# Direct and efficient estimation of bilinear forms in staggered tensor panels

Alberto Bordino[1], Thomas B. Berrett[1], and Olga Klopp[2]

[1]Department of Statistics, University of Warwick,
{alberto.bordino,tom.berrett}@warwick.ac.uk
[2]ESSEC Business School, kloppolga@math.cnrs.fr

**Abstract**

We study the estimation of bilinear forms from noisy, partially observed tensor data. The signal follows a Tucker2 model, with shared unit and time factors across tensor layers and slice-specific cores. The missingness pattern is structured and motivated by staggered adoption designs, which are common in causal inference and related applications. We first analyse the four-block missingness pattern, the basic building block for general staggered adoption, and propose a spectral algorithm that pools information across layers and targets the functional directly, rather than completing the entire tensor. We prove a non-asymptotic mean squared error bound that exhibits a phase transition in the number of layers, showing when pooling improves estimation, and match it with a local minimax lower bound up to constants. We then extend the construction to general staggered adoption designs via an anchored four-block reduction, and derive analogous theoretical guarantees. Finally, we validate our theoretical findings through experiments on both simulated and real-world datasets.

## 1 Introduction

Causal inference can naturally be formulated as a missing data problem. For each unit, only the potential outcome associated with the realised treatment path is observed, while the potential outcomes under alternative treatment paths are counterfactual and remain unobserved. This is what Holland (1986) calls *the fundamental problem of causal inference*. In applications, the resulting missingness mechanism may take different forms. In controlled or well-randomised settings, treatment assignment may be approximately independent of the potential outcomes. In observational settings, however, treatment assignment often depends on measured or unmeasured factors that are also related to the outcome. The missing untreated outcomes are then plausibly missing not at random (MNAR), as their absence is induced by treatment, and treatment timing may itself carry information about the latent untreated trajectory. Nevertheless, the missingness pattern is usually not arbitrary, since adoption typically induces a structure in the observed entries.

One important example, and the one studied in this paper, is staggered adoption (Athey and Imbens, 2022). For each given treatment, units are observed over multiple periods, and each unit may begin treatment

at its own adoption time. Once treatment has begun, it is irreversible in the sense that the unit remains treated in all subsequent periods. From the perspective of untreated potential outcomes, the data are therefore observed for all units before their adoption times and missing for treated units after adoption, hence, if units are ordered by adoption time and periods are ordered chronologically, the resulting untreated-outcome matrix exhibits a staircase pattern. The simplest form of such a missingness structure is the four-block setting illustrated in Figure 1 and analysed in Section 2. This design is common in policy evaluation; for instance, the COVID-19 policy tracker, available on GitHub, shows that many interventions, such as international travel controls or income-support policies, are implemented in a staggered fashion. This dataset is also an example where missingness is likely MNAR, as policies aimed at containing the virus are more likely to be adopted in places where the disease burden is higher.

A common strategy in causal panel data is to impute the missing untreated potential outcomes and use the completed panel to estimate causal quantities such as average treatment effects or contrasts. Low-rank matrix completion provides a natural framework for this task. Although such methodologies were developed mainly for missing completely at random observation patterns (e.g., Candès and Recht, 2009; Keshavan et al., 2010; Negahban and Wainwright, 2012; Koltchinskii et al., 2011; Klopp, 2014; Chi et al., 2019), they can be repurposed in causal panels by treating unobserved untreated outcomes as missing entries of an approximately low-rank matrix. Athey et al. (2021) formalised this connection by relating low-rank matrix completion to two classical approaches in causal panel analysis, unconfoundedness-based methods (e.g., Imbens and Rubin, 2015) and synthetic-control methods (e.g., Abadie, 2021), and proposed estimating the missing counterfactual entries through nuclear-norm penalised least squares.

A subsequent literature has developed matrix-completion methods for MNAR settings with structured missingness. Choi and Yuan (2026) study staggered adoption designs and extend the nuclear-norm approach of Athey et al. (2021) by partitioning the missing entries into groups and applying convex relaxation within each group. They prove $\ell_\infty$ estimation error bounds that improve on the Frobenius-norm bound obtained in Athey et al. (2021). Similarly, Agarwal et al. (2026) study a related problem with row and column side information, providing Frobenius-norm guarantees for an estimator based on sieve projection and nuclear-norm penalisation. Alongside optimisation-based approaches, spectral methods emerge as a parallel line of work for MNAR matrix completion. In this regard, Yan and Wainwright (2024) consider panels with staggered adoption and propose a spectral algorithm based on singular value decomposition and prove non-asymptotic entrywise guarantees as well as Gaussian approximations. Related factor-based approaches, including Bai and Ng (2021) and Cahan et al. (2023), exploit tall and wide observed blocks to estimate latent factors and impute missing panel entries. Finally, as a third line of research, Agarwal et al. (2023) developed a completion method based on synthetic nearest neighbours for a broad class of MNAR patterns, with $\ell_\infty$ error bounds and asymptotic normality.

The primary target in much of this literature, however, remains recovery of the missing matrix, either as a whole or entry by entry. The problem of estimating general bilinear forms is briefly mentioned by Xia et al. (2024, Appendix C), but no theoretical guarantees are provided for this target. Instead, existing error bounds are typically stated for full-matrix recovery, for instance in Frobenius norm or entrywise $\ell_\infty$ norm. These results are valuable, but they are not tailored to the objectives that often arise in applications, where the parameter of interest is a lower dimensional causal functional, such as an average treatment effect or a policy-weighted aggregate over a target population. Estimating the full matrix and then applying the desired

functional is a natural plug-in approach, but it need not be statistically or computationally efficient for the functional itself. This motivates the study of direct methodologies for estimating causal functionals under structured missingness.

Furthermore, many causal inference applications involve multiple treatments. The COVID-19 policy setting provides a simple example where several interventions, such as school closures, travel restrictions, and income-support measures, may be observed for the same time periods. Analysing each policy separately ignores common structure across treatments, while flattening all dimensions into a matrix can obscure treatment-specific effects. A tensor representation is therefore a natural generalisation when the latent potential-outcome object is indexed not only by unit and time, but also by treatment or policy. Recent work has begun to develop tensor methods for causal inference, but the theory remains less developed than in the matrix case. Auerbach et al. (2022) arrange multivariate longitudinal outcomes as a unit-by-time-by-outcome tensor and use nuclear-norm penalisation to impute the missing entries and study COVID-19 mandates. Agarwal et al. (2025) extend synthetic-control ideas to multiple treatments using a low-rank tensor factor model. Mandal and Parkes (2019) and Gao et al. (2025) consider tensor formulations for longitudinal causal problems, where treatment histories are stacked along an additional tensor mode. In particular, Gao et al. (2025) estimate the latent tensor using an inverse-probability-weighted low-rank Tucker formulation, implemented by projected gradient descent. Their main guarantee is a non-asymptotic Frobenius-norm bound for tensor recovery (Theorem 1). In Remark 2, they relate their framework to Athey et al. (2021), noting that a special case reduces to a staggered-adoption panel setting with two potential-outcome matrices.

These contributions show that tensor-valued potential-outcome models arise naturally when one allows for multiple outcomes or sequential regimes. At the same time, as in the MNAR matrix-completion literature, existing tensor-completion theory in causal settings is still largely centred on the recovery of the latent tensor, rather than on direct estimation of specified functionals under structured missingness. The goal of this paper is to address these two issues jointly. First, we use a tensor model to accommodate multiple treatments, policy regimes, or outcomes, with shared latent structure across slices and slice-specific cores that capture heterogeneity across the third dimension. Second, we estimate functionals of the missing counterfactual object directly, rather than taking full completion as the primary inferential goal. The functionals we study are bilinear forms, which include several causal estimands of interest as special cases, such as average and individual counterfactual components, as well as linear trends over time or across units. This target-specific approach can improve statistical efficiency and reduce computational cost when only these summaries are required. The main results are presented in Section 2 to follow.

We conclude the introduction with notation used throughout the paper. Given a third-order tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and a sequence of indices $I^{(t)} = \{i_1^{(t)}, \dots, i_{|I^{(t)}|}^{(t)}\} \subseteq [n_t]$, $t \in \{1,2,3\}$, we let $\mathcal{X}_{I^{(1)},I^{(2)},I^{(3)}} \in \mathbb{R}^{|I^{(1)}| \times |I^{(2)}| \times |I^{(3)}|}$ denote the subtensor obtained by selecting indices in each mode according to the corresponding index set. That is, for all $t \in \{1,2,3\}$, $1 \le k_t \le |I^{(t)}|$ we set $(\mathcal{X}_{I^{(1)},I^{(2)},I^{(3)}})_{k_1,k_2,k_3} = \mathcal{X}_{i_{k_1}^{(1)},i_{k_2}^{(2)},i_{k_3}^{(3)}}$. We use the symbol $\bullet$ in a subscript to denote the full index set in the corresponding mode. For example, $\mathcal{X}_{\bullet,I^{(2)},\bullet} := \mathcal{X}_{[n_1],I^{(2)},[n_3]}$. When an index set is a singleton, say $I^{(t)} = \{i\}$, we simply write $i$ in the corresponding mode. We use the same indexing notation for matrices and vectors. We also denote by $\mathbf{0}_d$ the null vector in dimension $d$, by $\mathbf{1}_d$ the all-one vector, by $I_d$ the identity matrix of dimension $d$, and by $\boldsymbol{e}_j^{(d)}$ the $j$-th canonical basis vector of $\mathbb{R}^d$. We will often omit the dependence on $d$ and simply use $\boldsymbol{e}_j$ when the ambient dimension is clear from the context. We also define $\boldsymbol{O}_{d_1 \times d_2} := \mathbf{0}_{d_1}\mathbf{0}_{d_2}^\top$ and $\mathbf{1}_{d_1 \times d_2} := \mathbf{1}_{d_1}\mathbf{1}_{d_2}^\top$.

For symmetric matrices $A, B$ of dimension $d$, we write $A \succeq 0$ if $A$ is positive semi-definite, and $A \succeq B$ if $A - B \succeq 0$. We denote the trace of $A$ with $\mathrm{tr}(A)$, and use $\mathrm{diag}(v)$ for a vector $v = (v_1, \ldots, v_d)$ to indicate a diagonal matrix with diagonal elements equal to $v_i$. We denote the minimum and maximum eigenvalues of a symmetric matrix $A$ by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively, and $\lambda_j(A)$ for its $j$-th largest eigenvalue. For a general matrix $B$, we write $\sigma_{\min}(B)$ and $\sigma_{\max}(B)$ for its smallest and largest singular values, and $\sigma_j(B)$ for its $j$-th largest singular value. We use $P_\Omega(M) := \Omega \odot M$ for the projection operator, where $\odot$ is the Hadamard product of two matrices. Also, $\mathrm{SVD}_r(A)$ denotes the rank-$r$ truncated singular value decomposition of $A$, returning $(U, \Sigma, V)$, where $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$ contain the top-$r$ left and right singular vectors, and $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal with the largest $r$ singular values. The Moore–Penrose pseudoinverse of $A = U\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r)V^\top$, with $\sigma_i > 0$, is $A^\dagger = V\,\mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1})\,U^\top$. We use $\|\cdot\|_p$ for the $\ell_p$-norm of a vector, and $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ for the spectral and Frobenius norms of a matrix, respectively. We write $\langle \cdot, \cdot \rangle$ for the Euclidean inner product of two vectors. The unit sphere in $\mathbb{R}^d$ is $\mathbb{B}_2(d) := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Finally, we use $\mathcal{O}(x)$ to denote a quantity whose absolute value is bounded above by $Cx$ for some universal constant $C > 0$.

# 2 Estimation of bilinear forms with four-block missingness

## 2.1 Statistical setting and main result

In this section, we present our main result on the estimation of bilinear forms in the four-block tensor setting illustrated in Figure 1. This four-block pattern provides the simplest nontrivial setting and serves as the key building block for the general staggered-adoption design discussed in Section 5.

For fixed dimensions $N, T, K \geq 1$, we define the deterministic missingness-pattern tensor $\Omega \in \mathbb{R}^{N \times T \times K}$ where, for all $j \in [K]$,

$$\Omega_{\bullet,\bullet,j} = \begin{pmatrix} \mathbf{1}_{N_{1j} \times T_{1j}} & \mathbf{1}_{N_{1j} \times T_{2j}} \\ \mathbf{1}_{N_{2j} \times T_{1j}} & \boldsymbol{O}_{N_{2j} \times T_{2j}} \end{pmatrix} \in \mathbb{R}^{N \times T}, \tag{1}$$

with $N = N_{1j} + N_{2j}$ and $T = T_{1j} + T_{2j}$. The tensor $\Omega$ will be fixed throughout this section. We observe $\mathcal{Y} \in \mathbb{R}^{N \times T \times K}$ with

$$\mathcal{Y}_{\bullet,\bullet,j} := P_{\Omega_{\bullet,\bullet,j}}(\mathcal{M}_{\bullet,\bullet,j} + \mathcal{E}_{\bullet,\bullet,j}) = \begin{pmatrix} \mathcal{M}^{(a)}_{\bullet,\bullet,j} + \mathcal{E}^{(a)}_{\bullet,\bullet,j} & \mathcal{M}^{(b)}_{\bullet,\bullet,j} + \mathcal{E}^{(b)}_{\bullet,\bullet,j} \\ \mathcal{M}^{(c)}_{\bullet,\bullet,j} + \mathcal{E}^{(c)}_{\bullet,\bullet,j} & \texttt{NA} \end{pmatrix}, \tag{2}$$

where $a, b, c$ refer to the observed blocks in (1), and $\mathcal{E} \in \mathbb{R}^{N \times T \times K}$ is such that $\mathcal{E}_{i,t,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for all $(i, t, j) \in [N] \times [T] \times [K]$. Figure 1 illustrates the observed tensor $\mathcal{Y}$ for $K = 3$. The requirement that the bottom-right block is missing in every layer is not crucial: what is essential is that, within each layer, every row is either fully observed or has missing entries beginning at a common time, which may vary across slices. Nonetheless, in this and the following sections we present our theory and methodology under the four-block design in (1), as this notation substantially simplifies the exposition. We refer the reader to Section 5 for the extension to more general staggered missingness designs.

As for the signal tensor $\mathcal{M} \in \mathbb{R}^{N \times T \times K}$, for $r \geq 1$ with $r \leq \min(N, T)$, we assume that $\mathcal{M}$ admits

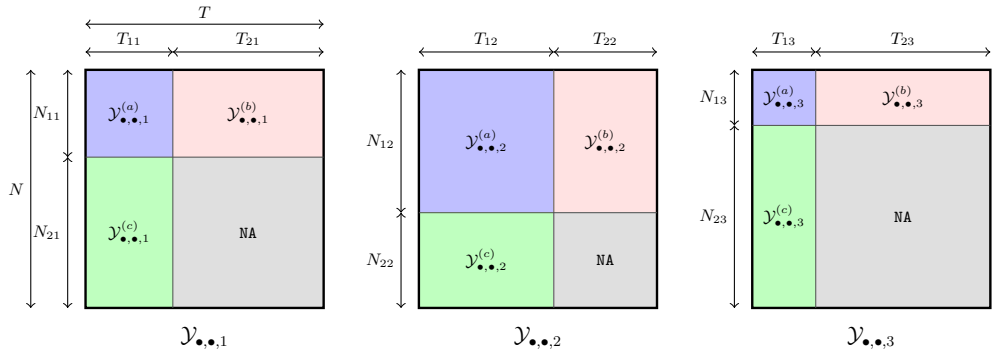


Figure 1: Layer-specific four-block structure for $K = 3$. Each slice $\mathcal{Y}_{\bullet,\bullet,j}$ contains $N_{1j}$ fully observed rows and $N_{2j}$ rows that are observed only in the first $T_{1j}$ columns. The total number of rows and columns are denoted by $N$ and $T$, respectively. We use $a, b, c$ to denote the observed blocks and $d$ to denote the missing block.

a Tucker2 decomposition (Kolda and Bader, 2009, Section 4) of rank $(r, r, K)$, meaning that there exist factor matrices $U \in \mathbb{R}^{N\times r}$ and $V \in \mathbb{R}^{T\times r}$ satisfying $U^\top U = V^\top V = I_r$, and a core tensor $\mathcal{C} \in \mathbb{R}^{r\times r\times K}$ such that $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$. Equivalently, each layer-specific signal matrix has rank at most $r$ and admits the factorisation $\mathcal{M}_{\bullet,\bullet,j} = U\,\mathcal{C}_{\bullet,\bullet,j}\,V^\top \in \mathbb{R}^{N\times T}$. A proof of the equivalence between the tensor formulation and this matrix-slice representation is given in Proposition 20 in Appendix D.3. This condition permits heterogeneity across slices while borrowing strength through shared latent row and column spaces. Similar modelling assumptions, often referred to as *common-subspace models*, have been studied in statistical settings under complete observation (Agterberg, 2026; Arroyo et al., 2021) and are motivated by biological applications, including neuroscience and single-cell RNA sequencing (Semedo et al., 2019; Ma and Ma, 2026).

It will also be convenient to partition $U = (U_{1j}\,;\,U_{2j})$ and $V = (V_{1j}\,;\,V_{2j})$ according to the $(N_{1j}, N_{2j})$ and $(T_{1j}, T_{2j})$ splits induced by slice $j$, where the semicolon denotes vertical stacking. Formally, for each $j \in [K]$ we define $U_{1j} = U_{[N_{1j}],\bullet} \in \mathbb{R}^{N_{1j}\times r}$ and $U_{2j} = U_{\{N_{1j}+1,\ldots,N\},\bullet} \in \mathbb{R}^{N_{2j}\times r}$; we define $V_{1j}$ and $V_{2j}$ analogously. Under this notation, the unobserved bottom-right block satisfies $\mathcal{M}^{(d)}_{\bullet,\bullet,j} = U_{2j}\,\mathcal{C}_{\bullet,\bullet,j}\,V_{2j}^\top$.

Our focus is on estimating general bilinear forms of the missing $d$-blocks. Formally, fix $k \in [K]$ and unit vectors $x \in \mathbb{B}_2(N_{2k})$, $y \in \mathbb{B}_2(T_{2k})$, and define

$$\mu^{(k)}_{xy} := x^\top \mathcal{M}^{(d)}_{\bullet,\bullet,k}\, y. \tag{3}$$

In words, the goal is to estimate a bilinear functional of the unobserved block in the $k$-th layer of the signal tensor $\mathcal{M}$. This class of targets includes several causal estimands of interest. For example, if $x$ and $y$ are constant vectors, $\mu^{(k)}_{xy}$ is proportional to an average counterfactual component over the missing block. If they are canonical basis vectors, it corresponds to an individual counterfactual component.

Our proposed estimator, presented in Algorithm 1 in Section 3, learns to predict the missing $d$-block from the $b$-block by regressing the $c$-block on the $a$-block using a spectral procedure that exploits the shared-subspace assumptions of the Tucker2 model. In this section, we present its theoretical analysis. To establish our results, we will need some assumptions. First, because the masks $\Omega_{\bullet,\bullet,j}$'s may be chosen adversarially and can induce pathological missing-not-at-random patterns, assumptions are needed to ensure that $\mu^{(k)}_{xy}$ is identifiable from the observed data. We build on prior work on MNAR matrix completion (Bai and Ng,

2021; Yan and Wainwright, 2024; Choi and Yuan, 2026; Agarwal et al., 2026) and impose a condition that quantitatively controls the spectra of the restricted Gram matrices $U_{1j}^\top U_{1j}$ and $V_{1j}^\top V_{1j}$.

**Assumption A1.** *There exist $0 \le c_\ell \le c_u$ such that for all $j \in [K]$ we have*

$$c_\ell \frac{N_{1j}}{N} I_r \preceq U_{1j}^\top U_{1j} \preceq c_u \frac{N_{1j}}{N} I_r, \qquad c_\ell \frac{T_{1j}}{T} I_r \preceq V_{1j}^\top V_{1j} \preceq c_u \frac{T_{1j}}{T} I_r.$$

In addition to this, our theoretical result requires the following three conditions. Throughout, $c_0 > 0$ and $c_{\mathrm{blk}} > 0$ denote sufficiently small absolute constants. In what follows and later sections we will also use the additional notation summarised in Table 1 in Section 3.

**Assumption A2.** *We have $r+\zeta \le c_{\mathrm{blk}} \min\big(N-r,\, T-r,\, N_{1k}, T_{1k}\big)$, $N-r \ge c_{\mathrm{blk}}\, N$, and $\min(\zeta_N, \zeta_T) \le c_{\mathrm{blk}}\, r$.*

**Assumption A3.** *Define the signal-to-noise ratio quantity $\theta := \sigma\, \gamma_{\min}^{-1} \max\Big(\sqrt{N},\, \sqrt{T},\, \sqrt{N/\rho_T},\, \sqrt{NT/N_{1k}}\Big)$, and assume that $\theta \le c_0$.*

**Assumption A4.** *We define the incoherence parameters $\nu_x := \sqrt{N/r}\, \|U_{2k}^\top x\|_2$ and $\nu_y := \sqrt{T/r}\, \|V_{2k}^\top y\|_2$, and assume they are of constant order.*

A detailed discussion of these assumptions is deferred to Section 2.3. Finally, motivated by the preceding conditions, we collect all admissible signal tensors into the following class. For fixed $r, N, T, K, \gamma_{\min}, \gamma_{\max}, \Omega$ and $0 \le c_\ell \le c_u$, we define

$$\begin{aligned}\mathcal{F}(c_\ell, c_u) := \big\{\, \mathcal{M} \in \mathbb{R}^{N\times T\times K} : \; & \mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K, \\ & \mathcal{C} \in \mathbb{R}^{r\times r\times K}, \quad U \in \mathbb{R}^{N\times r}, \quad V \in \mathbb{R}^{T\times r}, \\ & U^\top U = V^\top V = I_r, \\ & 0 < \gamma_{\min} \le \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \le \gamma_{\max} < \infty \quad \text{for all } j \in [K], \\ & \text{Assumption (A1) holds with constants } c_\ell, c_u \text{ for the fixed design } \Omega \,\big\}.\end{aligned}$$

We can now prove our main result. In line with previous literature, we will restrict attention to $c_\ell > 0$, as justified by Propositions 18 and 19 in Appendix D.2. Throughout the following, we will write $\kappa := \gamma_{\max}/\gamma_{\min}$ for the condition number, and assume it is of constant order.

**Theorem 1.** *Fix absolute constants $0 < c_\ell \le c_u$, a tensor $\mathcal{M} \in \mathcal{F}(c_\ell, c_u)$, an index $k \in [K]$, and unit vectors $x \in \mathbb{B}_2(N_{2k}), y \in \mathbb{B}_2(T_{2k})$. Let $\mu_{xy}^{(k)}$ be as in (3), and define $\hat{\mu}_{xy}^{(k)}$ to be the output of Algorithm 1 run with $0 < \tau \le \frac{c_\ell N_{1k}}{2N}$. Assume (A2), (A3), and (A4) with $\nu_x \ne 0, \nu_y \ne 0$. Let*

$$\Upsilon_{xy} := \frac{\sigma^2(r+\zeta_N)}{\rho_N} \|U_{2k}^\top x\|_2^2 + \frac{\sigma^2(r+\zeta_T)}{\rho_T} \|V_{2k}^\top y\|_2^2 + \frac{\sigma^2 N}{N_{1k}} \|U_{2k}^\top x\|_2^2\, \|V_{2k}^\top y\|_2^2,$$

*and further suppose that*

$$\frac{\gamma_{\max}^2}{\tau} \frac{N_{1k}}{N} \left(p_N^{-10} + p_T^{-10}\right) + \frac{\sigma^2}{\tau} \left(N_{1k} + T\right) \left(p_N^{-5} + p_T^{-5}\right) \le c_0 \Upsilon_{xy}. \tag{4}$$

*There exists a constant $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa, \nu_x, \nu_y) > 0$ such that $\mathbb{E}_{\mathcal{M}}[\{\hat{\mu}_{xy}^{(k)} - \mu_{xy}^{(k)}\}^2] \le c_1\, \Upsilon_{xy}$.*

All proofs are deferred to Appendix A. We first observe that the term $\sigma^2 (N/N_{1k}) \|U_{2k}^\top x\|_2^2 \|V_{2k}^\top y\|_2^2$ is asymmetric because Algorithm 1 uses vertical regression, predicting the missing $d$-block from the observed $b$-block; see Section 3 for a complete discussion of this. However, applying the same construction to the transposed tensor gives the analogous term with $T/T_{1k}$ in place of $N/N_{1k}$, hence taking the better of the two orientations yields the symmetric quantity

$$\sigma^2 \min\left(\frac{N}{N_{1k}}, \frac{T}{T_{1k}}\right) \|U_{2k}^\top x\|_2^2 \|V_{2k}^\top y\|_2^2.$$

In light of this, Theorem 1 gives a precise upper bound on the estimation error that reveals the effect of pooling. In particular, treating $r$ and $\zeta$ as constant-order quantities, the mean squared error is bounded by a term of the order $\sigma^2 \|U_{2k}^\top x\|_2^2 \rho_N^{-1} + \sigma^2 \|V_{2k}^\top y\|_2^2 \rho_T^{-1}$ in the small-$K$ regime, and by a term of order $\sigma^2 \min(N/N_{1k}, T/T_{1k}) \|U_{2k}^\top x\|_2^2 \|V_{2k}^\top y\|_2^2$ in the large-$K$ regime. As for their interpretations, the first two terms arise from estimating $U$ and $V$, and decrease with $K$ because these factors are common across layers. By contrast, the third term does not decrease with $K$, and can be interpreted as an irreducible layer-specific error, reflecting the difficulty of estimating $\mathcal{C}_{\bullet,\bullet,k}$. This phase transition in the rate as a function of $K$ is illustrated by the simulation study in Figure 3, and is complemented with local minimax lower bounds in Theorems 2 and 3.

## 2.2 Comparison with existing literature

Our work is closely related to Yan and Wainwright (2024), which studies entrywise inference for causal panel data under staggered adoption and corresponds to the special case $K = 1$, $x = \boldsymbol{e}_i$, and $y = \boldsymbol{e}_t$. In this case, both methods build on the spectral approach of Bai and Ng (2021), but differ in their inferential target and estimation procedure. Yan and Wainwright (2024, Algorithm 1) estimate missing entries of $M_d$ via completion of the full missing $d$-block, whereas we estimate a general bilinear form directly. More generally, however, our setting allows $K \geq 1$, and one of our contributions is to extend this direct regress-then-denoise approach to tensor data with four-block missingness.

The theoretical comparison is also transparent in the matrix entrywise case. When $K = 1$, $x = \boldsymbol{e}_i$, and $y = \boldsymbol{e}_t$, the upper bound in Theorem 1 matches their Equation 4.7 up to constants, apart from the term proportional to $\|U_{2k}^\top x\|_2^2 \|V_{2k}^\top y\|_2^2$. Since $\|V_{2k}^\top y\|_2^2 \leq 1$, this term is lower order and can be absorbed into the second one. In our tensor setting, however, it is important to keep this term explicit, as it identifies the component of the error that does not decrease under pooling across layers and thereby characterises the phase transition in the rate as $K$ grows.

At the proof level, both the leave-one-block-out method used in Yan and Wainwright (2024) and our method yield first-order expansions of $\hat{\mu}_{xy}^{(k)} - \mu_{xy}^{(k)}$ into Gaussian terms and remainders. The main difference lies in how the remainder terms are controlled. While their approach could in principle be adapted to the present tensor setting, it gives remainders that are negligible only under signal-to-noise conditions deteriorating with $K$. The proof of Theorem 1 instead proceeds via the Haar compression bounds outlined in Appendix E. More precisely, we apply these bounds to a centred version of $Y_{\text{left}}^{\text{p}} (Y_{\text{left}}^{\text{p}})^\top$, following the argument in the opening paragraph of the proof of Lemma 6 in Appendix C. This yields an expansion of the same form under weaker assumptions. Furthermore, compared with Yan and Wainwright (2024, Equation B.4), and

aside from allowing general unit vectors and extending the result to our tensor setting, Lemma 6 yields $\|(\hat{U}_{\text{left}} H_U - U)^\top x\|_2 \lesssim \sigma\gamma_{\min}^{-1}\sqrt{r+\zeta_T}\,\rho_T^{-1/2} + \sigma^2\gamma_{\min}^{-2} N\rho_T^{-1}\|U^\top x\|_2$, rather than $\|(\hat{U}_{\text{left}} H_U - U)^\top x\|_2 \lesssim \sigma\gamma_{\min}^{-1}\sqrt{r+\zeta_T}\,\rho_T^{-1/2} + \sigma^2\gamma_{\min}^{-2}(N+T_{1,\text{p}})\,\rho_T^{-1}\|U^\top x\|_2$. This sharper dependence on the dimensions in the second-order term is an independent contribution of interest, and is crucial for ensuring that the signal-to-noise requirement improves with $K$.

### 2.3 Discussion of the Assumptions

We now comment on the four assumptions needed in Theorem 1. Assumption (A1) requires $U_{1j}^\top U_{1j}$ and $V_{1j}^\top V_{1j}$ to be uniformly well-conditioned across $j \in [K]$, with eigenvalues proportional to the corresponding block fractions $N_{1j}/N$ and $T_{1j}/T$. This condition is trivially satisfied with $c_\ell = 0$, $c_u = \max(N/\min_{j\in[K]} N_{1j}, T/\min_{j\in[K]} T_{1j})$. Nevertheless, Propositions 18 and 19 in Appendix D.2 establish that restricting to $c_\ell > 0$ is necessary for (3) to be identifiable. We observe that, when $c_\ell > 0$, we also get $r \le \min(N_{1j}, T_{1j})$ for all $j \in [K]$, which implies $r \le \min(N, T, N_{1,\text{p}}, T_{1,\text{p}})$, which is the minimal dimensional requirement for the rank-$r$ SVDs used in Algorithm 1.

Assumption (A2) consists of mild dimension-regularity conditions, which are introduced to simplify the statement of the final result and make it more transparent, while Assumption (A4) is a standard incoherence condition adapted to the directions $x$ and $y$ of interest.

Finally, (A3) is a signal-to-noise condition requiring the noise level to be small relative to the relevant population signal strength. An analogous condition appears in Agterberg (2026) for estimating $U$ in a shared-subspace model with complete observations, where the error is measured by $\|\sin\Theta(\widehat{U}, U)\|_F^2$. By analogy, their Theorem 1 suggests that estimating (3) would be information-theoretically impossible without (A3); this heuristic is further strengthened by the additional difficulty introduced by missing observations in our setting. Furthermore, compared to Yan and Wainwright (2024, Assumption 4.3), (A3) becomes progressively less stringent as the number of layers $K$ increases, reflecting the benefit of pooling information across layers. This improvement continues until the requirement saturates at a local term which cannot be further reduced by pooling, thereby highlighting that a sufficiently strong slice-specific signal is still needed to learn $\mathcal{C}_{\bullet,\bullet,k}$.

## 3 Proposed methodology for four-block missingness

We now introduce Algorithm 1 to estimate (3) for a fixed $\mathcal{M} \in \mathcal{F}(c_\ell, c_u)$ with $c_\ell > 0$. To simplify the presentation of the method and its analysis, we introduce some notation, summarised in Table 1. We recall that parentheses with a space denote horizontal concatenation; semicolons denote vertical concatenation.

We now provide some intuition by considering the noiseless case. The key observation is that $\mathcal{M}^{(d)}_{\bullet,\bullet,k} = U_{2k}\big(U_{1k}^\top U_{1k}\big)^{-1} U_{1k}^\top\, \mathcal{M}^{(b)}_{\bullet,\bullet,k}$, so the missing $d$-block can be recovered by mapping the observed $b$-block through the linear operator $U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top$. For arbitrary unit vectors $x$ and $y$, this yields

$$\mu_{xy}^{(k)} = x^\top \mathcal{M}^{(d)}_{\bullet,\bullet,k}\, y = \Big\langle U_{2k}^\top x,\ \big(U_{1k}^\top U_{1k}\big)^{-1} U_{1k}^\top\, \mathcal{M}^{(b)}_{\bullet,\bullet,k}\, y \Big\rangle.$$

This motivates a two-step procedure that leverages the shared-subspace assumption, where we first form a pooled left matrix $M_{\text{left}}^{\text{p}} = U\, W_{\text{left}}^\top$ to learn the relevant left singular subspace and hence the associated

| Notation | Definition | Notation | Definition |
|---|---|---|---|
| $M_{\text{up}}^{(j)}$ | $(\mathcal{M}_{\bullet,\bullet,j}^{(a)}\ \mathcal{M}_{\bullet,\bullet,j}^{(b)}) \in \mathbb{R}^{N_{1j}\times T}$ | $M_{\text{left}}^{(j)}$ | $(\mathcal{M}_{\bullet,\bullet,j}^{(a)}\,;\,\mathcal{M}_{\bullet,\bullet,j}^{(c)}) \in \mathbb{R}^{N\times T_{1j}}$ |
| $E_{\text{up}}^{(j)}$ | $(\mathcal{E}_{\bullet,\bullet,j}^{(a)}\ \mathcal{E}_{\bullet,\bullet,j}^{(b)}) \in \mathbb{R}^{N_{1j}\times T}$ | $E_{\text{left}}^{(j)}$ | $(\mathcal{E}_{\bullet,\bullet,j}^{(a)}\,;\,\mathcal{E}_{\bullet,\bullet,j}^{(c)}) \in \mathbb{R}^{N\times T_{1j}}$ |
| $Y_{\text{up}}^{(j)}$ | $(\mathcal{Y}_{\bullet,\bullet,j}^{(a)}\ \mathcal{Y}_{\bullet,\bullet,j}^{(b)}) \in \mathbb{R}^{N_{1j}\times T}$ | $Y_{\text{left}}^{(j)}$ | $(\mathcal{Y}_{\bullet,\bullet,j}^{(a)}\,;\,\mathcal{Y}_{\bullet,\bullet,j}^{(c)}) \in \mathbb{R}^{N\times T_{1j}}$ |
| $M_{\text{up}}^{\text{p}}$ | $(M_{\text{up}}^{(1)}\,;\cdots;\,M_{\text{up}}^{(K)}) \in \mathbb{R}^{N_{1,\text{p}}\times T}$ | $M_{\text{left}}^{\text{p}}$ | $(M_{\text{left}}^{(1)}\ \cdots\ M_{\text{left}}^{(K)}) \in \mathbb{R}^{N\times T_{1,\text{p}}}$ |
| $E_{\text{up}}^{\text{p}}$ | $(E_{\text{up}}^{(1)}\,;\cdots;\,E_{\text{up}}^{(K)}) \in \mathbb{R}^{N_{1,\text{p}}\times T}$ | $E_{\text{left}}^{\text{p}}$ | $(E_{\text{left}}^{(1)}\ \cdots\ E_{\text{left}}^{(K)}) \in \mathbb{R}^{N\times T_{1,\text{p}}}$ |
| $Y_{\text{up}}^{\text{p}}$ | $(Y_{\text{up}}^{(1)}\,;\cdots;\,Y_{\text{up}}^{(K)}) \in \mathbb{R}^{N_{1,\text{p}}\times T}$ | $Y_{\text{left}}^{\text{p}}$ | $(Y_{\text{left}}^{(1)}\ \cdots\ Y_{\text{left}}^{(K)}) \in \mathbb{R}^{N\times T_{1,\text{p}}}$ |
| $W_{\text{up}}$ | $(U_{11}\,\mathcal{C}_{\bullet,\bullet,1}\,;\cdots;\,U_{1K}\,\mathcal{C}_{\bullet,\bullet,K}) \in \mathbb{R}^{N_{1,\text{p}}\times r}$ | $W_{\text{left}}$ | $\big(V_{11}\mathcal{C}_{\bullet,\bullet,1}^{\top};\cdots;V_{1K}\mathcal{C}_{\bullet,\bullet,K}^{\top}\big) \in \mathbb{R}^{T_{1,\text{p}}\times r}$ |
| $(U_{\text{up}},\Sigma_{\text{up}},V_{\text{up}})$ | $\text{SVD}_r(M_{\text{up}}^{\text{p}})$ | $(U_{\text{left}},\Sigma_{\text{left}},V_{\text{left}})$ | $\text{SVD}_r(M_{\text{left}}^{\text{p}})$ |
| $(\hat{U}_{\text{up}},\hat{\Sigma}_{\text{up}},\hat{V}_{\text{up}})$ | $\text{SVD}_r(Y_{\text{up}}^{\text{p}})$ | $(\hat{U}_{\text{left}},\hat{\Sigma}_{\text{left}},\hat{V}_{\text{left}})$ | $\text{SVD}_r(Y_{\text{left}}^{\text{p}})$ |
| $N_{1,\text{p}}$ | $\sum_{j=1}^{K} N_{1j}$ | $T_{1,\text{p}}$ | $\sum_{j=1}^{K} T_{1j}$ |
| $\rho_N$ | $N_{1,\text{p}}/N$ | $\rho_T$ | $T_{1,\text{p}}/T$ |
| $p_N$ | $\max\{N_{1,\text{p}},T\}$ | $p_T$ | $\max\{N,T_{1,\text{p}}\}$ |
| $\hat{U}_{1k}$ | $(\hat{U}_{\text{left}})_{[N_{1k}],\bullet} \in \mathbb{R}^{N_{1k}\times r}$ | $\hat{U}_{2k}$ | $(\hat{U}_{\text{left}})_{\{N_{1k}+1,\ldots,N\},\bullet} \in \mathbb{R}^{N_{2k}\times r}$ |
| $\hat{U}_{\text{up}}^{(k)}$ | $(\hat{U}_{\text{up}})_{\{s_k+1,\ldots,s_k+N_{1k}\},\bullet} \in \mathbb{R}^{N_{1k}\times r}$ | $\hat{V}_{2k}$ | $(\hat{V}_{\text{up}})_{\{T_{1k}+1,\ldots,T\},\bullet} \in \mathbb{R}^{T_{2k}\times r}$ |
| $\zeta_N$ | $\log(N_{1,\text{p}}+T)$ | $\zeta_T$ | $\log(N+T_{1,\text{p}})$ |
| $s_k$ | $\sum_{j=1}^{k-1} N_{1j}$ | $\zeta$ | $\max(\zeta_N,\zeta_T)$ |

Table 1: Notation used throughout the paper.

least-squares map, and then form a pooled upper matrix $M_{\text{up}}^{\text{p}} = W_{\text{up}}\,V^{\top}$ to construct a low-rank denoised estimate of the $b$-block, but only through its action on $y$.

More precisely, if $W_{\text{left}}$ has rank $r$, we have $\text{SVD}_r(M_{\text{left}}^{\text{p}}) = (U_{\text{left}},\Sigma_{\text{left}},V_{\text{left}})$ with $U_{\text{left}} = UQ_{\text{left}}$ for some orthogonal $Q_{\text{left}} \in \mathbb{R}^{r\times r}$. Importantly, the operator $U_{2k}\big(U_{1k}^{\top}U_{1k}\big)^{-1}U_{1k}^{\top}$ is rotationally invariant. One sufficient set of conditions ensuring that $W_{\text{left}}$ is full rank is Assumption (A1) with fixed $c_\ell > 0$, together with $\sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \geq \gamma_{\min} > 0$ for all $j \in [K]$; see Lemma 5 in Appendix C. On the other hand, if $W_{\text{up}}$ has rank $r$, for the pooled upper matrix we have $\text{SVD}_r(M_{\text{up}}^{\text{p}}) = (U_{\text{up}},\Sigma_{\text{up}},V_{\text{up}})$, where $V_{\text{up}} = VQ_{\text{up}}$ for some orthogonal $Q_{\text{up}} \in \mathbb{R}^{r\times r}$. Writing $U_{\text{up}} = \big(U_{\text{up}}^{(1)};\ \cdots;\ U_{\text{up}}^{(K)}\big)$ with $U_{\text{up}}^{(j)} \in \mathbb{R}^{N_{1j}\times r}$, we have $U_{\text{up}}\Sigma_{\text{up}} = W_{\text{up}}Q_{\text{up}}$ and $U_{\text{up}}^{(k)}\Sigma_{\text{up}} = U_{1k}\mathcal{C}_{\bullet,\bullet,k}Q_{\text{up}}$, hence $U_{\text{up}}^{(k)}\Sigma_{\text{up}}(V_{\text{up}})_{\{T_{1k}+1,\ldots,T\},\bullet}^{\top} = U_{1k}\mathcal{C}_{\bullet,\bullet,k}Q_{\text{up}}Q_{\text{up}}^{\top}V_{2k}^{\top} = U_{1k}\mathcal{C}_{\bullet,\bullet,k}V_{2k}^{\top} = \mathcal{M}_{\bullet,\bullet,k}^{(b)}$. This shows that also this quantity is rotationally invariant, and further ensures that no cross-alignment between the two SVDs is required.

In the noisy setting, Algorithm 1 follows the same principle, but applied to the observed tensor $\mathcal{Y}$ rather than the signal tensor $\mathcal{M}$. Our method computes the rank-$r$ truncated SVDs of $Y_{\text{left}}^{\text{p}}$ and $Y_{\text{up}}^{\text{p}}$, formed by horizontally stacking the blue and green blocks and vertically stacking the blue and pink blocks in Figure 1, respectively. This step, often referred to as *Stack-SVD*, exploits the singular subspaces shared across panels in order to improve subspace estimation; see Ma and Ma (2026); Baharav et al. (2025) for theoretical guarantees and comparisons with alternative aggregation schemes.

Another novelty of our method is the use of a clipped spectral inverse to estimate $(U_{1k}^{\top}U_{1k})^{-1}$. Specifically, after computing $\hat{H}_k = \widehat{U}_{1k}^{\top}\widehat{U}_{1k} = Q\,\text{diag}(\lambda_1,\ldots,\lambda_r)Q^{\top}$, we define $\hat{H}_{k,\tau}^{\text{inv}} := Q\,\text{diag}\big\{(\lambda_i \vee \tau)^{-1}\big\}_{i=1}^{r}Q^{\top}$, where $\tau > 0$ is a tuning parameter. We show in (24) in Appendix C that, if (A1) holds and $\tau \leq c_\ell N_{1k}/(2N)$, we have $\hat{H}_{k,\tau}^{\text{inv}} = (\hat{U}_{1k}^{\top}\hat{U}_{1k})^{-1}$ with high probability. However, on the complementary low-probability event,

**Algorithm 1** BilinearTensor4Block for the estimation of $\mu_{xy}^{(k)} = x^\top \mathcal{M}_{\bullet,\bullet,k}^{(d)} y$ in slice $k$ of a tensor with four-block missingness

**Require:** integer $k \in [K]$, rank $r$, unit vectors $x \in \mathbb{B}_2(N_{2k})$, $y \in \mathbb{B}_2(T_{2k})$, data $\mathcal{Y}$, block sizes $\{(N_{1j}, N_{2j}, T_{1j}, T_{2j})\}_{j=1}^K$ satisfying $N = N_{1j} + N_{2j}$ and $T = T_{1j} + T_{2j}$ for all $j \in [K]$, parameter $\tau > 0$.

1: Form pooled left matrix $Y_{\mathrm{left}}^{\mathrm{p}} \leftarrow (Y_{\mathrm{left}}^{(1)} \ \cdots \ Y_{\mathrm{left}}^{(K)}) \in \mathbb{R}^{N \times T_{1,\mathrm{p}}}$.

2: Compute rank-$r$ truncated singular value decomposition $(\hat{U}_{\mathrm{left}}, \hat{\Sigma}_{\mathrm{left}}, \hat{V}_{\mathrm{left}}) \leftarrow \mathrm{SVD}_r(Y_{\mathrm{left}}^{\mathrm{p}})$.

3: Set $\hat{U}_{1k} \leftarrow (\hat{U}_{\mathrm{left}})_{[N_{1k}],\bullet}$ and $\hat{U}_{2k} \leftarrow (\hat{U}_{\mathrm{left}})_{\{N_{1k}+1,\ldots,N\},\bullet}$.

4: Compute $\hat{H}_k \leftarrow \hat{U}_{1k}^\top \hat{U}_{1k} \in \mathbb{R}^{r \times r}$, take the eigendecomposition $\hat{H}_k = Q \,\mathrm{diag}(\lambda_1, \ldots, \lambda_r) Q^\top$, and set

$$\hat{H}_{k,\tau}^{\mathrm{inv}} \leftarrow Q \,\mathrm{diag}\left( \left\{ \frac{1}{\max[\lambda_i, \tau]} \right\}_{i=1}^{r} \right) Q^\top.$$

5: Compute $\hat{\alpha}_x^{(k)} \leftarrow \hat{U}_{2k}^\top x \in \mathbb{R}^r$.

6: Form pooled upper matrix $Y_{\mathrm{up}}^{\mathrm{p}} \leftarrow (Y_{\mathrm{up}}^{(1)} \ ; \ \cdots \ ; \ Y_{\mathrm{up}}^{(K)}) \in \mathbb{R}^{N_{1,\mathrm{p}} \times T}$.

7: Compute rank-$r$ truncated singular value decomposition $(\hat{U}_{\mathrm{up}}, \hat{\Sigma}_{\mathrm{up}}, \hat{V}_{\mathrm{up}}) \leftarrow \mathrm{SVD}_r(Y_{\mathrm{up}}^{\mathrm{p}})$.

