

Estimating Heterogeneous Teacher Effects with Grouped Three-Way Interactive Fixed Effects

Sangmyung Ha

October 19, 2026

Abstract

We develop a grouped three-way interactive fixed-effects model for estimating heterogeneous teacher effects in multidimensional panel data. We establish aggregate consistency of the global least-squares estimator, derive stochastic expansions and limiting distributions for the slope estimator and aligned factor estimates, and characterize three mode-specific incidental-parameter bias terms. Under the maintained identification and strong factor conditions, slope and aggregate factor consistency permit one dimension to remain fixed when the other two diverge. An application to matched student–teacher achievement data reveals that teacher–group matching accounts for approximately 27% of fitted relative teacher value added variation. Using adaptive LASSO, we identify six CLASS, MQI, or teacher survey items that explain the teacher-level effect, and one teacher survey item that explains the matching effect. We also find that teacher rankings are sensitive to classroom assignment and that substantial gains are potentially attainable from improved teacher– student matching.

JEL: C33, C55, I21

1 Introduction

Conventional teacher value-added specifications often summarize teacher effects using a scalar coefficient, even though effects may differ across student types and years. For example, one teacher may be especially effective with students who enter the year with low prior achievement, while another may be more effective with initially high-achieving students. Recent studies, such as Ahn et al. (2025), Bates et al. (2025), Biasi et al. (2026), and Delgado (2025), have begun to investigate heterogeneous teacher value added. Estimating this heterogeneity remains challenging because teacher–student–type–year cells often contain few observations. This paper develops a grouped three-way interactive fixed-effects model to estimate heterogeneous teacher effects, and uses these estimates to study matching effects between teachers and students. The grouped three-way interactive structure in the fixed effects is especially useful in this setting because it pools information across different teacher–student–type–year cells by imposing economically meaningful restrictions.

We use math scores from fourth- and fifth-grade classrooms from the National Center for Teacher Effectiveness to estimate the teacher value-added.¹ To study teacher-student matching effects, we divide students into five distinct groups based on prior academic achievements.

¹A causal value-added interpretation requires teacher assignment to be conditionally independent of unobserved determinants of achievement.

We find that the matching effect accounts for approximately 27% of fitted teacher value-added variation, which is much higher than the value 12% found in Ahn et al. (2025).²

To examine the characteristics that explain the variations in the teacher’s level and matching effects, we use adaptive LASSO on 55 candidate regressors. We use the estimated teacher level effects and the difference in matching effects between the lowest and highest prior-achievement groups as the dependent variables. The difference in matching effects captures the comparative advantage of a teacher in instructing students in the lower prior achievement group. We find that higher teacher-level effects are associated with classroom behavior management, mathematical precision, cognitively engaging instruction, and not reducing demanding mathematics problems in order to prepare for state tests. For example, a one-point increase in the teacher’s skill in behavior management (a CLASS item measured on a 1–7 scale) is associated with 0.0328σ increase in level effects, and a one-point increase in mathematical language or notation imprecision (an MQI item measured on a 1–3 scale) is associated with a 0.0655σ decrease in level effects.

On the other hand, we find that higher comparative advantage in teaching the lower-achieving students is associated with the frequency of assigning problems requiring a few minutes of thought. One possible interpretation is that assigning such problems provides opportunities for deliberate mathematical reasoning that are particularly well matched to lower-achieving students’ learning needs.

Using counterfactual exercises, we assess the sensitivity of teacher rankings to classroom assignment. We find that 27.59% of the teachers that were ranked at the bottom 5% in teaching effectiveness escape the bottom 5% when reassigned to their best-matched classrooms within the same school. Even when every teacher is assigned to their individual best classrooms and the bottom 5% threshold is recalculated, 17.24% of the teachers in the original bottom 5% would no longer be the bottom 5%.

We also quantify the potential gains from reallocating teachers to different classrooms, holding classroom rosters fixed. We find that there are substantial potential gains from improved teacher–student matching.³ The students in the lowest prior achievement group have the most potential gains. For example, by reassigning only 25% of teachers to better-matched classrooms within the same school, students’ standardized test scores can increase by 0.037σ on average. For the students in the lowest quintile, the improvement is 0.097σ on average.

We also find that imposing the three-way separability structure within an empirical Bayes framework can modestly improve out-of-sample forecasts of teacher-level effects relative to an unrestricted time-varying empirical Bayes specification. We find no consistent improvement in forecasting matching effects. Out-of-sample R^2 for matching effects is low across specifications, although these values should be interpreted cautiously because the benchmark targets are unrestricted group-specific teacher-effect estimates that may contain substantial sampling error.

Aside from the empirical results described above, this paper makes four theoretical contributions. First, we develop a grouped three-way separable specification in the fixed effects in panel models and establish consistency of the least-squares estimator. The sample growth

²This comparison is descriptive. Differences in samples, model specifications, and variance definitions limit direct comparability. Our statistic uses plug-in sample variances of the estimated teacher-level and matching components and may therefore reflect estimation noise. In contrast, Ahn et al. (2025) estimate population variance components using a Gaussian-mixture model that accounts for sampling error in the teacher-coefficient estimates. The time-invariant teacher coefficients in Ahn et al. (2025) may also understate matching effects when the true effects vary over time.

³These findings are consistent with Ahn et al. (2025)

condition needed for consistency permits one dimension to remain fixed when the other two diverge.

Second, we derive stochastic expansions and limiting distributions for the slope estimator and aligned factor estimates. The slope expansion identifies three mode-specific incidental-parameter bias components. Grouping and separability reduce the order of the incidental-parameter bias term associated with the cross-sectional units. We also construct a feasible bias correction when the errors are conditionally uncorrelated.

Third, we propose an information criterion for selecting the number of trilinear components in the fixed effects, following the penalized-fit approach of Bai and Ng (2002). Fourth, we extend the least squares consistency analysis to the inclusion of additive fixed effects. While demeaning observations across each mode removes the additive fixed effects, estimating the additive fixed effects directly using least squares can avoid information loss and improve precision.

The framework studied in this paper can also be applied to other settings with observed types and multidimensional data, including worker-type–firm, consumer-type–products, product-type–market, and patient-type–hospital panels. Its usefulness depends on whether grouping and separability provide an appropriate representation of the latent heterogeneity.

Related literature. This paper is related to four strands of literature. The first strand studies heterogeneous teacher effectiveness and student–teacher matching. Biasi et al. (2026) divide students into low- and high-prior-achievement groups according to whether lagged achievement lies below or above the grade-specific state median. They estimate two teacher-specific value-added measures using OLS and empirical Bayes shrinkage and use them as teacher attributes in a structural model of district wage setting and teacher sorting. Ahn et al. (2025) do not impose a single mutually exclusive grouping. Instead, they allow teacher value added to depend on continuous and discrete student attributes, e.g., prior achievement, gender, or free or reduced-price lunch status, by allowing teacher-specific slope coefficients. They estimate the model using a two-step fixed-effects and Gaussian-mixture procedure. Bates et al. (2025) separate economically disadvantaged and advantaged students, construct best linear predictions of group-specific teacher value added from prior-year outcomes, and incorporate these measures into a two-sided matching model. Delgado (2025) estimates reliability-weighted teacher effects for Black and non-Black students, with extensions based on race-by-income and race-by-gender groups.

These studies estimate subgroup-specific teacher effects or teacher-specific coefficients on observed student attributes. In contrast, we represent the full student-group–teacher–year effect using a fully interactive, or three-way separable, structure. This structure pools information across student groups, teachers, and years. We estimate this component jointly with the slope coefficients and provide asymptotic and rank-selection theory for the least squares estimator.

The second studies interactive fixed effects and multifactor error structures in panel data. Pesaran (2006) proposes the common correlated effects estimator for heterogeneous panels with unobserved common factors, while Bai (2009) develops least-squares estimation and inference for large- (N) , large- (T) panels. Moon and Weidner (2015) study inference when the number of factors used in estimation may exceed the true number. More recently, Freeman (2026) develops transformation-based slope estimators for panels with three or more dimensions under separable interactive fixed effects, treating the latent effects as nuisance parameters. Our focus is the grouped separable restriction and the joint least-squares estimation, including the estimation of the latent component itself. We establish asymptotic

theory for both the slope coefficients and the fixed effects. Our analysis can naturally be extended to models with more dimensions as in Freeman (2026).

The third strand studies grouped and structured heterogeneity. Bonhomme and Manresa (2015) assign units to latent groups whose time-varying fixed effects are common within groups but differ across groups. Our model likewise pools observations through a group structure, but differs in two respects. First, the grouped heterogeneity enters multiplicatively through a low-rank tensor component rather than through group-specific additive time effects. Second, group membership is observed and treated as known in our analysis.

The fourth strand is the literature on tensor factor models with Canonical Polyadic (CP) structure on the common component. Chen et al. (2026) and Chang et al. (2026) develop CP tensor factor models for high-dimensional tensor data where the common component is identified using a pervasiveness assumption on the common component. Our contribution differs by embedding the CP structure in a regression model with observed covariates, allowing the covariates to be correlated with the latent component, and focusing on estimation and inference for the slope coefficient as well as the grouped interactive effects.

The remainder of the paper is organized as follows. Section 2 introduces the grouped three-way interactive fixed effects model. Section 3 defines the least-squares estimator and describes its alternating-least-squares implementation. Section 4 states the assumptions and Section 5 develops the asymptotic theory. Section 6 proposes an information criterion for selecting the number of trilinear components. Section 7 extends the model to the inclusion of additive fixed effects. Section 8 compares finite-sample performance with alternative estimators. Section 9 presents the teacher-value-added application, and Section 10 concludes.

Notations and definitions of different norms, products, and matricizations can be found in Appendix 11.1.

2 Model

We consider a balanced three-way array indexed by $i = 1, \dots, N$, $j = 1, \dots, M$, and $t = 1, \dots, T$. The indices may represent units, locations, products, cohorts, or other observational positions, depending on the application. The observed outcome satisfies

$$y_{ijt} = \delta' x_{ijt} + \pi_{ijt} + \varepsilon_{ijt}, \quad (1)$$

where $x_{ijt} \in \mathbb{R}^p$ is a vector of observed covariates, $\delta \in \mathbb{R}^p$ is the corresponding coefficient vector, ε_{ijt} is an idiosyncratic error, and π_{ijt} is an unobserved interactive fixed-effect component. We specify the interactive component as

$$\pi_{ijt} = \sum_{k=1}^{\ell} u_{g(i),k} v_{jk} w_{tk}, \quad (2)$$

where $g : \{1, \dots, N\} \rightarrow \{1, \dots, G\}$ assigns each first-mode unit i to one of G groups. Group membership is observed and treated as known throughout the analysis.⁴ Accordingly, units belonging to the same group share a common loading vector of first-mode factor loadings. This restriction reduces the number of first-mode loading parameters from $N\ell$ to $G\ell$, yielding

⁴The analysis conditions on the group assignments. When group membership is latent, it may instead be estimated jointly with the remaining parameters using a grouped fixed-effects procedure such as that of Bonhomme and Manresa (2015). Analyzing the asymptotic consequences of estimating group membership is beyond the scope of this paper.

substantial parsimony when $G \ll N$. The specification in (2) is multiplicatively separable across the three modes and therefore admits a CP representation with at most ℓ rank-one components. The grouping restriction imposes repeated rows in the first-mode loading matrix. In the special case $G = N$ and $g(i) = i$ for every i , each first-mode unit has its own loading vector, so the grouping restriction disappears and (2) reduces to $\pi_{ijt} = \sum_{k=1}^{\ell} u_{ik} v_{jk} w_{tk}$.

Collecting the observations into tensors, let

$$Y = (y_{ijt}), \quad E = (\varepsilon_{ijt}) \in \mathbb{R}^{N \times M \times T},$$

and let

$$X = (x_{ijqt}) \in \mathbb{R}^{N \times M \times T \times p}, \quad x_{ijt} = (x_{ijqt})_{q=1}^p.$$

Suppressing the singleton fourth mode, the mode-4 product of X and δ' satisfies⁵

$$X \times_4 \delta' := (\delta' x_{ijt})_{i,j,t=1}^{N,M,T} \in \mathbb{R}^{N \times M \times T}.$$

Define the membership matrix $D \in \{0, 1\}^{N \times G}$ by $D_{ih} = 1\{g(i) = h\}$, $i = 1, \dots, N$, $h = 1, \dots, G$. Let $N_g = \sum_{i=1}^N 1\{g(i) = g\}$, $g = 1, \dots, G$ denote the number of samples in group g . We assume that the group membership is exclusive so that

$$D \iota_G = \iota_N, \quad D' D = \text{diag}(N_1, \dots, N_G),$$

where ι_G, ι_N are G and N dimensional vectors of ones. For $k = 1, \dots, \ell$, collect the unobserved fixed effects in each mode as

$$\begin{aligned} u_k &= (u_{1k}, \dots, u_{Gk})' \in \mathbb{R}^G \\ v_k &= (v_{1k}, \dots, v_{Mk})' \in \mathbb{R}^M \\ w_k &= (w_{1k}, \dots, w_{Tk})' \in \mathbb{R}^T. \end{aligned}$$

The model (1) can be written in tensor form as

$$Y = X \times_4 \delta' + \sum_{k=1}^{\ell} (Du_k \circ v_k \circ w_k) + E,$$

where fixed effects are written as outer products, reflecting a multiplicatively separable structure across the three modes. Let $\otimes$ denote the Kronecker product. Then, the model also admits the vectorized representation

$$y = X'_{(4)} \delta + \sum_{k=1}^{\ell} (Du_k \otimes v_k \otimes w_k) + e,$$

where $y = \text{vec}(Y)$, $e = \text{vec}(E)$, and $X_{(4)}$ denotes mode-4 matricization of X satisfying $\text{vec}(X \times_4 \delta') = X'_{(4)} \delta$.⁶ In particular, $X'_{(4)} \in \mathbb{R}^{NMT \times p}$ collects $(x_{ijt})_{i,j,t}$ in its rows.

⁵The definitions and notations used for different products used in this paper are described in Appendix 11.1.

⁶The definitions and notations used for different matricizations used in this paper are described in Appendix 11.1.

Example: Matched student–teacher data. Let $j = 1, \dots, M$ index teachers and $t = 1, \dots, T$ index academic periods. We interpret each teacher–year cell (j, t) as a classroom taught by teacher j in year t . In the balanced benchmark, each classroom contains N students, indexed locally by $i = 1, \dots, N$. The index i denotes a within-classroom position rather than a persistent student identifier. Consequently, the student occupying position i may differ across classrooms and periods, and the $N \times M \times T$ array does not require the same students to be followed across teachers or over time. The grouped specification is particularly useful in this setting because it links otherwise incomparable local student positions through observed prior-achievement groups.

We group the students into G observed groups according to prior achievement. Note that the group assignment $g : \{1, \dots, N\} \rightarrow \{1, \dots, G\}$ does not depend on (j, t) . Let N_g denote the number of students in group g within each classroom, where $\sum_{g=1}^G N_g = N$. The homogeneous group assignment assumes that every teacher–period classroom contains exactly N_g students from prior-achievement group g .⁷

Interpretation of the separable structure. Suppose $\pi_{gjt} = \sum_{k=1}^{\ell} u_{gk} v_{jk} w_{tk}$ is interpreted as teacher j ’s value-added for a given group g in year t . A natural mechanism that can impose this separable structure in the teaching effects is common changes in the productivity or importance of teaching inputs, combined with stable group and teacher responses. Suppose there are ℓ teaching inputs or activities. Teacher j ’s effective provision of activity k in year t is

$$e_{jtk} = v_{jk} w_{tk},$$

where v_{jk} is the teacher’s proficiency and w_{tk} is a common change in the availability, productivity, or importance of that activity. If group g ’s return to the activity is u_{gk} , then

$$\pi_{gjt} = \sum_{k=1}^{\ell} u_{gk} e_{jtk} = \sum_{k=1}^{\ell} u_{gk} v_{jk} w_{tk}.$$

For example, suppose:

- One activity is remedial instruction: lower-achieving groups benefit more, teachers differ in their remedial teaching skills, and a common policy changes its intensity over time.
- Another is advanced problem-solving instruction: groups have different returns, teachers have different proficiencies, and curriculum changes alter its importance.
- Digital resources provide another possible channel: common changes in technology interact with teacher proficiency and group-specific benefits.

Comparison with a flattened model. The proposed structure differs from flattening (i, j) into a single cross-sectional index and applying the two-way interactive fixed-effects framework of Bai (2009). The flattened specification is

$$y_{ijt} = x'_{ijt} \delta + \sum_{k=1}^{\ell} \lambda_{ij,k} f_{tk} + \varepsilon_{ijt},$$

⁷When classroom sizes or group compositions vary across teacher–period cells, the resulting sample is unbalanced. While the benchmark model assumes a balanced sample, we discuss the unbalanced case in Appendix 11.10.

with unrestricted loadings across the NM pairs. Our specification imposes $\lambda_{ij,k} = u_{g(i),k}v_{jk}$. For fixed rank, the number of parameters therefore scales with $G + M + T$, compared with $NM + T$ in the unrestricted flattened model. The restriction also links the fitted effects across economically meaningful dimensions. Group loadings u_{gk} describe how student types respond to each latent dimension k , teacher loadings v_{jk} capture teachers' contributions along that dimension, and year loadings w_{tk} allow these interactions to evolve over time.

The grouped separable structure also changes the number of observations contributing to each loading estimate. Let N_g denote the number of units i in the group g with $\sum_{g=1}^G N_g = N$. In estimating u_g , we use information in the variations of $N_g MT$ samples. Similarly, in estimating v_j and w_t , we use variations of NT and NM samples respectively. In contrast, an unrestricted flattened loading λ_{ij} is estimated using observations along the time dimension alone, with only variations in T samples. The grouped separable structure can therefore be particularly useful when the panel has a short time dimension.

3 Estimation

For a given number ℓ of interactive components, we estimate the model by minimizing the average squared residual

$$\begin{aligned} S_{NMT}(\delta, U, V, W) &= \frac{1}{NMT} \left\| Y - X \times_4 \delta' - \sum_{k=1}^{\ell} (Du_k \circ v_k \circ w_k) \right\|_F^2 \\ &= \frac{1}{NMT} \sum_{i=1}^N \sum_{j=1}^M \sum_{t=1}^T \left(y_{ijt} - x'_{ijt} \delta - \sum_{k=1}^{\ell} u_{g(i),k} v_{jk} w_{tk} \right)^2 \end{aligned} \quad (3)$$

with respect to (δ, U, V, W) .

Provided that the relevant Gram matrices are nonsingular, the first-order condition with respect to δ can be written as

$$\hat{\delta} = \left(X_{(4)} X'_{(4)} \right)^{-1} X_{(4)} \left(y - \sum_{k=1}^{\ell} (D\hat{u}_k \otimes \hat{v}_k \otimes \hat{w}_k) \right). \quad (4)$$

Similarly, the first-order conditions with respect to the factor matrices can be expressed as

$$\begin{aligned} \hat{U} &= (D' D)^{-1} D' \left(Y - X \times_4 \hat{\delta}' \right)_{(1)} (\hat{V} * \hat{W}) \left((\hat{V} * \hat{W})' (\hat{V} * \hat{W}) \right)^{-1}, \\ \hat{V} &= \left(Y - X \times_4 \hat{\delta}' \right)_{(2)} (D\hat{U} * \hat{W}) \left((D\hat{U} * \hat{W})' (D\hat{U} * \hat{W}) \right)^{-1}, \\ \hat{W} &= \left(Y - X \times_4 \hat{\delta}' \right)_{(3)} (D\hat{U} * \hat{V}) \left((D\hat{U} * \hat{V})' (D\hat{U} * \hat{V}) \right)^{-1}. \end{aligned} \quad (5)$$

Equations (4) and (5) jointly characterize a stationary point of the least-squares criterion.

We compute a solution to these simultaneous first-order conditions using an alternating least squares (ALS) algorithm. Starting from initial values, ALS successively applies the blockwise solution maps in (4) and (5), using the most recently available values of the remaining parameter blocks. Because each block update minimizes the criterion while holding the other blocks fixed, the sequence of objective values is nonincreasing and therefore converges. Under suitable boundedness and regularity conditions, every accumulation

point of the parameter sequence is a stationary point of the least-squares criterion. Because the criterion is nonconvex, global optimality is not guaranteed. We therefore use multiple initializations and retain the solution with the smallest residual sum of squares.

Projected representation. For given (U, V, W) , define

$$\mathcal{T}(U, V, W) = \left\{ \sum_{k=1}^{\ell} (Da_k \otimes v_k \otimes w_k + Du_k \otimes b_k \otimes w_k + Du_k \otimes v_k \otimes c_k) : \right. \\ \left. a_k \in \mathbb{R}^G, \quad b_k \in \mathbb{R}^M, \quad c_k \in \mathbb{R}^T, \quad k = 1, \dots, \ell \right\}. \quad (6)$$

The minimizer $\hat{\delta}$ of (3) can also be characterized as

$$\hat{\delta} = \left(X_{(4)} M_{\hat{\mathcal{T}}} X'_{(4)} \right)^{-1} X_{(4)} M_{\hat{\mathcal{T}}} y, \quad \hat{\mathcal{T}} = \mathcal{T}(\hat{U}, \hat{V}, \hat{W}).$$

This is a Frisch–Waugh–Lovell-type representation that projects out the variations in the unobserved fixed effects.⁸

4 Assumptions

Let

$$\mathcal{F} = \sigma(X, D, U, V, W), \quad \mathcal{I} = \{1, \dots, N\} \times \{1, \dots, M\} \times \{1, \dots, T\}.$$

For $a = (i_a, j_a, t_a) \in \mathcal{I}$, write $\varepsilon_a = \varepsilon_{i_a j_a t_a}$. Suppose that the following conditions hold.

Assumption 1 (Asymptotic Framework, Normalization, and Estimation). *Along an asymptotic sequence satisfying*

$$\min\{NMT/G, NT, NM\} \rightarrow \infty,$$

the dimensions p and ℓ remain fixed. The membership matrix $D \in \{0, 1\}^{N \times G}$ has full column rank, and

$$1 \leq \ell \leq \min\{G, M, T\}.$$

For a constant $c_U > 0$ independent of the panel dimensions, let Θ_ℓ be the set of triples $(U, V, W) \in \mathbb{R}^{G \times \ell} \times \mathbb{R}^{M \times \ell} \times \mathbb{R}^{T \times \ell}$ satisfying

$$\text{diag} \left(\frac{1}{M} V' V \right) = \text{diag} \left(\frac{1}{T} W' W \right) = \iota_\ell, \quad \lambda_{\max} \left(\frac{1}{N} U' D' D U \right) \leq c_U, \quad (7)$$

where ι_ℓ is the ℓ -dimensional vector of ones. The true factor matrices admit a representation in Θ_ℓ . The least-squares estimator satisfies

$$(\hat{\delta}, \hat{U}, \hat{V}, \hat{W}) \in \underset{\delta \in \mathbb{R}^p, (U, V, W) \in \Theta_\ell}{\text{argmin}} S_{NMT}(\delta, U, V, W).$$

⁸The derivation is in Appendix 11.2

Assumption 2 (Strong Factors and Stochastic Bounds). *The regressor x_{ijt} has uniformly bounded fourth moment conditional on group assignment, i.e., there exists a constant $c_x < \infty$ independent of the panel dimensions,*

$$\sup_{(i,j,t) \in \mathcal{I}} \mathbb{E}[\|x_{ijt}\|^4 | D] \leq c_x$$

almost surely, uniformly along the asymptotic sequence. The grouped-error satisfies

$$\|E^G\| = O_p(\sqrt{G + M + T}),$$

where

$$E^G = E \times_1 (D'D)^{-1/2} D' = \left(\frac{1}{\sqrt{N_g}} \sum_{\substack{i=1 \\ g(i)=g}}^N \varepsilon_{ijt} \right)_{g,j,t=1}^{G,M,T}.$$

For constants $0 < c_0 \leq 1 \leq c_1 < c_U < \infty$, with probability approaching one,

$$\begin{aligned} c_0 I_\ell &\leq \frac{1}{N} U' D' D U \leq c_1 I_\ell \\ c_0 I_\ell &\leq \frac{1}{M} V' V \leq c_1 I_\ell \\ c_0 I_\ell &\leq \frac{1}{T} W' W \leq c_1 I_\ell. \end{aligned}$$

Assumption 3. *The errors have uniformly finite fourth moments and $\mathbb{E}[\varepsilon_a | \mathcal{F}] = 0$. Let $\Gamma_{ab} = \mathbb{E}[\varepsilon_a \varepsilon_b | \mathcal{F}]$ for $a, b \in \mathcal{I}$. For some finite constant c_ε , with probability approaching one,*

$$\sup_{a \in \mathcal{I}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \leq c_\varepsilon.$$

Define

$$\mathcal{R}^D = \{(a, b) \in \mathcal{I}^2 : g(i_a) = g(i_b) \text{ or } j_a = j_b \text{ or } t_a = t_b\}, \quad z_{ab} = \varepsilon_a \varepsilon_b - \Gamma_{ab}.$$

For some finite constant c_4 , with probability approaching one,

$$\sup_{(a,b) \in \mathcal{R}^D} \sum_{(c,d) \in \mathcal{R}^D} |\text{Cov}(z_{ab}, z_{cd} | \mathcal{F})| \leq c_4.$$

Assumption 4 (Identification of δ). *There exists a constant $c_\delta > 0$ such that, with probability approaching one,*

$$\inf_{\tilde{U} \in \mathbb{R}^{G \times \ell}, \tilde{V} \in \mathbb{R}^{M \times \ell}} \lambda_{\min} \left[\frac{1}{NMT} X_{(4)} (M_{D\tilde{U}*} \tilde{V} \otimes M_W) X'_{(4)} \right] \geq c_\delta,$$

where W denotes the true factors.

Assumption 1 describes the asymptotic framework. In particular, not all N, M, T need to diverge as long as $\min\{NMT/G, NT, NM\} \rightarrow \infty$. Additional rate conditions will be imposed when necessary. We define the estimator as a global minimizer satisfying the constraints in Θ_ℓ . We show in Lemma 1 that the constraints do not bind so that the first-order conditions (5) hold with probability approaching one.

Assumption 2 implies that the factors are strong and the within-group averaged idiosyncratic errors are non-pervasive. In particular, the tensor spectral norm bound on E^G can be shown to hold when the idiosyncratic errors are independent with uniformly bounded fourth moments. Assumption 3 imposes weak serial and cross-sectional correlation. For conditionally centered Gaussian errors, the Gaussian fourth-moment identity, e.g., fourth-order case of Isserlis' formula, gives

$$\text{Cov}(z_{ab}, z_{cd} \mid \mathcal{F}) = \Gamma_{ac}\Gamma_{bd} + \Gamma_{ad}\Gamma_{bc}$$

It follows that

$$\begin{aligned} \sum_{(c,d) \in \mathcal{R}^D} |\text{Cov}(z_{ab}, z_{cd} \mid \mathcal{F})| &\leq \sum_{c,d} (|\Gamma_{ac}\Gamma_{bd}| + |\Gamma_{ad}\Gamma_{bc}|) \\ &\leq 2c_\varepsilon^2. \end{aligned}$$

Thus, for this class, the fourth-order condition follows from the second-order row-sum bound.

Assumption 4 requires sufficient regressor variation to remain after projecting the (i, j) -mode away from an arbitrary candidate structured loading space $(D\tilde{U}*\tilde{V})$ and projecting the time mode away from the true factor space W . Equivalently, with probability approaching one,

$$\frac{1}{NMT} \left\| M_W (X \times_4 a')_{(3)} M_{D\tilde{U}*\tilde{V}} \right\|_F^2 \geq c_\delta \|a\|^2$$

uniformly over $a \in \mathbb{R}^p$, $\tilde{U} \in \mathbb{R}^{G \times \ell}$, and $\tilde{V} \in \mathbb{R}^{M \times \ell}$. Thus, the assumption rules out changes in δ that can be absorbed by changing the candidate interactive component. Uniformity over $(\tilde{U}, \tilde{V})$ ensures that the lower bound holds at the jointly estimated loading space.

Remark: Comparing identification of δ with a flattened two-way interactive fixed effects model. Suppose instead that the indices (i, j) are enumerated into a single cross-sectional index and the model is estimated as the unrestricted two-way interactive fixed-effects model. Because the two-way least-squares problem is symmetric in its cross-sectional and time dimensions, the analogue of Bai (2009)'s Assumption A after interchanging the cross-sectional and time dimensions is given by

$$\inf_{H \in \mathbb{R}^{NM \times \ell}} \lambda_{\min} \left[\frac{1}{NMT} X_{(4)} (M_H \otimes M_W) X'_{(4)} \right] > 0$$

with probability approaching one, where W denotes the true factors. This condition is stronger than Assumption 4. Indeed, every structured loading matrix $D\tilde{U}*\tilde{V}$ is an element of $\mathbb{R}^{NM \times \ell}$, whereas an unrestricted $H \in \mathbb{R}^{NM \times \ell}$ need not have separable or grouped columns. To illustrate, consider a scalar regressor of the form

$$x_{ijt} = q_{ij}\tau_t.$$

Let

$$Q = (q_{ij})_{i,j=1}^{N,M} \in \mathbb{R}^{N \times M}, \quad q = \text{vec}(Q) \in \mathbb{R}^{NM}, \quad \tau = (\tau_1, \dots, \tau_T)' \in \mathbb{R}^T.$$

For any candidate cross-sectional loading matrix H ,

$$\frac{1}{NMT} X_{(4)} (M_H \otimes M_W) X'_{(4)} = \left(\frac{q' M_H q}{NM} \right) \left(\frac{\tau' M_W \tau}{T} \right).$$

If $H = D\tilde{U} * \tilde{V}$, every vector in the column space of H , when reshaped as an $N \times M$ matrix, has matrix rank at most ℓ . Therefore, by the Eckart–Young inequality,

$$\begin{aligned}
q' M_{D\tilde{U} * \tilde{V}} q &= \|q - P_{D\tilde{U} * \tilde{V}} q\|_F^2 \\
&= \inf_{a \in \mathbb{R}^\ell} \left\| Q - \sum_{k=1}^{\ell} a_k (D\tilde{u}_k) \tilde{v}_k' \right\|_F^2 \\
&\geq \inf_{\text{rank}(A) \leq \ell} \|Q - A\|_F^2 \\
&= \sum_{r=\ell+1}^{\min(N,M)} \sigma_r^2(Q).
\end{aligned}$$

Consequently, Assumption 4 holds in this example if

$$\liminf_{N,M \rightarrow \infty} \frac{1}{NM} \sum_{r=\ell+1}^{\min(N,M)} \sigma_r^2(Q) > 0, \quad \liminf_{T \rightarrow \infty} \frac{\tau' M_W \tau}{T} > 0.$$

By contrast, under the unrestricted two-way condition, one may choose $H \in \mathbb{R}^{NM \times \ell}$ with $M_H q = 0$ so that

$$\inf_{H \in \mathbb{R}^{NM \times \ell}} \frac{1}{NMT} X_{(4)} (M_H \otimes M_W) X_{(4)}' = 0.$$

Thus, the grouped separable structure can identify δ for regressors that may not be identified after flattening the (i, j) indices and allowing an unrestricted $NM \times \ell$ loading matrix.

Remark: Implication for the teacher value-added model. Suppose we estimate a conventional teacher value-added model with teacher-year fixed effects,

$$y_{ijt} = x_{ijt}' \delta + v_{jt} + \varepsilon_{ijt}.$$

A well-known feature of this model is that teacher-year characteristics, such as teaching experience, cannot be included as regressors because they are absorbed into the teacher-year fixed effects.⁹ Under the three-way separable model, teacher covariates are identified in general, as long as they cannot be absorbed into the separable loading structure, and can directly enter as regressors.

5 Asymptotic Theory

The following lemma gives the consistency and convergence rate of the interactive fixed effect components (U, V, W) .

Lemma 1 (Normalized factor alignment and asymptotic inactivity of the inequality constraint). *Suppose that Assumptions 1, 2, 3, and 4 hold. Let (U, V, W) denote the true factor matrices in the normalized representation satisfying (7). Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be the global least-squares estimator defined in Assumption 1. There exist possibly random $\ell \times \ell$ permutation*

⁹Kwon (2026) instead considers using teacher covariates to construct a target location for shrinkage estimators.

and diagonal sign matrices Π_ℓ, S_v, S_w , with the diagonal entries of S_v and S_w belonging to $\{-1, 1\}$, such that

$$\frac{\|D(\hat{U}\Pi_\ell S_v S_w - U)\|_F}{\sqrt{N}} + \frac{\|\hat{V}\Pi_\ell S_v - V\|_F}{\sqrt{M}} + \frac{\|\hat{W}\Pi_\ell S_w - W\|_F}{\sqrt{T}} = O_p\left(\sqrt{\frac{G+M+T}{NMT}}\right). \quad (8)$$

Define the residual tensor

$$\hat{E} = Y - X \times_4 \hat{\delta}' - \sum_{k=1}^{\ell} (D\hat{u}_k \circ \hat{v}_k \circ \hat{w}_k).$$

With probability approaching one, the estimator satisfies the unrestricted normal equations

$$\begin{aligned} X_{(4)} \text{vec}(\hat{E}) &= 0, & D' \hat{E}_{(1)} (\hat{V} * \hat{W}) &= 0, \\ \hat{E}_{(2)} (D\hat{U} * \hat{W}) &= 0, & \hat{E}_{(3)} (D\hat{U} * \hat{V}) &= 0. \end{aligned} \quad (9)$$

Lemma 1 shows that the regression equations (5) hold with probability approaching one. Also, the estimated fixed effect components $(\hat{U}, \hat{V}, \hat{W})$ are consistent up to permutations and sign changes. This is consistent with the inherent indeterminacies of the CP decomposition as studied by Kruskal (1977). The following Proposition 1 gives the stochastic expansion for $\hat{\delta}$ needed to apply the central limit theorem.

