign and robust for inference.

As a statistical application, Theorem 2.1 is especially useful for inference with over-estimated factors, as formalized below.

Corollary 2.1 ($r \geq 1$)**.** *Suppose the assumptions of Theorem 2.1 hold. Let G_N and G_T be any $N \times K$ and $T \times K$ matrices with $K = O(1)$, $\|G_N\|_{\mathrm{F}} + \|G_T\|_{\mathrm{F}} = O_P(\sqrt{T})$, and G_N, G_T independent of U. Then*

$$\frac{1}{NT}\|\widehat{B}'UG_T\| + \frac{1}{NT}\|\widehat{F}'U'G_N\| = O_P(T^{-1/2}\,\nu_M^{-1}).$$

To see that the rate in Corollary 2.1 is sharp, consider the strong-factor case $\nu_M = \sqrt{T}$. Then $(NT)^{-1}\|\widehat{F}'UG_N\| = O_P(T^{-1})$, which is the same rate as if the true factors were used $(NT)^{-1}\|F'UG_N\| = O_P(T^{-1})$. So the price of replacing the population factor by its PCA estimator does not introduce extra first order variance. Such a statistical corollary plays a central role in Section 3 for factor-augmented inference.

Lastly, the corollary below specifies the special case that $r = 0$:

Corollary 2.2 (Boundary case $r = 0$)**.** *Suppose Assumption 2.1 holds, $N/T \to \phi \neq 1$, and $r = 0$, so that $X = U = \Sigma_e^{1/2}E$. Then for every bounded integer R and any unit-norm vectors $\eta \in \mathbb{R}^N, \zeta \in \mathbb{R}^T$ independent of U,*

$$\|\eta'\widehat{\Xi}_R\| \;\prec\; T^{-1/2}, \qquad \|\zeta'\widehat{V}_R\| \;\prec\; T^{-1/2}.$$

Each of the leading R empirical singular vectors of X is therefore incoherent with respect to any direction independent of U at the standard noise-only rate. The bound follows from a more general result, which we state as Proposition C.4 in the appendix: when $r = 0$, the empirical singular vectors of X coincide with those of U, and no factor-strength condition is needed.

2.4 Asymptotic Behavior of the Overestimated Factor Space

We next turn to estimation of the low-rank signal and the factor space.

Theorem 2.2 (Low-rank recovery, $r \geq 1$)**.** *Under the assumptions of Theorem 2.1, for each fixed $R \geq r$ with $R - r$ bounded,*

$$\frac{1}{\sqrt{NT}} \|\widehat{M} - M\|_{\mathrm{F}} = O_P\big(T^{-1/2}\big).$$

The bound is uniform over $R \in \{r, r+1, \ldots, \bar{R}\}$ for any fixed $\bar{R} \geq r$.

Theorem 2.2 shows that the low-rank component is recovered at the standard parametric rate, uniformly over any bounded working dimension $R \geq r$. Two features of this statement are worth noting. The rate $T^{-1/2}$ matches the optimal rate achievable by a hypothetical oracle estimator that knows r: overestimation by a bounded amount does not slow down the rate of convergence for low-rank recovery. The extra components $\widehat{\Xi}_{-r}\widehat{L}_{-r}\widehat{V}'_{-r}$ included in $\widehat{M}$ have Frobenius norm of order $\sqrt{T}$, but Theorem 2.1(iii) ensures that this contribution is aligned with the noise direction rather than the signal direction. While alternative estimators based on hard-thresholding, soft-thresholding, or nuclear-norm minimization can also achieve parametric recovery, fixed-order PCA does so without any tuning parameter beyond the integer R itself; this is useful in settings where cross-validation is unstable or computationally expensive.

We now address estimation of the factor space itself. In the classical case $R = r$, PCA consistently recovers F up to an invertible $r \times r$ rotation. When $R > r$, the notion of rotation must be generalized, and we introduce two natural extensions.

Expanded rotation. Define the $r \times R$ matrix

$$H' = \frac{1}{N} B' \widehat{B}.$$

The model $X = BF' + U$ immediately yields

$$\widehat{F} = FH' + \frac{1}{N} U' \widehat{B}. \tag{2.2}$$

Up to the statistical error $N^{-1}U'\widehat{B}$, the true factors F are *expanded* by H' into the larger space spanned by $\widehat{F}$.

Compressed rotation. Let $H^+ = (HH')^+ H$ denote the $R \times r$ Moore–Penrose inverse

of H. Post-multiplying (2.2) by H^+ and using $H'H^+ = I_r$, valid on the high-probability event $\{\lambda_r(H) > 0\}$, gives

$$\widehat{F}H^+ = F + \frac{1}{N} U' \widehat{B}\, H^+. \tag{2.3}$$

Thus $\widehat{F}$ is *compressed* by H^+ back to an object of the true dimension.

The two rotations serve different purposes. The expanded rotation H is appropriate when the goal is to align the columns of $\widehat{F}$ with F *without* forcing a dimension match; this formulation applies when $\widehat{F}$ enters downstream as a linear regressor, since regression projects onto a span and is invariant to the choice of basis within that span. Let $\mathrm{col}(F)$ denote the linear space spanned by the columns of F. The identity (2.2) exhibits the geometric content of the expansion: the r-dimensional factor span $\mathrm{col}(F)$ is approximately contained in the R-dimensional fitted span $\mathrm{col}(\widehat{F})$, with H' mapping a basis of the former into the latter.

The compressed rotation H^+ is appropriate when the goal is to identify the r true factor directions *within* the larger R-dimensional fitted space; this formulation applies when the downstream object is an r-column object explicitly referencing the true factor dimension rather than the working dimension R. The identity (2.3) performs the inverse extraction of $\mathrm{col}(F)$ from inside $\mathrm{col}(\widehat{F})$; The two notions coincide when $R = r$, in which case both reduce to the standard $r \times r$ rotation matrix of the classical PCA literature.

The next proposition records the non-degeneracy scale of H uniformly over $R \geq r$. We write $\nu_{\min} := \lambda_r(S_B)$ when $r \geq 1$, so that $\nu_{\min} \asymp \nu_M^2/N$.

Proposition 2.3 (Non-degeneracy of the rotation)**.** *Under the assumptions of Theorem 2.1 with $r \geq 1$, there exists a constant $c_0 > 0$ depending only on (c, C, ϕ) such that, for every bounded $R \geq r$,*

$$\mathrm{P}\{\lambda_r(H) \geq c_0 \nu_{\min}^{1/2}\} \to 1, \qquad \|H^+\| = O_P(\nu_{\min}^{-1/2}).$$

Under strong factors, $\nu_{\min} \asymp 1$, so Proposition 2.3 recovers the familiar bounded-inverse behavior $\|H^+\| = O_P(1)$. Under weak factors, the inverse rotation carries the additional scale $\nu_{\min}^{-1/2}$; this scale slows down the compressed-rotation convergence rate, as stated below.

Theorem 2.4 (Factor space, $r \geq 1$)**.** *Suppose the assumptions of Theorem 2.1 hold and there exists $\varepsilon > 0$ with $T^\varepsilon = o(\nu_M)$. Then for every bounded $R \geq r$:*

(i) Expanded rotation.

$$\frac{1}{\sqrt{T}} \|\widehat{F} - FH'\| = O_P(T^{-1/2}), \qquad \frac{1}{T} \left\| F'(\widehat{F} - FH') \right\| \prec T^{-1/2}\, \nu_M^{-1}.$$

(ii) Compressed rotation.

$$\frac{1}{\sqrt{T}}\|\widehat{F}H^{+}-F\| = O_P(\nu_M^{-1}), \qquad \frac{1}{T}\|F'(\widehat{F}H^{+}-F)\| \prec \nu_M^{-2}.$$

(iii) Inverse factor covariance.

$$H'\left(\frac{1}{T}\widehat{F}'\widehat{F}\right)^{-1}H = \left(\frac{1}{T}F'F\right)^{-1} + o_P(1).$$

We use the ν_M form here for direct comparison with (ii) and Bai (2003). When $\nu_M = \sqrt{T}$ (strong factors) both $\frac{1}{\sqrt{T}}\|\widehat{F} - FH'\|$ and $\frac{1}{\sqrt{T}}\|\widehat{F}H^{+} - F\|$ converge at $T^{-1/2}$. When ν_M is slower than $\sqrt{T}$ (weaker factors), the compressed rate in (ii) is slower than the expanded rate in (i), illustrating that recovering the true factor space is a more difficult problem than aligning the true factor space in the over-estimated space.

In addition, Part (iii) of Theorem 2.4 shows that the inverse estimated factor covariance, properly sandwiched by H, is consistent for the true inverse factor covariance regardless of the working dimension $R \geq r$. These inverse covariance matrices are often used for factor-augmented inference when estimated factors are being used as regressors. Hence, result (iii) is the property on which the robust factor-augmented inference of the next section relies.

3 Application to Treatment-Effect Inference

As an illustrative application of the spectral results in Section 2, we consider inference on a treatment coefficient β in the factor-augmented system

$$\begin{aligned} y_t &= \mu_y + \beta\, g_t + \rho' f_t + \eta_t, \\ g_t &= \mu_g + \alpha_g' f_t + \varepsilon_{g,t}, \\ z_t &= \mu_z + \alpha_z' f_t + \varepsilon_{z,t}, \end{aligned} \tag{3.1}$$

where g_t is the treatment variable of interest. We consider the application where g_t is correlated with η_t (so that it is endogenous). Then we include an observed instrumental variable (IV) z_t, and $\mathbb{E}(\eta_t|\varepsilon_{z,t}, f_t) = 0$. The latent factors f_t are not directly observed; instead we observe the high-dimensional panel of controls

$$x_t = Bf_t + u_t, \tag{3.2}$$

exactly as in Section 2.1, so that f_t can be recovered by PCA from X. The intercepts (μ_y, μ_g, μ_z) are unrestricted nuisance parameters that absorb marginal means of (y_t, g_t, z_t) at no asymptotic cost. The innovations $(\eta_t, \varepsilon_{g,t})$ are allowed to be mutually correlated within a period.

This setup places the application within the partialling-out / Frisch–Waugh–Lovell / Neyman-orthogonal residualized-regression tradition of DML (Chernozhukov et al., 2016), with the distinguishing feature that the high-dimensional nuisance is the latent factor space spanned by F, recovered by PCA from the panel X, rather than a function of observed controls. The new content of this section is to show that fixed-order PCA is a valid nuisance estimator, delivering $\sqrt{T}$-asymptotic normality with the unadjusted Eicker–White sandwich variance, without sample splitting or data-driven selection of R.

Let $P_A = A(A'A)^{-1}A'$ be the projection matrix of A. First, we apply PCA to X to extract R factors, whose $T \times R$ matrix is denoted by $\widehat{F}$, as in Section 2.2. Then we use the factor-augmented IV estimator of β, which is based on the residualized just-identified IV regression:

$$\widehat{\beta} \;=\; \left(\widehat{\varepsilon}_z'\,\widehat{\varepsilon}_g\right)^{-1}\widehat{\varepsilon}_z'\,\widehat{\varepsilon}_y, \tag{3.3}$$

with $\widehat{\varepsilon}_y = (I - P_{[1_T,\widehat{F}]})Y$, $\widehat{\varepsilon}_g = (I - P_{[1_T,\widehat{F}]})G$, and $\widehat{\varepsilon}_z = (I - P_{[1_T,\widehat{F}]})Z$ denoting residuals from regressing the outcome $Y = (y_1, \ldots, y_T)'$, the treatment $G = (g_1, \ldots, g_T)'$, and the instrument $Z = (z_1, \ldots, z_T)'$ on a constant and the PCA factor estimator. The projection on $[1_T, \widehat{F}]$ absorbs the unobserved intercepts together with the latent-factor nuisance $\rho' f_t$, and $\widehat{\beta}$ is the second-stage IV slope on the residualized variables.

The key innovation of our estimator is that our analysis does not require consistent estimation of the true factor number; it suffices that $R \geq r$ so that the span of $\widehat{F}$ cover the span of F. The key structural assumption is that $(\eta_t, \varepsilon_{g,t}, \varepsilon_{z,t})$ is i.i.d. across t and independent of (F, U), formalized in Assumption 3.1(i) below. Under i.i.d. errors, the autocovariances of the score $\varepsilon_{z,t}\eta_t$ vanish, so the long-run variance reduces to the contemporaneous expectation and the heteroskedasticity-only Eicker–White (HC_0) sandwich is the natural variance estimator; the analysis allows arbitrary contemporaneous heteroskedasticity in the score, but not serial dependence or factor-driven conditional variance dynamics.

Remark 1 (OLS as a special case)**.** In the special case that the treatment variable g_t is uncorrelated with η_t, our method collapses to the residual based on OLS estimator by setting $z_t = g_t$. Then $\widehat{\beta}$ is similar to the OLS estimator analyzed in Belloni et al. (2014), but with the high-dimensional Lasso step replaced by the PCA step.

Assumption 3.1. *The regression and instrument errors and signal strength satisfy:*

(i) $(\eta_t, \varepsilon_{g,t}, \varepsilon_{z,t})$ *is independent and identically distributed across* t*, independent of* (F, U)*, with* $\mathbb{E}[\eta_t] = \mathbb{E}[\varepsilon_{g,t}] = \mathbb{E}[\varepsilon_{z,t}] = 0$*, the exclusion restriction* $\mathbb{E}[\eta_t\,\varepsilon_{z,t}] = 0$*, and the relevance condition* $\gamma := \mathbb{E}[\varepsilon_{g,t}\,\varepsilon_{z,t}] \neq 0$*.*

(ii) $\sqrt{T} = o(\nu_M^2)$*.*

(iii) $\mathbb{E}|\eta_t|^8 + \mathbb{E}|\varepsilon_{g,t}|^8 + \mathbb{E}|\varepsilon_{z,t}|^8 \leq C$*, with* $\mathbb{E}[\eta_t^2] > 0$ *and* $\mathbb{E}[\varepsilon_{z,t}^2] > 0$*.*

Condition (ii) is the factor-strength requirement relevant for inference; it is equivalent to $\lambda_r(B'B) \gg T^{1/2}$ and is weaker than the standard strong-factor assumption. Condition (iii) is a uniform moment bound that supports both Lyapunov's central limit theorem for $\sqrt{T}(\widehat{\beta} - \beta)$ and consistency of the heteroskedasticity-robust sandwich variance estimator. When $z_t = g_t$, the relevance condition $\gamma = \mathbb{E}[\varepsilon_{g,t}^2] \neq 0$ holds automatically and the exclusion $\mathbb{E}[\eta_t \, \varepsilon_{z,t}] = 0$ becomes the OLS exogeneity $\mathbb{E}[\eta_t \, \varepsilon_{g,t}] = 0$.

Theorem 3.1. *Suppose Assumptions 2.1, 2.2, and 3.1 hold, with the convention that Assumption 3.1(ii) imposes no condition when $r = 0$ since ν_M is undefined in that case. Define the residual $\widehat{\eta}_t = \widehat{\varepsilon}_{y,t} - \widehat{\beta}\,\widehat{\varepsilon}_{g,t}$ and the Eicker–White (HC_0) variance estimator*

$$\widehat{\sigma}^2 \;=\; \big(\widehat{\varepsilon}_z'\widehat{\varepsilon}_g/T\big)^{-1} \left(\tfrac{1}{T}\sum_{t=1}^{T} \widehat{\varepsilon}_{z,t}^2\, \widehat{\eta}_t^2\right) \big(\widehat{\varepsilon}_z'\widehat{\varepsilon}_g/T\big)^{-1}.$$

Then for every fixed (bounded) integer R with $0 \le r \le R$,

$$\sqrt{T}\,\widehat{\sigma}^{-1}(\widehat{\beta} - \beta) \;\xrightarrow{\;d\;}\; \mathcal{N}(0,1).$$

The estimator $\widehat{\sigma}^2$ is consistent for the just-identified IV sandwich variance

$$\sigma^2 \;=\; \gamma^{-2}\,\mathbb{E}[\varepsilon_{z,t}^2\eta_t^2], \qquad \gamma \;=\; \mathbb{E}[\varepsilon_{g,t}\,\varepsilon_{z,t}].$$

In the OLS specialization $z_t = g_t$, this reduces to $\sigma^2 = \mathbb{E}[\varepsilon_{g,t}^2]^{-2}\,\mathbb{E}[\varepsilon_{g,t}^2\eta_t^2]$, the partialled-out OLS sandwich.

We briefly discuss the variance-vs-bias tradeoff behind Theorem 3.1. Including $R - r$ extra principal components removes $R - r$ extra degrees of freedom from the residualized regression, creating a finite-sample degrees-of-freedom cost of the familiar order $T/(T - R - 1) = 1 + O(R/T)$; this is asymptotically negligible for fixed R but should be read as a finite-sample variance cost when R is moderate relative to T, as documented in Section 4. Overestimation does not induce first-order bias: by Theorem 2.1(iii), the overestimated principal directions are asymptotically orthogonal to the factor space, so they do not remove confounding signal.

Moving on to the variance, note that if F were observed, the score would be $T^{-1/2}\sum_t \varepsilon_{z,t}\eta_t$. Replacing F by $\widehat{F}$ introduces additional terms involving the residual errors and the estimated projection $P_{\widehat{F}}$. Because $\widehat{F}$ is a function of (F, U) and the regression errors are independent of (F, U), these terms can be controlled conditionally on the factor-estimation sample. The residual bounds in the supplement show that the effect of estimating the factor space is $o_P(1)$ after $\sqrt{T}$ normalization whenever $\sqrt{T} = o(\nu_M^2)$. Thus the feasible residualized score has the same first-order limit as the infeasible score based on the true factor space. Therefore, the net effect is no first-order bias and only a mild finite-sample efficiency cost relative to the consequences of underestimation.