8: Compute $s_k \leftarrow \sum_{j=1}^{k-1} N_{1j}$, and extract $\hat{U}_{\mathrm{up}}^{(k)} \leftarrow (\hat{U}_{\mathrm{up}})_{\{s_k+1,\ldots,s_k+N_{1k}\},\bullet}$, $\hat{V}_{2k} \leftarrow (\hat{V}_{\mathrm{up}})_{\{T_{1k}+1,\ldots,T\},\bullet}$.

9: Compute $T_y \leftarrow \hat{V}_{2k}^\top y \in \mathbb{R}^r$, $W_y \leftarrow \hat{\Sigma}_{\mathrm{up}} T_y \in \mathbb{R}^r$, and $X_y \leftarrow \hat{U}_{\mathrm{up}}^{(k)} W_y \in \mathbb{R}^{N_{1k}}$.

10: Compute $\hat{\beta}_y^{(k)} \leftarrow \hat{H}_{k,\tau}^{\mathrm{inv}} \hat{U}_{1k}^\top X_y \in \mathbb{R}^r$.

11: **return** $\hat{\mu}_{xy}^{(k)} \leftarrow \langle \hat{\alpha}_x^{(k)}, \hat{\beta}_y^{(k)} \rangle$.

spectral thresholding stabilises the inverse, since $\|\hat{H}_{k,\tau}^{\mathrm{inv}} \hat{U}_{1k}^\top\|_{\mathrm{op}}^2 = \max_{i \in [r]} \lambda_i / (\lambda_i \vee \tau)^2 \leq \tau^{-1}$, and allows returning a nontrivial output; this may be useful to practitioners. We comment more on the role of $\tau$ in Figure 4 in Section 6.1.

Finally, we further elaborate on the computational complexity of our procedure. The dominant cost is given by the rank-$r$ truncated singular value decompositions of $Y_{\mathrm{left}}^{\mathrm{p}} \in \mathbb{R}^{N \times T_{1,\mathrm{p}}}$ and $Y_{\mathrm{up}}^{\mathrm{p}} \in \mathbb{R}^{N_{1,\mathrm{p}} \times T}$, which are of the order $\mathcal{O}(NT_{1,\mathrm{p}} r + N_{1,\mathrm{p}} T r)$. After these decompositions are computed, the bilinear form is targeted directly. Indeed, Steps 9 and 10 compute $T_y = \hat{V}_{2k}^\top y$, $W_y = \hat{\Sigma}_{\mathrm{up}} T_y$, $X_y = \hat{U}_{\mathrm{up}}^{(k)} W_y$, and $\hat{\beta}_y^{(k)}$ without ever materialising the full matrix. This avoids an $\mathcal{O}(N_{1k} T_{2k} r)$ block-construction cost, and computes the required action on $y$ in $\mathcal{O}((T_{2k} + N_{1k}) r)$ time; including the computation of $\hat{\alpha}_x^{(k)} = \hat{U}_{2k}^\top x$ in Step 5, the total cost per query is $\mathcal{O}((N_{2k} + T_{2k} + N_{1k}) r)$. Furthermore, for the clipped inverse in Step 4, forming $\hat{H}_k = \hat{U}_{1k}^\top \hat{U}_{1k}$ costs $\mathcal{O}(N_{1k} r^2)$, while its eigendecomposition and computing $\hat{H}_{k,\tau}^{\mathrm{inv}}$ cost $\mathcal{O}(r^3)$. Taken together, the runtime analyses of these steps also indicate that our algorithm is well suited to caching. In particular, for fixed $k$, once the pooled singular value decompositions and the slice-specific regression factorisation have been computed, each additional query costs only $\mathcal{O}((N_{2k} + T_{2k} + N_{1k}) r)$. When $k$ varies, the pooled singular value decompositions can still be reused, and the only additional slice-specific computation is the $\mathcal{O}(N_{1k} r^2)$ regression factorisation, giving a total cost of $\mathcal{O}(N_{1k} r^2 + (N_{2k} + T_{2k} + N_{1k}) r)$.

# 4 Local minimax lower bounds

We next complement the upper bound in Theorem 1 by establishing local minimax lower bounds in neighbourhoods of fixed tensors $\mathcal{M}_0$ that satisfy suitable conditions. The first such result corresponds to the large-$K$ regime. Here, for $\mathcal{M} \in \mathcal{F}(c_\ell, c_u)$, we write $\mu_{xy}^{(k)}(\mathcal{M}) := x^\top \mathcal{M}_{\bullet,\bullet,k}^{(d)} y$ and $Z_\Omega := \{ \mathcal{M}_{itj} + \mathcal{E}_{itj} :$

$\Omega_{i,t,j} = 1,\ (i,t,j) \in [N] \times [T] \times [K]\}$ for the observed entries, with the mask $\Omega$ fixed and known, and use $\mathbb{P}_{\mathcal{M}}$ and $\mathbb{E}_{\mathcal{M}}$ for probability and expectation under the law of $Z_\Omega$.

**Theorem 2.** *Fix $k \in [K]$ and unit vectors $x \in \mathbb{B}_2(N_{2k}), y \in \mathbb{B}_2(T_{2k})$. Let $\mathcal{M}_0 = \mathcal{C}_0 \times_1 U_0 \times_2 V_0 \times_3 I_K \in \mathcal{F}(c_\ell, c_u)$. Assume that the $k$-th core matrix is separated from the boundary of the admissible singular-value interval, in the sense that $\delta_{\gamma,k} := \min\{\sigma_{\min}((\mathcal{C}_0)_{\bullet,\bullet,k}) - \gamma_{\min}, \gamma_{\max} - \sigma_{\max}((\mathcal{C}_0)_{\bullet,\bullet,k})\} > 0$. For $\varsigma > 0$ define $\mathcal{F}_{\mathrm{loc}}(\mathcal{M}_0, \varsigma) := \{\mathcal{M} \in \mathcal{F}(c_\ell, c_u) : \|\mathcal{M} - \mathcal{M}_0\|_F \le \varsigma\}$. There exists a constant $c \equiv c(c_u) > 0$ such that*

$$\inf_{\phi} \sup_{\mathcal{M} \in \mathcal{F}_{\mathrm{loc}}(\mathcal{M}_0,\varsigma)} \mathbb{E}_{\mathcal{M}} \left[ \{\phi(Z_\Omega) - \mu^{(k)}_{xy}(\mathcal{M})\}^2 \right] \ge c \min \left\{ \sigma^2 \min\left(\frac{N}{N_{1k}}, \frac{T}{T_{1k}}\right), \varsigma^2, \delta^2_{\gamma,k} \right\} \|U_{0,2k}^\top x\|_2^2 \, \|V_{0,2k}^\top y\|_2^2,$$

*where the infimum is over all Borel-measurable functions $\phi$ of the observed entries $Z_\Omega$.*

For the second result, which concerns the small-$K$ regime, we suppose for simplicity that $N_{1j} = N_1$ and $T_{1j} = T_1$ for all $j \in [K]$. Under this assumption, $U_{0,1j_1} = U_{0,1j_2}$ and $U_{0,2j_1} = U_{0,2j_2}$ for all $j_1, j_2 \in [K]$. We therefore denote these common matrices by $U_{0,1}$ and $U_{0,2}$, respectively, and adopt the analogous convention for $V_{0,1}$ and $V_{0,2}$. We define the projections $P_{V_{0,2}} := V_{0,2}\left(V_{0,2}^\top V_{0,2}\right)^{-1} V_{0,2}^\top$, $P_{U_{0,2}} := U_{0,2}\left(U_{0,2}^\top U_{0,2}\right)^{-1} U_{0,2}^\top$, $P^\perp_{V_{0,2}} := I_{T_2} - P_{V_{0,2}}$, and $P^\perp_{U_{0,2}} := I_{N_2} - P_{U_{0,2}}$. The inverses are well defined when (A1) holds with $c_u \max(N_1/N, T_1/T) < 1$. We also introduce $\omega_V := \|P^\perp_{V_{0,2}} y\|_2$, $\omega_U := \|P^\perp_{U_{0,2}} x\|_2 \in [0,1]$, which measure the components of $y$ and $x$ orthogonal to the column spaces of $V_{0,2}$ and $U_{0,2}$, respectively.

**Theorem 3.** *Fix $k \in [K]$ and unit vectors $x \in \mathbb{B}_2(N_{2k}), y \in \mathbb{B}_2(T_{2k})$. Set $N_{1j} = N_1$ and $T_{1j} = T_1$ for all $j \in [K]$. Let $\mathcal{M}_0 = \mathcal{C}_0 \times_1 U_0 \times_2 V_0 \times_3 I_K \in \mathcal{F}(c_\ell, c_u)$, and suppose that (A1) holds with margin $0 < \delta_{\mathrm{A1}} < (c_u - c_\ell)/2$, in the sense that*

$$(c_\ell + \delta_{\mathrm{A1}}) \frac{N_1}{N} I_r \preceq U_{0,1}^\top U_{0,1} \preceq (c_u - \delta_{\mathrm{A1}}) \frac{N_1}{N} I_r, \qquad (c_\ell + \delta_{\mathrm{A1}}) \frac{T_1}{T} I_r \preceq V_{0,1}^\top V_{0,1} \preceq (c_u - \delta_{\mathrm{A1}}) \frac{T_1}{T} I_r.$$

*Also assume that $c_u \max(N_1/N, T_1/T) < 1$. There exists a constant $c \equiv c(c_u, \gamma_{\min}, \gamma_{\max}) > 0$ such that*

$$\inf_{\phi} \sup_{\mathcal{F}_{\mathrm{loc}}(\mathcal{M}_0,\varsigma)} \mathbb{E}_{\mathcal{M}} \left[ \left\{ \phi(Z_\Omega) - \mu^{(k)}_{xy}(\mathcal{M}) \right\}^2 \right] \ge c\,\omega_V^2 \min\left(\frac{\sigma^2 N}{K N_1}, \varepsilon_V^2\right) \|U_{0,2}^\top x\|_2^2 + c\,\omega_U^2 \min\left(\frac{\sigma^2 T}{K T_1}, \varepsilon_U^2\right) \|V_{0,2}^\top y\|_2^2,$$

*where $\varepsilon_U := \min(\omega_U/2, \sqrt{T_1/T}, \varsigma\gamma_{\max}^{-1} K^{-1/2}, \sqrt{\delta_{\mathrm{A1}}/c_\ell}$, $\varepsilon_V := \min(\omega_V/2, \sqrt{N_1/N}, \varsigma\gamma_{\max}^{-1} K^{-1/2}, \sqrt{\delta_{\mathrm{A1}}/c_\ell})$, and the infimum is over all Borel-measurable functions $\phi$ of the observed entries $Z_\Omega$.*

For any $\mathcal{M}_0$ covered by both sets of assumptions, the combined results of Theorems 2 and 3 show the necessity of the elbow behaviour in the rate as a function of $K$. In particular, when $\sigma^2$ is sufficiently small and $\omega_V, \omega_U$ are bounded away from zero, taking the maximum of the respective right-hand sides yields a lower bound of the order

$$\frac{\sigma^2 N}{K N_1} \|U_{0,2}^\top x\|_2^2 + \frac{\sigma^2 T}{K T_1} \|V_{0,2}^\top y\|_2^2 + \sigma^2 \min\left(\frac{N}{N_1}, \frac{T}{T_1}\right) \|U_{0,2k}^\top x\|_2^2 \, \|V_{0,2k}^\top y\|_2^2.$$

This matches the upper bound in Theorem 1 up to constants, rank factors, and logarithmic factors. Moreover,

although the results above are local and stated around a fixed $\mathcal{M}_0$, they also imply global minimax lower bounds by taking $\varsigma$ sufficiently large so that it contains the entire parameter space.

We also observe that we are most interested in the regime where $\omega_V$ and $\omega_U$ are not small, which is crucial for Theorem 3 to be nontrivial. To see why, start by noticing that $(1 - c_u N_1/N)(1 - \omega_U^2) \leq \sigma_{\min}^2(U_{0,2})\,(1 - \omega_U^2) \leq \|U_{0,2}^\top x\|_2^2 \leq \sigma_{\max}^2(U_{0,2})\,(1 - \omega_U^2) \leq 1 - \omega_U^2$. An analogous statement holds for $\omega_V$ and $\|V_{0,2}^\top y\|_2$. Assuming $c_u \max(N_1/N, T_1/T) \leq 1 - \kappa$ for $\kappa > 0$, we thus get $\|U_{0,2}^\top x\|_2^2 \asymp_\kappa 1 - \omega_U^2$ and $\|V_{0,2}^\top y\|_2^2 \asymp_\kappa 1 - \omega_V^2$. Under (A4), we have $\|U_{0,2}^\top x\|_2^2 \asymp r/N \ll 1$ and $\|V_{0,2}^\top y\|_2^2 \asymp r/T \ll 1$, hence comparisons with Theorem 1 are most natural in the regime where $\omega_U$ and $\omega_V$ are of constant order, in fact close to one.

# 5 Estimation of bilinear forms under staggered adoption

## 5.1 Proposed methodology for staggered missingness

We now extend the four-block setting to general staggered-adoption designs. Our methodology reduces the staggered missingness problem to simpler four-block missingness patterns by constructing pooled upper and left matrices in the same spirit as before. The main additional challenge is to accommodate the more complex missingness structure induced by staggered adoption. In particular, for a fixed layer $j \in [K]$, staggered adoption means that missingness is irreversible, i.e. for each unit $i \in [N]$, there is an adoption time $A_{ij}$ such that $\Omega_{i,t,j} = \mathbb{1}\{t < A_{ij}\}$. For completeness, we set $A_{ij} = \infty$ for never-adopters.

It is useful to note that each layer has its own natural row ordering under which the corresponding missingness mask is a staircase; these orderings need not agree across layers. Since our target will be a bilinear form in layer $k$, we use the adoption-time ordering of the target layer as the common row ordering for all slices. In other words, we permute the rows so that $A_{1k} \geq \cdots \geq A_{Nk}$. This entails no loss of generality, since applying a common row permutation to all layers preserves the Tucker2 structure. Under this convention, the mask $\Omega_{\bullet,\bullet,k}$ admits an equivalent staircase characterisation: there exists an integer $o_k \geq 2$ and ordered non-empty contiguous partitions $[N] = R_{1k} \cup \cdots \cup R_{o_k,k}$ and $[T] = C_{1k} \cup \cdots \cup C_{o_k,k}$, with $|R_{ak}| = N_{ak}$ and $|C_{bk}| = T_{bk}$, such that $\Omega_{i,t,k} = \mathbb{1}\{(i,t) \in R_{ak} \times C_{bk}$ for some $a, b$ with $a + b \leq o_k + 1\}$. This block representation will be useful in what follows, as it allows us to describe the observed and missing regions of the target layer in terms of the staircase partitions $\{R_{ak}\}_{a=1}^{o_k}$ and $\{C_{bk}\}_{b=1}^{o_k}$.

As is apparent from the staircase representation above, the only assumption we make on $\Omega$ is that the missingness pattern for the target slice $\Omega_{\bullet,\bullet,k}$ contains fully observed rows, corresponding to never-adopting units, as well as an initial time period during which no unit in that slice has adopted. This is the basic requirement that allows us to reuse the methodology developed for the four-block design. An example of a staggered-adoption pattern covered by our framework is illustrated in Figure 2.

Having specified the structure of the missingness masks, we now introduce the signal and noise model. As in the previous sections, we assume that the signal tensor $\mathcal{M}$ admits a Tucker2 decomposition of rank $(r, r, K)$, so that $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$, where $U \in \mathbb{R}^{N \times r}$ and $V \in \mathbb{R}^{T \times r}$ have orthonormal columns. The noise tensor $\mathcal{E}$ has independent Gaussian entries with mean zero and variance $\sigma^2$, and for each $j \in [K]$ we observe $\mathcal{Y}_{\bullet,\bullet,j} = P_{\Omega_{\bullet,\bullet,j}}(\mathcal{M}_{\bullet,\bullet,j} + \mathcal{E}_{\bullet,\bullet,j})$. Our goal here is to estimate bilinear forms over all missing entries in layer $k \in [K]$. For simplicity, we first focus on a specific missing block, since this is the key step needed for the general case. In this regard, choose indices $(a, b)$ such that $a + b > o_k + 1$, so that the block $R_{ak} \times C_{bk}$

is unobserved. For unit vectors $x \in \mathbb{B}_2(N_{ak})$ and $y \in \mathbb{B}_2(T_{bk})$, our target estimand is the bilinear form

$$\mu_{xy}^{(k,a,b)} := x^\top \mathcal{M}_{\bullet,\bullet,k}^{(a,b)} y, \tag{5}$$

where $\mathcal{M}_{\bullet,\bullet,k}^{(a,b)} := U_{ak}\, \mathcal{C}_{\bullet,\bullet,k} V_{bk}^\top \in \mathbb{R}^{N_{ak} \times T_{bk}}$, with $U_{ak} := U_{R_{ak},\bullet} \in \mathbb{R}^{N_{ak} \times r}$ and $V_{bk} := V_{C_{bk},\bullet} \in \mathbb{R}^{T_{bk} \times r}$ denoting the restrictions of the Tucker2 factors to specified row and column blocks.

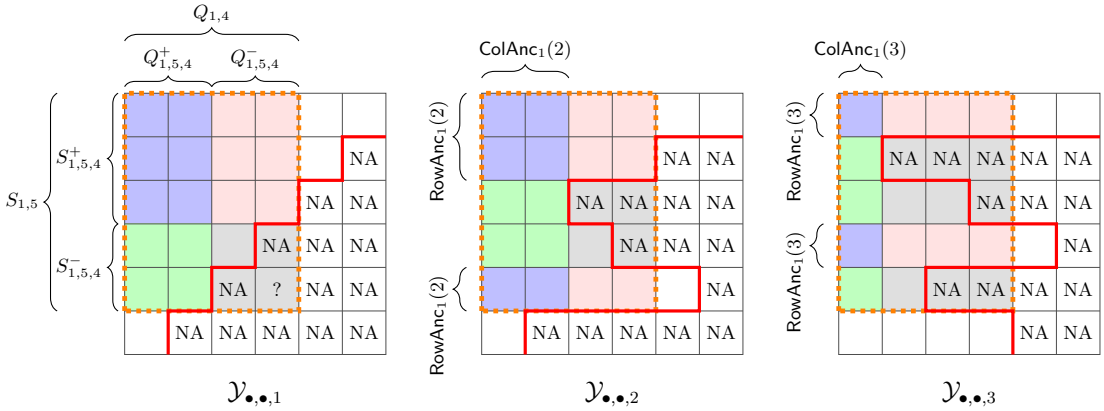


Figure 2: Layer-specific staggered adoption for $K = 3$, target layer $k = 1$, and target block $(a, b) = (5, 4)$. Panel 1 orders the target layer by its own rows, yielding a staircase missingness pattern; the orange dotted rectangle identifies the target rows $S_{1,5}$ and columns $Q_{1,4}$ used to form auxiliary matrices. Panels 2–3 show non-target layers under the same target-layer ordering, with general staggered missingness. These layers provide fully observed anchor rows $\mathsf{RowAnc}_1(j)$ and anchor columns $\mathsf{ColAnc}_1(j)$: blue/pink blocks form the upper pooled matrix, blue/green blocks form the left pooled matrix, and grey blocks are discarded.

The following definitions are needed to present our algorithm. We set

$$S^+_{k,a,b} := \bigcup_{a'=1}^{o_k+1-b} R_{a'k}, \quad S^-_{k,a,b} := \bigcup_{a'=o_k+2-b}^{a} R_{a'k}, \quad Q^+_{k,a,b} := \bigcup_{b'=1}^{o_k+1-a} C_{b'k}, \quad Q^-_{k,a,b} := \bigcup_{b'=o_k+2-a}^{b} C_{b'k}.$$

Since $a + b > o_k + 1$, the lower limits $o_k + 2 - b$ and $o_k + 2 - a$ are at most $a$ and $b$, respectively, so the sets $S^-_{k,a,b}$ and $Q^-_{k,a,b}$ are non-empty and contain $R_{ak}$ and $C_{bk}$. We also write $S_{k,a} := S^+_{k,a,b} \cup S^-_{k,a,b}$, $Q_{k,b} := Q^+_{k,a,b} \cup Q^-_{k,a,b}$, and observe that $S_{k,a}$ and $Q^+_{k,a,b}$ depend only on $a$, whereas $Q_{k,b}$ and $S^+_{k,a,b}$ depend only on $b$.

The four index sets in display define a four-block structure in which the target missing block $(a, b)$ is contained in $S^-_{k,a,b} \times Q^-_{k,a,b}$. Figure 2 illustrates this construction, with the target block marked by a question mark. Some entries of $S^-_{k,a,b} \times Q^-_{k,a,b}$ may be observed under the original staggered pattern, but we discard them to obtain a literal four-block construction. Furthermore, to leverage the shared subspaces $U$ and $V$ in the Tucker2 model, we define anchor sets that identify auxiliary fully observed rows and columns. For $j \neq k$, we write $\mathsf{ColAnc}_k(j) := \{t \in Q_{k,b} : \Omega_{i,t,j} = 1 \text{ for all } i \in S_{k,a}\}$ and $\mathsf{RowAnc}_k(j) := \{i \in S_{k,a} : \Omega_{i,t,j} = 1 \text{ for all } t \in Q_{k,b}\}$. These correspond to periods in $Q_{k,b}$ that are observed for all units in $S_{k,a}$, and to units in $S_{k,a}$ that are observed throughout all periods in $Q_{k,b}$, respectively. For completeness, we also set $\mathsf{RowAnc}_k(k) := S^+_{k,a,b}$ and $\mathsf{ColAnc}_k(k) := Q^+_{k,a,b}$. These sets identify the auxiliary submatrices used to construct the upper and left pooled matrices, as outlined in the following algorithm.

**Algorithm 2** BILINEARTENSORSTAGGERED for the estimation of $\mu_{xy}^{(k,a,b)} = x^\top \mathcal{M}_{\bullet,\bullet,k}^{(a,b)} y$ for a fixed missing block $(a,b)$ in slice $k$ of a tensor with staggered adoption missingness

**Require:** index $k \in [K]$, missing block $(a,b)$ with $a+b > o_k+1$, rank $r$, unit vectors $x \in \mathbb{B}_2(N_{ak})$ and $y \in \mathbb{B}_2(T_{bk})$, data $\mathcal{Y}$, parameter $\tau > 0$.

1: Permute rows so that $\Omega_{\bullet,\bullet,k}$ is in staircase form.

2: Compute $S_{k,a,b}^+, S_{k,a}, Q_{k,a,b}^+, Q_{k,b}$, and $\mathsf{RowAnc}_k(j), \mathsf{ColAnc}_k(j)$ for all $j \in [K]$.

3: Form pooled left matrix $Y_{\text{left}}^{\text{p}} \leftarrow (\mathcal{Y}_{S_{k,a},\mathsf{ColAnc}_k(1),1} \ \cdots \ \mathcal{Y}_{S_{k,a},\mathsf{ColAnc}_k(K),K}) \in \mathbb{R}^{|S_{k,a}| \times \sum_{j=1}^K |\mathsf{ColAnc}_k(j)|}$.

4: Compute rank-$r$ truncated singular value decomposition $(\hat{U}_{\text{left}}, \hat{\Sigma}_{\text{left}}, \hat{V}_{\text{left}}) \leftarrow \mathrm{SVD}_r(Y_{\text{left}}^{\text{p}})$.

5: Set $\hat{U}_{+k} \leftarrow (\hat{U}_{\text{left}})_{S_{k,a,b}^+,\bullet}$ and $\hat{U}_{ak} \leftarrow (\hat{U}_{\text{left}})_{R_{ak},\bullet}$.

6: Compute $\hat{H}_k \leftarrow \hat{U}_{+k}^\top \hat{U}_{+k} \in \mathbb{R}^{r\times r}$, take the eigendecomposition $\hat{H}_k = Q\,\mathrm{diag}(\lambda_1,\dots,\lambda_r)\,Q^\top$, and set

$$\hat{H}_{k,\tau}^{\text{inv}} \leftarrow Q\,\mathrm{diag}\left(\left\{\frac{1}{\max[\lambda_i,\tau]}\right\}_{i=1}^r\right) Q^\top.$$

7: Compute $\hat{\alpha}_x^{(k,a,b)} \leftarrow \hat{U}_{ak}^\top x \in \mathbb{R}^r$.

8: Form pooled upper matrix $Y_{\text{up}}^{\text{p}} \leftarrow (\mathcal{Y}_{\mathsf{RowAnc}_k(1),Q_{k,b},1} ; \cdots ; \mathcal{Y}_{\mathsf{RowAnc}_k(K),Q_{k,b},K}) \in \mathbb{R}^{\sum_{j=1}^K |\mathsf{RowAnc}_k(j)| \times |Q_{k,b}|}$.

9: Compute rank-$r$ truncated singular value decomposition $(\hat{U}_{\text{up}}, \hat{\Sigma}_{\text{up}}, \hat{V}_{\text{up}}) \leftarrow \mathrm{SVD}_r(Y_{\text{up}}^{\text{p}})$.

10: Let $s_k \leftarrow \sum_{j=1}^{k-1} |\mathsf{RowAnc}_k(j)|$, and extract $\hat{U}_{\text{up}}^{(k)} \leftarrow (\hat{U}_{\text{up}})_{\{s_k+1,\dots,s_k+|S_{k,a,b}^+|\},\bullet}$, $\hat{V}_{bk} \leftarrow (\hat{V}_{\text{up}})_{C_{bk},\bullet}$.

11: Compute $T_y \leftarrow \hat{V}_{bk}^\top y \in \mathbb{R}^r$, $W_y \leftarrow \hat{\Sigma}_{\text{up}} T_y \in \mathbb{R}^r$, and $X_y \leftarrow \hat{U}_{\text{up}}^{(k)} W_y \in \mathbb{R}^{|S_{k,a,b}^+|}$.

12: Compute $\hat{\beta}_y^{(k,a,b)} \leftarrow \hat{H}_{k,\tau}^{\text{inv}} \hat{U}_{+k}^\top X_y \in \mathbb{R}^r$.

13: **return** $\hat{\mu}_{xy}^{(k,a,b)} \leftarrow \langle \hat{\alpha}_x^{(k,a,b)}, \hat{\beta}_y^{(k,a,b)} \rangle$.

Algorithm 2 extends Algorithm 1 to the staggered-adoption setting, and reduces to it when the missingness pattern has four-block form. For a fixed missing block $(a,b)$ in layer $k$, the algorithm restricts attention to $\mathcal{Y}_{S_{k,a},Q_{k,b},\bullet}$, and uses the observations lying in $\bigcup_{j=1}^K (S_{k,a} \times \mathsf{ColAnc}_k(j) \times \{j\})$ and $\bigcup_{j=1}^K (\mathsf{RowAnc}_k(j) \times Q_{k,b} \times \{j\})$ to construct an auxiliary four-block problem, discarding all remaining entries. In particular, the left pooled matrix is formed from the anchor-column blocks $\mathcal{Y}_{S_{k,a},\mathsf{ColAnc}_k(j),j}$, while the upper pooled matrix is formed from the anchor-row blocks $\mathcal{Y}_{\mathsf{RowAnc}_k(j),Q_{k,b},j}$. This construction is illustrated in Figure 2.

Our procedure generalises Yan and Wainwright (2024, Algorithm 2) to the tensor setting and targets the bilinear form directly, rather than reconstructing the entire missing block. Related denoising techniques, using anchor sets and combined with PCA, were employed in Liu et al. (2026) for a different statistical problem, where the goal is to recover the global left subspace from matrix data with blockwise missingness, with error measured in Frobenius norm.

The auxiliary four-block construction also imposes a basic dimensional feasibility condition. In applications, the working rank must satisfy $r \le \min(\sum_{j\in[K]} |\mathsf{RowAnc}_k(j)|, \sum_{j\in[K]} |\mathsf{ColAnc}_k(j)|, |S_{k,a}|, |Q_{k,b}|)$, so that the two rank-$r$ truncated SVDs are well defined. This is only a minimal requirement for running the procedure. Even when this condition holds, additional assumptions are needed to guarantee that the resulting estimator is accurate; the theoretical analysis of Algorithm 2 is the object of the next section.

Finally, when aggregate quantities over multiple missing blocks of the same slice are required, such as those introduced in Appendix B.1, the most direct strategy is to apply Algorithm 2 separately to each missing block and then aggregate the resulting estimates. This blockwise implementation recomputes two rank-$r$ SVDs for every missing block, leading to the quadratic-cost procedure described in Algorithm 3 in

Appendix B.1. We also propose a reduced-anchor variant that reuses computations by caching the pooled left SVD once for each active row block $a$, and the pooled upper SVD once for each active column block $b$. This yields the linear-SVD-cost procedure in Algorithm 4, at the price of some loss in statistical efficiency. Appendix B.1 provides an extensive discussion of this computational–statistical tradeoff, and Figure 5 in Section 6.1 compares the two procedures in simulation.

## 5.2 Theoretical analysis

Algorithm 2 inherits the desirable properties of Algorithm 1 under analogous assumptions, with particular care needed in adapting Assumption (A1) to the auxiliary four-block reduction. To state these assumptions and the resulting corollary, we suppress the dependence on $(k, a, b)$ for simplicity, and set $\mathfrak{n} := |S|$ and $\mathfrak{t} := |Q|$. For the target layer, define $\mathfrak{n}_{1k} := |S^+|$ and $\mathfrak{t}_{1k} := |Q^+|$. For each $j \neq k$, define $\mathfrak{n}_{1j} := |\mathsf{RowAnc}_k(j)|$ and $\mathfrak{t}_{1j} := |\mathsf{ColAnc}_k(j)|$, and set $\mathfrak{n}_{2j} := \mathfrak{n} - \mathfrak{n}_{1j}$ and $\mathfrak{t}_{2j} := \mathfrak{t} - \mathfrak{t}_{1j}$. Finally, let $\mathfrak{n}_{1,\mathrm{p}} := \sum_{j=1}^{K} \mathfrak{n}_{1j}$, $\mathfrak{t}_{1,\mathrm{p}} := \sum_{j=1}^{K} \mathfrak{t}_{1j}$, $\rho_{\mathfrak{n}} := \mathfrak{n}_{1,\mathrm{p}}/\mathfrak{n}$, $\rho_{\mathfrak{t}} := \mathfrak{t}_{1,\mathrm{p}}/\mathfrak{t}$, $p_{\mathfrak{n}} := \max(\mathfrak{n}_{1,\mathrm{p}}, \mathfrak{t})$, $p_{\mathfrak{t}} := \max(\mathfrak{n}, \mathfrak{t}_{1,\mathrm{p}})$, $\zeta_{\mathfrak{n}} := \log(\mathfrak{n}_{1,\mathrm{p}} + \mathfrak{t})$, $\zeta_{\mathfrak{t}} := \log(\mathfrak{n} + \mathfrak{t}_{1,\mathrm{p}})$, $\widetilde{\gamma}_{\min} := c_\ell\, \gamma_{\min} \sqrt{\mathfrak{n}\mathfrak{t}/NT}$ and $\widetilde{\gamma}_{\max} := c_u\, \gamma_{\max} \sqrt{\mathfrak{n}\mathfrak{t}/NT}$. We assume the following.

**Assumption A5.** *There exist constants $0 < c_\ell \leq c_u$ such that*

$$c_\ell \frac{\mathfrak{n}_{1k}}{N} I_r \preceq U_{S^+}^\top U_{S^+} \preceq c_u \frac{\mathfrak{n}_{1k}}{N} I_r, \qquad c_\ell \frac{\mathfrak{n}}{N} I_r \preceq U_S^\top U_S \preceq c_u \frac{\mathfrak{n}}{N} I_r,$$

*and, for every $j \neq k$,*

$$c_\ell \frac{\mathfrak{n}_{1j}}{N} I_r \preceq U_{\mathsf{RowAnc}_k(j)}^\top U_{\mathsf{RowAnc}_k(j)} \preceq c_u \frac{\mathfrak{n}_{1j}}{N} I_r.$$

*We require the column factors satisfy the analogous conditions with $V, Q, Q^+, \mathsf{ColAnc}_k(j), \mathfrak{t}_{1j}$ and $T$ in place of $U, S, S^+, \mathsf{RowAnc}_k(j), \mathfrak{n}_{1j}$ and $N$, respectively.*

**Assumption A6.** *We have $r + \max(\zeta_{\mathfrak{n}}, \zeta_{\mathfrak{t}}) \leq c_{\mathrm{blk}} \min(\mathfrak{n} - r, \mathfrak{t} - r, \mathfrak{n}_{1k}, \mathfrak{t}_{1k})$, $\mathfrak{n} - r \geq c_{\mathrm{blk}}\mathfrak{n}$, and $\min(\zeta_{\mathfrak{n}}, \zeta_{\mathfrak{t}}) \leq c_{\mathrm{blk}} r$ for sufficiently small constants $c_0, c_{\mathrm{blk}} > 0$. Also, the noise level satisfies*

$$\widetilde{\theta} := \frac{\sigma}{\widetilde{\gamma}_{\min}} \max\left\{ \sqrt{\mathfrak{n}}, \sqrt{\mathfrak{t}}, \sqrt{\frac{\mathfrak{n}}{\rho_{\mathfrak{t}}}}, \sqrt{\frac{\mathfrak{n}\mathfrak{t}}{\mathfrak{n}_{1k}}} \right\} \leq c_0.$$

*Finally, we write $\widetilde{U} := U_S (U_S^\top U_S)^{-1/2}$ and $\widetilde{V} := V_Q (V_Q^\top V_Q)^{-1/2}$, and assume that $\widetilde{\nu}_x := \sqrt{\mathfrak{n}/r}\, \|\widetilde{U}_{R_{ak}}^\top x\|_2$ and $\widetilde{\nu}_y := \sqrt{\mathfrak{t}/r}\, \|\widetilde{V}_{C_{bk}}^\top y\|_2$ are of constant order.*

Assumption (A6) is the adaptation of (A2), (A3) and (A4) for staggered designs. Moreover, the first and third conditions in (A5) are the direct analogues of (A1), and require the observed row and column blocks retained in the auxiliary four-block problem to contain all $r$ latent directions in a well-conditioned way. The middle condition is the restricted analogue of the orthonormality condition $U^\top U = V^\top V = I_r$, and requires that the Gram matrices associated to $U_S$ and $V_Q$ remain full rank and well conditioned.

**Corollary 4.** *Consider a tensor $\mathcal{M}$ satisfying $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$, where $U \in \mathbb{R}^{N \times r}$ and $V \in \mathbb{R}^{T \times r}$ have orthonormal columns, and the core tensor is such that $0 < \gamma_{\min} \leq \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \leq \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \leq \gamma_{\max} < \infty$. Choose $k \in [K]$, indices $(a, b)$ such that $a + b > o_k + 1$, and unit vectors $x \in \mathbb{B}_2(N_{ak})$ and $y \in \mathbb{B}_2(T_{bk})$. Let $\mu_{xy}^{(k,a,b)}$ be as in (5), and define $\hat{\mu}_{xy}^{(k,a,b)}$ to be the output of Algorithm 2 run with $\tau \leq \frac{c_\ell \mathfrak{n}_{1k}}{2c_u \mathfrak{n}}$. Fix also absolute*

*constants* $0 < c_\ell \le c_u < \infty$, *and assume* (A5), (A6) *with* $\tilde{\nu}_x \neq 0, \tilde{\nu}_y \neq 0$. *Let*

$$\widetilde{\Upsilon}_{xy} := \frac{\sigma^2(r+\zeta_{\mathfrak{n}})}{\rho_{\mathfrak{n}}}\|\widetilde{U}_{R_{ak}}^\top x\|_2^2 + \frac{\sigma^2(r+\zeta_{\mathfrak{t}})}{\rho_{\mathfrak{t}}}\|\widetilde{V}_{C_{bk}}^\top y\|_2^2 + \frac{\sigma^2\mathfrak{n}}{\mathfrak{n}_{1k}}\|\widetilde{U}_{R_{ak}}^\top x\|_2^2\|\widetilde{V}_{C_{bk}}^\top y\|_2^2,$$

*and further assume that*

$$\frac{\widetilde{\gamma}_{\max}^2}{\tau}\frac{\mathfrak{n}_{1k}}{\mathfrak{n}}\left(p_{\mathfrak{n}}^{-10}+p_{\mathfrak{t}}^{-10}\right) + \frac{\sigma^2}{\tau}(\mathfrak{n}_{1k}+\mathfrak{t})\left(p_{\mathfrak{n}}^{-5}+p_{\mathfrak{t}}^{-5}\right) \le c_0\widetilde{\Upsilon}_{xy}. \tag{6}$$

*There exists a constant* $c_1 = c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa, \widetilde{\nu}_x, \widetilde{\nu}_y) < \infty$ *such that* $\mathbb{E}_{\mathcal{M}}[\{\widehat{\mu}_{xy}^{(k,a,b)} - \mu_{xy}^{(k,a,b)}\}^2] \le c_1\,\widetilde{\Upsilon}_{xy}$.

This result follows directly from Theorem 1, with the original dimensions and block sizes replaced by their auxiliary counterparts. To see why, observe that $U_S^\top U_S$ and $V_Q^\top V_Q$ are invertible under (A5), hence $\widetilde{U} := U_S(U_S^\top U_S)^{-1/2}$, $\widetilde{V} := V_Q(V_Q^\top V_Q)^{-1/2}$, and $\widetilde{\mathcal{C}}_{\bullet,\bullet,j} := (U_S^\top U_S)^{1/2}\mathcal{C}_{\bullet,\bullet,j}(V_Q^\top V_Q)^{1/2}$ give an orthonormal Tucker2 representation of the auxiliary signal $\mathcal{M}_{S,Q,j} = \widetilde{U}\widetilde{\mathcal{C}}_{\bullet,\bullet,j}\widetilde{V}^\top$. The auxiliary core tensor satisfies $0 < \tilde{\gamma}_{\min} \le \sigma_{\min}(\widetilde{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\widetilde{C}_{\bullet,\bullet,j}) \le \tilde{\gamma}_{\max}$, while $\tilde{U}$ and $\tilde{V}$ satisfy (A1) with $c_\ell/c_u$ and $c_u/c_\ell$ in place of $c_\ell$ and $c_u$, respectively; see the proof of Corollary 4 for the precise details. This, together with (A6), implies that the pooled upper and left matrices obey the same conditions as in the four-block setting, with $N, T, N_{1k}, T_{1k}, \rho_N, \rho_T, p_N, p_T, \zeta_N, \zeta_T, \gamma_{\min}, \gamma_{\max}, c_\ell, c_u$ replaced by $\mathfrak{n}, \mathfrak{t}, \mathfrak{n}_{1k}, \mathfrak{t}_{1k}, \rho_{\mathfrak{n}}, \rho_{\mathfrak{t}}, p_{\mathfrak{n}}, p_{\mathfrak{t}}, \zeta_{\mathfrak{n}}, \zeta_{\mathfrak{t}}, \widetilde{\gamma}_{\min}, \widetilde{\gamma}_{\max}$, $c_\ell/c_u, c_u/c_\ell$. This is precisely what is needed to apply Theorem 1, even though the auxiliary observation pattern need not consist of four contiguous blocks, and with these substitutions the stated result follows.

# 6 Simulations

## 6.1 Synthetic data

Code and dataset access for reproducing the simulations are available at https://github.com/abordino/FunctionalCausalTensor. In this subsection we empirically validate our theoretical claims on synthetic data. In particular, we verify that pooling improves performance for moderate values of $K$, while saturation occurs for large $K$, thereby confirming the phase transitions predicted by Theorem 1. We also include robustness checks with respect to rank misspecification, SNR levels, vector inputs, and mask dimensions. For staggered adoption designs, we show that pooling across layers reduces statistical error and demonstrate how to lower computational costs when computing an average counterfactual component over multiple blocks.

In Fig. 3(a), we fix $N = 100$, $T = 80$, $r = 6$ and vary $K \in \{1, 2, 5, 10, 20, 50, 200\}$. The matrices $U$ and $V$ are generated by drawing Gaussian random matrices and orthonormalising their columns. For each slice $j \in [K]$, we generate $\mathcal{C}_{\bullet,\bullet,j} = O_j\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r)\,\tilde{O}_j^\top$, where $O_j, \tilde{O}_j \in \mathbb{R}^{r\times r}$ are independent random orthonormal matrices, and the singular values are fixed at $(\sigma_1, \ldots, \sigma_r) = (2, 1.72, 1.44, 1.16, 0.88, 0.60)$. The observations are generated according to (2), with noise variance $\sigma^2$ chosen so that $\mathrm{SNR}^{-1} := \sigma^2 N/\sigma_r^2$ and $\mathrm{SNR} = 1$, and with block sizes $N_{1j} = 70$ and $T_{1j} = 60$ for all $j \in [K]$. We fix $k = 1$ and consider unit vectors $x \in \mathbb{R}^{N_{21}}$ and $y \in \mathbb{R}^{T_{21}}$ independently drawn from standard Gaussian distributions and then normalised to have unit norm. We compare five procedures for estimating (3):

1. Estimated pooled stands for Algorithm 1 with $\tau = 0.01$;

2. ESTIMATED NO-POOL stands for Algorithm 1 with $\tau = 0.01$, applied only to the target slice $\mathcal{Y}_{\bullet,\bullet,k}$;

3. ORACLE POOLED returns

$$\mathcal{M}^{(d)}_{\bullet,\bullet,k} + x^\top U_{2k}\, \mathcal{C}_{\bullet,\bullet,k}\, (W_{\text{up}}^\top W_{\text{up}})^{-1} W_{\text{up}}^\top\, (E^{\text{p}}_{\text{up}})_{\bullet,\{T_{1k}+1,\ldots,T\}}\, y \\ + x^\top (E^{\text{p}}_{\text{left}})_{\{N_{1k}+1,\ldots,N\},\bullet}\, W_{\text{left}} (W_{\text{left}}^\top W_{\text{left}})^{-1} \mathcal{C}_{\bullet,\bullet,k}\, V_{2k}^\top\, y;$$

4. ORACLE NO-POOL is the layer-specific counterpart of ORACLE POOLED and returns

$$\mathcal{M}^{(d)}_{\bullet,\bullet,k} + x^\top U_{2k} (U_{1k}^\top U_{1k})^{-1}\, U_{1k}^\top \mathcal{E}^{(b)}_{\bullet,\bullet,k}\, y + x^\top \mathcal{E}^{(c)}_{\bullet,\bullet,k} V_{1k} (V_{1k}^\top V_{1k})^{-1} V_{2k}^\top\, y;$$

5. ORACLE LOCAL returns $\mathcal{M}^{(d)}_{\bullet,\bullet,k} + x^\top U_{2k} (U_{1k}^\top U_{1k})^{-1}\, U_{1k}^\top (E^{\text{p}}_{\text{up}})_{\{\sum_{j=1}^{k-1} N_{1j}+1,\ldots,\sum_{j=1}^{k} N_{1j}\},\bullet}\, V\, V_{2k}^\top\, y.$

The oracle quantities correspond to the Gaussian terms in the expansion of $\hat{\mu}^{(k)}_{xy} - \mu^{(k)}_{xy}$ given in Lemma 14 in Appendix C, and capture the leading contribution to the stochastic error of our procedure. For each value of $K$ and for each of these estimators, we run 500 replications and report the average squared error. The results show that the NO-POOL estimators do not decay with $K$, as expected. By contrast, the two POOLED estimators exhibit an approximate $1/K$ decay and appear to approach the line corresponding to ORACLE LOCAL. This line is much lower than the others because it is related to both $\|U_{2k}^\top x\|_2$ and $\|V_{2k}^\top y\|_2$, yielding an additional factor of roughly $\sqrt{r/T} = \sqrt{6/80} = 0.075$. To better understand the relationship among these latter three methods, in Fig. 3(b) we repeat the same simulation study for $K \in \{50, 150, 300, 500, 1000\}$. The results show that ORACLE POOLED continues to decay with $K$, whereas Algorithm 1 saturates at approximately the level of ORACLE LOCAL. This is in accordance with Theorem 1.

Fig. 3(c) uses the same setup as Fig. 3(a), with a rank-6 signal tensor, but draws the mask dimensions randomly over $N_{1j} \in \{30, \ldots, 70\}$ and $T_{1j} \in \{30, \ldots, 60\}$. The query vectors are chosen to be aligned with the leading eigenspaces of $U_{2k}$ and $V_{2k}$, thereby increasing $\|U_{2k}^\top x\|_2$ and $\|V_{2k}^\top y\|_2$ and deliberately violating the incoherence condition in (A4). In addition, all estimators are run with misspecified rank $r + 5 = 11$. The results show that ESTIMATED POOLED still exhibits a similar phase transition as in Fig. 3(a), despite the misspecification and incoherence violation. By contrast, ESTIMATED NO-POOL is more sensitive to both sources of misspecification and has substantially larger error than ORACLE NO-POOL.

Finally, Fig. 3(d) uses the same setup as Fig. 3(b), but varies SNR $\in \{1, 10^{-2}, 10^{-4}\}$ to examine sensitivity under increasingly noisy regimes. The results indicate that the proposed methodology remains reasonably stable even at lower signal levels, with ESTIMATED POOLED appearing close to ORACLE LOCAL across the considered SNR values.