Proposition 1. Suppose that Assumptions 1, 2, 3, and 4 hold. Let $\mathcal{T}$ be the tangent space defined in (6) evaluated at the true parameters (U, V, W) . Let

$$\Gamma_{NMT} = \frac{1}{NMT} X_{(4)} M_{\mathcal{T}} X'_{(4)}, \quad \mathcal{Z}_{NMT} = \frac{1}{\sqrt{NMT}} X_{(4)} M_{\mathcal{T}} \text{vec}(E).$$

The least squares estimator $\hat{\delta}$ defined in Assumption 1 satisfies

$$\Gamma_{NMT} \sqrt{NMT} (\hat{\delta} - \delta) = \mathcal{Z}_{NMT} - \mathcal{B}_1 - \mathcal{B}_2 - \mathcal{B}_3 + R_{NMT}, \quad (10)$$

where the remainder term R_{NMT} has the bound $O_p\left(\sqrt{\frac{G+M+T}{NMT}} + \frac{(G+M+T)^{3/2}}{NMT}\right)$ and the mode specific bias terms are given by

$$\begin{aligned} \mathcal{B}_1 &= \frac{1}{\sqrt{NMT}} X_{(4)} P_{\mathcal{T}} Q \text{vec}(\bar{U}(\bar{U}'\bar{U})^{-1}((V * W)'(V * W))^{-1}(V * W)'\Omega_1^G M_{V*W}) \\ \mathcal{B}_2 &= \frac{1}{\sqrt{NMT}} X_{(4)} P_{\mathcal{T}} Q K_{1,2}^G \text{vec}(V(V'V)^{-1}((\bar{U} * W)'(\bar{U} * W))^{-1}(\bar{U} * W)'\Omega_2^G M_{\bar{U}*W}) \\ \mathcal{B}_3 &= \frac{1}{\sqrt{NMT}} X_{(4)} P_{\mathcal{T}} Q K_{1,3}^G \text{vec}(W(W'W)^{-1}((\bar{U} * V)'(\bar{U} * V))^{-1}(\bar{U} * V)'\Omega_3^G M_{\bar{U}*V}). \end{aligned}$$

Here, $\bar{U} = (D'D)^{1/2}U$ and $\Omega_m^G = \mathbb{E}[(E_{(m)}^G)'E_{(m)}^G \mid \mathcal{F}]$ for $m = 1, 2, 3$. The matrix Q is defined as

$$Q_D = D(D'D)^{-1/2}, \quad Q = Q_D \otimes I_M \otimes I_T$$

and $P_{\mathcal{T}}$ is projection on (6). and $K_{1,m}^G$ denotes the mode-permutation matrix for tensors of dimensions $G \times M \times T$.¹⁰ All factor-dependent matrices and projectors in the displayed bias terms are evaluated at the true parameters. Moreover,

$$\mathcal{B}_1 = O_p(G/\sqrt{NMT}), \quad \mathcal{B}_2 = O_p(M/\sqrt{NMT}), \quad \mathcal{B}_3 = O_p(T/\sqrt{NMT}).$$

¹⁰For example, the matrix $K_{1,2}^G$ satisfies $(u \otimes v \otimes w) = K_{1,2}^G(v \otimes u \otimes w)$ for all u, v, w of dimension G, M , and T respectively. A general definition of the mode-permutation matrices is in Appendix 11.1.

Remark: Dimension conditions for negligible expansion remainders The remainder in (10) has the bound

$$R_{NMT} = O_p \left(\sqrt{\frac{G + M + T}{NMT}} + \frac{(G + M + T)^{3/2}}{NMT} \right).$$

In particular, this allows one dimension to remain fixed, as long as the other two dimensions diverge at the same rate. Let any two of N, M, T have the same order as $L \rightarrow \infty$ while the remaining dimension is $O(L)$. Since $G \leq N$ and all dimensions are positive integers, we have

$$\frac{G + M + T}{NMT} = O(L^{-1}) \rightarrow 0, \quad \frac{(G + M + T)^{3/2}}{NMT} = O(L^{-1/2}) \rightarrow 0.$$

This is regardless of which two dimensions diverge.

Remark: Higher-order extensions Consider a model with grouped four-way separable fixed effects,

$$y_{ij\ell t} = x'_{ij\ell t} \delta + \sum_{k=1}^r u_{g(i),k} v_{jk} h_{\ell k} w_{tk} + \varepsilon_{ij\ell t},$$

with total sizes (N, M, L, T) in each dimension. Under suitable extensions to the assumptions given in Section 4, we may derive a similar stochastic expansion as in Proposition 1 with four mode-specific bias terms, and the bound on the remainder term given by¹¹

$$R_{NMLT} = O_p \left(\sqrt{\frac{G + M + L + T}{NMLT}} + \frac{(G + M + L + T)^{3/2}}{NMLT} \right).$$

In particular, the remainder term is asymptotically negligible when two dimension sizes grow at a common rate, and the remaining dimension sizes are fixed. Thus, a fully separable specification on the fixed effects permits a valid limit distribution for the slope parameter without requiring every dimension to diverge.

Remark: Bias reduction from the grouped separable structure The three bias components in (10) are incidental-parameter bias terms associated with estimation of the three sets of factor parameters. More precisely, first-order perturbations of the mode-specific projections M_{V*W} , $M_{\bar{U}*W}$, and $M_{\bar{U}*V}$ generate quadratic error terms whose conditional expectations produce the displayed covariance-based biases. Their bounds imply

$$\mathcal{B}_1 + \mathcal{B}_2 = O_p \left(\frac{G + M}{\sqrt{NMT}} \right), \quad \mathcal{B}_3 = O_p \left(\frac{T}{\sqrt{NMT}} \right).$$

All three bias components are $o_p(1)$ in the $\sqrt{NMT}$ -normalized expansion if

$$\frac{G^2}{NMT} \rightarrow 0, \quad \frac{M}{NT} \rightarrow 0, \quad \frac{T}{NM} \rightarrow 0.$$

These sufficient conditions are equivalent to $(G + M + T)^2/(NMT) \rightarrow 0$. They hold, in particular, when all N , M , and T grows at a comparable rate.

¹¹The relevant assumptions and stochastic expansion can be found in Appendix 11.9.

For comparison, stack (i, j) into a cross-sectional index of dimension NM . The true loading matrix in the resulting two-way representation is $DU * V$. Let $\hat{\delta}_{\text{flat}}$ denote the estimator obtained by allowing an unrestricted $NM \times \ell$ loading matrix. This estimator differs from $\hat{\delta}$, which imposes both grouping and separability. The conventional loading-side and time-factor-side bias scales associated with Bai (2009) are, respectively,

$$O_p \left(\sqrt{\frac{NM}{T}} \right) = O_p \left(\frac{NM}{\sqrt{NMT}} \right), \quad O_p \left(\sqrt{\frac{T}{NM}} \right) = O_p \left(\frac{T}{\sqrt{NMT}} \right).$$

Relative to these scales, the grouped separable structure reduces the loading-side dimensional bias bound from $NM/\sqrt{NMT}$ to $(G+M)/\sqrt{NMT}$, while retaining the same dimensional bound for the time-factor-side bias.

Remark: Conditionally uncorrelated and homoskedastic case If the errors are additionally conditionally uncorrelated and homoskedastic with

$$\mathbb{E}[ee' \mid \mathcal{F}] = \sigma^2 I_{NMT},$$

then

$$\Omega_1^G = G\sigma^2 I_{MT}, \quad \Omega_2^G = M\sigma^2 I_{GT}, \quad \Omega_3^G = T\sigma^2 I_{GM}.$$

The mode-specific bias formulas give

$$\mathcal{B}_1 = \mathcal{B}_2 = \mathcal{B}_3 = 0.$$

Thus conditional uncorrelatedness and homoskedasticity eliminate these leading population bias terms.

To derive the stochastic expansion of the estimated factors and factor loadings, we impose the following assumption.

Assumption 5 (Slice-averaged weighted covariance summability). *Let u_h , v_j , and w_t denote the transposes of the corresponding rows of U , V , and W , respectively. For $b = (i_b, j_b, t_b) \in \mathcal{I}$, define*

$$\omega_b = \|u_{g(i_b)}\|^2 \|v_{j_b}\|^2 + \|u_{g(i_b)}\|^2 \|w_{t_b}\|^2 + \|v_{j_b}\|^2 \|w_{t_b}\|^2.$$

Define the source slices

$$\begin{aligned} \mathcal{I}_\tau^{(t)} &= \{a \in \mathcal{I} : t_a = \tau\}, & |\mathcal{I}_\tau^{(t)}| &= NM, \\ \mathcal{I}_h^{(g)} &= \{a \in \mathcal{I} : g(i_a) = h\}, & |\mathcal{I}_h^{(g)}| &= N_h MT, \\ \mathcal{I}_{j_0}^{(j)} &= \{a \in \mathcal{I} : j_a = j_0\}, & |\mathcal{I}_{j_0}^{(j)}| &= NT. \end{aligned}$$

For every fixed admissible τ , h , and j_0 , suppose that

$$\begin{aligned} \frac{1}{NM} \sum_{a \in \mathcal{I}_\tau^{(t)}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \omega_b &= O_p(1), \\ \frac{1}{N_h MT} \sum_{a \in \mathcal{I}_h^{(g)}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \omega_b &= O_p(1), \\ \frac{1}{NT} \sum_{a \in \mathcal{I}_{j_0}^{(j)}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \omega_b &= O_p(1). \end{aligned}$$

These conditions are pointwise in the slice indices; no uniform bound over τ , h , or j_0 is imposed.

Remark: Role of slice-averaged weighted summability Assumption 5 controls covariance magnitudes weighted by the factor norms that enter higher-order error contractions. Its additional content is control on the selected source slices. To see the distinction from global averaging, write $a_U = \text{tr}(U'D'DU/N)$, $a_V = \text{tr}(V'V/M)$, and $a_W = \text{tr}(W'W/T)$. Note that Assumption 2 and Assumption 3 implies

$$\frac{1}{NMT} \sum_{a \in \mathcal{I}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \omega_b \leq c_\varepsilon \frac{1}{NMT} \sum_{b \in \mathcal{I}} \omega_b = c_\varepsilon (a_U a_V + a_U a_W + a_V a_W) = O_p(1).$$

Thus, global weighted summability already follows from existing assumptions, whereas Assumption 5 prevents the weighted covariance mass from concentrating on the particular group, teacher, or year under consideration. A stronger sufficient condition that can replace Assumption 5 is $\max_{b \in \mathcal{I}} \omega_b = O_p(1)$, which follows, for example, from uniformly bounded factors. Assumption 5 requires only the slice averages for each fixed index.

Proposition 2 (Fixed-index expansions with mode-specific remainders). *Suppose that Assumptions 1, 2, 3, 4, and 5 hold. Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be the global least-squares estimator defined in Assumption 1, and let H_u, H_v, H_w be the alignment matrices defined as*

$$\begin{aligned} H'_u &= (V * W)'(\hat{V} * \hat{W}) \left[(\hat{V} * \hat{W})'(\hat{V} * \hat{W}) \right]^{-1}, \\ H'_v &= (DU * W)'(D\hat{U} * \hat{W}) \left[(D\hat{U} * \hat{W})'(D\hat{U} * \hat{W}) \right]^{-1}, \\ H'_w &= (DU * V)'(D\hat{U} * \hat{V}) \left[(D\hat{U} * \hat{V})'(D\hat{U} * \hat{V}) \right]^{-1} \end{aligned}$$

and let $\Sigma_{VW} = (\frac{1}{M}V'V) \odot (\frac{1}{T}W'W)$, $\Sigma_{UW} = (\frac{1}{N}U'D'DU) \odot (\frac{1}{T}W'W)$, and $\Sigma_{UV} = (\frac{1}{N}U'D'DU) \odot (\frac{1}{M}V'V)$.

For each fixed admissible group g , teacher j , and period t , define

$$\begin{aligned} S_{u,g} &= \frac{1}{\sqrt{N_g M T}} \sum_{\substack{1 \leq i \leq N \\ g(i)=g}} \sum_{j'=1}^M \sum_{t'=1}^T \varepsilon_{ij't'}(v_{j'} \odot w_{t'}), \\ S_{v,j} &= \frac{1}{\sqrt{N T}} \sum_{i=1}^N \sum_{t'=1}^T \varepsilon_{ijt'}(u_{g(i)} \odot w_{t'}), \\ S_{w,t} &= \frac{1}{\sqrt{N M}} \sum_{i=1}^N \sum_{j'=1}^M \varepsilon_{ij't}(u_{g(i)} \odot v_{j'}), \end{aligned}$$

and the $\ell \times p$ cross-moment matrices

$$\begin{aligned} Q_{u,g} &= \frac{1}{N_g MT} \sum_{\substack{1 \leq i \leq N \\ g(i)=g}} \sum_{j'=1}^M \sum_{t'=1}^T (v_{j'} \odot w_{t'}) x'_{ij't'}, \\ Q_{v,j} &= \frac{1}{NT} \sum_{i=1}^N \sum_{t'=1}^T (u_{g(i)} \odot w_{t'}) x'_{ij't'}, \\ Q_{w,t} &= \frac{1}{NM} \sum_{i=1}^N \sum_{j'=1}^M (u_{g(i)} \odot v_{j'}) x'_{ij't}. \end{aligned}$$

Let $\theta_{NMT} = \sqrt{NMT}(\hat{\delta} - \delta)$. Then

$$\begin{aligned} \sqrt{NM} H_w^{-1}(\hat{w}_t - H_w w_t) &= \Sigma_{UV}^{-1} \left[S_{w,t} - \frac{1}{\sqrt{T}} Q_{w,t} \theta_{NMT} \right] + R_{w,t,NMT}, \\ \sqrt{NT} H_v^{-1}(\hat{v}_j - H_v v_j) &= \Sigma_{UW}^{-1} \left[S_{v,j} - \frac{1}{\sqrt{M}} Q_{v,j} \theta_{NMT} \right] + R_{v,j,NMT}, \\ \sqrt{N_g MT} H_u^{-1}(\hat{u}_g - H_u u_g) &= \Sigma_{VW}^{-1} \left[S_{u,g} - \sqrt{\frac{N_g}{N}} Q_{u,g} \theta_{NMT} \right] + R_{u,g,NMT}. \end{aligned}$$

The remainder vectors satisfy

$$\begin{aligned} \|R_{w,t,NMT}\| &= O_p \left(\rho + \frac{\sqrt{N}}{T\sqrt{M}} + \frac{\sqrt{M}}{T\sqrt{N}} \right), \\ \|R_{v,j,NMT}\| &= O_p \left(\rho + \frac{\sqrt{N}}{M\sqrt{T}} + \frac{\sqrt{T}}{M\sqrt{N}} \right), \\ \|R_{u,g,NMT}\| &= O_p \left(\rho + \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}} + \sqrt{N_g MT} \rho^4 \right), \end{aligned}$$

where $\rho = \sqrt{\frac{G+M+T}{NMT}}$. If $N_g G/N = O_p(1)$, we have

$$\|R_{u,g,NMT}\| = O_p \left(\rho + \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}} \right).$$

The group-size condition holds, in particular, when G remains fixed or when $N_g = O_p(N/G)$ for the group under consideration.

Remark: When slope estimation affects the leading expansion The normalizations in Proposition 2 reflect the numbers of observations entering the respective normal equations: $N_g MT$ for a group loading, NT for a teacher loading, and NM for a year factor.

The Assumption 2 imply $Q_{u,g} = O_p(1)$, $Q_{v,j} = O_p(1)$, and $Q_{w,t} = O_p(1)$ by the Cauchy–Schwarz inequality. If, in addition, $\theta_{NMT} = O_p(1)$, the slope-estimation terms satisfy

$$\frac{1}{\sqrt{T}} Q_{w,t} \theta_{NMT} = O_p(T^{-1/2}), \quad \frac{1}{\sqrt{M}} Q_{v,j} \theta_{NMT} = O_p(M^{-1/2}), \quad \sqrt{\frac{N_g}{N}} Q_{u,g} \theta_{NMT} = O_p \left(\sqrt{\frac{N_g}{N}} \right).$$

Consequently, when the respective remainders are $o_p(1)$, the teacher and year expansions have the same first-order score representations as if the slope were known whenever $T \rightarrow \infty$ and $M \rightarrow \infty$, respectively. The group expansion has this oracle property if $N_g/N \rightarrow_p 0$. By contrast, if G is fixed and $N_g/N \rightarrow_p \pi_g > 0$, the slope correction generally remains in the leading group expansion. The limiting distribution must then account jointly for $S_{u,g}$ and θ_{NMT} , including their covariance.

Remark: Mode-specific dimension conditions In addition to $\rho \rightarrow 0$, sufficient conditions for the displayed remainder bounds to be $o_p(1)$ are

$$\begin{aligned} \frac{N_g M}{N^2 T} + \frac{N_g T}{N^2 M} &\rightarrow_p 0 && \text{for the group expansion under the upper group-size condition.} \\ \frac{N}{M^2 T} + \frac{T}{M^2 N} &\rightarrow 0 && \text{for the teacher expansion,} \\ \frac{N}{M T^2} + \frac{M}{N T^2} &\rightarrow 0 && \text{for the year expansion.} \end{aligned}$$

These restrictions are mode-specific and need not follow from $\rho \rightarrow 0$. For example, if N is fixed and $M, T \rightarrow \infty$ grow at a comparable rate, the year and teacher remainder bounds vanish, whereas the group remainder bound need not vanish.

When all N, M, T has the same order as $L \rightarrow \infty$, the restriction $G \leq N$ gives $\rho = O(L^{-1})$, and

$$\|R_{w,t,NMT}\| + \|R_{v,j,NMT}\| = O_p(L^{-1}), \quad \|R_{u,g,NMT}\| = O_p\left(L^{-1} + \frac{\sqrt{N_g}}{L}\right).$$

In particular, when N_g grows at the same rate as N/G , the group remainder is $O_p(L^{-1} + (LG)^{-1/2})$ and is negligible for both fixed and growing G .

Assumption 6 (Joint fixed-index and slope-score central limit theorem). *Let $S_{u,g}, S_{v,j}, S_{w,t}$, the cross-moment matrices $Q_{u,g}, Q_{v,j}, Q_{w,t}$, and the normalized loading Gram matrices $\Sigma_{VW}, \Sigma_{UW}, \Sigma_{UV}$ be as defined in Proposition 2. Let $\mathcal{Z}_{NMT}$, Γ_{NMT} , and $\mathcal{B}_m$, $m = 1, 2, 3$, be as defined in Proposition 1. All scores, cross-moment matrices, Gram matrices, and bias terms are evaluated at the same normalized representation of the true factors. Along the asymptotic sequence in Assumption 1, the following conditions hold.*

The slope Gram matrix and bias terms satisfy

$$\Gamma_{NMT} \xrightarrow{p} \Gamma_\delta > 0, \quad \mathcal{B}_m \xrightarrow{p} b_m, \quad m = 1, 2, 3,$$

where $\Gamma_\delta \in \mathbb{R}^{p \times p}$ and $b_m \in \mathbb{R}^p$. We define $b_\delta = b_1 + b_2 + b_3$. The normalized loading Gram matrices satisfy

$$\begin{aligned} \Sigma_{VW} &\xrightarrow{p} \bar{\Sigma}_{VW} > 0, \\ \Sigma_{UW} &\xrightarrow{p} \bar{\Sigma}_{UW} > 0, \\ \Sigma_{UV} &\xrightarrow{p} \bar{\Sigma}_{UV} > 0, \end{aligned}$$

where $\bar{\Sigma}_{VW}, \bar{\Sigma}_{UW}, \bar{\Sigma}_{UV} \in \mathbb{R}^{\ell \times \ell}$. For every fixed admissible group g , teacher j , and period t , the group share satisfies

$$\frac{N_g}{N} \xrightarrow{p} \pi_g, \quad \pi_g \in [0, 1],$$

and the cross-moment matrices satisfy

$$Q_{u,g} \xrightarrow{p} \bar{Q}_{u,g}, \quad Q_{v,j} = O_p(1), \quad Q_{w,t} = O_p(1),$$

where $Q_{u,g}, Q_{v,j}, Q_{w,t} \in \mathbb{R}^{\ell \times p}$ and $\bar{Q}_{u,g} \in \mathbb{R}^{\ell \times p}$. These matrices need not be zero. We define the limiting group slope coefficient $C_{u,g} = \sqrt{\pi_g} \bar{Q}_{u,g}$ so that

$$\sqrt{\frac{N_g}{N}} Q_{u,g} \xrightarrow{p} C_{u,g}.$$

For every fixed admissible triple (g, j, t) , the loading scores and the oracle slope score satisfy the joint central limit theorem

$$\begin{pmatrix} S_{u,g} \\ S_{v,j} \\ S_{w,t} \\ Z_{NMT} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma_{gjt}),$$

where $\Sigma_{gjt} \in \mathbb{R}^{(3\ell+p) \times (3\ell+p)}$ is symmetric and positive semidefinite, with block representation

$$\Sigma_{gjt} = \begin{pmatrix} \Gamma_g^{vw} & \Xi_{uv,gj} & \Xi_{uw,gt} & \Xi_{gZ} \\ \Xi'_{uv,gj} & \Gamma_j^{uw} & \Xi_{vw,jt} & \Xi_{jZ} \\ \Xi'_{uw,gt} & \Xi'_{vw,jt} & \Gamma_t^{uv} & \Xi_{tZ} \\ \Xi'_{gZ} & \Xi'_{jZ} & \Xi'_{tZ} & \Omega_Z \end{pmatrix}.$$

The diagonal covariance blocks satisfy

$$\Gamma_g^{vw}, \Gamma_j^{uw}, \Gamma_t^{uv} \in \mathbb{R}^{\ell \times \ell}, \quad \Gamma_g^{vw} > 0, \quad \Gamma_j^{uw} > 0, \quad \Gamma_t^{uv} > 0,$$

and

$$\Omega_Z \in \mathbb{R}^{p \times p}, \quad \Omega_Z > 0.$$

Theorem 1 (Joint asymptotic normality of the slope and aligned factor errors). *Suppose that Assumptions 1, 2, 3, 4, 5, and 6 hold. Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be the global least-squares estimator defined in Assumption 1. Suppose that*

$$\frac{(G + M + T)^{3/2}}{NMT} \rightarrow 0. \quad (11)$$

Let $b_\delta = b_1 + b_2 + b_3$, and use the limiting matrices and vectors defined in Assumption 6.

The estimated factors are expressed in the true-factor column labeling and sign orientation. Specifically, among the equivalent least-squares triples

$$(\hat{U}\Pi_\ell S_v S_w, \hat{V}\Pi_\ell S_v, \hat{W}\Pi_\ell S_w),$$

where Π_ℓ ranges over $\ell \times \ell$ permutation matrices and S_v, S_w range over diagonal matrices with entries in $\{-1, 1\}$, choose a representative minimizing

$$\frac{1}{M} \|\hat{V}\Pi_\ell S_v - V\|_F^2 + \frac{1}{T} \|\hat{W}\Pi_\ell S_w - W\|_F^2.$$

Retain the hat notation for the selected representative, and define H_u, H_v, H_w from it as in Proposition 2. The followings hold.

1. (Identity limits of the alignment matrices.)

The matrices H_u, H_v, H_w are nonsingular with probability approaching one and satisfy

$$H_u \xrightarrow{p} I_\ell, \quad H_v \xrightarrow{p} I_\ell, \quad H_w \xrightarrow{p} I_\ell.$$

2. (Slope central limit theorem.)

The slope estimator satisfies

$$\sqrt{NMT}(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(-\Gamma_\delta^{-1}b_\delta, \Gamma_\delta^{-1}\Omega_Z\Gamma_\delta^{-1}).$$

3. (Joint fixed-index central limit theorem.)

Fix an admissible triple (g, j, t) and suppose additionally that

$$\frac{N+M}{T\sqrt{NM}} \rightarrow 0, \quad \frac{N+T}{M\sqrt{NT}} \rightarrow 0, \quad \frac{\sqrt{N_g}(M+T)}{N\sqrt{MT}} \xrightarrow{p} 0. \quad (12)$$

Let

$$C_{u,g} = \sqrt{\pi_g} \bar{Q}_{u,g}.$$

Then the aligned factor errors satisfy

$$\begin{pmatrix} \sqrt{N_g MT}(\hat{u}_g - H_u u_g) \\ \sqrt{NT}(\hat{v}_j - H_v v_j) \\ \sqrt{NM}(\hat{w}_t - H_w w_t) \\ \sqrt{NMT}(\hat{\delta} - \delta) \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mu_{gjt}, \mathcal{A}_{gjt} \Sigma_{gjt} \mathcal{A}'_{gjt}), \quad (13)$$

where

$$\mathcal{A}_{gjt} = \begin{pmatrix} \bar{\Sigma}_{VW}^{-1} & 0 & 0 & -\bar{\Sigma}_{VW}^{-1} C_{u,g} \Gamma_\delta^{-1} \\ 0 & \bar{\Sigma}_{UW}^{-1} & 0 & 0 \\ 0 & 0 & \bar{\Sigma}_{UV}^{-1} & 0 \\ 0 & 0 & 0 & \Gamma_\delta^{-1} \end{pmatrix}, \quad \mu_{gjt} = \begin{pmatrix} \bar{\Sigma}_{VW}^{-1} C_{u,g} \Gamma_\delta^{-1} b_\delta \\ 0_\ell \\ 0_\ell \\ -\Gamma_\delta^{-1} b_\delta \end{pmatrix}.$$

For completeness, define

$$\begin{aligned} V_{u,g} &= \bar{\Sigma}_{VW}^{-1} \left[\Gamma_g^{vw} + C_{u,g} \Gamma_\delta^{-1} \Omega_Z \Gamma_\delta^{-1} C'_{u,g} - \Xi_{gZ} \Gamma_\delta^{-1} C'_{u,g} - C_{u,g} \Gamma_\delta^{-1} \Xi'_{gZ} \right] \bar{\Sigma}_{VW}^{-1}, \\ V_{v,j} &= \bar{\Sigma}_{UW}^{-1} \Gamma_j^{uw} \bar{\Sigma}_{UW}^{-1}, \\ V_{w,t} &= \bar{\Sigma}_{UV}^{-1} \Gamma_t^{uv} \bar{\Sigma}_{UV}^{-1}. \end{aligned}$$

The marginal fixed-index limits are

$$\begin{aligned} \sqrt{N_g MT}(\hat{u}_g - H_u u_g) &\xrightarrow{d} \mathcal{N}(\bar{\Sigma}_{VW}^{-1} C_{u,g} \Gamma_\delta^{-1} b_\delta, V_{u,g}), \\ \sqrt{NT}(\hat{v}_j - H_v v_j) &\xrightarrow{d} \mathcal{N}(0, V_{v,j}), \\ \sqrt{NM}(\hat{w}_t - H_w w_t) &\xrightarrow{d} \mathcal{N}(0, V_{w,t}). \end{aligned} \quad (14)$$

Remark: Bias limits for groups with a positive limiting share If $\pi_g > 0$, the group loading rate condition can be written as

$$\sqrt{\frac{N_g}{N}} \frac{M+T}{\sqrt{NMT}} \rightarrow_p 0.$$

It therefore implies $M/\sqrt{NMT} \rightarrow 0$ and $T/\sqrt{NMT} \rightarrow 0$. The bias bounds in Proposition 1 yield $\mathcal{B}_2, \mathcal{B}_3 = o_p(1)$, so $b_2 = b_3 = 0$. If G is also fixed, $\mathcal{B}_1 = O_p(G/\sqrt{NMT}) = o_p(1)$, and hence $b_\delta = 0$. All limiting means in part (iii) then vanish. Nevertheless, $C_{u,g}$ may remain nonzero, so slope estimation can affect the group variance even when it produces no asymptotic mean shift.

Remark: Alignment matrices and the centering of factor errors The identity limits of H_u, H_v, H_w allow their inverse premultipliers to be removed from the normalized expansions because the corresponding normalized errors are stochastically bounded. For example,

$$\sqrt{N_g MT}(\hat{u}_g - H_u u_g) = \sqrt{N_g MT} H_u^{-1}(\hat{u}_g - H_u u_g) + o_p(1).$$

However, we may not replace $H_u u_g$ by u_g . Such a replacement would require the additional condition

$$\sqrt{N_g MT}(H_u - I_\ell)u_g = o_p(1),$$

with analogous conditions for the teacher and year factors.

5.1 Bias Corrected Estimator

In this section, we show that the three bias terms in (10) can be corrected, as in the framework of Hahn and Kuersteiner (2002) and Bai (2009).

Suppose the errors are conditionally uncorrelated, i.e., $\mathbb{E}[\varepsilon_a \varepsilon_b \mid \mathcal{F}] = 0$ for $a \neq b$. We define

$$\hat{\varepsilon}_{ijt} = y_{ijt} - x'_{ijt} \hat{\delta} - \sum_{k=1}^{\ell} \hat{u}_{g(i),k} \hat{v}_{jk} \hat{w}_{tk}, \quad \hat{\sigma}_{gjt}^2 = \frac{1}{N_g} \sum_{i:g(i)=g} \hat{\varepsilon}_{ijt}^2.$$

The grouped conditional variance is $\sigma_{gjt}^{G,2} = N_g^{-1} \sum_{i:g(i)=g} \mathbb{E}[\varepsilon_{ijt}^2 \mid \mathcal{F}]$, and each Ω_m^G is diagonal. We may use the corresponding residual-based estimators given by

$$\hat{\Omega}_1^G = \text{diag} \left(\left(\sum_{g=1}^G \hat{\sigma}_{gjt}^2 \right)_{(j,t)} \right), \quad \hat{\Omega}_2^G = \text{diag} \left(\left(\sum_{j=1}^M \hat{\sigma}_{gjt}^2 \right)_{(g,t)} \right), \quad \hat{\Omega}_3^G = \text{diag} \left(\left(\sum_{t=1}^T \hat{\sigma}_{gjt}^2 \right)_{(g,j)} \right), \quad (15)$$

with the pair indices ordered according to the corresponding unfolding convention.

Proposition 3. *Suppose that Assumptions 1, 2, 3, and 4 hold. Suppose also that $\mathbb{E}[\varepsilon_a \varepsilon_b \mid \mathcal{F}] = 0$ for all distinct $a, b \in \mathcal{I}$. The true factor rows satisfy*

$$\frac{1}{N} \sum_{g=1}^G N_g \|u_g\|^4 + \frac{1}{M} \sum_{j=1}^M \|v_j\|^4 + \frac{1}{T} \sum_{t=1}^T \|w_t\|^4 = O_p(1)$$

and the group sizes satisfies

$$\sum_{g=1}^G \frac{1}{N_g} = O_p \left(\frac{G^2}{N} \right). \quad (16)$$

Let $\mathcal{T}$ be the tangent space defined in (6) evaluated at the true parameters (U, V, W) and let Γ_{NMT} and $\mathcal{Z}_{NMT}$ be as defined in Proposition 1. Set $\hat{\hat{U}} = (D'D)^{1/2}\hat{U}$ and

$$(\hat{F}_1, \hat{F}_2, \hat{F}_3) = (\hat{\hat{U}}, \hat{V}, \hat{W}), \quad (\hat{C}_1, \hat{C}_2, \hat{C}_3) = (\hat{V} * \hat{W}, \hat{\hat{U}} * \hat{W}, \hat{\hat{U}} * \hat{V}).$$

Let $\hat{\Omega}_m^G$, $m = 1, 2, 3$, be data-based estimators of the conditional covariance defined in (15), of dimensions $MT \times MT$, $GT \times GT$, and $GM \times GM$, respectively. Define

$$\hat{\mathcal{B}}_m = \frac{1}{\sqrt{NMT}} X_{(4)} P_{\hat{\mathcal{T}}} Q K_{1,m}^G \text{vec} \left(\hat{F}_m (\hat{F}_m' \hat{F}_m)^{-1} (\hat{C}_m' \hat{C}_m)^{-1} \hat{C}_m' \hat{\Omega}_m^G M_{\hat{C}_m} \right), \quad m = 1, 2, 3.$$

The bias-corrected estimator $\hat{\delta}_{bc}$ is defined as

$$\hat{\delta}_{bc} = \hat{\delta} + \frac{1}{\sqrt{NMT}} \hat{\Gamma}_{NMT}^{-1} \left(\hat{\mathcal{B}}_1 + \hat{\mathcal{B}}_2 + \hat{\mathcal{B}}_3 \right), \quad \hat{\Gamma}_{NMT} = \frac{1}{NMT} X_{(4)} M_{\hat{\mathcal{T}}} X_{(4)}'$$

Then, we have

$$\Gamma_{NMT} \sqrt{NMT} (\hat{\delta}_{bc} - \delta) = \mathcal{Z}_{NMT} + O_p \left(\sqrt{\frac{G+M+T}{NMT}} + \frac{(G+M+T)^{3/2}}{NMT} \right). \quad (17)$$

Remark: Centered asymptotic normality. Suppose additionally that

$$\frac{(G+M+T)^{3/2}}{NMT} \rightarrow 0, \quad \Gamma_{NMT} \xrightarrow{p} \Gamma_\delta > 0, \quad \mathcal{Z}_{NMT} \xrightarrow{d} \mathcal{N}(0, \Omega_Z).$$

The bias-corrected expansion and Slutsky's theorem imply

$$\sqrt{n}(\hat{\delta}_{bc} - \delta) \xrightarrow{d} \mathcal{N}(0, \Gamma_\delta^{-1} \Omega_Z \Gamma_\delta^{-1}).$$

The limiting covariance is determined by the oracle slope score. This conclusion does not require the individual bias terms $\mathcal{B}_m$ to converge to finite limits: the correction estimates and removes their sum at the required rate.

Remark: Aggregate control of group sizes. The reciprocal-size condition can be written as

$$\frac{1}{G} \sum_{g=1}^G \frac{N/G}{N_g} = \frac{N}{G^2} \sum_{g=1}^G \frac{1}{N_g} = O_p(1).$$

Cauchy-Schwarz implies that this quantity is at least one, with equality for equal group sizes. The condition therefore controls the average reciprocal group size relative to the equal-size benchmark. A uniform lower bound $N_g \geq cN/G$ is sufficient, but the aggregate condition allows some substantially smaller groups. For example, if $N = L^2$, $G = L$, $N_1 = 1$, and $N_g = L + 1$ for $g \geq 2$, then

$$\frac{N}{G^2} \sum_g \frac{1}{N_g} = 1 + \frac{L-1}{L+1} \rightarrow 2,$$

although $N_1/(N/G) = 1/L \rightarrow 0$.

6 Selecting the Number of Interactive Components

In this section, we let ℓ_0 denote the true number of interactive components and $\ell \in \{0, \dots, \ell_{\max}\}$ denote a candidate number, where $\ell_{\max}$ is fixed and $1 \leq \ell_0 \leq \ell_{\max}$. Assumptions 1 and 2 concerning the true factor matrices are understood with $\ell = \ell_0$. For each candidate rank, we jointly minimize the residual sum of squares over the slope coefficient and the grouped interactive component. Following the penalized-fit approach of Bai and Ng (2002), we select the rank by adding a complexity penalty to the resulting criterion.