Theorem 3.1 does not prescribe how large R should be in finite samples; we defer concrete guidance to Section 4, where variance inflation as a function of R/T is examined empirically. A working dimension chosen slightly above the value returned by an information criterion or an eigenvalue-ratio test is a natural conservative default. In addition, we allow $R > r = 0$ as a special case, which is the case when it is an unknown matter of fact to statisticians that there are no confounding factors. The result shows that it does not hurt to estimate $R > 0$ "factors" in this case.

Remark 2 (Connection to related work)**.** In canonical DML, cross-fitting is used to weaken the dependence between the estimated nuisance and the score evaluated on the same observations. Here that dependence is structurally absent: $\widehat{F}$ is a measurable function of (F, U). By Assumption 3.1, (F, U), and therefore $\widehat{F}$, are independent of the regression and instrument errors, so no sample splitting is needed.

The closest non-DML antecedent is Moon and Weidner (2015), who establish robustness-to-overspecification for interactive-fixed-effects panel models via Kato perturbation of a profile-likelihood objective. Two further contrasts are worth noting beyond the rate comparison given in Section 2.3. First, the operator-perturbation route is more flexible across estimator families (e.g. quasi-MLE, GMM), while the local-law route specializes to PCA but yields sharper rates when applicable. Second, the variance-sandwich consistency in Theorem 2.4(iii) that underlies our unadjusted-sandwich CLT does not appear to be available from operator-perturbation alone.

4 Simulations

We report Monte Carlo evidence on the finite-sample behavior of the fixed-order PCA inference of Section 3, focusing on the OLS specialization $z_t = g_t$ (Remark 1). The data-generating process has $i = 1, \ldots, N$ and $t = 1, \ldots, T$, intercepts $\mu_g = 2$ and $\mu_y = 3$, and $\beta = 0$. Loadings are generated sparsely:

$$b_{i,k} \sim \mathcal{N}(0,1) \text{ with probability } p_N = N^{-\alpha}, \qquad b_{i,k} = 0 \text{ otherwise},$$

so a non-zero $\alpha \in [0, 1)$ controls the strength of the latent factors: at $\alpha = 0$ the loadings are dense and $\nu_M \asymp \sqrt{N}$ (the classical strong-factor case of Bai (2003)); as α grows the non-zero proportion $N^{-\alpha}$ shrinks and $\nu_M \asymp N^{(1-\alpha)/2}$ approaches the boundary $\nu_M \to \infty$ admitted by Assumption 2.2(iii). Idiosyncratic errors are $u_t = \Sigma_e^{1/2} e_t$ with $e_{i,t}$ i.i.d. $\mathcal{N}(0,1)$ across both i and t, exactly as required by Assumption 2.1(i), and $\Sigma_e = \mathrm{diag}(D)$ with $D_{ii} \sim \mathrm{Uniform}(0.5, 1.5)$. The regression errors $\varepsilon_{g,t}$, η_t, the factors $f_{k,t}$, the loadings ρ_k, and $\alpha_{g,k}$ are all i.i.d. standard normal and mutually independent.

For each replication, $\widehat{\beta}$ is computed by partialling out $[1_T, \widehat{F}]$ from y_t and g_t and running OLS on the resulting residuals, with $\widehat{F}$ from the SVD of X as defined in Section 3;

standard errors are Eicker–White HC_0. We report the standardized statistic

$$t_\beta = \sqrt{T}\,\frac{\widehat{\beta} - \beta}{\widehat{\sigma}},$$

and assess the closeness of its empirical distribution under the null $\beta = 0$ to the standard-normal benchmark implied by Theorem 3.1. The tables below report, across 1,000 Monte Carlo replications per cell, the sample mean and sample standard deviation (sd) of t_β, its 0.025 and 0.975 quantiles, and the Kolmogorov–Smirnov (KS) p-value against $\mathcal{N}(0,1)$. The Monte Carlo standard error of the sample sd at this resolution is roughly $1/\sqrt{2{,}000} \approx 0.022$.

Choice of (N, T, α) The theoretical conditions of Section 3 impose three joint requirements: (a) Assumption 2.2(ii) requires $N/T \to \phi$ with $\phi \in (0, \infty)$ and $\phi \neq 1$, so N and T should grow proportionally and the aspect ratio should be bounded away from one; (b) Assumption 3.1(ii) imposes the rate condition $\sqrt{T} = o(\nu_M^2)$, which under the sparse-loading design becomes $T = o(N^{2-2\alpha})$ and binds non-trivially when α is close to 1/2; (c) the working dimension R should be a fixed bounded overestimate of the true r. We therefore organize the evidence around three axes that map to (a)–(c), with $N = 200$ throughout:

- Experiment 1 fixes $\alpha = 0$ (strong factors, so the rate condition is slack) and varies $T \in \{100, 400, 800\}$, giving $N/T \in \{2, 0.5, 0.25\}$, all bounded away from one. This isolates the role of the aspect ratio.

- Experiment 2 fixes $T = 400$ (so $N/T = 0.5$, well inside the theory) and varies $\alpha \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$, sweeping factor strength from the classical regime toward the rate boundary.

- Experiment 3 sets $r = 0$ (no factors), so the inference reduces to the partialled-out boundary case, and varies $T \in \{100, 400, 800\}$.

Experiment 1: varying the aspect ratio N/T Table 1 reports the diagnostics for $R \in \{1, 2, 3, 6, 12, 30\}$, spanning under-specification ($R < r$), correct specification ($R = r$), and over-specification. The contrast across R is sharp. When $R < r$, the sd of t_β explodes to between 4 and 13 and KS rejects at every T. Once $R \geq r$, the sd drops to between 1.01 and 1.13 and KS p-values typically exceed 0.4; at $T \in \{400, 800\}$ the standard-normal approximation is quite accurate across the whole $R \geq r$ range, and at $T = 100$ it degrades only once R approaches $T/3$, consistent with the intuition of variance-inflation. The aspect-ratio behaviour is symmetric: the $T = 100$ ($N/T = 2$) and $T = 400$ ($N/T = 0.5$) columns behave equivalently, in line with the symmetry of Assumption 2.2(ii) about $\phi = 1$. Figure 1 displays the corresponding histograms for

$R \in \{2, 3, 12\}$, with the $R = 2$ column ($R < r$) wildly diffuse and most mass off the plotting window.

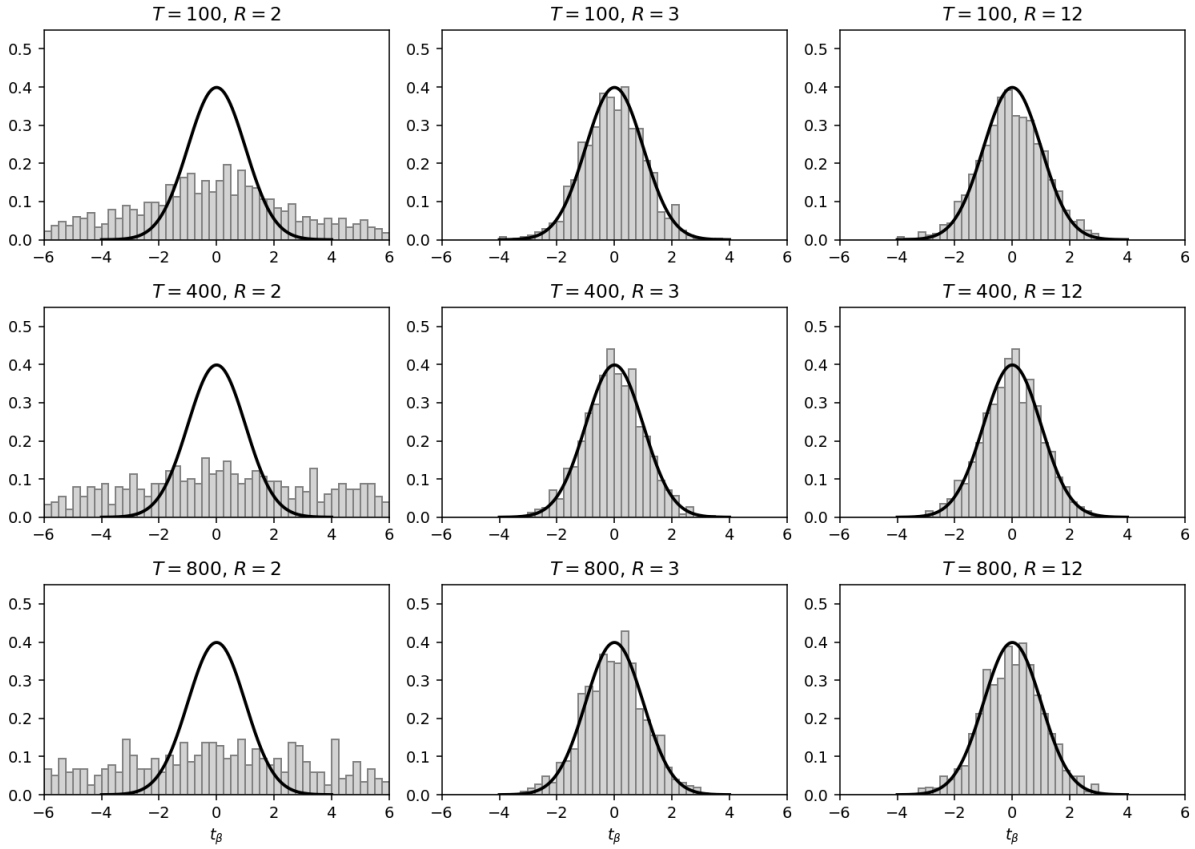

Figure 1: Empirical density of t_β across 1,000 Monte Carlo replications under the design of Experiment 1 ($r = 3$, $\alpha = 0$, $N = 200$). Rows: $T \in \{100, 400, 800\}$. Columns: $R \in \{2, 3, 12\}$. Black curves are the standard-normal density. The $R = 2$ column ($R < r$) is wildly diffuse and most of its mass lies outside the plotting window.

Experiment 2: varying the factor strength α Table 2 reports the diagnostics across α and $R \in \{1, 2, 3, 6, 12\}$. As α grows, $\nu_M \asymp N^{(1-\alpha)/2}$ shrinks, so the rate condition $\sqrt{T} = o(\nu_M^2)$ becomes $T = o(N^{2-2\alpha})$; the rate-condition margin $N^{2-2\alpha}/T$ decays as α increases. The simulations track this scale sharply. Under-specified rows ($R \in \{1, 2\}$) behave as in Experiment 1, with sd of t_β between 8 and 10 across all α. Once $R \geq r$, the sd becomes a clean monotone function of α: 1.07–1.08 at $\alpha \in \{0.0, 0.1\}$ (rate condition slack, KS $p \gtrsim 0.25$), 1.20 at $\alpha = 0.2$, 1.31 at $\alpha = 0.3$, and 1.83 at $\alpha = 0.4$, with 0.025/0.975 quantiles widening to ± 3.7. The role of the rate condition $\sqrt{T} = o(\nu_M^2)$ is therefore not a proof artifact: it sharply governs the finite-sample quality of the standard-normal approximation.

Experiment 3: boundary case ($r = 0$) Theorem 3.1 is stated for $r \geq 1$; Table 3 shows that the same Gaussian behavior extends to the no-factor boundary. At $T \in \{400, 800\}$ the standard-normal approximation is essentially exact for every $R \in$

Table 1: Experiment 1, $r = 3$, $\alpha = 0$. Each entry is the sample mean and sample standard deviation (sd) of t_β across 1,000 Monte Carlo replications, the 0.025 and 0.975 sample quantiles ($\alpha_{.025}$, $\alpha_{.975}$), and the p-value of the Kolmogorov–Smirnov test of the empirical distribution of t_β against $\mathcal{N}(0,1)$. The standard-normal benchmarks for the quantiles are ± 1.96.

T	R	mean	sd	$\alpha_{.025}$	$\alpha_{.975}$	KS p
100	1	-0.11	5.19	-10.09	10.25	$< 10^{-4}$
	2	-0.21	4.25	-9.50	9.17	$< 10^{-4}$
	3	-0.02	1.06	-2.09	2.12	0.475
	6	-0.02	1.09	-2.18	2.21	0.335
	12	-0.02	1.12	-2.26	2.16	0.153
	30	-0.01	1.28	-2.68	2.52	0.001
400	1	$+0.07$	9.72	-20.06	18.15	$< 10^{-4}$
	2	-0.05	8.62	-19.42	18.14	$< 10^{-4}$
	3	-0.00	1.01	-2.06	1.96	0.947
	6	$+0.00$	1.01	-2.08	1.95	0.943
	12	$+0.00$	1.02	-2.06	1.97	0.675
	30	-0.00	1.05	-2.02	2.02	0.600
800	1	$+0.29$	13.07	-26.75	24.60	$< 10^{-4}$
	2	$+0.47$	12.05	-24.92	25.19	$< 10^{-4}$
	3	$+0.01$	1.07	-2.24	2.09	0.521
	6	$+0.01$	1.07	-2.20	2.08	0.500
	12	$+0.01$	1.07	-2.25	2.11	0.603
	30	$+0.01$	1.08	-2.21	2.11	0.387

Table 2: Experiment 2, $r = 3$, $T = 400$. Each entry is the sample mean and sample standard deviation (sd) of t_β across 1,000 Monte Carlo replications, the 0.025 and 0.975 sample quantiles, and the KS p-value against $\mathcal{N}(0,1)$.

α	R	mean	sd	$\alpha_{.025}$	$\alpha_{.975}$	KS p
0.0	1	-0.04	9.66	-18.24	19.42	$< 10^{-4}$
	2	-0.06	8.49	-18.47	17.09	$< 10^{-4}$
	3	-0.00	1.07	-2.10	2.10	0.315
	6	$+0.00$	1.07	-2.03	2.14	0.393
	12	$+0.01$	1.08	-2.09	2.13	0.364
0.1	1	$+0.10$	9.66	-20.06	18.72	$< 10^{-4}$
	2	$+0.35$	8.55	-16.75	18.48	$< 10^{-4}$
	3	$+0.04$	1.07	-2.00	2.17	0.253
	6	$+0.04$	1.07	-2.01	2.20	0.076
	12	$+0.04$	1.07	-2.04	2.13	0.077
0.2	1	$+0.21$	9.69	-18.65	19.64	$< 10^{-4}$
	2	$+0.25$	8.33	-15.62	18.41	$< 10^{-4}$
	3	-0.02	1.19	-2.42	2.27	0.027
	6	-0.02	1.20	-2.40	2.30	0.018
	12	-0.02	1.20	-2.47	2.23	0.016
0.3	1	-0.04	9.31	-18.85	17.78	$< 10^{-4}$
	2	$+0.17$	8.02	-17.29	15.20	$< 10^{-4}$
	3	$+0.04$	1.31	-2.66	2.45	$< 10^{-4}$
	6	$+0.03$	1.31	-2.71	2.51	$< 10^{-4}$
	12	$+0.03$	1.31	-2.76	2.54	$< 10^{-4}$
0.4	1	-0.54	9.52	-19.34	19.33	$< 10^{-4}$
	2	-0.16	8.31	-17.13	17.06	$< 10^{-4}$
	3	-0.06	1.83	-3.69	3.70	$< 10^{-4}$
	6	-0.06	1.83	-3.66	3.71	$< 10^{-4}$
	12	-0.06	1.82	-3.65	3.70	$< 10^{-4}$

$\{0, 3, 6, 12, 30\}$ (sd within 1.02–1.10, KS $p \gtrsim 0.4$); at $T = 100$, the approximation deteriorates only when R approaches $T/3$.

Table 3: Experiment 3, $r = 0$. Each entry is the sample mean and sample standard deviation (sd) of t_β across 1,000 Monte Carlo replications, the 0.025 and 0.975 sample quantiles, and the KS p-value against $\mathcal{N}(0, 1)$.

T	R	mean	sd	$\alpha_{.025}$	$\alpha_{.975}$	KS p
100	0	+0.02	1.06	−1.90	2.14	0.795
	3	+0.02	1.09	−1.97	2.17	0.664
	6	+0.01	1.11	−1.94	2.23	0.464
	12	+0.01	1.13	−2.02	2.34	0.088
	30	−0.02	1.28	−2.47	2.73	$< 10^{-3}$
400	0	−0.02	1.05	−2.14	2.06	0.711
	3	−0.02	1.05	−2.16	2.10	0.363
	6	−0.02	1.06	−2.17	2.10	0.698
	12	−0.02	1.07	−2.18	2.09	0.545
	30	−0.02	1.10	−2.22	2.19	0.459
800	0	−0.01	1.02	−2.08	1.98	0.935
	3	−0.01	1.02	−2.07	2.00	0.984
	6	−0.01	1.03	−2.09	2.01	0.980
	12	−0.01	1.04	−2.12	2.02	0.958
	30	−0.01	1.05	−2.16	2.08	0.962

Two operational takeaways follow. Since ν_M is unobserved in practice, the leading singular values of X are a natural diagnostic for the rate condition $\sqrt{T} = o(\nu_M^2)$: if they do not visibly separate from the bulk, the inflation in Table 2 should be expected. And since under-specification is catastrophic while modest over-specification is essentially free, a natural conservative default is to choose R slightly above the value returned by an information criterion or eigenvalue-ratio test, and to interpret unstable estimates across $R \geq \widehat{r}$ as evidence that the factor-strength regime is unfavorable rather than that R is too small.

5 Empirical Application

We illustrate fixed-order PCA on a single cross-section from the Health and Retirement Study (Sonnega et al., 2014), the panel survey at the empirical core of the structural retirement literature including French and Jones (2011), with respondents treated as i.i.d. draws. The substantive question is whether *labor supply at the intensive-and-extensive margin*—the choice variable N_t in the canonical life-cycle model of French and Jones (2011)—affects depressive symptoms after controlling for a low-dimensional latent state of vitality, socioeconomic resources, and underlying mental-health propensity that drives both labor-supply choices and contemporaneous well-being. We apply the partialled-out residualized regression of Section 3.