We also empirically evaluate Algorithm 2 in a synthetic staggered-adoption design. We generate a rank-5 Tucker2 signal tensor with $N = 150$, $T = 200$, and $K = 10$, where $U \in \mathbb{R}^{N\times r}$ and $V \in \mathbb{R}^{T\times r}$ are orthonormal and generated as before, and the entries of the core matrices $\mathcal{C}_{\bullet,\bullet,j} \in \mathbb{R}^{r\times r}$ are independent standard normal variables. We observe $\mathcal{Y}_{\bullet,\bullet,j} = P_{\Omega_{\bullet,\bullet,j}}(\mathcal{M}_{\bullet,\bullet,j} + \mathcal{E}_{\bullet,\bullet,j})$, where the entries of $\mathcal{E}$ are independent $\mathcal{N}(0, 0.03^2)$. The missingness masks $\Omega_{\bullet,\bullet,j}$ are generated by an irreversible adoption process in which, for each unit $i$ and layer $j$, the adoption time $A_{ij}$ is sampled independently from a layer-specific grid of adoption times, with probability 0.20 of never adopting. We then set $\Omega_{i,t,j} = \mathbb{1}\{t < A_{ij}\}$. We take $k = 1$ as the target layer and use the adoption ordering in this slice to reorder the units across all layers. Under this ordering,

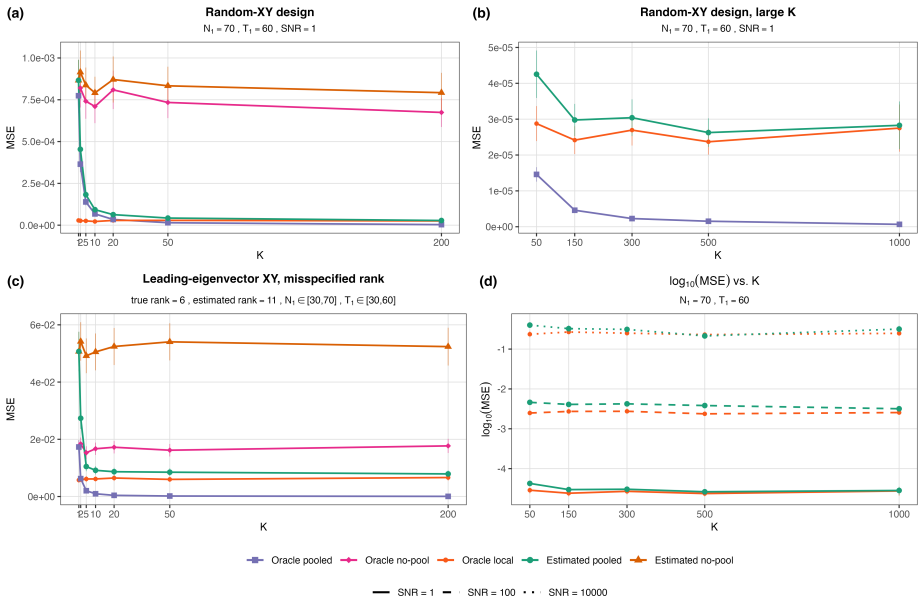


Figure 3: (a) Mean squared error of the five estimators as a function of $K \in \{1, 2, 5, 10, 20, 50, 200\}$, with $N = 100$, $T = 80$, $r = 6$, SNR $= 1$, and block sizes $N_{1j} = 70$, $T_{1j} = 60$ for all $j \in [K]$. Results are averaged over 500 replications. (b) Mean squared error of the pooled and local estimators for $K \in \{50, 150, 300, 500, 1000\}$ under the same simulation setting. (c) Analogue of (a) with violation of incoherence, rank-misspecification and masks sizes $N_{1j} \in \{30, \ldots, 70\}$ and $T_{1j} \in \{30, \ldots, 60\}$. (d) Sensitivity analysis with different SNR values. Error bars show $\pm 1.96$ standard error of the estimates.

$\Omega_{\bullet,\bullet,1}$ is a staircase and satisfies $\Omega_{i,t,1} = \mathbb{1}\{(i,t) \in R_{a1} \times C_{b1}$ for some $a, b$ with $a + b \leq 5\}$. We then target all six missing blocks and, for each of them, we generate 100 independent query pairs $x \in \mathbb{B}_2(N_{a1})$ and $y \in \mathbb{B}_2(T_{b1})$ by drawing independent standard normal vectors and normalising them to have unit norm. For each query we estimate (5) using both the tensor-pooled estimator outlined in Algorithm 2, and its matrix-only counterpart, which runs the procedure on $\mathcal{Y}_{\bullet,\bullet,k}$ only and does not borrow information from the other layers. In particular, this algorithm uses $\mathcal{Y}_{S_{k,a},Q^+_{k,a,b},k}$ and $\mathcal{Y}_{S^+_{k,a,b},Q_{k,b},k}$ in place of $Y^{\mathrm{p}}_{\mathrm{left}}$ and $Y^{\mathrm{p}}_{\mathrm{up}}$, respectively. Both procedures are run with $\tau \in \left\{10^{-0.5+0.025j} : j = 0, \ldots, 20\right\}$.

Figure 4 reports the mean absolute estimation error over the 100 random bilinear queries for each selected block and each value of $\tau$. The tensor-pooled estimator has substantially lower estimation error than the matrix-only estimator. This is expected, since the tensor method exploits the shared row and column subspaces across layers, whereas the matrix-only method uses only the two anchor blocks available in $\mathcal{Y}_{\bullet,\bullet,k}$. In general, the fact that certain blocks have larger errors than others can be attributed to differences in the size of the missing blocks: smaller missing blocks may have a lower signal-to-noise ratio, which can in turn reduce estimation accuracy. Furthermore, we see that both estimators are stable across values of $\tau$, particularly when $\tau \ll 1$. This agrees with our theory: under Assumption (A1) with $c_\ell > 0$, choosing $\tau$ small enough ensures that clipping is inactive with high probability, while stabilising the inverse on the complementary event. Thus, $\tau$ acts primarily as a safeguard rather than a tuning parameter, and careful tuning appears unnecessary.

Finally, we compare Algorithm 3 and Algorithm 4 in terms of statistical efficiency and runtime when

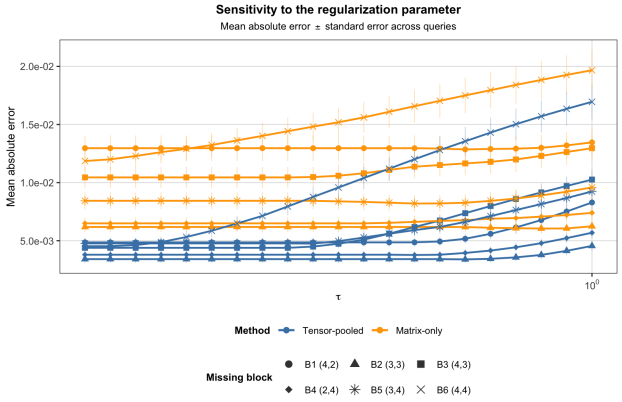


Figure 4: Mean absolute estimation error over 100 random bilinear queries for each target block. Algorithm 2 and its matrix counterpart are run for varying $\tau \in \{10^{-0.5+0.025j} : j = 0, \ldots, 20\}$.

estimating the aggregate ATE functional; this quantity and the two procedures are presented in detail in Appendix B.1. Informally, this estimand is a weighted average of the bilinear forms in (5) over all missing blocks in the target slice. We generate a rank-5 Tucker2 signal tensor with $N = 150$, $T = 200$, and $K = 5$, using random orthonormal matrices $U \in \mathbb{R}^{N\times r}$ and $V \in \mathbb{R}^{T\times r}$ and independent standard normal entries in each core matrix $\mathcal{C}_{\bullet,\bullet,j}$. The error tensor $\mathcal{E}$ has independent $\mathcal{N}(0, 0.01^2)$ entries. The target layer is $k = 1$ and has staircase missingness with $o_k \in \{4, 6, 8, 10, 12, 15, 20\}$; all non-target layers are fully observed. For each value of $o_k$, we run both procedures with $\tau = 10^{-3}$ over 1000 replications and report the average runtime divided by $o_k$ and the mean absolute deviation from the true value.

The results are consistent with the discussion in Appendix B.1. The runtime divided by $o_k$ appears approximately linear for Algorithm 3, reflecting its quadratic dependence on the number of target blocks, while it is nearly constant for Algorithm 4, reflecting the fact that the dominant SVD cost is linear in $o_k$. Also, Algorithm 3 has better statistical accuracy, as expected, since Algorithm 4 uses reduced anchor sets to improve runtime and therefore sacrifices some statistical efficiency.

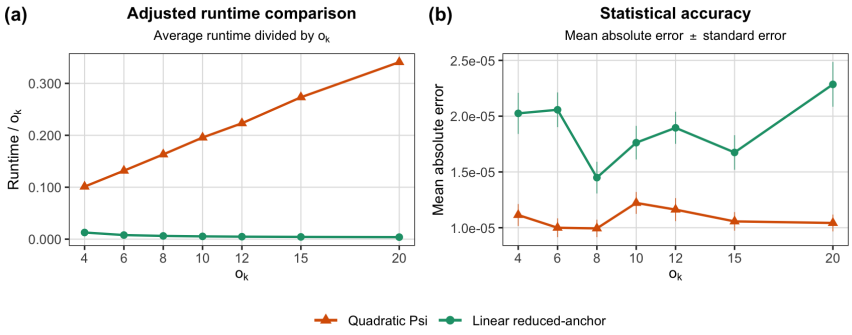


Figure 5: Runtime (left) and accuracy (right) comparison for Algorithms 3 and 4 when estimating the aggregate ATE functional in a synthetic staggered-adoption design. Here $o_k \in \{4, 6, 8, 10, 12, 15, 20\}$ denotes the number of row/time blocks induced by the staircase-adoption pattern in the target slice.

## 6.2 Real-data application: Castle Doctrine data

In this and the next section, we consider real-data applications motivated by causal inference. We work with two signal tensors, $\mathcal{M}(0)$ and $\mathcal{M}(1)$, corresponding to the untreated and treated responses, respectively. Following the potential-outcomes framework (Rubin, 1974), the entries of the fully observed data tensor $\mathcal{Y}$ satisfy $\mathcal{Y}_{itj} = \Omega_{itj}\,\mathcal{Y}_{itj}(0) + (1-\Omega_{itj})\,\mathcal{Y}_{itj}(1)$, where $\mathcal{Y}(0)$ denotes the untreated potential outcome, which is observed on $\{(i,t,j) : \Omega_{itj} = 1\}$ and missing on the complementary set, and $\mathcal{Y}(1)$ denotes the treated potential outcome, which is observed only over $\{(i,t,j) : \Omega_{itj} = 0\}$. Here, we focus on estimating bilinear forms of $\mathcal{M}(0)$ and $\mathcal{M}(1)$ over the treated region. The latter problem is straightforward because $\mathcal{Y}(1)$ is observed on this region, so simple plug-in estimators can be used. By contrast, $\mathcal{Y}(0)$ is unobserved, hence bilinear functionals of $\mathcal{M}(0)$ require different approaches such as those introduced in the previous sections.

The empirical study considered here is based on the Castle Doctrine data from the `PolicyEval` repository, available on GitHub, which records U.S. state-level public-safety outcomes together with the adoption of laws expanding the legal right to use force in self-defense, often referred to as Castle Doctrine or Stand Your Ground laws. The dataset is a standard staggered-adoption benchmark in the difference-in-differences literature (e.g. Cheng and Hoekstra, 2013).

After processing the data as described in Appendix B.2, we obtain two tensors, $\mathcal{Y}(0) \in \mathbb{R}^{50\times 11\times 4}$ and $\mathcal{Y}(1) \in \mathbb{R}^{50\times 11\times 4}$, representing policy-off and policy-on potential outcomes, respectively. The policy-off tensor $\mathcal{Y}(0)$ follows a staggered-adoption missingness pattern, while $\mathcal{Y}(1)$ is observed on the complementary policy-on region where $\mathcal{Y}(0)$ is missing. Figure 9 in Appendix B.2 shows the resulting observation patterns. Here, rows represent U.S. states, columns represent calendar years from 2000 to 2010, and slices represent the four logged crime-rate outcomes `l_motor`, `l_robbery`, `l_assault`, and `l_homicide`; these correspond respectively to log-transformed motor theft, robbery, aggravated assault, and homicide rates. Importantly, the tensor slices correspond to different outcomes rather than different policy regimes. As a result, since policy adoption is common across outcomes, the staggered-adoption missingness pattern is shared by all layers of $\mathcal{Y}(0)$.

To assess policy efficacy, we fix a target outcome slice $k \in [4]$ and sort its rows so that the adoption pattern is a staircase with $\Omega_{i,t,k} = \mathbb{1}\{(i,t) \in R_{ak} \times C_{bk}$ for $a, b$ with $a+b \le o_k+1\}$ for some integer $o_k \ge 2$. Let $\mathcal{D}_k := \{(a,b) : a+b > o_k+1\}$ be the set of policy-on target blocks. For $c \in \{0,1\}$, we consider $\Psi_c^{(h)}(k) \propto \sum_{(a,b)\in\mathcal{D}_k} x_{a,h}^\top \mathcal{M}_{\bullet,\bullet,k}^{(a,b)}(c)\, y_{b,h}$, where $\mathcal{M}_{\bullet,\bullet,k}^{(a,b)}(c)$ is the block of $\mathcal{M}(c)$ restricted to rows $R_{ak}$ and columns $C_{bk}$, and $x_{a,h}, y_{b,h}$ are the query vectors for four specific bilinear forms $h \in \{\text{ATE}, \textsc{RowHet}, \textsc{Local}\text{-}i_0, \textsc{Trend}\}$. These summaries are, respectively, an average potential outcome, a signed row contrast, a row-specific average, and a within-block temporal slope; see Appendix B.1 for the formal definition of these quantities. At the sample level, we estimate the $\Psi_0^{(h)}(k)$'s by applying Algorithm 3 to $\mathcal{Y}(0)$, as well as its matrix counterpart, which runs the same procedure on $\mathcal{Y}_{\bullet,\bullet,k}$ only; both methods are run with $\tau = 10^{-2}$ and $r = 3$. We prefer this approach to Algorithm 4 because of its greater statistical accuracy, especially given that the number of missing blocks in this application is relatively small. By contrast, the $\Psi_1^{(h)}(k)$'s are easier to estimate because $\mathcal{Y}(1)$ is fully observed over $\mathcal{D}_k$, so we use plug-in estimators for them. We also consider the induced policy effects $\Delta^{(h)}(k) := \Psi_1^{(h)}(k) - \Psi_0^{(h)}(k)$ and estimate them by subtracting the corresponding estimators.

Table 2 shows the estimates for ATE, RowHet, Local-$i_0$, Trend in the target slice $k = 2$ corresponding

to `l_robbery`. For LOCAL-$i_0$, we consider three values of $i_0$ associated to Florida, Montana and Texas. For the ROWHET functional, we set $\eta_i = +1$ for states that voted Republican in the 2000 presidential election and $\eta_i = -1$ for states that voted Democratic, so that ROWHET should be interpreted as a contrast between these two groups of states. Confidence intervals are computed using a bootstrap procedure with $B = 500$ samples. Specifically, we sample rows with replacement from the target layer while keeping the other layers fixed, making the comparison with the matrix estimator fairer since the latter uses only $\mathcal{Y}_{\bullet,\bullet,k}$ and is therefore unaffected by sampling uncertainty in the additional slices. For each bootstrap sample, we recompute the estimators, take the standard deviation of the resulting estimates as the standard error, and report confidence intervals as the point estimate $\pm 1.96$ standard errors.

The results show little evidence of an average or local-level effect, as for both estimators the ATE, local summaries, and ROWHET confidence intervals all include zero. The main exception is the TREND functional, which is significantly negative under both the pooled tensor estimator, $\widehat{\Delta}^{(\text{TREND})} = -0.1359$ with confidence interval $(-0.2165, -0.0554)$, and the matrix analogue, $\widehat{\Delta}^{(\text{TREND})}_{\text{mat}} = -0.1098$ with confidence interval $(-0.2069, -0.0126)$. Thus, the clearest signal is a negative post-adoption trend rather than an average or state-specific level effect, suggesting that U.S. states that adopted Castle Doctrine laws were more likely to experience a mild and gradual decline in robbery rates over the post-adoption period, rather than a sharp immediate drop at the time of adoption. This underscores the importance of considering functionals beyond simple averages.

Table 2: Estimates of $\Delta^{(h)}(2)$ using the pooled tensor estimator of Algorithm 3 and its matrix counterpart. Entries report point estimates, with 95% confidence intervals in parentheses below. These are computed by resampling rows from the target layer only, while keeping the other layers fixed; bootstrap standard errors are then used to report intervals as the point estimate $\pm 1.96$ standard errors.

| | ATE | LOCAL-FLORIDA | LOCAL-MONTANA | LOCAL-TEXAS | ROWHET | TREND |
|---|---|---|---|---|---|---|
| $\widehat{\Delta}^{(h)}$ | -0.0241<br>(-0.0948, 0.0466) | 0.0621<br>(-0.2061, 0.3303) | -0.1958<br>(-0.4880, 0.0964) | -0.0197<br>(-0.2333, 0.1939) | -0.0171<br>(-0.0844, 0.0503) | -0.1359<br>(-0.2165, -0.0554) |
| $\widehat{\Delta}^{(h)}_{\text{mat}}$ | 0.0337<br>(-0.0270, 0.0945) | -0.0094<br>(-0.2854, 0.2667) | 0.0104<br>(-0.2133, 0.2341) | 0.0356<br>(-0.1459, 0.2170) | 0.0347<br>(-0.0244, 0.0939) | -0.1098<br>(-0.2069, -0.0126) |

Figure 6 also reports estimates of $\Psi_0^{(h)}(2)$ under more restrictive missingness patterns. Specifically, we rerun the simulations for three additional versions of $\mathcal{Y}(0)$, where additional missing entries are introduced in layer $k = 2$ by retaining only the first 3 columns and the first 5, 10, and 15 rows, respectively; all other layers are left unchanged. The fourth panel shows results for the original staggered missingness pattern. The point estimates remain close across methods, but the tensor confidence intervals are smaller in the first panel, where the target layer contains the least information. This suggests that, conditional on the additional layers, the tensor method reduces uncertainty relative to the matrix method by borrowing information across slices. As more rows are retained, the target layer becomes more informative, and the confidence intervals of the two methods become comparable.

## 6.3 Real-data application: COVID-19 data

For this application we construct a COVID-19 panel by merging policy-response indicators from the Oxford COVID-19 Government Response Tracker with epidemiological outcomes from the Our World in Data COVID-19 dataset.

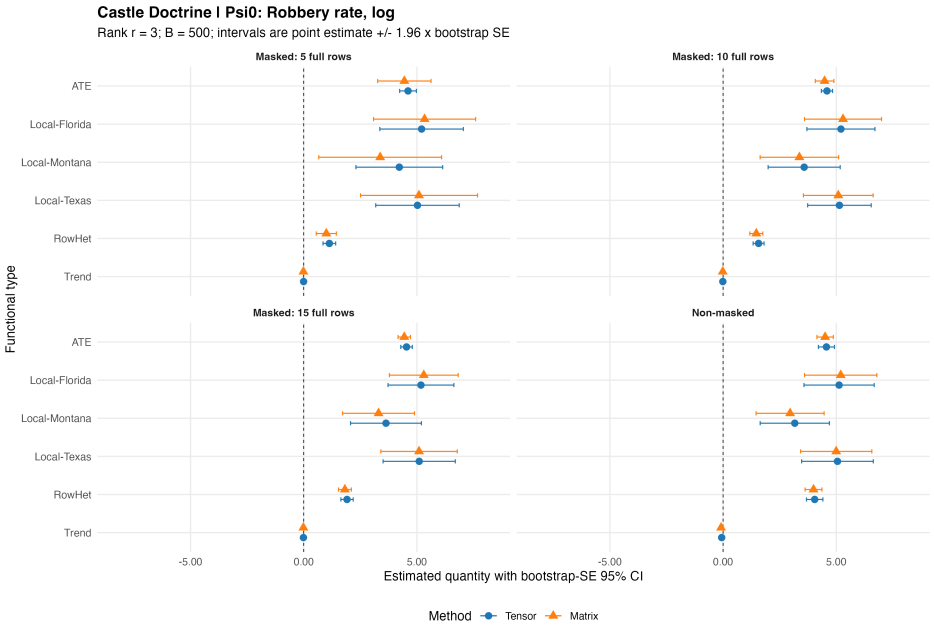


Figure 6: Estimates of $\Psi_0^{(h)}(2)$ using the pooled tensor estimator of Algorithm 3 and its matrix counterpart. Confidence intervals are computed as in Table 2. The first three panels correspond to the masked version of $\mathcal{Y}(0)$ where additional missingness is introduced by retaining only the first 3 columns and the first 5, 10, and 15 rows, respectively; all other layers are left unchanged. The fourth panel shows results for the original staggered missingness pattern.

We focus on two policies, `C6_stay_at_home_requirements` and `H3_contact_tracing`, over the window from March 05, 2020, to April 05, 2020. We then retain 18 countries, mostly European, where adoption of both policies was irreversible within this one-month period. This yields the tensor dataset $\mathcal{Y} \in \mathbb{R}^{18\times 32\times 2}$, where the modes correspond to countries, days, and policies. The first slice $\mathcal{Y}_{\bullet,\bullet,1}$ is associated with the 28-day delayed outcome `new_deaths_smoothed_per_million` and the stay-at-home policy, while the second slice $\mathcal{Y}_{\bullet,\bullet,2}$ is associated with the 28-day delayed outcome `new_cases_smoothed_per_million` and the contact-tracing policy. The 28-day delay accounts for the time between policy adoption and its potential effect on reported cases or deaths. Also, this pairing seems the most natural since stay-at-home requirements may affect downstream mortality, while contact tracing more directly targets contagion and hence reported cases.

The outcome tensor is plotted in Fig. 7. In each panel, blue entries indicate periods in which the policy is inactive, red entries indicate periods in which it is active, and colour intensity reflects the magnitude of the corresponding outcome. As before, policy status determines the staggered-adoption pattern: $\mathcal{Y}(0)$ contains only blue entries and therefore has staggered missingness, while $\mathcal{Y}(1)$ contains only red entries.

Figure 8 reports the estimates of $\Psi_0^{(h)}(1)$ and $\Delta^{(h)}(1)$ with $h \in \{\text{ATE}, \text{TREND}\}$ for the target layer $k = 1$ corresponding to `new_deaths_smoothed_per_million` under `C6_stay_at_home_requirements`. Pointwise estimates are computed with Algorithm 3 and its matrix-only counterpart, both run with $\tau = 10^{-4}$ and $r = 3$; confidence intervals are calculated as in Section 6.2 with $B = 500$. For both methods, the estimated effect for the trend functional is close to zero, indicating little evidence of a systematic change in the post-adoption trend over this period. The estimates of the average treatment effect, however, differ across the two methods. In particular, the matrix confidence interval includes zero, whereas the tensor estimator

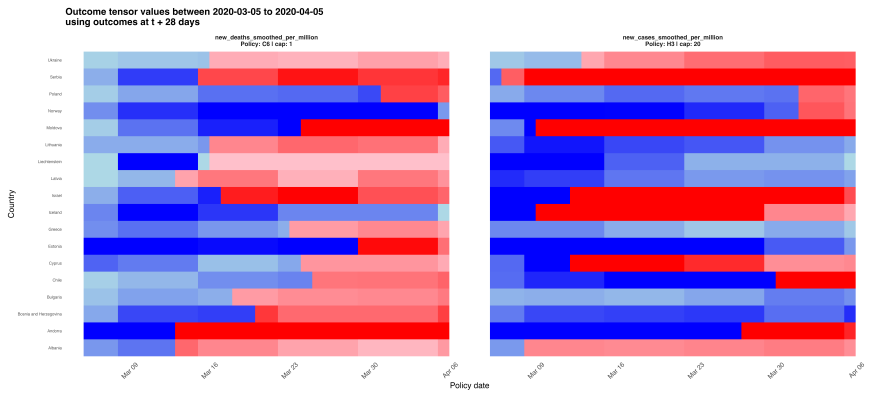


Figure 7: Outcome tensor from the merged OxCGRT–OWID panel. Rows are countries, columns are dates from March 05 to April 05, 2020, and panels are outcome–policy pairs. Outcomes are measured at a 28-day delay, so date $t$ shows the value at $t + 28$. Blue denotes policy-off entries, red policy-on entries, and darker shades larger outcome values.

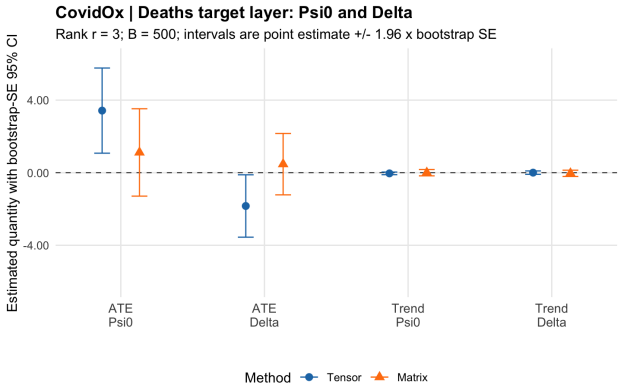


Figure 8: Estimates of $\Psi_0^{(h)}(1)$ and $\Delta^{(h)}(1)$ using the pooled tensor estimator of Algorithm 3 and its matrix counterpart. Confidence intervals are computed as in Section 6.2 with $B = 500$.

gives a significantly negative model-based estimate. Under the maintained low-rank counterfactual model and the assumed interpretation of policy timing, this is consistent with a reduction in new deaths associated with stay-at-home requirements. The difference between the two estimators, and in particular the counterintuitive conclusion of the matrix estimator that stay-at-home policies did not help reduce deaths, can be attributed to the fact that in the target layer only two rows, corresponding to Norway and Iceland, are fully observed. This makes estimation more challenging. On the other hand, the tensor method overcomes this difficulty by borrowing information on the latent unit and time factors from the second layer, which contains `new_cases_smoothed_per_million` under `H3_contact_tracing`. This application therefore provides a setting in which our methodology offers a clear advantage over standard matrix methods.

## Acknowledgements

The research of the second author was supported by European Research Council Starting Grant 101163546.

## References

Alberto Abadie. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal Matrix Completion. In *Proceedings of the 36th Annual Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3821–3826, 2023.

Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic Interventions: Extending Synthetic Controls to Multiple Treatments. *Operations Research*, 74(2):840–859, 2025.

Anish Agarwal, Jungjun Choi, and Ming Yuan. Robust Matrix Estimation with Side Information, 2026. URL https://arxiv.org/abs/2603.24833.

Joshua Agterberg. Statistically and Computationally Optimal Estimation and Inference of Common Subspaces, 2026. URL https://arxiv.org/abs/2606.06483.

Jesús Arroyo, Avanti Athreya, Joshua Cape, Guodong Chen, Carey E. Priebe, and Joshua T. Vogelstein. Inference for Multiple Heterogeneous Networks with a Common Invariant Subspace. *Journal of Machine Learning Research*, 22(142):1–49, 2021.

Susan Athey and Guido W. Imbens. Design-based analysis in Difference-In-Differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, 116(536): 1716–1730, 2021.

Jonathan Auerbach, Martin Slawski, and Shixue Zhang. Tensor Completion for Causal Inference with Multivariate Longitudinal Data: A Reevaluation of COVID-19 Mandates, 2022. URL https://arxiv.org/abs/2203.04689.

Tavor Z. Baharav, Phillip B. Nicol, Rafael A. Irizarry, and Rong Ma. Stacked SVD or SVD stacked? A Random Matrix Theory perspective on data integration, 2025. URL https://arxiv.org/abs/2507.22170.

Jushan Bai and Serena Ng. Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data. *Journal of the American Statistical Association*, 116(536):1746–1763, 2021.

Ercument Cahan, Jushan Bai, and Serena Ng. Factor-Based Imputation of Missing Values and Covariances in Panel Data of Large Dimensions. *Journal of Econometrics*, 233(1):113–131, 2023.

Emmanuel J. Candès and Benjamin Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral Methods for Data Science: A Statistical Perspective. *Foundations and Trends in Machine Learning*, 14(5):566–806, 2021.

Cheng Cheng and Mark Hoekstra. Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine. *Journal of Human Resources*, 48(3):821–854, 2013.

Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Yasuko Chikuse. *Statistics on Special Manifolds*, volume 174 of *Lecture Notes in Statistics*. Springer, New York, 2003.

Jungjun Choi and Ming Yuan. Matrix Completion When Missing Is Not at Random and Its Applications in Causal Panel Data Models. *Journal of the American Statistical Association*, 2026. To appear.

Kenneth R. Davidson and Stanislaw J. Szarek. Local Operator Theory, Random Matrices and Banach Spaces. In *Handbook of the Geometry of Banach Spaces*, volume 1, pages 317–366. Elsevier, 2001.

Chenyin Gao, Han Chen, Anru R. Zhang, and Shu Yang. Causal Inference on Sequential Treatments via Tensor Completion, 2025. URL https://arxiv.org/abs/2511.15866.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59 – 79, 1995.

David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices, 2010. URL https://arxiv.org/abs/1001.2738.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396): 945–960, 1986.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, 2015.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3): 455–500, 2009.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Ziqi Liu, Ye Tian, and Weijing Tang. Representation Learning with Blockwise Missingness and Signal Heterogeneity, 2026. URL https://arxiv.org/abs/2602.11511.

Zhengchi Ma and Rong Ma. Optimal estimation of shared singular subspaces across multiple noisy matrices. *IEEE Transactions on Information Theory*, 72(5):3277–3300, 2026.

Debmalya Mandal and David Parkes. Weighted Tensor Completion for Time-Series Causal Inference, 2019. URL https://arxiv.org/abs/1902.04646.

Francesco Mezzadri. How to Generate Random Matrices from the Classical Compact Groups. *Notices of the American Mathematical Society*, 54(5):592–604, 2007.

Sahand N. Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

João D. Semedo, Amin Zandvakili, Christian K. Machens, Byron M. Yu, and Adam Kohn. Cortical Areas Interact through a Communication Subspace. *Neuron*, 102(1):249–259.e4, 2019.

G. W. Stewart. The Efficient Generation of Random Orthogonal Matrices with an Application to Condition Estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409, 1980.

Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2019.

Martin J Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 2019.

Eric Xia, Yuling Yan, and Martin J. Wainwright. Inference under Staggered Adoption: Case Study of the Affordable Care Act, 2024. URL https://arxiv.org/abs/2412.09482.

Yuling Yan and Martin J. Wainwright. Entrywise Inference for Missing Panel Data: A Simple and Instance-Optimal Approach, 2024. URL https://arxiv.org/abs/2401.13665.

# A Proofs

## A.1 Proofs for Section 2

*Proof of Theorem 1.* The proof of this result relies on the matrix denoising theory developed in Appendix C. Start by writing $\hat{\mu}_{xy}^{(k)} - \mu_{xy}^{(k)} = Z_{xy}^{(1)} + Z_{xy}^{(2)} + Z_{xy}^{(3)} + Z_{xy}^{(4)} + \Delta_{xy} =: Z_{xy} + \Delta_{xy}$, where

$$
\begin{aligned}
Z_{xy}^{(1)} &:= x^\top (E_{\text{left}}^{\text{p}})_{\mathcal{I}_k,\bullet} W_{\text{left}} (W_{\text{left}}^\top W_{\text{left}})^{-1} \mathcal{C}_{\bullet,\bullet,k} (W_{\text{up}}^\top W_{\text{up}})^{-1} W_{\text{up}}^\top (E_{\text{up}}^{\text{p}})_{\bullet,\mathcal{J}_k} y, \\
Z_{xy}^{(2)} &:= x^\top U_{2k} \mathcal{C}_{\bullet,\bullet,k} (W_{\text{up}}^\top W_{\text{up}})^{-1} W_{\text{up}}^\top (E_{\text{up}}^{\text{p}})_{\bullet,\mathcal{J}_k} y, \\
Z_{xy}^{(3)} &:= x^\top U_{2k} (U_{1k}^\top U_{1k})^{-1} U_{1k}^\top (E_{\text{up}}^{\text{p}})_{\mathcal{I}_k^{\text{up}},\bullet} V V_{2k}^\top y, \\
Z_{xy}^{(4)} &:= x^\top (E_{\text{left}}^{\text{p}})_{\mathcal{I}_k,\bullet} W_{\text{left}} (W_{\text{left}}^\top W_{\text{left}})^{-1} \mathcal{C}_{\bullet,\bullet,k} V_{2k}^\top y.
\end{aligned}
$$

Under the assumptions of Theorem 1, Lemma 15 gives $\mathbb{E}[Z_{xy}^2] \leq c_1 \, \Upsilon_{xy}$ for a sufficiently large constant $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\text{blk}}, \kappa, \nu_x, \nu_y) > 0$. Moreover, the same result ensures that there exists an event $\mathcal{G}_1$ with $\mathbb{P}(\mathcal{G}_1) \geq 1 - \mathcal{O}(p_N^{-10} + p_T^{-10})$ such that $\Delta_{xy}^2 \leq c_1 \, \Upsilon_{xy}$ under $\mathcal{G}_1$.

Possibly enlarging the constant $c_1$ and allowing it to change from line to line, and writing $\hat{\mu} \equiv \hat{\mu}_{xy}^{(k)}$, $\mu \equiv \mu_{xy}^{(k)}$ to simplify the notation, we then decompose the mean squared error as

$$
\begin{aligned}
\mathbb{E}\left[\{\hat{\mu} - \mu\}^2\right] &= \mathbb{E}\left[\{\hat{\mu} - \mu\}^2 \mathbb{1}_{\mathcal{G}_1}\right] + \mathbb{E}\left[\{\hat{\mu} - \mu\}^2 \mathbb{1}_{\mathcal{G}_1^{\complement}}\right] \\
&= \mathbb{E}\left[\{Z_{xy} + \Delta_{xy}\}^2 \mathbb{1}_{\mathcal{G}_1}\right] + \mathbb{E}\left[\{\hat{\mu} - \mu\}^2 \mathbb{1}_{\mathcal{G}_1^{\complement}}\right] \\
&\leq 2\mathbb{E}\left[Z_{xy}^2\right] + 2\mathbb{E}\left[\Delta_{xy}^2 \mathbb{1}_{\mathcal{G}_1}\right] + 2\,\mathbb{E}\left[\hat{\mu}^2 \mathbb{1}_{\mathcal{G}_1^{\complement}}\right] + 2\mu^2\,\mathbb{P}(\mathcal{G}_1^{\complement}) \\
&\leq 2\mathbb{E}[Z_{xy}^2] + 2c_1 \Upsilon_{xy} + 2\,\mathbb{E}\left[\hat{\mu}^2 \mathbb{1}_{\mathcal{G}_1^{\complement}}\right] + 2\mu^2\,\mathbb{P}(\mathcal{G}_1^{\complement}) \\
&\leq c_1\,\Upsilon_{xy} + 2\,\mathbb{E}\left[\hat{\mu}^2 \mathbb{1}_{\mathcal{G}_1^{\complement}}\right] + c_1\,\mu^2 (p_N^{-10} + p_T^{-10}),
\end{aligned}
$$

where the previous decomposition and the bound on the second moment of $Z_{xy}$ from Lemma 15 are used to control the contribution on the good event $\mathcal{G}_1$, while the probability bound for $\mathcal{G}_1^{\complement}$ is used in the last inequality. We next derive separate bounds for each of the remaining two terms. For the third term, we have $|\mu| \leq \|\mathcal{C}_{\bullet,\bullet,k}\|_{\text{op}} \|U_{2k}^\top x\|_2 \, \|V_{2k}^\top y\|_2 \leq \gamma_{\max}$, which yields

$$
\mu^2 \, (p_N^{-10} + p_T^{-10}) \leq \gamma_{\max}^2 \, (p_N^{-10} + p_T^{-10}) \leq c_1 \tau^{-1} \, \gamma_{\max}^2 \, (p_N^{-10} + p_T^{-10}) \, \frac{N_{1k}}{N}.
$$

The last inequality follows from $\tau \leq c_\ell N_{1k} / (2N)$. For the second term, we start by noticing that $\big\|\hat{H}_{k,\tau}^{\text{inv}} \, \widehat{U}_{1k}^\top\big\|_{\text{op}}^2 = \big\|\hat{H}_{k,\tau}^{\text{inv}} \hat{H}_k \hat{H}_{k,\tau}^{\text{inv}}\big\|_{\text{op}} = \max_{i \in [r]} \lambda_i / (\lambda_i \vee \tau)^2 \leq \tau^{-1}$. This allows showing

$$
\begin{aligned}
|\hat{\mu}| &= \left|x^\top \hat{U}_{2k} \hat{H}_{k,\tau}^{\text{inv}} \hat{U}_{1k}^\top \hat{U}_{\text{up}}^{(k)} \hat{\Sigma}_{\text{up}} \hat{V}_{2k}^\top y\right| \leq \big\|\hat{U}_{2k}^\top x\big\|_2 \big\|\hat{H}_{k,\tau}^{\text{inv}} \hat{U}_{1k}^\top\big\|_{\text{op}} \big\|\hat{U}_{\text{up}}^{(k)} \hat{\Sigma}_{\text{up}}\big\|_{\text{op}} \big\|\hat{V}_{2k}^\top y\big\|_2 \\
&\leq \tau^{-1/2} \, \|(Y_{\text{up}}^{\text{p}})_{\mathcal{I}_k^{\text{up}},\bullet} \hat{V}_{\text{up}}\|_{\text{op}} \leq \tau^{-1/2} \|(Y_{\text{up}}^{\text{p}})_{\mathcal{I}_k^{\text{up}},\bullet}\|_{\text{op}} = \tau^{-1/2} \|Y_{\text{up}}^{(k)}\|_{\text{op}} \\
&\leq \tau^{-1/2} \|M_{\text{up}}^{(k)}\|_{\text{op}} + \tau^{-1/2} \|E_{\text{up}}^{(k)}\|_{\text{op}} = \tau^{-1/2} \|U_{1k} \mathcal{C}_{\bullet,\bullet,k} V^\top\|_{\text{op}} + \tau^{-1/2} \|E_{\text{up}}^{(k)}\|_{\text{op}} \\
&\leq \tau^{-1/2} c_u^{1/2} \gamma_{\max} \sqrt{N_{1k}/N} + \tau^{-1/2} \big\|E_{\text{up}}^{(k)}\big\|_{\text{op}},
\end{aligned}
$$

where the last inequality follows from (A1). This, together with an application of the Cauchy–Schwarz inequality and Lemma 22 with $p = 4$, gives

$$\begin{aligned}\mathbb{E}\left[\hat{\mu}^2 \mathbb{1}_{\mathcal{G}_1^{\complement}}\right] &\le 2\tau^{-1} c_u \gamma_{\max}^2 \frac{N_{1k}}{N} \mathbb{P}\big(\mathcal{G}_1^{\complement}\big) + 2\tau^{-1} \mathbb{E}\left[\big\|E_{\mathrm{up}}^{(k)}\big\|_{\mathrm{op}}^2 \mathbb{1}_{\mathcal{G}_1^{\complement}}\right] \\ &\le c_1 \tau^{-1} \gamma_{\max}^2 \frac{N_{1k}}{N} \big(p_N^{-10} + p_T^{-10}\big) + c_1 \tau^{-1} \big(p_N^{-5} + p_T^{-5}\big) \mathbb{E}^{1/2}\left[\big\|E_{\mathrm{up}}^{(k)}\big\|_{\mathrm{op}}^4\right] \\ &\le c_1 \left\{\tau^{-1} \gamma_{\max}^2 \frac{N_{1k}}{N} \big(p_N^{-10} + p_T^{-10}\big) + \tau^{-1} \big(p_N^{-5} + p_T^{-5}\big) \sigma^2 \big(N_{1k} + T\big)\right\}.\end{aligned}$$

Combining the previous bounds and using $\tau^{-1}\,\gamma_{\max}^2\,(p_N^{-10} + p_T^{-10})\,\frac{N_{1k}}{N} \;+\; \tau^{-1}(p_N^{-5} + p_T^{-5})\,(\sigma^2 N_{1k} + \sigma^2 T) \le c_0\,\Upsilon_{xy}$ concludes the proof. □

## A.2 Proofs for Section 4

*Proof of Theorem 2.* We prove the result by reducing the problem to a one-dimensional parametric submodel. To this end, we assume that $U_{0,2k}^{\top} x \neq 0$ and $V_{0,2k}^{\top} y \neq 0$; if either of these conditions fails, the desired lower bound is trivially satisfied. Let

$$F_k := (U_{0,2k}^{\top}\, x)(V_{0,2k}^{\top}\, y)^{\top}, \qquad G_k := \frac{F_k}{\|F_k\|_F^2}.$$

Write $C_{0,j} := (\mathcal{C}_0)_{\bullet,\bullet,j}$ to simplify the notation. For $\theta \in \mathbb{R}$, define a perturbed core tensor $\mathcal{C}(\theta)$ by keeping all slices except the $k$-th one fixed, and setting $C_k(\theta) := C_{0,k} + \theta\, G_k$. We then set $\mathcal{M}(\theta) := \mathcal{C}(\theta) \times_1 U_0 \times_2 V_0 \times_3 I_K$. We next verify the range of $\theta$ for which the path remains in the local parameter space. Since $U_0$ and $V_0$ have orthonormal columns, we have

$$\|C_k(\theta) - C_{0,k}\|_{\mathrm{op}} \le \|C_k(\theta) - C_{0,k}\|_F = \frac{|\theta|}{\|F_k\|_F}, \qquad \|\mathcal{M}(\theta) - \mathcal{M}_0\|_F = \|U_0\{C_k(\theta) - C_{0,k}\}V_0^{\top}\|_F = \frac{|\theta|}{\|F_k\|_F}.$$

We thus deduce that the path $\{\mathcal{M}(\theta) : |\theta| \le h\}$ is contained in $\{\mathcal{M} \in \mathcal{F}(c_\ell, c_u) : \|\mathcal{M} - \mathcal{M}_0\|_F \le \varsigma\}$ whenever $h \le \|F_k\|_F \min(\varsigma, \delta_{\gamma,k})$. In particular, the singular values of $C_k(\theta)$ remain in $[\gamma_{\min}, \gamma_{\max}]$ by the operator-norm bound above and Weyl's inequality (Lemma 23). Assumption (A1) continues to hold along the path because $U_0$ and $V_0$ are fixed.

We next compute the induced change in the target functional. Since the bottom-right block of the $k$-th slice is $U_{0,2k} C_k(\theta) V_{0,2k}^{\top}$, we have

$$\mu(\theta) := \mu_{xy}^{(k)}(\mathcal{M}(\theta)) = x^{\top}\, U_{0,2k} C_k(\theta) V_{0,2k}^{\top}\, y = \mu_{xy}^{(k)}(\mathcal{M}_0) + \theta\, x^{\top} U_{0,2k}\, G_k V_{0,2k}^{\top}\, y = \mu_{xy}^{(k)}(\mathcal{M}_0) + \theta.$$

Hence estimating $\mu_{xy}^{(k)}(\mathcal{M}(\theta))$ along this path is equivalent to estimating the scalar parameter $\theta$, up to the known additive constant $\mu_{xy}^{(k)}(\mathcal{M}_0)$. It is also useful to compute the Fisher information for $\theta$. Using the Gaussianity of the error and the four-block structure of $\Omega_{\bullet,\bullet,k}$, the Fisher information is constant and equal to

$$\mathfrak{I}_\theta = \sigma^{-2} \left\|P_{\Omega_{\bullet,\bullet,k}}(U_0 G_k V_0^{\top})\right\|_F^2 = \sigma^{-2} \left\{\|U_{0,1k} G_k V_{0,1k}^{\top}\|_F^2 + \|U_{0,1k} G_k V_{0,2k}^{\top}\|_F^2 + \|U_{0,2k} G_k V_{0,1k}^{\top}\|_F^2\right\}$$

$$
\begin{aligned}
&= \sigma^{-2} \left\{ \|U_{0,1k} G_k\|_F^2 + \|G_k V_{0,1k}^\top\|_F^2 - \|U_{0,1k} G_k V_{0,1k}^\top\|_F^2 \right\} \\
&= \sigma^{-2} \|F_k\|_F^{-4} \left\{ \|U_{0,1k} F_k\|_F^2 + \|F_k V_{0,1k}^\top\|_F^2 - \|U_{0,1k} F_k V_{0,1k}^\top\|_F^2 \right\} \\
&\leq \sigma^{-2} \|F_k\|_F^{-4} \left\{ \|U_{0,1k} F_k\|_F^2 + \|F_k V_{0,1k}^\top\|_F^2 \right\} \\
&\leq \sigma^{-2} c_u \|F_k\|_F^{-2} \left\{ \frac{N_{1k}}{N} + \frac{T_{1k}}{T} \right\},
\end{aligned}
$$

where the last inequality follows from Assumption (A1). Note that only the $k$-th slice contributes to $\mathfrak{I}_\theta$, since the submodel is fixed along all other slices.