For $\ell \geq 1$, let Θ_ℓ impose the normalization and upper bound in Assumption 1, with ℓ columns and the same constant c_U for every candidate rank:

$$\Theta_\ell = \left\{ (\tilde{U}, \tilde{V}, \tilde{W}) \in \mathbb{R}^{G \times \ell} \times \mathbb{R}^{M \times \ell} \times \mathbb{R}^{T \times \ell} : \right. \\ \left. \text{diag} \left(\frac{1}{M} \tilde{V}' \tilde{V} \right) = \text{diag} \left(\frac{1}{T} \tilde{W}' \tilde{W} \right) = \iota_\ell, \lambda_{\max} \left(\frac{1}{N} \tilde{U}' D' D \tilde{U} \right) \leq c_U \right\}.$$

Define the admissible grouped common-component classes by

$$\mathcal{C}_{\ell, c_U}^D = \left\{ \sum_{k=1}^{\ell} (D \tilde{u}_k \circ \tilde{v}_k \circ \tilde{w}_k) : (\tilde{U}, \tilde{V}, \tilde{W}) \in \Theta_\ell \right\}, \quad \ell \geq 1, \\ \mathcal{C}_{0, c_U}^D = \{0\}.$$

Every tensor in $\mathcal{C}_{\ell, c_U}^D$ has CP rank at most ℓ with the parameter bound used in estimation. Appending a zero column to $\tilde{U}$ and columns with norms $\sqrt{M}$ and $\sqrt{T}$ to $\tilde{V}$ and $\tilde{W}$, respectively, preserves both the fitted tensor and the normalization restrictions. Therefore,

$$\mathcal{C}_{0, c_U}^D \subseteq \mathcal{C}_{1, c_U}^D \subseteq \dots \subseteq \mathcal{C}_{\ell_{\max}, c_U}^D.$$

For each candidate ℓ , define

$$\mathcal{V}(\ell) = \min_{\delta \in \mathbb{R}^p, \Pi \in \mathcal{C}_{\ell, c_U}^D} \frac{1}{NMT} \|Y - X \times_4 \delta' - \Pi\|_F^2.$$

Concentrating the criterion over δ gives

$$\mathcal{V}(\ell) = \min_{\Pi \in \mathcal{C}_{\ell, c_U}^D} \frac{1}{NMT} \left\| M_{X'_{(4)}} \text{vec}(Y - \Pi) \right\|^2.$$

Let $g(G, N, M, T) > 0$ be a deterministic penalty. Define

$$PC(\ell) = \mathcal{V}(\ell) + \ell g(G, N, M, T), \quad \hat{\ell} = \min_{0 \leq \ell \leq \ell_{\max}} \text{argmin} PC(\ell), \quad (18)$$

where the outer minimum specifies the tie-breaking rule.

Assumption 7 (Uniform identification over candidate ranks). *There exists a constant $c_{\text{rank}} > 0$ such that, with probability approaching one,*

$$\inf_{\substack{\tilde{U} \in \mathbb{R}^{G \times \ell_{\max}} \\ \tilde{V} \in \mathbb{R}^{M \times \ell_{\max}}}} \lambda_{\min} \left[\frac{1}{NMT} X_{(4)} (M_{D\tilde{U} * \tilde{V}} \otimes M_W) X'_{(4)} \right] \geq c_{\text{rank}}.$$

Here W is the true $T \times \ell_0$ factor matrix.

Proposition 4 (Consistency of the information criterion). *Suppose that Assumptions 1, 2, 3, and 7 hold, with the true-factor restrictions evaluated at rank ℓ_0 . Let $\ell_{\max}$ be fixed, with $1 \leq \ell_0 \leq \ell_{\max}$, and define $\hat{\ell}$ by (18). If the positive deterministic penalty satisfies*

$$g(G, N, M, T) \rightarrow 0, \quad \frac{NMT}{G + M + T} g(G, N, M, T) \rightarrow \infty, \quad (19)$$

then, along the asymptotic sequence in Assumption 1, we have

$$\mathbb{P}\{\hat{\ell} = \ell_0\} \rightarrow 1.$$

Remark: Penalty choice. The conditions in (19) require a penalty that vanishes but dominates the order $(G + M + T)/(NMT)$ of the improvement from fitting additional components. Natural admissible choices are

$$\begin{aligned} g_1(G, N, M, T) &= \frac{G + M + T}{NMT} \log \left(\frac{NMT}{G + M + T} \right) \\ g_2(G, N, M, T) &= \frac{G + M + T}{NMT} \log (C_{GNT}^2) \\ g_3(G, N, M, T) &= \frac{\log(C_{GNT}^2)}{C_{GNT}^2}, \end{aligned}$$

where $C_{GNT}^2 = \min(NMT/G, NT, NM)$. This choice remains admissible throughout the asymptotic framework of Assumption 1, without requiring the three panel dimensions to grow at comparable rates.

Remark: Zero true rank. The result also extends to $\ell_0 = 0$. In that case, set $\Pi = 0$, regard U, V, W as matrices with no columns, and set $M_W = I_T$. The positive-rank dimension restriction and true-factor normalization and Gram conditions in Assumptions 1 and 2 are then omitted; their remaining dimensional, membership, regressor, and error conditions are retained. There are no underfitted candidates, and the uniform criterion bound and overfitting argument prove $\mathbb{P}\{\hat{\ell} = 0\} \rightarrow 1$.

Remark: Scale-invariant information criterion The additive criterion

$$PC(\ell) = \mathcal{V}(\ell) + \ell g(G, N, M, T)$$

depends on the scale of the outcome variable because $\mathcal{V}(\ell)$ has the units of a residual variance, whereas $g(G, N, M, T)$ is dimensionless. A scale-invariant alternative is

$$IC(\ell) = \log \mathcal{V}(\ell) + \ell g(G, N, M, T).$$

Suppose, in addition, that

$$\mathcal{V}(\ell_0) \xrightarrow{p} \sigma_\varepsilon^2 > 0.$$

For overfitted models, we may show

$$\mathcal{V}(\ell_0) - \mathcal{V}(\ell) = O_p(\rho^2),$$

where $\rho = \sqrt{\frac{G+M+T}{NMT}}$. Thus, a first-order expansion of the logarithm gives

$$\log \mathcal{V}(\ell) - \log \mathcal{V}(\ell_0) = O_p(\rho^2).$$

For underfitted models, we may show that the residual-variance gap remains bounded away from zero. Therefore, the same penalty conditions imply consistency of

$$\hat{\ell}_{IC} = \underset{0 \leq \ell \leq \ell_{\max}}{\operatorname{argmin}} IC(\ell).$$

Another scale-adjusted alternative is

$$PC^*(\ell) = \mathcal{V}(\ell) + \ell \hat{\sigma}^2 g(G, N, M, T),$$

where

$$\hat{\sigma}^2 = \mathcal{V}(\ell_{\max})$$

or any other consistent estimator of the limiting idiosyncratic variance.

7 Inclusion of Additive Fixed Effects

Consider the model with both additive and interactive fixed effects

$$y_{ijt} = x'_{ijt}\delta + \kappa + \mu_{g(i)} + \nu_j + \xi_t + \pi_{ijt} + \varepsilon_{ijt}, \quad \pi_{ijt} = \sum_{k=1}^{\ell} u_{g(i),k} v_{jk} w_{tk}.$$

To select a unique intercept-group-teacher-year representation of the additive component, we impose

$$\iota'_N D\mu = \sum_{g=1}^G N_g \mu_g = 0, \quad \iota'_M \nu = 0, \quad \iota'_T \xi = 0. \quad (20)$$

Define the $N \times M \times T$ dimensional additive tensor by

$$\mathcal{A}(\kappa, \mu, \nu, \xi) = \kappa(\iota_N \circ \iota_M \circ \iota_T) + (D\mu \circ \iota_M \circ \iota_T) + (\iota_N \circ \nu \circ \iota_T) + (\iota_N \circ \iota_M \circ \xi).$$

Let $y = \operatorname{vec}(Y)$, $e = \operatorname{vec}(E)$, $a(\kappa, \mu, \nu, \xi) = \operatorname{vec}(\mathcal{A}(\kappa, \mu, \nu, \xi))$, and $\pi(U, V, W) = \operatorname{vec}(\Pi(U, V, W))$. Suppressing parameter arguments at the true values, the tensor and vector representations are

$$Y = X \times_4 \delta' + \mathcal{A} + \Pi + E, \quad y = X'_{(4)} \delta + a + \pi + e.$$

Define a global least-squares estimator by

$$\begin{aligned} & (\hat{\delta}, \hat{\kappa}, \hat{\mu}, \hat{\nu}, \hat{\xi}, \hat{U}, \hat{V}, \hat{W}) \\ & \in \underset{\substack{\delta \in \mathbb{R}^p, \kappa \in \mathbb{R}, \mu \in \mathbb{R}^G, \nu \in \mathbb{R}^M, \xi \in \mathbb{R}^T \\ (U, V, W) \in \Theta_\ell \\ \iota'_N D\mu=0, \iota'_M \nu=0, \iota'_T \xi=0}}{\operatorname{argmin}} \frac{1}{NMT} \left\| y - X'_{(4)} \delta - a(\kappa, \mu, \nu, \xi) - \pi(U, V, W) \right\|^2. \end{aligned} \quad (21)$$

To derive the profiled objective function from (21), let

$$\mathcal{A}_D = (\iota_N \otimes \iota_M \otimes \iota_T \quad D \otimes \iota_M \otimes \iota_T \quad \iota_N \otimes I_M \otimes \iota_T \quad \iota_N \otimes \iota_M \otimes I_T)$$

Although $\mathcal{A}_D$ contains redundant columns, its column space is precisely the space of additive fitted vectors satisfying the representation $\mathcal{A}(\kappa, \mu, \nu, \xi)$. Conditional on (δ, U, V, W) , minimization over the normalized additive coefficients is therefore orthogonal projection

of $y - X'_{(4)}\delta - \pi(U, V, W)$ onto $\mathcal{A}_D$. Since $\iota_N \in \text{col}(D)$, the matrix $P_D - P_{\iota_N}$ is the orthogonal projector onto the group-constant vectors with observation-weighted mean zero. Consequently,

$$P_{\mathcal{A}_D} = P_{\iota_n} + (P_D - P_{\iota_N}) \otimes P_{\iota_M} \otimes P_{\iota_T} + P_{\iota_N} \otimes M_{\iota_M} \otimes P_{\iota_T} + P_{\iota_N} \otimes P_{\iota_M} \otimes M_{\iota_T},$$

where $\iota_n = \iota_N \otimes \iota_M \otimes \iota_T$. The four summands are mutually orthogonal projectors onto the intercept, normalized group-effect, normalized teacher-effect, and normalized period-effect spaces. In particular,

$$\text{rank}(P_{\mathcal{A}_D}) = 1 + (G - 1) + (M - 1) + (T - 1) = G + M + T - 2.$$

Let $M_{\mathcal{A}_D} = I_n - P_{\mathcal{A}_D}$. Profiling the additive coefficients out of (21) gives

$$(\hat{\delta}, \hat{U}, \hat{V}, \hat{W}) \in \underset{\substack{\delta \in \mathbb{R}^p \\ (U, V, W) \in \Theta_\ell}}{\text{argmin}} \frac{1}{NMT} \left\| M_{\mathcal{A}_D} \{y - X'_{(4)}\delta - \pi(U, V, W)\} \right\|^2. \quad (22)$$

Conversely, any global minimizer of (22), completed with its normalized additive least-squares coefficients, is a global minimizer of (21). Writing $\hat{\Pi} = \Pi(\hat{U}, \hat{V}, \hat{W})$ and $\hat{\mathcal{A}} = \mathcal{A}(\hat{\kappa}, \hat{\mu}, \hat{\nu}, \hat{\xi})$, the fitted additive component satisfies

$$\text{vec}(\hat{\mathcal{A}}) = P_{\mathcal{A}_D}(y - X'_{(4)}\hat{\delta} - \text{vec}(\hat{\Pi})).$$

The individual additive estimates have explicit expressions. For an arbitrary residual tensor $R = (r_{ijt})$, define

$$\begin{aligned} \bar{r}_{...} &= \frac{1}{NMT} \sum_{i=1}^N \sum_{j=1}^M \sum_{t=1}^T r_{ijt}, & \bar{r}_{g..} &= \frac{1}{N_g MT} \sum_{i:g(i)=g} \sum_{j=1}^M \sum_{t=1}^T r_{ijt}, \\ \bar{r}_{.j.} &= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T r_{ijt}, & \bar{r}_{..t} &= \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M r_{ijt}. \end{aligned}$$

The projection and residual-making formulas, indexed according to the original tensor entries, are

$$\begin{aligned} [P_{\mathcal{A}_D} \text{vec}(R)]_{ijt} &= \bar{r}_{g(i)..} + \bar{r}_{.j.} + \bar{r}_{..t} - 2\bar{r}_{...}, \\ [M_{\mathcal{A}_D} \text{vec}(R)]_{ijt} &= r_{ijt} - \bar{r}_{g(i)..} - \bar{r}_{.j.} - \bar{r}_{..t} + 2\bar{r}_{...}. \end{aligned}$$

Thus, for $\hat{r}_{ijt} = y_{ijt} - x'_{ijt}\hat{\delta} - \hat{\pi}_{ijt}$, the normalized additive coefficients associated with the selected profiled minimizer are

$$\begin{aligned} \hat{\kappa} &= \bar{\hat{r}}_{...}, & \hat{\mu}_g &= \bar{\hat{r}}_{g..} - \bar{\hat{r}}_{...}, \\ \hat{\nu}_j &= \bar{\hat{r}}_{.j.} - \bar{\hat{r}}_{...}, & \hat{\xi}_t &= \bar{\hat{r}}_{..t} - \bar{\hat{r}}_{...}. \end{aligned} \quad (23)$$

The means in (23) are computed after subtracting both the fitted regression component and the fitted interactive component.

Assumption 8 (Additive fixed effects). *The following two assumptions hold.*

(i) Strong centered factors. Define

$$\begin{aligned}\bar{u}' &= \frac{1}{N} \iota_N' D U, & \bar{v}' &= \frac{1}{M} \iota_M' V, & \bar{w}' &= \frac{1}{T} \iota_T' W, \\ U^c &= U - \iota_G \bar{u}', & V^c &= V - \iota_M \bar{v}', & W^c &= W - \iota_T \bar{w}'.\end{aligned}$$

Assume $\ell \leq \min\{G-1, M-1, T-1\}$. For a constant $c_c > 0$ independent of the panel dimensions, with probability approaching one,

$$\begin{aligned}c_c I_\ell &\leq \frac{1}{N} (U^c)' D' D U^c \leq c_1 I_\ell, \\ c_c I_\ell &\leq \frac{1}{M} (V^c)' V^c \leq c_1 I_\ell, \\ c_c I_\ell &\leq \frac{1}{T} (W^c)' W^c \leq c_1 I_\ell.\end{aligned}$$

Here c_1 is as in Assumption 2. These conditions concern the true centered factors; neither the true nor the estimated raw factor matrices are required to have zero column means.

(ii) Identification after accounting for additive and interactive effects. For arbitrary $\tilde{U} \in \mathbb{R}^{G \times \ell}$ and $\tilde{V} \in \mathbb{R}^{M \times \ell}$, define

$$\tilde{C}_A(\tilde{U}, \tilde{V}) = [\iota_N \otimes \iota_M, D\tilde{U} * \tilde{V}], \quad W_A = [\iota_T, W].$$

There exists a constant $c_{\delta, A} > 0$ such that, with probability approaching one,

$$\inf_{\substack{\tilde{U} \in \mathbb{R}^{G \times \ell} \\ \tilde{V} \in \mathbb{R}^{M \times \ell}}} \lambda_{\min} \left[\frac{1}{NMT} X_{(4)} \left(M_{\tilde{C}_A(\tilde{U}, \tilde{V})} \otimes M_{W_A} \right) X_{(4)}' \right] \geq c_{\delta, A}.$$

This condition extends Assumption 4 to the inclusion of additive fixed effects.

Lemma 2 (Consistency with group, teacher, and period additive effects). *Suppose Assumptions 1, 2, 3, and 8 hold. Both the true and candidate additive effects satisfy the observation-weighted normalizations in (20). Define the global least-squares estimator by (21). Then the slope and the interactive component satisfy*

$$\|\hat{\delta} - \delta\| + \frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{n}} = O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right). \quad (24)$$

The uncentered joint loading spaces satisfy

$$\begin{aligned}\frac{1}{\sqrt{MT}} \|M_{\hat{V} * \hat{W}}(V * W)\|_F &= O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right), \\ \frac{1}{\sqrt{NT}} \|M_{D\hat{U} * \hat{W}}(DU * W)\|_F &= O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right), \\ \frac{1}{\sqrt{NM}} \|M_{D\hat{U} * \hat{V}}(DU * V)\|_F &= O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right).\end{aligned} \quad (25)$$

Moreover, the additive-effect estimators satisfy

$$\begin{aligned}
|\hat{\kappa} - \kappa| &= O_p \left(\frac{1}{\sqrt{NMT}} + \sqrt{\frac{G+M+T}{NMT}} \right), \\
\frac{1}{\sqrt{N}} \|D(\hat{\mu} - \mu)\| &= \left[\frac{1}{N} \sum_{g=1}^G N_g (\hat{\mu}_g - \mu_g)^2 \right]^{1/2} = O_p \left(\sqrt{\frac{G}{NMT}} + \sqrt{\frac{G+M+T}{NMT}} \right), \\
\frac{1}{\sqrt{M}} \|\hat{\nu} - \nu\| &= O_p \left(\frac{1}{\sqrt{NT}} + \sqrt{\frac{G+M+T}{NMT}} \right), \\
\frac{1}{\sqrt{T}} \|\hat{\xi} - \xi\| &= O_p \left(\frac{1}{\sqrt{NM}} + \sqrt{\frac{G+M+T}{NMT}} \right).
\end{aligned} \tag{26}$$

Remark: The interactive component after profiling. Additive profiling removes the intercept and the group, teacher, and year additive effects. To describe what remains in the interactive effects, let u_k^c, v_k^c, w_k^c denote the columns of U^c, V^c, W^c , and let $\bar{u}_k, \bar{v}_k, \bar{w}_k$ denote the corresponding column means. Expanding each rank-one component gives

$$\Pi^{\mathcal{A}_D} = \sum_{k=1}^{\ell} \left\{ Du_k^c \circ v_k^c \circ w_k^c + \bar{w}_k (Du_k^c \circ v_k^c \circ \iota_T) + \bar{v}_k (Du_k^c \circ \iota_M \circ w_k^c) + \bar{u}_k (\iota_N \circ v_k^c \circ w_k^c) \right\}.$$

The last three terms are, respectively, group–teacher interactions constant across years, group–years interactions constant across teachers, and teacher–years interactions constant across groups. Each is orthogonal to the additive space because its two varying factors have zero means. Thus additive profiling retains these two-way interactions together with the fully centered three-way interaction; full three-way centering retains only the first term. In other words, we have

$$M_{\mathcal{A}_D} \geq M_{\iota_N} \otimes M_{\iota_M} \otimes M_{\iota_T},$$

and fitting an unrestricted rank- ℓ CP tensor to the outcome and regressors transformed by $M_{\iota_N} \otimes M_{\iota_M} \otimes M_{\iota_T}$ would generally define a different estimator.

The fully three-way centered sample alone does not identify the raw factor means or the additive coefficients, whereas the profiling estimator defined in Lemma 2 identifies both the raw factor means and the additive effects. Consistency of the fully three-way centered slope estimator requires identification from the regressor variation surviving $M_{\iota_N} \otimes M_{\iota_M} \otimes M_{\iota_T}$, which is not generally implied by Assumption 8(ii). The additional contrasts retained by profiling may strengthen identification and improve precision.

Remark: Average consistency and the stochastic expansions of the model with additive fixed effects as in Lemma 1 and Propositions 1 and 2 can be derived from Lemma 2 using the same arguments as in the corresponding proofs under suitable additional assumptions. With additive effects, the relevant tangent space is

$$\mathcal{T}_A = \text{col}(\mathcal{A}_D) + \mathcal{T},$$

so that the slope Gram matrix and oracle score become

$$\Gamma_{NMT}^A = \frac{1}{NMT} X_{(4)} M_{\mathcal{T}_A} X'_{(4)}, \quad \mathcal{Z}_{NMT}^A = \frac{1}{\sqrt{NMT}} X_{(4)} M_{\mathcal{T}_A} e.$$

The bias formulas also require corresponding modifications. The details are omitted.

8 Monte Carlo Simulations

In this section, we assess finite-sample performance of the proposed grouped three-way estimator.

8.1 Data-Generating Process

We simulate data from the model

$$y_{ijt} = x'_{ijt}\delta + \kappa + \mu_{g(i)} + \nu_j + \xi_t + \pi_{ijt} + \varepsilon_{ijt}, \quad \pi_{ijt} = \sum_{k=1}^{\ell} u_{g(i),k} v_{jk} w_{tk}.$$

We set $p = 2$ and $\ell = 2$. We assume that G divides N and construct equally sized groups:

$$N_g = \frac{N}{G}, \quad g(i) = g \quad \text{for } i = (g-1)N_g + 1, \dots, gN_g, \quad g = 1, \dots, G.$$

The additive fixed effects μ_g, ν_j, ξ_t are drawn from a standard normal distribution and are demeaned. The grand mean is $\kappa = 1$. We also draw the fixed effects (u_g, v_j, w_t) from a standard normal distribution and rescale the terms without changing the interactive component so that columns of $\frac{1}{\sqrt{M}}V$ and $\frac{1}{\sqrt{T}}W$ are unit norm. No sample mean-zero restriction is imposed on the columns of U , V , or W .

The regressors are generated according to

$$x_{ijt,q} = m_q + c_q \pi_{ijt} + \eta_{ijt,q}, \quad q = 1, 2,$$

where $\eta_{ijt,q} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. We set

$$\delta = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \quad m_1 = m_2 = c_1 = c_2 = 1.$$

Thus, the regressors are correlated with the interactive fixed effects but contain independent variation through their idiosyncratic components.

We consider two error specifications. In the first specification,

$$\varepsilon_{ijt} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

In the second specification, we independently draw $\tilde{\varepsilon}_{ijt} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for $t = 0, \dots, T$ and define

$$\varepsilon_{ijt} = \frac{1}{\sqrt{2}} (\tilde{\varepsilon}_{ijt} + \tilde{\varepsilon}_{ij,t-1}), \quad t = 1, \dots, T.$$

This moving-average specification satisfies

$$\text{Var}(\varepsilon_{ijt}) = 1, \quad \text{Corr}(\varepsilon_{ijt}, \varepsilon_{ij,t-1}) = \frac{1}{2}, \quad t = 2, \dots, T,$$

with zero autocorrelations at lags greater than one and independence across distinct pairs (i, j) . Both error specifications therefore have the same marginal variance. All primitive Gaussian draws used to generate the additive effects, interactive factors, regressor innovations, and error innovations are mutually independent.

8.2 Estimation

Given the estimation model,

$$y_{ijt} = x'_{ijt}\delta + \kappa + \alpha_{ij} + \xi_t + \sum_{k=1}^{\ell} \lambda_{ij,k} f_{tk} + e_{ijt},$$

we consider two restrictions

$$\text{Separability } (R_1) : \quad \alpha_{ij} = \mu_i + \nu_j, \quad \lambda_{ij,k} = a_{ik}b_{jk},$$

$$\text{Grouping } (R_2) : \quad \mu_i = \mu_{g(i)}, \quad a_{ik} = a_{g(i),k}.$$

The four estimation methods implemented here are summarized in a table.

Methods	Fixed effects structure	Interactive parameters	Estimation
1	Grouped three-way (R_1, R_2)	$(G + M + T)\ell$	Joint least squares
2	Ungrouped three-way (R_1)	$(N + M + T)\ell$	Joint least squares
3	Ungrouped three-way (R_1)	$(N + M + T)\ell$	Freeman (2026)
4	Flattened two-way	$(NM + T)\ell$	Joint least squares

The first method uses the correctly specified grouped three-way interactive fixed effect model. The second and third methods do not impose the grouping restriction. The second method estimates the model using least squares; the third method estimates the model using a transformation method as in Freeman (2026). The final method estimates a flattened two-way interactive fixed effect model that does not impose the separability restriction on fixed effects for i and j . All four methods are correctly specified; differences arise from fixed effects parameter parsimony and estimation method.¹² For each configuration and each error specification, we generate $R = 1,000$ Monte Carlo replications.

¹²Details on implementation methods are in Appendix 11.3.

Table 1: Monte Carlo performance of the slope estimators

Estimator	$\hat{\delta}_1$				$\hat{\delta}_2$			
	Bias	SD	RMSE	Ratio	Bias	SD	RMSE	Ratio
<i>i.i.d. errors; N = 24, M = 24, T = 8, G = 4</i>								
Grouped three-way	-0.0004	0.0146	0.0146	1.0000	-0.0002	0.0148	0.0148	1.0000
Ungrouped three-way	-0.0001	0.0148	0.0147	1.0114	0.0001	0.0149	0.0149	1.0056
Freeman linear	-0.0007	0.0218	0.0218	1.4952	0.0005	0.0214	0.0214	1.4425
Flattened two-way	0.0013	0.0195	0.0195	1.3382	0.0021	0.0192	0.0193	1.3026
<i>i.i.d. errors; N = 48, M = 24, T = 8, G = 6</i>								
Grouped three-way	0.0001	0.0105	0.0105	1.0000	0.0003	0.0106	0.0106	1.0000
Ungrouped three-way	0.0002	0.0105	0.0105	1.0041	0.0003	0.0106	0.0106	1.0037
Freeman linear	-0.0003	0.0148	0.0148	1.4131	0.0001	0.0145	0.0145	1.3653
Flattened two-way	0.0003	0.0132	0.0132	1.2606	0.0006	0.0132	0.0132	1.2416
<i>i.i.d. errors; N = 48, M = 48, T = 8, G = 8</i>								
Grouped three-way	0.0003	0.0074	0.0074	1.0000	-0.0008	0.0075	0.0075	1.0000
Ungrouped three-way	0.0003	0.0074	0.0074	1.0038	-0.0007	0.0075	0.0075	1.0030
Freeman linear	0.0001	0.0099	0.0099	1.3360	-0.0007	0.0099	0.0100	1.3258
Flattened two-way	0.0004	0.0093	0.0093	1.2598	-0.0006	0.0096	0.0096	1.2842
<i>i.i.d. errors; N = 48, M = 48, T = 12, G = 8</i>								
Grouped three-way	-0.0002	0.0060	0.0060	1.0000	-0.0001	0.0061	0.0061	1.0000
Ungrouped three-way	-0.0001	0.0060	0.0060	1.0045	-0.0001	0.0061	0.0061	1.0042
Freeman linear	-0.0002	0.0076	0.0076	1.2608	0.0000	0.0075	0.0075	1.2310
Flattened two-way	-0.0001	0.0072	0.0072	1.1900	0.0001	0.0071	0.0071	1.1647

Notes: Ratio is the estimator RMSE divided by the grouped-estimator RMSE

Table 1 shows that the grouped three-way estimator has the lowest RMSE for both slope coefficients in every reported design. Its bias is small relative to its standard deviation, and its RMSE declines approximately at the $1/\sqrt{NMT}$ rate: the values of $\sqrt{NMT}$ times RMSE lie between approximately 0.99 and 1.02. The ungrouped three-way estimator performs almost identically, with RMSE only 0.3%–1.1% larger. Thus, in these designs, the additional gain from grouping is modest.

The larger precision gains arise when comparing the grouped three-way estimator with the Freeman linear and flattened two-way estimators. Their RMSEs are approximately 23%–50% and 16%–34% larger than those of the grouped three-way estimator, respectively. The flattened two-way estimator has lower RMSE than the Freeman linear estimator in every i.i.d. design. Because biases are small relative to standard deviations, these differences primarily reflect differences in sampling variability. The results support the precision benefits of exploiting the separable three-way structure.

Table 2: Monte Carlo performance of the slope estimators

Estimator	$\hat{\delta}_1$				$\hat{\delta}_2$			
	Bias	SD	RMSE	Ratio	Bias	SD	RMSE	Ratio
<i>MA(1) errors; $N = 24, M = 24, T = 8, G = 4$</i>								
Grouped three-way	-0.0006	0.0144	0.0145	1.0000	-0.0003	0.0147	0.0147	1.0000
Ungrouped three-way	0.0010	0.0247	0.0247	1.7060	0.0014	0.0241	0.0241	1.6334
Freeman linear	0.0010	0.0184	0.0184	1.2745	0.0017	0.0186	0.0186	1.2647
Flattened two-way	0.0721	0.0914	0.1164	8.0500	0.0726	0.0915	0.1168	7.9213
<i>MA(1) errors; $N = 48, M = 24, T = 8, G = 6$</i>								
Grouped three-way	0.0003	0.0101	0.0101	1.0000	0.0000	0.0104	0.0104	1.0000
Ungrouped three-way	0.0004	0.0102	0.0102	1.0019	0.0002	0.0104	0.0104	1.0037
Freeman linear	0.0004	0.0130	0.0130	1.2798	0.0002	0.0121	0.0121	1.1661
Flattened two-way	0.0723	0.0971	0.1210	11.9342	0.0724	0.0967	0.1208	11.6442
<i>MA(1) errors; $N = 48, M = 48, T = 8, G = 8$</i>								
Grouped three-way	0.0003	0.0076	0.0076	1.0000	-0.0005	0.0074	0.0074	1.0000
Ungrouped three-way	0.0003	0.0076	0.0076	0.9994	-0.0004	0.0074	0.0074	1.0015
Freeman linear	0.0004	0.0088	0.0088	1.1541	-0.0002	0.0087	0.0087	1.1837
Flattened two-way	0.0634	0.0873	0.1079	14.2169	0.0627	0.0887	0.1086	14.7104
<i>MA(1) errors; $N = 48, M = 48, T = 12, G = 8$</i>								
Grouped three-way	-0.0001	0.0058	0.0058	1.0000	-0.0001	0.0061	0.0061	1.0000
Ungrouped three-way	0.0000	0.0058	0.0058	1.0003	-0.0001	0.0061	0.0061	1.0049
Freeman linear	0.0000	0.0069	0.0069	1.1798	0.0000	0.0069	0.0069	1.1254
Flattened two-way	0.0139	0.0150	0.0204	3.5015	0.0139	0.0152	0.0206	3.3736

Notes: Ratio is the estimator RMSE divided by the grouped-estimator RMSE

Table 2 shows that the grouped three-way estimator maintains small bias and declining RMSE under serial dependence. Grouped and ungrouped three-way estimators have essentially identical RMSEs in the three larger designs; the ungrouped three-way estimator has a slightly lower RMSE for the first coefficient when $N = M = 48$ and $T = 8$. In the smallest design, however, ungrouped three-way RMSE is approximately 63%–71% larger than grouped three-way RMSE. Freeman linear RMSE is approximately 13%–28% larger than grouped three-way RMSE across the reported MA(1) designs.

The flattened two-way estimator exhibits substantially larger bias and dispersion, with RMSE approximately 3.4–14.7 times that of the grouped three-way estimator. Squared bias accounts for approximately 33%–46% of its MSE, so both bias and sampling variability contribute to the deterioration. At $T = 8$, increasing the cross-sectional dimensions produces little reduction in its bias. Increasing T from 8 to 12, holding $N = M = 48$ and $G = 8$, sharply reduces both bias and standard deviation. This pattern is consistent with difficulties associated with estimating unrestricted interactive effects under serial dependence when the time dimension is short. The Freeman linear estimator outperforms flattened two-way in every MA(1) design, reversing their ranking under i.i.d. errors.

9 Empirical Application: Heterogeneous Teacher Value Added

We use student-level records from the National Center for Teacher Effectiveness Main Study (ICPSR 36095). The estimation sample contains 11,390 student-year observations on 9,818 distinct students and 317 teachers, covering fourth and fifth grades in 53 schools across four districts. The data span three academic years, 2010–2011 through 2012–2013, with

4,814, 4,130, and 2,446 observations, respectively. The resulting panel is unbalanced¹³ and contains 583 observed teacher-year pairs.

Students are classified into five prior-achievement groups using their prior-year state mathematics test scores. Table 3 reports teachers' coverage across the three study years, while Table 4 reports achievement-group coverage within teacher-year pairs. A group is represented if at least one student from that group is observed for the teacher in that year.

Table 3: Number of years observed per teacher

Years observed	Teachers	Percent
One year	126	39.75
Two years	116	36.59
Three years	75	23.66
Total	317	100.00

Notes: Only 75 teachers appear on all three years.

Table 4: Achievement groups represented per teacher-year

Groups represented	2010–2011	2011–2012	2012–2013	Total	Percent
One	1	2	0	3	0.51
Two	12	8	3	23	3.95
Three	32	22	11	65	11.15
Four	77	62	34	173	29.67
Five	130	120	69	319	54.72
Total	252	214	117	583	100.00

Notes: Percentages use all 583 observed teacher-year pairs as the denominator. Group coverage is calculated after pooling a teacher's classrooms within each year.

Teacher-year pairs represent 4.34 achievement groups on average, with a median of five. Overall, 2,531 of the 2,915 possible group-teacher-year cells contain observed students, corresponding to coverage of 86.83%. Occupied group-teacher-year cells contain 4.50 students on average, with a median of four and a range of 1–24. Among these cells, 345 (13.63%) contain one student and 730 (28.84%) contain at most two students.

The sparse and uneven coverage in Tables 3 and 4 motivates imposing structure on the teacher effects. Unrestricted group-teacher-year effects rely on the observations within each cell, which can yield noisy estimates in small cells and provide no direct estimates for empty cells. The three-way separability restriction pools information across groups, teachers, and years: group loadings are shared across teachers and years, year factors are shared across teachers and groups, and each teacher's loadings are estimated using all of that teacher's observed groups and years. Provided the observed support contains sufficient identifying variation, this structure allows observations from other cells to contribute to estimation.

We use the following model with additive teacher specific effects and grouped three-way

¹³Discussion of estimation with unbalanced panel is in Appendix 11.10.

interactive fixed effects,

$$y_{ijt} = x'_{ijt}\delta + \nu_j + \pi_{g(i),jt} + \varepsilon_{ijt}, \quad \pi_{g(i),jt} = \sum_{k=1}^{\ell} u_{g(i),k}v_{j,k}w_{t,k}.$$

Student achievement y_{ijt} is measured using standardized state mathematics scores. The student controls x_{ijt} include prior standardized test score and dummies for grade, gender, race, free or reduced-price lunch eligibility, special education status, and limited English proficiency status. The unobserved component $\nu_j + \pi_{gjt}$ is the teacher j 's effects in year t for students in prior-achievement group g .