The key statistical insight from this application (as shown by results presented in Table 4) is the *robustness profile* of the first-order PCA estimator across R. The estimator stays positive and within $\pm 35\%$ of the saturating $R = N$ value across $R \in \{2, \ldots, 28\}$ — exactly the robustness-to-tuning-parameter property the fixed-order PCA framework is designed to deliver. In contrast, a procedure that demands consistent recovery of r would have to commit to a single $\widehat{r}$, whereas our framework certifies inference uniformly over any R in the upper-R tail.

We now provide detailed implementation of this application. The data are from the RAND HRS Longitudinal File 1992–2022 v1.0, restricted to the wave-14 (2018) interview. After complete-case filtering, the sample has $T = 14{,}672$ respondents. Treatment is annual hours worked,

$$g_t \;=\; (\text{hours/week})_t \times (\text{weeks/year})_t \;+\; (\text{2nd-job hours/week})_t \times (\text{2nd-job weeks/year})_t,$$

clipped to $[0, 5{,}000]$ and set to zero for non-workers. The marginal distribution is bimodal: a 62.3% atom at $g_t = 0$ (non-workers, including fully retired respondents and other labor-market non-participants) and a continuous mass concentrated near full-time, with $\mathrm{P}(g_t \in [1{,}000, 2{,}200)) = 20.5\%$ and $\mathrm{P}(g_t \geq 2{,}200) = 10.5\%$. The continuous specification strictly nests the binary retirement indicator: it preserves the extensive margin (the $g_t = 0$ atom) while resolving the intensive-margin variation among bridge-job holders, partial retirees, and full-time workers that the structural retirement literature treats as economically distinct states. The outcome $y_t = \text{cesd}_t$ is the eight-item Center for Epidemiologic Studies Depression score, integer in $[0, 8]$ with higher values indicating more depressive symptoms (sample mean 1.53, sample standard deviation 2.02). The control panel $x_t \in \mathbb{R}^N$ collects $N = 28$ standardized fundamentals, including sex, ethnicity, education, marital and veteran status, chronic-condition and smoking indicators, cognitive scores and household variables. Each measures a distinct dimension of the underlying biopsychosocial state.

A preliminary spectral decomposition of the standardized panel returns shows that there are no single dominant factor: the first three principal components capture roughly 27% of the total variation.

To bring the IV machinery of Section 3 to bear, we use the institutional cutoff at age 62 as the instrument:

$$z_t \;=\; \mathbf{1}\{\text{respondent } t \text{ is at least 62 years old}\}.$$

Age 62 is the early Social Security claiming age and the empirically dominant discontinuity in U.S. retirement hazards. We prefer it to the alternative cutoff at age 65 because it isolates the labor-supply / Social Security channel from the Medicare insurance channel that turns on at 65. Conditional on the latent state f_t, z_t shifts labor supply primarily

through this institutional channel rather than through individual mental-health propensity, supporting the exclusion restriction in Assumption 3.1(i). The first-stage relationship is unambiguously strong: at $R = 0$, regressing g_t on z_t gives a slope of -0.98 thousand hours/year ($t = -57.7$), and the residualized first-stage Wald statistic remains in the range $|t| \in [40, 55]$ across all working dimensions $R \in \{1, \ldots, 28\}$ reported below.[1]

Table 4 reports both the OLS estimator $\widehat{\beta}_{\text{OLS}}$ ($z_t = g_t$) and the IV estimator $\widehat{\beta}_{\text{IV}}$ ($z_t = \mathbf{1}\{\text{age}_t \geq 62\}$) of Theorem 3.1 across $R \in \{0, 1, 2, 3, 5, 7, 10, 15, 20, 28\}$, both with HC_0 standard errors and both reported per a 1,000-hours-per-year increment.

One scope qualification should be flagged before reading the numbers. The i.i.d.-across-respondents assumption used in our theory ignores household clustering present in HRS, where spouses appear as separate observations sharing many of the controls. We report the unweighted HC_0 inference for transparency with our theory but note that a household-clustered standard error would slightly inflate the reported standard errors. With this caveat, the IV estimator under the exclusion restriction in Assumption 3.1(i) does have a causal local-average-treatment-effect (LATE) interpretation in the sense of Imbens and Angrist (1994): $\widehat{\beta}_{\text{IV}}$ identifies the effect of labor supply on depressive symptoms among the subpopulation of compliers, namely respondents whose retirement timing responds to crossing the age-62 institutional threshold. Off-panel confounders that are absorbed neither by $\widehat{F}$ nor by z_t (for example, idiosyncratic preferences and unanticipated household shocks) cannot bias $\widehat{\beta}_{\text{IV}}$ to first order so long as they are uncorrelated with the age-62 cutoff after factor adjustment.

Three patterns are visible. (i) Omitted-factor bias at $R = 0$ is severe: OLS returns $\widehat{\beta} = -0.276$ ($t = -19.7$), the strong negative cross-sectional association between hours worked and depressive symptoms documented in the HRS-based literature (Dave et al., 2008; Mandal and Roe, 2008); once even one principal component is partialled out, the OLS estimate collapses to essentially zero (magnitude below 0.04 across all $R \geq 1$). The unconditioned cross-sectional correlation is almost entirely accounted for by a single factor direction in x_t, and OLS without an instrument provides no informative signal once that factor is removed. (ii) The IV LATE, by contrast, is uniformly *positive* and stable across all $R \geq 1$: $\widehat{\beta}_{\text{IV}}$ ranges between $+0.64$ and $+1.01$, taking the value $+0.660$ ($t = 12.95$) at the saturating $R = N = 28$. The sign flip relative to OLS at $R = 0$ is the textbook signature of selection bias: respondents who endogenously work more hours are mentally healthier on average for unobserved reasons that the controls do not span. Once the age-62 instrument purges this selection, the IV LATE points the other way: among respondents whose labor supply is shifted by the institutional cutoff, an additional 1,000 hours per year of work raises CES-D by roughly 0.66 points (about a third of the outcome's

[1] Estimating with the alternative cutoff $z_t = \mathbf{1}\{\text{age}_t \geq 65\}$ delivers qualitatively identical conclusions across all R: the same sign pattern, IV point estimates within roughly 5% of the age-62 values, and the same significance ordering. We report the age-62 specification as primary because the absence of the Medicare channel makes the just-identified exclusion restriction more defensible.

standard deviation), consistent in sign with the IV-based retirement literature (Bonsang et al., 2012; Coe and Zamarro, 2011; Mazzonna and Peracchi, 2012; Insler, 2014) that finds retirement to be protective for mental health among compliers. (iii) HC_0 standard errors and the first-stage Wald statistic are essentially flat across $R \geq 1$ for both columns.

Table 4: Real-data estimates of the effect of annual labor supply (hours per year, in thousands) on the eight-item CES-D depression score in HRS wave 14 (2018), partialling out the top-R principal components of the standardized $N = 28$-dimensional control panel. Both columns instantiate Theorem 3.1: the OLS column corresponds to the $z_t = g_t$ specialization, and the IV column to $z_t = \mathbf{1}\{\text{age}_t \geq 62\}$. HC_0 standard errors. $T = 14{,}672$.

R	OLS		IV	
	$\widehat{\beta}_{\text{OLS}}$	t	$\widehat{\beta}_{\text{IV}}$	t
0 (no controls)	−0.276	−19.72	+0.331	+8.89
1	−0.019	−1.38	+0.640	+16.27
2	−0.002	−0.15	+0.910	+18.12
3	+0.018	+1.23	+1.005	+19.15
5	+0.013	+0.84	+0.986	+18.87
7	−0.003	−0.23	+0.711	+14.90
10	−0.033	−2.21	+0.646	+13.19
15	−0.032	−2.17	+0.647	+13.28
20	−0.027	−1.81	+0.684	+13.33
28 (saturating)	−0.030	−1.98	+0.660	+12.95

6 Conclusion

This paper develops spectral theory for principal component analysis in factor models when the working number of factors R is fixed and weakly dominates the true, unknown factor dimension r. Leveraging anisotropic local laws from random matrix theory, we show that the overestimated empirical eigencomponents are noise-governed, incoherent, and near-orthogonal to the factor space; that the low-rank signal and factor space are recovered at the usual parametric rates under suitably generalized (expanded and compressed) rotations; and that factor-augmented inference on a treatment coefficient remains asymptotically valid for any bounded $R \geq r$ when $r \geq 1$. These results formally justify a common empirical practice (deliberately overestimating or adopting a conservative upper bound on the number of factors) and shift the analytical burden from consistent factor-number selection to the structurally milder requirement of bounding r from above.

Beyond the technical contributions, our results have a methodological implication for empirical practice. The dominant approach in factor-model inference treats consistent dimension selection as a logically prior step: estimate r, condition on it, and proceed. This paper shows that this step can be replaced by the weaker requirement of specifying an upper bound $R \geq r$. The benefit is robustness: an inferential procedure that depends only

on $R \geq r$ is insulated from the finite-sample volatility of $\widehat{r}$, which is well documented to be substantial in the signal regimes where applied researchers most need reliable inference. The cost is a variance inflation that scales with R/T. For empirically relevant settings such as factor-augmented treatment-effect inference, factor-based forecasting, and large-panel principal-component regression, this trade-off can be favorable.

Several directions merit further investigation, listed in approximate order of substantive importance. First, the factor-augmented regression we consider assumes serially-uncorrelated residuals. A parallel extension to weakly serially-dependent residuals appears tractable so long as cross-sectional independence in u_t is maintained. The genuinely difficult extension is to allow serial *and* cross-sectional dependence simultaneously in u_t, since that is precisely the regime in which the anisotropic local law of Knowles and Yin (2017) ceases to apply, and a fundamentally different random-matrix input would be required. Second, our analysis imposes $N/T \to \phi \neq 1$ to inherit incoherence from Knowles and Yin (2017); removing this gap may require new random-matrix tools, perhaps through a fluctuation-scale argument at the Marchenko–Pastur edge. Finally, our framework restricts attention to bounded $R - r$; the regime in which R grows slowly with T would require sharper control of the cumulative variance contribution of the overestimated components and may be relevant for sieve-type applications.

A Regularity Conditions of the MP law

We summarize the regularity conditions on the population covariance spectrum, following Knowles and Yin (2017), using our notation system. Given Σ_e with eigenvalues ordered $\sigma_1 \geq \ldots \geq \sigma_N > 0$, let

$$\pi := \frac{1}{N}\sum_{i=1}^{N} \delta_{\sigma_i}$$

denote the empirical spectral measure. Write $\phi := N/T$ for the dimensional ratio.

The MP law ϱ describes the typical spectral distribution of the rescaled noise matrix, $U = \frac{1}{\sqrt{T}}\Sigma_e^{1/2}E$, when $T \to \infty$, and E follows Assumption 2.1 (i). The Stieltjes transform of ϱ is the unique solution in $\mathbb{C}^+$ of the self-consistent equation

$$\frac{1}{m} = -z + \phi \int \frac{x}{1+mx}\,\pi(\mathrm{d}x), \qquad z \in \mathbb{C}^+. \tag{A.1}$$

Let $n := |\operatorname{supp}\pi \setminus \{0\}|$ be the number of distinct nonzero eigenvalues of Σ, and write $\operatorname{supp}\pi \setminus \{0\} = \{s_1 > s_2 > \cdots > s_n\}$. Setting $r_i := \phi\,\pi(\{s_i\})$, the equation (A.1) is equivalently written as

$$z = f(m), \qquad f(x) := -\frac{1}{x} + \sum_{i=1}^{n} \frac{r_i}{x + s_i^{-1}}. \tag{A.2}$$

The function f is smooth on each of the $n+1$ open intervals

$$I_1 := (-s_1^{-1}, 0), \quad I_i := (-s_i^{-1}, -s_{i-1}^{-1}) \ \ (i = 2, \ldots, n), \quad I_0 := \mathbb{R} \setminus \bigcup_{i=1}^{n} I_i,$$

of the real projective line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$.

Let $\mathcal{C} \subset \overline{\mathbb{R}}$ denote the multiset of critical points of f, where a nondegenerate critical point is counted once and a degenerate one twice, and where ∞ is counted as a nondegenerate critical point when $\phi = 1$. Some of the known properties of $\mathcal{C}$ include:

- $|\mathcal{C} \cap I_0| = |\mathcal{C} \cap I_1| = 1$ and $|\mathcal{C} \cap I_i| \in \{0, 2\}$ for $i = 2, \ldots, n$.
- $|\mathcal{C}| = 2p$ is even, for some integer $p \geq 1$.

We label the $2p-1$ critical points in $I_1 \cup \cdots \cup I_n$ as $x_1 \geq x_2 \geq \cdots \geq x_{2p-1}$, and denote by x_{2p} the unique critical point in I_0. The associated *critical values* are

$$a_k := f(x_k), \qquad k = 1, \ldots, 2p,$$

and satisfy $a_1 \geq a_2 \geq \cdots \geq a_{2p}$, with $a_k \in [0, C]$ for all k. Moreover, $x_k = m(a_k)$, where m is extended to $\mathbb{R}$ by continuity from $\mathbb{C}^+$.

The support of the asymptotic density ϱ in $(0,\infty)$ consists of exactly p connected components:

$$\operatorname{supp}\varrho\cap(0,\infty)=\bigcup_{k=1}^{p}[a_{2k},\,a_{2k-1}]\cap(0,\infty). \tag{A.3}$$

Thus the k-th *bulk component* of ϱ is the interval $[a_{2k}, a_{2k-1}]$, with left edge a_{2k} and right edge a_{2k-1}. The spectral edges are collectively indexed by $k = 1,\dots,2p$, where odd indices $k = 1,3,\dots,2p-1$ correspond to right edges and even indices $k = 2,4,\dots,2p$ correspond to left edges of the respective bulk components.

We now state the regularity conditions, which govern whether the density ϱ is well-behaved near a given spectral edge or in the interior of a given bulk component. Fix a small constant $\delta > 0$ throughout.

Definition A.1 (Regularity (Knowles and Yin, 2017, Definition 2.7))**.**

(i) ***Regular edge.*** *The edge $k \in \{1,\dots,2p\}$ is called* regular *if the following three conditions hold simultaneously:*

$$a_k \geq \delta, \qquad \min_{l\neq k}|a_k - a_l| \geq \delta, \qquad \min_{1\leq i\leq n}|x_k + s_i^{-1}| \geq \delta. \tag{A.4}$$

(ii) ***Regular bulk component.*** *The k-th bulk component $k \in \{1,\dots,p\}$ is called* regular *if, for every fixed $\delta' > 0$, there exists a constant $c \equiv c_{\delta,\delta'} > 0$ such that the density of ϱ is bounded below by c on the compactly interior interval $[a_{2k}+\delta',\, a_{2k-1}-\delta']$.*

B Canonical rotation construction

This subsection records the explicit construction of the rotations H_B and H_F used in identity (2.1) of the main paper, deferred from Section 2.1 for compactness. Let $M = \Xi_r L_r V_r'$ be the SVD of BF', write $S_B = N^{-1}B'B$ and $S_f = T^{-1}F'F$, and let $J \in \mathbb{R}^{r\times r}$ denote the diagonal matrix of common eigenvalues of $S_B^{1/2}S_fS_B^{1/2}$ and $S_f^{1/2}S_BS_f^{1/2}$. Let G_B be the $r\times r$ matrix whose columns are eigenvectors of $S_B^{1/2}S_fS_B^{1/2}$ corresponding to eigenvalues J, and similarly let G_F collect the eigenvectors of $S_f^{1/2}S_BS_f^{1/2}$. Define

$$H_B \;:=\; S_B^{-1/2}\,G_B, \qquad H_F \;:=\; S_f^{-1/2}\,G_F. \tag{B.1}$$

A direct computation gives

$$\frac{1}{NT}\,MM'\Big(\frac{1}{\sqrt{N}}BH_B\Big) \;=\; \Big(\frac{1}{\sqrt{N}}BH_B\Big)\,J, \qquad \Big(\frac{1}{\sqrt{N}}BH_B\Big)'\Big(\frac{1}{\sqrt{N}}BH_B\Big) \;=\; I_r,$$

so $N^{-1/2}BH_B = \Xi_r$ and equivalently $N^{-1/2}B = \Xi_r H_B^{-1}$. The parallel computation on the factor side, using $(NT)^{-1}M'M = T^{-1}FS_BF'$, gives

$$\frac{1}{NT}M'M\Big(\frac{1}{\sqrt{T}}FH_F\Big) = \Big(\frac{1}{\sqrt{T}}FH_F\Big)J, \qquad \Big(\frac{1}{\sqrt{T}}FH_F\Big)'\Big(\frac{1}{\sqrt{T}}FH_F\Big) = I_r,$$

so $T^{-1/2}FH_F = V_r$ and $T^{-1/2}F = V_rH_F^{-1}$. Substituting both identities into $M = BF'$ yields $M = \sqrt{NT}\,\Xi_r H_B^{-1}(H_F')^{-1}V_r'$; comparing with the SVD $M = \Xi_r L_r V_r'$ gives $(H_F'H_B)^{-1} = (NT)^{-1/2}L_r$. This proves identity (2.1).