We now apply the van Trees inequality (Gill and Levit, 1995, Equation 3) on the interval $[-h, h]$. Let $p$ be an arbitrary absolutely continuous prior density on $[-h, h]$ satisfying $p(-h) = p(h) = 0$. For any estimator $\phi(Z_\Omega)$, define $\widehat{\theta} := \phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M}_0)$. Since $\mu(\theta) = \mu_{xy}^{(k)}(\mathcal{M}_0) + \theta$, the risk for estimating the functional is exactly the risk for estimating $\theta$ along the submodel. Writing $J(p) := \int_{-h}^{h} \frac{\{p'(\theta)\}^2}{p(\theta)}\, d\theta$ and choosing $h = \|F_k\|_F \min(\varsigma, \delta_{\gamma,k})$, the van Trees inequality yields

$$
\begin{aligned}
\sup_{|\theta| \leq h} \mathbb{E}_\theta \left[ \{\phi(Z_\Omega) - \mu(\theta)\}^2 \right] &\geq \int_{-h}^{h} \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2]\, p(\theta)\, d\theta \\
&\geq \frac{1}{\int_{-h}^{h} \mathfrak{I}_\theta\, p(\theta)\, d\theta + \inf_{p:\, p(\pm h) = 0} J(p)} = \frac{1}{\int_{-h}^{h} \mathfrak{I}_\theta\, p(\theta)\, d\theta + \pi^2/h^2} \\
&\geq \left\{ \sigma^{-2} c_u \|F_k\|_F^{-2} \left( \frac{N_{1k}}{N} + \frac{T_{1k}}{T} \right) + \pi^2 \|F_k\|_F^{-2} \min^{-2}(\varsigma, \delta_{\gamma,k}) \right\}^{-1} \\
&= \|F_k\|_F^2 \left\{ \sigma^{-2} c_u \left( \frac{N_{1k}}{N} + \frac{T_{1k}}{T} \right) + \pi^2 \min^{-2}(\varsigma, \delta_{\gamma,k}) \right\}^{-1} \\
&\geq \|F_k\|_F^2 \left\{ 2 c_u \sigma^{-2} \max \left( \frac{N_{1k}}{N}, \frac{T_{1k}}{T} \right) + \pi^2 \min^{-2}(\varsigma, \delta_{\gamma,k}) \right\}^{-1} \\
&\geq \frac{1}{2} \min \left\{ \frac{\sigma^2}{2c_u} \min \left( \frac{N}{N_{1k}}, \frac{T}{T_{1k}} \right), \pi^{-2} \min^2(\varsigma, \delta_{\gamma,k}) \right\} \|F_k\|_F^2 \\
&= \min \left\{ \frac{\sigma^2}{4c_u} \min \left( \frac{N}{N_{1k}}, \frac{T}{T_{1k}} \right), \frac{\varsigma^2}{2\pi^2}, \frac{\delta_{\gamma,k}^2}{2\pi^2} \right\} \|F_k\|_F^2 \\
&\geq \min \left\{ \frac{1}{4c_u}, \frac{1}{2\pi^2} \right\} \min \left\{ \sigma^2 \min \left( \frac{N}{N_{1k}}, \frac{T}{T_{1k}} \right), \varsigma^2, \delta_{\gamma,k}^2 \right\} \|F_k\|_F^2 \\
&= \min \left\{ \frac{1}{4c_u}, \frac{1}{2\pi^2} \right\} \min \left\{ \sigma^2 \min \left( \frac{N}{N_{1k}}, \frac{T}{T_{1k}} \right), \varsigma^2, \delta_{\gamma,k}^2 \right\} \|U_{0,2k}^\top x\|_2^2 \|V_{0,2k}^\top y\|_2^2,
\end{aligned}
$$

where the first equality follows from the fact that $J(p)$ is minimised by $p(\theta) = h^{-1} \cos^2(\pi\, \theta / 2h)$, while the successive inequality follows from our previous bound on $\mathfrak{I}_\theta$. Since the path is contained in the local parameter space, this lower bound also applies to the local minimax risk over $\mathcal{F}_{\text{loc}}(\mathcal{M}_0, \varsigma)$. This concludes the proof. □

*Proof of Theorem 3.* We prove the first term in the lower bound by reducing the problem to a one-dimensional parametric submodel where only $V_0$ is perturbed. The second follows by the symmetric argument in which $U_0$ is perturbed instead of $V_0$. To this end, we will assume that $U_{0,2}^\top x \neq 0$ and $\omega_V > 0$; if either of these conditions are not met, the desired lower bound is trivially satisfied. Let $C_{0,j} := (\mathcal{C}_0)_{\bullet,\bullet,j}$ satisfying $\gamma_{\min} \leq \sigma_r(C_{0,j}) \leq$

$\sigma_1(C_{0,j}) \le \gamma_{\max}$ for all $j \in [K]$. Let $G_V := C_{0,k}^\top U_{0,2}^\top x\, y^\top P_{V_{0,2}}^\perp \neq 0$ since $U_{0,2}^\top x \neq 0$, $\sigma_r(C_{0,k}) \ge \gamma_{\min} > 0$, and $\omega_V > 0$. Since $G_V$ is rank one, write $G_V = dab^\top$, where

$$d := \|G_V\|_F = \omega_V \|C_{0,k}^\top U_{0,2}^\top x\|_2, \qquad a := \frac{C_{0,k}^\top U_{0,2}^\top x}{\|C_{0,k}^\top U_{0,2}^\top x\|_2}, \qquad b := \frac{P_{V_{0,2}}^\perp y}{\omega_V}.$$

Then $\|a\|_2 = \|b\|_2 = 1$ and $G_V b = da$. By definition of $b$ we also have $b \in \mathrm{Im}(P_{V_{0,2}}^\perp)$, hence $V_{0,2}^\top b = 0$ and $P_{V_{0,2}}^\perp b = b$. Setting $w := (\mathbf{0}_{T_1}^\top ; b^\top)^\top \in \mathbb{R}^T$, we obtain a vector supported only on the last $T_2$ entries that is orthogonal to $V_0$, meaning that $V_0^\top w = 0$.

We now introduce a one-dimensional submodel by perturbing $V_0$ only. For $\theta \in \mathbb{R}$, define $V(\theta) := (V_0 + \theta w a^\top)(I_r + \theta^2 a a^\top)^{-1/2} = (V_0 + \theta w a^\top)(I_r + [\{1+\theta^2\}^{-1/2} - 1]\, aa^\top)$. We have $V(\theta)^\top V(\theta) = I_r$ and

$$V'(\theta) = (1+\theta^2)^{-3/2}(w - \theta\, V_0 a)\, a^\top.$$

We leave $U_0$ and $\mathcal{C}_0$ untouched, and set $\mathcal{M}(\theta) := \mathcal{C}_0 \times_1 U_0 \times_2 V(\theta) \times_3 I_K$. We next verify the range of $\theta$ for which the path $\{\mathcal{M}(\theta) : |\theta| \le h\}$ remains in the local parameter space. As $\mathcal{C}_0$ and $U_0$ are fixed, the only conditions to check are the Frobenius-norm bound, and assumption (A1) for $V(\theta)$. As for the former, since $U_0^\top U_0 = I_r, V_0^\top w = 0, w^\top w = 1, a^\top a = 1$, we have

$$\begin{aligned}\|\mathcal{M}(\theta) - \mathcal{M}_0\|_F^2 &= \sum_{j=1}^K \left\|U_0\, C_{0,j}\, \{V(\theta) - V_0\}^\top\right\|_F^2 \le \gamma_{\max}^2 \sum_{j=1}^K \|V(\theta) - V_0\|_F^2 \\ &= \gamma_{\max}^2\, K\, \|V(\theta) - V_0\|_F^2 = 2\,\gamma_{\max}^2\, K \left\{1 - \left(1+\theta^2\right)^{-1/2}\right\} \le \gamma_{\max}^2\, K\, \theta^2 \le \gamma_{\max}^2\, K\, h^2,\end{aligned}$$

where in the penultimate inequality we used the standard bound $2\left\{1 - (1+x)^{-1/2}\right\} \le x$ for $x \ge 0$. This is upper bounded by $\varsigma^2$ for $h \le \varsigma\, \gamma_{\max}^{-1} K^{-1/2}$. As for (A1), start by observing that $(I_r + \theta^2 aa^\top)^{-1/2}$ is symmetric and has eigenvalues bounded between $(1+\theta^2)^{-1/2}$ and 1. As a result, for $V_1(\theta) := V(\theta)_{[T_1],\bullet} = V_{0,1}\,(I_r + \theta^2 aa^\top)^{-1/2}$ we have

$$\begin{aligned}V_1(\theta)^\top V_1(\theta) &= \{I_r + \theta^2 aa^\top\}^{-1/2}\, V_{0,1}^\top V_{0,1}\, \{I_r + \theta^2 aa^\top\}^{-1/2} \\ &\preceq (c_u - \delta_{\mathrm{A1}})\, \frac{T_1}{T}\, \{I_r + \theta^2 aa^\top\}^{-1} \preceq (c_u - \delta_{\mathrm{A1}})\, \frac{T_1}{T} I_r \preceq c_u\, \frac{T_1}{T} I_r.\end{aligned}$$

Similarly,

$$V_1(\theta)^\top V_1(\theta) \succeq (c_\ell + \delta_{\mathrm{A1}})\, \frac{T_1}{T}\, \{I_r + \theta^2 aa^\top\}^{-1} \succeq \frac{c_\ell + \delta_{\mathrm{A1}}}{1+\theta^2}\, \frac{T_1}{T} I_r,$$

which is lower bounded by $c_\ell T_1 T^{-1} I_r$ whenever $h^2 \le \delta_{\mathrm{A1}}\, c_\ell^{-1}$. We thus deduce that the path $\{\mathcal{M}(\theta) : |\theta| \le h\}$ is contained in $\{\mathcal{M} \in \mathcal{F}(c_\ell, c_u) : \|\mathcal{M} - \mathcal{M}_0\|_F \le \varsigma\}$ whenever $h \le \min(\varsigma\, \gamma_{\max}^{-1} K^{-1/2}, \delta_{\mathrm{A1}}^{1/2}\, c_\ell^{-1/2})$. In order to simplify the computations below while remaining in the Frobenius neighbourhood of $\mathcal{M}_0$, we will set $h = \min(\omega_V/2,\ \sqrt{N_1/N},\ \varsigma\, \gamma_{\max}^{-1} K^{-1/2},\ \delta_{\mathrm{A1}}^{1/2}\, c_\ell^{-1/2}) =: \varepsilon_V$.

We next compute the induced change in $\mu(\theta) := \mu_{xy}^{(k)}(\mathcal{M}(\theta)) = x^\top U_{0,2}\, C_{0,k} V_2(\theta)^\top y$. Using $G_V = C_{0,k}^\top U_{0,2}^\top x\, y^\top P_{V_{0,2}}^\perp = dab^\top$, $P_{V_{0,2}}^\perp b = b$, $(x^\top U_{0,2} C_{0,k} a)(b^\top y) = d$, and the expression for $V'(\theta)$ derived above,

we have

$$\begin{aligned}
\mu'(\theta) &= x^\top U_{0,2}\, C_{0,k} V_2'(\theta)^\top y \\
&= (1+\theta^2)^{-3/2} x^\top U_{0,2} C_{0,k}\, a\, (b - \theta V_{0,2} a)^\top\, y \\
&= d(1+\theta^2)^{-3/2} \left\{1 - \theta d^{-1} a^\top (C_{0,k}^\top U_{0,2}^\top x\, y^\top V_{0,2})\, a\right\} \\
&\geq d(1+\theta^2)^{-3/2} \left\{1 - |\theta|\, d^{-1} |a^\top (C_{0,k}^\top U_{0,2}^\top x\, y^\top V_{0,2})\, a|\right\} \\
&\geq d\,(1+h^2)^{-3/2} (1 - h d^{-1} \left\|C_{0,k}^\top U_{0,2}^\top x\, y^\top V_{0,2}\right\|_{\mathrm{op}}) \\
&\geq d\,(1+h^2)^{-3/2} \left(1 - h\omega_V^{-1}\right) \geq 2^{-5/2}\, d \\
&= 2^{-5/2}\, \omega_V \left\|C_{0,k}^\top U_{0,2}^\top x\right\|_2 \geq 2^{-5/2} \gamma_{\min}\, \omega_V \|U_{0,2}^\top x\|_2 > 0,
\end{aligned}$$

where in the penultimate inequality we used $h \leq \min(\sqrt{N_1/N},\, \omega_V/2) \leq \min(1, \omega_V/2)$. Coming now to the Fisher information for this model, it is useful to compute $\|V_2'(\theta)\|_F^2 \leq \|V'(\theta)\|_F^2 = \left(1+\theta^2\right)^{-2} \leq 1$. Similarly, we can show that $\|V_1'(\theta)\|_F^2 = \theta^2(1+\theta^2)^{-3}\|V_{0,1} a\|_2^2 \leq \theta^2 (c_u - \delta_{\mathrm{A1}})\, \frac{T_1}{T} (1+\theta^2)^{-3}$. We thus have

$$\begin{aligned}
\mathfrak{I}_\theta &= \sigma^{-2} \sum_{j=1}^K \|P_{\Omega_{\bullet,\bullet,j}}(U_0\, C_{0,j}[V'(\theta)]^\top)\|_F^2 = \sigma^{-2} \sum_{j=1}^K \left\{\|V_1'(\theta) C_{0,j}^\top U_0^\top\|_F^2 + \|V_2'(\theta) C_{0,j}^\top U_{0,1}^\top\|_F^2\right\} \\
&\leq \gamma_{\max}^2\, \sigma^{-2}\, K \left\{\|V_1'(\theta)\|_F^2 + (c_u - \delta_{\mathrm{A1}}) \frac{N_1}{N} \|V_2'(\theta)\|_F^2\right\} \leq \gamma_{\max}^2\, \sigma^{-2}\, K \left\{\|V_1'(\theta)\|_F^2 + (c_u - \delta_{\mathrm{A1}}) \frac{N_1}{N}\right\} \\
&\leq \gamma_{\max}^2\, \sigma^{-2}\, K (c_u - \delta_{\mathrm{A1}}) \left\{\theta^2 (1+\theta^2)^{-3} \frac{T_1}{T} + \frac{N_1}{N}\right\} \\
&\leq \gamma_{\max}^2\, \sigma^{-2}\, K\, (c_u - \delta_{\mathrm{A1}}) \left\{h^2 + \frac{N_1}{N}\right\} \leq 2\, \gamma_{\max}^2\, \sigma^{-2}\, (c_u - \delta_{\mathrm{A1}}) \frac{K N_1}{N},
\end{aligned}$$

where in the last inequality we used $h^2 \leq N_1/N$.

We now apply the van Trees inequality (Gill and Levit, 1995, Equation 4) on the interval $[-h, h]$ with $h = \varepsilon_V$. Let $p$ be an arbitrary absolutely continuous prior density on $[-h, h]$ satisfying $p(-h) = p(h) = 0$. Writing $J(p) := \int_{-h}^h \frac{\{p'(\theta)\}^2}{p(\theta)}\, d\theta$, and setting $c_V := 2^{-6} \gamma_{\min}^2 \min\left\{2^{-1} \gamma_{\max}^{-2}\, c_u^{-1}, \pi^{-2}\right\}$, the van Trees inequality yields

$$\begin{aligned}
\sup_{|\theta| \leq h} \mathbb{E}_\theta \left[\{\phi(Z_\Omega) - \mu(\theta)\}^2\right] &\geq \int_{-h}^h \mathbb{E}_\theta \left[\{\widehat{\mu} - \mu(\theta)\}^2\right] p(\theta)\, d\theta \geq \frac{\left\{\int_{-h}^h \mu'(\theta)\, p(\theta)\, d\theta\right\}^2}{\int_{-h}^h \mathfrak{I}_\theta\, p(\theta)\, d\theta + \inf_{p:\, p(\pm h) = 0} J(p)} \\
&= \frac{\left\{\int_{-h}^h \mu'(\theta) p(\theta)\, d\theta\right\}^2}{\int_{-h}^h \mathfrak{I}_\theta\, p(\theta)\, d\theta + \pi^2/h^2} \geq \frac{2^{-5} \gamma_{\min}^2 \omega_V^2 \|U_{0,2}^\top x\|_2^2}{2\gamma_{\max}^2 \sigma^{-2} (c_u - \delta_{\mathrm{A1}}) K N_1/N + \pi^2/h^2} \\
&= \frac{2^{-5} \gamma_{\min}^2 \omega_V^2 \|U_{0,2}^\top x\|_2^2}{2\gamma_{\max}^2 (c_u - \delta_{\mathrm{A1}}) \sigma^{-2} K N_1/N + \pi^2 h^{-2}} \\
&\geq \frac{2^{-5} \gamma_{\min}^2 \omega_V^2 \|U_{0,2}^\top x\|_2^2}{2 \max\left(2\gamma_{\max}^2 (c_u - \delta_{\mathrm{A1}}) \sigma^{-2} K N_1/N, \pi^2 h^{-2}\right)} \\
&= 2^{-6} \gamma_{\min}^2 \omega_V^2 \|U_{0,2}^\top x\|_2^2\, \min\left(\frac{\sigma^2 N}{2\gamma_{\max}^2 (c_u - \delta_{\mathrm{A1}}) K N_1}, \frac{h^2}{\pi^2}\right)
\end{aligned}$$

$$\geq 2^{-6}\gamma_{\min}^2 \min\left\{\frac{1}{2\gamma_{\max}^2(c_u-\delta_{\mathrm{A1}})}, \frac{1}{\pi^2}\right\}\omega_V^2 \min\left(\frac{\sigma^2 N}{KN_1}, h^2\right)\|U_{0,2}^\top x\|_2^2$$
$$\geq c_V\,\omega_V^2 \min\left(\frac{\sigma^2 N}{KN_1}, \varepsilon_V^2\right)\|U_{0,2}^\top x\|_2^2,$$

where the first equality follows from the fact that $J(p)$ is minimised by $p(\theta) = h^{-1}\cos^2(\pi\,\theta/2h)$, while the successive inequalities follow from our previous bounds on $\mu'(\theta)$, and $\mathfrak{I}_\theta$. Since the path is contained in the local parameter space, this lower bound also applies to the local minimax risk over $\mathcal{F}_{\mathrm{loc}}(\mathcal{M}_0, \varsigma)$.

Interchanging the roles of $U_0$ and $V_0$ and applying the same argument to $P_{U_{0,2}}^\perp x\, y^\top V_{0,2} C_{0,k}^\top$, yields the second term in the lower bound. In particular, we will assume that $V_{0,2}^\top y \neq 0$ and $\omega_U > 0$; if either of these conditions are not met, the desired lower bound is trivially satisfied. We then define $G_U := P_{U_{0,2}}^\perp x\, y^\top V_{0,2} C_{0,k}^\top$. Since $y^\top V_{0,2} C_{0,k}^\top \neq 0$ and $\omega_U = \|P_{U_{0,2}}^\perp x\|_2 > 0$, this matrix is rank one. Writing $G_U = d_U a_U b_U^\top$, where

$$d_U = \omega_U \|C_{0,k} V_{0,2}^\top y\|_2, \qquad a_U = \frac{P_{U_{0,2}}^\perp x}{\omega_U}, \qquad b_U = \frac{C_{0,k} V_{0,2}^\top y}{\|C_{0,k} V_{0,2}^\top y\|_2},$$

and perturbing $U_0$ along $(\mathbf{0}_{N_1}^\top\,;\, a_U^\top)^\top$ gives an analogous one-dimensional path $U(\theta)$ with $V_0$ and $\mathcal{C}_0$ fixed. The same calculations, with $N_1/N$ and $T_1/T$ interchanged, yield

$$\inf_{\phi}\sup_{\mathcal{M}\in\mathcal{F}_{\mathrm{loc}}(\mathcal{M}_0,\varsigma)} \mathbb{E}_{\mathcal{M}}\left[\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M})\}^2\right] \geq c_V\,\omega_U^2 \min\left(\frac{\sigma^2 T}{KT_1}, \varepsilon_U^2\right)\|V_{0,2}^\top y\|_2^2,$$

where $\varepsilon_U := \min(\omega_U/2,\, \sqrt{T_1/T},\, \varsigma\gamma_{\max}^{-1}K^{-1/2},\, \delta_{\mathrm{A1}}^{1/2} c_\ell^{-1/2})$. Combining this with the lower bound obtained from the $V_0$-perturbation, and using that the maximum of two lower bounds is at least their average, gives

$$\inf_{\phi}\sup_{\mathcal{M}\in\mathcal{F}_{\mathrm{loc}}(\mathcal{M}_0,\varsigma)} \mathbb{E}_{\mathcal{M}}\left[\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M})\}^2\right] \geq \frac{c_V}{2}\,\omega_V^2 \min\left(\frac{\sigma^2 N}{KN_1}, \varepsilon_V^2\right)\|U_{0,2}^\top x\|_2^2$$
$$+ \frac{c_V}{2}\,\omega_U^2 \min\left(\frac{\sigma^2 T}{KT_1}, \varepsilon_U^2\right)\|V_{0,2}^\top y\|_2^2.$$

Setting $c = c_V/2$ yields the desired bound. □

## A.3 Proofs for Section 5

*Proof of Corollary 4.* We verify that the auxiliary problem obtained by restricting to $S \times Q$ satisfies the hypotheses of Theorem 1. Although, for $j \neq k$, the missingness masks $\Omega_{S,Q,j}$ are not necessarily in four-block form, this is not an essential requirement. What is needed in order to apply Theorem 1, and in particular Lemma 14 in Appendix C, is the relevant subblock conditioning assumption for the rows and columns corresponding to fully observed rows and columns, respectively. These are exactly the submatrices that enter the definitions of the upper and left pooled matrices, and they are what enable improved estimation of the shared subspaces.

We start by verifying the analogue of (A1). Let $G_U := U_S^\top U_S$ and $G_V := V_Q^\top V_Q$. By Assumption (A5) we have

$$c_\ell \frac{\mathfrak{n}}{N} I_r \preceq G_U \preceq c_u \frac{\mathfrak{n}}{N} I_r, \qquad c_\ell \frac{\mathfrak{t}}{T} I_r \preceq G_V \preceq c_u \frac{\mathfrak{t}}{T} I_r,$$

hence $G_U$ and $G_V$ are nonsingular. Letting $\widetilde{U} := U_S G_U^{-1/2}$, $\widetilde{V} := V_Q G_V^{-1/2}$ and $\widetilde{\mathcal{C}}_{\bullet,\bullet,j} := G_U^{1/2} \mathcal{C}_{\bullet,\bullet,j} G_V^{1/2}$, we have that $\widetilde{U}^\top \widetilde{U} = I_r$, $\widetilde{V}^\top \widetilde{V} = I_r$ and $\mathcal{M}_{S,Q,j} = U_S \mathcal{C}_{\bullet,\bullet,j} V_Q^\top = \widetilde{U} \widetilde{\mathcal{C}}_{\bullet,\bullet,j} \widetilde{V}^\top$. This shows that the restriction of the signal to $S \times Q$ admits an orthonormal Tucker2 representation with row and column dimensions $\mathfrak{n}$ and $\mathfrak{t}$, respectively.

We next check that the restricted Gram matrices are well conditioned. For the target layer, we have $\widetilde{U}_{S^+}^\top \widetilde{U}_{S^+} = G_U^{-1/2} U_{S^+}^\top U_{S^+} G_U^{-1/2}$, hence, using (A5) and the preceding bounds on $G_U$ gives

$$\frac{c_\ell}{c_u} \frac{\mathfrak{n}_{1k}}{\mathfrak{n}} I_r \preceq \widetilde{U}_{S^+}^\top \widetilde{U}_{S^+} \preceq \frac{c_u}{c_\ell} \frac{\mathfrak{n}_{1k}}{\mathfrak{n}} I_r.$$

Similarly, for each $j \neq k$ we have $\widetilde{U}_{\mathsf{RowAnc}_k(j)}^\top \widetilde{U}_{\mathsf{RowAnc}_k(j)} = G_U^{-1/2} U_{\mathsf{RowAnc}_k(j)}^\top U_{\mathsf{RowAnc}_k(j)} G_U^{-1/2}$, thus

$$\frac{c_\ell}{c_u} \frac{\mathfrak{n}_{1j}}{\mathfrak{n}} I_r \preceq \widetilde{U}_{\mathsf{RowAnc}_k(j)}^\top \widetilde{U}_{\mathsf{RowAnc}_k(j)} \preceq \frac{c_u}{c_\ell} \frac{\mathfrak{n}_{1j}}{\mathfrak{n}} I_r.$$

Since the same argument applies to the column factors, we can conclude that the auxiliary sub-block conditioning assumption holds with $c_\ell/c_u, c_u/c_\ell$ in place of $c_\ell, c_u$, respectively.

It remains to identify the signal strengths in the auxiliary model. Since $\widetilde{\mathcal{C}}_{\bullet,\bullet,j} = G_U^{1/2} \mathcal{C}_{\bullet,\bullet,j} G_V^{1/2}$, standard bounds on the singular values of a matrix product yields

$$\widetilde{\gamma}_{\min} := c_\ell \, \gamma_{\min} \sqrt{\frac{\mathfrak{n}\mathfrak{t}}{NT}} \leq \sigma_{\min}(\widetilde{C}_j) \leq \sigma_{\max}(\widetilde{C}_j) \leq c_u \gamma_{\max} \sqrt{\frac{\mathfrak{n}\mathfrak{t}}{NT}} =: \widetilde{\gamma}_{\max},$$

thereby showing that the effective lower and upper signals are $\widetilde{\gamma}_{\min}$ and $\widetilde{\gamma}_{\max}$.

The noise distribution is unchanged by restriction since, on the observed auxiliary coordinates, the errors are still independent centred Gaussian variables with variance $\sigma^2$. Moreover, Assumption (A6) is the analogue of (A2), (A3) and (A4) after replacing $N, T, N_{1k}, T_{1k}, \rho_N, \rho_T, p_N, p_T, \zeta_N, \zeta_T, \gamma_{\min}, \gamma_{\max}$ by $\mathfrak{n}, \mathfrak{t}, \mathfrak{n}_{1k}, \mathfrak{t}_{1k}, \rho_{\mathfrak{n}}, \rho_{\mathfrak{t}}, p_{\mathfrak{n}}, p_{\mathfrak{t}}, \zeta_{\mathfrak{n}}, \zeta_{\mathfrak{t}}, \widetilde{\gamma}_{\min}, \widetilde{\gamma}_{\max}$. Also, (6) is the analogue of (4), again with the same substitution of auxiliary dimensions and signal strengths. We can thus conclude that all hypotheses of Theorem 1 hold for the auxiliary problem, hence applying this result gives $\mathbb{E}_{\mathcal{M}}[\{\widehat{\mu}_{xy}^{(k,a,b)} - \mu_{xy}^{(k,a,b)}\}^2] \leq c_1 \widetilde{\Upsilon}_{xy}$. This concludes the proof. □

# B Additional details on the simulation studies

## B.1 Target estimands used in the empirical applications

We provide more details on the target estimands used in Sections 6.2 and 6.3. In these applications, we work with two signal tensors, $\mathcal{M}(0)$ and $\mathcal{M}(1)$, corresponding to the untreated and treated responses, respectively. The staggered-adoption mask $\Omega$ is constructed from the treatment variable, with $\Omega_{itj} = 1$ if entry $(i,t,j) \in [N] \times [T] \times [K]$ lies in the untreated region, and $\Omega_{itj} = 0$ otherwise. The fully observed tensor $\mathcal{Y}$ satisfies $\mathcal{Y}_{itj} = \Omega_{itj} \, \mathcal{Y}_{itj}(0) + (1 - \Omega_{itj}) \, \mathcal{Y}_{itj}(1)$, where $\mathcal{Y}(0)$ denotes the untreated potential outcome, which is observed on $\{(i,t,j) : \Omega_{itj} = 1\}$ and missing on the complementary set, and $\mathcal{Y}(1)$ denotes the treated potential outcome, which is observed only over $\{(i,t,j) : \Omega_{itj} = 0\}$.

We now introduce the four functionals used in the empirical applications: ATE, RowHet, Local-$i_0$,

and TREND. Fix a target slice $k$ with staircase adoption, and let $\mathcal{D}_k = \{(a,b) : a+b > o_k+1\}$ for some integer $o_k \geq 2$ be the collection of policy-on target blocks. For all $(a,b) \in \mathcal{D}_k$ and $c \in \{0,1\}$, we also denote with $\mathcal{M}^{(a,b)}_{\bullet,\bullet,k}(c)$ the restriction of $\mathcal{M}(c)$ to rows $R_{ak}$ and columns $C_{bk}$. For $i \in R_{ak}$ and $t \in C_{bk}$, define the local-index maps $\ell_a(i) := \mathrm{pos}_{R_{ak}}(i)$ and $\ell_b(t) := \mathrm{pos}_{C_{bk}}(t)$, so that $\ell_a(i)$ is the position of row $i$ within $R_{ak}$, and $\ell_b(t)$ is the position of column $t$ within $C_{bk}$. For the ROWHET functional, choose a sign vector $\eta = (\eta_1, \ldots, \eta_N)^\top \in \{\pm 1\}^N$. For the LOCAL-$i_0$ functional, choose a row-block index $a_0$ such that $\{b : (a_0, b) \in \mathcal{D}_k\} \neq \varnothing$, and then fix a row index $i_0 \in R_{a_0 k}$. We also define $\mathcal{D}^{\mathrm{tr}}_k := \{(a,b) \in \mathcal{D}_k : T_{bk} \geq 2\}$, and assume $\mathcal{D}^{\mathrm{tr}}_k \neq \varnothing$ whenever the TREND functional is considered. For $h \in \{\mathrm{ATE}, \mathrm{RowHet}, \mathrm{Local}\text{-}i_0, \mathrm{Trend}\}$, we write

$$\mathcal{D}_{k,h} := \begin{cases} \mathcal{D}_k, & h \in \{\mathrm{ATE}, \mathrm{RowHet}\}, \\ \{(a_0, b) : (a_0, b) \in \mathcal{D}_k\}, & h = \mathrm{Local}\text{-}i_0, \\ \mathcal{D}^{\mathrm{tr}}_k, & h = \mathrm{Trend} \end{cases}$$

for the active block set, and, for fixed $(a,b) \in \mathcal{D}_{k,h}$, we consider the query directions

$$\begin{aligned} \mathrm{ATE}: &\quad x_{a,h} = N_{ak}^{-1/2} \mathbf{1}_{N_{ak}}, &\quad y_{b,h} = T_{bk}^{-1/2} \mathbf{1}_{T_{bk}}, \\ \mathrm{RowHet}: &\quad x_{a,h} = N_{ak}^{-1/2} \eta_{R_{ak}}, &\quad y_{b,h} = T_{bk}^{-1/2} \mathbf{1}_{T_{bk}}, \\ \mathrm{Local}\text{-}i_0: &\quad x_{a,h} = \boldsymbol{e}_{\ell_a(i_0)}, &\quad y_{b,h} = T_{bk}^{-1/2} \mathbf{1}_{T_{bk}}, \\ \mathrm{Trend}: &\quad x_{a,h} = N_{ak}^{-1/2} \mathbf{1}_{N_{ak}}, &\quad y_{b,h} = \frac{z_b - \bar{z}_b \mathbf{1}_{T_{bk}}}{\|z_b - \bar{z}_b \mathbf{1}_{T_{bk}}\|_2}, \end{aligned}$$

with $z_b = (1, \ldots, T_{bk})^\top$ and $\bar{z}_b = (T_{bk}+1)/2$. Based on this, for $h \in \{\mathrm{ATE}, \mathrm{RowHet}, \mathrm{Local}\text{-}i_0, \mathrm{Trend}\}$ and $(a,b) \in \mathcal{D}_{k,h}$, we define the block-level bilinear forms

$$\mu_h^{(k,a,b)}(c) := x_{a,h}^\top \mathcal{M}^{(a,b)}_{\bullet,\bullet,k}(c)\, y_{b,h}.$$

As for the interpretation of these functionals, ATE averages all entries in the block, ROWHET averages a signed row contrast over the block's columns, LOCAL-$i_0$ averages only row $i_0$ over the block's columns, while TREND averages over rows and contrasts later columns with earlier columns. In particular, this latter bilinear form also has a simple slope interpretation. For $t \in C_{bk}$ and $c \in \{0,1\}$, define the row-averaged trajectory $\bar{m}_t^{(a,b)}(c) := N_{ak}^{-1} \sum_{i \in R_{ak}} \mathcal{M}^{(a,b)}_{i,t,k}(c)$. If this trajectory is linear in local time, i.e. $\bar{m}_t^{(a,b)}(c) = \alpha^c_{a,b} + \beta^c_{a,b}\, \ell_b(t)$, then

$$x_{a,\mathrm{Trend}}^\top \mathcal{M}^{(a,b)}_{\bullet,\bullet,k}(c)\, y_{b,\mathrm{Trend}} = \sqrt{N_{ak}}\, \beta^c_{a,b} \left\{ \frac{T_{bk}(T_{bk}^2 - 1)}{12} \right\}^{1/2},$$

thus showing that TREND recovers the slope of the row-averaged trajectory up to a known normalisation.

We then aggregate these block-level summaries over all missing blocks in the target slice by

$$\Psi_c^{(h)}(k) := \{W_h(k)\}^{-1} \sum_{(a,b) \in \mathcal{D}_{k,h}} c_{ab}^{(h)} \mu_h^{(k,a,b)}(c),$$

where the choices of weights and normalising constants, together with the simplified form of each estimand,

are given below:

| $h$ | $c_{ab}^{(h)}$ | $W_h(k)$ | $\Psi_c^{(h)}(k)$ |
|---|---|---|---|
| ATE | $\sqrt{N_{ak}T_{bk}}$ | $\sum_{(a,b)\in\mathcal{D}_k} N_{ak}T_{bk}$ | $\frac{\sum_{(a,b)\in\mathcal{D}_k}\sum_{i\in R_{ak}}\sum_{t\in C_{bk}}\mathcal{M}_{i,t,k}^{(a,b)}(c)}{\sum_{(a,b)\in\mathcal{D}_k} N_{ak}T_{bk}}$ |
| RowHet | $\sqrt{N_{ak}T_{bk}}$ | $\sum_{(a,b)\in\mathcal{D}_k} N_{ak}T_{bk}$ | $\frac{\sum_{(a,b)\in\mathcal{D}_k}\sum_{i\in R_{ak}}\sum_{t\in C_{bk}}\eta_i\,\mathcal{M}_{i,t,k}^{(a,b)}(c)}{\sum_{(a,b)\in\mathcal{D}_k} N_{ak}T_{bk}}$ |
| Local-$i_0$ | $\sqrt{T_{bk}}$ | $\sum_{b:\,(a_0,b)\in\mathcal{D}_k} T_{bk}$ | $\frac{\sum_{b:\,(a_0,b)\in\mathcal{D}_k}\sum_{t\in C_{bk}}\mathcal{M}_{i_0,t,k}^{(a_0,b)}(c)}{\sum_{b:\,(a_0,b)\in\mathcal{D}_k} T_{bk}}$ |
| Trend | $\frac{1}{\sqrt{N_{ak}T_{bk}(T_{bk}^2-1)/12}}$ | $\lvert\mathcal{D}_k^{\mathrm{tr}}\rvert$ | $\frac{1}{\lvert\mathcal{D}_k^{\mathrm{tr}}\rvert}\sum_{(a,b)\in\mathcal{D}_k^{\mathrm{tr}}}\beta_{a,b}^{c}.$ |

(7)

The final expression for $\Psi_c^{(\mathrm{Trend})}(k)$ uses the linearity condition on the row-averaged trajectory and the assumption $\mathcal{D}_k^{\mathrm{tr}} \neq \varnothing$. These four quantities have the following interpretations: $\Psi_c^{(\mathrm{ATE})}(k)$ is the average potential outcome over the missing entries in slice $k$, $\Psi_c^{(\mathrm{RowHet})}(k)$ is the corresponding signed row contrast, $\Psi_c^{(\mathrm{Local}\text{-}i_0)}(k)$ is the average potential outcome for row $i_0$ over the missing target blocks containing that row, and $\Psi_c^{(\mathrm{Trend})}(k)$ is the average within-block slope of the row-averaged trajectory.

Based on these potential-outcome summaries, we also define the aggregate policy effect for functional $h$ by $\Delta^{(h)}(k) := \Psi_1^{(h)}(k) - \Psi_0^{(h)}(k)$.

Coming now to the estimation of these aggregate quantities, it is useful to recall that $\mathcal{Y}_{\bullet,\bullet,k}(1)$ is observed over $\mathcal{D}_k$, since these are exactly the policy-on target blocks in which $\mathcal{Y}_{\bullet,\bullet,k}(0)$ has missing entries. As a result, functionals with $c = 1$ are easier to target and can be estimated by simple plug-in estimators. On the other hand, quantities such as $\mu_h^{(k,a,b)}(0)$ require an alternative approach, and can be estimated using Algorithm 2. This immediately leads to a naive estimator of $\Psi_0^{(h)}(k)$ that applies Algorithm 2 separately to each missing target block, and then aggregates the resulting block-level estimates using the weights $c_{ab}^{(h)}$ and normalising constants $W_h(k)$ defined in (7).

**Algorithm 3** QuadraticStaggeredAggregate for estimating $\Psi_0^{(h)}(k)$

**Require:** target slice $k \in [K]$, functional $h \in \{\mathrm{ATE}, \mathrm{RowHet}, \mathrm{Local}\text{-}i_0, \mathrm{Trend}\}$, rank $r$, data $\mathcal{Y}$, staircase partitions $\{R_{ak}\}_{a=1}^{o_k}$ and $\{C_{bk}\}_{b=1}^{o_k}$, parameter $\tau > 0$, and, when needed, sign vector $\eta$ and row index $i_0$.

1: Initialize $S_h \leftarrow 0$.
2: **for** $(a, b) \in \mathcal{D}_{k,h}$ **do**
3: Construct $x_{a,h}$ and $y_{b,h}$ according to the definitions above.
4: Run Algorithm 2 with inputs $(k, a, b, r, x_{a,h}, y_{b,h}, \mathcal{Y}, \tau)$, and denote its output by $\widehat{\mu}_h^{(k,a,b)}(0)$.
5: Update

$$S_h \leftarrow S_h + c_{ab}^{(h)}\widehat{\mu}_h^{(k,a,b)}(0).$$

6: **end for**
7: **return** $\widehat{\Psi}_0^{(h)}(k) \leftarrow \{W_h(k)\}^{-1}S_h$.

The blockwise plug-in estimator in Algorithm 3 applies Algorithm 2 separately to every target block

$(a,b) \in \mathcal{D}_{k,h} \subseteq \mathcal{D}_k$, hence it recomputes two rank-$r$ singular value decompositions for each missing block. In the ATE and RowHet cases, we have $|\mathcal{D}_{k,h}| = |\mathcal{D}_k| = o_k(o_k-1)/2$, so the cost is quadratic in $o_k$.

This cost can be reduced by trading some statistical efficiency for computational savings through a reduced-anchor construction. In particular, for fixed $a$, we keep the target-slice column anchor $\mathsf{ColAnc}_{k,a,b}(k) = Q^+_{k,a,b}$ but replace $\mathsf{ColAnc}_{k,a,b}(j)$ by $\mathsf{ColAnc}_{k,a,b}(j) \cap Q_{k,1}$ for each $j \neq k$. Because $Q^+_{k,a,b}$ depends only on $a$, and because $Q_{k,1} \subseteq Q_{k,b}$ for all $b$, the resulting pooled left matrix depends only on $a$. Similarly, for fixed $b$, we keep the target-slice row anchor $\mathsf{RowAnc}_{k,a,b}(k) = S^+_{k,a,b}$ but replace $\mathsf{RowAnc}_{k,a,b}(j)$ by $\mathsf{RowAnc}_{k,a,b}(j) \cap S_{k,1}$ for each $j \neq k$. Because $S^+_{k,a,b}$ depends only on $b$, and because $S_{k,1} \subseteq S_{k,a}$ for all $a$, the resulting pooled upper matrix depends only on $b$.

With this reduced-anchor construction, the SVDs of the the pooled left and upper matrices can be cached and reused, as illustrated in the following algorithm. We will use the shorthand $Q^+_{k,a,b} \equiv Q^+_{k,a}$ and $S^+_{k,a,b} \equiv S^+_{k,b}$ to emphasise that these sets depend only on $a$ and $b$, respectively.

---

**Algorithm 4** LinearStaggeredAggregate for estimating $\Psi_0^{(h)}(k)$

---

**Require:** target slice $k \in [K]$, functional $h \in \{\mathrm{ATE}, \textsc{RowHet}, \textsc{Local}\text{-}i_0, \textsc{Trend}\}$, rank $r$, data $\mathcal{Y}$, staircase partitions $\{R_{ak}\}_{a=1}^{o_k}$ and $\{C_{bk}\}_{b=1}^{o_k}$, parameter $\tau > 0$, and, when needed, sign vector $\eta$ and row index $i_0$.

Let $\mathcal{A}_h := \{a : \exists b \text{ such that } (a,b) \in \mathcal{D}_{k,h}\}$ and $\mathcal{B}_h := \{b : \exists a \text{ such that } (a,b) \in \mathcal{D}_{k,h}\}$.
**for** $a \in \mathcal{A}_h$ **do**
  Set $\mathsf{ColAnc}^{\mathrm{red}}_{k,a}(k) := Q^+_{k,a}$ and, for $j \neq k$, $\mathsf{ColAnc}^{\mathrm{red}}_{k,a}(j) := \{t \in Q_{k,1} : \Omega_{i,t,j} = 1 \text{ for all } i \in S_{k,a}\}$.
  Form reduced pooled left matrix $Y^{\mathrm{red}}_{\mathrm{left},a} \leftarrow (\mathcal{Y}_{S_{k,a},\mathsf{ColAnc}^{\mathrm{red}}_{k,a}(1),1} \ \cdots \ \mathcal{Y}_{S_{k,a},\mathsf{ColAnc}^{\mathrm{red}}_{k,a}(K),K})$.
  Compute rank-$r$ truncated singular value decomposition $(\widehat{U}^{\mathrm{red}}_{\mathrm{left},a}, \widehat{\Sigma}^{\mathrm{red}}_{\mathrm{left},a}, \widehat{V}^{\mathrm{red}}_{\mathrm{left},a}) \leftarrow \mathrm{SVD}_r(Y^{\mathrm{red}}_{\mathrm{left},a})$.
  Construct $x_{a,h}$, and cache $\widehat{\alpha}^{(k,a,h)}_{\mathrm{red}} \leftarrow (\widehat{U}^{\mathrm{red}}_{\mathrm{left},a})^\top_{R_{ak},\bullet}\, x_{a,h} \in \mathbb{R}^r$.
**end for**
**for** $b \in \mathcal{B}_h$ **do**
  Set $\mathsf{RowAnc}^{\mathrm{red}}_{k,b}(k) := S^+_{k,b}$ and, for $j \neq k$, $\mathsf{RowAnc}^{\mathrm{red}}_{k,b}(j) := \{i \in S_{k,1} : \Omega_{i,t,j} = 1 \text{ for all } t \in Q_{k,b}\}$.
  Form reduced pooled upper matrix $Y^{\mathrm{red}}_{\mathrm{up},b} \leftarrow (\mathcal{Y}_{\mathsf{RowAnc}^{\mathrm{red}}_{k,b}(1),Q_{k,b},1} \ ; \ \cdots \ ; \ \mathcal{Y}_{\mathsf{RowAnc}^{\mathrm{red}}_{k,b}(K),Q_{k,b},K})$.
  Compute rank-$r$ truncated singular value decomposition $(\widehat{U}^{\mathrm{red}}_{\mathrm{up},b}, \widehat{\Sigma}^{\mathrm{red}}_{\mathrm{up},b}, \widehat{V}^{\mathrm{red}}_{\mathrm{up},b}) \leftarrow \mathrm{SVD}_r(Y^{\mathrm{red}}_{\mathrm{up},b})$.
  Set $s_{k,b} := \sum_{j=1}^{k-1} |\mathsf{RowAnc}^{\mathrm{red}}_{k,b}(j)|$, $\widehat{U}^{(k)}_{\mathrm{up},b} \leftarrow (\widehat{U}^{\mathrm{red}}_{\mathrm{up},b})_{\{s_{k,b}+1,\ldots,s_{k,b}+|S^+_{k,b}|\},\bullet}$, $\widehat{V}_{bk} \leftarrow (\widehat{V}^{\mathrm{red}}_{\mathrm{up},b})_{C_{bk},\bullet}$.
  Construct $y_{b,h}$, and compute $T_{b,h} \leftarrow \widehat{V}_{bk}^\top\, y_{b,h}$, $W_{b,h} \leftarrow \widehat{\Sigma}^{\mathrm{red}}_{\mathrm{up},b} T_{b,h}$, $X_{b,h} \leftarrow \widehat{U}^{(k)}_{\mathrm{up},b} W_{b,h}$.
**end for**
Set $S_h \leftarrow 0$.
**for** $(a,b) \in \mathcal{D}_{k,h}$ **do**
  Set $\widehat{U}^{(a,b)}_{+k} \leftarrow (\widehat{U}^{\mathrm{red}}_{\mathrm{left},a})_{S^+_{k,b},\bullet}$.
  Compute $\widehat{H}_{a,b} \leftarrow (\widehat{U}^{(a,b)}_{+k})^\top \widehat{U}^{(a,b)}_{+k}$ and $\widehat{H}_{a,b} = Q_{a,b}\,\mathrm{diag}(\lambda_{a,b,1},\ldots,\lambda_{a,b,r})\,Q_{a,b}^\top$. Then set

$$\widehat{H}^{\mathrm{inv}}_{a,b,\tau} \leftarrow Q_{a,b}\,\mathrm{diag}\left(\left\{\frac{1}{\max[\lambda_{a,b,i},\tau]}\right\}_{i=1}^{r}\right) Q_{a,b}^\top.$$

  Compute $\widehat{\beta}^{(k,a,b,h)}_{\mathrm{red}} \leftarrow \widehat{H}^{\mathrm{inv}}_{a,b,\tau}(\widehat{U}^{(a,b)}_{+k})^\top X_{b,h} \in \mathbb{R}^r$.
  Update $S_h \leftarrow S_h + c^{(h)}_{ab}\langle \widehat{\alpha}^{(k,a,h)}_{\mathrm{red}}, \widehat{\beta}^{(k,a,b,h)}_{\mathrm{red}}\rangle$.
**end for**
**return** $\widehat{\Psi}^{(h)}_{0,\mathrm{lin}}(k) \leftarrow \{W_h(k)\}^{-1} S_h$.