We plot the estimated teacher components $(\hat{u}_g, \hat{v}_j, \hat{w}_t)$ estimated using $\ell = 2$ in the following Figure 1.¹⁴¹⁵

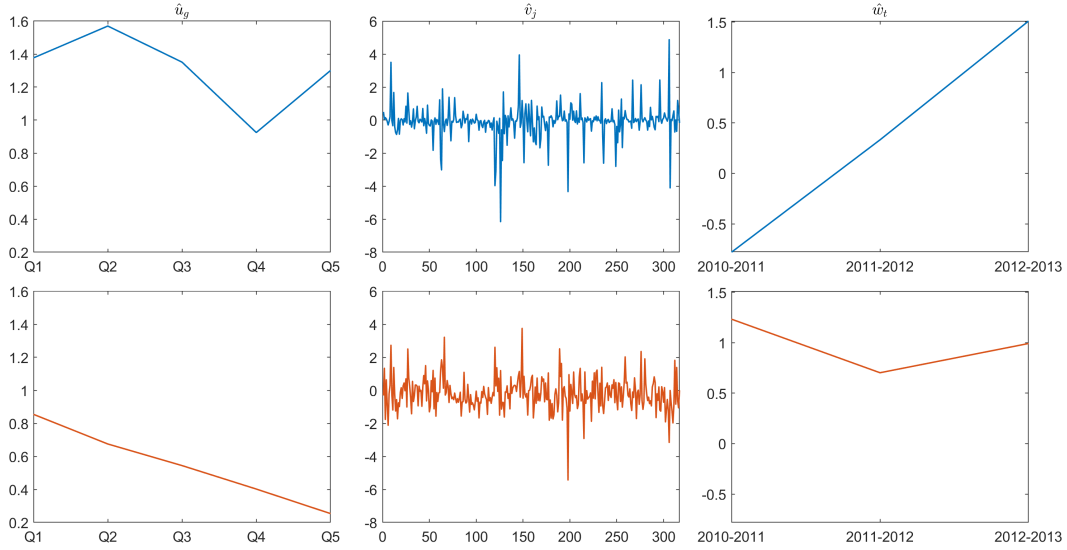


Figure 1: The estimated teacher components $(\hat{u}_g, \hat{v}_j, \hat{w}_t)$ with $\ell = 2$. The first row plots $k = 1$ and the second row plots $k = 2$.

Write $\hat{\pi}_{gjt}^{(k)} = \hat{u}_{gk}\hat{v}_{jk}\hat{w}_{tk}$. The first component combines positive group loadings with an increasing year factor. For a fixed group g , teacher j and two years $t < t'$, its change is

$$\hat{\pi}_{gjt'}^{(1)} - \hat{\pi}_{gjt}^{(1)} = \hat{u}_{g1}\hat{v}_{j1}(\hat{w}_{t'1} - \hat{w}_{t1}).$$

Since $(\hat{w}_{t'1} - \hat{w}_{t1}) > 0$ is shared across teachers, the first component $\hat{\pi}_{gjt}^{(1)}$ works to increase the teacher effect over time for teachers with $\hat{v}_{j1} > 0$. The magnitude of the increase varies across achievement groups through $\hat{u}_{g1}$.

¹⁴Assumption 8 implicitly requires $T \geq \ell + 2$ because full column rank in W^c implies $\text{rank}([\iota_T, W]) = \ell + 1$. This contradicts $\ell = 2$ because we only have $T = 3$. However, when only teacher-specific additive fixed effects are included, a uniform eigenvalue lower bound on $\frac{1}{NMT}X_{(4)}(M_{C_\nu}(\tilde{U}, \tilde{V}) \otimes M_W)X'_{(4)}$ where $C_\nu = (\iota_N \otimes I_M, D\tilde{U} * \tilde{V})$ is enough so that only $T \geq \ell + 1$ is required.

¹⁵We also report the results with $\ell = 1$ in Appendix 11.7. The results are qualitatively similar, with a smaller proportion of variation in the relative teacher value added explained by the matching effects.

For the second component, the declining group loadings and positive year factors imply that, for $\hat{v}_{j2} > 0$, this component contributes more to predicted achievement among initially lower-achieving students. For a lower-achievement group g_L and a higher-achievement group g_H , we have

$$\hat{\pi}_{g_L j t}^{(2)} - \hat{\pi}_{g_H j t}^{(2)} = (\hat{u}_{g_L, 2} - \hat{u}_{g_H, 2}) \hat{v}_{j2} \hat{w}_{t2}.$$

Note that $(\hat{u}_{g_L, 2} - \hat{u}_{g_H, 2}) > 0$ is shared across teachers. For teachers with $\hat{v}_{j2} > 0$, the second component $\hat{\pi}_{g j t}^{(2)}$ works to raise the teachers' effectiveness for the lower-achievement group more than that for the higher-achievement group, increasing this teacher's comparative advantage over the lower-achievement group. The differential reverses for teachers with negative $\hat{v}_{j2}$. The estimated second year factor remains positive and relatively stable throughout the sample. Heatmaps visualizing the estimated relative value added and matching effects are in Appendix 11.4.

9.1 Variance Decomposition of Teacher Effectiveness

Let $\tau_{g j t} = \nu_j + \sum_{k=1}^{\ell} u_{gk} v_{jk} w_{tk}$ denote the fitted teacher component. Let $\mathcal{J}_t$ be the set of teachers observed in period t , with $m_t = |\mathcal{J}_t|$. Using equal reference weights across the $G = 5$ prior-achievement groups, we define

$$\bar{\tau}_{j t} = \frac{1}{G} \sum_{g=1}^G \tau_{g j t}, \quad \bar{\tau}_{g t} = \frac{1}{m_t} \sum_{j \in \mathcal{J}_t} \tau_{g j t}, \quad \bar{\tau}_t = \frac{1}{m_t} \sum_{j \in \mathcal{J}_t} \bar{\tau}_{j t}, \quad \bar{\tau} = \frac{\sum_{t=1}^T m_t \bar{\tau}_t}{\sum_{t=1}^T m_t}.$$

For each observed teacher-year cell, we evaluate $\tau_{g j t}$ for every achievement group, including groups absent from that cell. The fitted teacher component admits the orthogonal decomposition

$$\begin{aligned} \tau_{g j t} - \bar{\tau} &= \underbrace{\bar{\tau}_t - \bar{\tau}}_{\text{year effect}} + \underbrace{\bar{\tau}_{g t} - \bar{\tau}_t}_{\text{group level effect}} + \underbrace{\bar{\tau}_{j t} - \bar{\tau}_t}_{\text{teacher level effect}} + \underbrace{\tau_{g j t} - \bar{\tau}_{j t} - \bar{\tau}_{g t} + \bar{\tau}_t}_{\text{matching effect}} \\ &= T_t + G_{g t} + L_{j t} + M_{g j t}. \end{aligned}$$

The year effect common to every teacher-group is captured in T_t . The group-level effect $G_{g t}$ captures effects that are common to teachers, i.e., how effective average teachers are on group g . The teacher-level effect $L_{j t}$ measures teacher j 's general effectiveness relative to the average active teacher in period t . The matching effect $M_{g j t}$ captures the remaining interaction and measures the teacher's comparative advantage for achievement group g .

Under equal weights across observed teacher-year cells and groups, the four components are pairwise orthogonal. Consequently, the sample variance can be decomposed as

$$V_{\tau} = V_T + V_G + V_L + V_M,$$

where

$$\begin{aligned}
V_\tau &= \frac{1}{G \sum_{t=1}^T m_t} \sum_{t=1}^T \sum_{j \in \mathcal{J}_t} \sum_{g=1}^G (\tau_{gjt} - \bar{\tau})^2, \\
V_T &= \frac{1}{\sum_{t=1}^T m_t} \sum_{t=1}^T m_t T_t^2, & V_G &= \frac{1}{\sum_{t=1}^T m_t} \sum_{t=1}^T m_t \left(\frac{1}{G} \sum_{g=1}^G G_{gt}^2 \right), \\
V_L &= \frac{1}{\sum_{t=1}^T m_t} \sum_{t=1}^T \sum_{j \in \mathcal{J}_t} L_{jt}^2, & V_M &= \frac{1}{G \sum_{t=1}^T m_t} \sum_{t=1}^T \sum_{j \in \mathcal{J}_t} \sum_{g=1}^G M_{gjt}^2.
\end{aligned}$$

Teacher value added relative to the average active teacher for group g in period t is defined as

$$\text{VA}_{gjt} = L_{jt} + M_{gjt},$$

The estimated variance components are

$$(\hat{V}_T, \hat{V}_G, \hat{V}_L, \hat{V}_M) = (0.000, 0.003, 0.097, 0.036).$$

We tabulate the variance share of the matching effect estimated using this method and estimated in Ahn et al. (2025).

	Grouped Three-Way		Ahn, Aucejo, and James (2025): Math	
	Variance	Share	Variance	Share
Relative teacher value added	0.133	100%	0.056	100%
Level	0.097	73%	0.049	88%
Matching	0.036	27%	0.007	12%

The matching component accounts for approximately 27% of relative teacher-value-added variation in the grouped three-way model, compared with 12% in the math estimates of Ahn et al. (2025).¹⁶

9.2 Explaining the Teacher Effects

To understand what explains the differences in the teacher-level and matching effects, we run two adaptive LASSO regressions with the estimated teacher-level effect, $\hat{L}_{jt}$, and the matching difference, $D_{jt} = \hat{M}_{1jt} - \hat{M}_{5jt}$, as dependent variables. We use 55 regressors that comprise 12 CLASS variables (DS0001); 13 MQI variables and the overall five-point lesson rating (DS0002); teaching experience and an indicator of elementary mathematics certification (DS0006); 25 survey items covering mathematics instruction, classroom climate, differentiated assignments, and test-related instructional changes (DS0007); and ACCURACY and KOSM (DS0008).¹⁷ Both regressions use the same sample of 483 teacher-years from 279 teachers.

¹⁶When the model $\ell = 1$ is used, 16% of the variation in the estimated relative teacher value added is explained by the matching effects.

¹⁷Detailed descriptions of the variables used are available in Appendix 11.5

The level regression selects six predictors (Table 5), whereas the matching-difference regression selects only MIS04 (Table 6).¹⁸ Tables below report penalized coefficients in original units, without post-selection OLS refitting. All remaining slopes are zero.

Table 5: Adaptive LASSO: Teacher-level effect $\hat{L}_{jt}$

Predictor	Coefficient
<i>CLASS measures (DS0001)</i>	
Behavior management (CLBM)	0.0328
<i>MQI measures (DS0002)</i>	
Imprecision in mathematical language or notation (LANGIMP)	-0.0655
Enacted task cognitive activation (ETCA)	0.0188
<i>Teacher survey items (DS0007)</i>	
Comparing different ways of solving procedural problems (MIR06)	0.0018
Instructional time lost to student misbehavior (CLC06)	-0.0302
Reducing demanding problems that challenge advanced students (TPC05)	-0.0089
Selected penalty, λ	0.0010
Teacher-year observations	483
Teachers	279

Table 5 reports coefficients in the regressors' original units. A one-point increase in the teacher-year average behavior-management rating (CLBM), measured on a 1–7 scale, is associated with a 0.0328σ increase in $\hat{L}_{jt}$. For the MQI items, measured on 1–3 scales, a one-point increase in enacted task cognitive activation¹⁹ (ETCA) is associated with a 0.0188σ increase in $\hat{L}_{jt}$, whereas a one-point increase in mathematical language or notation imprecision (LANGIMP) is associated with a 0.0655σ decrease. The remaining selected items use 1–5 scales. A one-category increase in the reported frequency of comparing different ways of solving procedural problems (MIR06) is associated with a 0.0018σ increase in $\hat{L}_{jt}$, while a one-category increase in instructional time lost to misbehavior (CLC06) is associated with a 0.0302σ decrease. A one-category increase in the reported extent of reducing demanding problems for advanced students in preparation for state-test (TPC05) is associated with a 0.0089σ decrease.

These patterns associate higher teacher-level effects with classroom management, mathematical precision, cognitively engaging instruction, and not reducing demanding problems in preparation for the state test.

¹⁸The adaptive penalty uses inverse-absolute OLS coefficient weights with $\gamma = 1$. We choose the penalty separately for each outcome by five-fold cross-validation, keeping all observations of a teacher in the same fold. Predictor standardization and adaptive weights are recomputed within each training fold. The selected penalties are $\lambda_L = 0.0010$ and $\lambda_D = 0.0034$.

¹⁹This measures the ability of the teacher to challenge students to actively process information

Table 6: Adaptive LASSO: Matching difference $D_{jt} = \widehat{M}_{1jt} - \widehat{M}_{5jt}$

Predictor	Coefficient
<i>Teacher survey item (DS0007)</i>	
Assigning problems requiring a few minutes of thought (MIS04)	0.0368
Selected penalty, λ	0.0034
Teacher-year observations	483
Teachers	279

Notes: The penalty is selected by five-fold cross-validation, with folds grouped by teacher. The model considers the same 55 predictors and uses the same sample as the level regression. Only the nonzero slope is shown. The coefficient is reported in original units, without post-selection OLS refitting.

In Table 6, MIS04 measures the frequency of assigning problems requiring a few minutes of thought, using a scale from 1 (rarely or never) to 5 (always). A one-category increase is associated with a 0.0368σ increase in $D_{jt} = \widehat{M}_{1jt} - \widehat{M}_{5jt}$, implying a more favorable matching for lower-achieving students. One possible interpretation is that assigning such problems provides opportunities for deliberate mathematical reasoning that are particularly well matched to lower-achieving students' learning needs.

9.3 Counterfactual Teacher Reallocation

We conduct two counterfactual exercises. First, we evaluate each teacher's relative value added when assigned to their individually best-matched classroom within their school or district. Second, we find the assignment that maximizes aggregate value added subject to every classroom receiving exactly one teacher and every teacher teaching exactly one classroom. Throughout both exercises, classroom rosters, student characteristics, and the estimated teacher-effect surface remain fixed.

For each year t , let $\mathcal{C}_t$ denote the included classrooms and $\mathcal{J}_t$ the included teachers. We restrict attention to teacher-classroom pairs for which the observed assignment is one-to-one, so $|\mathcal{C}_t| = |\mathcal{J}_t|$. Let $j_t^0(c)$ denote classroom c 's observed teacher and $c_t^0(j)$ denote teacher j 's observed classroom.

Let n_{gct} be the number of students from achievement group g in classroom c , and define

$$n_{ct} = \sum_{g=1}^G n_{gct}, \quad p_{gct} = \frac{n_{gct}}{n_{ct}}, \quad n_t = \sum_{c \in \mathcal{C}_t} n_{ct}.$$

The predicted matching contribution $M_t(j, c)$ and average relative value added $VA_t(j, c)$ of teacher j in classroom c are

$$M_t(j, c) = \sum_{g=1}^G p_{gct} \widehat{M}_{gjt},$$

$$VA_t(j, c) = \sum_{g=1}^G p_{gct} \widehat{VA}_{gjt} = \widehat{L}_{jt} + M_t(j, c).$$

Thus, classroom value added depends on the teacher's general effectiveness and their matching effects evaluated at the classroom's actual achievement-group composition. We retain the original estimates $\widehat{L}_{jt}$ and $\widehat{M}_{gjt}$ throughout the counterfactuals.

Let $s(c)$, $d(c)$, and $r(c)$ denote classroom c 's school, district, and grade. Define teacher j 's eligible classrooms under within-school and within-district reallocation as

$$\begin{aligned}\mathcal{C}_{jt}^S &= \{c \in \mathcal{C}_t : s(c) = s(c_t^0(j)), r(c) = r(c_t^0(j))\}, \\ \mathcal{C}_{jt}^D &= \{c \in \mathcal{C}_t : d(c) = d(c_t^0(j)), r(c) = r(c_t^0(j))\}.\end{aligned}$$

Teachers remain in their original grade and year. Within-school reallocation preserves their school, whereas within-district reallocation permits movement across schools in their original district. The observed classroom belongs to both menus, and $\mathcal{C}_{jt}^S \subseteq \mathcal{C}_{jt}^D$.

Exercise 1: Each teacher's individually best-matched classroom. For $b \in \{S, D\}$, define

$$c_{jt}^{\text{best},b} \in \operatorname{argmax}_{c \in \mathcal{C}_{jt}^b} M_t(j, c).$$

Because $\hat{L}_{jt}$ is fixed for teacher j , maximizing the matching contribution also maximizes that teacher's average relative value added. The resulting increase in value added from the observed classroom is

$$\Delta_{jt}^{\text{best},b} = M_t(j, c_{jt}^{\text{best},b}) - M_t(j, c_t^0(j))$$

Consequently,

$$0 \leq \Delta_{jt}^{\text{best},S} \leq \Delta_{jt}^{\text{best},D}.$$

The mean gain across teacher-year observations is

$$\bar{\Delta}^{\text{best},b} = \frac{\sum_{t=1}^T \sum_{j \in \mathcal{J}_t} \Delta_{jt}^{\text{best},b}}{\sum_{t=1}^T |\mathcal{J}_t|}.$$

This exercise evaluates each teacher separately. Several teachers may select the same classroom, while other classrooms may receive no selections. Its results therefore measure individual opportunities rather than gains from a simultaneously feasible allocation.

Exercise 2: The optimal feasible assignment. Let $a_{jct} = 1$ indicate that teacher j is assigned to classroom c in year t , and let $A = (a_{jct})_{j \in \mathcal{J}_t, c \in \mathcal{C}_t}$. For $b \in \{S, D\}$, define the feasible assignment set

$$\mathcal{A}_t^b = \left\{ A : a_{jct} \in \{0, 1\}, \sum_{c \in \mathcal{C}_t} a_{jct} = 1 \text{ for every } j, \sum_{j \in \mathcal{J}_t} a_{jct} = 1 \text{ for every } c, a_{jct} = 0 \text{ if } c \notin \mathcal{C}_{jt}^b \right\}.$$

The row constraints assign every teacher to exactly one classroom. The column constraints give every classroom exactly one teacher. The eligibility constraints impose the school or district restriction.

For a feasible assignment, total predicted relative value added across students is

$$\widehat{\text{TVA}}_t(A) = \sum_{c \in \mathcal{C}_t} \sum_{j \in \mathcal{J}_t} a_{jct} n_{ct} \text{VA}_t(j, c).$$

Define the model-implied first-best assignment within the specified restrictions by

$$A_t^{\text{FB},b} \in \operatorname{argmax}_{A \in \mathcal{A}_t^b} \widehat{\text{TVA}}_t(A).$$

Let A_t^0 denote the observed assignment. The maximum total value added and the average gain per student are

$$\begin{aligned}\widehat{\text{TVA}}_t^{\text{FB},b} &= \widehat{\text{TVA}}_t(A_t^{\text{FB},b}), \\ \Delta_t^{\text{FB},b} &= \frac{\widehat{\text{TVA}}_t(A_t^{\text{FB},b}) - \widehat{\text{TVA}}_t(A_t^0)}{n_t}.\end{aligned}$$

Because the observed assignment is feasible and the district assignment set contains the school assignment set,

$$0 \leq \Delta_t^{\text{FB},S} \leq \Delta_t^{\text{FB},D}.$$

The pooled average gain across years is

$$\Delta^{\text{FB},b} = \frac{\sum_{t=1}^T \left\{ \widehat{\text{TVA}}_t(A_t^{\text{FB},b}) - \widehat{\text{TVA}}_t(A_t^0) \right\}}{\sum_{t=1}^T n_t}.$$

Computationally, we construct the teacher–classroom payoff matrix with entries $n_{ct}\text{VA}_t(j, c)$ and solve a maximum-weight one-to-one assignment problem separately for each school–grade–year or district–grade–year pool. The optimization accounts for the opportunity cost of assigning a teacher to a classroom that could instead receive another teacher. It can therefore assign some teachers to classrooms other than their individually preferred choices.

Holding $\hat{L}_{jt}$ fixed preserves each teacher’s estimated general effectiveness. Nevertheless, with unequal classroom sizes, the student-weighted level contribution

$$\sum_{c \in \mathcal{C}_t} \sum_{j \in \mathcal{J}_t} a_{jct} n_{ct} \hat{L}_{jt}$$

can change as teachers move between classrooms. The total-value-added optimum can therefore exploit both matching and the allocation of teacher-level effectiveness across classrooms of different sizes.

To isolate the matching channel, define

$$\hat{H}_t(A) = \sum_{c \in \mathcal{C}_t} \sum_{j \in \mathcal{J}_t} a_{jct} n_{ct} M_t(j, c), \quad A_t^{\text{match},b} \in \operatorname{argmax}_{A \in \mathcal{A}_t^b} \hat{H}_t(A).$$

The corresponding average matching gain is

$$\Delta_t^{\text{match},b} = \frac{\hat{H}_t(A_t^{\text{match},b}) - \hat{H}_t(A_t^0)}{n_t}.$$

This statistic measures the improvement in the matching component alone. With unequal classroom sizes, its optimizing assignment can differ from the assignment maximizing total value added.

Let $j_t^{\text{FB},b}(c)$ be classroom c ’s teacher under the optimal feasible assignment. The predicted gain for achievement group g in that classroom is

$$\Delta_{gct}^{\text{FB},b} = \hat{L}_{j_t^{\text{FB},b}(c),t} - \hat{L}_{j_t^0(c),t} + \hat{M}_{g,j_t^{\text{FB},b}(c),t} - \hat{M}_{g,j_t^0(c),t}.$$

The average gain for group g is

$$\Delta_g^{\text{FB},b} = \frac{\sum_{t=1}^T \sum_{c \in \mathcal{C}_t} n_{gct} \Delta_{gct}^{\text{FB},b}}{\sum_{t=1}^T \sum_{c \in \mathcal{C}_t} n_{gct}},$$

provided that its denominator is positive. Although the optimized aggregate gain is non-negative, individual classrooms or achievement groups can experience losses. We therefore report both aggregate gains and their distribution across groups and classrooms.

These counterfactuals change teacher assignments while preserving student rosters, classroom composition, and the fitted teacher effect surface. The first-best is conditional on the estimated model and the specified eligibility restrictions. A causal interpretation additionally requires teacher effects to remain valid after reassignment; cross-school moves require their effectiveness to transfer across school environments. Estimated optimal gains should also be understood with the qualification that optimized assignments could partly reflect estimation error.

Empirical results. The reallocation sample contains 566 teacher–classroom–year pairs and 10,621 student-year observations over 2011–2013. We exclude 17 teacher-year pairs associated with multiple classrooms and retain the original fitted effects and reference weights throughout the exercises.²⁰ Table 7 reports counterfactual gains in standardized state mathematics-score units and individual-best matching diagnostics.

Teachers’ individual opportunities are unevenly distributed. The mean gain from a teacher’s individually best classroom is 0.0644 within schools and 0.1499 within districts, whereas the corresponding medians are 0.0138 and 0.0885. Observed assignments already maximize predicted value added within the eligible menu for 206 of the 566 teacher-year observations (36.40%) under school restrictions and 26 (4.59%) under district restrictions, counting tied optima as already best matched. These opportunities are evaluated separately for each teacher and need not form a feasible joint allocation.

We also examine whether individual best matching changes the classification of teachers initially in the bottom tail of observed classroom-average relative value added. Defining the bottom 5% separately within each year and including cutoff ties yields 29 teacher-year observations. When all teachers are evaluated in their individual best classrooms and the within-year 5th-percentile cutoffs are recalculated, 5 of these observations (17.24%) leave the bottom tail under school restrictions and 8 (27.59%) under district restrictions. Relative to the fixed original within-year cutoffs, 8 observations (27.59%) improve above the threshold under school restrictions and 16 (55.17%) do so under district restrictions. The larger fixed-cutoff escape rates show that some teachers clear the original benchmark while remaining in the counterfactual bottom tail as other teachers also improve.

Under the optimal feasible allocation, pooled average total value added increases by 0.0538 per student within schools and 0.1352 within districts. Matching contributes 0.0429 and 0.0916 to these gains, respectively, accounting for approximately 80% and 68% of the total improvement. The remaining contributions, 0.0109 and 0.0436, reflect the allocation of fixed teacher-level effectiveness across classrooms of different sizes. Separately maximizing the matching component yields matching gains of 0.0476 within schools and 0.0989 within districts, with corresponding total value-added gains of 0.0422 and 0.0939. Thus, the two objectives favor different allocations when classroom enrollment is unequal.

Under the total-value-added optimum, pooled average gains are positive in all five achievement groups. Figure 2 shows that gains are largest in the lowest-achievement group, followed by the highest-achievement group, and smallest in the middle group. Nevertheless, classroom-average gains are negative for 27.0% of classroom-year observations under school

²⁰For the 12 included classrooms containing students from more than one grade, eligibility is determined by the unique modal student grade.

reallocation and 36.7% under district reallocation. Positive aggregate gains therefore coexist with losses in some classrooms.

Table 7: Counterfactual gains and matching diagnostics

	Within school	Within district
<i>Panel 1A: Individually best classroom</i>		
Mean gain	0.0644	0.1499
Median gain	0.0138	0.0885
<i>Panel 1B: Individual-best matching diagnostics</i>		
Already in a best classroom	206 (36.40%)	26 (4.59%)
Escape after re-ranking	5 (17.24%)	8 (27.59%)
Exceed original 5th-percentile cutoff	8 (27.59%)	16 (55.17%)
<i>Panel 2A: Feasible total-value-added optimum</i>		
Total value-added gain	0.0538	0.1352
Matching contribution	0.0429	0.0916
Level contribution	0.0109	0.0436
<i>Panel 2B: Feasible matching-only optimum</i>		
Total value-added gain	0.0422	0.0939
Matching contribution	0.0476	0.0989
Level contribution	-0.0054	-0.0050
<i>Panel 2C: Group gains under the total-value-added optimum</i>		
Q1 (lowest prior achievement)	0.1007	0.2245
Q2	0.0539	0.1138
Q3	0.0197	0.0649
Q4	0.0343	0.0862
Q5 (highest prior achievement)	0.0594	0.1829
<i>Panel 2D: Group gains under the matching-only optimum</i>		
Q1 (lowest prior achievement)	0.1090	0.1992
Q2	0.0490	0.0830
Q3	-0.0030	0.0086
Q4	0.0191	0.0290
Q5 (highest prior achievement)	0.0363	0.1459

Notes: Except for Panels 1B, all report changes from the observed assignment in standardized state mathematics-score units. Panels 2A and 2B report average gains over students and years. Panel 2C and 2D report average gains over students and years for each group. Panel 2B and 2D maximizes matching alone. Panel 1B reports counts with percentages in parentheses. Already-best percentages use all 566 teacher-year observations; escape percentages use the 29 originally bottom-tail teacher-year observations. Bottom-tail membership is defined within each year, including cutoff ties. Re-ranking uses the restriction-specific individual-best value-added distribution; the original-cutoff measure holds the observed threshold fixed.

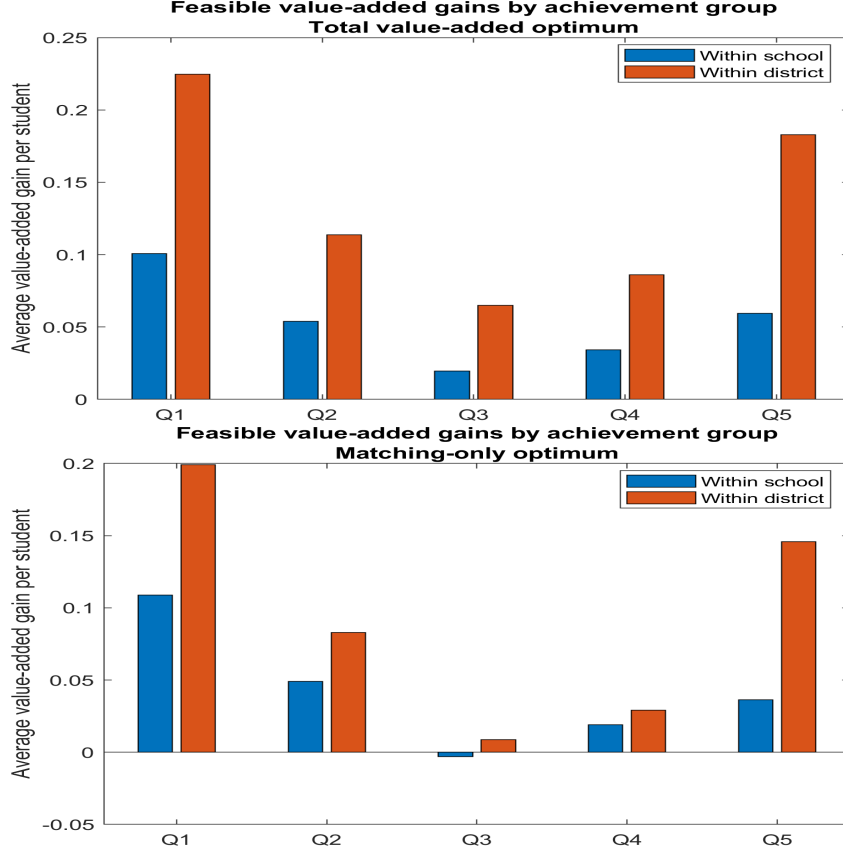


Figure 2: Under reassignment based on total value-added, gains are largest in the lowest-achievement group, followed by the highest-achievement group, and smallest in the middle group. Under within-school reassignment based on matching effects, gains are largest in the lowest-achievement group, followed by the lower-achievement group.

Reassigning a large share of teachers may be difficult to implement in practice. We therefore examine allocations that allow at most a fraction $\rho \in \{0.10, 0.25, 0.50, 1\}$ of included teachers to change classrooms in each year. Let $c_t^0(j)$ denote teacher j 's observed classroom and let a_{jct} indicate assignment to classroom c . We impose $\sum_{j \in \mathcal{J}_t} [1 - a_{j, c_t^0(j), t}] \leq \rho |\mathcal{J}_t|$, which limits the fraction of teachers changing classrooms to at most ρ . The cap is shared across all included teachers within a year, while assignments remain subject to the relevant school or district restriction, preservation of grade, and one-to-one teacher-classroom matching. The specification with $\rho = 1$ reproduces the unrestricted baseline.

Table 8: Gains from feasible teacher reallocation by movement cap

Move cap	Within school				Within district			
	10%	25%	50%	100%	10%	25%	50%	100%
<i>Panel A: Total-value-added optimum</i>								
Average gain	0.028	0.044	0.053	0.054	0.046	0.083	0.117	0.135
Q1 (lowest)	0.065	0.084	0.099	0.101	0.069	0.131	0.188	0.225
Q2	0.028	0.043	0.052	0.054	0.052	0.078	0.098	0.114
Q3	0.012	0.019	0.020	0.020	0.021	0.043	0.053	0.065
Q4	0.010	0.024	0.036	0.034	0.017	0.046	0.067	0.086
Q5 (highest)	0.024	0.047	0.057	0.059	0.070	0.117	0.177	0.183
<i>Panel B: Matching-only optimum</i>								
Average gain	0.026	0.037	0.043	0.042	0.038	0.064	0.085	0.094
Q1 (lowest)	0.067	0.097	0.107	0.109	0.086	0.142	0.184	0.199
Q2	0.029	0.038	0.050	0.049	0.036	0.061	0.072	0.083
Q3	0.011	0.009	0.004	-0.003	0.004	0.005	0.016	0.009
Q4	0.004	0.015	0.019	0.019	0.003	0.016	0.022	0.029
Q5 (highest)	0.017	0.028	0.035	0.036	0.061	0.093	0.130	0.146

Notes: All entries report gains in total relative value added, including level and matching effects, in standardized state mathematics-score units per student. Movement caps apply jointly across all included teachers within each year. The 100% specification imposes no additional movement constraint. Panel B selects assignments by maximizing matching effects alone but reports total value-added gains.

Under the total-value-added optimum, restricting teacher movement to at most 25% per year retains approximately 81.4% and 61.6% of the average gains under unrestricted within-school and within-district reallocation, respectively. Thus, a substantial share of the model-implied gains remains attainable under teacher mobility limited at 25%.

9.4 Forecasting Teacher Effects in the Randomized Sample

The National Center for Teacher Effectiveness (NCTE) study included a randomized-assignment component in its final academic year, 2012–2013. Participating schools first constructed classroom rosters through their usual procedures, after which NCTE staff worked with school leaders to assign participating fourth- and fifth-grade mathematics teachers to those rosters by lottery within eligible school-grade groups.

We compare unrestricted time-varying empirical Bayes estimation with grouped three-way empirical Bayes estimation, which imposes the restrictions on the population mean and covariance described in Appendix 11.8. All models are estimated using 8,944 usable student observations from 306 teachers in 2010–2011 and 2011–2012. For the grouped three-way specification, we consider ranks $\ell \in \{1, 2, 3, 4\}$. We apply two forecast rules to the posterior teacher effects estimated from both training years: carrying forward the 2011–2012 effect and averaging the effects from the two training years.

The target teacher effects are estimated using an unrestricted group-teacher model. In other words, we estimate the benchmark model²¹

$$y_{ijT} = x'_{ijT}\delta^B + \tau_{g(i),j}^B + \varepsilon_{ijT}^B.$$

²¹Achievement groups and the regressors are the same as in the main analysis. In the randomized-roster sample, the grade dummy is absorbed by the teacher-group fixed effects and is not included.

Here, T denotes the final study year, 2012–2013, and j indexes the student’s observed teacher. We decompose the fitted benchmark effects and the forecast effects into level and matching components on the same common eligible teacher–group cells.

Table 9: Out-of-sample R^2 for teacher-level and matching effects: actual-classroom benchmark

Specification	Carry-forward		Two-year average	
	Level	Matching	Level	Matching
Unrestricted time-varying EB	0.3025	0.0423	0.3328	0.0314
Three-way EB, rank 1	0.2051	−0.0059	0.2417	−0.0123
Three-way EB, rank 2	0.3057	0.0194	0.3506	0.0130
Three-way EB, rank 3	0.3110	0.0309	0.3427	0.0264
Three-way EB, rank 4	0.3024	0.0401	0.3328	0.0339

Notes: All forecasting models are estimated using only 2010–2011 and 2011–2012 data. The unrestricted time-varying one-step EB model is estimated by Gaussian EM with monotone Anderson acceleration. The unrestricted 2013 benchmark uses the same 968 original randomized-roster observations, attributing outcomes to observed teachers rather than originally assigned teachers. Evaluation requires at least one benchmark observation per teacher–group cell and uses identical eligible cells across methods: 68 teachers and 256 cells, containing 853 observations. Benchmark and forecast components are decomposed on these common cells. Out-of-sample R^2 is $1 - \sum_k (B_k - P_k)^2 / \sum_k (B_k - \bar{B})^2$, where B_k is the estimated benchmark component, P_k its forecast, and $\bar{B}$ the evaluation-sample benchmark mean. Level calculations weight teachers equally; matching calculations weight eligible cells equally. Negative values indicate greater squared error than the benchmark-mean predictor. Actual-classroom assignments may reflect nonrandom switching.

Table 9 shows that grouped three-way specifications with ranks two and three yield modestly higher teacher-level forecast R^2 than unrestricted time-varying empirical Bayes estimation under both forecast rules. Rank four performs very similarly to the unrestricted specification, with slightly lower carry-forward performance and approximately equal performance under the two-year-average rule. The highest teacher-level forecast R^2 is 0.3506, obtained by the rank-two specification using the two-year average. Forecast R^2 is substantially lower for the estimated matching components. Unrestricted empirical Bayes estimation performs best for matching under carry-forward, whereas the rank-four specification performs best under the two-year-average rule. Within the structured specifications, matching performance improves as rank increases from one to four, suggesting that additional flexibility can help forecast matching effects. However, the targets are estimated unrestricted benchmark effects, and their sampling error, especially in sparsely populated cells, may partly account for the low matching forecast R^2 .

10 Conclusion

This paper develops a grouped three-way interactive fixed-effects model for panel data that can be used to estimate latent matching effects. We establish the asymptotic properties of the least-squares estimator and provide an alternating-least-squares implementation that works well with unbalanced samples and inclusion of additive fixed effects. In particular, we show that consistency of the slope parameter and aggregate consistency of the factor estimates require only two dimensions to diverge. This result is useful in many applications where data has multiple indices but has short samples for many indices.