C General Theorems

C.1 Preliminaries: notation and probabilistic tools

Throughout the supplement we use the convention $X, M, U \in \mathbb{R}^{N\times T}$, with left singular vectors lying in $\mathbb{R}^N$ and right singular vectors in $\mathbb{R}^T$. If $r \geq 1$, let $BF' = \Xi_r L_r V_r'$ be the SVD of BF', where $\Xi_r \in \mathbb{R}^{N\times r}$, $L_r \in \mathbb{R}^{r\times r}$, and $V_r \in \mathbb{R}^{T\times r}$. Let $\nu_{\min}$ denote the minimum nonzero eigenvalue of $\frac{1}{N}B'B$. The singular values of BF' are of order

$$\nu_1(L_r) \asymp \nu_r(BF') \asymp \nu_1(BF') \asymp \sqrt{NT}\nu_{\min}^{1/2} \asymp \nu_M\sqrt{T},$$

so that

$$\nu_{\min}^{-1} \asymp N\nu_M^{-2}.$$

For the empirical spectrum of X, let $\widehat{\xi}_l \in \mathbb{R}^N$ denote the left singular vector of X corresponding to the l-th largest singular value $\widehat{\lambda}_l$, and assemble $\widehat{\Xi}_R = (\widehat{\xi}_1, \ldots, \widehat{\xi}_R) \in \mathbb{R}^{N\times R}$ and $\widehat{\Xi}_{-r} = (\widehat{\xi}_{r+1}, \ldots, \widehat{\xi}_R) \in \mathbb{R}^{N\times(R-r)}$. The PCA loading estimators are $\widehat{B}_r = \sqrt{N}\,\widehat{\Xi}_r$ and $\widehat{B}_{-r} = \sqrt{N}\,\widehat{\Xi}_{-r}$, and we write $\widehat{B} = \sqrt{N}\,\widehat{\Xi}_R$ for the full $N \times R$ block. Right singular vectors of X are denoted $\widehat{v}_l \in \mathbb{R}^T$, with $\widehat{V}_R$ and $\widehat{V}_{-r}$ defined analogously. For the noise matrix $U = \Sigma_e^{1/2}E$, $u_{0,k} \in \mathbb{R}^N$ and $w_{0,k} \in \mathbb{R}^T$ denote the k-th left and right singular vectors.

We will use the following result, which is Lemma S.5 of Hu and Wang (2024).

Lemma C.1. *Suppose $A \in \mathcal{R}^{n\times K}, B \in \mathcal{R}^{K\times p}$ and $\operatorname{rank}(A) = \operatorname{rank}(B) = K$, then*

$$\lambda_K(AB) \geq \lambda_K(A)\lambda_K(B), \quad \lambda_1(AB) \leq \lambda_1(A)\lambda_1(B).$$

Proof. See Hu and Wang (2024), Lemma S.5. ∎

The next lemma controls the alignment between the rank-r signal subspace and the leading k noise singular vectors. Together with Lemma C.3 below, it provides the "orthogonality budget" that drives the proof of Theorem 2.1(i).

Lemma C.2. *Under Assumption 2.1, with $\Xi_r \in \mathbb{R}^{N\times r}$ the left singular vectors of M and $U_k \in \mathbb{R}^{N\times k}$ the leading k left singular vectors of $\Sigma_e^{1/2}E$, for every fixed $r, k = O(1)$ and*

every $\varepsilon > 0$,

$$\|U_k' \, \Xi_r\| \;\prec\; T^{\varepsilon - 1/2}.$$

The same bound holds for $\|W_k' \, V_r\|$ *on the right side, where* $V_r \in \mathbb{R}^{T \times r}$ *are the right singular vectors of* M *and* $W_k \in \mathbb{R}^{T \times k}$ *the leading* k *right singular vectors of* $\Sigma_e^{1/2} E$.

Proof. By Proposition C.4(3) applied entrywise: each entry $u_{0,j}' \, \xi_i$ is bounded by $CT^{\varepsilon - 1/2}$ with high probability, since ξ_i is determined by M and hence independent of both E and $\Sigma_e^{1/2}$. The operator-norm bound follows by a union bound over the $rk = O(1)$ entries and the equivalence between operator norm and maximum entry for matrices of fixed dimension. The right-side bound is symmetric. ∎

We also need an incoherence-type bound on noise-side singular vectors.

Lemma C.3. *Under Assumption 2.1, for every fixed* $k \le R - r$ *and every* $\varepsilon > 0$, *the* k*-th right singular vector* $w_k \in \mathbb{R}^T$ *of* $\Sigma_e^{1/2} E$ *satisfies*

$$|w_k' \, v| \;\prec\; T^{\varepsilon - 1/2} \qquad \textit{for any deterministic unit vector } v \in \mathbb{R}^T \textit{ independent of } E.$$

Consequently, with $W_k = [w_1, \dots, w_k] \in \mathbb{R}^{T \times k}$ *and* $V_r \in \mathbb{R}^{T \times r}$ *the right singular vectors of the rank-*r *signal* M,

$$\|V_r' \, W_k\| \;\prec\; T^{\varepsilon - 1/2},$$

and for the signal matrix $M \in \mathbb{R}^{N \times T}$ *with* $\|M\| \asymp \nu_M \sqrt{T}$,

$$\|M \, W_k\| \;\le\; \|L_r\| \, \|V_r' \, W_k\| \;\prec\; \nu_M T^{\varepsilon}.$$

Proof. The first claim is the anisotropic delocalization bound for w_k given in Theorem 3.12 (and Remark 3.13) of Knowles and Yin (2017); see Proposition C.4 (claim 3) below for the form used here. The bound on $\|V_r' \, W_k\|$ follows by applying the first claim with v ranging over the $r = O(1)$ columns of V_r (which are E-independent because V_r is determined by M) and assembling via a union bound. The bound on $\|MW_k\|$ then uses the SVD $M = \Xi_r L_r V_r'$ and $\|L_r\| \asymp \nu_M \sqrt{T}$. ∎

C.2 Proof of Theorem 2.1

The proof uses local laws for the noise matrix under factor models, stated later in Proposition C.4.

Claim (i): singular values. Throughout, $\xi_1, \dots, \xi_r$ denote the columns of $\Xi_r \in \mathbb{R}^{N \times r}$ (the left singular vectors of M), $u_1, \dots, u_k$ the columns of $U_k \in \mathbb{R}^{N \times k}$ (the leading k left singular vectors of $\Sigma_e^{1/2} E$), $w_1, \dots, w_k$ the columns of $W_k \in \mathbb{R}^{T \times k}$ (the corresponding right singular vectors), and $\Lambda_k \in \mathbb{R}^{k \times k}$ the diagonal matrix of the leading k singular values of $\Sigma_e^{1/2} E$. The rank-k truncated SVD identity $\Sigma_e^{1/2} E \, W_k = U_k \Lambda_k$, equivalently $(\Sigma_e^{1/2} E)' \, U_k = W_k \Lambda_k$, will be used repeatedly.

Upper bound. The bound $\widehat{\lambda}_{r+k} \le \lambda_k(\Sigma_e^{1/2}E)$ is immediate from Weyl's inequality for singular values applied to $X = M + \Sigma_e^{1/2}E$:

$$\widehat{\lambda}_{r+k} = \sigma_{r+k}(M + \Sigma_e^{1/2}E) \;\le\; \sigma_{r+1}(M) + \sigma_k(\Sigma_e^{1/2}E) \;=\; 0 + \lambda_k(\Sigma_e^{1/2}E),$$

since $\mathrm{rank}(M) = r$ implies $\sigma_{r+1}(M) = 0$.

Lower bound. Consider the subspace $S = \mathrm{span}\{\xi_1, \dots, \xi_r, u_1, \dots, u_k\} \subset \mathbb{R}^N$, of dimension $r + k$ with high probability by Lemma C.2. By Courant–Fischer,

$$\widehat{\lambda}_{r+k} \ge \min_{z \in S, \|z\|=1} \|(M + \Sigma_e^{1/2}E)'z\|.$$

Every $z \in S$ admits the representation $z = \Xi_r x + U_k y$ with $x \in \mathbb{R}^r$, $y \in \mathbb{R}^k$. Meanwhile, consider $P_r = \Xi_r\Xi_r', P_c = I - P_r$, which are the projection onto the subspace of Ξ_r and its complement respectively. We note that

$$z = \Xi_r x + U_k y = \Xi_r x + (\Xi_r\Xi_r')U_k y + P_c U_k y = \Xi_r(x + \Xi_r' U_k y) + P_c U_k y.$$

In other words, we can write all $z \in S$ as $z = \Xi_r x + P_c U_k y$ for some x and y.

Step 1: orthogonality budget. By Lemma C.2, $\|U_k'\Xi_r\| \prec T^{\varepsilon-1/2}$, so

$$1 = \|z\|^2 = \|\Xi_r x + P_c U_k y\|^2 = \|x\|^2 + \|P_c U_k y\|^2.$$

Meanwhile, we note that by Proposition C.4, $\|P_r U_k\| = \|\Xi_r U_k\| = O_p(T^{\varepsilon-1/2})$

$$\|P_c U_k y\| \ge \|U_k y\| - \|P_r U_k y\| \ge (1 - CT^{\varepsilon-1/2})\|y\|$$

which yields

$$\big|\|x\|^2 + \|y\|^2 - 1\big| \;\le\; C\,T^{\varepsilon-1/2}\,\|y\|^2 \;\le\; C\,T^{\varepsilon-1/2}. \tag{C.1}$$

Step 2: lower bound on $\|X'z\|^2$. Decompose

$$\begin{aligned} X'z &= (M + \Sigma_e^{1/2}E)'(\Xi_r x + P_c U_k y - P_r U_k y) \\ &= \underbrace{(M + \Sigma_e^{1/2}E)'\Xi_r x}_{A \in \mathbb{R}^T} + \underbrace{W_k\Lambda_k y}_{D \in \mathbb{R}^T} - \underbrace{(\Sigma_e^{1/2}E)'P_r U_k y}_{B \in \mathbb{R}^T}, \end{aligned}$$

where we used that $M'(I - P_r) = 0$ and the second piece uses $(\Sigma_e^{1/2}E)'U_k = W_k\Lambda_k$. We bound each piece.

(a) $\|A\|$. Using $\Xi_r'M = L_r V_r'$, $M'\Xi_r = V_r L_r$, and $\|(\Sigma_e^{1/2}E)'\Xi_r\| \le \|\Sigma_e^{1/2}E\| \asymp \sqrt{T}$:

$$\|A\| \;\ge\; \|V_r L_r x\| - \|(\Sigma_e^{1/2}E)'\Xi_r\, x\| \;\ge\; \nu_M\sqrt{T}\,\|x\| - C\sqrt{T}\,\|x\| \;\ge\; c\,\nu_M\sqrt{T}\,\|x\|$$

for ν_M large, hence $\|A\|^2 \geq c\,\nu_M^2 T\,\|x\|^2$. Meanwhile, $\|A\| \leq \|x\|\|M+\Sigma_e^{1/2}E\| = O(T\|x\|)$.

(b) $\|D\|$. Since W_k has orthonormal columns, $\|D\|^2 = \|\Lambda_k y\|^2 \geq \lambda_k(\Sigma_e^{1/2}E)^2\,\|y\|^2$.

(c) $\|B\|$. By Lemma C.2, $\|B\| \leq \|\Sigma_e^{1/2}E\|\|P_r U_k\|\,\|y\| = O(T^\varepsilon\|y\|)$.

(d) Cross terms. The cross terms involving D are controlled using Lemma C.3 ($\|V_r' W_k\| \prec T^{-1/2}$) and Lemma C.2:

$$|\langle A, D\rangle| = |x'\Xi_r'(M+\Sigma_e^{1/2}E)W_k\Lambda_k y| \leq \|x\|\,\|y\|\,\|(M+\Sigma_e^{1/2}E)W_k\|\,\|\Lambda_k\|,$$
$$\|(M+\Sigma_e^{1/2}E)W_k\| \leq \|MW_k\| + \|\Sigma_e^{1/2}E\,W_k\| \leq \nu_M + \sqrt{T} \leq C\sqrt{T}\,T^\varepsilon,$$

where $\|MW_k\| \prec \nu_M$ is from Lemma C.3. Hence $|\langle A, D\rangle| \leq CT^{1+\varepsilon}\|x\|\|y\|$. Meanwhile, we have

$$|\langle A, B\rangle| \leq \|A\|\|B\| = O(T^{1+\varepsilon}\|x\|\|y\|).$$

Moving on, using $W_k'(\Sigma_e^{1/2}E)' = (U_k\Lambda_k)'$ we find

$$\langle D, B\rangle = y'\Lambda_k W_k'(\Sigma_e^{1/2}E)'P_r U_k y = y'\Lambda_k^2 U_k'\Xi_r\Xi_r' U_k y.$$

Therefore by Lemma C.2,

$$|\langle D, B\rangle| \leq \|\Lambda_k\|^2\|U_k'\Xi_r\|^2\|y\|^2 \prec T\cdot T^{2\varepsilon-1}\|y\|^2 = T^{2\varepsilon}\|y\|^2.$$

Combining (a)–(d), since $X'z = A + D - B$,

$$\begin{aligned}\|X'z\|^2 &= \|A\|^2 + \|D\|^2 + \|B\|^2 - 2\langle A, B\rangle + 2\langle A, D\rangle - 2\langle D, B\rangle \\ &\geq \nu_M^2 T\,\|x\|^2 + \lambda_k(\Sigma_e^{1/2}E)^2\,\|y\|^2 + 0 - CT^{1+\varepsilon}\,\|x\|\|y\| - CT^{2\varepsilon}\|y\|^2.\end{aligned}$$

Step 3: minimization. Let $a = \|x\|$, $l_k = \lambda_k(\Sigma_e^{1/2}E) \asymp \sqrt{T}$; using (C.1), that is, lower-bound $\|y\|^2$ by $1-a^2$, upper-bound $\|y\|^2$ by $1-a^2+CT^{2\varepsilon-1}$ when it appears with negative sign,

$$\|X'z\|^2 \geq \nu_M^2 T a^2 + l_k^2(1-a^2) - CT^{1+\varepsilon}\,a\sqrt{1+CT^{2\varepsilon-1}-a^2} - CT^{2\varepsilon}(1+CT^{2\varepsilon-1}-a^2).$$

Define

$$f(a) = (\nu_M^2 T - l_k^2 + CT^{2\varepsilon})a^2 + l_k^2 - CT^{1+\varepsilon}a\sqrt{1+CT^{2\varepsilon-1}-a^2} - CT^{2\varepsilon}(1+CT^{2\varepsilon-1}).$$

Note that f has derivative

$$f'(a) = 2a(\nu_M^2 T - l_k^2 + CT^{2\varepsilon}) - CT^{1+\varepsilon}\frac{1+CT^{2\varepsilon-1}-2a^2}{\sqrt{1+CT^{2\varepsilon-1}-a^2}}.$$

Since $\nu_M^2 - l_k^2 + CT^{2\varepsilon} \asymp \nu_M^2 T$ for $\nu_M \to \infty$, $f'(a) > 0$ whenever $a > \nu_M^{-2}T^\varepsilon$; hence the

minimizer $a^* \le \nu_M^{-2} T^{\varepsilon}$, and

$$\widehat{\lambda}_{r+k}^2 \ge f(a^*) \ge l_k^2 - C\nu_M^{-2} T^{1+2\varepsilon} - 2CT^{2\varepsilon} \ge \lambda_k(\Sigma_e^{1/2} E)^2 - O_{\prec}(\nu_M^{-2} T),$$

since $T^{2\varepsilon} \prec 1 \prec \nu_M^{-2} T$ throughout the range $\nu_M = O(\sqrt{N})$ (the inequality $1 \le \nu_M^{-2} T$ holds at $\nu_M \le \sqrt{T}$, with equality at the strong-factor boundary). Combined with the Weyl upper bound, this proves

$$0 \le \lambda_k(\Sigma_e^{1/2} E)^2 - \widehat{\lambda}_{r+k}^2 \prec \nu_M^{-2} T,$$

uniformly over $\nu_M = O(\sqrt{N})$, including the strong-factor boundary $\nu_M \asymp \sqrt{N}$.

Claim (iii): near-orthogonality, left singular vectors. We first prove claim (iii) and then claim (ii), since the proof of the latter uses the former.

Step 1. Block transformation of the eigenvector equation. Recall the SVD $M = \Xi_r L_r V_r'$ from the Preliminaries; we write $B_r := L_r V_r' \in \mathbb{R}^{r \times T}$ for compactness, so $M = \Xi_r B_r$ and $\|B_r\| = \|L_r\| \asymp \sqrt{T}\,\nu_M$. Since B shares its left singular space with Ξ_r and all r eigenvalues of $B'B/N$ are of order $\nu_{\min} \asymp \nu_M^2/N$, showing $\|B\|_{\mathrm{F}}^{-1} \|B' \widehat{\Xi}_{-r}\| = O(T^{\varepsilon} \nu_M^{-2})$ is equivalent to showing $\|\Xi_r' \widehat{\xi}_k\| = O(T^{\varepsilon} \nu_M^{-2})$ for all $k \in [r+1, R]$.

Let $\Xi_c \in \mathbb{R}^{N \times (N-r)}$ be an orthonormal completion of Ξ_r, so $\Xi_r \Xi_r' + \Xi_c \Xi_c' = I_N$. Decompose $\widehat{\xi}_k = \Xi_r x_k + \Xi_c y_k$ with

$$x_k = \Xi_r' \widehat{\xi}_k \in \mathbb{R}^r, \quad y_k = \Xi_c' \widehat{\xi}_k \in \mathbb{R}^{N-r}.$$

Then $\|x_k\|^2 + \|y_k\|^2 = 1$ by orthonormality of $\widehat{\xi}_k$ and the identity $\Xi_r \Xi_r' + \Xi_c \Xi_c' = I_N$. The goal is to bound $\|x_k\|$ for $r+1 \le k \le R$.