---

Algorithm 4 avoids recomputing two spectral decompositions for each missing block by exploiting the reduced-anchor construction. Instead, it computes at most $o_k - 1$ left decompositions and at most $o_k - 1$ upper decompositions, thus making the dominant spectral cost linear in $o_k$ rather than quadratic. This reduction is obtained at the expense of statistical efficiency. While the blockwise auxiliary construction uses all anchor rows and columns available for each target block, the reduced-anchor construction uses a smaller common set of anchors, restricting the non-target-slice column anchors to $Q_{k,1}$ and the non-target-slice row anchors to $S_{k,1}$. The resulting pooled matrices may therefore contain less information, which can weaken the conditioning of the auxiliary Gram matrices and make the estimated shared row and column subspaces less accurate. Nevertheless, when $Q_{k,1}$ and $S_{k,1}$ are sufficiently large and well conditioned, Algorithm 4 provides a computationally cheaper alternative to Algorithm 3. A numerical comparison between Algorithm 3 and Algorithm 4 is illustrated in Fig. 5.

### B.2 Castle Doctrine data

We provide more details on the simulation setup of Section 6.2. We use the Castle Doctrine data from the `PolicyEval` repository, available on GitHub. The data contain state identifiers, calendar years, a Castle Doctrine treatment variable, and several state-level public-safety and socioeconomic variables. We use four logged crime-rate outcomes corresponding to the log motor-theft rate, log robbery rate, log aggravated-assault rate, and log murder rate.

We organise the observed outcomes into a fully observed tensor $\mathcal{Y} \in \mathbb{R}^{50\times 11\times 4}$, whose modes correspond to U.S. states, calendar years, and crime outcomes. Thus, each entry $\mathcal{Y}_{itj}$ records outcome $j$ for state $i$ in year $t$. The four outcome slices are `l_motor`, `l_robbery`, `l_assault`, and `l_homicide`, respectively.

We construct the staggered-adoption mask $\Omega$ from the Castle Doctrine treatment variable. For all $j \in [K]$, we set $\Omega_{itj} = 1$ if entry $(i, t, j)$ lies in the untreated region and zero otherwise. Because treatment status is common across crime outcomes, the same treatment pattern applies to each outcome slice; equivalently, $\Omega_{\bullet,\bullet,1} = \Omega_{\bullet,\bullet,2} = \Omega_{\bullet,\bullet,3} = \Omega_{\bullet,\bullet,4}$. Rows are ordered with never-adopting states at the top. Among adopting states, rows are arranged from later to earlier adopters, so that the treatment boundary moves smoothly across the panel. The resulting observation patterns are shown in Figure 9. Blue cells indicate untreated observations, red cells indicate treated observations, and darker shades correspond to larger logged crime-rate values. In the potential-outcomes notation $\mathcal{Y}_{itj} = \Omega_{itj}\,\mathcal{Y}_{itj}(0) + (1-\Omega_{itj})\,\mathcal{Y}_{itj}(1)$, the blue cells are therefore the observed entries of $\mathcal{Y}(0)$, while the corresponding treated entries are treated as missing. Conversely, the red cells are the observed entries of $\mathcal{Y}(1)$, with the corresponding untreated entries treated as missing.

## C Matrix denoising in the pooled four-block setting

This appendix collects matrix denoising results specialised to the tensor four-block framework introduced in Section 2. Throughout, we allow the constant $c_1$ to vary from line to line, while still depending only on $c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa$. We use the notation introduced in Table 1. We also denote by $\mathbb{O}(d) := \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = QQ^\top = I_d\}$ the set of orthogonal matrices of dimension $d$, and write $\mathcal{I}_k := \{N_{1k}+1, \dots, N\}$, $\mathcal{J}_k := \{T_{1k}+1, \dots, T\}$ and $\mathcal{I}_k^{\mathrm{up}} := \{s_k+1, \dots, s_k+N_{1k}\}$, where we recall $s_k = \sum_{j=1}^{k-1} N_{1j}$.

The main results of this section are Lemmas 14 and 15. Their proofs rely on the intermediate results

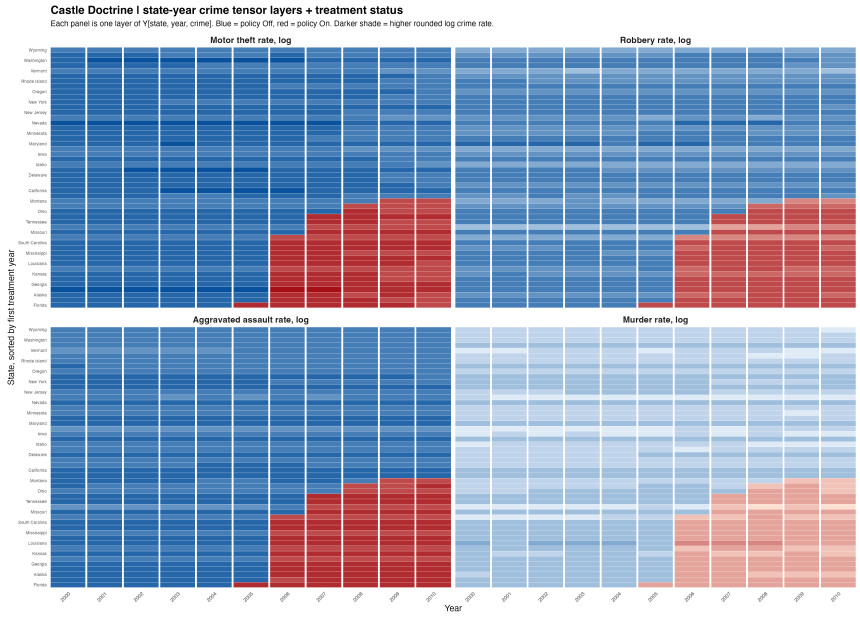


Figure 9: Castle Doctrine state-year-crime tensor used in the real-data simulation. Panels show logged crime rates for motor theft, robbery, aggravated assault, and murder. Rows are U.S. states, with never-adopters first and adopters ordered from later to earlier adoption years; columns are calendar years from 2000 to 2010. Blue cells are untreated, red cells are treated, and darker shades indicate larger logged crime-rate values.

presented below, which provide first-order expansions of some relevant quantities appearing in the definition of $\hat{\mu}_{xy}^{(k)}$ in terms of the matrices $W_{\text{left}}$, $W_{\text{up}}$, $E_{\text{left}}^{\text{p}}$, $E_{\text{up}}^{\text{p}}$. The first step in applying these results to $Y_{\text{left}}^{\text{p}}$ and $Y_{\text{up}}^{\text{p}}$ is to characterise the spectrum of the corresponding signal matrices $M_{\text{left}}^{\text{p}}$ and $M_{\text{up}}^{\text{p}}$.

**Lemma 5.** *Grant assumption (A1) with fixed constants $c_\ell, c_u$ satisfying $0 < c_\ell \le c_u < \infty$, and suppose $0 < \gamma_{\min} \le \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \le \gamma_{\max} < \infty$ for all $j \in [K]$. We have*

$$\begin{cases} c_\ell^{1/2}\gamma_{\min}\,\rho_T^{1/2} \le \sigma_r(M_{\text{left}}^{\text{p}}) \le \sigma_1(M_{\text{left}}^{\text{p}}) \le c_u^{1/2}\gamma_{\max}\,\rho_T^{1/2}, \\ c_\ell^{1/2}\gamma_{\min}\,\rho_N^{1/2} \le \sigma_r(M_{\text{up}}^{\text{p}}) \le \sigma_1(M_{\text{up}}^{\text{p}}) \le c_u^{1/2}\gamma_{\max}\,\rho_N^{1/2}. \end{cases}$$

*Proof.* Assumption (A1) implies

$$c_\ell \sum_{j=1}^{K} \frac{T_{1j}}{T}\,\mathcal{C}_{\bullet,\bullet,j}\mathcal{C}_{\bullet,\bullet,j}^\top \preceq W_{\text{left}}^\top W_{\text{left}} = \sum_{k=1}^{K} \mathcal{C}_{\bullet,\bullet,k}\,V_{1k}^\top V_{1k}\,\mathcal{C}_{\bullet,\bullet,k}^\top \preceq c_u \sum_{j=1}^{K} \frac{T_{1j}}{T}\,\mathcal{C}_{\bullet,\bullet,j}\mathcal{C}_{\bullet,\bullet,j}^\top,$$

$$c_\ell \sum_{j=1}^{K} \frac{N_{1j}}{N}\,\mathcal{C}_{\bullet,\bullet,j}^\top\mathcal{C}_{\bullet,\bullet,j} \preceq W_{\text{up}}^\top W_{\text{up}} = \sum_{k=1}^{K} \mathcal{C}_{\bullet,\bullet,k}^\top\,U_{1k}^\top U_{1k}\,\mathcal{C}_{\bullet,\bullet,k} \preceq c_u \sum_{j=1}^{K} \frac{N_{1j}}{N}\,\mathcal{C}_{\bullet,\bullet,j}^\top\mathcal{C}_{\bullet,\bullet,j}.$$

Since $M_{\text{left}}^{\text{p}} = UW_{\text{left}}^\top$, $M_{\text{up}}^{\text{p}} = W_{\text{up}}V^\top$ and $U, V$ have orthonormal columns, we also have $\sigma_j(M_{\text{left}}^{\text{p}}) = \sigma_j(W_{\text{left}})$

and $\sigma_j\big(M^{\mathrm{p}}_{\mathrm{up}}\big) = \sigma_j(W_{\mathrm{up}})$ for all $j \in [r]$. We thus get

$$\sqrt{c_\ell\,\lambda_r\left(\sum_{j=1}^{K}\frac{T_{1j}}{T}\,\mathcal{C}_{\bullet,\bullet,j}\mathcal{C}_{\bullet,\bullet,j}^\top\right)} \;\le\; \sigma_r(M^{\mathrm{p}}_{\mathrm{left}}) \;\le\; \sigma_1(M^{\mathrm{p}}_{\mathrm{left}}) \le \sqrt{c_u\,\lambda_1\left(\sum_{j=1}^{K}\frac{T_{1j}}{T}\,\mathcal{C}_{\bullet,\bullet,j}\mathcal{C}_{\bullet,\bullet,j}^\top\right)},$$

$$\sqrt{c_\ell\,\lambda_r\left(\sum_{j=1}^{K}\frac{N_{1j}}{N}\,\mathcal{C}_{\bullet,\bullet,j}^\top\mathcal{C}_{\bullet,\bullet,j}\right)} \;\le\; \sigma_r\big(M^{\mathrm{p}}_{\mathrm{up}}\big) \;\le\; \sigma_1\big(M^{\mathrm{p}}_{\mathrm{up}}\big) \le \sqrt{c_u\,\lambda_1\left(\sum_{j=1}^{K}\frac{N_{1j}}{N}\,\mathcal{C}_{\bullet,\bullet,j}^\top\mathcal{C}_{\bullet,\bullet,j}\right)}.$$

Combining this with $0 < \gamma_{\min} \le \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \le \gamma_{\max} < \infty$ concludes the proof. □

The following lemma provides an upper bound on the estimation error of $\widehat{U}_{\mathrm{left}}$ relative to $U$ measured by the operator norm of the projected error $\Pi_N^\top(\widehat{U}_{\mathrm{left}}H_U - U)$. The proof relies on tools from Haar compression and properties of the Stiefel manifold, as outlined in Appendix E.

**Lemma 6.** *Grant Assumptions (A1) with fixed constants $0 < c_\ell \le c_u$, (A2) and (A3). Suppose further that $0 < \gamma_{\min} \le \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \le \gamma_{\max} < \infty$ for all $j \in [K]$, and let $\kappa := \gamma_{\max}/\gamma_{\min}$. Write $Y^{\mathrm{p}}_{\mathrm{left}} = M^{\mathrm{p}}_{\mathrm{left}} + E^{\mathrm{p}}_{\mathrm{left}}$, with $M^{\mathrm{p}}_{\mathrm{left}} = UW^\top_{\mathrm{left}}$, and set $\Lambda := W^\top_{\mathrm{left}}W_{\mathrm{left}}$. Let $(\hat{U}_{\mathrm{left}}, \hat{\Sigma}_{\mathrm{left}}, \hat{V}_{\mathrm{left}}) := \mathrm{SVD}_r(Y^{\mathrm{p}}_{\mathrm{left}})$ and $H_U := \mathrm{sgn}(\hat{U}^\top_{\mathrm{left}}U)$, and define the centred empirical eigenvalue matrix $\hat{\Lambda}_c := H_U^\top(\hat{\Sigma}^2_{\mathrm{left}} - \sigma^2 T_{1,\mathrm{p}} I_r)H_U$. Also fix $1 \le p \le N$ and $\Pi_N \in \mathbb{R}^{N\times p}$ with $\Pi_N^\top\Pi_N = I_p$. There exists a constant $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ such that, with probability at least $1 - \mathcal{O}(p_T^{-10})$, the following statements hold:*

*(i) The centred empirical eigenvalue matrix is well-conditioned, i.e.*

$$\lambda_r(\hat{\Lambda}_c) \ge \frac{3}{4}\,\lambda_r(\Lambda), \qquad \|\hat{\Lambda}_c^{-1}\|_{\mathrm{op}} \le \frac{4}{3}\,\lambda_r(\Lambda)^{-1}. \tag{8}$$

*(ii) We have*

$$\|\Pi_N^\top(\hat{U}_{\mathrm{left}}H_U - U)\|_{\mathrm{op}} \le c_1\frac{\sigma\sqrt{p + r + \zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + c_1\,\frac{\sigma^2 N}{\gamma^2_{\min}\,\rho_T}\|\Pi_N^\top U\|_{\mathrm{op}}. \tag{9}$$

*Proof.* For readability, only in this proof we write $Y = Y^{\mathrm{p}}_{\mathrm{left}}$, $M = M^{\mathrm{p}}_{\mathrm{left}} = UW^\top_{\mathrm{left}}$, and $E = E^{\mathrm{p}}_{\mathrm{left}}$. Define $\hat{S} := YY^\top - \sigma^2 T_{1,\mathrm{p}} I_N$, $S_0 := MM^\top = U\Lambda U^\top$, and $\Xi := \hat{S} - S_0$. Let $U_\perp \in \mathbb{R}^{N\times(N-r)}$ be such that $[U \;\; U_\perp] \in \mathbb{O}(N)$.

Since $YY^\top$ and $\hat{S}$ differ by a scalar multiple of the identity, they have the same eigenvectors. Hence $\hat{U}_{\mathrm{left}}$ is also the top-$r$ eigenspace of $\hat{S}$, and

$$\hat{S}\hat{U}_{\mathrm{left}}H_U = \hat{U}_{\mathrm{left}}H_U\hat{\Lambda}_c. \tag{10}$$

Set $G_1 := U^\top E \in \mathbb{R}^{r\times T_{1,\mathrm{p}}}$ and $G_2 := U_\perp^\top E \in \mathbb{R}^{(N-r)\times T_{1,\mathrm{p}}}$. Since $E$ has independent $\mathcal{N}(0,\sigma^2)$ entries, rotational invariance ensures that $G_1$ and $G_2$ are independent Gaussian matrices with i.i.d. $\mathcal{N}(0,\sigma^2)$ entries.

Expanding $\Xi$ gives $\Xi = UW_{\text{left}}^\top E^\top + EW_{\text{left}}U^\top + (EE^\top - \sigma^2 T_{1,\text{p}} I_N)$, and therefore

$$\begin{aligned} U^\top \Xi U &= W_{\text{left}}^\top G_1^\top + G_1 W_{\text{left}} + (G_1 G_1^\top - \sigma^2 T_{1,\text{p}} I_r), \\ U_\perp^\top \Xi U &= G_2 W_{\text{left}} + G_2 G_1^\top = G_2 K, \qquad K := W_{\text{left}} + G_1^\top, \\ U_\perp^\top \Xi U_\perp &= G_2 G_2^\top - \sigma^2 T_{1,\text{p}} I_{N-r}. \end{aligned} \tag{11}$$

Write the aligned empirical eigenspace as $\hat{U}_{\text{left}} H_U = UC + U_\perp S$, where $C := U^\top \hat{U}_{\text{left}} H_U$ and $S := U_\perp^\top \hat{U}_{\text{left}} H_U$. Since $H_U$ is the Procrustes alignment, $C$ is symmetric and satisfies $C \succeq 0$, $C^\top C + S^\top S = I_r$. We can thus write

$$\Pi_N^\top (\hat{U}_{\text{left}} H_U - U) = \Pi_N^\top U_\perp S + \Pi_N^\top U (C - I_r), \tag{12}$$

which shows that it is enough to control the two terms on the right-hand side.

More precisely, the preceding decomposition shows that sharp control of some projection of $\hat{U}_{\text{left}} H_U - U$ reduces mainly to controlling the off-subspace component $S = U_\perp^\top \hat{U}_{\text{left}} H_U$. Indeed, $C = (I_r - S^\top S)^{1/2}$ and $\|C - I_r\|_{\text{op}} \le \|S\|_{\text{op}}^2$, so the term involving $C - I_r$ is second order. A direct application of Wedin's theorem (Chen et al., 2021, Section 2.4) would control only the global subspace error $\|S\|_{\text{op}} = \|U_\perp^\top \hat{U}_{\text{left}}\|_{\text{op}} = \|\sin\Theta(\hat{U}_{\text{left}}, U)\|_{\text{op}} \lesssim \|E_{\text{left}}^{\text{p}}\|_{\text{op}} / \sigma_r(M_{\text{left}}^{\text{p}})$, which is governed by an ambient noise norm and hence scales with the full row dimension $N$. When combined with the triangle inequality $\|\Pi_N^\top (\hat{U}_{\text{left}} H_U - U)\|_{\text{op}} \le \|\Pi_N^\top U_\perp\|_{\text{op}} \|S\|_{\text{op}} + \|\Pi_N^\top U\|_{\text{op}} \|S\|_{\text{op}}^2$, this would not exploit the fact that $\Pi_N$ has only $p$ columns. Instead, we apply the Haar–Stiefel compression bounds from Appendix E, which allow us to use the fixed projection $\Pi_N^\top U_\perp$ to reduce the random off-subspace component by a factor of order $\sqrt{(p + r + \zeta_T)/N}$, as shown in (20). This is the key idea that turns an ambient subspace perturbation estimate into the projected bound needed here.

• **Conditioning of $\hat{\Lambda}_c$.** The eigenvalues of $\hat{\Lambda}_c$ are the top $r$ eigenvalues of $\hat{S}$. By the Courant–Fischer formula restricted to $\text{col}(U)$, we have $\lambda_r(\hat{\Lambda}_c) = \lambda_r(\hat{S}) \ge \lambda_r(U^\top \hat{S} U) = \lambda_r(\Lambda + U^\top \Xi U)$. Weyl's inequality (Lemma 23) then gives

$$\lambda_r(\hat{\Lambda}_c) \ge \lambda_r(\Lambda) - \|U^\top \Xi U\|_{\text{op}}. \tag{13}$$

Define the events

$$\|G_1 W_{\text{left}}\|_{\text{op}} \le c_1 \sigma \|W_{\text{left}}\|_{\text{op}} \sqrt{r + \zeta_T}, \qquad \|G_1 G_1^\top - \sigma^2 T_{1,\text{p}} I_r\|_{\text{op}} \le c_1 \sigma^2 \left\{ \sqrt{T_{1,\text{p}}(r + \zeta_T)} + r + \zeta_T \right\}.$$

By Lemma 21 in Appendix E and a standard Wishart concentration bound (e.g. Vershynin, 2019, Theorem 4.6.1), the probability that at least one of the two displayed events fails is at most $\mathcal{O}(p_T^{-10})$. If these bounds hold, the first display in (11) implies $\|U^\top \Xi U\|_{\text{op}} \le c_1 \sigma \|W_{\text{left}}\|_{\text{op}} \sqrt{r + \zeta_T} + c_1 \sigma^2 \{ \sqrt{T_{1,\text{p}}(r + \zeta_T)} + r + \zeta_T \}$. By Lemma 5 we have $\lambda_r(\Lambda) \ge c_\ell \, \gamma_{\min}^2 \rho_T$ and $\|W_{\text{left}}\|_{\text{op}} = \sigma_1(M_{\text{left}}^{\text{p}}) \le c_u^{1/2} \, \gamma_{\max} \rho_T^{1/2}$, which yield

$$\frac{\|U^\top \Xi U\|_{\text{op}}}{\lambda_r(\Lambda)} \le c_1 \kappa \frac{\sigma}{\gamma_{\min}} \sqrt{\frac{T(r + \zeta_T)}{T_{1,\text{p}}}} + c_1 \frac{\sigma^2 T}{\gamma_{\min}^2} \left( \sqrt{\frac{r + \zeta_T}{T_{1,\text{p}}}} + \frac{r + \zeta_T}{T_{1,\text{p}}} \right)$$

$$\leq c_1 \frac{\sigma}{\gamma_{\min}} \sqrt{\frac{NT}{T_{1,\mathrm{p}}}} + c_1 \frac{\sigma^2 T}{\gamma_{\min}^2} \left( \sqrt{\frac{N}{T_{1,\mathrm{p}}}} + \frac{N}{T_{1,\mathrm{p}}} \right) \leq c_1 \frac{\sigma}{\gamma_{\min}} \sqrt{\frac{NT}{T_{1,\mathrm{p}}}} \leq c_1 \theta \leq \frac{1}{4}, \tag{14}$$

where the last inequality follows from Assumptions (A2) and (A3), provided that the absolute constants $c_0 > 0$ and $c_{\mathrm{blk}} > 0$ are chosen sufficiently small. Combining this with (13) proves (i).

• **Reduction to the range of $G_2$.** Using (10) and $\hat{U}_{\mathrm{left}} H_U = UC + U_\perp S$, and then left-multiplying by $U_\perp^\top$, we obtain

$$G_2 KC + (G_2 G_2^\top - \sigma^2 T_{1,\mathrm{p}} I_{N-r}) S = S \hat{\Lambda}_c. \tag{15}$$

Let $P_2 := \mathrm{Proj}_{\mathrm{col}(G_2)}$ and $P_2^\perp := I_{N-r} - P_2$. Since $P_2^\perp G_2 K = 0$, multiplying (15) by $P_2^\perp$ gives $P_2^\perp S(\hat{\Lambda}_c + \sigma^2 T_{1,\mathrm{p}} I_r) = 0$. On the event where $\hat{\Lambda}_c \succ 0$, we also have that $\hat{\Lambda}_c + \sigma^2 T_{1,\mathrm{p}} I_r$ is invertible, hence $P_2^\perp S(\hat{\Lambda}_c + \sigma^2 T_{1,\mathrm{p}} I_r) = 0$ implies $P_2^\perp S = 0$, and therefore $S = P_2 S$. Defining $D_2 := (G_2 G_2^\top - \sigma^2 T_{1,\mathrm{p}} I_{N-r}) P_2$, we may rewrite (15) as

$$G_2 KC + D_2 S = S \hat{\Lambda}_c. \tag{16}$$

The inclusion of the projection $P_2$ in the definition of $D_2$ is crucial: although $G_2 G_2^\top - \sigma^2 T_{1,\mathrm{p}} I_{N-r}$ may be large on $\mathrm{col}(G_2)^\perp$, the identity $S = P_2 S$ ensures that only its restriction to $\mathrm{col}(G_2)$ is relevant.

• **Restricted centered-Wishart bound for $D_2$.** Let $q := \mathrm{rank}(G_2) = \min(N - r, T_{1,\mathrm{p}})$ almost surely. Since $D_2$ acts on $\mathrm{col}(G_2)$, we have $\|D_2\|_{\mathrm{op}} = \max_{1 \leq i \leq q} |\sigma_i(G_2)^2 - \sigma^2 T_{1,\mathrm{p}}|$. By the standard two-sided singular value bound for Gaussian matrices (e.g. Vershynin, 2019, Theorem 4.6.1) applied to $G_2^\top / \sigma$, and using (A2) to absorb the terms involving $\zeta_T$ into the right-hand side, we have $\|D_2\|_{\mathrm{op}} \leq c_1 \sigma^2 (\sqrt{N T_{1,\mathrm{p}}} + N + \zeta_T)$ with probability at least $1 - \mathcal{O}(p_T^{-10})$. This also implies

$$\|D_2\|_{\mathrm{op}} \|\hat{\Lambda}_c^{-1}\|_{\mathrm{op}} \leq c_1 \frac{\sigma^2 T}{\gamma_{\min}^2} \left( \sqrt{\frac{N}{T_{1,\mathrm{p}}}} + \frac{N + \zeta_T}{T_{1,\mathrm{p}}} \right) \leq c_1 \, \theta^2 \leq \frac{1}{4}, \tag{17}$$

where the last inequality follows from (A3).

• **Control of $G_2 K$.** We now control $G_2 K \hat{\Lambda}_c^{-1}$ both globally and after projection by $\Pi_N^\top U_\perp$. Conditional on $G_1$, the matrix $K = W_{\mathrm{left}} + G_1^\top$ is deterministic, independent of $G_2$, and $\mathrm{rank}(K) \leq r$. Hence Lemma 21 gives

$$\|\Pi_N^\top U_\perp G_2 K\|_{\mathrm{op}} \leq c_1 \sigma \|K\|_{\mathrm{op}} \sqrt{p + r + \zeta_T}, \qquad \|G_2 K\|_{\mathrm{op}} \leq c_1 \sigma \|K\|_{\mathrm{op}} \sqrt{N + r + \zeta_T} \tag{18}$$

with probability at least $1 - \mathcal{O}(p_T^{-10})$. The same bounds hold unconditionally with the same probability. Moreover, on the event $\|G_1\|_{\mathrm{op}} \leq c_1 \sigma (\sqrt{T_{1,\mathrm{p}}} + \sqrt{r + \zeta_T})$, which has probability at least $1 - \mathcal{O}(p_T^{-10})$ again by Lemma 21, we have $\|K\|_{\mathrm{op}} \leq \|W_{\mathrm{left}}\|_{\mathrm{op}} + \|G_1\|_{\mathrm{op}} \leq c_1 (\gamma_{\max} \rho_T^{1/2} + \sigma \sqrt{T_{1,\mathrm{p}}} + \sigma \sqrt{r + \zeta_T})$, where the second inequality follows from Lemma 5. Combining this bound with (18) and (i) gives, on an event of probability

at least $1-\mathcal{O}(p_T^{-10})$, we have

$$\begin{aligned}\|\Pi_N^\top U_\perp G_2 K\|_{\mathrm{op}}\|\hat{\Lambda}_c^{-1}\|_{\mathrm{op}} &\le c_1\frac{\sigma\sqrt{p+r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + c_1\frac{\sigma^2 T}{\gamma_{\min}^2}\left(\sqrt{\frac{p+r+\zeta_T}{T_{1,\mathrm{p}}}} + \frac{\sqrt{(p+r+\zeta_T)(r+\zeta_T)}}{T_{1,\mathrm{p}}}\right)\\ &\le c_1\frac{\sigma\sqrt{p+r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}},\end{aligned}$$

$$\begin{aligned}\|G_2 K\|_{\mathrm{op}}\|\hat{\Lambda}_c^{-1}\|_{\mathrm{op}} &\le c_1\frac{\sigma\sqrt{N+r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + c_1\frac{\sigma^2 T}{\gamma_{\min}^2}\left(\sqrt{\frac{N+r+\zeta_T}{T_{1,\mathrm{p}}}} + \frac{\sqrt{(N+r+\zeta_T)(r+\zeta_T)}}{T_{1,\mathrm{p}}}\right)\\ &\le c_1\frac{\sigma\sqrt{N}}{\gamma_{\min}\,\rho_T^{1/2}},\end{aligned} \tag{19}$$

where the final inequalities in both displays follow from (A2), (A3).

• **Haar-measure step.** We now aim to control $\|\Pi_N^\top U_\perp S\|_{\mathrm{op}}$ using Lemma 25. This is achieved through a Haar-measure argument; see Appendix E for the statement of the lemma and the relevant background material on the Stiefel manifold. The first step is to rewrite (16) in coordinates adapted to $\mathrm{col}(G_2)$. In this regard, let $q = \mathrm{rank}(G_2)$. Choose $(V_2, \Sigma_2)$ measurably from the eigendecomposition of $G_2^\top G_2 = V_2\Sigma_2^2 V_2^\top$, with the positive eigenvalues sorted in decreasing order, and define $Q := G_2 V_2 \Sigma_2^{-1} \in \mathrm{St}(N-r, q)$, so that $G_2 = Q\Sigma_2 V_2^\top$. Since $P_2$ is the orthogonal projector onto $\mathrm{col}(G_2)$, we have $P_2 = QQ^\top$. Thus $S = P_2 S$ implies $S = QR$, where $R := Q^\top S \in \mathbb{R}^{q\times r}$. Substituting $G_2 = Q\Sigma_2 V_2^\top$ and $S = QR$ into (16), and using $D_2 = Q(\Sigma_2^2 - \sigma^2 T_{1,\mathrm{p}} I_q)Q^\top$, gives $\Sigma_2 V_2^\top KC + (\Sigma_2^2 - \sigma^2 T_{1,\mathrm{p}} I_q)R = R\hat{\Lambda}_c$ after multiplying by $Q^\top$.

Now, let $Q_\perp$ be chosen measurably so that $[Q\ Q_\perp] \in \mathbb{O}(N-r)$, with the last block omitted if $q = N-r$, and set $O := [\,U\ \ U_\perp Q\ \ U_\perp Q_\perp\,]$. The following calculations allow us to show that $\hat{S}$ has a simple block form with respect to the basis given by $O$. Using (11), $\widehat{S} = U\Lambda U^\top + \Xi$ and $G_2 = Q\Sigma_2 V_2^\top$, we get

$$U^\top \widehat{S}(U_\perp Q) = (Q^\top U_\perp^\top \widehat{S} U)^\top = (Q^\top G_2 K)^\top = (\Sigma_2 V_2^\top K)^\top = K^\top V_2 \Sigma_2,$$

$$(U_\perp Q)^\top \widehat{S} U = Q^\top U_\perp^\top \widehat{S} U = Q^\top G_2 K = \Sigma_2 V_2^\top K,$$

$$(U_\perp Q)^\top \widehat{S}(U_\perp Q) = Q^\top (G_2 G_2^\top - \sigma^2 T_{1,\mathrm{p}} I_{N-r}) Q = \Sigma_2^2 - \sigma^2 T_{1,\mathrm{p}} I_q.$$

Moreover, since $\mathrm{col}(G_2) = \mathrm{col}(Q)$, we have $Q_\perp^\top G_2 = 0$. Hence $U^\top \widehat{S}(U_\perp Q_\perp) = 0$, $(U_\perp Q)^\top \widehat{S}(U_\perp Q_\perp) = 0$, $(U_\perp Q_\perp)^\top \widehat{S} U = 0$, $(U_\perp Q_\perp)^\top \widehat{S}(U_\perp Q) = 0$, and $(U_\perp Q_\perp)^\top \widehat{S}(U_\perp Q_\perp) = Q_\perp^\top (G_2 G_2^\top - \sigma^2 T_{1,\mathrm{p}} I_{N-r}) Q_\perp = -\sigma^2 T_{1,\mathrm{p}} I_{N-r-q}$. We therefore get

$$O^\top \hat{S} O = \begin{pmatrix} \Lambda + U^\top \Xi U & K^\top V_2 \Sigma_2 & 0 \\ \Sigma_2 V_2^\top K & \Sigma_2^2 - \sigma^2 T_{1,\mathrm{p}} I_q & 0 \\ 0 & 0 & -\sigma^2 T_{1,\mathrm{p}} I_{N-r-q} \end{pmatrix}.$$

This shows that, on the event that $\hat{S}$ has at least $r$ positive eigenvalues, the top-$r$ eigenspace of $\hat{S}$ is contained in the column space of $(U, U_\perp Q)$. Formally, using $S = QR$, we have $\hat{U}_{\mathrm{left}} H_U = UC + U_\perp QR = O(C^\top, R^\top, 0)^\top$,

and since $\hat{U}_{\text{left}} H_U$ is the aligned top-$r$ eigenspace of $\hat{S}$, we can write

$$(C^\top, R^\top, 0)^\top \hat{\Lambda}_c = O^\top \hat{U}_{\text{left}} H_U \hat{\Lambda}_c = O^\top \hat{S} \hat{U}_{\text{left}} H_U = O^\top \hat{S} O (C^\top, R^\top, 0)^\top.$$

Comparing the first two block rows gives

$$\begin{pmatrix} \Lambda + U^\top \Xi U & K^\top V_2 \Sigma_2 \\ \Sigma_2 V_2^\top K & \Sigma_2^2 - \sigma^2 T_{1,\mathrm{p}} I_q \end{pmatrix} \begin{pmatrix} C \\ R \end{pmatrix} = \begin{pmatrix} C \\ R \end{pmatrix} \hat{\Lambda}_c.$$

On the event $\lambda_r(\hat{\Lambda}_c) > 0$, the $r$ largest eigenvalues of $\hat{S}$ are positive and therefore cannot arise from the negative block $-\sigma^2 T_{1,\mathrm{p}} I_{N-r-q}$, hence they correspond precisely to the $r$ largest eigenvalues of the reduced block above. We thus get that $(C^\top, R^\top)^\top$ is the top-$r$ eigenspace of the reduced block, and $\hat{\Lambda}_c$ is the associated eigenvalue matrix.

Now, this reduced block depends on $G_2$ only through $(\Sigma_2, V_2)$, and not through $Q$. Moreover, since $K = W_{\text{left}} + G_1^\top$ and $U^\top \Xi U$ is a function of $G_1$, the reduced block is measurable with respect to $\mathcal{F} := \sigma(G_1, \Sigma_2, V_2)$, hence, after fixing deterministic measurable choices of eigenspaces and of the Procrustes alignment, $(C, R, \hat{\Lambda}_c)$ is $\mathcal{F}$-measurable. Furthermore, writing $R = H_R \Omega_R J_R^\top$ for the compact singular value decomposition of $R$ and $\ell = \operatorname{rank}(R) \le r$, also $H_R$ is $\mathcal{F}$-measurable. For completeness, observe that we may assume $\ell \ge 1$; if $\ell = 0$ the desired bound is immediate.

On the other hand, since $G_2$ has i.i.d. Gaussian entries, its law is left-orthogonally invariant, in the sense that for every deterministic $O_0 \in \mathbb{O}(N-r)$ we have $O_0 G_2 \overset{d}{=} G_2$. Moreover, $(O_0 G_2)^\top (O_0 G_2) = G_2^\top G_2$, so left multiplication changes only the left singular subspace, from $Q$ to $O_0 Q$, while leaving $(\Sigma_2, V_2)$ unchanged. It follows that the conditional law of $Q$ given $(\Sigma_2, V_2)$ is left-invariant on $\mathrm{St}(N-r, q)$, and by uniqueness of the left-orthogonally invariant probability measure on the Stiefel manifold (e.g. Chikuse, 2003, Theorem 1.2.2 and Section 1.3.1), it is Haar. Furthermore, since this conditional law does not depend on the value of $(\Sigma_2, V_2)$, then $Q$ is independent of $\sigma(\Sigma_2, V_2)$, and, since $G_1$ is independent of $G_2$, we also have $Q \perp\!\!\!\perp \mathcal{F}$.

We have therefore shown that, conditional on $\mathcal{F}$, the matrix $H_R$ is fixed, while $Q$ is Haar-distributed on $\mathrm{St}(N-r, q)$ and independent of $\mathcal{F}$. Hence, by Lemma 25, $QH_R$ is Haar-distributed on $\mathrm{St}(N-r, \ell)$ conditionally on $\mathcal{F}$. Applying the high probability bound in the same lemma and using (A2), we obtain $\|\Pi_N^\top U_\perp Q H_R\|_{\mathrm{op}} \le c_1 \sqrt{(p+r+\zeta_T)/(N-r)} \le c_1 \sqrt{(p+r+\zeta_T)/N}$ with conditional probability at least $1 - \mathcal{O}(p_T^{-10})$. The same bound also holds unconditionally. Since $S = QR = QH_R \Omega_R J_R^\top$ and $\|\Omega_R\|_{\mathrm{op}} = \|R\|_{\mathrm{op}} = \|QR\|_{\mathrm{op}} = \|S\|_{\mathrm{op}}$, we obtain

$$\|\Pi_N^\top U_\perp S\|_{\mathrm{op}} \le \|\Pi_N^\top U_\perp Q H_R\|_{\mathrm{op}} \, \|\Omega_R\|_{\mathrm{op}} \, \|J_R^\top\|_{\mathrm{op}} \le c_1 \sqrt{\frac{p+r+\zeta_T}{N}} \|S\|_{\mathrm{op}}. \tag{20}$$

The prefactor $\sqrt{(p+r+\zeta_T)/N}$ is precisely the projection factor that we aimed to obtain, as discussed at the beginning of the proof.

• **Conclusion.** It remains to bound $\|S\|_{\mathrm{op}}$. From (16) we get $S = G_2 K C \hat{\Lambda}_c^{-1} + D_2 S \hat{\Lambda}_c^{-1}$, and, since $\|C\|_{\mathrm{op}} \le 1$, the triangle inequality gives $\|S\|_{\mathrm{op}} \le \|G_2 K\|_{\mathrm{op}} \|\hat{\Lambda}_c^{-1}\|_{\mathrm{op}} + \|D_2\|_{\mathrm{op}} \|\hat{\Lambda}_c^{-1}\|_{\mathrm{op}} \|S\|_{\mathrm{op}}$. The preceding

Wishart bound for $D_2$ in (17) allows reordering this inequality, and gives

$$\|S\|_{\mathrm{op}} \le c_1 \|G_2 K\|_{\mathrm{op}} \|\hat{\Lambda}_c^{-1}\|_{\mathrm{op}} \le c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}} \tag{21}$$

by (19). This, together with (20), yields $\|\Pi_N^\top U_\perp S\|_{\mathrm{op}} \le c_1 \sigma \gamma_{\min}^{-1} \sqrt{(p+r+\zeta_T)/\rho_T}$. Finally, since $C = (I_r - S^\top S)^{1/2}$ and $\|C - I_r\|_{\mathrm{op}} \le \|S\|_{\mathrm{op}}^2$, we have $\|\Pi_N^\top U (C - I_r)\|_{\mathrm{op}} \le \|\Pi_N^\top U\|_{\mathrm{op}} \|S\|_{\mathrm{op}}^2 \le c_1 \frac{\sigma^2 N}{\gamma_{\min}^2 \rho_T} \|\Pi_N^\top U\|_{\mathrm{op}}$. This proves (9), and concludes the proof. □

Many useful corollaries can be derived from (9), yielding bounds that hold with probability at least $1 - \mathcal{O}(p_T^{-10})$. In particular, for any fixed $x \in \mathbb{B}_2(N)$, by choosing $\Pi_N = x$ we have

$$\|(\hat{U}_{\mathrm{left}} H_U - U)^\top x\|_2 \le c_1 \frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + c_1 \frac{\sigma^2 N}{\gamma_{\min}^2\,\rho_T} \|U^\top x\|_2 \tag{22}$$

with high probability, which gives an estimation bound directly for the action of the error $\hat{U}_{\mathrm{left}} H_U - U$ along any fixed unit direction. Furthermore, for fixed $k \in [K]$ the same bounds hold blockwise, in the sense that, for every fixed $1 \le p \le N_{1k}$ and $\Pi_N \in \mathbb{R}^{N_{1k} \times p}$ with $\Pi_N^\top \Pi_N = I_p$, we have

$$\|\Pi_N^\top (\hat{U}_{1k} H_U - U_{1k})\|_{\mathrm{op}} \le c_1 \frac{\sigma\sqrt{p+r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + c_1 \frac{\sigma^2 N}{\gamma_{\min}^2\,\rho_T} \|\Pi_N^\top U_{1k}\|_{\mathrm{op}}. \tag{23}$$

As a result, choosing $\Pi_N = I_{N_{1k}}$ and using (A1), (A2) give $\|\hat{U}_{1k}\, H_U - U_{1k}\|_{\mathrm{op}} \le c_1\, \sigma\, \gamma_{\min}^{-1} \sqrt{N_{1k}/\rho_T}$ and, as a byproduct, $\hat{U}_{1k}^\top \hat{U}_{1k}$ is invertible, satisfying $\frac{c_\ell}{2} \frac{N_{1k}}{N} I_r \preceq \hat{U}_{1k}^\top \hat{U}_{1k} \preceq 2c_u \frac{N_{1k}}{N} I_r$. This follows from Weyl's (Lemma 23), along similar lines to the proof of (8). This is particularly useful in our setting, as it ensures that the matrix $\hat{H}_{k,\tau}^{\mathrm{inv}}$ used in Algorithm 1 coincides with $(\hat{U}_{1k}^\top \hat{U}_{1k})^{-1}$ with high probability whenever the algorithm is run with $\tau \le \frac{c_\ell\, N_{1k}}{2\, N}$. Indeed, in this case we have

$$\begin{aligned}
\left\{ \frac{c_\ell}{2} \frac{N_{1k}}{N} I_r \preceq \hat{H}_k \preceq 2c_u \frac{N_{1k}}{N} I_r \right\} &= \left\{ \frac{c_\ell}{2} \frac{N_{1k}}{N} I_r \preceq \hat{H}_k \right\} \bigcap \left\{ \hat{H}_k \preceq 2c_u \frac{N_{1k}}{N} I_r \right\} \\
&\subseteq \left\{ \frac{c_\ell}{2} \frac{N_{1k}}{N} I_r \preceq \hat{H}_k \right\} \subseteq \left\{ \tau\, I_r \preceq \hat{H}_k \right\},
\end{aligned} \tag{24}$$

which further implies that $\mathbb{P}(\hat{H}_k \succeq \tau\, I_r) \ge 1 - \mathcal{O}(p_T^{-10})$. On this event, all eigenvalues of $\hat{H}_k$ are at least $\tau$, hence $\lambda_i \vee \tau = \lambda_i$ for every $i \in [r]$, and $\hat{H}_{k,\tau}^{\mathrm{inv}} = \hat{H}_k^{-1} = (\hat{U}_{1k}^\top \hat{U}_{1k})^{-1}$.

**Lemma 7.** *Let $\Lambda, \hat{\Lambda}_c, C$ be as in Lemma 6, and suppose that the assumptions of Lemma 6 hold. There exists a constant $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ such that, with probability at least $1 - \mathcal{O}(p_T^{-10})$, we have*

$$\left\| C \widehat{\Lambda}_c^{-1} - \Lambda^{-1} \right\|_{\mathrm{op}} \le c_1 \sigma \gamma_{\min}^{-3} \sqrt{N}\, \rho_T^{-3/2}.$$

*Proof.* All the bounds below will hold on the events from Lemma 6, which have probability at least $1 - \mathcal{O}(p_T^{-10})$. Using $U_\perp^\top \Xi U = G_2 K$, $C\widehat{\Lambda}_c - \Lambda = \Lambda(C - I_r) + U^\top \Xi U\, C + U^\top \Xi U_\perp S$, $\Lambda - \widehat{\Lambda}_c = \Lambda - C\widehat{\Lambda}_c + (C - I_r)\widehat{\Lambda}_c$,

and $C\widehat{\Lambda}_c^{-1} - \Lambda^{-1} = (C - I_r)\widehat{\Lambda}_c^{-1} + \widehat{\Lambda}_c^{-1}(\Lambda - \widehat{\Lambda}_c)\Lambda^{-1}$ from Lemma 6, we can write

$$\begin{aligned}
\|C\widehat{\Lambda}_c^{-1} - \Lambda^{-1}\|_{\mathrm{op}} &\leq \|C - I_r\|_{\mathrm{op}}\|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}} + \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|\Lambda - \widehat{\Lambda}_c\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} \\
&\leq \|C - I_r\|_{\mathrm{op}}\|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}} + \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|\Lambda - C\widehat{\Lambda}_c\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} + \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|C - I_r\|_{\mathrm{op}}\|\widehat{\Lambda}_c\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} \\
&\leq \|C - I_r\|_{\mathrm{op}}\|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}} + \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|\Lambda\|_{\mathrm{op}}\|C - I_r\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} + \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|U^\top \Xi U\|_{\mathrm{op}}\|C\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} \\
&\quad + \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|U^\top \Xi U_\perp S\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} + \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|C - I_r\|_{\mathrm{op}}\|\widehat{\Lambda}_c\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}}. \qquad (25)
\end{aligned}$$

We now bound the five terms on the right-hand side individually. For the first one, combining $\|C - I_r\|_{\mathrm{op}} \leq \|S\|_{\mathrm{op}}^2$, $\|\Lambda^{-1}\|_{\mathrm{op}} \lesssim \gamma_{\min}^{-2}\rho_T^{-1}$, (21) and (8) gives

$$\|C - I_r\|_{\mathrm{op}}\|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}} \leq c_1 \frac{\sigma^2 N}{\gamma_{\min}^4 \rho_T^2} = c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}} \frac{\sigma\sqrt{N}}{\gamma_{\min}^3\rho_T^{3/2}} \leq c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}^3\rho_T^{3/2}},$$

where the last inequality follows from (A3). The second term can be controlled by combining the previous bound with $\|\Lambda\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} \leq c_1\gamma_{\max}^2\rho_T\,\gamma_{\min}^{-2}\rho_T^{-1} = \kappa^2 c_1$ and incorporating the constant $\kappa^2$ into $c_1$. A similar argument applies to the fifth term, using the bound $\|\widehat{\Lambda}_c\|_{\mathrm{op}} \lesssim \gamma_{\max}^2\rho_T$.