The proposed least-squares estimator for the interactive fixed-effect model is shown to outperform that of Freeman (2026) in estimating the slope coefficient in the Monte Carlo designs considered. Furthermore, the least-squares estimator can also consistently estimate the interactive fixed effects themselves, in contrast with Freeman (2026), which treats the fixed effects as nuisance parameters.

The empirical application estimates heterogeneous teacher value added using matched student–teacher achievement data. The application illustrates how scalar measures of teacher effectiveness can conceal comparative advantage across student groups. Holding classroom rosters fixed, the fitted model predicts that reallocating at most 25% of teachers within their schools to maximize matching gains increases average achievement by 0.037 standard deviations, with a gain of 0.097 standard deviations for students in the lowest prior-achievement quintile.

We also find that imposing the three-way separability structure within an empirical Bayes framework can modestly improve out-of-sample forecasts of teacher-level effects relative to an unrestricted time-varying empirical Bayes specification.

This framework is relevant to other settings with naturally three- or higher-dimensional data, including worker-firm, consumer-products, product-market, trade, and health care panels. Possible future extensions include estimating latent group structures and estimating dynamic or nonlinear panels with grouped three-way interactive fixed effects.

11 Appendix

11.1 Notations

We use $\circ$, $*$, $\otimes$, $\odot$ to denote the outer, Khatri-Rao, Kronecker, and Hadamard products, respectively. In particular, for

$$A = (a_1, \dots, a_R) \in \mathbb{R}^{I \times R} \quad \text{and} \quad B = (b_1, \dots, b_R) \in \mathbb{R}^{J \times R},$$

the Khatri-Rao product is defined as

$$A * B = (a_1 \otimes b_1, \dots, a_R \otimes b_R) \in \mathbb{R}^{IJ \times R}.$$

For arbitrary matrix $A \in \mathbb{R}^{m \times r}$, define the projection and annihilator matrices

$$P_A = A(A'A)^+ A' = AA^+, \quad M_A = I_m - P_A,$$

where A^+ is the Moore-Penrose inverse. We use $\|\cdot\|_F$ and $\|\cdot\|$ to denote the Frobenius and spectral norms, respectively, for vectors, matrices, and tensors. For tensors $A, B \in \mathbb{R}^{N_1 \times \dots \times N_d}$, define the inner product

$$\langle A, B \rangle = \sum_{i_1=1}^{N_1} \dots \sum_{i_d=1}^{N_d} a_{i_1 \dots i_d} b_{i_1 \dots i_d},$$

so that $\|A\|_F = \langle A, A \rangle^{1/2}$. The spectral norm of A is defined as

$$\|A\| = \max_{\substack{\|u^{(m)}\|_2=1 \\ m=1, \dots, d}} \left| \langle A, u^{(1)} \circ \dots \circ u^{(d)} \rangle \right|,$$

where $u^{(m)} \in \mathbb{R}^{N_m}$ for $m = 1, \dots, d$. For vectors, the Frobenius and spectral norms both coincide with the Euclidean norm, while for matrices the above definition of $\|\cdot\|$ coincides with the usual matrix spectral norm. In particular, for vectors $u^{(m)} \in \mathbb{R}^{N_m}$, $m = 1, \dots, d$,

$$\|u^{(1)} \circ \dots \circ u^{(d)}\| = \|u^{(1)} \circ \dots \circ u^{(d)}\|_F = \prod_{m=1}^d \|u^{(m)}\|_2.$$

For $Y \in \mathbb{R}^{N_1 \times \dots \times N_d}$, define the ℓ -th generalized singular value of a tensor as

$$\delta_\ell(Y) = \inf_{(a_s^{(i)})_{i,k=1}^{d,\ell-1}} \left\| Y - \sum_{s=1}^{\ell-1} (a_s^{(1)} \circ \dots \circ a_s^{(d)}) \right\|.$$

For matrices, this residual equals the ℓ -th singular value by the Eckart–Young theorem.

The mode- n product of a tensor $Y \in \mathbb{R}^{N_1 \times \dots \times N_d}$ with a matrix $A \in \mathbb{R}^{J \times N_n}$ is denoted by

$$Y \times_n A \in \mathbb{R}^{N_1 \times \dots \times N_{n-1} \times J \times N_{n+1} \times \dots \times N_d}$$

and has entries

$$(Y \times_n A)_{i_1 \dots i_{n-1} j i_{n+1} \dots i_d} = \sum_{r=1}^{N_n} y_{i_1 \dots i_{n-1} r i_{n+1} \dots i_d} a_{jr}.$$

The mode- n matricization of Y , denoted by

$$Y_{(n)} \in \mathbb{R}^{N_n \times \prod_{r \neq n} N_r},$$

is the matrix whose columns are the mode- n fibers of Y . We order the mode- n fibers by backward, or last-index-fastest, enumeration: among the modes other than n , the highest-numbered mode varies fastest. Formally, let $(r_1, \dots, r_{d-1})$ be the increasing sequence obtained by removing n from $(1, \dots, d)$. Then

$$(Y_{(n)})_{i_n, k} = y_{i_1, \dots, i_d}, \quad k = 1 + \sum_{q=1}^{d-1} (i_{r_q} - 1) \prod_{h=q+1}^{d-1} N_{r_h},$$

where an empty product is understood to equal one.²²

Likewise, $\text{vec}(Y)$ denotes the vectorization of Y induced by backward enumeration. Specifically,

$$[\text{vec}(Y)]_k = y_{i_1, \dots, i_d}, \quad k = 1 + \sum_{q=1}^d (i_q - 1) \prod_{h=q+1}^d N_h.$$

For matrices, $\text{vec}(\cdot)$ therefore coincides with row-major vectorization.

The main reason we adopt backward enumeration is that it preserves the natural ordering of the modes in the Kronecker and Khatri–Rao representations of outer-product tensors. In particular,

$$\text{vec}(a^{(1)} \circ \dots \circ a^{(d)}) = a^{(1)} \otimes \dots \otimes a^{(d)}.$$

Moreover, let

$$Y = \sum_{s=1}^{\ell} (a_s^{(1)} \circ \dots \circ a_s^{(d)}), \quad A^{(m)} = (a_1^{(m)}, \dots, a_{\ell}^{(m)}) \in \mathbb{R}^{N_m \times \ell}.$$

Then its mode- n matricization is

$$Y_{(n)} = A^{(n)} \left(A^{(r_1)} * \dots * A^{(r_{d-1})} \right)',$$

where $(r_1, \dots, r_{d-1})$ is the increasing sequence of mode indices other than n .

For a tensor $Y \in \mathbb{R}^{N_1 \times \dots \times N_d}$ and $i, j \in \{1, \dots, d\}$, let $K_{i,j} \in \{0, 1\}^{\mathcal{N} \times \mathcal{N}}$, where $\mathcal{N} = \prod_{m=1}^d N_m$, denote the mode-permutation matrix satisfying

$$K_{i,j} \text{vec}(Y_{(j)}) = \text{vec}(Y_{(i)}).$$

More explicitly, let $(r_1^{(n)}, \dots, r_{d-1}^{(n)})$ be the increasing sequence obtained by removing n from $(1, \dots, d)$. Then, for arbitrary vectors $a^{(m)} \in \mathbb{R}^{N_m}$, $m = 1, \dots, d$,

$$K_{i,j} \left(a^{(j)} \otimes a^{(r_1^{(j)})} \otimes \dots \otimes a^{(r_{d-1}^{(j)})} \right) = a^{(i)} \otimes a^{(r_1^{(i)})} \otimes \dots \otimes a^{(r_{d-1}^{(i)})}.$$

The mode-permutation matrices satisfy

$$K_{i,i} = I_{\mathcal{N}}, \quad K_{i,j}^{-1} = K_{j,i} = K'_{i,j}, \quad K_{i,k} = K_{i,j} K_{j,k}.$$

Since $\text{vec}(Y) = \text{vec}(Y_{(1)})$ under our backward-enumeration convention,

$$K_{1,j} \text{vec}(Y_{(j)}) = \text{vec}(Y), \quad j = 1, \dots, d.$$

²²This ordering differs from the convention of Kolda and Bader (2009), under which the lowest-numbered remaining mode varies fastest, i.e., first-index-fastest.

11.2 Derivation of the FWL representation

Since the mode-3 matricization of the interactive component is $W(DU * V)'$, concentrating over W gives

$$\min_{W \in \mathbb{R}^{T \times \ell}} S_{NMT}(\delta, U, V, W) = \frac{1}{NMT} \|(Y - X \times_4 \delta')_{(3)} M_{DU * V}\|_F^2$$

Consequently, the joint normal equations for δ and W imply

$$\hat{\delta} = \left[X_{(4)} (M_{D\hat{U} * \hat{V}} \otimes I_T) X'_{(4)} \right]^{-1} X_{(4)} (M_{D\hat{U} * \hat{V}} \otimes I_T) y$$

This is a Frisch–Waugh–Lovell-type representation conditional on the estimated loadings $D\hat{U}, \hat{V}$: the outcome vector and the columns of $X'_{(4)}$ are projected onto the orthogonal complement of $\text{col}(D\hat{U} * \hat{V}) \otimes \mathbb{R}^T$.

A representation incorporating all three factor normal equations can be obtained. For given (U, V, W) , define

$$\mathcal{T}(U, V, W) = \left\{ \sum_{k=1}^{\ell} (Da_k \otimes v_k \otimes w_k + Du_k \otimes b_k \otimes w_k + Du_k \otimes v_k \otimes c_k) : \right. \\ \left. a_k \in \mathbb{R}^G, \quad b_k \in \mathbb{R}^M, \quad c_k \in \mathbb{R}^T, \quad k = 1, \dots, \ell \right\}.$$

Since D has full column rank, let

$$Q_D = D(D'D)^{-1/2}, \quad \bar{U} = (D'D)^{1/2}U, \quad Q = Q_D \otimes I_M \otimes I_T.$$

Then $DU = Q_D \bar{U}$ and $Q'Q = I_{GMT}$. Let $K_{i,j}^G$ denote the mode-permutation matrices defined in the notation section for tensors of dimensions $G \times M \times T$. Suppressing dependence on (U, V, W) , define the grouped block projectors

$$P_1^G = I_G \otimes P_{V * W}, \\ P_2^G = K_{1,2}^G (I_M \otimes P_{\bar{U} * W}) K_{2,1}^G, \\ P_3^G = K_{1,3}^G (I_T \otimes P_{\bar{U} * V}) K_{3,1}^G.$$

These project onto the grouped perturbation spaces generated by changes in U , V , and W , respectively. Set

$$S^G = P_1^G + P_2^G + P_3^G, \quad \mathcal{T}^G = \text{range}(S^G).$$

Because the P_m^G are orthogonal projectors,

$$\mathcal{T}^G = \text{range}(P_1^G) + \text{range}(P_2^G) + \text{range}(P_3^G), \quad \mathcal{T}(U, V, W) = Q\mathcal{T}^G.$$

The block spaces generally overlap, so S^G need not itself be an orthogonal projector. The orthogonal projectors onto $\mathcal{T}^G$ and $\mathcal{T} = \mathcal{T}(U, V, W)$ are instead

$$P_{\mathcal{T}^G} = S^G (S^G)^+, \quad P_{\mathcal{T}} = Q P_{\mathcal{T}^G} Q'.$$

The corresponding annihilator in the original observation space is

$$M_{\mathcal{T}} = I_{NMT} - Q P_{\mathcal{T}^G} Q' \\ = M_D \otimes I_M \otimes I_T + Q (I_{GMT} - P_{\mathcal{T}^G}) Q'.$$

In particular, this annihilator retains variation orthogonal to the group-membership space. Evaluate the tangent space at the joint estimates and write

$$\hat{\mathcal{T}} = \mathcal{T}(\hat{U}, \hat{V}, \hat{W}), \quad \hat{\pi} = \sum_{k=1}^{\ell} (D\hat{u}_k \otimes \hat{v}_k \otimes \hat{w}_k), \quad \hat{e} = y - X'_{(4)}\hat{\delta} - \hat{\pi}.$$

The normal equations for the three factor blocks imply that $\hat{e}$ is orthogonal to every perturbation in (6), whereas the normal equation for δ gives $X_{(4)}\hat{e} = 0$. Thus,

$$P_{\hat{\mathcal{T}}}\hat{e} = 0, \quad X_{(4)}\hat{e} = 0.$$

Moreover, $\hat{\pi} \in \hat{\mathcal{T}}$, as follows by taking $a_k = \hat{u}_k$ and $b_k = c_k = 0$ in (6). Consequently,

$$\hat{e} = M_{\hat{\mathcal{T}}} \left(y - X'_{(4)}\hat{\delta} \right).$$

Combining this identity with $X_{(4)}\hat{e} = 0$ yields

$$\hat{\delta} = \left[X_{(4)}M_{\hat{\mathcal{T}}}X'_{(4)} \right]^{-1} X_{(4)}M_{\hat{\mathcal{T}}}y,$$

provided that $X_{(4)}M_{\hat{\mathcal{T}}}X'_{(4)}$ is nonsingular.

11.3 Details on Monte Carlo Simulations

Below are the details of the four estimation methods.

1. Grouped three-way least squares.

The grouped three-way least squares estimator uses the observed group assignment and jointly fits

$$y_{ijt} = x'_{ijt}\delta + \kappa + \mu_{g(i)} + \nu_j + \xi_t + \sum_{k=1}^{\ell} a_{g(i),k}b_{jk}f_{tk} + e_{ijt}. \quad (27)$$

We denote the slope component of the joint least-squares estimator by $\hat{\delta}_G$. We absorb the mean of each additive effect in κ so that the additive effects satisfy

$$\sum_{g=1}^G N_g \mu_g = \sum_{j=1}^M \nu_j = \sum_{t=1}^T \xi_t = 0.$$

The factor scales are normalized so that

$$\frac{1}{M} \sum_{j=1}^M b_{jk}^2 = \frac{1}{T} \sum_{t=1}^T f_{tk}^2 = 1, \quad k = 1, \dots, \ell,$$

with the corresponding scales absorbed into the first-mode loadings. No mean-zero restrictions are imposed on the factor columns. The additive effects and trilinear component are estimated jointly, or equivalently the additive effects are profiled from the joint least-squares criterion.

2. Ungrouped three-way least squares.

The second estimator ignores the group restriction and jointly fits

$$y_{ijt} = x'_{ijt}\delta + \kappa + \mu_i + \nu_j + \xi_t + \sum_{k=1}^{\ell} a_{ik}b_{jk}f_{tk} + e_{ijt}. \quad (28)$$

We denote its slope estimator by $\hat{\delta}_U$. This model contains the DGP under

$$\mu_i = \mu_{g(i)}, \quad a_{ik} = u_{g(i),k}, \quad b_{jk} = v_{jk}, \quad f_{tk} = w_{tk}.$$

The estimator uses the same factor scale normalizations as the grouped estimator, while the additive effects satisfy

$$\sum_{i=1}^N \mu_i = \sum_{j=1}^M \nu_j = \sum_{t=1}^T \xi_t = 0.$$

Relative to the grouped specification, this model estimates $N\ell$ rather than $G\ell$ first-mode loadings and $N - 1$ rather than $G - 1$ first-mode additive effects. It therefore introduces $(N - G)(\ell + 1)$ additional first-mode parameters. The comparison between $\hat{\delta}_G$ and $\hat{\delta}_U$ measures the gains from imposing the observed grouping on both the additive effects and the interactive loadings.

3. Freeman linear projection.

The third estimator implements the linear-projection construction in Section 5.3 of Freeman (2026), using separate preliminary matrix interactive fixed effects regressions to estimate the three loading spaces. It does not use the group assignment or jointly fit a trilinear decomposition.

Let $X_q \in \mathbb{R}^{N \times M \times T}$ denote the tensor containing regressor q . We first center the outcome and each regressor in all three modes:

$$\begin{aligned} Y^c &= Y \times_1 M_{\iota_N} \times_2 M_{\iota_M} \times_3 M_{\iota_T}, \\ X_q^c &= X_q \times_1 M_{\iota_N} \times_2 M_{\iota_M} \times_3 M_{\iota_T}, \quad q = 1, \dots, p. \end{aligned} \quad (29)$$

This removes the additive effects while preserving a trilinear representation of the remaining interactive component:

$$\pi_{ijt}^c = \sum_{k=1}^{\ell} (u_{g(i),k} - \bar{u}_k)(v_{jk} - \bar{v}_k)(w_{tk} - \bar{w}_k),$$

where $\bar{u}_k$, $\bar{v}_k$, and $\bar{w}_k$ are the averages over groups, the second mode, and the third mode, respectively.

Write $n = NMT$, $d_1 = N$, $d_2 = M$, and $d_3 = T$. For each $m = 1, 2, 3$, let $Y_{(m)}^c$ and $X_{q,(m)}^c$ denote the mode- m unfoldings, with dimension $d_m \times (n/d_m)$. The preliminary estimates solve

$$(\tilde{\delta}_m, \hat{B}_m) \in \underset{\theta \in \mathbb{R}^p, B: \text{rank}(B) \leq \ell}{\text{argmin}} \frac{1}{n} \left\| Y_{(m)}^c - \sum_{q=1}^p \theta_q X_{q,(m)}^c - B \right\|_F^2,$$

where $B \in \mathbb{R}^{d_m \times (n/d_m)}$. For each mode, form

$$\hat{R}_m = Y_{(m)}^c - \sum_{q=1}^p \tilde{\delta}_{m,q} X_{q,(m)}^c.$$

Let $\hat{L}_m \in \mathbb{R}^{d_m \times \ell}$ contain the leading ℓ left singular vectors of $\hat{R}_m$, normalized by $\hat{L}_m' \hat{L}_m = I_\ell$, and define

$$\hat{M}_m = I_{d_m} - \hat{L}_m \hat{L}_m', \quad m = 1, 2, 3.$$

The singular vectors are extracted before subtracting the fitted low-rank component $\hat{B}_m$. We then transform the centered data:

$$\begin{aligned} \check{Y} &= Y^c \times_1 \hat{M}_1 \times_2 \hat{M}_2 \times_3 \hat{M}_3, \\ \check{X}_q &= X_q^c \times_1 \hat{M}_1 \times_2 \hat{M}_2 \times_3 \hat{M}_3, \quad q = 1, \dots, p. \end{aligned}$$

Writing $\check{x}_{ijt} = ((\check{X}_1)_{ijt}, \dots, (\check{X}_p)_{ijt})'$, the final slope estimator is

$$\hat{\delta}_{FW} = \left(\sum_{i=1}^N \sum_{j=1}^M \sum_{t=1}^T \check{x}_{ijt} \check{x}_{ijt}' \right)^{-1} \sum_{i=1}^N \sum_{j=1}^M \sum_{t=1}^T \check{x}_{ijt} \check{y}_{ijt}.$$

No intercept is included in this final regression.

This comparator is the full-sample linear-projection point estimator, without kernel smoothing or sample splitting. The preliminary centering in (29) also removes pairwise additive variation that is constant in one mode and is therefore stronger than profiling only the additive effects specified in (27) or (28). The two joint three-way least-squares estimators are fitted to the original data and do not use this preliminary three-mode centering.

4. Flattened matrix IFE.

The fourth estimator treats (i, j) as a single cross-sectional index and fits

$$y_{ijt} = x'_{ijt} \delta + \kappa + \alpha_{ij} + \xi_t + \sum_{k=1}^{\ell} \lambda_{ij,k} f_{tk} + e_{ijt}.$$

The model permits unrestricted pair-specific intercepts and loadings, while retaining a rank- ℓ factor structure over time. It contains the DGP under

$$\alpha_{ij} = \mu_{g(i)} + \nu_j, \quad \lambda_{ij,k} = u_{g(i),k} v_{jk}, \quad f_{tk} = w_{tk}.$$

Unlike the ungrouped trilinear specification, it does not impose the additive decomposition of α_{ij} or the separability of $\lambda_{ij,k}$ across i and j .

Let $Y_{(3)}, (X_q)_{(3)} \in \mathbb{R}^{T \times NM}$ denote the mode-3 unfolding of outcome and regressor matrices $X_q \in \mathbb{R}^{N \times M \times T}$ containing regressor q . Profiling out the pair-specific intercepts and time effects gives the slope estimator

$$\hat{\delta}_F \in \operatorname{argmin}_{\theta \in \mathbb{R}^p} \min_{\operatorname{rank}(B) \leq \ell} \frac{1}{NMT} \left\| M_{\ell_T} \left(Y_{(3)} - \sum_{q=1}^p \theta_q (X_q)_{(3)} \right) M_{\ell_{NM}} - B \right\|_F^2.$$

Here $B \in \mathbb{R}^{T \times NM}$ represents the interactive component after profiling out the additive effects.

The grouped, ungrouped, and flattened specifications impose successively weaker restrictions on the conditional mean. Freeman’s estimator differs from these joint least-squares estimators in both its nuisance estimation and the variation retained by its transformations. Consequently, differences in finite-sample performance reflect not only the imposed grouping and separability restrictions but also the estimation procedure.

Numerical Implementation and Performance Measures We consider the sample-size configurations

$$(N, M, T, G) \in \{(24, 24, 8, 4), (48, 24, 8, 6), (48, 48, 8, 8), (48, 48, 12, 8)\}.$$

For each configuration and each error specification, we generate $R = 1,000$ Monte Carlo replications. All four estimators are applied to the same simulated observations within each replication.

The grouped and ungrouped three-way estimators are computed by alternating least squares. Conditional on the current slope and additive effects, we update the three loading matrices sequentially by linear least squares. Conditional on the updated interactive component, we jointly update the slope and additive effects by profiling out the appropriate additive-effect space. The second- and third-mode loading columns are rescaled to have sample mean square one, with compensating scales absorbed into the first-mode loadings. These normalizations preserve the fitted interactive component, and no sample mean-zero restrictions are imposed on the CP loading columns.

Each CP regression is computed from 20 starting values. For the first start, we fit the regression containing only the slope and the specified additive effects and extract the leading ℓ left singular vectors from the second- and third-mode unfoldings of its residual tensor. These vectors initialize the corresponding loading matrices, while the first-mode loadings are initialized by conditional least squares. The remaining starts use orthonormalized independent Gaussian draws for the second- and third-mode loading matrices, followed by the same conditional first-mode update. Only the grouped estimator uses the observed group assignment.

The flattened matrix IFE estimator and each of Freeman’s three preliminary matrix IFE regressions alternate between a slope update and a rank- ℓ truncated singular value decomposition of the regression residual matrix. Each matrix regression also uses 20 starts: the first uses the ordinary least-squares slope from the relevant transformed data, the second uses a zero slope vector, and the remaining starts use Gaussian perturbations of the ordinary least-squares slope, scaled by the residual and regressor root mean squares. Freeman’s final slope is obtained by a single least-squares regression after applying the estimated loading-space projections; no kernel bandwidth or sample splitting is used.

Stopping rule and numerical convergence. Each start is allowed at most 10,000 complete update cycles. Let $Q^{(s)}$ denote the mean squared residual, $\delta^{(s)}$ the slope vector, and $\hat{Y}^{(s)}$ the fitted outcome array after cycle s . For the matrix regressions, these quantities are evaluated on the transformed data entering the relevant regression. We declare numerical

convergence when all three conditions hold:

$$\begin{aligned}\frac{|Q^{(s)} - Q^{(s-1)}|}{\max\{1, |Q^{(s-1)}|\}} &\leq 10^{-8}, \\ \frac{\|\delta^{(s)} - \delta^{(s-1)}\|}{\max\{1, \|\delta^{(s-1)}\|\}} &\leq 10^{-7}, \\ \frac{\|\hat{Y}^{(s)} - \hat{Y}^{(s-1)}\|_F}{\max\{1, \|\hat{Y}^{(s-1)}\|_F\}} &\leq 10^{-5}.\end{aligned}$$

All 1,000 replications yielded finite, converged estimates for every estimator in each sample-size configuration and error specification.

Slope-estimation performance. Let $\hat{\delta}_a^{(r)}$ denote estimator $a \in \{G, U, FW, F\}$ in replication r . For coefficient $q = 1, 2$, let

$$\bar{\delta}_{a,q} = \frac{1}{R} \sum_{r=1}^R \hat{\delta}_{a,q}^{(r)}.$$

We report the empirical bias, standard deviation, and root mean squared error:

$$\begin{aligned}\text{Bias}_{a,q} &= \frac{1}{R} \sum_{r=1}^R \left(\hat{\delta}_{a,q}^{(r)} - \delta_q \right), \\ \text{SD}_{a,q} &= \left\{ \frac{1}{R-1} \sum_{r=1}^R \left(\hat{\delta}_{a,q}^{(r)} - \bar{\delta}_{a,q} \right)^2 \right\}^{1/2}, \\ \text{RMSE}_{a,q} &= \left\{ \frac{1}{R} \sum_{r=1}^R \left(\hat{\delta}_{a,q}^{(r)} - \delta_q \right)^2 \right\}^{1/2}.\end{aligned}$$

11.4 Heatmaps of Teacher Relative Value Added and Matching Effects

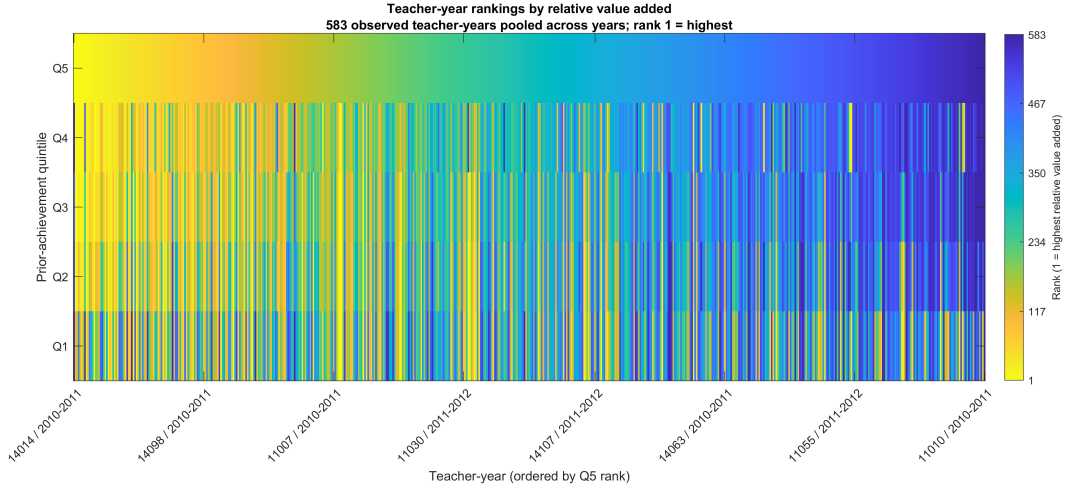


Figure 3: Heatmap of teacher-year rankings of relative value added $\hat{L}_{jt} + \hat{M}_{gjt}$ for each group g , ordered by their Q5 rank.

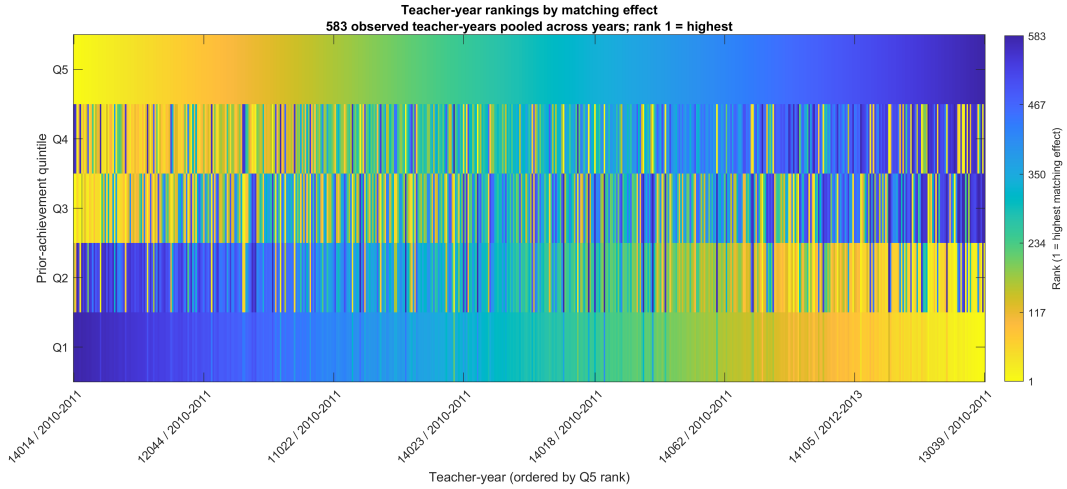


Figure 4: Heatmap of teacher-year rankings of matching effects $\hat{M}_{gjt}$ for each group g , ordered by their Q5 rank. The matching effects add up to zero across groups and across teachers for each year by construction.

In each heatmap, columns represent observed teacher-years and are ordered separately by their estimated ranking for Q5, the highest prior-achievement quintile. For relative value added, the broad Q5 ordering remains visible in Q4 and, to a lesser extent, Q3 and Q2, whereas Q1 exhibits considerable reordering. The common teacher-level component $\hat{L}_{jt}$

coexists with group-specific matching effects that can change the ranking of teacher-years across student groups.

The matching-effect heatmap reveals a sharper contrast. Teacher-years with high estimated matching effects for Q5 generally rank low for Q1, and vice versa. The zero-sum normalization imposes a trade-off across groups, but it does not require the observed near reversal in rankings between Q5 and Q1. This near reversal illustrates the estimated comparative advantage across the prior-achievement distribution. Because the heatmaps display ranks, they describe differences in ordering rather than the magnitude of the underlying effects.

11.5 Regressors in the adaptive LASSO analyses

The regressors used in adaptive LASSO combine 12 CLASS variables (DS0001), 14 MQI measures (DS0002), two teacher characteristics (DS0006), 25 teacher-survey items (DS0007), and two knowledge-of-students scores (DS0008).

CLASS classroom-observation measures (DS0001). We include the following 12 variables from the Classroom Assessment Scoring System (CLASS). These measures describe observed classroom interactions, instructional practices, and student engagement.

CLPC: Positive climate. Positive emotional tone, warmth, respect, and positive relationships in classroom interactions.

CLTS: Teacher sensitivity. The teacher’s sensitivity and responsiveness to students’ academic and emotional needs.

CLRSP: Regard for student perspectives. Attention to students’ ideas and perspectives, including opportunities for autonomy, meaningful choices, and participation.

CLNC: Negative climate. Negative emotional tone and hostile or disrespectful interactions.

CLBM: Behavior management. Management of student behavior and behavioral expectations.

CLPRDT: Productivity. Classroom organization and productive use of instructional time.

CLILF: Instructional learning formats. Instructional formats and activities that support students’ involvement in learning.

CLCU: Content understanding. Instruction that develops understanding of content and connections among ideas, including attention to prior knowledge and misconceptions.

CLAPS: Analysis and inquiry. Opportunities for students to engage in analysis and inquiry.

CLQF: Quality of feedback. Feedback that supports students’ learning and understanding.

CLINSTD: Instructional dialogue. Dialogue and exchanges focused on instructional content.

CLSTENG: Student engagement. Students' observed attention to and involvement in classroom learning.

The underlying CLASS ratings use a seven-point scale: 1–2 indicates low levels, 3–5 middle levels, and 6–7 high levels of the rated dimension. We retain the original direction of each variable. In particular, higher CLNC indicates a more negative classroom climate.

Mathematical Quality of Instruction measures (DS0002). We use 13 MQI variables that are included in the four categories described in Hill et al. (2012). We include these items individually, together with the overall lesson rating, MQI5.

The constituent items are grouped as follows.

- **Richness of mathematics:** five items. LINK measures connections among mathematical ideas; EXPL, mathematical explanations; MMETH, multiple procedures or alternative solution methods; MLANG, use of mathematical language; and MGEN, development of mathematical generalizations.
- **Working with students and mathematics:** two items. USEPROD measures how instruction responds to and uses students' mathematical work or contributions. REMED measures attention to students' mathematical errors, misunderstandings, and difficulties.
- **Errors and imprecision:** three items. MAJERR records major mathematical errors; LANGIMP, imprecision in mathematical language or notation; and LCP, lack of clarity in presenting mathematical content.
- **Student mathematical meaning-making and reasoning:** three items. STEXPL measures students' provision of mathematical explanations; SMQR, students' mathematical questioning and reasoning; and ETCA, cognitive activation of the mathematical task as enacted in the classroom.

We also use MQI5, the observer's overall judgment of mathematical quality across the entire lesson.

Teacher background characteristics (DS0006). We include two background characteristics.

EXPERIENCE. Reported years of teaching experience.

ELMATHCERT. An indicator for self-reported elementary mathematics certification, coded one when the qualification is reported and zero otherwise.

Teacher instructional-practice survey items (DS0007). We include 25 separate raw items from the fall teacher questionnaires. The selected items have questionnaire mappings in all three study years, corresponding to spring-year identifiers 2011, 2012, and 2013. The variables used are

Mathematics instruction: six MIR items. MIR02: explaining why mathematical procedures work; MIR03: using concrete materials to explain why procedures work; MIR05: asking questions about why procedures work; MIR06: comparing different ways of solving procedural problems; MIR08: assigning procedural practice to improve speed,

accuracy, or ease of use; and MIR09: using flashcards, games, or computers to improve recall or skill. Thus, the included MIR items cover both conceptual explanations and procedural practice.

Student reasoning and solution methods: four MIS items. MIS02: inviting students to offer different solution methods; MIS03: assigning problems without specifying a solution method; MIS04: assigning problems requiring a few minutes of thought; and MIS05: asking students to compare or evaluate solution methods.

Classroom climate: eight CLC items. CLC01: students getting along with one another; CLC02: reprimanding students to control the class; CLC04: feeling disrespected by students; CLC05: friendly rapport between teacher and students; CLC06: instructional time lost to misbehavior; CLC07: respectful language and listening; CLC08: students following rules without constant reminders; and CLC09: harsh voices and negative language in class.

Differentiated assignments: one FAS item. FAS03: differentiating mathematics assignments to address individual students' learning needs.

Test-related instructional practices: six TPC items. TPC01: de-emphasizing mathematics topics rarely or never tested; TPC02: focusing on topics carrying greater test weight; TPC03: limiting in-depth discussion of mathematical concepts; TPC04: limiting mathematics projects or hands-on activities; TPC05: reducing demanding problems that challenge advanced students; and TPC07: sequencing topics so tested content is covered before the test.

Teachers' knowledge of students (DS0008). We include two derived scores from the spring teacher questionnaires.

ACCURACY. Accuracy in predicting the percentage of students in the teacher's own class who would answer particular mathematics test items correctly. Higher scores indicate more accurate judgments of class performance.

KOSM. Knowledge of students' mathematical misconceptions, assessed through teachers' predictions of the most common incorrect answers among fourth- or fifth-grade students generally. Higher scores indicate greater knowledge of these misconceptions.