By claim 1, with high probability $\widehat{\lambda}_k \asymp \sqrt{T}$. We write $E_r = \Xi_r' \Sigma_e^{1/2} E, E_c = \Xi_c' \Sigma_e^{1/2} E$. Then we have

$$(M + \Sigma_e^{1/2} E)(M + \Sigma_e^{1/2} E)'(\Xi_r x_k + \Xi_c y_k) = \widehat{\lambda}_k^2 \cdot (\Xi_r x_k + \Xi_c y_k)$$

Because $\Xi_r \perp \Xi_c$, respectively left multiply by Ξ_r and Ξ_c to both sides leads to:

$$Q_r := (B_r + E_r), \quad \begin{bmatrix} Q_r Q_r' & Q_r E_c' \\ E_c Q_r' & E_c E_c' \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} \widehat{\lambda}_k^2 x_k \\ \widehat{\lambda}_k^2 y_k \end{bmatrix} \tag{C.2}$$

First row gives us

$$x_k = -(Q_r Q_r' - \widehat{\lambda}_k^2 I)^{-1} Q_r E_c' y_k \tag{C.3}$$

The inversion is feasible because with high probability, the following holds with some constant

$$\lambda_r(Q_r) \ge \lambda_r(B_r) - \|E_r\| \ge \frac{1}{C} \nu_M \sqrt{T} - C\sqrt{T} \ge \frac{1}{2C} \nu_M \sqrt{T}$$

for some constant c, while $\widehat{\lambda}_k \le \|\Sigma_e^{1/2} E\| = C\sqrt{T}$ with high probability. So $Q_r Q_r' - \widehat{\lambda}_k^2 I \succ 0$ with high probability. Substituting (C.3) into the second row gives

$$\widehat{\lambda}_k^2 y_k = E_c Q_r' x_k + E_c E_c' y_k = -E_c Q_r' (Q_r Q_r' - \widehat{\lambda}_k^2 I)^{-1} Q_r E_c' y_k + E_c E_c' y_k.$$

Step 2: the case $k = r+1$. Left multiply y' on both sides, and rearrange,

$$y_k' E_c Q_r' (Q_r Q_r' - \widehat{\lambda}_k^2 I)^{-1} Q_r E_c' y_k = y_k' E_c E_c' y_k - \widehat{\lambda}_k^2 \|y_k\|^2 \tag{C.4}$$

We observe that the left-hand side is non-negative because $(Q_r Q_r' - \widehat{\lambda}_k^2 I)^{-1}$ is positive definite with high probability. Moreover, since $\lambda_1(Q_r Q_r' - \widehat{\lambda}_k^2 I) \le CT\nu_M^2$ with high probability, we have the positive-semidefinite inequality $(Q_r Q_r' - \widehat{\lambda}_k^2 I)^{-1} \succeq \frac{1}{CT\nu_M^2} I$, whence

$$\frac{1}{CT\nu_M^2} \|Q_r E_c' y_k\|^2 \le y_k' E_c E_c' y_k - \widehat{\lambda}_k^2 \|y_k\|^2. \tag{C.5}$$

We argue by induction, beginning with the case $k = r+1$. Thus, by claim (i), we have

$$\frac{1}{CT\nu_M^2} \|Q_r E_c' y_{r+1}\|^2 \le \lambda_1(\Sigma_e^{1/2} E)^2 - \widehat{\lambda}_{r+1}^2 \le O(T^{1+2\varepsilon} \nu_M^{-2}).$$

So $\|Q_r E_c' y_{r+1}\| \le CT^{1+\varepsilon}$.

$$x_{r+1} = -(Q_r Q_r' - \widehat{\lambda}_{r+1}^2 I)^{-1} Q_r E_c' y_{r+1} = O(T^{\varepsilon} \nu_M^{-2}),$$

which proves the claim for $r+1$.

Step 3: induction to other cases. Throughout this step the index i ranges over $1, \dots, R-r$. Since $R - r = O(1)$ by assumption, all accumulated Gram–Schmidt errors satisfy $C_R \cdot (\prec \nu_M^{-2})$ with C_R a constant depending only on the bounded difference $R - r$; we absorb C_R into the implicit constant in $\prec$ below. Next, suppose the bound holds through $k = r + i - 1$. For $k = r + i$, orthonormality implies that, for any $j \le i - 1$,

$$0 = \widehat{\eta}_{r+j}' \widehat{\eta}_{r+i} = y_{r+j}' y_{r+i} + x_{r+j}' x_{r+i} \Rightarrow y_{r+j}' y_{r+i} = -x_{r+j}' x_{r+i} = O(T^{\varepsilon} \nu_M^{-2}),$$

where we used the induction $\|x_{r+j}\| = O(T^{\varepsilon} \nu_M^{-2})$ and $\|x_{r+i}\| \le 1$. In other words, y_{r+j} is close to orthogonal to y_{r+i}. Consider a Gram-Schmidt transform, and let

$$\tilde{y}_{r+i} = \frac{y_{r+i} - \sum_{j \le i-1} \langle y_{r+i}, y_{r+j} \rangle y_{r+j}}{\|y_{r+i} - \sum_{j \le i-1} \langle y_{r+i}, y_{r+j} \rangle y_{r+j}\|}.$$

Then $\|\tilde{y}_{r+i} - y_{r+i}\| \le O(T^{\varepsilon} \nu_M^{-2})$, and

$$\|E_c' \tilde{y}_{r+i}\|^2 - \|E_c' y_{r+i}\|^2 \le \|E_c' y_{r+i}\| \|E_c\| T^{\varepsilon} \nu_M^{-2} \le \|\Sigma_e^{1/2} E\|^2 T^{\varepsilon} \nu_M^{-2} = O(T^{1+\varepsilon} \nu_M^{-2}).$$

Since $i \le R - r = O(1)$, let $\tilde{y}_{r+1} = y_{r+1}$. Then $\tilde{y}_{r+1:r+i}$ consists of i orthonormal vectors, so using claim (i),

$$\|E_c' y_{r+1:r+i}\|^2 - O(T^{1+\varepsilon}\nu_M^{-2}) \le \|E_c'\tilde{y}_{r+1:r+i}\|^2 \le \sum_{j\le i}\lambda_{j+r}(E_c)^2 \lesssim \sum_{j\le i}\lambda_{j+r}(\Sigma_e^{1/2}E)^2 \le \sum_{j\le i}\widehat{\lambda}_{j+r}^2 + O(T^{1+2\varepsilon}\nu_M^{-2}).$$

So we have

$$\|E_c' y_{r+1:r+i}\|^2 - \sum_{j\le i}\widehat{\lambda}_{j+r}^2 = O(T^{1+2\varepsilon}\nu_M^{-2})$$

Meanwhile,

$$\begin{aligned}\sum_{j\le i}\|x_{r+j}\|^2 &= \sum_{j\le i}\|(Q_rQ_r' - \widehat{\lambda}_{r+j}^2 I)^{-1}Q_rE_c'y_{r+j}\|^2 \\ &\lesssim \frac{1}{T^2\nu_M^4}\sum_{j\le i}\|Q_rE_c'y_{r+j}\|^2 \quad \text{(by (C.5))} \\ &\lesssim \frac{1}{T\nu_M^2}\sum_{j\le i}(\|E_c'y_{r+j}\|^2 - \widehat{\lambda}_{r+j}^2) \\ &= \frac{1}{T\nu_M^2}\left(\|E_c'y_{r+1:r+i}\|^2 - \sum_{j\le i}\widehat{\lambda}_{r+j}^2\right) \lesssim T^{2\varepsilon}\nu_M^{-4}.\end{aligned}$$

From this we conclude that $\|x_{r+i}\| = O(T^{\varepsilon}\nu_M^{-2})$.

Claim (iii): near-orthogonality, right singular vectors. The bound $\|F\|_{\mathrm{F}}^{-1}\|F'\widehat{V}_{-r}\| \prec \nu_M^{-2}$ for the right side follows by symmetry. Specifically, applying the argument above to the transposed matrix $X' = M' + (\Sigma_e^{1/2}E)' \in \mathbb{R}^{T\times N}$ exchanges the roles of left and right singular vectors, mapping $\Xi_r \mapsto V_r$, $V_r \mapsto \Xi_r$, $U_k \mapsto W_k$, and $W_k \mapsto U_k$. The orthogonality budget $\|W_k'V_r\| \prec T^{\varepsilon-1/2}$ is the right-side counterpart of Lemma C.2, also supplied by Proposition C.4 (3). All local-law inputs from Knowles and Yin (2017) are symmetric in the left/right convention under our setting $N \asymp T$, $\phi \neq 1$, so the entire argument transposes verbatim, yielding $\|V_r'\widehat{v}_k\| \prec T^{\varepsilon}\nu_M^{-2}$ for each $k \in [r+1, R]$ and consequently $\|F\|_{\mathrm{F}}^{-1}\|F'\widehat{V}_{-r}\| \prec \nu_M^{-2}$ via the analogue of the equivalence $\|B\|_{\mathrm{F}}^{-1}\|B'\widehat{\Xi}_{-r}\| \Leftrightarrow \|\Xi_r'\widehat{\xi}_k\|$ established in Step 1.

Claim (ii): incoherence of extra eigenvectors. Let $u_1, u_2, \ldots \in \mathbb{R}^N$ and $w_1, w_2, \ldots \in \mathbb{R}^T$ denote the left and right singular vectors of $\Sigma_e^{1/2}E$ ordered by decreasing singular value. We first prove the bound $\|\eta'\widehat{\xi}_k\| \prec \nu_M^{-1}$ for the left singular vectors of X; the symmetric bound for $\widehat{v}_k$ follows by transposition (cf. Step 3 below).

Step 1: expansion of $\widehat{\xi}_k$ in terms of u_i. The column space of $X = M + \Sigma_e^{1/2}E$ is contained in the sum of the column spaces of M (which equals $\mathrm{span}(\Xi_r)$) and of $\Sigma_e^{1/2}E$ (which is spanned by $u_1, \ldots, u_{\min(N,T)}$). Hence each empirical left singular vector $\widehat{\xi}_k \in \mathbb{R}^N$

admits the decomposition $\widehat{\xi}_k = \Xi_r x_k + \Xi_c y_k$ from claim (iii), with

$$\Xi_c y_k = \sum_{i \in [\min(N,T)]} a_{k,i}\, u_i, \qquad a_{k,i} = \langle u_i, \Xi_c y_k \rangle.$$

Recall from the induction used in the claim (iii), where we have that

$$x_i = O(T^{\varepsilon}\nu_M^{-2}), \quad y_i' y_j = y_i' \Xi_c \Xi_c' y_j = O(T^{\varepsilon}\nu_M^{-2}), \quad \|y_i - \tilde{y}_i\| = O(T^{\varepsilon}\nu_M^{-2}), \quad i,j \in \{r+1,\dots,R\}.$$

In other words, $\{\Xi_c' y_k\}_{k\in[R]}$ is within distance $O(T^{\varepsilon}\nu_M^{-2})$ of a collection of orthonormal vectors. As a consequence, if we write $A_i = [a_{1,i},\dots,a_{R,i}]'$, then the following holds for all i with high probability:

$$\begin{aligned} \|A_i\|^2 = \sum_{j=1}^{R} \langle u_i, \Xi_c' y_j \rangle^2 &\le \sum_{j=1}^{R} \langle u_i, \Xi_c' \tilde{y}_j \rangle^2 + O(T^{\varepsilon}\nu_M^{-2}) \\ &\le \|u_i\|^2 + O(T^{\varepsilon}\nu_M^{-2}) = 1 + O(T^{\varepsilon}\nu_M^{-2}). \end{aligned} \tag{C.6}$$

Moreover,

$$\sum_{i=1}^{N} \|A_i\|^2 = \sum_{j=1}^{R}\sum_{i=1}^{N} \langle u_i, \Xi_c' y_j \rangle^2 = \sum_{j=1}^{R} \|\Xi_c' y_j\|^2 = R \pm O(T^{\varepsilon}\nu_M^{-2}). \tag{C.7}$$

To continue, recall from (C.5) ,

$$y_j' \Xi_c \Sigma_e^{1/2} E E' \Sigma_e^{1/2} \Xi_c' y_j = y_j' E_c E_c' y_j \ge \widehat{\lambda}_j^2 \|y_j\|^2.$$

Next, recall claim (i) and $\|y_j\|^2 = 1 - \|x_j\|^2 \ge 1 - O(T^{2\varepsilon}\nu_M^{-4})$. We obtain

$$\begin{aligned} \sum_{i\in[N]} a_{j,i}^2 \lambda_i^2(\Sigma_e^{1/2} E) = y_j' \Xi_c \Sigma_e^{1/2} E E' \Sigma_e^{1/2} \Xi_c' y_j &\ge \widehat{\lambda}_j^2 \|y_j\|^2 \ge \lambda_j^2(\Sigma_e^{1/2} E)\|y_j\|^2 - O(T^{1+2\varepsilon}\nu_M^{-2}) \\ &\ge \lambda_j^2(\Sigma_e^{1/2} E) - O(T^{1+2\varepsilon}\nu_M^{-2}). \end{aligned}$$

Summing these equations over $j \le R$ gives

$$\sum_{i\in[N]} \|A_i\|^2 \lambda_i^2(\Sigma_e^{1/2} E) \ge \sum_{j=1}^{R} \lambda_j^2(\Sigma_e^{1/2} E) - O(T^{1+2\varepsilon}\nu_M^{-2}).$$

We rewrite the left-hand side as $\sum_{i\le R} \|A_i\|^2 \lambda_i^2(\Sigma_e^{1/2} E) + \sum_{i>R} \|A_i\|^2 \lambda_i^2(\Sigma_e^{1/2} E)$, and move these terms to the right-hand side:

$$O(T^{1+2\varepsilon}\nu_M^{-2}) \ge \sum_{j=1}^{R} \lambda_j^2(\Sigma_e^{1/2} E) - \sum_{i=1}^{R} \|A_i\|^2 \lambda_i^2(\Sigma_e^{1/2} E) - \sum_{i=R+1}^{N} \|A_i\|^2 \lambda_i^2(\Sigma_e^{1/2} E)$$

$$
\begin{aligned}
&= \sum_{i=1}^{R}(1-\|A_i\|^2)\lambda_i^2(\Sigma_e^{1/2}E) - \sum_{i=R+1}^{N}\|A_i\|^2\lambda_i^2(\Sigma_e^{1/2}E) \text{ (by (C.6) and } \lambda_i \geq \lambda_R) \\
&\geq \lambda_R^2(\Sigma_e^{1/2}E)\sum_{i=1}^{R}(1-\|A_i\|^2) - \sum_{i=R+1}^{N}\|A_i\|^2\lambda_i^2(\Sigma_e^{1/2}E) - O(T^{1+2\varepsilon}\nu_M^{-2}) \\
&\geq \lambda_R^2(\Sigma_e^{1/2}E)\sum_{i=R+1}^{T}\|A_i\|^2 - \sum_{i=R+1}^{N}\|A_i\|^2\lambda_i^2(\Sigma_e^{1/2}E) - O(T^{1+2\varepsilon}\nu_M^{-2}) \text{ (by (C.7))} \\
&= -O(T^{1+2\varepsilon}\nu_M^{-2}) + \sum_{i=R+1}^{N}(\lambda_R^2(\Sigma_e^{1/2}E) - \lambda_i^2(\Sigma_e^{1/2}E))\|A_i\|^2.
\end{aligned}
$$

By Proposition C.4 regularity condition (2), we know that with high probability,

$$\lambda_R(\Sigma_e^{1/2}E)^2 - \lambda_k(\Sigma_e^{1/2}E)^2 \geq ck, \quad k > R.$$

Since R is fixed and $T^\varepsilon \to \infty$, for T large enough $T^\varepsilon > R$, so this gap inequality applies on the entire range $k \geq T^\varepsilon$ used below. Therefore

$$O(T^{1+2\varepsilon}\nu_M^{-2}) \geq \sum_{i\geq K^\varepsilon}\|A_i\|^2(\lambda_R(\Sigma_e^{1/2}E)^2 - \lambda_i^2(\Sigma_e^{1/2}E)) \geq c\sum_{k\geq T^\varepsilon}\|A_k\|^2 k,$$

which by Cauchy–Schwarz further leads to

$$\sum_{k=T^\varepsilon+1}^{N}\|A_k\| \leq \sqrt{(\sum_{k=T^\varepsilon+1}^{N}\|A_k\|^2 k)(\sum_{k=T^\varepsilon+1}^{N}\frac{1}{k})} \leq O(T^{2\varepsilon+1/2}\nu_M^{-1}). \tag{C.8}$$

Step 2: delocalization for $\widehat{\xi}_k$. Recall that

$$\widehat{\xi}_k = \Xi_c' y_k + \Xi_r' x_k = \Xi_r' x_k + \sum_{i\in[N]} a_{k,i}u_i.$$

Therefore,

$$\eta'\widehat{\xi}_k = \eta'\Xi_r' x_k + \sum_{i\leq T^\varepsilon} a_{k,i}\eta' u_i + \sum_{i\geq T^\varepsilon+1} a_{k,i}\eta' u_i.$$

Theorem 2.1 indicates that $\|x_k\| = O(T^\varepsilon\nu_M^{-2})$. So $|\eta'\Xi_r' x_k| \leq O(T^\varepsilon\nu_M^{-2})$. Meanwhile, by Proposition C.4 and $a_{k,i} \leq 1$,

$$\sum_{i\leq T^\varepsilon} a_{k,i}\eta' u_i \lesssim T^{\varepsilon-1/2}\sum_{i\leq T^\varepsilon}|a_{k,i}| \lesssim T^{2\varepsilon-1/2}.$$

Moreover, by (C.8)

$$\sum_{i\geq T^\varepsilon+1} a_{k,i}\eta' u_i \lesssim T^{\varepsilon-1/2}\sum_{i\geq T^\varepsilon+1}|a_{k,i}| \lesssim T^{\varepsilon-1/2}\sum_{i\geq T^\varepsilon+1}\|A_i\| = O(T^{3\varepsilon}\nu_M^{-1}).$$

Combining the preceding bounds proves the claim for $\widehat{\xi}_k$.

Step 3: delocalization for $\widehat{v}_k$. The bound $\|\zeta'\widehat{v}_k\| \prec \nu_M^{-1}$ on the right side follows by the same transposition argument used for claim (iii). Replacing the model $X = M+\Sigma_e^{1/2}E$ by its transpose X' exchanges left and right singular vectors, mapping the decomposition $\widehat{\xi}_k = \Xi_r x_k + \sum_i a_{k,i}u_i$ to $\widehat{v}_k = V_r x_k + \sum_i a_{k,i}w_i$, and Steps 1–2 carry over verbatim with $\eta \in \mathbb{R}^N$ replaced by $\zeta \in \mathbb{R}^T$ and Lemma C.2's right-side bound $\|W_k'V_r\| \prec T^{\varepsilon-1/2}$ supplying the orthogonality budget.