For the third term in (25), we can use $\|C\|_{\mathrm{op}} \leq 1$, $\|\Lambda^{-1}\|_{\mathrm{op}} \lesssim \gamma_{\min}^{-2}\rho_T^{-1}$, (8) and the second line in (14) to write

$$\|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|U^\top \Xi U\|_{\mathrm{op}}\|C\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} \leq c_1 \frac{\|U^\top \Xi U\|_{\mathrm{op}}}{\lambda_r(\Lambda)} \frac{1}{\lambda_r(\Lambda)} \leq c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}} \frac{1}{\gamma_{\min}^2\rho_T} = c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}^3\rho_T^{3/2}}.$$

For the fourth piece, (11) gives $U_\perp^\top \Xi U = G_2 K$ hence $U^\top \Xi U_\perp S = K^\top G_2^\top S$. As a result, combining this with (21) and (19) therefore implies

$$\begin{aligned}
\|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|U^\top \Xi U_\perp S\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} &\leq \|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}}\|G_2 K\|_{\mathrm{op}}\|S\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} \leq c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}}\|S\|_{\mathrm{op}}\|\Lambda^{-1}\|_{\mathrm{op}} \\
&\leq c_1 \frac{\sigma^2 N}{\gamma_{\min}^2\rho_T} \frac{1}{\gamma_{\min}^2\rho_T} = c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}} \frac{\sigma\sqrt{N}}{\gamma_{\min}^3\rho_T^{3/2}} \leq c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}^3\rho_T^{3/2}},
\end{aligned}$$

where the last inequality follows from (A3). Combining the five bounds concludes the proof. □

We next provide a representation of $\widehat{U}_{\mathrm{left}} H_U - U$ in terms of a Gaussian term and a remainder.

**Corollary 8.** *Use the assumptions and notation of Lemma 6. Set* $\Psi_U := \widehat{U}_{\mathrm{left}} H_U - U - E_{\mathrm{left}}^{\mathrm{p}} W_{\mathrm{left}} (W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}$, *and for fixed* $k \in [K]$ *define* $\Psi_{U,2k} := (\Psi_U)_{\mathcal{I}_k,\bullet}$, *where* $\mathcal{I}_k = \{N_{1k} + 1, \ldots, N\}$. *For every fixed* $x \in \mathbb{B}_2(N_{2k})$, *there exists a constant* $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ *such that, with probability at least* $1 - \mathcal{O}(p_T^{-10})$, *we have*

$$\|(\Psi_{U,2k})^\top x\|_2 \leq c_1 \frac{\sigma^2\sqrt{(N + T_{1,\mathrm{p}})(r + \zeta_T)}}{\gamma_{\min}^2\rho_T} + c_1 \left( \frac{\sigma^2 N}{\gamma_{\min}^2\rho_T} + \frac{\sigma\sqrt{r + \zeta_T}}{\gamma_{\min}\rho_T^{1/2}} \right) \|U_{2k}^\top x\|_2. \qquad (26)$$

*Proof.* Let $\bar{x} \in \mathbb{B}_2(N)$ be the zero extension of $x$ to the coordinates $\mathcal{I}_k$, i.e. $\bar{x}_i = x_{i - N_{1k}}\,\mathbb{1}\{i \in \mathcal{I}_k\}$ for all $i \in [N]$. Then $\|(\Psi_{U,2k})^\top x\|_2 = \|\Psi_U^\top \bar{x}\|_2$ and $\|U^\top \bar{x}\|_2 = \|U_{2k}^\top x\|_2$.

We now recall some important facts that were already stated and justified in the proof of Lemma 6. For readability write $E := E^{\mathrm{p}}_{\mathrm{left}}$, $Y := Y^{\mathrm{p}}_{\mathrm{left}}$, $W := W_{\mathrm{left}}$, $M := M^{\mathrm{p}}_{\mathrm{left}} = UW^\top$, and $\Lambda := W^\top W$. Let $U_\perp \in \mathbb{R}^{N\times(N-r)}$ be such that $[U\ U_\perp] \in \mathbb{O}(N)$, and define $G_1 := U^\top E$ and $G_2 := U_\perp^\top E$. Since $E$ has independent $\mathcal{N}(0,\sigma^2)$ entries, rotational invariance gives that $G_1$ and $G_2$ are independent Gaussian matrices with independent $\mathcal{N}(0,\sigma^2)$ entries. Set $\widehat{S} := YY^\top - \sigma^2 T_{1,\mathrm{p}} I_N$ and $\widehat{\Lambda}_c := H_U^\top(\widehat{\Sigma}^2_{\mathrm{left}} - \sigma^2 T_{1,\mathrm{p}} I_r)H_U$. Since subtracting $\sigma^2 T_{1,\mathrm{p}} I_N$ does not change eigenvectors, $\widehat{U}_{\mathrm{left}}H_U$ satisfies $\widehat{S}\widehat{U}_{\mathrm{left}}H_U = \widehat{U}_{\mathrm{left}}H_U\widehat{\Lambda}_c$. Write $\widehat{U}_{\mathrm{left}}H_U = UC + U_\perp S$, where $C := U^\top\widehat{U}_{\mathrm{left}}H_U$ and $S := U_\perp^\top\widehat{U}_{\mathrm{left}}H_U$. Because $H_U$ is the Procrustes alignment, $C$ is symmetric positive semidefinite and $C = (I_r - S^\top S)^{1/2}$. Expanding $\widehat{S} - U\Lambda U^\top$ gives $\Xi := \widehat{S} - U\Lambda U^\top = UW^\top E^\top + EWU^\top + (EE^\top - \sigma^2 T_{1,\mathrm{p}} I_N)$. Therefore $U_\perp^\top \Xi U = G_2 W + G_2 G_1^\top = G_2(W + G_1^\top)$, and projecting onto $U_\perp$ gives

$$S = G_2(W + G_1^\top)C\widehat{\Lambda}_c^{-1} + D_2 S\widehat{\Lambda}_c^{-1}, \qquad D_2 := (G_2G_2^\top - \sigma^2 T_{1,\mathrm{p}} I_{N-r})P_2, \tag{27}$$

where $P_2$ denotes the orthogonal projector onto the column space of $G_2$. In particular, this follows from $S = P_2 S$, which holds under $\widehat{\Lambda}_c \succ 0$. Also, since $EW = UG_1W + U_\perp G_2 W$, the definition of $\Psi_U$ gives $\Psi_U = U_\perp\{S - G_2W\Lambda^{-1}\} + U\{C - I_r - G_1W\Lambda^{-1}\}$, consequently we have

$$\|\Psi_U^\top \bar{x}\|_2 \le \|\{S - G_2W\Lambda^{-1}\}^\top U_\perp^\top \bar{x}\|_2 + \|(C - I_r)U^\top\bar{x}\|_2 + \|\{G_1W\Lambda^{-1}\}^\top U^\top \bar{x}\|_2. \tag{28}$$

We now bound these three terms separately. First, subtracting $G_2W\Lambda^{-1}$ from the first equation in (27) yields

$$S - G_2W\Lambda^{-1} = G_2W(C\widehat{\Lambda}_c^{-1} - \Lambda^{-1}) + G_2G_1^\top C\widehat{\Lambda}_c^{-1} + D_2S\widehat{\Lambda}_c^{-1}. \tag{29}$$

We recall that on the high-probability event from Lemma 6 we have $\lambda_r(\widehat{\Lambda}_c) \ge 3\lambda_r(\Lambda)/4$, $\|\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}} \le c_1\lambda_r(\Lambda)^{-1}$, $\|S\|_{\mathrm{op}} \le c_1\sigma\gamma_{\min}^{-1}\sqrt{N/\rho_T}$, and $\|C - I_r\|_{\mathrm{op}} \le \|S\|^2_{\mathrm{op}}$. Moreover, arguing as in the paragraph after (13) and simplifying some bounds using (A2), the same event also gives $\|G_2W\|_{\mathrm{op}} \le c_1\sigma\|W\|_{\mathrm{op}}\sqrt{N}$, $\|G_1W\|_{\mathrm{op}} \le c_1\sigma\|W\|_{\mathrm{op}}\sqrt{r+\zeta_T}$, and $\|G_2G_1^\top\|_{\mathrm{op}} \le c_1\sigma^2\sqrt{NT_{1,\mathrm{p}}}$. Now, using $\lambda_r(\Lambda) \ge c_1\gamma^2_{\min}\rho_T$, the middle term of (29) above satisfies

$$\|G_2G_1^\top C\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}} \le c_1\frac{\sigma^2\sqrt{NT_{1,\mathrm{p}}}}{\gamma^2_{\min}\rho_T}.$$

We now deal with the other two terms. First, combining $\|C\widehat{\Lambda}_c^{-1} - \Lambda^{-1}\|_{\mathrm{op}} \le c_1\sigma\gamma_{\min}^{-3}\sqrt{N}\rho_T^{-3/2}$ from Lemma 7 with the above bound for $\|G_2W\|_{\mathrm{op}}$ gives $\|G_2W(C\widehat{\Lambda}_c^{-1} - \Lambda^{-1})\|_{\mathrm{op}} \le c_1\sigma^2\gamma_{\min}^{-2}N/\rho_T$. Similarly, the restricted Wishart bound (17) for $D_2$ and the bound on $\|S\|_{\mathrm{op}}$ in (21) give $\|D_2S\widehat{\Lambda}_c^{-1}\|_{\mathrm{op}} \le c_1\sigma^2\gamma_{\min}^{-2}N/\rho_T$, thereby implying

$$\|S - G_2W\Lambda^{-1}\|_{\mathrm{op}} \le c_1\frac{\sigma^2\sqrt{T_{1,\mathrm{p}}N}}{\gamma^2_{\min}\rho_T} + c_1\frac{\sigma^2 N}{\gamma^2_{\min}\rho_T}.$$

It remains to convert this operator bound into a Euclidean norm bound. Let $G_2 = Q\Sigma_2V_2^\top$ be its singular value decomposition. Arguing as in the proof of Lemma 6, conditional on $\sigma(G_1,\Sigma_2,V_2)$, the factor $Q$ is Haar-distributed on the appropriate Stiefel manifold. Since $S = QR$ for an $\sigma(G_1,\Sigma_2,V_2)$-measurable matrix $R$ and $G_2W\Lambda^{-1} = Q\Sigma_2V_2^\top W\Lambda^{-1}$, we can write $S - G_2W\Lambda^{-1} = Q\widetilde{R}$, where $\widetilde{R}$ is $\sigma(G_1,\Sigma_2,V_2)$-measurable and has rank at most $r$. Now, let $\ell := \mathrm{rank}(\widetilde{R}) \le r$, and consider the compact SVD of $\widetilde{R} = L_{\widetilde{R}}D_{\widetilde{R}}M_{\widetilde{R}}^\top$

with $L_{\widetilde{R}}^\top L_{\widetilde{R}} = I_\ell$, where the factors may be chosen $\sigma(G_1, \Sigma_2, V_2)$-measurable. Conditional on $\sigma(G_1, \Sigma_2, V_2)$, the matrix $L_{\widetilde{R}}$ is fixed and $Q$ is Haar-distributed on $\mathrm{St}(N-r, q)$. As a result, Lemma 25, applied with $d = N - r$, $H = L_{\widetilde{R}}$, $A = \bar{x}^\top U_\perp$, and $t^2 \asymp \zeta_T$, gives $\|L_{\widetilde{R}}^\top Q^\top U_\perp^\top \bar{x}\|_2 \le c_1 \|\bar{x}^\top U_\perp\|_2 \sqrt{(r+\zeta_T)/(N-r)} \le c_1 \|\bar{x}^\top U_\perp\|_2 \sqrt{(r+\zeta_T)/N}$, where the last inequality follows from (A2). Finally, using $\|\bar{x}^\top U_\perp\|_2 \le \|\bar{x}\|_2 = 1$, we obtain

$$\|\widetilde{R}^\top Q^\top U_\perp^\top \bar{x}\|_2 = \|M_{\widetilde{R}} D_{\widetilde{R}} L_{\widetilde{R}}^\top Q^\top U_\perp^\top \bar{x}\|_2 \le \|D_{\widetilde{R}}\|_{\mathrm{op}} \|L_{\widetilde{R}}^\top Q^\top U_\perp^\top \bar{x}\|_2 = \|\widetilde{R}\|_{\mathrm{op}} \|L_{\widetilde{R}}^\top Q^\top U_\perp^\top \bar{x}\|_2 \le c_1 \sqrt{\frac{r+\zeta_T}{N}} \|\widetilde{R}\|_{\mathrm{op}}$$

with probability at least $1 - \mathcal{O}(p_T^{-10})$, which therefore implies

$$\|\{S - G_2 W \Lambda^{-1}\}^\top U_\perp^\top \bar{x}\|_2 \le c_1 \frac{\sigma^2 \sqrt{(N + T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}^2 \rho_T}.$$

This concludes the analysis for the first and more problematic term in (28). For the second one, using $\|C - I_r\|_{\mathrm{op}} \le \|S\|_{\mathrm{op}}^2$ and $\|S\|_{\mathrm{op}} \le c_1 \sigma \gamma_{\min}^{-1} \sqrt{N/\rho_T}$ from proof of Lemma 6, we get $\|(C - I_r) U^\top \bar{x}\|_2 \le c_1 \sigma^2 \gamma_{\min}^{-2} N \rho_T^{-1} \|U^\top \bar{x}\|_2 = c_1 \sigma^2 \gamma_{\min}^{-2} N \rho_T^{-1} \|U_{2k}^\top x\|_2$. For the third term, Lemma 21 gives $\|G_1 W\|_{\mathrm{op}} \le c_1 \sigma \|W\|_{\mathrm{op}} \sqrt{r + \zeta_T}$ with probability at least $1 - \mathcal{O}(p_T^{-10})$. Also, Lemma 5 gives $\|W\|_{\mathrm{op}} \le c_u^{1/2} \gamma_{\max} \rho_T^{1/2}$ and $\|\Lambda^{-1}\|_{\mathrm{op}} = \lambda_r(\Lambda)^{-1} = \sigma_r(W)^{-2} \le c_\ell^{-1} \gamma_{\min}^{-2} \rho_T^{-1}$. As a result, we have $\|G_1 W \Lambda^{-1}\|_{\mathrm{op}} \le \|G_1 W\|_{\mathrm{op}} \|\Lambda^{-1}\|_{\mathrm{op}} \le c_1 \sigma \sqrt{r+\zeta_T}\, \gamma_{\max} \rho_T^{1/2} (\gamma_{\min}^2 \rho_T)^{-1} \le c_1 \sigma \gamma_{\min}^{-1} \sqrt{(r+\zeta_T)/\rho_T}$, which further implies $\|\{G_1 W \Lambda^{-1}\}^\top U^\top \bar{x}\|_2 \le c_1 \sigma \gamma_{\min}^{-1} \sqrt{(r+\zeta_T)/\rho_T}\, \|U_{2k}^\top x\|_2$.

Combining the three bounds concludes the proof. □

As a sanity check, combining (26) and $\widehat{U}_{\mathrm{left}} H_U - U = \Psi_U + E_{\mathrm{left}}^{\mathrm{p}} W_{\mathrm{left}} (W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}$ with Lemmas 5 and 21 and Assumptions (A2), (A3), allows proving a bound for $\|(\widehat{U}_{\mathrm{left}} H_U - U)^\top x\|_2$ that agrees with (22).

Moreover, by applying Lemmas 6 and 8 to $Y_{\mathrm{up}}^\top$ we get the following corollary.

**Corollary 9.** *Grant Assumptions (A1) with fixed constants $0 < c_\ell \le c_u$, (A2) and (A3). Suppose further that $0 < \gamma_{\min} \le \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \le \gamma_{\max} < \infty$ for all $j \in [K]$, and let $\kappa := \gamma_{\max}/\gamma_{\min}$. Write $Y_{\mathrm{up}}^{\mathrm{p}} = M_{\mathrm{up}}^{\mathrm{p}} + E_{\mathrm{up}}^{\mathrm{p}}$, with $M_{\mathrm{up}}^{\mathrm{p}} = W_{\mathrm{up}} V^\top$. Let $(\hat{U}_{\mathrm{up}}, \hat{\Sigma}_{\mathrm{up}}, \hat{V}_{\mathrm{up}}) := \mathrm{SVD}_r(Y_{\mathrm{up}}^{\mathrm{p}})$ and $H_V := \mathrm{sgn}(\hat{V}_{\mathrm{up}}^\top V)$, and define $\Psi_V := \widehat{V}_{\mathrm{up}} H_V - V - (E_{\mathrm{up}}^{\mathrm{p}})^\top W_{\mathrm{up}} (W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}$. For fixed $k \in [K]$ we also set $\Psi_{V,2k} := (\Psi_V)_{\mathcal{J}_k,\bullet}$, where $\mathcal{J}_k = \{T_{1k}+1, \ldots, T\}$. Fix also $\bar{y} \in \mathbb{B}_2(T)$ and $y \in \mathbb{B}_2(T_{2k})$. There exists a constant $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ such that, with probability at least $1 - \mathcal{O}(p_N^{-10})$, the following statements hold:*

*(i) We have*

$$\|(\hat{V}_{\mathrm{up}} H_V - V)^\top \bar{y}\|_2 \le c_1 \frac{\sigma \sqrt{r+\zeta_N}}{\gamma_{\min}\, \rho_N^{1/2}} + c_1 \frac{\sigma^2 T}{\gamma_{\min}^2\, \rho_N} \|V^\top \bar{y}\|_2. \tag{30}$$

*(ii) We can bound the operator norm as*

$$\|\hat{V}_{\mathrm{up}} H_V - V\|_{\mathrm{op}} \le c_1 \frac{\sigma \sqrt{T}}{\gamma_{\min}\, \rho_N^{1/2}}. \tag{31}$$

*(iii) We have*

$$\|\Psi_{V,2k}^\top y\|_2 \le c_1 \frac{\sigma^2\sqrt{(T+N_{1,\mathrm{p}})(r+\zeta_N)}}{\gamma_{\min}^2\rho_N} + c_1\left(\frac{\sigma^2 T}{\gamma_{\min}^2\rho_N} + \frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\rho_N^{1/2}}\right)\|V_{2k}^\top y\|_2. \tag{32}$$

*Proof.* The proof follows by applying Lemmas 6 and 8 to $(Y_{\mathrm{up}}^{\mathrm{p}})^\top = VW_{\mathrm{up}}^\top + (E_{\mathrm{up}}^{\mathrm{p}})^\top$. In particular, (30) follows from (22), (31) from (9) with $\Pi_T = I_T$ and (A3), and (32) from (26). □

Returning to theoretical guarantees for quantities obtained from the SVD of $Y_{\mathrm{left}}$, the following result controls the estimation error for $(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top$, a key object for predicting the $c$-block from the $a$-block.

**Lemma 10.** *Adopt the assumptions and notation of Lemma 6. Fix $k \in [K]$, write $H_k := U_{1k}^\top U_{1k}$ and $\widehat{H}_k := (\hat{U}_{1k}H_U)^\top \hat{U}_{1k}H_U$, and set $D_k := \widehat{H}_k^{-1}(\hat{U}_{1k}H_U)^\top - H_k^{-1}U_{1k}^\top$. There exists $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ such that, with probability at least $1 - \mathcal{O}(p_T^{-10})$, we have*

$$\|D_k\|_{\mathrm{op}} \le c_1 \frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}}\sqrt{\frac{N}{N_{1k}}}. \tag{33}$$

*Proof.* We showed in (23) and the discussion thereafter that $\hat{U}_{1k}H_U$ has full column rank on the high-probability event of Lemma 6. On this event, letting $\Delta_k := \hat{U}_{1k}H_U - U_{1k}$ and using $D_k = \widehat{H}_k^{-1}(\hat{U}_{1k}H_U)^\top - H_k^{-1}U_{1k}^\top = (\hat{U}_{1k}H_U)^\dagger - U_{1k}^\dagger$, Lemma 24 gives

$$D_k = -U_{1k}^\dagger\Delta_k(\hat{U}_{1k}H_U)^\dagger + U_{1k}^\dagger(U_{1k}^\dagger)^\top\Delta_k^\top\{I_{N_{1k}} - (\hat{U}_{1k}H_U)(\hat{U}_{1k}H_U)^\dagger\}.$$

As a result, we have

$$\|D_k\|_{\mathrm{op}} \le \|U_{1k}^\dagger\|_{\mathrm{op}}\|\Delta_k\|_{\mathrm{op}}\|(\hat{U}_{1k}H_U)^\dagger\|_{\mathrm{op}} + \|U_{1k}^\dagger\|_{\mathrm{op}}^2\|\Delta_k\|_{\mathrm{op}} \le (\sqrt{2}+1)c_\ell^{-1}\frac{N}{N_{1k}}\|\Delta_k\|_{\mathrm{op}} \le c_1\frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}}\sqrt{\frac{N}{N_{1k}}}, \tag{34}$$

where the penultimate inequality follows from (A1) and $\frac{c_\ell}{2}\frac{N_{1k}}{N}I_r \preceq \hat{U}_{1k}^\top\hat{U}_{1k} \preceq 2c_u\frac{N_{1k}}{N}I_r$, while the last one follows from the discussion right after (23). This completes the proof. □

**Lemma 11.** *Grant the assumptions of Lemmas 6. Fix $k \in [K]$, and recall $H_k = U_{1k}^\top U_{1k}$, $\hat{H}_k = H_U^\top\hat{U}_{1k}^\top\hat{U}_{1k}H_U$. On the high probability event where $\hat{H}_k$ is invertible, define $L := \hat{U}_{2k}H_U\hat{H}_k^{-1}H_U^\top\hat{U}_{1k}^\top - U_{2k}H_k^{-1}U_{1k}^\top$. Writing*

$$\Delta_L := L\,U_{1k}\mathcal{C}_{\bullet,\bullet,k} - (E_{\mathrm{left}}^{\mathrm{p}})_{\{N_{1k}+1,\ldots,N\},\bullet}W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}\mathcal{C}_{\bullet,\bullet,k}, \tag{35}$$

*for every fixed $x \in \mathbb{B}_2(N_{2k})$ there exists $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ such that, with probability at least $1 - \mathcal{O}(p_T^{-10})$, we have*

$$\|x^\top\Delta_L\|_2 \le c_1\frac{\sigma^2\sqrt{(N+T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}\rho_T} + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\|U_{2k}^\top x\|_2. \tag{36}$$

*Proof.* Using $H_k^{-1}U_{1k}^\top U_{1k} = I_r$, by definition of $L$, we have

$$\begin{aligned} LU_{1k}\mathcal{C}_{\bullet,\bullet,k} &= \hat{U}_{2k}H_U\hat{H}_k^{-1}H_U^\top\hat{U}_{1k}^\top U_{1k}\mathcal{C}_{\bullet,\bullet,k} - U_{2k}H_k^{-1}U_{1k}^\top U_{1k}\mathcal{C}_{\bullet,\bullet,k} \\ &= \big(\hat{U}_{2k}H_U - U_{2k}\big)\mathcal{C}_{\bullet,\bullet,k} + \hat{U}_{2k}H_U\left(\hat{H}_k^{-1}H_U^\top\hat{U}_{1k}^\top - H_k^{-1}U_{1k}^\top\right)U_{1k}\mathcal{C}_{\bullet,\bullet,k}, \end{aligned}$$

Also, the definition of $\Psi_{U,2k}$ in Corollary 8 gives $\hat{U}_{2k}H_U - U_{2k} = (E^{\mathrm{p}}_{\mathrm{left}})_{\{N_{1k}+1,\dots,N\},\bullet}W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1} + \Psi_{U,2k}$. Substituting this identity into the above display and using the definition of $\Delta_L$ in (35) give $\Delta_L = \Psi_{U,2k}\mathcal{C}_{\bullet,\bullet,k} + \hat{U}_{2k}H_U(\hat{H}_k^{-1}H_U^\top\hat{U}_{1k}^\top - H_k^{-1}U_{1k}^\top)U_{1k}\mathcal{C}_{\bullet,\bullet,k}$.

We will use this to control the Euclidean norm of $x^\top\Delta_L$ by bounding the norm of each of the two term separately. For the first one, combining (26) with $\|\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}} \le \gamma_{\max} = \kappa\gamma_{\min}$ gives

$$\begin{aligned} \|x^\top\Psi_{U,2k}\mathcal{C}_{\bullet,\bullet,k}\|_2 &\le \|\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}}\,\|\Psi_{U,2k}^\top x\|_2 \\ &\le c_1\frac{\sigma^2\sqrt{(N+T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}\rho_T} + c_1\left(\frac{\sigma^2 N}{\gamma_{\min}\,\rho_T} + \frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\right)\|U_{2k}^\top x\|_2. \end{aligned}$$

Second, recalling $D_k = \widehat{H}_k^{-1}(\hat{U}_{1k}H_U)^\top - H_k^{-1}U_{1k}^\top$ from Lemma 10, we have

$$\begin{aligned} \|x^\top\hat{U}_{2k}H_U\,D_k\,U_{1k}\mathcal{C}_{\bullet,\bullet,k}\|_2 &\le \|H_U^\top\hat{U}_{2k}^\top x\|_2\,\|\,D_k\,U_{1k}\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}} \\ &\le \Big(\|(\hat{U}_{2k}H_U - U_{2k})^\top x\|_2 + \|U_{2k}^\top x\|_2\Big)\,\|\,D_k\,U_{1k}\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}} \\ &\le c_1\left(\frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + \frac{\sigma^2 N}{\gamma_{\min}^2\,\rho_T}\|U_{2k}^\top x\|_2 + \|U_{2k}^\top x\|_2\right)\|\,D_k\,U_{1k}\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}} \\ &\le c_1\left(\frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + \|U_{2k}^\top x\|_2\right)\|\,D_k\,U_{1k}\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}} \\ &\le c_1\left(\frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\,\rho_T^{1/2}} + \|U_{2k}^\top x\|_2\right)\|\,D_k\|_{\mathrm{op}}\,\|U_{1k}\|_{\mathrm{op}}\,\|\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}} \\ &\le c_1\frac{\sigma^2\sqrt{N(r+\zeta_T)}}{\gamma_{\min}\,\rho_T} + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\,\|U_{2k}^\top x\|_2, \end{aligned}$$

where the third inequality follows from (22), and the last one from (33), (A1) and $\|\mathcal{C}_{\bullet,\bullet,k}\|_{\mathrm{op}} \le \gamma_{\max} = \kappa\gamma_{\min}$. Combining the two displays and using (A2), (A3) concludes the proof. □

**Lemma 12.** *Grant Assumptions (A1) with fixed constants $0 < c_\ell \le c_u$, (A2) and (A3). Suppose that $0 < \gamma_{\min} \le \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \le \gamma_{\max} < \infty$ for all $j \in [K]$. Write $Y^{\mathrm{p}}_{\mathrm{up}} = M^{\mathrm{p}}_{\mathrm{up}} + E^{\mathrm{p}}_{\mathrm{up}}$, where $M^{\mathrm{p}}_{\mathrm{up}} = W_{\mathrm{up}}V^\top$. Also recall $(\widehat{U}_{\mathrm{up}}, \widehat{\Sigma}_{\mathrm{up}}, \widehat{V}_{\mathrm{up}}) = \mathrm{SVD}_r(Y^{\mathrm{p}}_{\mathrm{up}})$, $H_V = \mathrm{sgn}(\widehat{V}_{\mathrm{up}}^\top V)$. Writing*

$$\Phi_{\mathrm{up}} := \widehat{U}_{\mathrm{up}}\widehat{\Sigma}_{\mathrm{up}}\widehat{V}_{\mathrm{up}}^\top - M^{\mathrm{p}}_{\mathrm{up}} - E^{\mathrm{p}}_{\mathrm{up}}VV^\top - W_{\mathrm{up}}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top E^{\mathrm{p}}_{\mathrm{up}}, \tag{37}$$

*for every fixed* $g \in \mathbb{R}^{N_{1,\mathrm{p}}}$ *and* $y \in \mathbb{B}_2(T)$ *there exists a constant* $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ *such that*

$$|g^\top \Phi_{\mathrm{up}}\, y| \le c_1 \frac{\sigma^2\sqrt{T(r+\zeta_N)}}{\gamma_{\min}\,\rho_N^{1/2}}\,\|g\|_2 + c_1\,\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}}\|W_{\mathrm{up}}^\top g\|_2 + c_1\frac{\sigma^2 T}{\gamma_{\min}\,\rho_N^{1/2}}\,\|g\|_2\,\|V^\top y\|_2 + c_1\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}\|W_{\mathrm{up}}^\top g\|_2\,\|V^\top y\|_2 \tag{38}$$

*with probability at least* $1 - \mathcal{O}(p_N^{-10})$.

*Proof.* Write $\widehat{M}_{\mathrm{up}}^{\mathrm{p}} := \widehat{U}_{\mathrm{up}}\widehat{\Sigma}_{\mathrm{up}}\widehat{V}_{\mathrm{up}}^\top$, $\delta_V := \widehat{V}_{\mathrm{up}}H_V - V$, and $\Psi_V := \delta_V - (E_{\mathrm{up}}^{\mathrm{p}})^\top W_{\mathrm{up}}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}$. Combining this with $\widehat{U}_{\mathrm{up}}\widehat{\Sigma}_{\mathrm{up}}H_V = Y_{\mathrm{up}}^{\mathrm{p}}(V+\delta_V)$ gives

$$\begin{aligned}
\widehat{M}_{\mathrm{up}}^{\mathrm{p}} - M_{\mathrm{up}}^{\mathrm{p}} &= Y_{\mathrm{up}}^{\mathrm{p}}(V+\delta_V)(V+\delta_V)^\top - W_{\mathrm{up}}V^\top = (W_{\mathrm{up}}V^\top + E_{\mathrm{up}}^{\mathrm{p}})(V+\delta_V)(V+\delta_V)^\top - W_{\mathrm{up}}V^\top \\
&= E_{\mathrm{up}}^{\mathrm{p}}VV^\top + E_{\mathrm{up}}^{\mathrm{p}}V\delta_V^\top + Y_{\mathrm{up}}^{\mathrm{p}}\delta_V V^\top + Y_{\mathrm{up}}^{\mathrm{p}}\delta_V\delta_V^\top + W_{\mathrm{up}}\delta_V^\top \\
&= E_{\mathrm{up}}^{\mathrm{p}}VV^\top + W_{\mathrm{up}}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top E_{\mathrm{up}}^{\mathrm{p}} + E_{\mathrm{up}}^{\mathrm{p}}V\delta_V^\top + Y_{\mathrm{up}}^{\mathrm{p}}\delta_V V^\top + Y_{\mathrm{up}}^{\mathrm{p}}\delta_V\delta_V^\top + W_{\mathrm{up}}\Psi_V^\top.
\end{aligned}$$

This, together with the definition of $\Phi_{\mathrm{up}}$ in (37), gives $g^\top\Phi_{\mathrm{up}}y = g^\top E_{\mathrm{up}}^{\mathrm{p}}V\delta_V^\top y + g^\top Y_{\mathrm{up}}^{\mathrm{p}}\delta_V V^\top y + g^\top Y_{\mathrm{up}}^{\mathrm{p}}\delta_V\delta_V^\top y + g^\top W_{\mathrm{up}}\Psi_V^\top y$.

We now bound each of the four terms in $g^\top\Phi_{\mathrm{up}}\,y$ individually. By (30) and (31) in Corollary 9, with probability at least $1-\mathcal{O}(p_N^{-10})$ we have

$$\|\delta_V^\top y\|_2 \le c_1\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}} + c_1\frac{\sigma^2 T}{\gamma_{\min}^2\,\rho_N}\|V^\top y\|_2, \qquad \|\delta_V\|_{\mathrm{op}} \le c_1\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}. \tag{39}$$

Combining the first bound with Lemma 21 and (A2) gives

$$\begin{aligned}
|g^\top E_{\mathrm{up}}^{\mathrm{p}}V\delta_V^\top y| &\le \|(E_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2\|\delta_V^\top y\|_2 \le c_1\sigma\sqrt{T+\zeta_N}\,\|g\|_2\,\|\delta_V^\top y\|_2 \\
&\le c_1\frac{\sigma^2\sqrt{T(r+\zeta_N)}}{\gamma_{\min}\,\rho_N^{1/2}}\,\|g\|_2 + c_1\frac{\sigma^3 T^{3/2}}{\gamma_{\min}^2\,\rho_N}\,\|g\|_2\,\|V^\top y\|_2.
\end{aligned}$$

Next, using $\|(Y_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2 \le \|(M_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2 + \|(E_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2 \le \|W_{\mathrm{up}}^\top g\|_2 + c_1\sigma\sqrt{T}\,\|g\|_2$, we also get

$$\begin{aligned}
|g^\top Y_{\mathrm{up}}^{\mathrm{p}}\delta_V V^\top y| &\le \|(Y_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2\,\|\delta_V\|_{\mathrm{op}}\,\|V^\top y\|_2 \le c_1\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}\|(Y_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2\,\|V^\top y\|_2 \\
&\le c_1\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}\left(\|W_{\mathrm{up}}^\top g\|_2 + \sigma\sqrt{T}\,\|g\|_2\right)\,\|V^\top y\|_2 \\
&\le c_1\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}\|W_{\mathrm{up}}^\top g\|_2\,\|V^\top y\|_2 + c_1\frac{\sigma^2 T}{\gamma_{\min}\,\rho_N^{1/2}}\,\|g\|_2\,\|V^\top y\|_2.
\end{aligned}$$

Similarly, for the third term we have

$$|g^\top Y_{\mathrm{up}}^{\mathrm{p}}\delta_V\delta_V^\top y| \le \|(Y_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2\,\|\delta_V\|_{\mathrm{op}}\,\|\delta_V^\top y\|_2 \le c_1\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}\|(Y_{\mathrm{up}}^{\mathrm{p}})^\top g\|_2\,\|\delta_V^\top y\|_2$$

$$
\begin{aligned}
&\le c_1 \frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}} \left( \|W_{\mathrm{up}}^\top g\|_2 + \sigma\sqrt{T}\,\|g\|_2 \right) \|\delta_V^\top y\|_2 \\
&\le c_1 \frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}} \left( \|W_{\mathrm{up}}^\top g\|_2 + \sigma\sqrt{T}\,\|g\|_2 \right) \left( \frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}} + \frac{\sigma^2 T}{\gamma_{\min}^2\,\rho_N} \|V^\top y\|_2 \right) \\
&\le c_1 \frac{\sigma^2\sqrt{T(r+\zeta_N)}}{\gamma_{\min}^2\,\rho_N} \|W_{\mathrm{up}}^\top g\|_2 + c_1 \frac{\sigma^3 T\sqrt{r+\zeta_N}}{\gamma_{\min}^2\,\rho_N} \|g\|_2 + c_1 \frac{\sigma^3 T^{3/2}}{\gamma_{\min}^3\,\rho_N^{3/2}} \|W_{\mathrm{up}}^\top g\|_2\, \|V^\top y\|_2 \\
&\qquad + c_1 \frac{\sigma^4 T^2}{\gamma_{\min}^3\,\rho_N^{3/2}} \|g\|_2\, \|V^\top y\|_2 .
\end{aligned}
$$

It remains to control the term involving $\Psi_V$. Using the definition of $\Psi_V$ and Lemmas 5 and 21 we can write

$$
\begin{aligned}
|g^\top W_{\mathrm{up}} \Psi_V^\top y| &\le |g^\top W_{\mathrm{up}} \delta_V^\top y| + |g^\top W_{\mathrm{up}} (W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1} W_{\mathrm{up}}^\top E_{\mathrm{up}}^{\mathrm{p}}\, y| \\
&\le \|W_{\mathrm{up}}^\top g\|_2 \, \left( \|\delta_V^\top y\|_2 + \|(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}\|_{\mathrm{op}}\, \|W_{\mathrm{up}}^\top E_{\mathrm{up}}^{\mathrm{p}}\, y\|_2 \right) \\
&\le c_1 \|W_{\mathrm{up}}^\top g\|_2 \left( \frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}} + c_1 \frac{\sigma^2 T}{\gamma_{\min}^2\,\rho_N} \|V^\top y\|_2 + \frac{1}{\gamma_{\min}^2 \rho_N}\, \sigma \|W_{\mathrm{up}}\|_{\mathrm{op}} \sqrt{r+\zeta_N} \right) \\
&\le c_1 \frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}} \|W_{\mathrm{up}}^\top g\|_2 + c_1 \frac{\sigma^2 T}{\gamma_{\min}^2\,\rho_N}\, \|W_{\mathrm{up}}^\top g\|_2\, \|V^\top y\|_2 .
\end{aligned}
$$

Combining the four preceding bounds and simplifying them further using (A2), (A3) proves (38). □

Lemma 12 provides an upper bound on the approximation error of the entire pooled upper matrix $M_{\mathrm{up}}^{\mathrm{p}} = W_{\mathrm{up}} V^\top$. By restricting $\Phi_{\mathrm{up}}$ in (37) to the subsets $\mathcal{I}_k^{\mathrm{up}} = \{s_k + 1, \dots, s_k + N_{1k}\}$ and $\mathcal{J}_k = \{T_{1k} + 1, \dots, T\}$, computations similar to those in the previous proof allow us to quantify the approximation error of $\mathcal{M}_{\bullet,\bullet,k}^{(b)}$.

**Corollary 13.** *Suppose the assumptions of Lemma 12 are satisfied, and use the notation introduced there. Also define* $\hat{M}_b^{(k)} := (\widehat{U}_{\mathrm{up}} \widehat{\Sigma}_{\mathrm{up}} \widehat{V}_{\mathrm{up}}^\top)_{\mathcal{I}_k^{\mathrm{up}}, \mathcal{J}_k}$ *and recall* $\mathcal{M}_{\bullet,\bullet,k}^{(b)} = (M_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}}, \mathcal{J}_k} = U_{1k} \mathcal{C}_{\bullet,\bullet,k}\, V_{2k}^\top$. *Writing*

$$
\Phi_k := \widehat{M}_b^{(k)} - \mathcal{M}_{\bullet,\bullet,k}^{(b)} - (E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\, V V_{2k}^\top - U_{1k}\, \mathcal{C}_{\bullet,\bullet,k}\, (W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1} W_{\mathrm{up}}^\top (E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}, \tag{40}
$$

*for fixed* $x \in \mathbb{B}_2(N_{2k})$, $y \in \mathbb{B}_2(T_{2k})$ *there exists a constant* $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$ *such that*

$$
|x^\top U_{2k} (U_{1k}^\top U_{1k})^{-1} U_{1k}^\top\, \Phi_k\, y| \le c_1\, \frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}} \|U_{2k}^\top x\|_2 + c_1 \frac{\sigma\sqrt{T}}{\rho_N^{1/2}}\, \|U_{2k}^\top x\|_2\, \|V_{2k}^\top y\|_2, \tag{41}
$$

$$
\left\| (U_{1k}^\top U_{1k})^{-1/2}\, U_{1k}^\top\, \Phi_k\, y \right\|_2 \le c_1 \frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}} \sqrt{\frac{N_{1k}}{N}} + c_1 \frac{\sigma\sqrt{T}}{\rho_N^{1/2}} \sqrt{\frac{N_{1k}}{N}} \|V_{2k}^\top y\|_2, \tag{42}
$$

$$
\|(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1}\, U_{1k}^\top)\, \Phi_k\, y\|_2 \le c_1 \frac{\sigma^3 T\sqrt{r+\zeta_N}}{\gamma_{\min}^2 \rho_N} + c_1 \frac{\sigma^2\sqrt{N_{1k}(r+\zeta_N)}}{\gamma_{\min}\,\rho_N^{1/2}} + c_1 \frac{\sigma^2\sqrt{T(N_{1k}+T)}}{\gamma_{\min}\,\rho_N^{1/2}} \|V_{2k}^\top y\|_2 \tag{43}
$$

*with probability at least* $1 - O(p_N^{-10})$.

*Proof.* Restricting (37) to the subsets $\mathcal{I}_k^{\mathrm{up}}$ and $\mathcal{J}_k$ and using the definition of $\Phi_k$ in (40) immediately yield $\Phi_k = (\Phi_{\mathrm{up}})_{\mathcal{I}_k^{\mathrm{up}},\mathcal{J}_k}$. In order to prove (41), set $B_k := U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top$ and define $g \in \mathbb{R}^{N_{1,\mathrm{p}}}$ and $\bar{y} \in \mathbb{B}_2(T)$ to be the vectors with entries $g_i = (B_k^\top x)_{i-s_k}\,\mathbb{1}\{i \in \mathcal{I}_k^{\mathrm{up}}\}$ and $\bar{y}_t = y_{t-T_{1k}}\,\mathbb{1}\{t \in \mathcal{J}_k\}$, respectively. This ensures that $V^\top \bar{y} = V_{2k}^\top y$ and that $|g^\top \Phi_{\mathrm{up}} \bar{y}|$ is equal to the left-hand side of (41). Furthermore, from (A1) we have $\|g\|_2 = \|B_k^\top x\|_2 = \|U_{1k}(U_{1k}^\top U_{1k})^{-1}U_{2k}^\top x\|_2 \le \|U_{1k}(U_{1k}^\top U_{1k})^{-1}\|_{\mathrm{op}}\|U_{2k}^\top x\|_2 \le c_\ell^{-1/2}\sqrt{N/N_{1k}}\,\|U_{2k}^\top x\|_2$, and $\|W_{\mathrm{up}}^\top g\|_2 = \|\mathcal{C}_{\bullet,\bullet,k}^\top U_{1k}^\top B_k^\top x\|_2 = \|\mathcal{C}_{\bullet,\bullet,k}^\top U_{2k}^\top x\|_2 \le \gamma_{\max}\|U_{2k}^\top x\|_2$. Combining these with (38) and further simplifying the resulting bound using (A2), (A3) proves (41).

It remains to prove the last two bounds. Using the expression for $\Phi_{\mathrm{up}}$ from the proof of Lemma 12 we get

$$\Phi_k\, y = (E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet} V \delta_V^\top \bar{y} + (Y_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\delta_V V^\top\,\bar{y} + (Y_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\delta_V\delta_V^\top\bar{y} + U_{1k}\mathcal{C}_{\bullet,\bullet,k}\,\Psi_V^\top\,\bar{y}. \tag{44}$$

We next bound the norms of the four terms separately under the action of $(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top$. As for the first one, using Lemma 21 and the bounds for $\|\delta_V^\top\bar{y}\|_2$ and $\|\delta_V\|_{\mathrm{op}}$ in (39), we get

$$\begin{aligned}\|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}V\delta_V^\top\bar{y}\|_2 &\le \|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}V\|_{\mathrm{op}}\,\|\delta_V^\top\bar{y}\|_2 \le c_1\sigma\sqrt{r+\zeta_N}\,\|\delta_V^\top\bar{y}\|_2\\ &\le c_1\sigma\sqrt{r+\zeta_N}\left(\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}} + \frac{\sigma^2 T}{\gamma_{\min}^2\,\rho_N}\,\|V_{2k}^\top y\|_2\right).\end{aligned}$$

Similarly, for the second and third terms in (44) we have

$$\begin{aligned}&\|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top\,(Y_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\,\delta_V\,(V^\top\bar{y}+\delta_V^\top\bar{y})\|_2\\ &\qquad\le \|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top\,\{(M_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet} + (E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\}\,\delta_V\,(V^\top\bar{y}+\delta_V^\top\bar{y})\|_2\\ &\qquad\le \|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top\,\{(M_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet} + (E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\}\|_{\mathrm{op}}\|\delta_V\|_{\mathrm{op}}\|V^\top\bar{y}+\delta_V^\top\bar{y}\|_2\\ &\qquad= \|(U_{1k}^\top U_{1k})^{1/2}\mathcal{C}_{\bullet,\bullet,k}\,V^\top + (U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top E_{\mathrm{up}}^{(k)}\|_{\mathrm{op}}\|\delta_V\|_{\mathrm{op}}\|V^\top\bar{y}+\delta_V^\top\bar{y}\|_2\\ &\qquad\le c_1\left(\kappa\,\gamma_{\min}\sqrt{\frac{N_{1k}}{N}} + \sigma\sqrt{T}\right)\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}\left(\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\rho_N^{1/2}} + \|V_{2k}^\top y\|_2\right).\end{aligned}$$

Finally, using (32) in Corollary 9 to bound $\|\Psi_V^\top\bar{y}\|_2 = \|\Psi_{V,2k}^\top\,y\|_2$, with probability at least $1 - \mathcal{O}(p_N^{-10})$ we have

$$\begin{aligned}\|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top\,U_{1k}\mathcal{C}_{\bullet,\bullet,k}\,\Psi_V^\top\,\bar{y}\|_2 &\le c_1\gamma_{\max}\sqrt{\frac{N_{1k}}{N}}\,\|\Psi_V^\top\bar{y}\|_2 = c_1\kappa\gamma_{\min}\sqrt{\frac{N_{1k}}{N}}\,\|\Psi_{V,2k}^\top\,y\|_2\\ &\le c_1\gamma_{\min}\sqrt{\frac{N_{1k}}{N}}\left\{\frac{\sigma^2\sqrt{(T+N_{1,\mathrm{p}})(r+\zeta_N)}}{\gamma_{\min}^2\rho_N} + \left(\frac{\sigma^2 T}{\gamma_{\min}^2\rho_N} + \frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\rho_N^{1/2}}\right)\|V_{2k}^\top y\|_2\right\}.\end{aligned}$$

Combining the last three displays and further simplifying the bound using (A2), (A3) proves (42).