11.6 Cross Validation Errors

Although adaptive LASSO cross-validation for tuning the parameter in matching difference selects $\lambda = 0.0034$, the cross-validation error does not increase dramatically as the tuning parameter decreases to $\lambda = 0.0020$.

Below are the cross-validation error plots for the adaptive LASSO regressions.

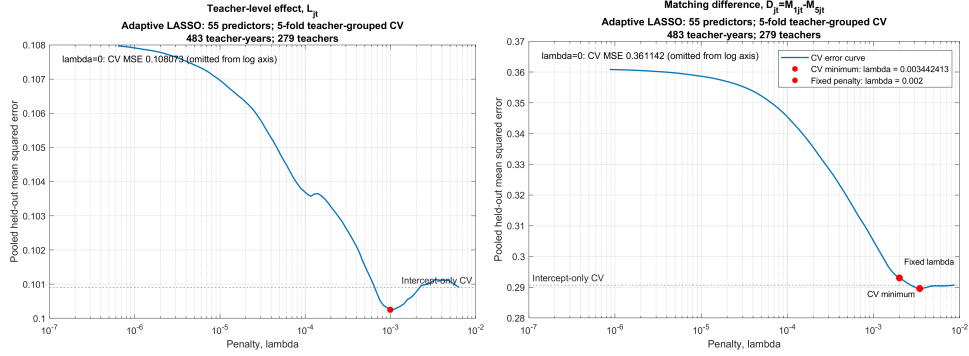


Figure 5: Cross-validation errors of level effects (Left) and matching difference (Right)

We also report the adaptive LASSO result obtained using this penalty parameter.

Table 10: Adaptive LASSO: Matching differential with a fixed penalty

Predictor	Coefficient
<i>MQI measure (DS0002)</i>	
Lack of clarity in presenting mathematical content (LCP)	0.2089
<i>Teacher survey items (DS0007)</i>	
Assigning procedural practice for speed, accuracy, or ease of use (MIR08)	0.0018
Assigning problems requiring a few minutes of thought (MIS04)	0.0620
Fixed penalty, λ	0.0020
Teacher-year observations	483
Teachers	279

Notes: The dependent variable is the matching differential, $D_{jt} = \widehat{M}_{1jt} - \widehat{M}_{5jt}$.

Table 10 uses $\lambda = 0.0020$. A one-point increase in the teacher-year average lack-of-clarity rating (LCP), measured on a 1–3 scale, is associated with a 0.2089σ increase in D_{jt} .²³ The following survey items use 1–5 frequency scales, from rarely or never to always. A one-category increase in assigning procedural practice (MIR08) is associated with a 0.0018σ increase in D_{jt} , while a one-category increase in assigning problems requiring thought (MIS04) is associated with a 0.0620σ increase. These positive coefficients describe shifts in relative matching advantage toward the lowest achievement group.

Below we report the in-sample R^2 of the penalized adaptive LASSO and an unpenalized OLS refit using its selected predictors, together with the cross-validation R^2 of the penalized model.

²³A one-standard-deviation increase in LCP corresponds to about 0.0225σ increase in the state math test score in the matching differential.

Table 11: Fit and cross-validation performance of the adaptive LASSO

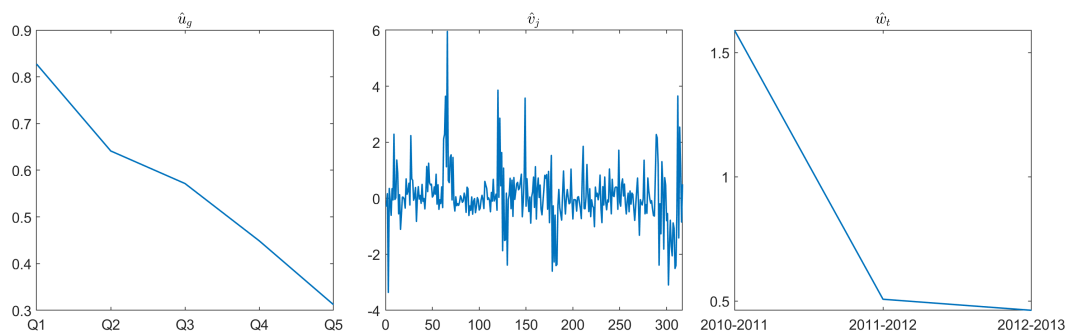
Specification	Penalized R^2	OLS R^2	CV R^2
Level, CV-selected penalty	0.0523	0.0957	0.0065
Matching difference, CV-selected penalty ($\lambda \approx 0.0034$)	0.0134	0.0224	0.0039
Matching difference, fixed penalty ($\lambda = 0.002$)	0.0265	0.0363	-0.0080

Notes: All specifications use 483 teacher-year observations from 279 teachers. Penalized R^2 uses the adaptive LASSO estimates. OLS R^2 is obtained by refitting the selected predictors and an intercept without penalization on the same sample. Cross-validated R^2 evaluates the penalized model and is defined as $R_{CV}^2 = 1 - \text{MSE}_{CV} / \text{MSE}_{CV,0}$, where $\text{MSE}_{CV,0}$ is the intercept-only benchmark. Both MSEs use the same five teacher-grouped folds, with benchmark predictions given by the corresponding training-fold outcome mean. Negative CV R^2 indicates worse predictive performance than this benchmark.

Although teachers differ substantially in their contributions to student achievement, predicting these contributions from observable characteristics remains difficult. Bau and Das (2020) find that observed teacher characteristics account for less than 5% of in-sample teacher-value-added variation, while Rockoff et al. (2011) find that expanding conventional credentials to include cognitive and noncognitive measures yields predictions capturing only about 10% of benchmark teacher-effect variation. Explaining teacher-student matching poses an additional challenge because it requires identifying characteristics associated with teachers' comparative advantage across student groups, rather than their general effectiveness. Consistent with this difficulty, Delgado (2025) finds that measured teacher characteristics and teaching practices explain little variation in comparative advantage.

11.7 Results with one fitted component

Here, we report the results with $\ell = 1$ fitted component.



When $\ell = 1$ is used, the matching effect accounts for 16% of the variation in the relative teacher-value added.

Table 12: Estimated variance components ($\ell = 1$)

Component	Variance	Share of total (%)
Common year, V_T	0.000011	0.01
Group level, V_G	0.000297	0.30
Teacher level, V_L	0.082019	83.86
Matching, V_M	0.015478	15.83
Total teacher component, V_τ	0.097804	100.00

Notes: The relative-value-added variance is $V_L + V_M = 0.097497$, of which matching accounts for 15.88%.

We use adaptive LASSO to regress the teacher-level effect on the 55 regressors. The selected variables for the teacher-level effect are given below. Five-fold cross-validation does not select any variables for the matching difference.

Table 13: Adaptive LASSO: Teacher-level effect $\hat{L}_{jt}$ with $\ell = 1$

Predictor	Coefficient
<i>MQI measure (DS0002)</i>	
Enacted task cognitive activation (ETCA)	0.1061
<i>Teacher survey items (DS0007)</i>	
Comparing different ways of solving procedural problems (MIR06)	0.0161
Asking students to compare or evaluate solution methods (MIS05)	0.0010
Instructional time lost to student misbehavior (CLC06)	-0.0199
Selected penalty, λ	0.0008993
Teacher-year observations	483
Teachers	279

Notes: The adaptive LASSO considers 55 predictors and selects the penalty using five-fold cross-validation, keeping all observations of a teacher in the same fold. Coefficients are reported in the regressors' original units, without post-selection OLS refitting. The intercept is -0.1695.

The tables below report results from counterfactual reassignment of teachers to different classrooms.

Table 14: Counterfactual gains and matching diagnostics ($\ell = 1$)

	Within school	Within district
<i>Panel 1A: Individually best classroom</i>		
Mean gain	0.0321	0.0878
Median gain	0.0043	0.0442
<i>Panel 1B: Individual-best matching diagnostics</i>		
Already in a best classroom	217 (38.34%)	29 (5.12%)
Escape after re-ranking	6 (20.69%)	6 (20.69%)
Exceed original 5th-percentile cutoff	8 (27.59%)	12 (41.38%)
<i>Panel 2A: Feasible total-value-added optimum</i>		
Total value-added gain	0.0288	0.0931
Matching contribution	0.0141	0.0467
Level contribution	0.0147	0.0464
<i>Panel 2B: Feasible matching-only optimum</i>		
Total value-added gain	0.0192	0.0488
Matching contribution	0.0177	0.0531
Level contribution	0.0016	-0.0043
<i>Panel 2C: Group gains under the total-value-added optimum</i>		
Q1 (lowest prior achievement)	0.0461	0.1516
Q2	0.0328	0.0826
Q3	0.0175	0.0632
Q4	0.0203	0.0619
Q5 (highest prior achievement)	0.0272	0.1048
<i>Panel 2D: Group gains under the matching-only optimum</i>		
Q1 (lowest prior achievement)	0.0549	0.1271
Q2	0.0264	0.0460
Q3	0.0033	0.0036
Q4	0.0017	-0.0016
Q5 (highest prior achievement)	0.0100	0.0674

Notes: Gains are changes from the observed assignment in standardized state mathematics-score units. Panel 1A weights the 566 teacher-year observations equally. Panels 2A and 2B report pooled gains weighted by the 10,621 student-year observations; Panels 2C and 2D use student-year weights within each achievement group. Each feasible allocation assigns exactly one teacher to each classroom and preserves teachers' grade and year, student rosters, and fitted effects. Panels 2B and 2D select the allocation by maximizing matching alone; total value-added gains and all group gains include both level and matching effects. Panel 1B reports counts with percentages in parentheses. Already-best percentages use all 566 teacher-years; escape percentages use the 29 originally bottom-tail teacher-years. Bottom-tail membership is defined within each year, including cutoff ties. Re-ranking uses the restriction-specific individual-best value-added distribution; the original-cutoff measure holds the observed threshold fixed. Tied optima count as already best. Individual-best choices are evaluated separately and need not form a feasible joint allocation.

Table 15: Gains from feasible teacher reallocation by movement cap ($\ell = 1$)

Move cap	Within school				Within district			
	10%	25%	50%	100%	10%	25%	50%	100%
<i>Panel A: Total-value-added optimum</i>								
Average gain	0.014	0.023	0.028	0.029	0.031	0.058	0.081	0.093
Q1 (lowest)	0.025	0.041	0.048	0.046	0.039	0.082	0.126	0.152
Q2	0.014	0.024	0.032	0.033	0.026	0.052	0.076	0.083
Q3	0.012	0.015	0.017	0.018	0.022	0.043	0.059	0.063
Q4	0.007	0.013	0.021	0.020	0.026	0.039	0.046	0.062
Q5 (highest)	0.012	0.023	0.025	0.027	0.042	0.070	0.097	0.105
<i>Panel B: Matching-only optimum</i>								
Average gain	0.011	0.015	0.018	0.019	0.021	0.036	0.043	0.049
Q1 (lowest)	0.031	0.040	0.053	0.055	0.054	0.089	0.113	0.127
Q2	0.008	0.019	0.024	0.026	0.017	0.028	0.036	0.046
Q3	0.004	0.007	0.005	0.003	0.005	0.009	0.015	0.004
Q4	0.005	0.002	0.003	0.002	0.000	-0.002	-0.003	-0.002
Q5 (highest)	0.007	0.009	0.008	0.010	0.030	0.056	0.053	0.067

Notes: Results use the $\ell = 1$ model. All entries report gains in total relative value added, including level and matching effects, in standardized state mathematics-score units per student. Movement caps apply jointly across all included teachers within each year; assignments preserve grade and remain within the indicated school or district. The maximum number of movers is $\rho|\mathcal{I}_t|$. The 100% specification imposes no additional movement constraint. Panel B selects assignments by maximizing matching effects alone but reports total value-added gains.

The magnitude of the variation in the relative teacher value added explained by the matching effects and the magnitude of teacher reassignment gains decrease substantially when $\ell = 1$ is used instead of $\ell = 2$. However, the qualitative results stay similar. We see that three out of six variables selected in adaptive LASSO for $\ell = 2$ are still selected. Also, the distribution of the potential gains across students from teacher reassignment is also similar as well, with the lowest-performing students gaining the most.

We impose $\ell = 2$ so that the teacher effects change meaningfully across time. Under $\ell = 1$, every teacher-group combination follows a scalar multiple of the same temporal movement. Also, if the true teacher effects have rank one, both models with $\ell = 1$ and $\ell = 2$ would consistently estimate the teacher effects, and adding an additional component should not produce a persistent increase in the estimated matching variance. Thus, the decrease in the magnitude of the matching effects when $\ell = 1$ can be seen as a sign that the true rank is greater than one.

11.8 Empirical Bayes Estimation and Grouped Three-Way Separability

We formulate empirical Bayes estimation of time-varying group-teacher effects and introduce the grouped three-way specification as interpretable restrictions on the population mean and covariance of these effects. The Gaussian prior pools information across teachers and exploits cross-group and temporal dependence. The estimated teacher-effect tensor can subsequently be decomposed into teacher-level and matching components.

Consider the model

$$y_{ijt} = x'_{ijt}\delta + \theta_{g(i),jt} + \varepsilon_{ijt},$$

where $j = 1, \dots, M$ indexes teachers, $t = 1, \dots, T$ indexes academic years, and $g(i) \in \{1, \dots, G\}$ denotes the observed student group. Define

$$\theta_{jt} = (\theta_{1,jt}, \dots, \theta_{G,jt})', \quad \theta_j = (\theta'_{j1}, \dots, \theta'_{jT})' \in \mathbb{R}^{GT}.$$

Conditional on the observed regressors and observation design, assume

$$\theta_j \sim \mathcal{N}_{GT}(\mu, \Lambda), \quad \Lambda \geq 0,$$

independently across teachers. Here,

$$\mu = (\mu'_1, \dots, \mu'_T)', \quad \Lambda = (\Lambda_{ts})_{t,s=1}^T,$$

where $\mu_t \in \mathbb{R}^G$ contains the population mean effects in year t , and

$$\Lambda_{ts} = \text{Cov}(\theta_{jt}, \theta_{js}) \in \mathbb{R}^{G \times G}.$$

The diagonal blocks describe cross-group variation within years, while the off-diagonal blocks describe dependence across groups and years.

The common conditional prior restricts dependence between teacher-effect trajectories and observed student characteristics or observation patterns. Such dependence may be accommodated through covariate-dependent prior means. Below, we consider population means that are common across teachers.

For teacher j , stack the n_j observed student outcomes across years:

$$Y_j = X_j \delta + Z_j \theta_j + \varepsilon_j,$$

where Z_j is an $n_j \times GT$ selection matrix. Let e_{gt} denote the coordinate vector in $\mathbb{R}^{GT}$ with a one in position $(t-1)G + g$. Each row of Z_j equals the appropriate $e'_{g(i),t}$. This representation accommodates unobserved group-year cells.

Conditional on the teacher effects, observed regressors, and observation design, assume that the student-level errors are independent Gaussian variables with variance $\sigma^2 > 0$. Thus,

$$\varepsilon_j \mid \theta_j, X_j, \text{observation design} \sim \mathcal{N}(0, \sigma^2 I_{n_j}),$$

with conditional independence across teachers.

For mean identification, let X and Z denote the matrices obtained by vertically stacking X_j and Z_j , respectively, and assume that $[X, Z]$ has full column rank. We let μ absorb common group-year means and exclude an overall intercept and other regressors determined entirely by group and year from X . For the unrestricted covariance specification, we also assume sufficient within-cell replication to identify σ^2 and sufficient joint observation of group-year cells across teachers to identify Λ .

Integrating out θ_j gives

$$Y_j \mid X_j, \text{observation design} \sim \mathcal{N}(X_j \delta + Z_j \mu, \mathcal{V}_j), \quad \mathcal{V}_j = Z_j \Lambda Z_j' + \sigma^2 I_{n_j}.$$

Define the negative marginal log-likelihood, up to an additive constant and a positive multiplicative factor, as

$$\mathcal{L}(\delta, \mu, \Lambda, \sigma^2) = \sum_{j=1}^M [\log |\mathcal{V}_j| + (Y_j - X_j \delta - Z_j \mu)' \mathcal{V}_j^{-1} (Y_j - X_j \delta - Z_j \mu)].$$

Unrestricted marginal maximum likelihood minimizes $\mathcal{L}(\delta, \mu, \Lambda, \sigma^2)$ over $\delta, \mu, \Lambda \geq 0$, and $\sigma^2 > 0$.

Grouped three-way separability restrictions. The grouped three-way specification restricts teacher-effect trajectories to

$$\theta_{g,jt} = \nu_j + \sum_{k=1}^{\ell} u_{gk} v_{jk} w_{tk}. \quad (30)$$

Define

$$a_j = (\nu_j, v'_j)', \quad h_{gt} = (1, (u_g \odot w_t)')'.$$

Let $B(U, W) \in \mathbb{R}^{GT \times (\ell+1)}$ have row h'_{gt} in position $(t-1)G + g$. Then

$$\theta_j = B(U, W) a_j.$$

Impose the Gaussian coefficient distribution

$$a_j \sim \mathcal{N}_{\ell+1}(m, \Omega), \quad \Omega \geq 0, \quad (31)$$

independently across teachers, conditional on the observed regressors and observation design. This induces the restrictions

$$\mu = B(U, W)m, \quad \Lambda = B(U, W)\Omega B(U, W)'. \quad (32)$$

Thus, the grouped three-way model restricts both population parameters of the general Gaussian model through the same group-year loading matrix.

To interpret these restrictions, partition

$$m = \begin{pmatrix} m_\nu \\ m_v \end{pmatrix}, \quad \Omega = \begin{pmatrix} \omega_{\nu\nu} & \omega'_{\nu v} \\ \omega_{\nu v} & \Omega_{vv} \end{pmatrix}.$$

The population mean for group g and year t becomes

$$\mu_{gt} = m_\nu + (u_g \odot w_t)' m_v = m_\nu + \sum_{k=1}^{\ell} u_{gk} m_{v,k} w_{tk},$$

while the covariance between a teacher's effects for (g, t) and (h, s) is

$$\Lambda_{(g,t),(h,s)} = \omega_{\nu\nu} + (u_g \odot w_t)' \omega_{\nu v} + \omega'_{\nu v} (u_h \odot w_s) + (u_g \odot w_t)' \Omega_{vv} (u_h \odot w_s).$$

Here, $\omega_{\nu\nu}$ generates covariance across every group-year pair because the teacher baseline ν_j enters all effects. The two terms involving $\omega_{\nu v}$ allow teachers' baseline effectiveness to correlate with their responses to the interactive components. $(u_g \odot w_t)' \Omega_{vv} (u_h \odot w_s)$ generates covariance through teachers' correlated responses to the shared group-year patterns.

Writing

$$H_j(U, W) = Z_j B(U, W),$$

the restricted marginal model has

$$\mathbb{E}[Y_j \mid X_j, \text{observation design}] = X_j \delta + H_j m, \quad \mathcal{V}_j = H_j \Omega H_j' + \sigma^2 I_{n_j}.$$

The restricted estimator therefore solves

$$\min_{\delta, U, W, m, \Omega \geq 0, \sigma^2 > 0} \mathcal{L}(\delta, B(U, W)m, B(U, W)\Omega B(U, W)', \sigma^2),$$

subject to the factor rank and identification constraints. The regression coefficients, shared factors, coefficient distribution, and noise variance are estimated jointly.

Posterior estimation of teacher effects. The same Gaussian posterior formulas apply to both specifications. Let $\psi = (\delta, \mu, \Lambda, \sigma^2)$ denote the marginal parameters, with (μ, Λ) estimated either freely or under (32). Conditional on the fitted parameters, the posterior mean is

$$\hat{\theta}_j = \hat{\mu} + \hat{\Lambda} Z_j' \hat{\mathcal{V}}_j^{-1} (Y_j - X_j \hat{\delta} - Z_j \hat{\mu}), \quad (33)$$

and its covariance is

$$\hat{C}_j^\theta = \hat{\Lambda} - \hat{\Lambda} Z_j' \hat{\mathcal{V}}_j^{-1} Z_j \hat{\Lambda}. \quad (34)$$

The empirical Bayes estimate and conditional posterior variance for an individual group–teacher–year effect are

$$\hat{\theta}_{g,jt}^{\text{EB}} = e_{gt}' \hat{\theta}_j, \quad \widehat{\text{Var}}(\theta_{g,jt} | Y, \psi) = e_{gt}' \hat{C}_j^\theta e_{gt}.$$

For the grouped three-way specification, these estimates can equivalently be computed through the lower-dimensional teacher coefficients. Writing $\hat{a}_j = (\hat{\nu}_j, \hat{v}_j')'$, their posterior mean is

$$\hat{a}_j = \hat{m} + \hat{\Omega} \hat{H}_j' \hat{\mathcal{V}}_j^{-1} (Y_j - X_j \hat{\delta} - \hat{H}_j \hat{m}), \quad (35)$$

and their posterior covariance is

$$\hat{C}_j^a = \hat{\Omega} - \hat{\Omega} \hat{H}_j' \hat{\mathcal{V}}_j^{-1} \hat{H}_j \hat{\Omega}. \quad (36)$$

Consequently,

$$\hat{\theta}_j = \hat{B} \hat{a}_j, \quad \hat{C}_j^\theta = \hat{B} \hat{C}_j^a \hat{B}',$$

and

$$\hat{\theta}_{g,jt}^{\text{EB}} = \hat{\nu}_j + \sum_{k=1}^{\ell} \hat{u}_{gk} \hat{v}_{jk} \hat{w}_{tk}.$$

The resulting posterior trajectories preserve the grouped three-way structure. All posterior variances condition on the fitted shared parameters and exclude uncertainty from estimating them.

Numerical implementation. For the unrestricted specification, an EM algorithm treats θ_j as the latent vector. At iteration q , its E-step computes

$$\bar{\theta}_j^{(q)} = \mathbb{E}_{\psi^{(q)}}[\theta_j | Y], \quad C_j^{\theta, (q)} = \text{Var}_{\psi^{(q)}}(\theta_j | Y)$$

using (33) and (34). Holding these moments fixed, define

$$R_q^\theta(\delta) = \sum_{j=1}^M \left[\|Y_j - X_j \delta - Z_j \bar{\theta}_j^{(q)}\|^2 + \text{tr} \left\{ Z_j C_j^{\theta, (q)} Z_j' \right\} \right].$$

The M-step updates are

$$\begin{aligned}\delta^{(q+1)} &= \left(\sum_j X_j' X_j \right)^{-1} \sum_j X_j' (Y_j - Z_j \bar{\theta}_j^{(q)}), \\ \mu^{(q+1)} &= \frac{1}{M} \sum_j \bar{\theta}_j^{(q)}, \\ \Lambda^{(q+1)} &= \frac{1}{M} \sum_j \left[C_j^{\theta, (q)} + \left(\bar{\theta}_j^{(q)} - \mu^{(q+1)} \right) \left(\bar{\theta}_j^{(q)} - \mu^{(q+1)} \right)' \right], \\ (\sigma^2)^{(q+1)} &= \frac{R_q^\theta(\delta^{(q+1)})}{\sum_j n_j}.\end{aligned}$$

For the grouped three-way specification, we instead implement EM using a_j as the latent vector. At iteration q , the E-step computes

$$\bar{a}_j^{(q)} = \mathbb{E}[a_j \mid Y; \psi_3^{(q)}], \quad C_j^{a, (q)} = \text{Var}(a_j \mid Y; \psi_3^{(q)}),$$

where $\psi_3 = (\delta, U, W, m, \Omega, \sigma^2)$, using (35) and (36). Holding these moments fixed, define

$$R_q^a(\delta, U, W) = \sum_{j=1}^M \left[\left\| Y_j - X_j \delta - H_j(U, W) \bar{a}_j^{(q)} \right\|^2 + \text{tr} \left\{ H_j(U, W) C_j^{a, (q)} H_j(U, W)' \right\} \right].$$

This objective follows from the expected complete-data log-likelihood. Its trace term accounts for posterior uncertainty in the teacher coefficients and must be retained when updating the factors and noise variance.

We update δ, U, W by alternating least squares to reduce R_q^a , yielding a generalized EM algorithm. The remaining updates are

$$\begin{aligned}m^{(q+1)} &= \frac{1}{M} \sum_j \bar{a}_j^{(q)}, \\ \Omega^{(q+1)} &= \frac{1}{M} \sum_j \left[C_j^{a, (q)} + \left(\bar{a}_j^{(q)} - m^{(q+1)} \right) \left(\bar{a}_j^{(q)} - m^{(q+1)} \right)' \right], \\ (\sigma^2)^{(q+1)} &= \frac{R_q^a(\delta^{(q+1)}, U^{(q+1)}, W^{(q+1)})}{\sum_j n_j}.\end{aligned}$$

For the structured specification, factor scales are normalized after all M-step updates have been completed. For nonzero columns, define

$$\begin{aligned}s_{u,k} &= \left(\frac{1}{G} \sum_g u_{gk}^2 \right)^{1/2}, \quad s_{w,k} = \left(\frac{1}{T} \sum_t w_{tk}^2 \right)^{1/2}, \\ A &= \text{diag}(1, s_{u,1} s_{w,1}, \dots, s_{u,\ell} s_{w,\ell}).\end{aligned}$$

Apply the transformations

$$u_{gk}^* = \frac{u_{gk}}{s_{u,k}}, \quad w_{tk}^* = \frac{w_{tk}}{s_{w,k}}, \quad m^* = A m, \quad \Omega^* = A \Omega A'.$$

Since $B^* = BA^{-1}$, these transformations preserve Bm and $B\Omega B'$, and hence preserve the marginal likelihood. The noise-variance update is completed before normalization, and the next E-step uses the normalized parameters. Signs and component labels are fixed by convention.

Both implementations use positive-definite starting values for the latent covariance matrix and a positive starting noise variance. Convergence is assessed using the observed marginal likelihood, and multiple initializations can be used to assess sensitivity to local optima.

11.9 Higher-order extension

Consider the grouped four-way separable fixed effects model

$$y_{ij\ell t} = x'_{ij\ell t}\delta + \sum_{k=1}^r u_{g(i),k}v_{jk}h_{\ell k}w_{tk} + \varepsilon_{ij\ell t}.$$

Let

$$\mathcal{F} = \sigma(X, D, U, V, H, W), \quad \mathcal{I} = \{1, \dots, N\} \times \{1, \dots, M\} \times \{1, \dots, L\} \times \{1, \dots, T\}.$$

For $a = (i_a, j_a, \ell_a, t_a) \in \mathcal{I}$, write $\varepsilon_a = \varepsilon_{i_a j_a \ell_a t_a}$. Suppose that the following conditions hold.

Assumption 9 (Asymptotic Framework, Normalization, and Estimation). *Along an asymptotic sequence satisfying*

$$\min\{NMLT/G, NLT, NMT, NML\} \rightarrow \infty,$$

the dimensions p and r remain fixed. The membership matrix $D \in \{0, 1\}^{N \times G}$ has full column rank, and

$$1 \leq r \leq \min\{G, M, L, T\}.$$

For a constant $c_U > 0$ independent of the panel dimensions, let Θ_r be the set of quadruples $(U, V, H, W) \in \mathbb{R}^{G \times r} \times \mathbb{R}^{M \times r} \times \mathbb{R}^{L \times r} \times \mathbb{R}^{T \times r}$ satisfying

$$\text{diag}\left(\frac{1}{M}V'V\right) = \text{diag}\left(\frac{1}{T}W'W\right) = \text{diag}\left(\frac{1}{L}H'H\right) = \iota_r, \quad \lambda_{\max}\left(\frac{1}{N}U'D'DU\right) \leq c_U,$$

where ι_r is the r -dimensional vector of ones. The true factor matrices admit a representation in Θ_r . The least-squares estimator satisfies

$$(\hat{\delta}, \hat{U}, \hat{V}, \hat{H}, \hat{W}) \in \underset{\delta \in \mathbb{R}^p, (U, V, H, W) \in \Theta_r}{\text{argmin}} \frac{1}{NMLT} \sum_{(i,j,\ell,t) \in \mathcal{I}} \left(y_{ij\ell t} - x'_{ij\ell t}\delta - \sum_{k=1}^r u_{g(i),k}v_{jk}h_{\ell k}w_{tk} \right)^2.$$

Assumption 10 (Strong Factors and Stochastic Bounds). *The regressor $x_{ij\ell t}$ has uniformly bounded fourth moment conditional on group assignment, i.e., there exists a constant $c_x < \infty$ independent of the panel dimensions,*

$$\sup_{(i,j,\ell,t) \in \mathcal{I}} \mathbb{E}[\|x_{ij\ell t}\|^4 | D] \leq c_x$$

almost surely, uniformly along the asymptotic sequence. The grouped-error satisfies

$$\|E^G\| = O_p(\sqrt{G + M + L + T}),$$

where

$$E^G = E \times_1 (D'D)^{-1/2} D' = \left(\frac{1}{\sqrt{N_g}} \sum_{\substack{i=1 \\ g(i)=g}}^N \varepsilon_{ij\ell t} \right)_{g,j,\ell,t=1}^{G,M,L,T}.$$

For constants $0 < c_0 \leq 1 \leq c_1 < c_U < \infty$, with probability approaching one,

$$\begin{aligned} c_0 I_r &\leq \frac{1}{N} U' D' D U \leq c_1 I_r \\ c_0 I_r &\leq \frac{1}{M} V' V \leq c_1 I_r \\ c_0 I_r &\leq \frac{1}{L} H' H \leq c_1 I_r \\ c_0 I_r &\leq \frac{1}{T} W' W \leq c_1 I_r. \end{aligned}$$

Assumption 11. The errors have uniformly finite fourth moments and $\mathbb{E}[\varepsilon_a \mid \mathcal{F}] = 0$. Let $\Gamma_{ab} = \mathbb{E}[\varepsilon_a \varepsilon_b \mid \mathcal{F}]$ for $a, b \in \mathcal{I}$. For some finite constant c_ε , with probability approaching one,

$$\sup_{a \in \mathcal{I}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \leq c_\varepsilon.$$

Define

$$\mathcal{R}^D = \{(a, b) \in \mathcal{I}^2 : g(i_a) = g(i_b) \text{ or } j_a = j_b \text{ or } \ell_a = \ell_b \text{ or } t_a = t_b\}, \quad z_{ab} = \varepsilon_a \varepsilon_b - \Gamma_{ab}.$$

For some finite constant c_4 , with probability approaching one,

$$\sup_{(a,b) \in \mathcal{R}^D} \sum_{(c,d) \in \mathcal{R}^D} |\text{Cov}(z_{ab}, z_{cd} \mid \mathcal{F})| \leq c_4.$$

Assumption 12 (Identification of δ). There exists a constant $c_\delta > 0$ such that, with probability approaching one,

$$\inf_{\tilde{U} \in \mathbb{R}^{G \times r}, \tilde{V} \in \mathbb{R}^{M \times r}, \tilde{H} \in \mathbb{R}^{L \times r}} \lambda_{\min} \left[\frac{1}{NMLT} X_{(5)} (M_{D\tilde{U}*} \tilde{V} *_H \otimes M_W) X'_{(5)} \right] \geq c_\delta.$$

Proposition 5. Suppose that Assumptions 9, 10, 11, and 12 hold. Let $\mathcal{T}$ be the tangent space defined as

$$\begin{aligned} \mathcal{T}(U, V, H, W) = \left\{ \sum_{k=1}^r (Da_k \otimes v_k \otimes h_k \otimes w_k + Du_k \otimes b_k \otimes h_k \otimes w_k \right. \\ \left. + Du_k \otimes v_k \otimes c_k \otimes w_k + Du_k \otimes v_k \otimes h_k \otimes d_k) : \right. \\ \left. a_k \in \mathbb{R}^G, \quad b_k \in \mathbb{R}^M, \quad c_k \in \mathbb{R}^L, \quad d_k \in \mathbb{R}^T, \quad k = 1, \dots, r \right\}. \end{aligned}$$

evaluated at the true parameters (U, V, H, W) . Let

$$\Gamma_{NMLT} = \frac{1}{NMLT} X_{(5)} M_{\mathcal{T}} X'_{(5)}, \quad \mathcal{Z}_{NMLT} = \frac{1}{\sqrt{NMLT}} X_{(5)} M_{\mathcal{T}} \text{vec}(E).$$

The least squares estimator $\hat{\delta}$ defined in Assumption 9 satisfies

$$\Gamma_{NMLT}\sqrt{NMLT}(\hat{\delta} - \delta) = \mathcal{Z}_{NMLT} - \mathcal{B}_1 - \mathcal{B}_2 - \mathcal{B}_3 - \mathcal{B}_4 + R_{NMLT},$$

where $R_{NMLT} = O_p\left(\sqrt{\frac{G+M+L+T}{NMLT}} + \frac{(G+M+L+T)^{3/2}}{NMLT}\right)$. Here,

$$(F_1, F_2, F_3, F_4) = ((D'D)^{1/2}U, V, H, W), \quad (d_1, d_2, d_3, d_4) = (G, M, L, T).$$

Let C_m be the Khatri–Rao product of the three F_a with $a \neq m$, retaining increasing mode order. Also define

$$Q = D(D'D)^{-1/2} \otimes I_M \otimes I_L \otimes I_T, \quad \Omega_m^G = \mathbb{E}\left[(E_{(m)}^G)'E_{(m)}^G \mid \mathcal{F}\right].$$

The biases are given by

$$\mathcal{B}_m = \frac{1}{\sqrt{NMLT}} X_{(5)} P_T Q K_{1,m}^G \text{vec}(F_m(F_m' F_m)^{-1} (C_m' C_m)^{-1} C_m' \Omega_m^G M_{C_m}), \quad m = 1, 2, 3, 4.$$

Here $K_{1,m}^G$ converts the mode- m unfolding's vectorization into the natural grouped tensor order, and acts on tensors of dimensions $G \times M \times L \times T$. Their corresponding bounds are

$$\mathcal{B}_1 = O_p(G/\sqrt{NMLT}), \mathcal{B}_2 = O_p(M/\sqrt{NMLT}), \mathcal{B}_3 = O_p(L/\sqrt{NMLT}), \mathcal{B}_4 = O_p(T/\sqrt{NMLT}).$$

Proof outline. Write $n = NMLT$, $s = G + M + L + T$, and $\rho = \sqrt{s/n}$. Let Π and $\hat{\Pi}$ denote the true and estimated interactive components.

First, global least-squares optimality, uniform identification, and the grouped-error spectral bound give

$$\|\hat{\delta} - \delta\| + \frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{n}} = O_p(\rho).$$

The strong-factor conditions then yield normalized factor errors of order $O_p(\rho)$ after permutation and sign alignment. Since $c_1 < c_U$, the inequality constraint is inactive with probability approaching one, and the unrestricted slope and factor normal equations hold.