C.3 Local laws of noise matrix

Recall that $X = M + U$ with $U = \Sigma_e^{1/2}E$, and that under Assumption 2.1 the entries of E are independent with all moments uniformly bounded, while Σ_e is positive definite with empirical spectral distribution π_U supported on $[c, 1]$.

In Proposition C.4 below, claim 3 (anisotropic delocalization of $u_{0,k}$ and $w_{0,k}$) is a direct consequence of Theorem 3.12 and Remark 3.13 of Knowles and Yin (2017) and Bloemendal et al. (2014). Claims 1 and 2 (a quantitative lower bound on the top-R singular values and an edge-spacing inequality) are not stated explicitly in Knowles and Yin (2017) and require a Stieltjes-transform argument; the bound on $a_1 \geq 1 + \delta/2$ in the proof below makes essential use of Assumption 2.1(ii). The three claims together provide, respectively, the spectral lower bound used in Step 2 of the proof of Theorem 2.1(i), the eigenvalue spacing used to control $\sum_{k\geq T^\varepsilon}\|A_k\|^2$ in claim (ii), and the entry-level delocalization used throughout (Lemmas C.3 and C.2).

Proposition C.4. *Suppose $N/T \to \phi \in \mathbb{R}\backslash\{1\}$. Under assumption 2.1, when T is sufficiently large, the following statements hold with high probability:*

1. *$\lambda_j(\Sigma_e^{1/2}E) > (1 + c_0)\sqrt{T}$ for all $j \leq R$.*
2. *$|\lambda_j(\Sigma_e^{1/2}E)^2 - \lambda_k(\Sigma_e^{1/2}E)^2| \geq c_0 k$ for all $j < R, k > T^\varepsilon$.*
3. *Let $u_{0,k}$ and $w_{0,k}$ be the k-th left and right singular vectors of U. For any two groups of norm-1 vectors $\mathcal{A} \subset \mathbb{R}^N, \mathcal{B} \subset \mathbb{R}^T$ that are independent of both E and $\Sigma_e^{1/2}$ (but may depend on M), suppose their cardinality $|\mathcal{A}| + |\mathcal{B}| \leq N^D$ for some fixed power D, then*

$$\max_{\eta\in\mathcal{A}} |\eta' u_{0,j}| \leq CT^{\varepsilon-1/2}, \quad \max_{\zeta\in\mathcal{B}} |\zeta' w_{0,j}| \leq CT^{\varepsilon-1/2}, \quad \forall j \leq \min\{T, N\}.$$

Proof. To obtain the first two claims, we use results from Knowles and Yin (2017). The limiting singular-value distribution of $\Sigma_e^{1/2}E/\sqrt{T}$ converges to a distribution $\rho_T(x)$, which can be defined through its Stieltjes transform:

$$m(z) := \int \frac{1}{x - z}\rho_T(dx),$$

which is defined on $\mathbb{C}/\text{supp}(\rho)$. Meanwhile, it is also the $\mathbb{C}_+$ solution to

$$z = f(m(z)) := -\frac{1}{m(z)} + \int \frac{\phi}{m(z) + x^{-1}} \hat{\mu}_T(dx).$$

where $\hat{\mu}_T$ is the empirical distribution of σ_i. The support of ρ_T is on the union of intervals $[a_{2p}, a_{2p-1}] \cup \cdots \cup [a_2, a_1]$ with $a_1 > a_2 > \ldots > a_{2p} \geq 0$ being the critical points. It is known the $x_1 = m(a_1)$ is the unique critical point of f in $(-\sigma_{(1)}^{-1}, 0)$, i.e.

$$0 = f'(x_1) = \frac{1}{x_1^2} - \int \frac{\phi}{(x_1 + x^{-1})^2} \hat{\mu}_T(dx)$$

Here $\sigma_{(1)}$ is the largest σ_i. Under the bulk regularity assumption, there exists $\tau^{-1}/4 > \delta > 0$ such that,

$$\hat{\mu}_T([\tau^{-1} - \delta, \tau^{-1} - \delta/2]) \geq 4\delta^2/\phi$$

Then we have

$$\frac{1}{x_1^2} = \int \frac{\phi}{(x_1 + x^{-1})^2} \hat{\mu}_T(dx) \geq \frac{4\delta^2}{(x_1 + (\tau^{-1} - \delta)^{-1})^2} \Rightarrow x_1 \geq \frac{1}{(1 + 2\delta)(\tau^{-1} - \delta)} \geq -\frac{\tau}{1 + \frac{1}{2}\delta}.$$

Therefore

$$a_1 = f(x_1) = -\frac{1}{x_1} + \int \frac{\phi}{x_1 + x^{-1}} \hat{\mu}_T(dx) \geq -\frac{1}{x_1} \geq \tau^{-1}(1 + \frac{1}{2}\delta).$$

Also note that for $m \in (-1, 0)$,

$$2mf'(m) + m^2 f''(m) = (m^2 f'(m))' = \int \frac{2\phi m^{-1} x^{-2}}{(1 + (xm)^{-1})^3} \hat{\mu}_T(dx) < 0$$

In particular we have

$$-f''(x_1) = \frac{1}{|x_1|^3} \int_\tau^{\tau^{-1}} \frac{2\phi x^{-2}}{(1 + (xx_1)^{-1})^3} \hat{\mu}_T(dx),$$

which is a positive number with uniform (in T) lower bound. Since $f(m)$ is analytic in $(-\sigma_{(1)}, 0)$, $|f^{(k)}(x_1)|$ is also uniformly bounded. Thus, we have the expansion

$$f(m) = a_1 + f''(x_1)(m - x_1)^2 + O(|m - x_1|^3).$$

In particular, for $z = a_1 - \delta + \delta\eta i$, by equating $z = f(m)$, we find

$$m(z) = i\sqrt{\frac{\delta}{-f''(x_1)}} \exp(-i\frac{1}{2}\text{arctan}\eta) + O(\delta\sqrt{\delta})$$

Then the inverse Stieltjes transform gives

$$\rho_T(a_1-\delta)=\lim_{\eta\to 0+}\frac{\mathrm{Im}m(a_1-\delta+\delta\eta i)}{2\pi}=\sqrt{\frac{\delta}{-f''(x_1)}}+O(\delta\sqrt{\delta}). \tag{C.9}$$

Define the typical singular value location $\gamma_{T,k}$ as the $1-\frac{k}{T}+\frac{1}{2T}$ quantile of $\rho_T(x)$, i.e.

$$\int_{\gamma_{T,k}}^{\infty}\rho_T(x)dx=\frac{k}{T}-\frac{1}{2T}.$$

When the support of ρ consists of multiple components, i.e. $p>1$, each $\gamma_{T,k}$ will belong to one of the component $[a_{2j},a_{2j-1}]$. Denote the N_k to be the rank of $\gamma_{T,k}$ among $\{\gamma_{T,i},i\in[\min(N,T)]\}\cap[a_{2j},a_{2j-1}]$. Denote the N_k^- to be the reverse rank. Theorem 3.12 of Knowles and Yin (2017) shows that with high probability, the following holds for all

$$|\lambda_k(\Sigma_e^{1/2}E/\sqrt{T})-\gamma_{T,k}|\le\min\{N_k,N_k^-\}^{-1/3}T^{\varepsilon-2/3}.$$

To obtain our first claim, it suffices to consider $\gamma_{T,j}$ for $j\le T^{\varepsilon}$, which are given by $\gamma_{T,j}=a_1-\delta_j$ that satisfies

$$\begin{aligned}\frac{j}{T}-\frac{1}{2T}=\int_{\gamma_{n,j}}^{a_1}\rho(x)dx&=\int_0^{\delta_j}\sqrt{\frac{\delta}{-f''(x_1)}}d\delta+O(\delta_j^2\sqrt{\delta_j})\\&=\frac{2}{3}\sqrt{\frac{\delta_j^3}{-f''(x_1)}}+O(\delta_j^2\sqrt{\delta_j}).\end{aligned}$$

Solving this equation gives us a constant C so that

$$C(\frac{j}{T})^{\frac{2}{3}}\ge\delta_j\ge\frac{1}{C}(\frac{j}{T})^{\frac{2}{3}},\quad j\le T^{\varepsilon}.$$

In particular,

$$\gamma_{T,R}\ge a_1-\delta_R\ge a_1-C(\frac{R}{T})^{\frac{2}{3}}$$

And obtain claim 1:

$$\lambda_R(\Sigma_e^{1/2}E)\ge\sqrt{n}(a_1-C(\frac{R}{T})^{\frac{2}{3}}-CT^{\varepsilon-2/3})\ge\sqrt{T}a_1-CT^{\varepsilon-1/6}.$$

Meanwhile, if $T^{1/3}\ge j\ge T^{3\varepsilon}$, then $\gamma_{T,j}\in[a_2,a_1]$ with high probability, with

$$\lambda_j(\Sigma_e^{1/2}E)\le\sqrt{T}(a_1-\frac{1}{C}(\frac{j}{T})^{\frac{2}{3}}+Cj^{-1/3}T^{\varepsilon-2/3})\le\sqrt{T}a_1-\frac{1}{2C}j^{\frac{2}{3}}T^{-\frac{1}{6}}\le\sqrt{T}a_1-\frac{1}{2C}j^{\frac{2}{3}}T^{-\frac{1}{6}}$$

Therefore

$$\lambda_j(\Sigma_e^{1/2}E)^2\le Ta_1^2-\frac{1}{2C}j^{\frac{2}{3}}T^{\frac{1}{3}}\le Ta_1^2-\frac{1}{2C}j.$$

Meanwhile, if $j \geq T^{1/3}$, Lemma 4.10 of Knowles and Yin (2017) has shown that the density is bounded $\rho_T(x) \leq C$, so using

$$\frac{j}{T} - \frac{1}{2T} = \int_{\gamma_{T,j}}^{a_1} \rho_T(x)dx \leq C(a_1 - \gamma_{T,j})$$

we find that $a_1 - \gamma_{T,j} \geq \frac{j}{2CT}$. Furthermore

$$\lambda_j(\Sigma_e^{1/2}E) \leq \sqrt{T}(a_1 - \frac{j}{2CT} + Cj^{-1/3}T^{\varepsilon-2/3}) \leq \sqrt{T}a_1 - \frac{j}{4C\sqrt{T}}.$$

Therefore

$$\lambda_j(\Sigma_e^{1/2}E)^2 \leq Ta_1^2 - \frac{a_1}{4C}j.$$

Combining this with claim 1 yields claim 2.

As for claim 3, (anisotropic delocalization of $u_{0,k}$ and $w_{0,k}$): Remark 3.13 of Knowles and Yin (2017) along with Theorem 3.13 in Bloemendal et al. (2014) indicate that for any unit norm $\eta_i \in \mathbb{R}^N$ and $\zeta_i \in \mathbb{R}^T$ that are independent of E, any fixed $\varepsilon > 0$ and $D > 0$, the following take place

$$\mathbb{P}(|\langle \eta_i, u_{0,k}\rangle|^2 + |\langle \zeta_i, w_{0,k}\rangle|^2 \leq N^{\varepsilon-1}) \leq N^{-D-1}.$$

Our claim applies a union bound over all $\eta_i \in \mathcal{A}$ and $\zeta_i \in \mathcal{B}$. ∎

D Proofs for Section 2

D.1 Proof of Corollary 2.1

Lemma D.1 (if $r \geq 1$)**.** *Suppose $T \asymp N$ and $T^\varepsilon = o(\nu_M^2)$. Then*

(i) There is $r \times r$ matrix H_r so that $\frac{1}{\sqrt{N}}\|\widehat{B}_r - BH_r\| = O_P(\frac{1}{\nu_M})$

(ii) $\lambda_{\min}(H_r) \asymp \lambda_{\max}(H_r) = O_P(\nu_{\min}^{-1/2})$

(iii) $\frac{1}{TN}\|G_T'U'\widehat{B}_r\| = O_P(T^{-1}\nu_{\min}^{-1/2})$ for any $T \times 1$ vector G_T that is independent of U and that $\|G_T\| = O_P(\sqrt{T})$.

(iv) Recall $H := \frac{1}{N}\widehat{B}'B$, $R \times r$. Then $\lambda_{\min}(H) \asymp \lambda_{\max}(H) = O_P(\nu_{\min}^{1/2})$.

Proof. Because $\nu_{\min}^{-1} \asymp N\nu_M^{-2}$. and $T \asymp N$. Hence $T^\varepsilon = o(\nu_M^2)$ is equivalent to $T^{\varepsilon-1} = o(\nu_{\min})$.

(i) Let that $\widetilde{L}_r$ be the $r \times r$ diagonal matrix of top r eigenvalues of XX'/T. By definition,

$$XX'\widehat{B}_r = T\widehat{B}_r\widetilde{L}_r.$$

With $X = BF' + U$, we can define $H_r := \frac{1}{T}F'X'\widehat{B}_r\widetilde{L}_r^{-1}$. Then

$$\widehat{B}_r - BH_r = \frac{1}{T}UX'\widehat{B}_r\widetilde{L}_r^{-1}.$$

Classical PCA analysis shows $\widetilde{L}_r^{-1} = O_P(\nu_M^{-2})$, also $\|X\| = O_P(\sqrt{T}\nu_M)$, $\|U\| = O_P(T^{1/2})$. Thus

$$\frac{1}{\sqrt{N}}\|\widehat{B}_r - BH_r\| = O_P(\nu_M^{-1}).$$

(ii) The desired result is standard analysis the first r eignvectors, whose proof appears in many places when factors are strong. We provide a proof below for completeness, which also allows weaker factors. First, $\|\widetilde{L}_r^{-1}\| = O_P(\nu_M^{-2})$, and $\|X\| = O_P(\sqrt{T}\nu_M)$. With $\nu_{\min} \asymp \nu_M^2/N$ and $T \asymp N$, we have

$$\|H_r\| \leq \frac{\sqrt{T}}{T}\|F'X\|\|\widetilde{L}_r^{-1}\| = O_P(\nu_{\min}^{-1/2}).$$

To bound the minimum singular value of H_r. Note $\frac{1}{N}\widehat{B}_r'\widehat{B}_r = I_r = H_r'\frac{1}{N}B'BH_r + o_{\mathrm{P}}(1)$. Let x denote the eigenvector of $H_r'H_r$ corresponding to its minimum eigenvalue. Then

$$\begin{aligned}\lambda_{\min}(H_r)^2 &= \lambda_{\min}(H_r'H_r) = x'H_r'H_rx \geq \frac{x'H_r'\frac{1}{N}B'BH_rx}{\lambda_{\max}(\frac{1}{N}B'B)} \geq \frac{\lambda_{\min}(H_r'\frac{1}{N}B'BH_r)}{\lambda_{\max}(\frac{1}{N}B'B)} \\ &\geq \frac{N}{\nu_M^2}(1 - o_P(1)) \geq \nu_{\min}^{-1}(1 - o_P(1)).\end{aligned}$$

(iii) We bound the two pieces of the decomposition $\widehat{B}_r = BH_r + (\widehat{B}_r - BH_r)$:

$$\tfrac{1}{NT}\|\widehat{B}_r'UG_T\| \;\leq\; \tfrac{\|H_r\|}{NT}\,\|B'UG_T\| \;+\; \tfrac{1}{NT}\,\|\widehat{B}_r - BH_r\|\,\|UG_T\|.$$

For the first piece,

$$\mathbb{E}\|B'UG_T\|^2 = \sum_t \sum_k \mathbb{E}g_t^2 B_k' \mathbb{E}(u_t u_t'|g_t)B_k \leq CrT\|B\|^2 = O(\nu_M^2 T).$$

So with $\|H_r\| = O_P(\nu_{\min}^{-1/2})$ and $\nu_M \asymp \sqrt{N\nu_{\min}}$,

$$\tfrac{\|H_r\|}{NT}\,\|B'UG_T\| = O_P(\frac{1}{\sqrt{NT}}).$$

For the second piece, $\|UG_T\| \leq \|U\|\,\|G_T\| = O_P(T)$ and $\|\widehat{B}_r - BH_r\| = O_P(\sqrt{N}\,\nu_M^{-1})$ from (i), so

$$\tfrac{1}{NT}\,\|\widehat{B}_r - BH_r\|\,\|UG_T\| \;=\; O_P\Big(\tfrac{\sqrt{N}\,\nu_M^{-1}\cdot T}{NT}\Big) \;=\; O_P(T^{-1}\nu_{\min}^{-1/2}).$$

Combining the two pieces gives the claim.

(iv) By (iii), because $T^{-1} = O(\nu_{\min}^{1/2})$,

$$\frac{1}{N}B'\widehat{B}_r = \frac{1}{N}S_f^{-1}H_r\widehat{L}_r - \frac{1}{TN}S_f^{-1}F'U'\widehat{B}_r = \frac{1}{N}S_f^{-1}H_r\widehat{L}_r + O_P(T^{-1}\nu_{\min}^{-1/2}).$$

This shows $\lambda_{\min}(\frac{1}{N}B'\widehat{B}_r) \asymp \nu_{\min}^{1/2}$. Also note that for $A := \frac{1}{N}B'\widehat{B}_r(\frac{1}{N}B'\widehat{B}_r)'$,

$$H'H = A + B'\widehat{B}_{-r}\widehat{B}'_{-r}B = A + O_P(\frac{1}{N}H_r'^{-1}(H_r'B' - \widehat{B}'_r)\widehat{B}_{-r}) = A + o_P(\nu_{\min}).$$

This implies $\lambda_{\min}(H) \asymp \lambda_{\max}(H) = O_P(\nu_{\min}^{1/2})$.