In order to prove (43), we will make use of the fact that $I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1}\,U_{1k}^\top$ is the orthogonal projector onto the orthogonal complement of $\mathrm{col}(U_{1k})$. This also implies that $\|I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1}\,U_{1k}^\top\|_{\mathrm{op}} = 1$ and $\mathrm{rank}(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1}\,U_{1k}^\top) = N_{1k} - r$. This implies that the signal contribution from the second

and third terms in (44) vanishes, and we are left the error matrix only. More precisely, we have

$$\begin{aligned}
&\|(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1} U_{1k}^\top)\,(Y_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\,\delta_V\,(V^\top \bar{y} + \delta_V^\top \bar{y})\|_2 \\
&\qquad = \|(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1} U_{1k}^\top)\,(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\,\delta_V\,(V^\top \bar{y} + \delta_V^\top \bar{y})\|_2 \\
&\qquad = \|(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1} U_{1k}^\top)\,E_{\mathrm{up}}^{(k)}\,\delta_V\,(V^\top \bar{y} + \delta_V^\top \bar{y})\|_2 \\
&\qquad \le \|(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1} U_{1k}^\top)\,E_{\mathrm{up}}^{(k)}\|_{\mathrm{op}}\,\|\delta_V\|_{\mathrm{op}}\,\|V^\top \bar{y} + \delta_V^\top \bar{y}\|_2 \\
&\qquad \le c_1\sigma\sqrt{T + (N_{1k} - r) + \zeta_N}\;\|\delta_V\|_{\mathrm{op}}\,\|V^\top \bar{y} + \delta_V^\top \bar{y}\|_2 \\
&\qquad \le c_1\sigma\sqrt{N_{1k} + T}\,\frac{\sigma\sqrt{T}}{\gamma_{\min}\,\rho_N^{1/2}}\left(\frac{\sigma\sqrt{r + \zeta_N}}{\gamma_{\min}\rho_N^{1/2}} + \|V_{2k}^\top y\|_2\right).
\end{aligned}$$

The fourth term in (44) completely vanishes under the action of $I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1} U_{1k}^\top$, while the first one gives

$$\begin{aligned}
\|(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1} U_{1k}^\top)\,(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet} V \delta_V^\top \bar{y}\|_2 &\le \|(I_{N_{1k}} - U_{1k}\{U_{1k}^\top U_{1k}\}^{-1} U_{1k}^\top)\,(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet} V\|_{\mathrm{op}}\,\|\delta_V^\top \bar{y}\|_2 \\
&\le c_1\sigma\sqrt{r + (N_{1k} - r) + \zeta_N}\;\|\delta_V^\top \bar{y}\|_2 \le c_1\sigma\sqrt{N_{1k}}\left(\frac{\sigma\sqrt{r + \zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}} + \frac{\sigma^2 T}{\gamma_{\min}^2\,\rho_N}\,\|V_{2k}^\top y\|_2\right).
\end{aligned}$$

Combining the last three displays and further simplifying the bound using (A2), (A3) proves (43). This concludes the proof. □

We now present the main results of this section, which give a first-order expansion of $\hat{\mu}_{xy}^{(k)} - \mu_{xy}^{(k)}$ as the sum of Gaussian terms and remainder terms that can be suitably bounded from above. In the special case $K = 1$ with $x = \boldsymbol{e}_i$ and $y = \boldsymbol{e}_t$, analogous expansions were proved in Yan and Wainwright (2024). Our results therefore generalise these earlier expansions to arbitrary $K \ge 1$, $x \in \mathbb{B}_2(N_{2k})$, and $y \in \mathbb{B}_2(T_{2k})$, while also providing remainder bounds obtained via different techniques. In particular, the approach in Yan and Wainwright (2024) relies on a leave-one-block-out argument, which could also be adapted to the present setting. However, this approach becomes suboptimal when $K$ grows: in particular, the resulting remainder term is negligible only under a signal-to-noise ratio condition that deteriorates with $K$. We therefore instead rely on the preceding lemmas, which yield analogous results under weaker conditions.

**Lemma 14.** *Grant assumption (A1) with fixed constants $c_\ell, c_u$ satisfying $0 < c_\ell \le c_u < \infty$, (A2) and (A3). Suppose further that $0 < \gamma_{\min} \le \sigma_{\min}(\mathcal{C}_{\bullet,\bullet,j}) \le \sigma_{\max}(\mathcal{C}_{\bullet,\bullet,j}) \le \gamma_{\max} < \infty$ for all $j \in [K]$, and let $\kappa := \gamma_{\max}/\gamma_{\min}$. Fix $k \in [K]$, unit vectors $x \in \mathbb{B}_2(N_{2k})$, $y \in \mathbb{B}_2(T_{2k})$, and let $\hat{\mu}_{xy}^{(k)}$ be the output of Algorithm 1 run with $\tau \le \frac{c_\ell N_{1k}}{2N}$. Also write the decomposition $\hat{\mu}_{xy}^{(k)} - \mu_{xy}^{(k)} = Z_{xy}^{(1)} + Z_{xy}^{(2)} + Z_{xy}^{(3)} + Z_{xy}^{(4)} + \Delta_{xy} =: Z_{xy} + \Delta_{xy}$, where*

$$\begin{aligned}
Z_{xy}^{(1)} &:= x^\top\,(E_{\mathrm{left}}^{\mathrm{p}})_{\mathcal{I}_k,\bullet}\,W_{\mathrm{left}}\,(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}\mathcal{C}_{\bullet,\bullet,k}\,(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1} W_{\mathrm{up}}^\top (E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}\;y \\
Z_{xy}^{(2)} &:= x^\top\,U_{2k}\,\mathcal{C}_{\bullet,\bullet,k}\,(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1} W_{\mathrm{up}}^\top\,(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}\;y \\
Z_{xy}^{(3)} &:= x^\top\,U_{2k}(U_{1k}^\top U_{1k})^{-1}\,U_{1k}^\top (E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\,V\,V_{2k}^\top\;y \\
Z_{xy}^{(4)} &:= x^\top\,(E_{\mathrm{left}}^{\mathrm{p}})_{\mathcal{I}_k,\bullet}\;W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}\mathcal{C}_{\bullet,\bullet,k}\,V_{2k}^\top\;y.
\end{aligned} \tag{45}$$

*There exists an event $\mathcal{G}_1$ with $\mathbb{P}(\mathcal{G}_1) \geq 1 - \mathcal{O}(p_N^{-10} + p_T^{-10})$ such that, under $\mathcal{G}_1$, the remainder satisfies*

$$|\Delta_{xy}| \leq c_1 \frac{\sigma^2\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}} + c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\|V_{2k}^\top y\|_2 + c_1\frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2 + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2 + c_1\frac{\sigma\sqrt{T}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2 \tag{46}$$

*for a sufficiently large constant $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\mathrm{blk}}, \kappa) > 0$.*

*Proof.* Let $\mathcal{G}_1$ be the intersection of the high-probability events in Lemmas 6, 10, 11, 12, 21 and Corollaries 8, 9, 13, applied with the specific deterministic choices of projection matrices and vectors used below. By a union bound, we have $\mathbb{P}(\mathcal{G}_1) \geq 1 - \mathcal{O}(p_N^{-10} + p_T^{-10})$. In particular, arguing as in (24), we know that, under $\mathcal{G}_1$, the matrix $\hat{H}^{\mathrm{inv}}_{k,\tau}$ used in Algorithm 1 coincides with $(\hat{U}_{1k}^\top\hat{U}_{1k})^{-1}$ whenever the algorithm is run with $\tau \leq \frac{c_\ell N_{1k}}{2N}$. Recalling the notation $H_k = U_{1k}^\top U_{1k}$, $\hat{H}_k = H_U^\top\hat{U}_{1k}^\top\hat{U}_{1k}H_U$, $L = \hat{U}_{2k}H_U\hat{H}_k^{-1}H_U^\top\hat{U}_{1k}^\top - U_{2k}H_k^{-1}U_{1k}^\top$, and using the expression (35) and (40) for $\Delta_L$ and $\Phi_k$, respectively, we can write

$$\begin{aligned}
\hat{\mu}_{xy}^{(k)} - \mu_{xy}^{(k)} &= x^\top\hat{U}_{2k}(\hat{U}_{1k}^\top\hat{U}_{1k})^{-1}\hat{U}_{1k}^\top\hat{U}_{\mathrm{up}}^{(k)}\hat{\Sigma}_{\mathrm{up}}\hat{V}_{2k}^\top y - x^\top U_{2k}\mathcal{C}_{\bullet,\bullet,k}V_{2k}^\top y \\
&= x^\top\hat{U}_{2k}(\hat{U}_{1k}^\top\hat{U}_{1k})^{-1}\hat{U}_{1k}^\top\widehat{M}_b^{(k)}y - x^\top U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\mathcal{M}_{\bullet,\bullet,k}^{(b)}y \\
&= x^\top U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top(\widehat{M}_b^{(k)} - \mathcal{M}_{\bullet,\bullet,k}^{(b)})y + x^\top L\mathcal{M}_{\bullet,\bullet,k}^{(b)}y + x^\top L(\widehat{M}_b^{(k)} - \mathcal{M}_{\bullet,\bullet,k}^{(b)})y \\
&= x^\top U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Big[(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top + U_{1k}\mathcal{C}_{\bullet,\bullet,k}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k} + \Phi_k\Big]y \\
&\quad + x^\top LU_{1k}\mathcal{C}_{\bullet,\bullet,k}V_{2k}^\top y + x^\top L\Big[(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top + U_{1k}\mathcal{C}_{\bullet,\bullet,k}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k} + \Phi_k\Big]y \\
&= x^\top U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y + x^\top U_{2k}\mathcal{C}_{\bullet,\bullet,k}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y \\
&\quad + x^\top U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Phi_k y + x^\top LU_{1k}\mathcal{C}_{\bullet,\bullet,k}V_{2k}^\top y + x^\top L(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y \\
&\quad + x^\top LU_{1k}\mathcal{C}_{\bullet,\bullet,k}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y + x^\top L\Phi_k y \\
&= Z_{xy}^{(3)} + Z_{xy}^{(2)} + x^\top U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Phi_k y \\
&\quad + x^\top(E_{\mathrm{left}}^{\mathrm{p}})_{\mathcal{I}_k,\bullet}W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}\mathcal{C}_{\bullet,\bullet,k}V_{2k}^\top y + x^\top\Delta_L V_{2k}^\top y \\
&\quad + x^\top L(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y + x^\top(E_{\mathrm{left}}^{\mathrm{p}})_{\mathcal{I}_k,\bullet}W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}\mathcal{C}_{\bullet,\bullet,k}(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y \\
&\quad + x^\top\Delta_L(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y + x^\top L\Phi_k y \\
&= Z_{xy}^{(1)} + Z_{xy}^{(2)} + Z_{xy}^{(3)} + Z_{xy}^{(4)} + x^\top U_{2k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Phi_k y + x^\top\Delta_L V_{2k}^\top y \\
&\quad + x^\top\Delta_L(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y + x^\top L(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y + x^\top L\Phi_k y.
\end{aligned} \tag{47}$$

We will now bound each of the remainder terms individually. The first term is controlled directly by (41), while for the second term is enough to write $|x^\top\Delta_L V_{2k}^\top y| \leq \|x^\top\Delta_L\|_2\|V_{2k}^\top y\|_2$, and bound the first factor using (36). For the third one, start by observing that Lemma 21 gives $\|(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y\|_2 \leq c_1\sigma\,\gamma_{\min}^{-1}\sqrt{r+\zeta_N}\,\rho_N^{-1/2}$. Applying again (36) from Lemma 11 then gives

$$|x^\top\Delta_L(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y| \leq \|x^\top\Delta_L\|_2\,\|(W_{\mathrm{up}}^\top W_{\mathrm{up}})^{-1}W_{\mathrm{up}}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\bullet,\mathcal{J}_k}y\|_2$$

$$
\begin{aligned}
&\leq c_1 \frac{\sigma^2\sqrt{(N+T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}\rho_T}\,\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}} \;+\; c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\,\|U_{2k}^\top x\|_2\,\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}}\\
&= c_1 \frac{\sigma^3\sqrt{(N+T_{1,\mathrm{p}})(r+\zeta_N)(r+\zeta_T)}}{\gamma_{\min}^2\rho_N^{1/2}\rho_T} + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\,\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\,\rho_N^{1/2}}\|U_{2k}^\top x\|_2\\
&\leq c_1 \frac{\sigma^2\sqrt{(r+\zeta_N)(r+\zeta_T)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}} + c_1\,\frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2,
\end{aligned}
$$

where the last inequality follows from (A3).

We next control the fourth term in (47). By the definition of $L$, for any vector $w \in \mathbb{R}^{N_{1k}}$ we have $Lw = (\hat{U}_{2k}H_U - U_{2k})H_k^{-1}U_{1k}^\top w + \hat{U}_{2k}H_U D_k w$, where $D_k = \widehat{H}_k^{-1}(\hat{U}_{1k}H_U)^\top - H_k^{-1}U_{1k}^\top$. Applying this identity with $w = (E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y$ and using $\hat{U}_{2k}H_U - U_{2k} = (E_{\mathrm{left}}^{\mathrm{p}})_{\{N_{1k}+1,\ldots,N\},\bullet}W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1} + \Psi_{U,2k}$ with $\Psi_{U,2k}$ defined in Corollary 8, gives

$$
\begin{aligned}
x^\top L(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y &= x^\top(E_{\mathrm{left}}^{\mathrm{p}})_{\{N_{1k}+1,\ldots,N\},\bullet}W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}H_k^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y\\
&\quad + x^\top\Psi_{U,2k}H_k^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y + x^\top\hat{U}_{2k}H_U D_k(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y. \qquad (48)
\end{aligned}
$$

We bound these three pieces separately. Since $(E_{\mathrm{left}}^{\mathrm{p}})_{\{N_{1k}+1,\ldots,N\},\bullet}$ and $(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}$ are independent, conditionally on the upper-pooled noise $(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}$, the first term is a Gaussian random variable with conditional variance $\sigma^2\|W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}H_k^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y\|_2^2$ and mean zero. Combining this with Lemma 21 and a standard Gaussian tail bound gives

$$
\begin{aligned}
&|x^\top(E_{\mathrm{left}}^{\mathrm{p}})_{\{N_{1k}+1,\ldots,N\},\bullet}W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}H_k^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y|\\
&\quad\leq c_1\sigma\sqrt{\zeta_T}\,\left\|W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\,V\,V_{2k}^\top y\right\|_2\\
&\quad\leq c_1\sigma\sqrt{\zeta_T}\,\left\|W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}\,V\right\|_{\mathrm{op}}\,\left\|V_{2k}^\top y\right\|_2\\
&\quad\leq c_1\sigma^2\sqrt{\zeta_T\,(r+\zeta_N)}\,\left\|W_{\mathrm{left}}(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}\right\|_{\mathrm{op}}\,\left\|(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\right\|_{\mathrm{op}}\,\left\|V_{2k}^\top y\right\|_2\\
&\quad\leq c_1\frac{\sigma^2}{\gamma_{\min}}\sqrt{\frac{N\,\zeta_T\,(r+\zeta_N)}{N_{1k}\,\rho_T}}\,\left\|V_{2k}^\top y\right\|_2 \leq c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\,\left\|V_{2k}^\top y\right\|_2
\end{aligned}
$$

with probability at least $1-\mathcal{O}(p_N^{-10}+p_T^{-10})$, where in the penultimate inequality we used Lemma 5 to get $\|W_{\mathrm{left}}\,(W_{\mathrm{left}}^\top W_{\mathrm{left}})^{-1}\|_{\mathrm{op}} \leq \sigma_r^{-1}(M_{\mathrm{left}}^{\mathrm{p}}) \leq c_\ell^{-1/2}\gamma_{\min}^{-1}\,\rho_T^{-1/2}$, and (A1) to get $\left\|(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\right\|_{\mathrm{op}} \leq c_\ell^{-1/2}\sqrt{N/N_{1k}}$. Similarly, for the second piece in (48), Lemma 21, (A1), (A2), (A3) and (26) give

$$
\begin{aligned}
&\left|x^\top\Psi_{U,2k}H_k^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y\right|\\
&\quad\leq \|\Psi_{U,2k}^\top x\|_2\,\left\|H_k^{-1}U_{1k}^\top(E_{\mathrm{up}}^{\mathrm{p}})_{\mathcal{I}_k^{\mathrm{up}},\bullet}VV_{2k}^\top y\right\|_2 \leq c_1\,\sigma\sqrt{\frac{N(r+\zeta_N)}{N_{1k}}}\,\|V_{2k}^\top y\|_2\,\|\Psi_{U,2k}^\top x\|_2\\
&\quad\leq c_1\,\sigma\sqrt{\frac{N(r+\zeta_N)}{N_{1k}}}\,\|V_{2k}^\top y\|_2\,\left\{\frac{\sigma^2\sqrt{(N+T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}^2\rho_T} + \left(\frac{\sigma^2 N}{\gamma_{\min}^2\rho_T} + \frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\rho_T^{1/2}}\right)\|U_{2k}^\top x\|_2\right\}\\
&\quad\leq c_1\frac{\sigma^3\sqrt{N(r+\zeta_N)(N+T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}^2\rho_T\sqrt{N_{1k}}}\,\|V_{2k}^\top y\|_2
\end{aligned}
$$

$$
\begin{aligned}
&+ c_1\sigma\sqrt{\frac{N(r+\zeta_N)}{N_{1k}}}\left(\frac{\sigma^2 N}{\gamma_{\min}^2\rho_T} + \frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\rho_T^{1/2}}\right)\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2 \\
&\le c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\left\|V_{2k}^\top y\right\|_2 + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\left\|U_{2k}^\top x\right\|_2\left\|V_{2k}^\top y\right\|_2.
\end{aligned}
$$

For the third piece, (22), (A2), (A3), Lemmas 10 and 21 yield

$$
\begin{aligned}
\left|x^\top \hat{U}_{2k} H_U D_k (E_{\text{up}}^{\text{p}})_{\mathcal{I}_k^{\text{up}},\bullet} V V_{2k}^\top y\right| &\le \|H_U^\top \hat{U}_{2k}^\top x\|_2\,\|D_k\|_{\text{op}}\left\|(E_{\text{up}}^{\text{p}})_{\mathcal{I}_k^{\text{up}},\bullet} V V_{2k}^\top y\right\|_2 \\
&\le c_1\left(\frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\rho_T^{1/2}} + \|U_{2k}^\top x\|_2\right)\frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}}\sqrt{\frac{N}{N_{1k}}}\,\sigma\sqrt{N_{1k}}\,\|V_{2k}^\top y\|_2 \\
&\le c_1\frac{\sigma^3 N\sqrt{r+\zeta_T}}{\gamma_{\min}^2\rho_T}\|V_{2k}^\top y\|_2 + c_1\frac{\sigma^2 N}{\gamma_{\min}\rho_T^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2 \\
&\le c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\left\|V_{2k}^\top y\right\|_2 + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\left\|U_{2k}^\top x\right\|_2\left\|V_{2k}^\top y\right\|_2.
\end{aligned}
$$

Combining the last three displays leads to

$$
\left|x^\top L (E_{\text{up}}^{\text{p}})_{\mathcal{I}_k^{\text{up}},\bullet} V V_{2k}^\top y\right| \le c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\left\|V_{2k}^\top y\right\|_2 + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\left\|U_{2k}^\top x\right\|_2\left\|V_{2k}^\top y\right\|_2.
$$

Finally, we control the fifth term in (47). Letting $P_{1k} := U_{1k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top$, we can write

$$
x^\top L\Phi_k y = x^\top L P_{1k}\Phi_k y + x^\top L(I_{N_{1k}} - P_{1k})\Phi_k y. \tag{49}
$$

We will now bound the first projected component. Set $z := \mathcal{C}_{\bullet,\bullet,k}^{-1}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Phi_k y$, so that $P_{1k}\Phi_k y = U_{1k}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Phi_k y = U_{1k}\mathcal{C}_{\bullet,\bullet,k}z$. Using the definition of $\Delta_L$ in (35) we obtain

$$
\begin{aligned}
x^\top L P_{1k}\Phi_k y &= x^\top (E_{\text{left}}^{\text{p}})_{\{N_{1k}+1,\dots,N\},\bullet} W_{\text{left}}(W_{\text{left}}^\top W_{\text{left}})^{-1}\mathcal{C}_{\bullet,\bullet,k}z + x^\top\Delta_L z \\
&= x^\top (E_{\text{left}}^{\text{p}})_{\{N_{1k}+1,\dots,N\},\bullet} W_{\text{left}}(W_{\text{left}}^\top W_{\text{left}})^{-1}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Phi_k y + x^\top\Delta_L z.
\end{aligned}
$$

We bound these two pieces separately. For the first one, Lemma 21, (A1) and (42) give

$$
\begin{aligned}
&\left|x^\top (E_{\text{left}}^{\text{p}})_{\{N_{1k}+1,\dots,N\},\bullet} W_{\text{left}}(W_{\text{left}}^\top W_{\text{left}})^{-1}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top\Phi_k y\right| \\
&\le \left\|x^\top (E_{\text{left}}^{\text{p}})_{\{N_{1k}+1,\dots,N\},\bullet} W_{\text{left}}(W_{\text{left}}^\top W_{\text{left}})^{-1}(U_{1k}^\top U_{1k})^{-1/2}\right\|_2\left\|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top\Phi_k y\right\|_2 \\
&\le \frac{\sigma}{\gamma_{\min}}\sqrt{\frac{N(r+\zeta_T)}{N_{1k}\,\rho_T}}\left\{\frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}}\sqrt{\frac{N_{1k}}{N}} + \frac{\sigma\sqrt{T}}{\rho_N^{1/2}}\sqrt{\frac{N_{1k}}{N}}\|V_{2k}^\top y\|_2\right\} \\
&\le c_1\frac{\sigma^2\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}} + c_1\frac{\sigma^2\sqrt{T(r+\zeta_T)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}}\|V_{2k}^\top y\|_2 \\
&\le c_1\frac{\sigma^2\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}} + c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\|V_{2k}^\top y\|_2.
\end{aligned} \tag{50}
$$

For the second projected piece, using $\sigma_{\min}(\mathcal{C}_{\bullet,\bullet,k}) \geq \gamma_{\min} > 0$, (A1), and (42), we have

$$\begin{aligned}\|z\|_2 &= \left\|\mathcal{C}_{\bullet,\bullet,k}^{-1}(U_{1k}^\top U_{1k})^{-1}U_{1k}^\top \Phi_k y\right\|_2 \\ &\leq \gamma_{\min}^{-1}\left\|(U_{1k}^\top U_{1k})^{-1/2}\right\|_{\mathrm{op}}\left\|(U_{1k}^\top U_{1k})^{-1/2}U_{1k}^\top \Phi_k y\right\|_2 \leq c_1\frac{\sigma\sqrt{r+\zeta_N}}{\gamma_{\min}\rho_N^{1/2}} + c_1\frac{\sigma\sqrt{T}}{\gamma_{\min}\rho_N^{1/2}}\|V_{2k}^\top y\|_2.\end{aligned}$$

Combining this with (36) gives

$$\begin{aligned}|x^\top \Delta_L z| &\leq \|x^\top \Delta_L\|_2\|z\|_2 \leq c_1\frac{\sigma^2\sqrt{(N+T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}\rho_T}\|z\|_2 + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\|U_{2k}^\top x\|_2\|z\|_2 \\ &\leq c_1\frac{\sigma^3\sqrt{(N+T_{1,\mathrm{p}})(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}^2\rho_T\rho_N^{1/2}} + c_1\frac{\sigma^3\sqrt{T(N+T_{1,\mathrm{p}})(r+\zeta_T)}}{\gamma_{\min}^2\rho_T\rho_N^{1/2}}\|V_{2k}^\top y\|_2 \\ &\quad + c_1\frac{\sigma^2\sqrt{N(r+\zeta_N)}}{\gamma_{\min}\rho_T^{1/2}\rho_N^{1/2}}\|U_{2k}^\top x\|_2 + c_1\frac{\sigma^2\sqrt{NT}}{\gamma_{\min}\rho_T^{1/2}\rho_N^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2 \\ &\leq c_1\frac{\sigma^2\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}} + c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\|V_{2k}^\top y\|_2 + c_1\frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2 + c_1\frac{\sigma\sqrt{T}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2,\end{aligned} \tag{51}$$

and concludes the analysis of the first term in (49). It remains to control the component orthogonal to $U_{1k}$. In particular, since $U_{1k}^\top(I_{N_{1k}} - P_{1k})\Phi_k y = 0$, the population part in the decomposition of $L$ vanishes, hence

$$\begin{aligned}L(I_{N_{1k}} - P_{1k})\Phi_k y &= \hat{U}_{2k}H_U\hat{H}_k^{-1}H_U^\top\hat{U}_{1k}^\top(I_{N_{1k}} - P_{1k})\Phi_k y - U_{2k}H_k^{-1}U_{1k}^\top(I_{N_{1k}} - P_{1k})\Phi_k y \\ &= \hat{U}_{2k}H_U\left(\hat{H}_k^{-1}H_U^\top\hat{U}_{1k}^\top - H_k^{-1}U_{1k}^\top\right)(I_{N_{1k}} - P_{1k})\Phi_k y \\ &= \hat{U}_{2k}H_U D_k(I_{N_{1k}} - P_{1k})\Phi_k y.\end{aligned}$$

Therefore, (22), (33), (43) and (A3) give

$$\begin{aligned}|x^\top L(I_{N_{1k}} - P_{1k})\Phi_k y| &\leq \|H_U^\top\hat{U}_{2k}^\top x\|_2\|D_k\|_{\mathrm{op}}\|(I_{N_{1k}} - P_{1k})\Phi_k y\|_2 \\ &\leq c_1\left(\frac{\sigma\sqrt{r+\zeta_T}}{\gamma_{\min}\rho_T^{1/2}} + \|U_{2k}^\top x\|_2\right)\frac{\sigma\sqrt{N}}{\gamma_{\min}\rho_T^{1/2}}\sqrt{\frac{N}{N_{1k}}} \\ &\qquad \times\left(\frac{\sigma^3 T\sqrt{r+\zeta_N}}{\gamma_{\min}^2\rho_N} + \frac{\sigma^2\sqrt{N_{1k}(r+\zeta_N)}}{\gamma_{\min}\rho_N^{1/2}} + \frac{\sigma^2\sqrt{T(N_{1k}+T)}}{\gamma_{\min}\rho_N^{1/2}}\|V_{2k}^\top y\|_2\right) \\ &= c_1\frac{\sigma^5 NT\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}^4\rho_T\rho_N\sqrt{N_{1k}}} + c_1\frac{\sigma^4 N\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}^3\rho_T\rho_N^{1/2}} + c_1\frac{\sigma^4 N\sqrt{(r+\zeta_T)T(N_{1k}+T)}}{\gamma_{\min}^3\rho_T\rho_N^{1/2}\sqrt{N_{1k}}}\|V_{2k}^\top y\|_2 \\ &\quad + c_1\frac{\sigma^4 NT\sqrt{r+\zeta_N}}{\gamma_{\min}^3\rho_T^{1/2}\rho_N\sqrt{N_{1k}}}\|U_{2k}^\top x\|_2 + c_1\frac{\sigma^3 N\sqrt{r+\zeta_N}}{\gamma_{\min}^2\rho_T^{1/2}\rho_N^{1/2}}\|U_{2k}^\top x\|_2 + c_1\frac{\sigma^3 N\sqrt{T(N_{1k}+T)}}{\gamma_{\min}^2\rho_T^{1/2}\rho_N^{1/2}\sqrt{N_{1k}}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2 \\ &\leq c_1\frac{\sigma^2\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}} + c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\|V_{2k}^\top y\|_2 + c_1\frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2 + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2.\end{aligned} \tag{52}$$

Combining (49), (50), (51), and (52), we obtain

$$|x^\top L\Phi_k y| \le c_1 \frac{\sigma^2\sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min}\rho_N^{1/2}\rho_T^{1/2}} + c_1\frac{\sigma\sqrt{r+\zeta_T}}{\rho_T^{1/2}}\|V_{2k}^\top y\|_2 + c_1\frac{\sigma\sqrt{r+\zeta_N}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2 \\ + c_1\frac{\sigma\sqrt{N}}{\rho_T^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2 + c_1\frac{\sigma\sqrt{T}}{\rho_N^{1/2}}\|U_{2k}^\top x\|_2\|V_{2k}^\top y\|_2.$$

Combining all the previous inequalities gives a bound on $|\Delta_{xy}|$ and completes the proof. □

**Lemma 15.** *Consider the setting of Lemma 14, and further assume (A4) with $\nu_x \neq 0, \nu_y \neq 0$. Also let $\mathcal{G}_1$ be the event such that $\mathbb{P}(\mathcal{G}_1) \ge 1 - \mathcal{O}(p_N^{-10} + p_T^{-10})$ under which (46) holds. Define*

$$\Upsilon_{xy} := \frac{\sigma^2(r+\zeta_N)}{\rho_N}\|U_{2k}^\top x\|_2^2 + \frac{\sigma^2(r+\zeta_T)}{\rho_T}\|V_{2k}^\top y\|_2^2 + \frac{\sigma^2 N}{N_{1k}}\|U_{2k}^\top x\|_2^2\,\|V_{2k}^\top y\|_2^2.$$

*We have $\mathbb{E}[Z_{xy}^2] \le c_1\,\Upsilon_{xy}$ for a sufficiently large constant $c_1 \equiv c_1(c_\ell, c_u, c_0, c_{\text{blk}}, \kappa, \nu_x, \nu_y) > 0$. Furthermore, under $\mathcal{G}_1$, the remainder satisfies $\Delta_{xy}^2 \le c_1\Upsilon_{xy}$.*

*Proof.* We will use $\mathbb{E}[Z_{xy}^2] \le 4\sum_{i=1}^4 \mathbb{E}[(Z_{xy}^{(i)})^2]$, and bound the second moment of each $Z_{xy}^{(i)}$ in (45) separately. First, from $Z_{xy}^{(2)} \sim \mathcal{N}(0, \sigma^2\,\|x^\top U_{2k}\,\mathcal{C}_{\bullet,\bullet,k}\,(W_{\text{up}}^\top W_{\text{up}})^{-1}W_{\text{up}}^\top\|_2^2)$ we get

$$\mathbb{E}[(Z_{xy}^{(2)})^2] = \text{Var}(Z_{xy}^{(2)}) = \sigma^2\,\|x^\top U_{2k}\,\mathcal{C}_{\bullet,\bullet,k}\,(W_{\text{up}}^\top W_{\text{up}})^{-1}W_{\text{up}}^\top\|_2^2 \\ \le \sigma^2\,\|U_{2k}^\top x\|_2^2\,\|\mathcal{C}_{\bullet,\bullet,k}\,(W_{\text{up}}^\top W_{\text{up}})^{-1}W_{\text{up}}^\top\|_{\text{op}}^2 \le \frac{\kappa^2\sigma^2}{c_\ell\,\rho_N}\|U_{2k}^\top x\|_2^2,$$

where the second inequality follows from Lemma 5. Furthermore, similar computations allow showing that $\mathbb{E}[(Z_{xy}^{(4)})^2] = \text{Var}(Z_{xy}^{(4)}) \le \kappa^2\sigma^2\,c_\ell^{-1}\,\|V_{2k}^\top y\|_2^2\,\rho_T^{-1}$ and $\mathbb{E}[(Z_{xy}^{(3)})^2] = \text{Var}(Z_{xy}^{(3)}) \le \sigma^2\,c_\ell^{-1}\,(N/N_{1k})\,\|U_{2k}^\top x\|_2^2\,\|V_{2k}^\top y\|_2^2$. Finally, for the first term we have

$$\begin{aligned}
\mathbb{E}[(Z_{xy}^{(1)})^2] &\le \left\|(W_{\text{left}}^\top W_{\text{left}})^{-1/2}\mathcal{C}_{\bullet,\bullet,k}(W_{\text{up}}^\top W_{\text{up}})^{-1/2}\right\|_{\text{op}}^2 \\
&\quad\times \mathbb{E}\left[\left\|(W_{\text{left}}^\top W_{\text{left}})^{-1/2}W_{\text{left}}^\top(E_{\text{left}}^{\text{p}})_{\mathcal{I}_k,\bullet}^\top x\right\|_2^2\left\|(W_{\text{up}}^\top W_{\text{up}})^{-1/2}W_{\text{up}}^\top(E_{\text{up}}^{\text{p}})_{\bullet,\mathcal{J}_k}y\right\|_2^2\right] \\
&\le \left\|(W_{\text{left}}^\top W_{\text{left}})^{-1/2}\mathcal{C}_{\bullet,\bullet,k}(W_{\text{up}}^\top W_{\text{up}})^{-1/2}\right\|_{\text{op}}^2 \\
&\quad\times \left\{\mathbb{E}\left\|(W_{\text{left}}^\top W_{\text{left}})^{-1/2}W_{\text{left}}^\top(E_{\text{left}}^{\text{p}})_{\mathcal{I}_k,\bullet}^\top x\right\|_2^4\right\}^{1/2}\left\{\mathbb{E}\left\|(W_{\text{up}}^\top W_{\text{up}})^{-1/2}W_{\text{up}}^\top(E_{\text{up}}^{\text{p}})_{\bullet,\mathcal{J}_k}y\right\|_2^4\right\}^{1/2} \\
&= \sigma^4 r(r+2)\left\|(W_{\text{left}}^\top W_{\text{left}})^{-1/2}\mathcal{C}_{\bullet,\bullet,k}(W_{\text{up}}^\top W_{\text{up}})^{-1/2}\right\|_{\text{op}}^2 \\
&\le \frac{\gamma_{\max}^2}{c_\ell^2\gamma_{\min}^4}\frac{\sigma^4 r(r+2)}{\rho_T\rho_N} \le c_1\frac{\sigma^4 r^2}{\gamma_{\min}^2\rho_T\rho_N}.
\end{aligned}$$

The second bound follows from the Cauchy–Schwarz inequality, the first and only equality uses the fact that $\sigma^{-1}(W_{\text{left}}^\top W_{\text{left}})^{-1/2}W_{\text{left}}^\top(E_{\text{left}}^{\text{p}})_{\mathcal{I}_k,\bullet}^\top x$ and $\sigma^{-1}(W_{\text{up}}^\top W_{\text{up}})^{-1/2}W_{\text{up}}^\top(E_{\text{up}}^{\text{p}})_{\bullet,\mathcal{J}_k}y$ are standard normal vectors

in $\mathbb{R}^r$, and hence have fourth moment $r(r+2)$ in squared Euclidean norm, while the penultimate inequality uses Lemma 5 to obtain $\lambda_{\min}(W_{\text{left}}^\top W_{\text{left}}) \geq c_\ell \gamma_{\min}^2 \rho_T$ and $\lambda_{\min}(W_{\text{up}}^\top W_{\text{up}}) \geq c_\ell \gamma_{\min}^2 \rho_N$. Furthermore, we can show that this term is dominated by either one of the first two terms in $\Upsilon_{xy}$. Indeed, we have

$$\frac{\sigma^4 r^2}{\gamma_{\min}^2 \rho_T \rho_N} = \frac{\sigma^4 r^2}{\gamma_{\min}^2 \rho_T \rho_N} \frac{N}{\nu_x^2 r} \|U_{2k}^\top x\|_2^2 = \frac{\sigma^2 N}{\gamma_{\min}^2 \rho_T} \frac{1}{\nu_x^2} \frac{\sigma^2 r}{\rho_N} \|U_{2k}^\top x\|_2^2 \leq \frac{c_0^2}{\nu_x^2} \frac{\sigma^2 (r+\zeta_N)}{\rho_N} \|U_{2k}^\top x\|_2^2.$$

This, combined with the previous bounds and $\min(r+\zeta_N, r+\zeta_T) \geq 1$, gives $\mathbb{E}[Z_{xy}^2] \leq c_1 \, \Upsilon_{xy}$.

Coming now to bounding the remainder, we observe that the second and third terms in (46) appear in the definition of $\Upsilon_{xy}$. It thus remains to control the other three. Using the definition of $\nu_x$ in (A4), for the first one we get

$$\begin{aligned}\frac{\sigma^2 \sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min} \rho_N^{1/2} \rho_T^{1/2}} &= \frac{\sigma^2 \sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min} \rho_N^{1/2} \rho_T^{1/2}} \frac{\sqrt{N}}{\nu_x \sqrt{r}} \|U_{2k}^\top x\|_2 = \frac{\sigma \sqrt{N}}{\gamma_{\min} \rho_T^{1/2}} \frac{\sigma \sqrt{(r+\zeta_N)(1+\zeta_T/r)}}{\nu_x \, \rho_N^{1/2}} \|U_{2k}^\top x\|_2 \\ &\leq \frac{c_0}{\nu_x} \frac{\sigma \sqrt{(r+\zeta_N)(1+\zeta_T/r)}}{\rho_N^{1/2}} \|U_{2k}^\top x\|_2.\end{aligned}$$

Arguing by symmetry and using $\min(\zeta_N, \zeta_T) \leq c_{\text{blk}} \, r$, we can thus conclude

$$\frac{\sigma^2 \sqrt{(r+\zeta_T)(r+\zeta_N)}}{\gamma_{\min} \rho_N^{1/2} \rho_T^{1/2}} \leq c_0 \sqrt{1+c_{\text{blk}}} \max \left\{ \frac{1}{\nu_x} \frac{\sigma \sqrt{r+\zeta_N}}{\rho_N^{1/2}} \|U_{2k}^\top x\|_2, \, \frac{1}{\nu_y} \frac{\sigma \sqrt{r+\zeta_T}}{\rho_T^{1/2}} \|V_{2k}^\top y\|_2 \right\}.$$

Similarly, we can bound the fourth term in (46) by

$$\frac{\sigma \sqrt{N}}{\rho_T^{1/2}} \|U_{2k}^\top x\|_2 \, \|V_{2k}^\top y\|_2 = \frac{\sigma \sqrt{N}}{\rho_T^{1/2}} \frac{\nu_x \sqrt{r}}{\sqrt{N}} \|V_{2k}^\top y\|_2 = \nu_x \frac{\sigma \sqrt{r}}{\rho_T^{1/2}} \|V_{2k}^\top y\|_2.$$

An analogous bound holds for the fifth term, thereby showing that $|\Delta_{xy}| \leq c_1 \sqrt{\Upsilon_{xy}}$ on the event where (46) holds. This completes the proof. □

# D Additional results

## D.1 Sufficient conditions for (A1)

We comment further on (A1) by providing two sufficient conditions under which it holds, either deterministically or with high probability.

**Lemma 16.** *Let $U \in \mathbb{R}^{N \times r}$ have orthonormal columns and assume there exists $\mu \geq 2$ such that $\|U^\top \boldsymbol{e}_j\|_2^2 \leq \mu r / N$ for all $j \in [N]$. Let $U_1$ be any submatrix formed by selecting $N_1$ rows of $U$. If $N_1 \geq \frac{2\mu r}{2\mu r + 1} N$ then assumption (A1) holds with $c_\ell = \frac{1}{2}$ and $c_u = \frac{5}{4}$.*

*Proof.* Write the row vectors $u_j := U^\top \boldsymbol{e}_j \in \mathbb{R}^r$, so that $\sum_{j=1}^N u_j u_j^\top = U^\top U = I_r$. Let $S \subset [N]$ index the $N_1$

selected rows, and let $S^{\complement}$ be the complement with $N_2 := |S^{\complement}| = N - N_1$. Then

$$U_1^\top U_1 = \sum_{j \in S} u_j u_j^\top = I_r - \sum_{j \in S^{\complement}} u_j u_j^\top = I_r - U_2^\top U_2,$$

where $U_2$ is the submatrix of the remaining $N_2$ rows. Since $U_2^\top U_2 \succeq 0$ we immediately get $U_1^\top U_1 \preceq I_r$.

For the lower bound, combining $\lambda_{\max}(U_2^\top U_2) \le \mathrm{tr}(U_2^\top U_2) = \|U_2\|_F^2$ with incoherence gives

$$\|U_2\|_{\mathrm{op}}^2 \le \|U_2\|_F^2 = \sum_{j \in S^{\complement}} \|u_j\|_2^2 \le N_2 \frac{\mu r}{N},$$

which leads to

$$U_1^\top U_1 = I_r - U_2^\top U_2 \succeq \Big(1 - \frac{\mu r\, N_2}{N}\Big) I_r \succeq \frac{1}{2} I_r \succeq \frac{1}{2}\frac{N_1}{N} I_r.$$

In particular, the last step follows from $N_1/N \le 1$, while in the penultimate inequality we used $N_1 \ge \frac{2\mu r}{2\mu r+1} N$, which implies $N_2/N \le \frac{1}{2\mu r+1}$, and thus $1 - \frac{\mu r\, N_2}{N} \ge 1 - \frac{\mu r}{2\mu r+1} = \frac{\mu r+1}{2\mu r+1} \ge \frac{1}{2}$.

For the upper bound, it is useful to notice that $\frac{5}{4}\frac{N_1}{N} \ge \frac{5}{4}\frac{2\mu r}{2\mu r+1} \ge 1$, where the middle inequality follows from $\mu r \ge 2$. This, together with $U_1^\top U_1 \preceq I_r$, shows that $U_1^\top U_1 \preceq I_r \preceq \frac{5}{4}\frac{N_1}{N} I_r$, and concludes the proof. □

**Lemma 17.** *Let $U \in \mathbb{R}^{N \times r}$ have orthonormal columns and satisfy $\|U^\top \boldsymbol{e}_j\|_2^2 \le \mu r/N$ for all $j \in [N]$ for some $\mu \ge 1$. Let $U_1$ be formed by selecting $N_1$ rows from $U$ uniformly at random without replacement. For all $\varepsilon \in (0,1)$ we have*

$$\mathbb{P}\left\{(1-\varepsilon)\frac{N_1}{N} I_r \;\preceq\; U_1^\top U_1 \;\preceq\; (1+\varepsilon)\frac{N_1}{N} I_r\right\} \ge 1 - 2r \exp\left\{-\frac{N_1 \varepsilon^2}{3\mu r}\right\}.$$

*Proof.* For each $j \in [N]$, define $u_j := U^\top \boldsymbol{e}_j \in \mathbb{R}^r$. Let $S$ denote the random set of $N_1$ row indices sampled uniformly at random without replacement from $[N]$, so that $U_1^\top U_1 = \sum_{j \in S} u_j u_j^\top$. Equivalently, we may write $U_1^\top U_1 = \sum_{i=1}^{N_1} X_i$, where $X_1, \dots, X_{N_1}$ are sampled uniformly without replacement from $\{u_j u_j^\top : j \in [N]\}$.

Each $u_j u_j^\top$ is positive semidefinite, and by the incoherence assumption we have $\lambda_{\max}(u_j u_j^\top) = \|u_j\|_2^2 \le \mu r/N =: B$ for all $j \in [N]$. Moreover, each $X_i$ has marginal distribution uniform on $\{u_j u_j^\top : j \in [N]\}$, so

$$\mathbb{E} X_i = \frac{1}{N}\sum_{j=1}^{N} u_j u_j^\top = \frac{1}{N} U^\top U = \frac{1}{N} I_r.$$

Writing $\lambda_- := \lambda_{\min}(\sum_{i=1}^{N_1} \mathbb{E} X_i)$ and $\lambda_+ := \lambda_{\max}(\sum_{i=1}^{N_1} \mathbb{E} X_i)$, we thus have $\lambda_- = \lambda_+ = N_1/N$.