Next, expand the joint normal equations around the true parameters and eliminate the first-order factor perturbations using the tangent-space projector. A local parametrization of the estimated loading spaces and the structured error-contraction bounds reduce the quadratic terms to four mode-specific error Gram matrices. Consequently,

$$\Gamma_{NMLT}\sqrt{n}(\hat{\delta} - \delta) = \mathcal{Z}_{NMLT} - \sum_{m=1}^4 \mathcal{B}_m^s + O_p(\rho + \sqrt{n}\rho^3),$$

where $\mathcal{B}_m^s$ is obtained from $\mathcal{B}_m$ by replacing Ω_m^G with $(E_{(m)}^G)'E_{(m)}^G$. Third-order factor perturbations contribute $O_p(\sqrt{n}\rho^3)$, while fourth-order terms are absorbed into this bound because $\rho \rightarrow 0$.

Finally, let $(d_1, d_2, d_3, d_4) = (G, M, L, T)$. The restricted fourth-product covariance condition gives

$$\mathcal{B}_m^s - \mathcal{B}_m = O_p\left(\sqrt{\frac{d_m}{n}}\right), \quad m = 1, 2, 3, 4.$$

Summing these differences contributes $O_p(\rho)$. Substitution and the identity $\sqrt{n}\rho^3 = s^{3/2}/n$ establish the stated expansion. Conditional covariance summability and the strong-factor bounds additionally imply $\mathcal{B}_m = O_p(d_m/\sqrt{n})$. $\square$

11.10 Unbalanced Panel

In matched student–teacher data, only a subset of the potential student–teacher–year observations is available. Let Ω collect the observed indices, i.e.,

$$\mathcal{I} = \{1, \dots, N\} \times \{1, \dots, M\} \times \{1, \dots, T\}, \quad \Omega \subseteq \mathcal{I}, \quad n_\Omega = |\Omega| \geq 1.$$

The observation pattern may be random. For $(i, j, t) \in \Omega$, consider

$$y_{ijt} = x'_{ijt}\delta + \pi_{ijt} + \varepsilon_{ijt}, \quad \pi_{ijt} = \sum_{k=1}^{\ell} u_{g(i),k} v_{jk} w_{tk}.$$

Write $\Pi(U, V, W) = \sum_{k=1}^{\ell} (Du_k \circ v_k \circ w_k)$ and $\pi(U, V, W) = \text{vec}(\Pi(U, V, W))$. Let $R_\Omega \in \{0, 1\}^{n_\Omega \times NMT}$ select the distinct observed entries in the order induced by the vectorization, so $R_\Omega R'_\Omega = I_{n_\Omega}$. Define

$$z_\Omega = R_\Omega \text{vec}(Z),$$

for any third-order tensor Z of relevant dimensions so that we have $y_\Omega = R_\Omega \text{vec}(Y)$, $e_\Omega = R_\Omega \text{vec}(E)$. Also define $X_\Omega = R_\Omega X'_{(4)}$. Then, we have

$$y_\Omega = X_\Omega \delta + \pi_\Omega(U, V, W) + e_\Omega,$$

where

$$\pi_\Omega(U, V, W) = R_\Omega \pi(U, V, W).$$

We distinguish convergence of the fitted interactive component on Ω from recovery of the full latent tensor, including its unobserved entries.

For a constant $c_U > 0$ independent of the panel dimensions, use the normalized factor space

$$\Theta_\ell = \left\{ (\tilde{U}, \tilde{V}, \tilde{W}) \in \mathbb{R}^{G \times \ell} \times \mathbb{R}^{M \times \ell} \times \mathbb{R}^{T \times \ell} : \right. \\ \left. \text{diag} \left(\frac{1}{M} \tilde{V}' \tilde{V} \right) = \text{diag} \left(\frac{1}{T} \tilde{W}' \tilde{W} \right) = \iota_\ell, \lambda_{\max} \left(\frac{1}{N} \tilde{U}' D' D \tilde{U} \right) \leq c_U \right\}.$$

We assume that c_U is sufficiently large so that the true factor matrices admit a representation in Θ_ℓ . Define

$$S_\Omega(\delta, U, V, W) = \frac{1}{n_\Omega} \sum_{(i,j,t) \in \Omega} \left[y_{ijt} - x'_{ijt}\delta - \sum_{k=1}^{\ell} u_{g(i),k} v_{jk} w_{tk} \right]^2 \\ = \frac{1}{n_\Omega} \|y_\Omega - X_\Omega \delta - \pi_\Omega(U, V, W)\|^2,$$

and let

$$(\hat{\delta}, \hat{U}, \hat{V}, \hat{W}) \in \underset{\delta \in \mathbb{R}^p, (U,V,W) \in \Theta_\ell}{\text{argmin}} S_\Omega(\delta, U, V, W) \quad (37)$$

be a global minimizer. Concentrating over δ gives

$$\min_{\delta \in \mathbb{R}^p} \|y_\Omega - X_\Omega \delta - \pi_\Omega(U, V, W)\|^2 = \|M_{X_\Omega}(y_\Omega - \pi_\Omega(U, V, W))\|^2.$$

Whenever $X'_\Omega X_\Omega$ is nonsingular, the first order condition for $\hat{\delta}$ is given by

$$\hat{\delta} = \left(\sum_{(i,j,t) \in \Omega} x_{ijt} x'_{ijt} \right)^{-1} \sum_{(i,j,t) \in \Omega} x_{ijt} \left[y_{ijt} - \sum_{k=1}^{\ell} \hat{u}_{g(i),k} \hat{v}_{jk} \hat{w}_{tk} \right]. \quad (38)$$

Define the observed slices

$$\Omega_g^{(u)} = \{(i, j, t) \in \Omega : g(i) = g\}, \quad \Omega_j^{(v)} = \{(i, j', t) \in \Omega : j' = j\}, \quad \Omega_t^{(w)} = \{(i, j, t') \in \Omega : t' = t\}.$$

At a minimizer for which $\lambda_{\max}(\hat{U}' D' D \hat{U} / N) < c_U$, compensating column rescalings permit unrestricted factor perturbations while maintaining the normalizations. If the corresponding observed block Gram matrices are nonsingular, the factor normal equations consequently yield

$$\begin{aligned} \hat{u}_g &= \left[\sum_{(i,j,t) \in \Omega_g^{(u)}} (\hat{v}_j \odot \hat{w}_t) (\hat{v}_j \odot \hat{w}_t)' \right]^{-1} \sum_{(i,j,t) \in \Omega_g^{(u)}} (\hat{v}_j \odot \hat{w}_t) (y_{ijt} - x'_{ijt} \hat{\delta}), \\ \hat{v}_j &= \left[\sum_{(i,j,t) \in \Omega_j^{(v)}} (\hat{u}_{g(i)} \odot \hat{w}_t) (\hat{u}_{g(i)} \odot \hat{w}_t)' \right]^{-1} \sum_{(i,j,t) \in \Omega_j^{(v)}} (\hat{u}_{g(i)} \odot \hat{w}_t) (y_{ijt} - x'_{ijt} \hat{\delta}), \\ \hat{w}_t &= \left[\sum_{(i,j,t) \in \Omega_t^{(w)}} (\hat{u}_{g(i)} \odot \hat{v}_j) (\hat{u}_{g(i)} \odot \hat{v}_j)' \right]^{-1} \sum_{(i,j,t) \in \Omega_t^{(w)}} (\hat{u}_{g(i)} \odot \hat{v}_j) (y_{ijt} - x'_{ijt} \hat{\delta}). \end{aligned} \quad (39)$$

We iterate over (38) and (39) until convergence. We use multiple initial points and use the parameters with the smallest objective function as our estimates.

Assumption 13 (Uniform slope identification in the observed sample). *For $\tilde{U} \in \mathbb{R}^{G \times \ell}$ and $\tilde{V} \in \mathbb{R}^{M \times \ell}$, define*

$$B_\Omega(\tilde{U}, \tilde{V}) = R_\Omega \left[(D\tilde{U} * \tilde{V}) \otimes I_T, D \otimes I_M \otimes W \right],$$

where W is the true time-factor matrix. The observed regressors satisfy

$$\frac{1}{n_\Omega} \|X_\Omega\|_F^2 = O_p(1).$$

There exists a constant $c_{\delta, \Omega} > 0$, independent of the panel dimensions, such that, with probability approaching one,

$$\inf_{\substack{\tilde{U} \in \mathbb{R}^{G \times \ell} \\ \tilde{V} \in \mathbb{R}^{M \times \ell}}} \lambda_{\min} \left[\frac{1}{n_\Omega} X'_\Omega M_{B_\Omega(\tilde{U}, \tilde{V})} X_\Omega \right] \geq c_{\delta, \Omega}.$$

Assumption 14 (Observed-sample exogeneity and error non-pervasiveness). *Let*

$$\mathcal{F}_\Omega = \sigma(\Omega, D, U, V, W, X_\Omega).$$

The observed errors satisfy $\mathbb{E}[e_\Omega \mid \mathcal{F}_\Omega] = 0$. For a constant $c_\varepsilon < \infty$ independent of the panel dimensions, with probability approaching one,

$$\|\mathbb{E}[e_\Omega e'_\Omega \mid \mathcal{F}_\Omega]\| \leq c_\varepsilon.$$

For a positive integer r , define

$$\mathcal{L}_r^D = \left\{ \sum_{k=1}^r (Da_k \circ b_k \circ c_k) : a_k \in \mathbb{R}^G, b_k \in \mathbb{R}^M, c_k \in \mathbb{R}^T \right\}, \quad \|E\|_{\Omega, r} = \sup_{\substack{Z \in \mathcal{L}_r^D \\ \|z_\Omega\| > 0}} \frac{|e'_\Omega z_\Omega|}{\|z_\Omega\|},$$

where $z_\Omega = R_\Omega \text{vec}(Z)$. For the fixed rank ℓ , we have the following bound

$$\|E\|_{\Omega, 2\ell} = O_p\left(\sqrt{G + M + T}\right).$$

Assumption 13 supplies global slope separation from candidate interactive components in the observed sample. Assumption 14 controls a supremum in the observed norm and therefore applies to an estimated low-rank difference that depends on the errors. Assumption 14 provides a high-level condition that implies non-pervasiveness of the errors on the observed sample. Conditioning on Ω makes the exogeneity requirement conditional on the realized observation pattern. The conditional mean-zero restriction also conditions on the observation pattern. Neither assumption alone identifies interactive-component values outside the observed sample.

Assumption 15 (Sampling stability for the full interactive component). *There exists a constant $c_\Pi > 0$, independent of the panel dimensions, such that, with probability approaching one,*

$$\frac{1}{n_\Omega} \|R_\Omega(\pi(\tilde{U}, \tilde{V}, \tilde{W}) - \pi(U, V, W))\|^2 \geq \frac{c_\Pi}{NMT} \|\pi(\tilde{U}, \tilde{V}, \tilde{W}) - \pi(U, V, W)\|^2 \quad (40)$$

uniformly for all $(\tilde{U}, \tilde{V}, \tilde{W}) \in \Theta_\ell$.

Assumption 15 is imposed only for conclusions about the full latent component. It rules out feasible components that differ substantially from the truth but are indistinguishable, or nearly indistinguishable, on the observed matches. The condition concerns differences from the true component, not every tensor of a given rank.

Lemma 3 (Convergence with an unbalanced sample). *Maintain the model, membership conditions, normalized parameter space, and true-parameter feasibility specified above. Let p, ℓ be fixed, and suppose that*

$$\rho_\Omega = \sqrt{\frac{G + M + T}{n_\Omega}}, \quad \rho_\Omega \rightarrow_p 0. \quad (41)$$

Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be a global minimizer in (37), and suppose that Assumptions 13 and 14 hold.

(i) *Observed-sample convergence. The slope and observed interactive component satisfy*

$$\|\hat{\delta} - \delta\| + \frac{\|R_\Omega(\hat{\pi} - \pi)\|}{\sqrt{n_\Omega}} = O_p(\rho_\Omega). \quad (42)$$

(ii) Full-component convergence. If Assumption 15 also holds, then

$$\frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{NMT}} = O_p(\rho_\Omega). \quad (43)$$

(iii) Full joint loading-space convergence. Suppose Assumption 15 holds and, for a constant $c_f > 0$ independent of the panel dimensions, with probability approaching one,

$$\frac{1}{N}U'D'DU \geq c_f I_\ell, \quad \frac{1}{M}V'V \geq c_f I_\ell, \quad \frac{1}{T}W'W \geq c_f I_\ell. \quad (44)$$

Then

$$\begin{aligned} \frac{1}{\sqrt{MT}} \|M_{\hat{V}*\hat{W}}(V * W)\|_F &= O_p(\rho_\Omega), \\ \frac{1}{\sqrt{NT}} \|M_{D\hat{U}*\hat{W}}(DU * W)\|_F &= O_p(\rho_\Omega), \\ \frac{1}{\sqrt{NM}} \|M_{D\hat{U}*\hat{V}}(DU * V)\|_F &= O_p(\rho_\Omega). \end{aligned} \quad (45)$$

The observed-sample conclusion requires neither Assumption 15 nor the strong-factor bounds in (44). All conclusions are aggregate-norm statements, not uniform fixed-index consistency results.

Remark: Observed fit and full-tensor recovery. Part (i) establishes convergence of the interactive fitted values on the observed matches. Recovery of the full latent tensor additionally requires Assumption 15, which ensures that feasible departures from the true component remain detectable in the observed sample. This condition concerns the geometry of the matches as well as their number. Observing every group, teacher, and period does not by itself establish sampling stability, since disconnected match patterns can leave unobserved interactions unidentified.

Remark: Random matching and conditional exogeneity. The observation pattern may depend on the regressors and latent factors, provided $\mathbb{E}[e_\Omega \mid \mathcal{F}_\Omega] = 0$. Conditioning on Ω makes this exogeneity requirement explicit for the realized matches. The low-rank error supremum is a separate stochastic restriction: it controls error-dependent fitted tensor differences in the observed norm and must be imposed or established in addition to the conditional covariance bound.

11.11 Lemmas used in Proofs

Lemma 4 (Conditional contraction bounds). *Let $e = \text{vec}(E)$, with entries indexed by $\mathcal{I}$ in the order induced by vec . Let $z_{\mathcal{R}^D}$ stack the variables $\{z_{ab} : (a, b) \in \mathcal{R}^D\}$ in a fixed ordering of the relevant ordered pairs. Suppose Assumption 3 holds. The following statements hold.*

(i) For every conformable $\mathcal{F}$ -measurable matrix A , we have

$$\mathbb{E}[\|Ae\|^2 \mid \mathcal{F}] \leq c_\epsilon \|A\|_F^2 \quad (46)$$

with probability approaching one. Consequently,

$$\|Ae\| = O_p(\|A\|_F).$$

(ii) For every conformable $\mathcal{F}$ -measurable matrix B , we have

$$\mathbb{E}[\|Bz_{\mathcal{R}^D}\|^2 \mid \mathcal{F}] \leq c_4 \|B\|_F^2 \quad (47)$$

with probability approaching one. Consequently,

$$\|Bz_{\mathcal{R}^D}\| = O_p(\|B\|_F).$$

(iii) For every $\mathcal{F}$ -measurable square matrix K conformable with e , we have

$$\mathbb{E}[|e'Ke| \mid \mathcal{F}] \leq c_\varepsilon \|K\|_*$$

with probability approaching one, where $\|\cdot\|_*$ denotes the nuclear norm. Consequently,

$$|e'Ke| = O_p(\|K\|_*).$$

No symmetry or pair-support restriction on K is required.

The bounds in parts (i) and (iii), including their stochastic-order consequences, also hold with e replaced by $Q'e$ for any $\mathcal{F}$ -measurable matrix Q satisfying $Q'Q = I$. In particular, they apply to

$$e_G = \text{vec}(E^G) = Q'e, \quad Q = \{D(D'D)^{-1/2}\} \otimes I_M \otimes I_T.$$

The coefficient matrices and vectors, including their dimensions, may vary with the panel dimensions. All stochastic-order statements with a random nonnegative scale are interpreted using the convention $0/0 = 0$; in each application above, a zero scale implies that the corresponding numerator is zero.

Proof. Let $\mathcal{E}_2$, and $\mathcal{E}_4$ be the event that

$$\sup_{a \in \mathcal{I}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \leq c_\varepsilon.$$

and

$$\sup_{(a,b) \in \mathcal{R}^D} \sum_{(c,d) \in \mathcal{R}^D} |\text{Cov}(z_{ab}, z_{cd} \mid \mathcal{F})| \leq c_4$$

holds in Assumption 3.

Part (i). Let

$$\Gamma = \mathbb{E}[ee' \mid \mathcal{F}] = (\Gamma_{ab})_{a,b \in \mathcal{I}},$$

where the rows and columns follow the ordering of e . Since $\mathbb{E}[e \mid \mathcal{F}] = 0$, the matrix Γ is the conditional covariance matrix of e . In particular, it is symmetric and positive semidefinite. Its induced matrix 1-norm and ∞ -norm are equal, so

$$\|\Gamma\| \leq \{\|\Gamma\|_1 \|\Gamma\|_\infty\}^{1/2} = \sup_{a \in \mathcal{I}} \sum_{b \in \mathcal{I}} |\Gamma_{ab}| \leq c_\varepsilon \quad \text{on } \mathcal{E}_2.$$

Because A is $\mathcal{F}$ -measurable,

$$\begin{aligned} \mathbb{E}[\|Ae\|^2 \mid \mathcal{F}] &= \text{tr}(A\mathbb{E}[ee' \mid \mathcal{F}]A') \\ &= \text{tr}(A\Gamma A') \\ &\leq \|\Gamma\|_{\text{op}} \text{tr}(AA') \\ &\leq c_\varepsilon \|A\|_F^2 \quad \text{on } \mathcal{E}_2. \end{aligned}$$

This proves (46).

For any $K > 0$, conditional Markov's inequality yields

$$\mathbf{1}_{\mathcal{E}_2} \mathbb{P}(\|Ae\| > K\|A\|_F \mid \mathcal{F}) \leq \frac{c_\varepsilon}{K^2} \mathbf{1}_{\mathcal{E}_2}.$$

When $\|A\|_F = 0$, the event inside the conditional probability is empty. Taking expectations and bounding the probability of $\mathcal{E}_2^c$ gives

$$\mathbb{P}(\|Ae\| > K\|A\|_F) \leq \frac{c_\varepsilon}{K^2} + \mathbb{P}(\mathcal{E}_2^c).$$

Since $\mathbb{P}(\mathcal{E}_2^c) \rightarrow 0$, first taking the limit superior along the asymptotic sequence and then letting $K \rightarrow \infty$ proves $\|Ae\| = O_p(\|A\|_F)$.

Part (ii). By the definition of z_{ab} ,

$$\mathbb{E}[z_{ab} \mid \mathcal{F}] = 0, \quad (a, b) \in \mathcal{R}^D.$$

The finite fourth-moment condition ensures that these centered products have finite second moments. Let

$$\Sigma_{\mathcal{R}^D} = \mathbb{E}[z_{\mathcal{R}^D} z_{\mathcal{R}^D}' \mid \mathcal{F}].$$

Its entries are

$$[\Sigma_{\mathcal{R}^D}]_{(a,b),(c,d)} = \text{Cov}(z_{ab}, z_{cd} \mid \mathcal{F}).$$

The matrix $\Sigma_{\mathcal{R}^D}$ is symmetric and positive semidefinite. Consequently, Assumption 3 implies

$$\|\Sigma_{\mathcal{R}^D}\|_{\text{op}} \leq \sup_{(a,b) \in \mathcal{R}^D} \sum_{(c,d) \in \mathcal{R}^D} |\text{Cov}(z_{ab}, z_{cd} \mid \mathcal{F})| \leq c_4 \quad \text{on } \mathcal{E}_4.$$

Because B is $\mathcal{F}$ -measurable,

$$\begin{aligned} \mathbb{E}[\|Bz_{\mathcal{R}^D}\|^2 \mid \mathcal{F}] &= \text{tr}(B\Sigma_{\mathcal{R}^D}B') \\ &\leq \|\Sigma_{\mathcal{R}^D}\|_{\text{op}} \|B\|_F^2 \\ &\leq c_4 \|B\|_F^2 \quad \text{on } \mathcal{E}_4. \end{aligned}$$

This proves (47). As in part (i), localization permits arbitrary finite $\mathcal{F}$ -measurable B . Conditional Markov's inequality then gives, for every $K > 0$,

$$\mathbb{P}(\|Bz_{\mathcal{R}^D}\| > K\|B\|_F) \leq \frac{c_4}{K^2} + \mathbb{P}(\mathcal{E}_4^c).$$

Since $\mathbb{P}(\mathcal{E}_4^c) \rightarrow 0$, the asserted stochastic order follows.

Part (iii). On $\mathcal{E}_2$, write a singular value decomposition $K = \sum_h \sigma_h a_h b_h'$, where $\|a_h\| = \|b_h\| = 1$. Part (i) and conditional Cauchy-Schwarz give

$$\begin{aligned} \mathbb{E}[|e'Ke| \mid \mathcal{F}] &\leq \sum_h \sigma_h \{\mathbb{E}[(a_h'e)^2 \mid \mathcal{F}] \mathbb{E}[(b_h'e)^2 \mid \mathcal{F}]\}^{1/2} \\ &\leq c_\varepsilon \sum_h \sigma_h = c_\varepsilon \|K\|_*. \end{aligned}$$

Conditional Markov's inequality yields, for every $L > 0$,

$$\mathbb{P}(|e'Ke| > L\|K\|_*) \leq c_\varepsilon/L + \mathbb{P}(\mathcal{E}_2^c),$$

which proves the asserted stochastic order under the stated zero-scale convention.

For the coordinate extension, apply parts (i) and (iii) to AQ' and QKQ' , respectively. Since $Q'Q = I$,

$$\|AQ'\|_F = \|A\|_F, \quad \|QKQ'\|_* = \|K\|_*.$$

□

Lemma 5 (Low-rank tensor contraction bounds). *Let $E \in \mathbb{R}^{N_1 \times N_2 \times N_3}$. Fix $m \in \{1, 2, 3\}$, and let $r_1 < r_2$ be the remaining mode indices, ordered consistently with the Kronecker products below. Let $R, R_0 \in \mathbb{R}^{N_{r_1} \times \ell}$ and $S, S_0 \in \mathbb{R}^{N_{r_2} \times \ell}$. If $R * S$ has full column rank, then*

$$\|E_{(m)} P_{R*S}\|_F \leq \ell \|E\|, \quad (48)$$

$$\|E_{(m)} M_{R*S}(R_0 * S_0)\|_F \leq \sqrt{\ell + 1} \|E\| \|M_{R*S}(R_0 * S_0)\|_F. \quad (49)$$

For the grouped setting, let $E \in \mathbb{R}^{N \times M \times T}$ and let $D \in \{0, 1\}^{N \times G}$ have full column rank. Define

$$Q_D = D(D'D)^{-1/2}, \quad E^G = E \times_1 Q'_D, \quad \bar{U} = (D'D)^{1/2}U, \quad \bar{U}_0 = (D'D)^{1/2}U_0,$$

where $U, U_0 \in \mathbb{R}^{G \times \ell}$, $V, V_0 \in \mathbb{R}^{M \times \ell}$, and $W, W_0 \in \mathbb{R}^{T \times \ell}$. Let $(m, L, L_0) \in \{(2, W, W_0), (3, V, V_0)\}$. If $(DU * L)$ has full column rank, then

$$\|E_{(m)} P_{DU*L}\|_F = \|E_{(m)}^G P_{\bar{U}*L}\|_F \leq \ell \|E^G\|, \quad (50)$$

$$\|E_{(m)} M_{DU*L}(DU_0 * L_0)\|_F = \|E_{(m)}^G M_{\bar{U}*L}(\bar{U}_0 * L_0)\|_F \leq \sqrt{\ell + 1} \|E^G\| \|M_{DU*L}(DU_0 * L_0)\|_F. \quad (51)$$

All bounds are deterministic and uniform over the admissible loading matrices.

Let $\mathcal{L}_m \subseteq \mathbb{R}^{N_m}$ have dimension r_m , and let Q_m denote its orthogonal projector, for $m = 1, 2, 3$. Define

$$\mathcal{V} = (\mathbb{R}^{N_1} \otimes \mathcal{L}_2 \otimes \mathcal{L}_3) + (\mathcal{L}_1 \otimes \mathbb{R}^{N_2} \otimes \mathcal{L}_3) + (\mathcal{L}_1 \otimes \mathcal{L}_2 \otimes \mathbb{R}^{N_3}).$$

Its orthogonal projector is

$$P_{\mathcal{V}} = I_{N_1} \otimes Q_2 \otimes Q_3 + Q_1 \otimes (I_{N_2} - Q_2) \otimes Q_3 + Q_1 \otimes Q_2 \otimes (I_{N_3} - Q_3).$$

Writing $e = \text{vec}(E)$ and $c_{\mathcal{V}} = \sqrt{r_2 r_3 + r_1 r_3 + r_1 r_2}$, we have

$$\|P_{\mathcal{V}} e\| \leq c_{\mathcal{V}} \|E\|. \quad (52)$$

Consequently,

$$|e'z| \leq c_{\mathcal{V}} \|E\| \|z\|, \quad z \in \mathcal{V},$$

and every conformable matrix H satisfying $H = HP_{\mathcal{V}}$ obeys

$$\|He\| \leq c_{\mathcal{V}} \|H\| \|E\|.$$

These inequalities are deterministic and uniform over the subspaces and coefficient matrices. They therefore also apply when these objects depend on the errors.

Proof. For $q \in \mathbb{R}^{N_{r_1} N_{r_2}}$, let $\text{mat}_{-m}(q) \in \mathbb{R}^{N_{r_1} \times N_{r_2}}$ denote its reshape, with $\text{mat}_{-m}(a \otimes b) = ab'$. Let $r = \text{rank}(\text{mat}_{-m}(q))$, and write a singular value decomposition as

$$\text{mat}_{-m}(q) = \sum_{h=1}^r \sigma_h a_h b_h', \quad \|a_h\| = \|b_h\| = 1.$$

Then $q = \sum_{h=1}^r \sigma_h (a_h \otimes b_h)$. For every unit vector $c \in \mathbb{R}^{N_m}$, the scalar $c' E_{(m)}(a_h \otimes b_h)$ is the contraction of E against unit vectors in its three modes, so its absolute value is bounded by the tensor spectral norm $\|E\|$. Consequently,

$$\begin{aligned} \|E_{(m)}q\| &= \sup_{\|c\|=1} \left| \sum_{h=1}^r \sigma_h c' E_{(m)}(a_h \otimes b_h) \right| \\ &\leq \|E\| \sum_{h=1}^r \sigma_h = \|E\| \|\text{mat}_{-m}(q)\|_* \\ &\leq \sqrt{r} \|E\| \left(\sum_{h=1}^r \sigma_h^2 \right)^{1/2} = \sqrt{r} \|E\| \|q\|. \end{aligned} \tag{53}$$

Here $\|\cdot\|_*$ denotes the nuclear norm, the last inequality follows from Cauchy-Schwarz, and the last equality uses $\|\text{mat}_{-m}(q)\|_F = \|q\|$. The case $q = 0$ is immediate.

Let $Q = (q_1, \dots, q_\ell)$ be an orthonormal basis of the column space of $R * S$, and write $M_{R*S}(R_0 * S_0) = (d_1, \dots, d_\ell)$. Each q_k is a linear combination of the ℓ columns of $R * S$, so its reshape has rank at most ℓ . Each d_k is a rank-one column of $R_0 * S_0$ minus an element in the column space of $R * S$, so its reshape has rank at most $\ell + 1$. Since $P_{R*S} = QQ'$, (53) therefore implies

$$\begin{aligned} \|E_{(m)}P_{R*S}\|_F^2 &= \sum_{k=1}^{\ell} \|E_{(m)}q_k\|^2 \leq \ell \|E\|^2 \sum_{k=1}^{\ell} \|q_k\|^2 = \ell^2 \|E\|^2, \\ \|E_{(m)}M_{R*S}(R_0 * S_0)\|_F^2 &= \sum_{k=1}^{\ell} \|E_{(m)}d_k\|^2 \leq (\ell + 1) \|E\|^2 \sum_{k=1}^{\ell} \|d_k\|^2 = (\ell + 1) \|E\|^2 \|M_{R*S}(R_0 * S_0)\|_F^2. \end{aligned}$$

Taking square roots proves (48)–(49).

For either grouped case, let K be the number of rows of L , and set $Q = Q_D \otimes I_K$. Since $Q'_D Q_D = I_G$, the Khatri-Rao and matricization identities give

$$Q'Q = I_{GK}, \quad (DU * L) = Q(\bar{U} * L), \quad (DU_0 * L_0) = Q(\bar{U}_0 * L_0), \quad E_{(m)}^G = E_{(m)}Q.$$

In particular, $(DU * L)$ has full column rank, and

$$P_{DU*L} = QP_{\bar{U}*L}Q', \quad M_{DU*L}(DU_0 * L_0) = QM_{\bar{U}*L}(\bar{U}_0 * L_0).$$

Consequently,

$$E_{(m)}P_{DU*L} = E_{(m)}^G P_{\bar{U}*L}Q', \quad E_{(m)}M_{DU*L}(DU_0 * L_0) = E_{(m)}^G M_{\bar{U}*L}(\bar{U}_0 * L_0).$$

Because $Q'Q = I_{GK}$, multiplication by Q on the left or by Q' on the right preserves the Frobenius norm. Applying (48), (49) to E^G , $\bar{U} * L$, and $\bar{U}_0 * L_0$ now proves (50), (51).

The three summands defining $P_{\mathcal{V}}$ are mutually orthogonal projectors whose ranges sum to $\mathcal{V}$. Let $\{b_h\}_{h=1}^{r_2}$ and $\{c_k\}_{k=1}^{r_3}$ be orthonormal bases of $\mathcal{L}_2$ and $\mathcal{L}_3$. The definition of the tensor spectral norm gives

$$\|(I_{N_1} \otimes b'_h \otimes c'_k)e\| = \sup_{\|a\|=1} |(a' \otimes b'_h \otimes c'_k)e| \leq \|E\|.$$

Therefore,

$$\|(I_{N_1} \otimes Q_2 \otimes Q_3)e\|^2 \leq r_2 r_3 \|E\|^2.$$

The other two summands are bounded analogously by $r_1 r_3 \|E\|^2$ and $r_1 r_2 \|E\|^2$, since their additional complementary projectors cannot increase the norm. Orthogonality proves (52). The remaining bounds follow from $z = P_{\mathcal{V}}z$ and $H = HP_{\mathcal{V}}$. $\square$

Lemma 6 (Convergence of the generated regressors). *Under Assumptions 1, 2, 3, and 4, the global least-squares estimator defined in Assumption 1 satisfies*

$$\begin{aligned} \frac{1}{\sqrt{MT}} \|M_{\hat{V}*\hat{W}}(V * W)\|_F &= O_p \left(\sqrt{\frac{G+M+T}{NMT}} \right), \\ \frac{1}{\sqrt{NT}} \|M_{D\hat{U}*\hat{W}}(DU * W)\|_F &= O_p \left(\sqrt{\frac{G+M+T}{NMT}} \right), \\ \frac{1}{\sqrt{NM}} \|M_{D\hat{U}*\hat{V}}(DU * V)\|_F &= O_p \left(\sqrt{\frac{G+M+T}{NMT}} \right), \\ \|\hat{\delta} - \delta\| &= O_p \left(\sqrt{\frac{G+M+T}{NMT}} \right). \end{aligned}$$

Also, the common component satisfies

$$\frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{NMT}} = O_p \left(\sqrt{\frac{G+M+T}{NMT}} \right).$$

Proof. Let $n = NMT$, $s = G + M + T$, $\rho = \sqrt{s/n}$, $Q_D = D(D'D)^{-1/2}$, $E^G = E \times_1 Q'_D$, and $e = \text{vec}(E)$. Assumption 2 and Lemma 4(i) give

$$\frac{\|X\|_F}{\sqrt{n}} = O_p(1), \quad \left\| \frac{X_{(4)}e}{\sqrt{n}} \right\| = O_p(1), \quad \frac{\|E^G\|}{\sqrt{n}} = O_p(\rho). \quad (54)$$

Let Π and $\hat{\Pi}$ denote the true and estimated common-component tensors, and define

$$\hat{\gamma} = \hat{\delta} - \delta, \quad \Delta = \hat{\Pi} - \Pi, \quad R = X \times_4 \hat{\gamma}' + \Delta, \quad r_n = \frac{\|R\|_F}{\sqrt{n}}.$$

Global least-squares optimality implies $\|E - R\|_F^2 \leq \|E\|_F^2$, and hence

$$\|R\|_F^2 \leq 2\|\hat{\gamma}\| \|X_{(4)}e\| + 2|\langle E, \Delta \rangle|. \quad (55)$$

Let (U, V, W) and $(\hat{U}, \hat{V}, \hat{W})$ be the true and estimated fixed effects. Choose a full-column-rank submatrix K of $[(D\hat{U} * \hat{V}), (DU * V)]$ with the same column space. Since

$$\Delta_{(3)} = \hat{W}(D\hat{U} * \hat{V})' - W(DU * V)', \quad \Delta_{(3)}P_K = \Delta_{(3)},$$

and K contains at most 2ℓ grouped rank-one columns, Lemma 5 gives

$$\begin{aligned} |\langle E, \Delta \rangle| &= |\langle E_{(3)} P_K, \Delta_{(3)} \rangle| \\ &\leq \|E_{(3)} P_K\|_F \|\Delta\|_F \leq 2\ell \|E^G\| \|\Delta\|_F. \end{aligned} \quad (56)$$

The contraction bound is deterministic and uniform, so it applies to the data-dependent matrix K .