■

Proof of Corollary 2.1

Proof. First, Lemma D.1 shows $\frac{1}{TN}\|G_T'U'\widehat{B}_r\| = O_P(T^{-1}\nu_{\min}^{-1/2})$. Next,

$\|UG_T\| \leq O_P(T)$. In addition, let $\eta = G_T/\|G_T\|$. Let $\widehat{L}_{-r}$ be the $R-r$ diagonal matrix of X corresponding to the top $R-r$ singular values. Then by Theorem 2.1 (i)(ii),

$$\|\widehat{L}_{-r}\widehat{V}_{-r}\eta\| \leq \|\widehat{L}_{-r}\|\|\widehat{V}_{-r}\eta\| \prec T^{1/2}\nu_M^{-1}.$$

Finally, by Theorem 2.1(iii),

$$\begin{aligned}
\frac{1}{NT}\|\widehat{B}'_{-r}UG_T\| &= O_P(\frac{\sqrt{NT}}{NT})\|\widehat{\Xi}'_{-r}U\eta\| = O_P(\frac{\sqrt{NT}}{NT})\|\widehat{\Xi}'_{-r}X\eta\| + O_P(\frac{\sqrt{NT}}{NT})\|\widehat{\Xi}'_{-r}M\eta\| \\
&\leq O_P(\frac{\sqrt{NT}}{NT})\|\widehat{L}_{-r}\widehat{V}_{-r}\eta\| + O_P(\frac{\sqrt{NT}}{NT})\|\widehat{\Xi}'_{-r}BF'\eta\| \\
&= O_P\left(T^{\varepsilon-1/2}\nu_M^{-1} + \frac{\sqrt{T}\|B\|_{\mathrm{F}}}{T}\cdot\nu_M^{-2}\right) = O_P(T^{\varepsilon-1/2}\nu_M^{-1}).
\end{aligned}$$

■

D.2 Proof of Theorem 2.2

Proof. Write the SVD of $M = \Xi_r L_r V_r'$. By the Davis–Kahan $\sin\Theta$ theorem, for the SVD of X there exists a rotation O_r such that $\|\widehat{V}_r - V_r O_r\|_{\mathrm{F}} = O_P(\nu_M^{-1})$. Note that $\widehat{M} = X\widehat{V}_R\widehat{V}_R'$ and $MV_rV_r' = M$. We decompose

$$\begin{aligned}
X\widehat{V}_R\widehat{V}_R' - M &= X\widehat{V}_r\widehat{V}_r' - M + X\widehat{V}_{-r}\widehat{V}_{-r}' \\
&= M(\widehat{V}_r\widehat{V}_r' - V_rV_r') + U\widehat{V}_r\widehat{V}_r' + M\widehat{V}_{-r}\widehat{V}_{-r}' + U\widehat{V}_{-r}\widehat{V}_{-r}'.
\end{aligned}$$

We bound: $\|U\widehat{V}_{-r}\widehat{V}_{-r}'\|_{\mathrm{F}} = O_P(\sqrt{T})$,

$$\|M(\widehat{V}_r\widehat{V}_r' - V_rV_r')\|_{\mathrm{F}} = O_P(\sqrt{T}), \quad \|U\widehat{V}_{-r}\widehat{V}_{-r}'\|_{\mathrm{F}} \leq \|\widehat{V}_{-r}\|_{\mathrm{F}}^2\|U\| = O_P(\sqrt{T})$$

and using Theorem 2.1 claim(ii), we have $\|M\widehat{V}_{-r}\widehat{V}_{-r}'\|_{\mathrm{F}} \prec \nu_M^{-1}\sqrt{T}$. Together,

$$\frac{1}{\sqrt{NT}}\|\widehat{M} - M\|_{\mathrm{F}} = O_p(\frac{1}{\sqrt{NT}}(\sqrt{T} + \nu_M^{-1}T^{\varepsilon}\sqrt{T})) = O_P(T^{-1/2})$$

given $T \asymp N$. The same conclusion holds in the case $r = 0$ ($M = 0$), where the first three terms in $X\widehat{V}_R\widehat{V}_R' - M$ vanish and the remaining term $\|U\widehat{V}_{-r}\widehat{V}_{-r}'\|_{\mathrm{F}} \le O_P(\sqrt{T})$.

■

D.3 Proof of Theorem 2.4

Proof. In this theorem we assume for some $\varepsilon > 0$, $T^{\varepsilon-1} = o(\nu_{\min})$. Note $H'H^+ = I_r$.

(i)(ii) By Lemma D.1, $\|H\| \le O_P(\nu_{\min}^{1/2})$. In addition, the definition of $\widehat{F}$ and $X = BF' + U$ give

$$\widehat{F} = FH' + \frac{1}{N}U'\widehat{B}, \quad \widehat{F}H^+ = F + \frac{1}{N}U'\widehat{B}H^+.$$

Recalling $T \asymp N$, we have

$$\begin{aligned}
\frac{1}{\sqrt{T}}\|\widehat{F}H^+ - F\| &= \frac{1}{\sqrt{T}}\|\frac{1}{N}U'\widehat{B}H^+\| = \frac{1}{N}\|U\|O_P(\nu_{\min}^{-1/2}) = O_P(T^{-1/2}\nu_{\min}^{-1/2}) \\
\frac{1}{\sqrt{T}}\|\widehat{F} - FH'\| &\le \frac{1}{N}\|U\| \le O_P(T^{-1/2}).
\end{aligned}$$

Now set $G_T = F$ in Corollary 2.1 to reach

$$\|\frac{1}{TN}F'U'\widehat{B}\| = O_p(T^{\varepsilon-1}\nu_{\min}^{-1/2}).$$

Hence

$$\begin{aligned}
\frac{1}{T}F'(\widehat{F}H^+ - F) &\le \|\frac{1}{TN}F'U'\widehat{B}\|O_P(\nu_{\min}^{-1/2}) = O_P(T^{\varepsilon-1}\nu_{\min}^{-1}) \\
\frac{1}{T}F'(\widehat{F} - FH') &\le \|\frac{1}{TN}F'U'\widehat{B}\| = O_P(T^{\varepsilon-1}\nu_{\min}^{-1/2}).
\end{aligned}$$

(iii) From (2.1), $\frac{1}{\sqrt{N}}B = \Xi_r H_B^{-1}$. This implies $\Xi_r L_r V_r' = M = BF' = \sqrt{N}\Xi_r H_B^{-1}F'$, yielding

$$F' = N^{-1/2}H_B L_r V_r', \quad F'F = \frac{1}{N}H_B L_r^2 H_B', \quad (F'F)^{-1} = NH_B'^{-1}L_r^{-2}H_B^{-1}.$$

Meanwhile, $(\widehat{F}'\widehat{F})^{-1} = N\widehat{L}_R^{-2}$, and $H = \frac{1}{N}\widehat{B}'B = \widehat{\Xi}_R'\Xi_r H_B^{-1}$, yielding

$$H'\left(\widehat{F}'\widehat{F}\right)^{-1}H = NH_B'^{-1}\Xi_r'\widehat{\Xi}_R\widehat{L}_R^{-2}\widehat{\Xi}_R'\Xi_r H_B^{-1}.$$

Also, $\|H_B^{-1}\| \le \|S_B\|^{1/2} \lesssim \nu_{\min}^{1/2})$. This implies, for $\nu_M^2 \asymp N\nu_{\min}$,

$$\begin{aligned}
\|H'\left(\widehat{F}'\widehat{F}\right)^{-1}H - (F'F)^{-1}\| &\le N\|H_B^{-1}\|^2\|\Xi_r'\widehat{\Xi}_R\widehat{L}_R^{-2}\widehat{\Xi}_R'\Xi_r - L_r^{-2}\| \\
&\lesssim \nu_M^2\|\Xi_r'\widehat{\Xi}_r\widehat{L}_r^{-2}\widehat{\Xi}_r'\Xi_r' - L_r^{-2}\| + \nu_M^2\|\Xi_r'\widehat{\Xi}_{-r}\widehat{L}_{-r}^{-2}\widehat{\Xi}_{-r}'\Xi_r'\| \\
&\lesssim \nu_M^2\|\widehat{\Xi}_r\widehat{L}_r^{-2}\widehat{\Xi}_r' - \Xi_r L_r^{-2}\Xi_r'\| + \nu_M^2\|\Xi_r'\widehat{\Xi}_{-r}\|^2\|\widehat{L}_{-r}^{-2}\| \\
&\prec T^{-1}\nu_M^{-1} + T^{-1}\nu_M^{-2},
\end{aligned}$$

where the second last line uses Lemma C.1, and the last line first use Davis–Kahan theorem for the Ξ_r, L_r perturbation, and the second item follows from Theorem 2.1. Hence

$$H'\left(\frac{1}{T}\widehat{F}'\widehat{F}\right)^{-1}H - \left(\frac{1}{T}F'F\right)^{-1} \prec \nu_M^{-1} + \nu_M^{-2} = o_P(1)$$

given $T^{\varepsilon-1} = o(\nu_{\min})$.

■

D.4 Proof of Proposition 2.3

Proof. Recall $H' = \frac{1}{N}B'\widehat{B} = \frac{1}{\sqrt{N}}B'\widehat{\Xi}_R$ since $\widehat{B} = \sqrt{N}\widehat{\Xi}_R$. Using identity (2.1), $\frac{1}{\sqrt{N}}B = \Xi_r H_B^{-1}$, hence

$$H' \;=\; H_B^{-1'}\,\Xi_r'\widehat{\Xi}_R \;=\; H_B^{-1'}\,[\,\Xi_r'\widehat{\Xi}_r,\;\Xi_r'\widehat{\Xi}_{-r}\,].$$

The two blocks contribute on different scales: the leading $r \times r$ block $\Xi_r'\widehat{\Xi}_r$ is bounded below by Davis–Kahan, while the $r \times (R-r)$ overestimated block $\Xi_r'\widehat{\Xi}_{-r}$ is asymptotically negligible by Theorem 2.1(iii). Also $\frac{1}{\sqrt{N}}\|H_B'\|\|B\| = O(1)$ by (B.1). Then by Theorem 2.1(iii),

$$\|\Xi_r'\widehat{\Xi}_{-r}\| \le \frac{1}{\sqrt{N}}\|H_B'\|\|B\|\|B\|^{-1}\|B'\widehat{\Xi}_{-r}\| \prec \nu_M^{-2} = o_P(1).$$

By the Davis–Kahan $\sin\Theta$ theorem combined with the singular-value separation in Theorem 2.2, $\lambda_r(\Xi_r'\widehat{\Xi}_r) \ge 1 - O_P(\nu_M^{-1}) \ge \frac{1}{2}$ with probability tending to one. Together,

$$\lambda_r(\Xi_r'\widehat{\Xi}_R) \;\ge\; \lambda_r(\Xi_r'\widehat{\Xi}_r) - \|\Xi_r'\widehat{\Xi}_{-r}\| \;\ge\; \tfrac{1}{4}$$

with probability tending to one. Since Assumption 2.2(iii) implies that the singular values of H_B^{-1} are of order $\nu_{\min}^{1/2}$, it follows that

$$\lambda_r(H') \;\ge\; \lambda_r(H_B^{-1'})\lambda_r(\Xi_r'\widehat{\Xi}_R) \;\ge\; c_0\nu_{\min}^{1/2}$$

for a constant $c_0 > 0$ depending only on (c, C, ϕ) in Assumptions 2.1–2.2, with probability $1 - o(1)$. Finally, $\|H^+\| = 1/\lambda_r(H) \le c_0^{-1}\nu_{\min}^{-1/2} = O_P(\nu_{\min}^{-1/2})$, which proves the claim. ■

E Proofs for Section 3

Proof of Theorem 3.1

Proof. The estimator is the just-identified IV slope $\widehat{\beta} = (\widehat{\varepsilon}_z'\widehat{\varepsilon}_g)^{-1}\widehat{\varepsilon}_z'\widehat{\varepsilon}_y$ on the residualization of (Y, G, Z) on $[1_T, \widehat{F}]$. By the Frisch–Waugh–Lovell identity,

$$\widehat{\varepsilon}_a \;=\; (I - P_{[1_T,\widehat{F}]})A \;=\; (I - P_{\widehat{F}^c})\widetilde{A}, \qquad A \in \{Y, G, Z\},$$

where $\widetilde{A} := A - \bar{a}\,1_T$ is the sample-demeaned regressor and $\widehat{F}^c := (I - P_{1_T})\widehat{F}$ is the column-demeaned PCA factor estimator (with the orthonormal basis taken via QR). Substituting the model (3.1) kills the intercepts μ_y, μ_g, μ_z inside the demeaning,

$$\widetilde{Y} \;=\; F^c\alpha_y + \widetilde{\varepsilon}_y, \qquad \widetilde{G} \;=\; F^c\alpha_g + \widetilde{\varepsilon}_g, \qquad \widetilde{Z} \;=\; F^c\alpha_z + \widetilde{\varepsilon}_z,$$

where $F^c := (I - P_{1_T})F$, $\widetilde{\varepsilon}_a := \varepsilon_a - \bar{\varepsilon}_a\,1_T$ for $a \in \{y, g, z\}$, $\alpha_y = \beta\alpha_g + \rho$, and $\varepsilon_y = \beta\varepsilon_g + \eta$.

The column-demeaning $\widehat{F} \mapsto \widehat{F}^c$ subtracts the rank-one term $1_T\,(T^{-1}1_T'\widehat{F})$, whose operator norm is $O_P(1)$ since $T^{-1}1_T'\widehat{F} = O_P(T^{-1/2})$ by Theorem 2.4(i) combined with the central limit theorem $T^{-1/2}1_T'F = O_P(1)$ for the mean-zero f_t. The same is true of F^c relative to F. Hence Theorems 2.1–2.4 continue to hold with $\widehat{V}_r, \widehat{V}_R, \widehat{V}_{-r}$ replaced by their column-demeaned counterparts $\widehat{V}_r^c, \widehat{V}_R^c, \widehat{V}_{-r}^c$ (the right singular vectors of the column-demeaned panel). Below we work directly with $\widehat{V}_r^c, \widehat{V}_R^c, \widehat{V}_{-r}^c$, V_r^c, and write $P_{\widehat{F}^c} = \widehat{V}_R^c\widehat{V}_R^{c\prime}$, $P_{F^c} = V_r^cV_r^{c\prime}$.

Step 1: residual approximation. For each $a \in \{y, g, z\}$,

$$\widetilde{\varepsilon}_a - \widehat{\varepsilon}_a = (P_{\widehat{F}^c} - P_{F^c})\widetilde{A} + P_{F^c}\widetilde{\varepsilon}_a = (\widehat{V}_r^c\widehat{V}_r^{c\prime} - V_r^cV_r^{c\prime})\widetilde{A} + \widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{A} + P_{F^c}\widetilde{\varepsilon}_a. \tag{E.1}$$

By the sin-theta theorem, $\|\widehat{V}_r^c - V_r^c\| \le O_P(\nu_M^{-1})$, so

$$\frac{1}{\sqrt{T}}\|(\widehat{V}_r^c\widehat{V}_r^{c\prime} - V_r^cV_r^{c\prime})\widetilde{A}\|_{\mathrm{F}} \le O_P(\nu_M^{-1}). \tag{E.2}$$

Write $\widetilde{A} = F^c\alpha_a + \widetilde{\varepsilon}_a$. Meanwhile, by Theorem 2.1(iii) for the first term, and using the i.i.d. structure of Assumption 3.1(i) and that $\widetilde{\varepsilon}_a$ is independent of X to obtain $\|V_r^cV_r^{c\prime}\widetilde{\varepsilon}_a\| + \|\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{\varepsilon}_a\| \le O_P(1)$.

$$\begin{aligned}\frac{1}{\sqrt{T}}\|\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{A}\| \;&\le\; \frac{1}{\sqrt{T}}\|\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}F^c\alpha_a\| + \frac{1}{\sqrt{T}}\|\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{\varepsilon}_a\| \\ &\le\; O_P(\frac{1}{\sqrt{T}})\|\widehat{V}_{-r}^{c\prime}F^c\| + O_P(T^{-1/2}) \prec \nu_M^{-2} + T^{-1/2}.\end{aligned} \tag{E.3}$$

Also $\frac{1}{\sqrt{T}}\|P_{F^c}\widetilde{\varepsilon}_a\| \le O_P(T^{-1/2})$. Hence $\frac{1}{\sqrt{T}}\|\widetilde{\varepsilon}_a - \widehat{\varepsilon}_a\| \prec \nu_M^{-1} + \nu_M^{-2} + T^{-1/2}$, for $a \in \{y, g, z\}$. Then under $\sqrt{T} = o(\nu_M^2)$,

$$\frac{1}{\sqrt{T}}\|\widetilde{\varepsilon}_a - \widehat{\varepsilon}_a\|^2 \prec \nu_M^{-2}\sqrt{T} + \nu_M^{-4}\sqrt{T} + T^{-1/2} = o_P(1).$$

Step 2: cross-product approximation. We show $\frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'(\widetilde{\varepsilon}_b - \widehat{\varepsilon}_b) = o_P(1)$ for every $a, b \in \{y, g, z\}$. Decompose

$$\frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'(\widetilde{\varepsilon}_b - \widehat{\varepsilon}_b) \;=\; (I) + (II)$$

$$
\begin{aligned}
(I) &= \frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'(\widehat{V}_r^c\widehat{V}_r^{c\prime} - V_r^cV_r^{c\prime})F^c\alpha_b \\
(II) &= \frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'\widehat{V}_r^c\widehat{V}_r^{c\prime}\widetilde{\varepsilon}_b - \frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'V_r^cV_r^{c\prime}\widetilde{\varepsilon}_b + \frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{B} + \frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'P_{F^c}\widetilde{\varepsilon}_b
\end{aligned}
$$

where $\widetilde{B} \in \{\widetilde{Y}, \widetilde{G}, \widetilde{Z}\}$. Let $J := \frac{1}{\sqrt{T}}(\widehat{V}_r^c\widehat{V}_r^{c\prime} - V_r^cV_r^{c\prime})F^c\alpha_b$. By (E.2), $\|J\| = O_P(\nu_M^{-1})$, so $\mathbb{E}[(I)^2 \mid F, U] = J'\mathbb{E}[\widetilde{\varepsilon}_a\widetilde{\varepsilon}_a' \mid F, U]J \le C\|J\|^2 = O_P(\nu_M^{-2})$, hence $(I) = o_P(1)$. For (II), using $(\widehat{V}_r^c\widehat{V}_r^{c\prime})^2 = \widehat{V}_r^c\widehat{V}_r^{c\prime}$ and (E.3),

$$
\Big|\frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{B}\Big| \le \Big\|\frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\Big\|\,\|\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{B}\| \prec T^{-1/2}(1 + T^{1/2}\nu_M^{-2}) = o_P(1).
$$

For the remaining three projection terms, with $\|A\|_{(n)}$ denoting the nuclear norm and $A \in \{\widehat{V}_r^c, V_r^c, F^c\}$,

$$
\mathbb{E}\Big|\frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'P_A\widetilde{\varepsilon}_b\Big| = \mathrm{tr}\Big(\frac{1}{\sqrt{T}}P_A\mathbb{E}[\widetilde{\varepsilon}_b\widetilde{\varepsilon}_a' \mid F, U]\Big) \le \frac{C}{\sqrt{T}}\|P_A\|_{(n)} \le \frac{C}{\sqrt{T}}\,\mathrm{rank}(A) = O(T^{-1/2}),
$$

so $(II) = o_P(1)$. Combining,

$$
\frac{1}{\sqrt{T}}\widehat{\varepsilon}_a'\widehat{\varepsilon}_b = \frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'\widetilde{\varepsilon}_b + o_P(1) = \frac{1}{\sqrt{T}}\varepsilon_a'\varepsilon_b + o_P(1), \qquad a, b \in \{y, g, z\}, \tag{E.4}
$$

where the last equality uses the $O_P(T^{-1/2})$ rank-one correction noted in the preamble.