Although the matrices $X_i$ are dependent, the trace-moment argument in Gross and Nesme (2010) allows the usual matrix Chernoff bounds to be applied to sampling without replacement from a finite collection. Hence, using Tropp (2012, Theorem 1.1) with dimension $r$, norm bound $B = \mu r/N$, and mean eigenvalues $\lambda_- = \lambda_+ = N_1/N$, gives, for $\varepsilon \in (0,1)$,

$$\mathbb{P}\left\{\lambda_{\min}(U_1^\top U_1) \le (1-\varepsilon)\frac{N_1}{N}\right\} \le r\left\{\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right\}^{\lambda_-/B} = r\left\{\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right\}^{N_1/(\mu r)} \le r \exp\left\{-\frac{N_1 \varepsilon^2}{3\mu r}\right\},$$

$$\mathbb{P}\left\{\lambda_{\max}(U_1^\top U_1) \geq (1+\varepsilon)\frac{N_1}{N}\right\} \leq r\left\{\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right\}^{\lambda_+/B} = r\left\{\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right\}^{N_1/(\mu r)} \leq r\exp\left\{-\frac{N_1\varepsilon^2}{3\mu r}\right\}.$$

In the last inequalities we used the standard bounds

$$\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}} \leq \exp\left\{-\frac{\varepsilon^2}{2}\right\} \leq \exp\left\{-\frac{\varepsilon^2}{3}\right\}, \qquad \frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} \leq \exp\left\{-\frac{\varepsilon^2}{3}\right\},$$

both of which hold for $\varepsilon \in (0,1)$. A union bound over the lower- and upper-tail events concludes the proof. □

Note that both lemmas become noninformative as soon as the rank $r$ is of the same order as the sampled dimension $N_1$. In Lemma 16, the sufficient condition $N_1 \geq \frac{2\mu r}{2\mu r+1} N$ forces $N_1/N \approx 1$ when $r$ is large. Interpreted in the four-block setting used in the main body, this means that the fraction of rows containing missing entries must be exceptionally small. In Lemma 17, while $U_1^\top U_1$ concentrates around $(N_1/N)I_r$ when $N_1$ is much larger than $\mu r \log r$, a Marchenko–Pastur-type heuristic suggests that $\lambda_{\min}(U_1^\top U_1) \approx \frac{N_1}{N}(1-\sqrt{\gamma})^2$ as $r/N_1 \to \gamma \in (0,1)$, so the lower constant $c_\ell$ in (A1) deteriorates and can be arbitrarily small when $r$ is too large relative to $N_1$.

Returning to Lemma 16, the requirement that the missing block be small has close analogues in the MNAR causal-panel matrix-completion literature. For instance, horizontal regression (see, e.g., Athey et al., 2021, for a discussion) in the unconfoundedness literature is most appropriate when there are many control units relative to the number of periods, whereas vertical regression in the synthetic-control literature is most appropriate when there are many pre-treatment periods relative to the number of donor units. More precisely, horizontal regression is essentially feasible only when $N \gg T$, whereas vertical regression is viable when $T \gg N$. In either case, estimation is reliable only if the corresponding regression design matrices and the implied factor structure are sufficiently well-conditioned. Moreover, settings with limited missing values are also a central building block of Choi and Yuan (2026). They first show that nuclear-norm regularisation can accurately estimate the low-rank signal under MNAR when the total number of missing entries is sufficiently small (Assumption (iii) in Theorem 2.1). They then extend to more general MNAR patterns by partitioning the missing set into small groups so that each subproblem has few missing entries.

## D.2 Hardness results for $c_\ell = 0$

We illustrate why restricting our attention to $c_\ell > 0$ in (A1) is essential for (3) to be identifiable. Indeed, when $c_\ell = 0$, the restricted Gram matrices $U_{1j}^\top U_{1j}$ and $V_{1j}^\top V_{1j}$ are allowed to be singular. This creates directions in which the slice-specific core $\mathcal{C}_{\bullet,\bullet,k}$ can be perturbed so that only the unobserved $d$-block of $\mathcal{M}_{\bullet,\bullet,k}$ changes, while all observed entries across all slices remain the same. Therefore, two elements of the class can induce the same distribution while having different values of $\mu_{xy}^{(k)}$, so consistent estimation is impossible.

We recall $\mu_{xy}^{(k)}(\mathcal{M}) = x^\top \mathcal{M}_{\bullet,\bullet,k}^{(d)} y$ and $Z_\Omega = \{\, \mathcal{M}_{itj} + \mathcal{E}_{itj} : \Omega_{i,t,j} = 1,\ (i,t,j) \in [N]\times[T]\times[K] \,\}$, with the mask $\Omega$ fixed and known, and write $\mathbb{P}_{\mathcal{M}}$ and $\mathbb{E}_{\mathcal{M}}$ for probability and expectation under the law of $Z_\Omega$.

**Proposition 18.** *Fix an index $k \in [K]$, constants $\gamma_{\max} > \gamma_{\min} > 0$, and unit vectors $x \in \mathbb{B}_2(N_{2k})$,*

$y \in \mathbb{B}_2(T_{2k})$. *Let* $\bar{c}_u := \max(N/\min_{j\in[K]} N_{1j},\, T/\min_{j\in[K]} T_{1j})$. *Then*

$$\inf_{\phi} \sup_{\mathcal{M}\in\mathcal{F}(0,\bar{c}_u)} \mathbb{E}_{\mathcal{M}}\left[\left\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M})\right\}^2\right] \geq \frac{(\gamma_{\max}-\gamma_{\min})^2}{4},$$

*where the infimum is over all Borel-measurable functions* $\phi$ *of the observed entries* $Z_\Omega$.

*Proof.* Define $\bar{x} \in \mathbb{B}_2(N)$ and $\bar{y} \in \mathbb{B}_2(T)$ to be the vectors with entries $\bar{x}_i := x_{\,i-N_{1k}}\,\mathbb{1}\{i > N_{1k}\}$ and $\bar{y}_t := y_{\,t-T_{1k}}\,\mathbb{1}\{t > T_{1k}\}$, respectively. Choose vectors $\{u_2,\dots,u_r\} \subset \mathbb{R}^N$ and $\{v_2,\dots,v_r\} \subset \mathbb{R}^T$ such that $\{\bar{x},u_2,\dots,u_r\}$ and $\{\bar{y},v_2,\dots,v_r\}$ are orthonormal sets. In particular, this is possible since $r \leq \min(N,T)$. Define $U := \left(\bar{x} \mid u_2 \mid \cdots \mid u_r\right) \in \mathbb{R}^{N\times r}$, $V := \left(\bar{y} \mid v_2 \mid \cdots \mid v_r\right) \in \mathbb{R}^{T\times r}$ so that $U^\top U = V^\top V = I_r$. For every slice $j \neq k$ we take $C_{\bullet,\bullet,j} = \gamma_{\min} I_r$. For the specific slice $k$, writing $\boldsymbol{e}_1 = (1,0,\dots,0)^\top \in \mathbb{R}^r$ for the first vector of the canonical basis in $\mathbb{R}^r$, we set $\mathcal{C}^-_{\bullet,\bullet,k} = \gamma_{\min} I_r$ and $\mathcal{C}^+_{\bullet,\bullet,k} = \gamma_{\min} I_r + (\gamma_{\max}-\gamma_{\min})\,\boldsymbol{e}_1\boldsymbol{e}_1^\top$. Finally, we define $\mathcal{M}^\pm := \mathcal{C}^\pm \times_1 U \times_2 V \times_3 I_K$.

We first verify that $\mathcal{M}^+, \mathcal{M}^- \in \mathcal{F}(0,\bar{c}_u)$. The orthonormality constraints on $U$ and $V$ hold by construction. Moreover, for every $j \in [K]$,

$$0 \preceq U_{1j}^\top U_{1j} \preceq I_r \preceq \bar{c}_u \frac{N_{1j}}{N} I_r, \qquad 0 \preceq V_{1j}^\top V_{1j} \preceq I_r \preceq \bar{c}_u \frac{T_{1j}}{T} I_r,$$

by the definition of $\bar{c}_u$, hence Assumption (A1) holds with $c_\ell = 0$ and $c_u = \bar{c}_u$. Finally, for $j \neq k$, all singular values of $\mathcal{C}_{\bullet,\bullet,j}$ are equal to $\gamma_{\min}$; for $j = k$, the singular values of $\mathcal{C}^-_{\bullet,\bullet,k}$ are all equal to $\gamma_{\min}$ and those of $\mathcal{C}^+_{\bullet,\bullet,k}$ lie in $[\gamma_{\min},\gamma_{\max}]$. Therefore both tensors belong to $\mathcal{F}(0,\bar{c}_u)$.

Next, since $\bar{x}$ and $\bar{y}$ are supported on the row and column indices of the missing $d$-block of slice $k$, the rank-one perturbation $(\gamma_{\max}-\gamma_{\min})\bar{x}\bar{y}^\top$ is supported entirely on that missing block. Hence $P_{\Omega_{\bullet,\bullet,j}}(\mathcal{M}^+_{\bullet,\bullet,j}) = P_{\Omega_{\bullet,\bullet,j}}(\mathcal{M}^-_{\bullet,\bullet,j})$ for all $j \in [K]$. Since the noise distribution is the same under $\mathcal{M}^+$ and $\mathcal{M}^-$, it follows that the induced laws of the observed data coincide, that is $\mathbb{P}_{\mathcal{M}^+} = \mathbb{P}_{\mathcal{M}^-}$. On the other hand, the corresponding target parameters are separated. Indeed,

$$\mu_{xy}^{(k)}(\mathcal{M}^+) - \mu_{xy}^{(k)}(\mathcal{M}^-) = x^\top\left(\mathcal{M}^{+,(d)}_{\bullet,\bullet,k} - \mathcal{M}^{-,(d)}_{\bullet,\bullet,k}\right) y = (\gamma_{\max}-\gamma_{\min}) x^\top x\, y^\top y = \gamma_{\max}-\gamma_{\min}.$$

Now, since $\mathbb{P}_{\mathcal{M}^+} = \mathbb{P}_{\mathcal{M}^-}$, expectations under the two laws are identical for every measurable function $\phi$ of $Z_\Omega$. We can therefore bound

$$\begin{aligned}
&\mathbb{E}_{\mathcal{M}^+}\left[\left\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M}^+)\right\}^2\right] + \mathbb{E}_{\mathcal{M}^-}\left[\left\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M}^-)\right\}^2\right] \\
&\quad= \mathbb{E}_{\mathcal{M}^+}\left[\left\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M}^+)\right\}^2 + \left\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M}^-)\right\}^2\right] \\
&\quad\geq \frac{1}{2}\left\{\mu_{xy}^{(k)}(\mathcal{M}^+) - \mu_{xy}^{(k)}(\mathcal{M}^-)\right\}^2 = \frac{1}{2}(\gamma_{\max}-\gamma_{\min})^2,
\end{aligned}$$

where we used the elementary inequality $(a-b)^2 + (a-c)^2 \geq (b-c)^2/2$. Consequently, we have

$$\sup_{\mathcal{M}\in\mathcal{F}(0,\bar{c}_u)} \mathbb{E}_{\mathcal{M}}\left[\left\{\phi(Z_\Omega) - \mu_{xy}^{(k)}(\mathcal{M})\right\}^2\right]$$

$$\geq \frac{1}{2}\left(\mathbb{E}_{\mathcal{M}^+}\left[\left\{\phi(Z_\Omega)-\mu_{xy}^{(k)}(\mathcal{M}^+)\right\}^2\right]+\mathbb{E}_{\mathcal{M}^-}\left[\left\{\phi(Z_\Omega)-\mu_{xy}^{(k)}(\mathcal{M}^-)\right\}^2\right]\right)\geq\frac{(\gamma_{\max}-\gamma_{\min})^2}{4}.$$

Taking the infimum over all Borel-measurable functions $\phi$ gives the claimed lower bound. □

The choice of $\bar{c}_u$ is to make the construction feasible over arbitrary missingness patterns. The same lower bound can be proved for each fixed $c_u > 0$ when $N_{2k} \leq N_{2j}$ and $T_{2k} \leq T_{2j}$ for all $j \neq k$, using a similar construction.

The second result shows that this issue is not merely an artifact of allowing slice-specific heterogeneity in $\mathcal{C}_{\bullet,\bullet,j}$. Even if we impose the strongest possible homogeneity assumption, i.e. $\mathcal{C}_{\bullet,\bullet,1} = \cdots = \mathcal{C}_{\bullet,\bullet,K}$, which is equivalent to $\mathcal{M}_{\bullet,\bullet,1} = \cdots = \mathcal{M}_{\bullet,\bullet,K}$, setting $c_\ell = 0$ can still make the problem information-theoretically hard unless auxiliary layers provide enough complementary information on the missing structure relevant to the target functional. We write $\mathcal{F}_{\mathrm{id}}(c_\ell, c_u) := \{\mathcal{M} \in \mathcal{F}(c_\ell, c_u) : \mathcal{C}_{\bullet,\bullet,1} = \cdots = \mathcal{C}_{\bullet,\bullet,K}\}$.

**Proposition 19.** *Fix an index $k \in [K]$, constants $\gamma_{\max} \geq \gamma_{\min} > 0$, and unit vectors $x \in \mathbb{B}_2(N_{2k})$, $y \in \mathbb{B}_2(T_{2k})$. Also let $\bar{x} \in \mathbb{B}_2(N)$ and $\bar{y} \in \mathbb{B}_2(T)$ to be the vectors with entries $\bar{x}_i := x_{\,i-N_{1k}}\,\mathbb{1}\{i > N_{1k}\}$ and $\bar{y}_t := y_{t-T_{1k}}\,\mathbb{1}\{t > T_{1k}\}$, respectively. Define $S_k(x,y) := \sum_{j=1}^K \left\|P_{\Omega_{\bullet,\bullet,j}}(\bar{x}\bar{y}^\top)\right\|_F^2$, and $\bar{c}_u := \max(N/\min_{j\in[K]} N_{1j},\, T/\min_{j\in[K]} T_{1j})$. Then*

$$\inf_{\phi}\sup_{\mathcal{M}\in\mathcal{F}_{\mathrm{id}}(0,\bar{c}_u)}\mathbb{E}_{\mathcal{M}}\left[\left\{\phi(Z_\Omega)-\mu_{xy}^{(k)}(\mathcal{M})\right\}^2\right]\geq\max_{\gamma\in[\gamma_{\min},\gamma_{\max}]}\gamma^2\left[2-2\Phi\left(\frac{\gamma\sqrt{S_k(x,y)}}{\sigma}\right)\right],$$

*where the infimum is over all Borel-measurable functions $\phi$ of the observed entries $Z_\Omega$.*

*Proof.* Choose vectors $\{u_2,\ldots,u_r\} \subset \mathbb{R}^N$ and $\{v_2,\ldots,v_r\} \subset \mathbb{R}^T$ such that $\{\bar{x}, u_2,\ldots,u_r\}$ and $\{\bar{y}, v_2,\ldots,v_r\}$ are orthonormal sets. In particular, this is possible since $r \leq \min(N,T)$. Define $U^\pm := \big(\pm\,\bar{x} \mid u_2 \mid \cdots \mid u_r\big) \in \mathbb{R}^{N\times r}$, $V := \big(\,\bar{y} \mid v_2 \mid \cdots \mid v_r\,\big) \in \mathbb{R}^{T\times r}$ so that $(U^\pm)^\top U^\pm = V^\top V = I_r$. Also, for all $j \in [K]$ set $\mathcal{C}_{\bullet,\bullet,j} = \gamma I_r$, where $\gamma \in [\gamma_{\min}, \gamma_{\max}]$.

These choices induce $\mathcal{M}^\pm \in \mathcal{F}_{\mathrm{id}}(0,\bar{c}_u)$ with $\mathcal{M}^\pm_{\bullet,\bullet,j} = \gamma\, U^\pm V^\top$. The precise computation follows an argument similar to the construction used in the proof of Proposition 18. Furthermore, we have

$$\mu_{xy}^{(k)}(\mathcal{M}^\pm) = x^\top \mathcal{M}^{\pm,(d)}_{\bullet,\bullet,k} y = \bar{x}^\top \mathcal{M}^\pm_{\bullet,\bullet,k}\,\bar{y} = \gamma \bar{x}^\top U^\pm V^\top \bar{y} = \pm\gamma.$$

hence $\{\mu_{xy}^{(k)}(\mathcal{M}^+) - \mu_{xy}^{(k)}(\mathcal{M}^-)\}^2 = 4\gamma^2$. By Le Cam's two-point method (Tsybakov, 2009, Theorem 2.2), for any measurable function $\phi$, the minimax risk is lower bounded by $\gamma^2\{1 - \mathrm{TV}(\mathbb{P}_{\mathcal{M}^+}, \mathbb{P}_{\mathcal{M}^-})\}$, where $\mathbb{P}_{\mathcal{M}^\pm}$ denotes the law of the observed entries $Z_\Omega$ under the signal $\mathcal{M}^\pm$. In particular, under Gaussian noise with common variance $\sigma^2$, independence of the errors across slices implies that the joint law of all observed entries across all slices is multivariate Gaussian with mean vector equal to the vectorisation of $\{P_{\Omega_{\bullet,\bullet,j}}(\mathcal{M}^\pm_{\bullet,\bullet,j})\}_{j\in[K]}$ and covariance $\sigma^2 I$. This, combined with $\mathcal{M}^+_{\bullet,\bullet,j} - \mathcal{M}^-_{\bullet,\bullet,j} = \gamma\,(U^+ - U^-)V^\top = 2\gamma\,\bar{x}\bar{y}^\top$, gives

$$\mathrm{TV}(\mathbb{P}_{\mathcal{M}^+}, \mathbb{P}_{\mathcal{M}^-}) = 2\Phi\left(\frac{\sqrt{\sum_{j=1}^K \|P_{\Omega_{\bullet,\bullet,j}}(2\gamma\,\bar{x}\bar{y}^\top)\|_F^2}}{2\sigma}\right) - 1 = 2\Phi\left(\frac{\gamma\,\sqrt{S_k(x,y)}}{\sigma}\right) - 1,$$

and concludes the proof upon taking the maximum over $\gamma \in [\gamma_{\min}, \gamma_{\max}]$. □

Proposition 19 gives a lower bound on the minimax risk over $\mathcal{F}_{\mathrm{id}}(0, \bar{c}_u)$, where all slices are identical. In particular, the lower bound depends on $S_k(x,y) = \sum_{j=1}^{K} \left\| P_{\Omega_{\bullet,\bullet,j}}(\bar{x}\bar{y}^\top) \right\|_F^2$, which quantifies how often the rank-one pattern $\bar{x}\bar{y}^\top$, supported on slice $k$'s missing block, is observed across other layers. For example, when $x = N_{2k}^{-1/2}\mathbf{1}_{N_{2k}}$ and $y = T_{2k}^{-1/2}\mathbf{1}_{T_{2k}}$ we have

$$S_k(x,y) = \frac{1}{N_{2k}T_{2k}} \sum_{j=1}^{K} \left\| (1 - \Omega_{\bullet,\bullet,k}) \odot \Omega_{\bullet,\bullet,j} \right\|_0 ,$$

so $S_k(x,y)$ is the fraction of slice $k$'s missing block that is observed elsewhere. When a slice $j \neq k$ has a much smaller missing block, this overlap increases and drives the lower bound to zero. On the other hand, in general, if $S_k(x,y) \lesssim \sigma^2/\gamma^2$, the two alternatives in the proof remain statistically close, and the minimax risk is of constant order. The extreme case corresponds to $S_k(x,y) = 0$, which occurs when the entire $d$-block missing under slice $k$ is also missing under every other slice; for example, this holds when $N_{2k} \leq N_{2j}$ and $T_{2k} \leq T_{2j}$ for all $j \neq k$.

Taken together, Propositions 18 and 19 highlight two distinct failure modes when $c_\ell = 0$. Proposition 18 shows that with slice-specific cores, the target functional can be non-identifiable, as different parameter values can induce the same distribution of observed entries while yielding different $\mu_{xy}^{(k)}$. Proposition 19 shows that even if the slices share a common core, the target may still be hard to estimate, as the minimax risk can remain bounded away from zero unless the missing entries of slice $k$ are sufficiently observed in other slices.

## D.3 Background on tensors

We now provide a brief background on tensors to familiarise the reader with the notation used in our model $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$. Although the main body of the paper considers only order-3 tensors, we keep this section fairly general at first. We then state and prove a result connecting the Tucker2 model with standard low-rank matrix factorisations.

A tensor is a multidimensional array $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ with entries $x_{i_1 \ldots i_d}$, where $1 \leq i_j \leq n_j$ for $j \in [d]$; order-1 tensors are vectors and order-2 tensors are matrices. Fixing all indices except $i_j$ yields a mode-$j$ fiber, which is the higher-order analogue of rows and columns. Fixing all but two indices yields a slice, which forms a two-dimensional subarray, e.g. for an order-3 tensor $\mathcal{X}$ the frontal slice $\mathcal{X}_{\bullet,\bullet,j}$ fixes the third index.

To connect tensor algebra to matrix algebra, it is convenient to use matricisation and define the mode-$j$ unfolding $X_{(j)} \in \mathbb{R}^{n_j \times \prod_{k \neq j} N_{1k}}$ as a rearrangement of $\mathcal{X}$ so that the mode-$j$ fibers become the columns of a matrix. Under this representation, the $j$-mode product (Kolda and Bader, 2009, Section 2.5) reduces to ordinary matrix multiplication. Specifically, for $A \in \mathbb{R}^{n' \times n_j}$, the tensor $\mathcal{Y} = \mathcal{X} \times_j A$ has dimensions $n_1 \times \cdots \times n_{j-1} \times n' \times n_{j+1} \times \cdots \times n_d$ and entries

$$(\mathcal{X} \times_j A)_{i_1 \ldots i_{j-1} k \, i_{j+1} \ldots i_d} = \sum_{i_j=1}^{n_j} x_{i_1 \ldots i_d} \, a_{k i_j} .$$

Furthermore, its unfolding satisfies $(\mathcal{X} \times_j A)_{(j)} = A X_{(j)}$.

A core model in multilinear algebra is the Tucker decomposition, which represents a tensor as a low-

dimensional core transformed along each mode and is of the form $\mathcal{X} = \mathcal{G} \times_1 A^{(1)} \times_2 \cdots \times_d A^{(d)}$, where the core $\mathcal{G}$ encodes interactions among latent components and the factor matrices map these components to the ambient spaces (Kolda and Bader, 2009, Equations 4.1–4.2). Setting one factor matrix to the identity yields the Tucker2 model, introduced in Section 2; in our notation this gives $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$, so mode 3 is left unchanged, i.e. slices are not mixed across $k$, while $U$ and $V$ act along modes 1 and 2, respectively.

We next provide equivalent characterisation for this model.

**Proposition 20.** *The following are equivalent:*

1. $M^{(j)} = UR_jV^\top$ *for all* $j \in [K]$*, with common orthonormal* $U \in \mathbb{R}^{N\times r}, V \in \mathbb{R}^{T\times r}$ *and* $R_j \in \mathbb{R}^{r\times r}$*;*
2. *The column (resp. row) spaces of all* $M^{(j)}$ *are contained in a common subspace* $\mathcal{U}$ *(resp.* $\mathcal{V}$*) of dimension at most* $r$*;*
3. *Stacking the matrices* $M^{(j)}$*'s yields a tensor* $\mathcal{M} \in \mathbb{R}^{N\times T\times K}$ *that admits a Tucker2 decomposition* $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$*, with core* $\mathcal{C} \in \mathbb{R}^{r\times r\times K}$ *and shared mode-1/mode-2 orthonormal factor matrices* $U$ *and* $V$*.*

*Proof.* (1) $\Rightarrow$ (2): If $M^{(j)} = UR_jV^\top$ with $U \in \mathbb{R}^{N\times r}$, $V \in \mathbb{R}^{T\times r}$ column-orthonormal, then $\mathrm{col}(M^{(j)}) \subseteq \mathrm{col}(U) =: \mathcal{U}$ and $\mathrm{row}(M^{(j)}) \subseteq \mathrm{col}(V) =: \mathcal{V}$ for all $j \in [K]$, so (2) holds.

(2) $\Rightarrow$ (1): Let $\mathcal{U}, \mathcal{V}$ be subspaces containing all column and row spaces, with $\dim(\mathcal{U}) \le r$ and $\dim(\mathcal{V}) \le r$, and let $U, V$ be orthonormal bases of $\mathcal{U}, \mathcal{V}$. Denote the orthogonal projections by $P_U := UU^\top$ and $P_V := VV^\top$. For each $j \in [K]$, the assumptions imply $P_U M^{(j)} = M^{(j)}$ and $M^{(j)} P_V = M^{(j)}$, hence

$$M^{(j)} = P_U M^{(j)} P_V = U\big(U^\top M^{(j)} V\big)V^\top.$$

Setting $R_j := U^\top M^{(j)} V$ gives (1).

(1) $\Rightarrow$ (3): Stack the matrices as a tensor $\mathcal{M} \in \mathbb{R}^{N\times T\times K}$ with frontal slices $\mathcal{M}_{\bullet,\bullet,j} = M^{(j)}$. Define a core tensor $\mathcal{C} \in \mathbb{R}^{r\times r\times K}$ by $\mathcal{C}_{\bullet,\bullet,j} := R_j$. Then $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$, which is a Tucker2 decomposition with shared mode-1/mode-2 orthonormal factors $U, V$.

(3) $\Rightarrow$ (1): Conversely, suppose $\mathcal{M}$ has a Tucker decomposition $\mathcal{M} = \mathcal{C} \times_1 U \times_2 V \times_3 I_K$ with $U \in \mathbb{R}^{N\times r}$, $V \in \mathbb{R}^{T\times r}$ column-orthonormal and $\mathcal{C} \in \mathbb{R}^{r\times r\times K}$. Writing $\mathcal{C}_{\bullet,\bullet,j}$ for the $j$-th frontal slice of $\mathcal{C}$, the $j$-th slice of $\mathcal{M}$ is

$$\mathcal{M}_{\bullet,\bullet,j} = \sum_{k=1}^{K} \delta_{jk}\, U\, \mathcal{C}_{\bullet,\bullet,k}\, V^\top = U\Big(\sum_{k=1}^{K} \delta_{jk}\, \mathcal{C}_{\bullet,\bullet,k}\Big)V^\top = U\, \mathcal{C}_{\bullet,\bullet,j} V^\top.$$

This concludes the proof. □

# E Auxiliary results

In this appendix we collect some useful results that are used in the proofs of our main results.

We begin with two tail probability bounds. For $\sigma > 0$, a random variable $X$ with mean $\mu = \mathbb{E}[X]$ is said to be $\sigma$-subgaussian if $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$.

**Lemma 21.** *Let $E \in \mathbb{R}^{n_1\times n_2}$ be a random matrix with mean-zero independent $\sigma$-subgaussian entries. For any fixed matrices $X \in \mathbb{R}^{n_1\times p_1}$ and $Y \in \mathbb{R}^{n_2\times p_2}$, for all $\delta \in (0,1)$ there exists an absolute constant $c_1 > 0$ such that*

$$\|X^\top E Y\|_{\mathrm{op}} \;\le\; c_1\,\sigma\,\|X\|_{\mathrm{op}}\,\|Y\|_{\mathrm{op}}\,\sqrt{\mathrm{rank}(X)+\mathrm{rank}(Y)+\log(\delta^{-1})}$$

*with probability at least $1-\delta$.*

*Proof.* Write the compact singular value decompositions $X = U_X\Sigma_X V_X^\top$ and $Y = U_Y\Sigma_Y V_Y^\top$, where $U_X \in \mathbb{R}^{n_1\times r_X}, V_X \in \mathbb{R}^{p_1\times r_X}, U_Y \in \mathbb{R}^{n_2\times r_Y}, V_Y \in \mathbb{R}^{p_2\times r_Y}$ have orthonormal columns, $r_X := \mathrm{rank}(X)$ and $r_Y := \mathrm{rank}(Y)$, and $\|\Sigma_X\|_{\mathrm{op}} = \|X\|_{\mathrm{op}}$, $\|\Sigma_Y\|_{\mathrm{op}} = \|Y\|_{\mathrm{op}}$. Then, we can write $X^\top EY = V_X\Sigma_X\big(U_X^\top E U_Y\big)\Sigma_Y V_Y^\top$, and by the submultiplicativity and orthonormal invariance of the spectral norm, we have $\|X^\top EY\|_{\mathrm{op}} \le \|X\|_{\mathrm{op}}\,\|Y\|_{\mathrm{op}}\,\|U_X^\top E U_Y\|_{\mathrm{op}}$. It remains to bound $\|U_X^\top E U_Y\|_{\mathrm{op}}$. For any $x \in \mathbb{B}_2(r_X)$ and $y \in \mathbb{B}_2(r_Y)$, define

$$Z(x,y) := x^\top U_X^\top E U_Y y = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} E_{ij}\,(U_X x)_i\,(U_Y y)_j,$$

so that $\|U_X^\top E U_Y\|_{\mathrm{op}} = \sup_{x\in\mathbb{B}_2(r_X),y\in\mathbb{B}_2(r_Y)}|Z(x,y)|$. For every fixed pair $(x,y)$ and any $\delta \in (0,1)$, Hoeffding's inequality for sums of subgaussian random variables and the fact that $\|U_X\,x\|_2 = \|U_Y\,y\|_2 = 1$ imply that there exists an absolute constant $c_1 > 0$ such that

$$\mathbb{P}\Big\{\,|Z(x,y)| \;>\; c_1\sigma\sqrt{\log(\delta^{-1})}\Big\} \le \delta. \tag{53}$$

In order to deal with the supremum, we will combine the above display with a standard netting argument Wainwright (2019, Chapter 5). In particular, let $\mathcal{N}_X$ and $\mathcal{N}_Y$ be $\frac{1}{4}$-nets of $\mathbb{B}_2(r_X)$ and $\mathbb{B}_2(r_Y)$, respectively. Their cardinalities satisfy $|\mathcal{N}_X| \le 9^{r_X}$ and $|\mathcal{N}_Y| \le 9^{r_Y}$ (Vershynin, 2019, Equation 4.20). Now, for all $x_1 \in \mathbb{B}_2(r_X)$ and $y_1 \in \mathbb{B}_2(r_Y)$, we can find $x_2 \in \mathcal{N}_X$, $y_2 \in \mathcal{N}_Y$ such that $\|x_1 - x_2\|_2 \le 1/4$ and $\|y_1 - y_2\|_2 \le 1/4$. This, combined with $Z(x_1,y_1) = \{Z(x_1,y_1) - Z(x_2,y_1)\} + \{Z(x_2,y_1) - Z(x_2,y_2)\} + Z(x_2,y_2)$, allows showing that $\sup_{x\in\mathbb{B}_2(r_X),y\in\mathbb{B}_2(r_Y)}|Z(x,y)| \le 2\max_{x\in\mathcal{N}_X,y\in\mathcal{N}_Y}|Z(x,y)|$ after taking the supremum on both sides. We thus get

$$\begin{aligned}\|U_X^\top E U_Y\|_{\mathrm{op}} &= \sup_{x\in\mathbb{B}_2(r_X),\,y\in\mathbb{B}_2(r_Y)}|Z(x,y)| \;\le\; 2\max_{x\in\mathcal{N}_X,\,y\in\mathcal{N}_Y}|Z(x,y)| \\ &\le c_1\sigma\sqrt{\log(\,|\mathcal{N}_X|\,|\mathcal{N}_Y|\,/\,\delta)} \le c_1\,\sigma\sqrt{r_X + r_Y + \log(\delta^{-1})}\end{aligned}$$

with probability at least $1-\delta$, where the second inequality follows from an application of (53) with $\delta/(|\mathcal{N}_X||\mathcal{N}_Y|)$ in place of $\delta$, and a union bound over $\mathcal{N}_X \times \mathcal{N}_Y$. This concludes the proof. □

**Lemma 22.** *Let $E \in \mathbb{R}^{n_1\times n_2}$ have independent entries distributed as $\mathcal{N}(0,\sigma^2)$. For every $p \ge 1$ we have*

$$\big(\mathbb{E}\|E\|_{\mathrm{op}}^p\big)^{1/p} \le \sigma\big(\sqrt{n_1} + \sqrt{n_2} + c_1\sqrt{p}\big),$$

*where $c_1 > 0$ is an absolute constant.*

*Proof.* By homogeneity, it suffices to consider the case $\sigma^2 = 1$. Indeed, writing $E = \sigma H$ with independent $H_{ij} \sim \mathcal{N}(0,1)$, we have $\|E\|_{\mathrm{op}} = \sigma\|H\|_{\mathrm{op}}$, and therefore $\big(\mathbb{E}\,\|E\|_{\mathrm{op}}^p\big)^{1/p} = \sigma\big(\mathbb{E}\,\|H\|_{\mathrm{op}}^p\big)^{1/p}$. Thus it is enough to prove that $\big(\mathbb{E}\|H\|_{\mathrm{op}}^p\big)^{1/p} \le \sqrt{n_1} + \sqrt{n_2} + c_1\sqrt{p}$. In this regard, writing $c > 0$ for an absolute constant, we recall from Vershynin (2019, Theorem 7.3.1 and Corollary 7.3.2) that $\mathbb{E}\,\|H\|_{\mathrm{op}} \le \sqrt{n_1} + \sqrt{n_2}$, and $\mathbb{P}\big(\|H\|_{\mathrm{op}} \ge \sqrt{n_1} + \sqrt{n_2} + t\big) \le 2e^{-ct^2}$ for all $t \ge 0$. Furthermore, letting $Y := (\|H\|_{\mathrm{op}} - \sqrt{n_1} - \sqrt{n_2})_+$, we also have $\|H\|_{\mathrm{op}} \le \sqrt{n_1} + \sqrt{n_2} + Y$ and $(\mathbb{E}\,\|H\|_{\mathrm{op}}^p)^{1/p} \le \sqrt{n_1} + \sqrt{n_2} + (\mathbb{E}Y^p)^{1/p}$, where the latter bound follows from Minkowski's inequality.

It remains to bound $(\mathbb{E}Y^p)^{1/p}$. Using the layer-cake formula, the change of variables $u = ct^2$, and Stirling's approximation, we obtain

$$\begin{aligned}
\mathbb{E}Y^p &= \int_0^\infty p\,t^{p-1}\,\mathbb{P}(Y > t)\,dt = \int_0^\infty p\,t^{p-1}\,\mathbb{P}\big(\|H\|_{\mathrm{op}} > \sqrt{n_1} + \sqrt{n_2} + t\big)\,dt \\
&\le 2p\int_0^\infty t^{p-1}e^{-ct^2}\,dt = p\,c^{-p/2}\int_0^\infty u^{p/2-1}e^{-u}\,du \\
&= p\,c^{-p/2}\Gamma(p/2) = 2\,c^{-p/2}\,\Gamma(p/2+1) \le (c_1\sqrt{p})^p,
\end{aligned}$$

for some constant $c_1 > 0$ depending only on $c$. Combining the above bounds gives

$$(\mathbb{E}\,\|H\|_{\mathrm{op}}^p)^{1/p} \le \sqrt{n_1} + \sqrt{n_2} + (\mathbb{E}Y^p)^{1/p} \le \sqrt{n_1} + \sqrt{n_2} + c_1\sqrt{p},$$

thereby completing the proof. □

We next recall Weyl's inequality for singular values and eigenvalues (Chen et al., 2021, Lemmas 2.2–2.3).

**Lemma 23.** *Let $A, E \in \mathbb{R}^{n\times m}$. Then, for every $1 \le i \le \min(n,m)$, the $i$-th largest singular values of $A$ and $A+E$ satisfy*

$$|\sigma_i(A+E) - \sigma_i(A)| \le \|E\|_{\mathrm{op}}.$$

*Moreover, if $n = m$ and $A, E \in \mathbb{R}^{n\times n}$ are symmetric, then, for every $1 \le i \le n$, the $i$-th largest eigenvalues of $A$ and $A+E$ satisfy*

$$|\lambda_i(A+E) - \lambda_i(A)| \le \|E\|_{\mathrm{op}}.$$

We recall that the Moore–Penrose pseudoinverse of $A = U\,\mathrm{diag}(\sigma_1,\dots,\sigma_r)V^\top$, with column orthonormal $U \in \mathbb{R}^{n_1\times r}, V \in \mathbb{R}^{n_2\times r}$ and $\sigma_i > 0$, is $A^\dagger = V\,\mathrm{diag}(\sigma_1^{-1},\dots,\sigma_r^{-1})\,U^\top$. The following lemma gives an exact identity for how the Moore–Penrose inverse changes when a full-column-rank matrix $A$ is perturbed to $B = A + \Delta$, as well as a simple operator-norm bound and a useful formula for the action of $B^\dagger - A^\dagger$ on $A$.

**Lemma 24.** *Let $A, B \in \mathbb{R}^{n_1\times n_2}$ have full column rank, and let $\Delta := B - A$. Then $B^\dagger - A^\dagger = -A^\dagger\Delta B^\dagger + A^\dagger(A^\dagger)^\top\Delta^\top(I_{n_1} - BB^\dagger)$. Consequently,*

$$\|B^\dagger - A^\dagger\|_{\mathrm{op}} \le \|A^\dagger\|_{\mathrm{op}}\,\|\Delta\|_{\mathrm{op}}\,\|B^\dagger\|_{\mathrm{op}} + \|A^\dagger\|_{\mathrm{op}}^2\,\|\Delta\|_{\mathrm{op}}.$$

*Moreover, we have $(B^\dagger - A^\dagger)A = -B^\dagger\Delta$.*

*Proof.* Since $A$ and $B$ have full column rank, $A^\dagger A = I_{n_2}$, $B^\dagger B = I_{n_2}$, and $A^\dagger = (A^\top A)^{-1}A^\top$. Therefore, $B^\dagger - A^\dagger = (B^\dagger B - A^\dagger B)B^\dagger - A^\dagger(I_{n_1} - BB^\dagger) = -A^\dagger \Delta B^\dagger - A^\dagger(I_{n_1} - BB^\dagger)$. It remains to rewrite the last term. Since $BB^\dagger$ is the orthogonal projector onto $\mathrm{col}(B)$, we have $B^\top(I_{n_1} - BB^\dagger) = 0$. As $B = A + \Delta$, this gives $A^\top(I_{n_1} - BB^\dagger) = -\Delta^\top(I_{n_1} - BB^\dagger)$, and hence $-A^\dagger(I_{n_1} - BB^\dagger) = (A^\top A)^{-1}\Delta^\top(I_{n_1} - BB^\dagger)$. Using $A^\dagger(A^\dagger)^\top = (A^\top A)^{-1}$ proves the first identity.

The norm bound follows from this identity and $\|I_{n_1} - BB^\dagger\|_{\mathrm{op}} \le 1$. Finally, $(B^\dagger - A^\dagger)A = B^\dagger A - I_{n_2} = B^\dagger A - B^\dagger B = -B^\dagger \Delta$. □

Finally, we present the auxiliary results on the Stiefel manifold used in the proofs. These include standard facts on the Haar measure and its generation via Gaussian QR decompositions; see, for example, Stewart (1980); Mezzadri (2007); Chikuse (2003). For integers $1 \le q \le d$, the Stiefel manifold is

$$\mathrm{St}(d, q) := \{Q \in \mathbb{R}^{d\times q} : Q^\top Q = I_q\}.$$

Thus, $\mathrm{St}(d, q)$ is the set of $d \times q$ matrices whose columns are orthonormal. The special cases $q = 1$ and $q = d$ reduce to the unit sphere $\mathbb{B}_2(d)$ and the orthogonal group $\mathbb{O}(d)$, respectively.

Although $\mathrm{St}(d, q)$ is not a group when $q < d$, it carries a natural probability measure that is invariant under left multiplication by orthogonal matrices. A random matrix $Q \in \mathrm{St}(d, q)$ is said to be Haar-distributed or uniformly distributed on the Stiefel manifold if $OQ \overset{d}{=} Q$ for every deterministic $O \in \mathbb{O}(d)$. Such a left-orthogonally invariant probability measure on $\mathrm{St}(d, q)$ is unique (e.g. Chikuse, 2003, Theorem 1.2.2 and Section 1.3.1). In words, multiplying $Q$ by any deterministic rotation or reflection does not change its law, hence a Haar-distributed element of $\mathrm{St}(d, q)$ may be viewed as a uniformly random matrix with orthonormal columns. A standard construction of such matrices is obtained from a Gaussian matrix. If $G \in \mathbb{R}^{d\times q}$ has i.i.d. $\mathcal{N}(0, 1)$ entries and $G = QR$ is its thin QR decomposition with the diagonal entries of $R$ taken positive, then $Q \in \mathrm{St}(d, q)$ is Haar-distributed.

The following lemma formalises a stability property of Haar-distributed Stiefel matrices. If $Q$ is uniformly distributed on $\mathrm{St}(d, q)$, then multiplying it on the right by any fixed $H \in \mathrm{St}(q, \ell)$ produces an $\ell$-dimensional orthonormal system which is itself uniformly distributed on $\mathrm{St}(d, \ell)$. The lemma also gives a high-probability bound for the size of this random orthonormal system after applying a fixed linear map $A$. In particular, when $\ell$ is small relative to $d$, the operator norm $\|AQH\|_{\mathrm{op}}$ is at most of the order

$$\|A\|_{\mathrm{op}}\sqrt{\frac{\mathrm{rank}(A) + \ell}{d}},$$

with high probability, up to universal constants.

**Lemma 25.** *Let $d, q, p \in \mathbb{N}$, and let $1 \le \ell \le q \le d$. Let $Q \in \mathbb{R}^{d\times q}$ be Haar-distributed on $\mathrm{St}(d, q)$, and let $\mathcal{F}$ be a sigma-field independent of $Q$. Let $H \in \mathrm{St}(q, \ell)$ be $\mathcal{F}$-measurable, and let $A \in \mathbb{R}^{p\times d}$ be deterministic. Then, conditional on $\mathcal{F}$, the matrix $QH$ is Haar-distributed on $\mathrm{St}(d, \ell)$. Moreover, there is a universal constant $c_1 > 0$ such that, for every $t \ge 0$ with $\sqrt{\ell} + t < \sqrt{d}$, we have*

$$\mathbb{P}\left\{\|AQH\|_{\mathrm{op}} \le c_1\|A\|_{\mathrm{op}}\frac{\sqrt{\mathrm{rank}(A) + \ell + t^2}}{\sqrt{d} - \sqrt{\ell} - t} \;\middle|\; \mathcal{F}\right\} \ge 1 - 2e^{-t^2/2}. \tag{54}$$

*In particular, if* $\sqrt{\ell} + t \le \sqrt{d}/2$, *then*

$$\mathbb{P}\left\{ \|AQH\|_{\mathrm{op}} \le 2\, c_1 \|A\|_{\mathrm{op}} \sqrt{\frac{\mathrm{rank}(A) + \ell + t^2}{d}} \;\middle|\; \mathcal{F} \right\} \ge 1 - 2e^{-t^2/2}. \tag{55}$$

*Proof.* Since $Q^\top Q = I_q$ and $H^\top H = I_\ell$, we have $(QH)^\top (QH) = H^\top Q^\top Q H = I_\ell$. Hence $QH \in \mathrm{St}(d, \ell)$.

We now identify the conditional law of $QH$. Fix $O \in \mathbb{O}(d)$. Since $Q$ is Haar-distributed on $\mathrm{St}(d, q)$ and is independent of $\mathcal{F}$, its conditional law given $\mathcal{F}$ is still Haar. Hence $OQ \overset{d}{=} Q$ conditionally on $\mathcal{F}$. Since $H$ is $\mathcal{F}$-measurable, it is fixed after conditioning on $\mathcal{F}$, and therefore $OQH \overset{d}{=} QH$ conditionally on $\mathcal{F}$. Thus the conditional law of $QH$ is invariant under left multiplication by every deterministic orthogonal matrix $O \in \mathbb{O}(d)$. By uniqueness of the left-orthogonally invariant probability measure on $\mathrm{St}(d, \ell)$ (e.g. Chikuse, 2003, Theorem 1.2.2 and Section 1.3.1), this conditional law is the Haar measure on $\mathrm{St}(d, \ell)$.

Let $G \in \mathbb{R}^{d \times \ell}$ have independent $\mathcal{N}(0, 1)$ entries and be independent of $\mathcal{F}$. By the Gaussian representation of the Haar measure on the Stiefel manifold (Chikuse, 2003, Theorem 2.4.3), $QH \overset{d}{=} G(G^\top G)^{-1/2}$ conditionally on $\mathcal{F}$. Hence

$$\|AQH\|_{\mathrm{op}} \overset{d}{=} \|AG(G^\top G)^{-1/2}\|_{\mathrm{op}} \le \frac{\|AG\|_{\mathrm{op}}}{\sigma_\ell(G)}$$

conditionally on $\mathcal{F}$. By Lemma 21, applied with $X = A^\top$, $Y = I_\ell$, $\sigma = 1$, and $\delta = e^{-t^2/2}$, we have $\|AG\|_{\mathrm{op}} \le c_1 \|A\|_{\mathrm{op}} \sqrt{\mathrm{rank}(A) + \ell + t^2}$ with probability at least $1 - e^{-t^2/2}$. Also, the standard lower-tail bound for the smallest singular value of a Gaussian matrix (Davidson and Szarek, 2001, Theorem II.13) gives $\sigma_\ell(G) \ge \sqrt{d} - \sqrt{\ell} - t$ with probability at least $1 - e^{-t^2/2}$. On the intersection of these two events, which has probability at least $1 - 2e^{-t^2/2}$, the bound (54) follows.

Finally, if $\sqrt{\ell} + t \le \sqrt{d}/2$, then $\sqrt{d} - \sqrt{\ell} - t \ge \sqrt{d}/2$. Substituting this lower bound into (54) gives (55). □