Step 3: asymptotic distribution. Since $\varepsilon_y = \beta\varepsilon_g + \eta$, applying (E.4) with $(a, b) = (z, y)$ and $(a, b) = (z, g)$,

$$
\frac{1}{\sqrt{T}}\widehat{\varepsilon}_z'\widehat{\varepsilon}_y - \beta\,\frac{1}{\sqrt{T}}\widehat{\varepsilon}_z'\widehat{\varepsilon}_g = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_{z,t}\eta_t + o_P(1),
$$

and $T^{-1}\widehat{\varepsilon}_z'\widehat{\varepsilon}_g = T^{-1}\varepsilon_z'\varepsilon_g + o_P(1) \to_p \gamma$ by the law of large numbers and the relevance condition $\gamma = \mathbb{E}[\varepsilon_{g,t}\varepsilon_{z,t}] \neq 0$ in Assumption 3.1(i). Hence

$$
\sqrt{T}(\widehat{\beta} - \beta) = \gamma^{-1}\,\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_{z,t}\eta_t + o_P(1). \tag{E.5}
$$

By the exclusion restriction $\mathbb{E}[\eta_t\varepsilon_{z,t}] = 0$ in Assumption 3.1(i), the i.i.d. score $\varepsilon_{z,t}\eta_t$ has mean zero, and by Assumption 3.1(iii) and Hölder $\mathbb{E}[\varepsilon_{z,t}^2\eta_t^2] \le (\mathbb{E}\varepsilon_{z,t}^4)^{1/2}(\mathbb{E}\eta_t^4)^{1/2} < \infty$. Lyapunov's central limit theorem (using the eighth moment in (iii) for the Lyapunov ratio) yields

$$
T^{-1/2}\sum_{t=1}^{T}\varepsilon_{z,t}\eta_t \to^d \mathcal{N}\big(0, \mathbb{E}[\varepsilon_{z,t}^2\eta_t^2]\big),
$$

so $\sqrt{T}\sigma^{-1}(\widehat{\beta} - \beta) \to^d \mathcal{N}(0, 1)$, where $\sigma^2 := \gamma^{-2}\,\mathbb{E}[\varepsilon_{z,t}^2\eta_t^2] > 0$ is the population just-

identified IV sandwich variance.

Step 4: Entrywise bound. For the consistency of the HC_0 variance estimator, it will be useful to prove the following result. For a "tall" matrix v whose dimension is $T \times l$ with $l = O(1)$ with v_t' as the t th row, define $\|v\|_\infty = \max_{t \le T} \|v_t\|$. We prove $\|\widehat{\varepsilon}_a - \widetilde{\varepsilon}_a\|_\infty = o_P(1)$. From (E.1),

$$\|\widetilde{\varepsilon}_a - \widehat{\varepsilon}_a\|_\infty \le \|(\widehat{V}_r^c \widehat{V}_r^{c\prime} - V_r^c V_r^{c\prime})\widetilde{A}\|_\infty + \|\widehat{V}_{-r}^c \widehat{V}_{-r}^{c\prime} \widetilde{A}\|_\infty + \|P_{F^c}\widetilde{\varepsilon}_a\|_\infty.$$

By Assumption 2.2, and identity (2.1), $\|V_r^c\|_\infty = \frac{1}{\sqrt{T}}\|F^c\|_\infty \|H_F\| = O_P(T^{-1/2})$. Also by the entrywise control of first r eigenvectors (e.g., Theorem 1 of Fan et al. (2021); in their notation $\eta_N = O_P(\frac{N}{\sqrt{T}})$, $c_N = \sqrt{\frac{\log N}{T}}$, $g_N = \nu_M \sqrt{T}$, $\|S\| = O(1)$, $\|S\xi_d\|_\infty \le \|S\|_1 \|\xi_d\|_\infty \le O_P(\nu_M^{-1})\|B\|_\infty = O_P(\nu_M^{-1})$),

$$\|\widehat{V}_r^c - V_r^c\|_\infty \le O_P(\nu_M^{-2} + \sqrt{\frac{\log N}{T}}\nu_M^{-1}).$$

Then

$$\begin{aligned} & \|(\widehat{V}_r^c \widehat{V}_r^{c\prime} - V_r^c V_r^{c\prime})\widetilde{A}\|_\infty \le \|\widehat{V}_r^c - V_r^c\|_\infty \|\widehat{V}_r^{c\prime}\widetilde{A}\| + \|\widehat{V}_r^c\|_\infty \|(\widehat{V}_r^c - V_r^c)'\widetilde{A}\| \\ \le\ & \|\widehat{V}_r^c - V_r^c\|_\infty O_P(\sqrt{T}) + (\|\widehat{V}_r^c - V_r^c\|_\infty + \|V_r^c\|_\infty)\|\widehat{V}_r^c - V_r^c\| O_P(\sqrt{T}) \\ \le\ & O_P(\nu_M^{-2}\sqrt{T} + \sqrt{\log N}\nu_M^{-1}) = o_P(1). \end{aligned}$$

Next, by the incoherency of overestimated eigenvectors (Theorem 2.1(ii)),

$$\|\widehat{V}_{-r}^c \widehat{V}_{-r}^{c\prime}\widetilde{A}\|_\infty \le \|\widehat{V}_{-r}^c\|_\infty \|\widehat{V}_{-r}^{c\prime}\widetilde{A}\| \prec \nu_M^{-1}(1 + \sqrt{T}\nu_M^{-2}) = o_P(1).$$

Finally, $\|P_{F^c}\widetilde{\varepsilon}_a\|_\infty \le O_P(T^{-1})\|F^c\|_\infty \|F^{c\prime}\widetilde{\varepsilon}_a\| = O_P(T^{-1/2}) = o_P(1)$.

This proves $\|\widetilde{\varepsilon}_a - \widehat{\varepsilon}_a\|_\infty = o_P(1)$.

Step 5: HC_0 sandwich consistency. It remains to show that the Eicker–White (HC_0) estimator $\widehat{\sigma}^2 = (\widehat{\varepsilon}_z'\widehat{\varepsilon}_g/T)^{-2}\, T^{-1}\sum_t \widehat{\varepsilon}_{z,t}^2 \widehat{\eta}_t^2$ defined in Theorem 3.1 satisfies $\widehat{\sigma}^2 = \sigma^2 + o_P(1)$. The outer factor satisfies $\widehat{\varepsilon}_z'\widehat{\varepsilon}_g/T \to_p \gamma$ by Step 3. For the meat, write $\widetilde{\eta}_t := \widetilde{\varepsilon}_{y,t} - \beta\widetilde{\varepsilon}_{g,t}$ so that $\widehat{\eta}_t = \widehat{\varepsilon}_{y,t} - \widehat{\beta}\widehat{\varepsilon}_{g,t} = \widetilde{\eta}_t + (\widehat{\varepsilon}_{y,t} - \widetilde{\varepsilon}_{y,t}) - \widehat{\beta}(\widehat{\varepsilon}_{g,t} - \widetilde{\varepsilon}_{g,t}) + (\beta - \widehat{\beta})\widetilde{\varepsilon}_{g,t}$. The $\|\cdot\|_\infty$ consistency in Step 4 implies $\|\widehat{\eta} - \widetilde{\eta}\|_\infty = o_P(1)$. Also,

$$T^{-1}\sum_t \widehat{\varepsilon}_{z,t}^2 \widehat{\eta}_t^2 = T^{-1}\sum_t \widetilde{\varepsilon}_{z,t}^2 \widetilde{\eta}_t^2 + R_T,$$

where R_T collects cross-product terms of the form $T^{-1}\sum_t (\widehat{\varepsilon}_{z,t} - \widetilde{\varepsilon}_{z,t})\widehat{\varepsilon}_{z,t}\widehat{\eta}_t^2$, $T^{-1}\sum_t \widetilde{\varepsilon}_{z,t}^2(\widehat{\eta}_t - \widetilde{\eta}_t)\widehat{\eta}_t$, etc., together with the $(\widehat{\beta} - \beta)$ contribution which is $O_P(T^{-1/2})$ by Step 3 and is

multiplied by $T^{-1}\sum_t \widetilde{\varepsilon}_{z,t}^2|\widetilde{\varepsilon}_{g,t}| = O_P(1)$. By Cauchy–Schwarz and Step 1,

$$|R_T| \le \Big(T^{-1}\sum_t (\widehat{\varepsilon}_{z,t}-\widetilde{\varepsilon}_{z,t})^2\Big)^{1/2}\Big(T^{-1}\sum_t \widehat{\varepsilon}_{z,t}^2\widehat{\eta}_t^4\Big)^{1/2} + (\text{symmetric terms in } g \text{ and } \eta) + o_P(1).$$

The first factor $T^{-1}\sum_t(\widehat{\varepsilon}_{z,t}-\widetilde{\varepsilon}_{z,t})^2 = o_P(1)$ by Step 1. For the second factor, by the elementary inequality $\widehat{\varepsilon}_{z,t}^2 \le 2\varepsilon_{z,t}^2 + 2(\widehat{\varepsilon}_{z,t}-\varepsilon_{z,t})^2$ and $\widehat{\eta}_t^4 \le 8\eta_t^4 + 8(\widehat{\eta}_t-\eta_t)^4$,

$$\begin{aligned} T^{-1}\sum_t \widehat{\varepsilon}_{z,t}^2\widehat{\eta}_t^4 \;\le\; & C\,T^{-1}\sum_t \varepsilon_{z,t}^2\eta_t^4 + C\,T^{-1}\sum_t \varepsilon_{z,t}^2(\widehat{\eta}_t-\eta_t)^4 \\ & + C\,T^{-1}\sum_t (\widehat{\varepsilon}_{z,t}-\varepsilon_{z,t})^2\eta_t^4 + C\,T^{-1}\sum_t (\widehat{\varepsilon}_{z,t}-\varepsilon_{z,t})^2(\widehat{\eta}_t-\eta_t)^4. \end{aligned}$$

The first term is $O_P(1)$ by the law of large numbers and Assumption 3.1(iii) ($\mathbb{E}[\varepsilon_{z,t}^2\eta_t^4] \le (\mathbb{E}\varepsilon_{z,t}^8)^{1/4}(\mathbb{E}\eta_t^8)^{3/4} < \infty$ by Hölder). All the other terms are $o_P(1)$ by the $\|\cdot\|_\infty$-consistency of $\widehat{\epsilon}_a, \widehat{\eta}$. Hence $T^{-1}\sum_t \widehat{\varepsilon}_{z,t}^2\widehat{\eta}_t^4 = O_P(1)$. Thus $R_T = o_P(1)$, and using $T^{-1}\sum_t \widetilde{\varepsilon}_{z,t}^2\widetilde{\eta}_t^2 = T^{-1}\sum_t \varepsilon_{z,t}^2\eta_t^2 + O_P(T^{-1/2})$,

$$T^{-1}\sum_t \widehat{\varepsilon}_{z,t}^2\widehat{\eta}_t^2 = \mathbb{E}[\varepsilon_{z,t}^2\eta_t^2] + o_P(1).$$

Combining, $\widehat{\sigma}^2 = \sigma^2 + o_P(1)$, and hence $\sqrt{T}\widehat{\sigma}^{-1}(\widehat{\beta}-\beta) \to^d \mathcal{N}(0,1)$, as claimed.

OLS as the special case $z_t = g_t$. Setting $z_t = g_t$ throughout collapses $\varepsilon_z \equiv \varepsilon_g$, $\widehat{\varepsilon}_z \equiv \widehat{\varepsilon}_g$, the relevance constant $\gamma = \mathbb{E}[\varepsilon_{g,t}^2] > 0$ holds automatically, and $\widehat{\beta} = (\widehat{\varepsilon}_g'\widehat{\varepsilon}_g)^{-1}\widehat{\varepsilon}_g'\widehat{\varepsilon}_y$ is the partialled-out OLS estimator. The exclusion restriction $\mathbb{E}[\eta_t\varepsilon_{z,t}] = 0$ becomes the OLS exogeneity $\mathbb{E}[\eta_t\varepsilon_{g,t}] = 0$, and the asymptotic variance reduces to $\sigma^2 = \mathbb{E}[\varepsilon_{g,t}^2]^{-2}\mathbb{E}[\varepsilon_{g,t}^2\eta_t^2]$, the familiar partialled-out OLS sandwich.

$r = 0$ as a special case. When $r = 0$, $\widetilde{\varepsilon}_a - \widehat{\varepsilon}_a = P_{\widehat{F}}\widetilde{A} = \widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{\varepsilon}_a$.

$$\frac{1}{\sqrt{T}}\|\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{\varepsilon}_a\| \;\le\; O_P(T^{-1/2}). \tag{E.6}$$

So $\frac{1}{\sqrt{T}}\|\widetilde{\varepsilon}_a - \widehat{\varepsilon}_a\|^2 = o_P(1)$. Meanwhile $\frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'(\widetilde{\varepsilon}_b - \widehat{\varepsilon}_b) = \frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'\widehat{V}_{-r}^c\widehat{V}_{-r}^{c\prime}\widetilde{\varepsilon}_b = O_P(T^{-1/2})$. This implies

$$\frac{1}{\sqrt{T}}\widetilde{\varepsilon}_a'\widehat{\varepsilon}_b = \frac{1}{\sqrt{T}}\varepsilon_a'\varepsilon_b + o_P(1), \qquad a,b \in \{y,g,z\}.$$

The rest of the proof follows similarly.

■

References

Abbe, E., J. Fan, K. Wang, and Y. Zhong (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics 48*(3), 1452–1474.

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

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics 174*(1), 1–14.

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

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

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

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

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

Bloemendal, A., L. Erdős, A. Knowles, H.-T. Yau, and J. Yin (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electronic Journal of Probability 19*(33), 1–53.

Bonsang, E., S. Adam, and S. Perelman (2012). Does retirement affect cognitive functioning? *Journal of Health Economics 31*(3), 490–501.

Cape, J., M. Tang, and C. E. Priebe (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics 47*(5), 2405–2439.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey (2016). Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, Centre for Microdata Methods and Practice.

Choi, J., H. Kwon, and Y. Liao (2025). Inference for low-rank models without estimating the rank. *Journal of the American Statistical Association*, 1–12.

Coe, N. B. and G. Zamarro (2011). Retirement effects on health in Europe. *Journal of Health Economics 30*(1), 77–86.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*(3), 373–394.

Dave, D., I. Rashad, and J. Spasojevic (2008). The effects of retirement on physical and mental health outcomes. *Southern Economic Journal 75*(2), 497–523.

Fan, J., K. Li, and Y. Liao (2021). Recent developments in factor models and applications in econometric learning. *Annual Review of Financial Economics 13*(1), 401–430.

Fan, J. and Y. Liao (2022). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association 117*(538), 909–924.

Fan, J., W. Wang, and Y. Zhong (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research 18*(207), 1–42.

French, E. and J. B. Jones (2011). The effects of health insurance and self-insurance on retirement behavior. *Econometrica 79*(3), 693–732.

Freyaldenhoven, S. (2022). Factor models with local factors—determining the number of relevant factors. *Journal of Econometrics 229*(1), 80–102.

Giglio, S., D. Xiu, and D. Zhang (2023). Prediction when factors are weak. Technical report, University of Chicago.

Hansen, C. and Y. Liao (2018). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory*, 1–45.

Hu, Y. and W. Wang (2024). Network-adjusted covariates for community detection. *Biometrika*, asae011.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Insler, M. (2014). The health consequences of retirement. *Journal of Human Resources 49*(1), 195–233.

Kato, T. (1995). *Perturbation theory for linear operators*. Springer-Verlag Berlin Heidelberg.

Knowles, A. and J. Yin (2017). Anisotropic local laws for random matrices. *Probability Theory and Related Fields 169*, 257–352.

Mandal, B. and B. Roe (2008). Job loss, retirement and the mental health of older Americans. *Journal of Mental Health Policy and Economics 11*(4), 167–176.

Mazzonna, F. and F. Peracchi (2012). Ageing, cognitive abilities and retirement. *European Economic Review 56*(4), 691–710.

Moon, R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*, 1543–1579.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The 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.

Sonnega, A., J. D. Faul, M. B. Ofstedal, K. M. Langa, J. W. Phillips, and D. R. Weir (2014). Cohort profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology 43*(2), 576–585.

Uematsu, Y. and T. Yamagata (2023). Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics 41*(1), 213–227.

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

Wang, W. and J. Fan (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics 45*(3), 1342–1374.