Next, $M_W \Pi_{(3)} = 0$ and $\hat{\Pi}_{(3)} M_{D\hat{U}*\hat{V}} = 0$. Therefore, Assumption 4 implies that, with probability approaching one,

$$\begin{aligned} r_n^2 &\geq \frac{1}{n} \|M_W R_{(3)} M_{D\hat{U}*\hat{V}}\|_F^2 \\ &= \hat{\gamma}' \left[\frac{1}{n} X_{(4)} (M_{D\hat{U}*\hat{V}} \otimes M_W) X'_{(4)} \right] \hat{\gamma} \\ &\geq c_\delta \|\hat{\gamma}\|^2 \end{aligned}$$

for some $c_\delta > 0$. On this event,

$$\|\hat{\gamma}\| \leq c_\delta^{-1/2} r_n, \quad \frac{\|\Delta\|_F}{\sqrt{n}} \leq r_n + \frac{\|X\|_F}{\sqrt{n}} \|\hat{\gamma}\| \leq \left(1 + c_\delta^{-1/2} \frac{\|X\|_F}{\sqrt{n}}\right) r_n.$$

Substituting these inequalities and (56) into (55), and using (54) and $n^{-1/2} \leq \rho$, yields

$$r_n^2 \leq \frac{2}{\sqrt{n}} \left\| \frac{X_{(4)} e}{\sqrt{n}} \right\| \|\hat{\gamma}\| + 4\ell \frac{\|E^G\|}{\sqrt{n}} \frac{\|\Delta\|_F}{\sqrt{n}} \leq O_p(\rho) r_n.$$

Thus $r_n = O_p(\rho)$, and consequently

$$\|\hat{\delta} - \delta\| + \frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{n}} = O_p(\rho). \quad (57)$$

It remains to obtain the three span bounds. For $m = 1, 2, 3$, write $\Pi_{(m)} = A_m C'_m$, where

$$(A_m, C_m, s_m) = \begin{cases} (DU, V * W, N), & m = 1, \\ (V, DU * W, M), & m = 2, \\ (W, DU * V, T), & m = 3, \end{cases}$$

and let $\hat{A}_m, \hat{C}_m$ denote the corresponding estimated matrices. Since $\hat{\Pi}_{(m)} = \hat{A}_m \hat{C}'_m$,

$$\Delta_{(m)} M_{\hat{C}_m} = -A_m C'_m M_{\hat{C}_m}.$$

Assumption 2 bounds $\lambda_{\min}(A'_m A_m / s_m)$ away from zero with probability approaching one. Hence, for each m ,

$$\sqrt{\frac{s_m}{n}} \|M_{\hat{C}_m} C_m\|_F \leq \left[\lambda_{\min} \left(\frac{A'_m A_m}{s_m} \right) \right]^{-1/2} \frac{\|\Delta\|_F}{\sqrt{n}} = O_p(\rho).$$

The normalizations $\sqrt{s_m/n}$ are $1/\sqrt{MT}$, $1/\sqrt{NT}$, and $1/\sqrt{NM}$, respectively. Together with (57), these establish all four claims. $\square$

References

- Ahn, T., Aucejo, E. M., and James, J. (2025). The importance of student-teacher matching: A multidimensional value-added approach. *The Review of Economics and Statistics*, pages 1–45.
- Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.
- Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.
- Bates, M., Dinerstein, M., Johnston, A. C., and Sorkin, I. (2025). Teacher labor market policy and the theory of the second best. *The Quarterly Journal of Economics*, 140(2):1417–1469.
- Bau, N. and Das, J. (2020). Teacher value added in a low-income country. *American Economic Journal: Economic Policy*, 12(1):62–96.
- Biasi, B., Fu, C., and Stromme, J. (2026). Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage.
- Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.
- Chang, Y., Ha, S., and Park, J. (2026). A general factor model with common components identified by tensor structure.
- Chen, B., Han, Y., and Yu, Q. (2026). Estimation and inference for CP tensor factor models. *Journal of Econometrics*, 253:106167.
- Delgado, W. (2025). Disparate teacher effects, comparative advantage, and match quality. *Economics of Education Review*, 106:102648.
- Freeman, H. (2026). Linear multidimensional regression with interactive fixed-effects.
- Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica*, 70(4):1639–1657.
- Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A., Litke, E., and Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, 17(2-3):88–106.
- Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.
- Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138.
- Kwon, S. (2026). Optimal shrinkage estimation of fixed effects in linear panel data models. *Econometrica*, 94(2):663–677.

- Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.
- Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.
- Rockoff, J. E., Jacob, B. A., Kane, T. J., and Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, 6(1):43–74.

A Proofs for Section 5 (Asymptotic Theory)

Lemma 1 (Normalized factor alignment and asymptotic inactivity of the inequality constraint). *Suppose that Assumptions 1, 2, 3, and 4 hold. Let (U, V, W) denote the true factor matrices in the normalized representation satisfying (7). Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be the global least-squares estimator defined in Assumption 1. There exist possibly random $\ell \times \ell$ permutation and diagonal sign matrices Π_ℓ, S_v, S_w , with the diagonal entries of S_v and S_w belonging to $\{-1, 1\}$, such that*

$$\frac{\|D(\hat{U}\Pi_\ell S_v S_w - U)\|_F}{\sqrt{N}} + \frac{\|\hat{V}\Pi_\ell S_v - V\|_F}{\sqrt{M}} + \frac{\|\hat{W}\Pi_\ell S_w - W\|_F}{\sqrt{T}} = O_p\left(\sqrt{\frac{G + M + T}{NMT}}\right). \quad (8)$$

Define the residual tensor

$$\hat{E} = Y - X \times_4 \hat{\delta}' - \sum_{k=1}^{\ell} (D\hat{u}_k \circ \hat{v}_k \circ \hat{w}_k).$$

With probability approaching one, the estimator satisfies the unrestricted normal equations

$$\begin{aligned} X_{(4)} \text{vec}(\hat{E}) &= 0, & D' \hat{E}_{(1)} (\hat{V} * \hat{W}) &= 0, \\ \hat{E}_{(2)} (D\hat{U} * \hat{W}) &= 0, & \hat{E}_{(3)} (D\hat{U} * \hat{V}) &= 0. \end{aligned} \quad (9)$$

Proof. Write $n = NMT$, $s = G + M + T$, and $\rho = \sqrt{s/n}$. Assumption 1 implies $\rho \rightarrow 0$. Let Π and $\hat{\Pi}$ denote the true and estimated common-component tensors. Define

$$Q_D = D(D'D)^{-1/2}, \quad A_1 = \frac{(D'D)^{1/2}U}{\sqrt{N}}, \quad A_2 = \frac{V}{\sqrt{M}}, \quad A_3 = \frac{W}{\sqrt{T}},$$

with estimated counterparts $\hat{A}_m$, and set

$$\mathcal{P} = \frac{1}{\sqrt{n}} \Pi \times_1 Q'_D, \quad \hat{\mathcal{P}} = \frac{1}{\sqrt{n}} \hat{\Pi} \times_1 Q'_D, \quad h = \hat{\mathcal{P}} - \mathcal{P}.$$

Also, define

$$C_1 = A_2 * A_3, \quad \hat{C}_1 = \hat{A}_2 * \hat{A}_3.$$

Then

$$\mathcal{P}_{(1)} = A_1 C'_1, \quad \hat{\mathcal{P}}_{(1)} = \hat{A}_1 \hat{C}'_1.$$

The preliminary common-component bound established by the global least-squares comparison in Lemma 6 gives

$$\|h\|_F = \frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{n}} = O_p(\rho). \quad (58)$$

This bound uses global optimality and does not require the factor normal equations. The equality follows because $Q'_D Q_D = I_G$ and the first-mode fibers of both Π and $\hat{\Pi}$ belong to $\text{col}(D)$. All inverses below are used on the indicated events whose probabilities tend to one; the eventual permutation and sign matrices may be defined arbitrarily on the complementary events.

1. Conditioning of the estimated joint loading space. On the strong-factor event,

$$c_0 I_\ell \leq A'_m A_m \leq c_1 I_\ell, \quad m = 1, 2, 3.$$

For every $z \in \mathbb{R}^\ell$,

$$\begin{aligned} \|C_1 z\|^2 &= \|A_2 \text{diag}(z) A'_3\|_F^2 \\ &\geq \sigma_{\min}(A_2)^2 \sigma_{\min}(A_3)^2 \|\text{diag}(z)\|_F^2 \\ &\geq c_0^2 \|z\|^2. \end{aligned}$$

Thus $\sigma_{\min}(C_1) \geq c_0$, and

$$\sigma_\ell(\mathcal{P}_{(1)}) \geq \sigma_{\min}(A_1) \sigma_{\min}(C_1) \geq c_0^{3/2}.$$

By (58) and the singular-value perturbation inequality,

$$\sigma_\ell(\hat{\mathcal{P}}_{(1)}) \geq \sigma_\ell(\mathcal{P}_{(1)}) - \|h_{(1)}\| \geq \frac{c_0^{3/2}}{2}$$

with probability approaching one. Since $\hat{\mathcal{P}}_{(1)} = \hat{A}_1 \hat{C}'_1$ has rank at most ℓ , it follows that $\hat{A}_1$ and $\hat{C}_1$ have full column rank on this event. Moreover, feasibility gives $\|\hat{A}_1\| \leq \sqrt{c_U}$, so

$$\sigma_{\min}(\hat{C}_1) \geq \frac{\sigma_\ell(\hat{\mathcal{P}}_{(1)})}{\|\hat{A}_1\|} \geq b_*, \quad b_* := \frac{c_0^{3/2}}{2\sqrt{c_U}} > 0. \quad (59)$$

The identity

$$h_{(1)} M_{\hat{C}_1} = -A_1 C'_1 M_{\hat{C}_1}$$

and the lower bound on $\sigma_{\min}(A_1)$ imply

$$\|M_{\hat{C}_1} C_1\|_F \leq c_0^{-1/2} \|h\|_F.$$

Since $\text{col}(C_1)$ and $\text{col}(\hat{C}_1)$ both have dimension ℓ , their projector distance satisfies

$$\begin{aligned} \|P_{\hat{C}_1} - P_{C_1}\| &= \|M_{\hat{C}_1} C_1 (C'_1 C_1)^{-1/2}\| \\ &\leq \frac{\|M_{\hat{C}_1} C_1\|_F}{\sigma_{\min}(C_1)} \\ &\leq c_0^{-3/2} \|h\|_F = O_p(\rho). \end{aligned}$$

2. Normalization-preserving alignment. For $m = 2, 3$, let $d_2 = M$ and $d_3 = T$, and complete A_m to an invertible square matrix

$$B_m = [A_m, A_{m,\perp}],$$

where the columns of $A_{m,\perp}$ form an orthonormal basis of $\text{col}(A_m)^\perp$. Then

$$B'_m B_m = \begin{pmatrix} A'_m A_m & 0 \\ 0 & I_{d_m - \ell} \end{pmatrix}.$$

Because $c_0 \leq 1 \leq c_1$, the strong-factor event gives

$$\|B_m\| \leq \sqrt{c_1}, \quad \|B_m^{-1}\| \leq c_0^{-1/2}.$$

Let

$$L = B_2^{-1} \otimes B_3^{-1}.$$

Since $B_m^{-1}a_{m,k} = e_k^{(d_m)}$ for $k = 1, \dots, \ell$, the image of $\text{col}(C_1)$ under L is

$$\mathcal{H}_0 = \text{span}\{e_k^{(M)} \otimes e_k^{(T)} : k = 1, \dots, \ell\},$$

where $(e_k^{(M)})$ and $(e_k^{(T)})$ are elementary basis of dimension M and T respectively. Furthermore,

$$\|L\| \leq c_0^{-1}, \quad \|L^{-1}\| \leq c_1.$$

For any nonzero $z \in \text{col}(\hat{C}_1)$, we have $P_{\hat{C}_1}z = z$ and $LP_{C_1}z \in \mathcal{H}_0$. Consequently,

$$\text{dist}(Lz, \mathcal{H}_0) = \min_{z_0 \in \mathcal{H}_0} \|Lz - z_0\| \leq \|Lz - LP_{C_1}z\| = \|L(P_{\hat{C}_1} - P_{C_1})z\| \leq \|L\| \|P_{\hat{C}_1} - P_{C_1}\| \|z\|.$$

Also,

$$\|Lz\| \geq \frac{\|z\|}{\|L^{-1}\|}.$$

It follows that

$$\begin{aligned} \epsilon_n &:= \sup_{\substack{z \in \text{col}(\hat{C}_1) \\ \|z\|=1}} \frac{\text{dist}(Lz, \mathcal{H}_0)}{\|Lz\|} \\ &\leq \|L\| \|L^{-1}\| \|P_{\hat{C}_1} - P_{C_1}\| \\ &\leq \frac{c_1}{c_0} \|P_{\hat{C}_1} - P_{C_1}\| = O_p(\rho). \end{aligned} \tag{60}$$

For each estimated component $r = 1, \dots, \ell$, define

$$x_r = \frac{B_2^{-1}\hat{a}_{2,r}}{\|B_2^{-1}\hat{a}_{2,r}\|}, \quad y_r = \frac{B_3^{-1}\hat{a}_{3,r}}{\|B_3^{-1}\hat{a}_{3,r}\|},$$

where $a_{m,r}$ and $\hat{a}_{m,r}$ denote columns of A_m and $\hat{A}_m$. Let $\hat{c}_r = \hat{a}_{2,r} \otimes \hat{a}_{3,r}$ be column r of $\hat{C}_1$. The column normalizations imply $\|\hat{c}_r\| = 1$, and

$$q_r := x_r \otimes y_r = \frac{L\hat{c}_r}{\|L\hat{c}_r\|}.$$

Therefore, q_r is a unit vector in $L \text{col}(\hat{C}_1)$ and

$$\|M_{\mathcal{H}_0} q_r\| = \frac{\text{dist}(L\hat{c}_r, \mathcal{H}_0)}{\|L\hat{c}_r\|} \leq \epsilon_n.$$

The vectors $e_k^{(M)} \otimes e_k^{(T)}$, $k = 1, \dots, \ell$, are orthonormal, so

$$P_{\mathcal{H}_0} q_r = \sum_{k=1}^{\ell} x_{r,k} y_{r,k} (e_k^{(M)} \otimes e_k^{(T)}).$$

By orthogonality,

$$1 - \sum_{k=1}^{\ell} x_{r,k}^2 y_{r,k}^2 = \|M_{\mathcal{H}_0} q_r\|^2 \leq \epsilon_n^2.$$

Choose k_r maximizing $x_{r,k}^2$ over $k = 1, \dots, \ell$, with ties broken by a fixed rule. Since x_r and y_r are unit vectors,

$$1 - \epsilon_n^2 \leq \sum_{k=1}^{\ell} x_{r,k}^2 y_{r,k}^2 \leq x_{r,k_r}^2 \sum_{k=1}^{\ell} y_{r,k}^2 \leq x_{r,k_r}^2.$$

Moreover,

$$\sum_{k=1}^{\ell} x_{r,k}^2 y_{r,k}^2 \leq y_{r,k_r}^2 + \sum_{\substack{k \leq \ell \\ k \neq k_r}} x_{r,k}^2 \leq y_{r,k_r}^2 + 1 - x_{r,k_r}^2 \leq y_{r,k_r}^2 + \epsilon_n^2.$$

Hence

$$x_{r,k_r}^2 \geq 1 - \epsilon_n^2, \quad y_{r,k_r}^2 \geq 1 - 2\epsilon_n^2.$$

Choose signs $s_{v,r}, s_{w,r} \in \{-1, 1\}$ satisfying

$$s_{v,r} x_{r,k_r} = |x_{r,k_r}|, \quad s_{w,r} y_{r,k_r} = |y_{r,k_r}|.$$

Then

$$\begin{aligned} \|x_r - s_{v,r} e_{k_r}^{(M)}\|^2 &= 2(1 - |x_{r,k_r}|) \leq 2(1 - x_{r,k_r}^2) \leq 2\epsilon_n^2, \\ \|y_r - s_{w,r} e_{k_r}^{(T)}\|^2 &= 2(1 - |y_{r,k_r}|) \leq 2(1 - y_{r,k_r}^2) \leq 4\epsilon_n^2. \end{aligned}$$

To return to the original coordinates, note that for any invertible matrix B and unit vectors x, y ,

$$\left\| \frac{Bx}{\|Bx\|} - \frac{By}{\|By\|} \right\| \leq \frac{2\|B(x-y)\|}{\|Bx\|} \leq \frac{2\|B\|}{\sigma_{\min}(B)} \|x-y\|.$$

The estimated column normalizations give

$$\hat{a}_{2,r} = \frac{B_2 x_r}{\|B_2 x_r\|}, \quad \hat{a}_{3,r} = \frac{B_3 y_r}{\|B_3 y_r\|},$$

while the true column normalizations imply

$$\frac{B_m(s e_k^{(d_m)})}{\|B_m(s e_k^{(d_m)})\|} = s a_{m,k}, \quad s \in \{-1, 1\}.$$

The uniform bounds on B_m and B_m^{-1} therefore yield

$$\|\hat{a}_{2,r} - s_{v,r} a_{2,k_r}\| + \|\hat{a}_{3,r} - s_{w,r} a_{3,k_r}\| \leq K\epsilon_n, \quad r = 1, \dots, \ell, \quad (61)$$

where K is a finite constant depending only on c_0 and c_1 .

We next show that different estimated components select different coordinates. Let $c_k = a_{2,k} \otimes a_{3,k}$ and $\eta_r = s_{v,r} s_{w,r}$. Since all columns of $A_2, A_3, \hat{A}_2, \hat{A}_3$ have unit norm, (61) implies

$$\|\hat{c}_r - \eta_r c_{k_r}\| \leq K\epsilon_n.$$

If $k_r = k_s$ for some $r \neq s$, then

$$\|\hat{c}_r - \eta_r \eta_s \hat{c}_s\| \leq 2K\epsilon_n.$$

Consequently,

$$\sigma_{\min}(\hat{C}_1) \leq \left\| \hat{C}_1 \frac{e_r - \eta_r \eta_s e_s}{\sqrt{2}} \right\| = \frac{\|\hat{c}_r - \eta_r \eta_s \hat{c}_s\|}{\sqrt{2}} \leq \sqrt{2} K \epsilon_n.$$

This contradicts (59) whenever $\sqrt{2} K \epsilon_n < b_*$. Since $\epsilon_n = o_p(1)$, the map $r \mapsto k_r$ is a permutation with probability approaching one.

Choose a permutation matrix Π_ℓ that reorders the estimated columns according to this matching, and choose diagonal sign matrices S_v, S_w that remove the corresponding signs. Define

$$\check{A}_1 = \hat{A}_1 \Pi_\ell S_v S_w, \quad \check{A}_2 = \hat{A}_2 \Pi_\ell S_v, \quad \check{A}_3 = \hat{A}_3 \Pi_\ell S_w.$$

The transformation preserves the fitted tensor and the column normalizations. Since ℓ is fixed, (61) and (60) give

$$\|\check{A}_2 - A_2\|_F + \|\check{A}_3 - A_3\|_F = O_p(\rho).$$

Let

$$\check{C}_1 = \check{A}_2 * \check{A}_3 = \hat{C}_1 \Pi_\ell S_v S_w.$$

The unit column norms imply

$$\|\check{C}_1 - C_1\|_F \leq \|\check{A}_2 - A_2\|_F + \|\check{A}_3 - A_3\|_F = O_p(\rho).$$

Moreover, $\Pi_\ell S_v S_w$ is orthogonal, so (59) yields

$$\|\check{C}_1(\check{C}_1' \check{C}_1)^{-1}\| = \frac{1}{\sigma_{\min}(\check{C}_1)} = \frac{1}{\sigma_{\min}(\hat{C}_1)} \leq b_*^{-1} = \frac{2\sqrt{c_U}}{c_0^{3/2}}$$

with probability approaching one.

Because the fitted tensor is unchanged, $\hat{\mathcal{P}}_{(1)} = \check{A}_1 \check{C}_1'$. Therefore,

$$h_{(1)} + A_1(C_1 - \check{C}_1)' = (\check{A}_1 - A_1)\check{C}_1',$$

and hence

$$\check{A}_1 - A_1 = [h_{(1)} + A_1(C_1 - \check{C}_1)'] \check{C}_1(\check{C}_1' \check{C}_1)^{-1}.$$

Taking Frobenius norms and using $\|A_1\| \leq \sqrt{c_1}$ gives

$$\|\check{A}_1 - A_1\|_F \leq b_*^{-1} \{\|h\|_F + \sqrt{c_1} \|\check{C}_1 - C_1\|_F\} = O_p(\rho).$$

Thus

$$\sum_{m=1}^3 \|\check{A}_m - A_m\|_F = O_p(\rho). \quad (62)$$

Finally, for every conformable matrix H ,

$$\|(D'D)^{1/2} H\|_F = \|DH\|_F.$$

Consequently, (62) is exactly the weighted factor-alignment rate in (8).

3. Eigenvalue consistency and inactivity of the upper bound. For the inactivity of the upper bound in the parameter constraints, we show

$$\mathbb{P} \left\{ \lambda_{\max} \left(\frac{1}{N} \hat{U}' D' D \hat{U} \right) < c_U \right\} \rightarrow 1. \quad (63)$$

Since $\Pi_\ell S_v S_w$ is orthogonal,

$$\lambda_{\max}(\hat{A}'_1 \hat{A}_1) = \lambda_{\max}(\check{A}'_1 \check{A}_1).$$

Moreover, feasibility gives $\|\check{A}_1\| = \|\hat{A}_1\| \leq \sqrt{c_U}$. On the strong-factor event,

$$\begin{aligned} \|\check{A}'_1 \check{A}_1 - A'_1 A_1\| &\leq (\|\check{A}_1\| + \|A_1\|) \|\check{A}_1 - A_1\| \\ &\leq (\sqrt{c_U} + \sqrt{c_1}) \|\check{A}_1 - A_1\| \\ &= O_p(\rho). \end{aligned}$$

The largest-eigenvalue perturbation inequality therefore gives

$$\left| \lambda_{\max}(\hat{A}'_1 \hat{A}_1) - \lambda_{\max}(A'_1 A_1) \right| \leq \|\check{A}'_1 \check{A}_1 - A'_1 A_1\| = O_p(\rho),$$

which proves

$$\left| \lambda_{\max} \left(\frac{1}{N} \hat{U}' D' D \hat{U} \right) - \lambda_{\max} \left(\frac{1}{N} U' D' D U \right) \right| = O_p(\rho)$$

Since $\lambda_{\max}(A'_1 A_1) \leq c_1$ with probability approaching one and $c_U - c_1 > 0$ is fixed,

$$\begin{aligned} &\mathbb{P} \left\{ \lambda_{\max}(\hat{A}'_1 \hat{A}_1) > \frac{c_1 + c_U}{2} \right\} \\ &\leq \mathbb{P}\{\lambda_{\max}(A'_1 A_1) > c_1\} + \mathbb{P} \left\{ \left| \lambda_{\max}(\hat{A}'_1 \hat{A}_1) - \lambda_{\max}(A'_1 A_1) \right| > \frac{c_U - c_1}{2} \right\} \rightarrow 0. \end{aligned}$$

This proves (63).

4. Validity of the unrestricted normal equations. Work at the original estimated factors on the strict-slack event in (63), and evaluate all gradients below at $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$. Since δ is unrestricted and sufficiently small perturbations of U remain feasible,

$$\nabla_\delta S_{NMT} = 0, \quad \nabla_U S_{NMT} = 0.$$

These are the slope and U -block equations in (9).

The equality constraints $\|v_k\|^2 = M$ and $\|w_k\|^2 = T$ remain in force. First-order optimality along their tangent directions implies that, for each k , there exist scalars α_k, β_k such that

$$\nabla_{v_k} S_{NMT} = \alpha_k \hat{v}_k, \quad \nabla_{w_k} S_{NMT} = \beta_k \hat{w}_k.$$

Direct differentiation of the trilinear component gives

$$\begin{aligned} \hat{u}'_k \nabla_{u_k} S_{NMT} &= \hat{v}'_k \nabla_{v_k} S_{NMT} = \hat{w}'_k \nabla_{w_k} S_{NMT} \\ &= -\frac{2}{n} \left\langle \hat{E}, D \hat{u}_k \circ \hat{v}_k \circ \hat{w}_k \right\rangle_F. \end{aligned}$$

Since $\nabla_{u_k} S_{NMT} = 0$, all three radial derivatives vanish. Hence

$$0 = \hat{v}'_k \nabla_{v_k} S_{NMT} = \alpha_k \|\hat{v}_k\|^2 = \alpha_k M,$$

and

$$0 = \hat{w}'_k \nabla_{w_k} S_{NMT} = \beta_k \|\hat{w}_k\|^2 = \beta_k T.$$

Because $M, T > 0$, we obtain $\alpha_k = \beta_k = 0$ for every k . Thus the unrestricted gradients with respect to V and W also vanish. Differentiating the criterion with respect to these blocks gives

$$\hat{E}_{(2)}(D\hat{U} * \hat{W}) = 0, \quad \hat{E}_{(3)}(D\hat{U} * \hat{V}) = 0.$$

Together with the slope and U -block equations, these establish (9) on an event whose probability tends to one. $\square$

Proposition 1. *Suppose that Assumptions 1, 2, 3, and 4 hold. Let $\mathcal{T}$ be the tangent space defined in (6) evaluated at the true parameters (U, V, W) . Let*

$$\Gamma_{NMT} = \frac{1}{NMT} X_{(4)} M_{\mathcal{T}} X'_{(4)}, \quad \mathcal{Z}_{NMT} = \frac{1}{\sqrt{NMT}} X_{(4)} M_{\mathcal{T}} \text{vec}(E).$$

The least squares estimator $\hat{\delta}$ defined in Assumption 1 satisfies

$$\Gamma_{NMT} \sqrt{NMT} (\hat{\delta} - \delta) = \mathcal{Z}_{NMT} - \mathcal{B}_1 - \mathcal{B}_2 - \mathcal{B}_3 + R_{NMT}, \quad (10)$$

where the remainder term R_{NMT} has the bound $O_p\left(\sqrt{\frac{G+M+T}{NMT}} + \frac{(G+M+T)^{3/2}}{NMT}\right)$ and the mode specific bias terms are given by

$$\begin{aligned} \mathcal{B}_1 &= \frac{1}{\sqrt{NMT}} X_{(4)} P_{\mathcal{T}} Q \text{vec}(\bar{U}(\bar{U}'\bar{U})^{-1}((V * W)'(V * W))^{-1}(V * W)' \Omega_1^G M_{V * W}) \\ \mathcal{B}_2 &= \frac{1}{\sqrt{NMT}} X_{(4)} P_{\mathcal{T}} Q K_{1,2}^G \text{vec}(V(V'V)^{-1}((\bar{U} * W)'(\bar{U} * W))^{-1}(\bar{U} * W)' \Omega_2^G M_{\bar{U} * W}) \\ \mathcal{B}_3 &= \frac{1}{\sqrt{NMT}} X_{(4)} P_{\mathcal{T}} Q K_{1,3}^G \text{vec}(W(W'W)^{-1}((\bar{U} * V)'(\bar{U} * V))^{-1}(\bar{U} * V)' \Omega_3^G M_{\bar{U} * V}). \end{aligned}$$

Here, $\bar{U} = (D'D)^{1/2}U$ and $\Omega_m^G = \mathbb{E}[(E_{(m)}^G)'E_{(m)}^G \mid \mathcal{F}]$ for $m = 1, 2, 3$. The matrix Q is defined as

$$Q_D = D(D'D)^{-1/2}, \quad Q = Q_D \otimes I_M \otimes I_T$$

and $P_{\mathcal{T}}$ is projection on (6). and $K_{1,m}^G$ denotes the mode-permutation matrix for tensors of dimensions $G \times M \times T$.²⁴ All factor-dependent matrices and projectors in the displayed bias terms are evaluated at the true parameters. Moreover,

$$\mathcal{B}_1 = O_p(G/\sqrt{NMT}), \quad \mathcal{B}_2 = O_p(M/\sqrt{NMT}), \quad \mathcal{B}_3 = O_p(T/\sqrt{NMT}).$$

Proof. Write

$$n = NMT, \quad s = G+M+T, \quad \rho = \sqrt{\frac{s}{n}}, \quad d_1 = G, \quad d_2 = M, \quad d_3 = T, \quad d_{\times} = GMT.$$

Assumption 1 implies $\rho \rightarrow 0$, while p and ℓ remain fixed. All ordinary inverses and unrestricted normal equations below are used on events whose probabilities tend to one.

For $v \in \mathbb{R}^{d_{\times}}$, let $v_{(m)}$ denote the mode- m matricization of its tensor reshape, with the remaining modes ordered increasingly, and let $\mathbf{v}_m$ denote the inverse operation. Thus $\mathbf{v}_m(v_{(m)}) = v$. Put $J_m = K_{m,1}^G$, so that

$$J_m v = \text{vec}(v_{(m)}), \quad \mathbf{v}_m(B) = J_m' \text{vec}(B).$$

²⁴For example, the matrix $K_{1,2}^G$ satisfies $(u \otimes v \otimes w) = K_{1,2}^G(v \otimes u \otimes w)$ for all u, v, w of dimension G, M , and T respectively. A general definition of the mode-permutation matrices is in Appendix 11.1.

Define

$$Q_D = D(D'DD)^{-1/2}, \quad Q = Q_D \otimes I_M \otimes I_T, \quad \mathbf{X} = \frac{X_{(4)}}{\sqrt{n}}, \quad \mathbf{X}_G = \mathbf{X}Q.$$

Then $Q'Q = I_{d_x}$. Let

$$e = \text{vec}(E), \quad E^G = E \times_1 Q'_D, \quad e_G = Q'e = \text{vec}(E^G),$$

and introduce the normalized factors

$$A_1 = \frac{(D'D)^{1/2}U}{\sqrt{N}}, \quad A_2 = \frac{V}{\sqrt{M}}, \quad A_3 = \frac{W}{\sqrt{T}}.$$

Writing $a_{m,k}$ for column k of A_m , define

$$p_0 = \sum_{k=1}^{\ell} a_{1,k} \otimes a_{2,k} \otimes a_{3,k}, \quad \text{vec}(\Pi) = \sqrt{n} Q p_0,$$

and

$$C_1 = A_2 * A_3, \quad C_2 = A_1 * A_3, \quad C_3 = A_1 * A_2, \quad Y_m = (p_0)_{(m)} = A_m C'_m.$$

Estimated quantities are defined analogously.

For each mode, let

$$\mathcal{T}_m^G = \{v \in \mathbb{R}^{d_x} : v_{(m)} = BC'_m, \quad B \in \mathbb{R}^{d_m \times \ell}\},$$

with orthogonal projector

$$P_m = J'_m(I_{d_m} \otimes P_{C_m})J_m.$$

Let P^G project onto

$$\mathcal{T}^G = \mathcal{T}_1^G + \mathcal{T}_2^G + \mathcal{T}_3^G,$$

and put

$$M^G = I_{d_x} - P^G, \quad \mathcal{T} = Q\mathcal{T}^G, \quad P_{\mathcal{T}} = QP^GQ', \quad M_{\mathcal{T}} = I_n - P_{\mathcal{T}}.$$

Use hats for the corresponding estimated spaces and projectors, and write

$$\Delta^G = \hat{P}^G - P^G.$$

Finally, define

$$\Gamma_n = \mathbf{X}M_{\mathcal{T}}\mathbf{X}' = \Gamma_{NMT}, \quad \mathcal{Z}_n = \mathbf{X}M_{\mathcal{T}}e = \mathcal{Z}_{NMT}.$$

The pseudoinverse of Y_m is

$$\Psi_m = Y_m^+ = C_m(C'_m C_m)^{-1}(A'_m A_m)^{-1}A'_m.$$

In these coordinates, the bias vectors in the statement can be written as

$$\mathcal{B}_m = \frac{1}{\sqrt{n}} \mathbf{X}_G P^G \mathbf{v}_m(\Psi'_m \Omega_m^G M_{C_m}), \quad \Omega_m^G = \mathbb{E}[(E_{(m)}^G)' E_{(m)}^G \mid \mathcal{F}],$$

where we've used $P_{\mathcal{T}}Q = QP^G$. Write

$$\gamma = \hat{\delta} - \delta, \quad \theta = \sqrt{n}\gamma, \quad h = \hat{p} - p_0,$$

and define the residual vectors

$$\hat{e} = e - X'_{(4)}\gamma - \sqrt{n}Qh, \quad \hat{e}_G = Q'\hat{e} = e_G - X'_G\theta - \sqrt{n}h.$$

Let $\hat{E}^G$ denote the grouped residual tensor satisfying $\text{vec}(\hat{E}^G) = \hat{e}_G$, so that

$$\hat{E}_{(m)}^G = (\hat{e}_G)_{(m)}.$$

The projector $M_{DU*V} \otimes M_W$ annihilates each true tangent block, and hence $M_{DU*V} \otimes M_W \leq M_{\mathcal{T}}$. Assumption 4 therefore gives, with probability approaching one,

$$\Gamma_n \geq \frac{1}{n}X_{(4)}(M_{DU*V} \otimes M_W)X'_{(4)} \geq c_\delta I_p, \quad \|\Gamma_n^{-1}\| \leq c_\delta^{-1}.$$

Assumption 2 also gives

$$\frac{\|X_{(4)}\|_F}{\sqrt{n}} + \frac{\|X_{(4)}Q\|_F}{\sqrt{n}} = O_p(1), \quad \|\Gamma_n\| = O_p(1).$$

Throughout the proof, conditional linear and nuclear-norm contractions follow from Lemma 4(i), (iii), while contractions involving error-dependent structured spaces follow from Lemma 5.

1. Alignment, tangent-projector control, and first-order approximations. Apply Lemma 1 and replace the estimated factors by their jointly aligned signed-permutation representatives, retaining the hat notation. This transformation leaves the fitted tensor, residuals, and estimated loading and tangent spaces unchanged. The lemma gives

$$\eta := \sum_{m=1}^3 \|\hat{A}_m - A_m\|_F = O_p(\rho). \quad (64)$$

Trilinearity and the strong-factor bounds therefore imply

$$\|h\| + \sum_{m=1}^3 \|\hat{C}_m - C_m\|_F = O_p(\rho).$$

The singular values of the true factors and Khatri–Rao matrices are bounded above and away from zero on the strong-factor event. For example,

$$\|C_1 z\|^2 = \|A_2 \text{diag}(z) A_3\|_F^2 \geq c_0^2 \|z\|^2.$$

Hence their estimated counterparts have full column rank with probability approaching one, and

$$\max_{m=1,2,3} \left\{ \|(\hat{A}'_m \hat{A}_m)^{-1}\| + \|(\hat{C}'_m \hat{C}_m)^{-1}\| \right\} = O_p(1).$$

The equal-rank projector identity gives

$$\begin{aligned} \|P_{\hat{C}_m} - P_{C_m}\| &\leq \frac{\|M_{\hat{C}_m} C_m\|_F}{\sigma_{\min}(C_m)} \\ &\leq \frac{\|\hat{C}_m - C_m\|_F}{\sigma_{\min}(C_m)} = O_p(\rho), \quad m = 1, 2, 3. \end{aligned} \quad (65)$$

Lemma 1 also supplies the unrestricted normal equations. In normalized grouped coordinates, they imply

$$X_{(4)}\hat{e} = 0, \quad \hat{E}_{(m)}^G \hat{C}_m = 0, \quad m = 1, 2, 3, \quad \hat{P}^G \hat{e}_G = 0. \quad (66)$$

The last identity holds because $\hat{e}_G$ is orthogonal to each estimated tangent block. The slope equation can be written as

$$\mathbf{X}\mathbf{X}'\gamma = \frac{1}{\sqrt{n}}\mathbf{X}e - \mathbf{X}_G h.$$

Since $\mathbf{X}\mathbf{X}' \geq \Gamma_n \geq c_\delta I_p$, $\mathbf{X}e = O_p(1)$ by (46), and $n^{-1/2} \leq \rho$, we obtain

$$\|\gamma\| + \|h\| = O_p(\rho). \quad (67)$$

We next control the full tangent projector. Let

$$S = P_1 + P_2 + P_3.$$

Its range is $\mathcal{T}^G$, so $P^G = SS^+$. To bound S^+ , complete each A_m to an invertible square matrix $B_m = [A_m, A_{m,\perp}]$, with orthonormal complement columns, and put $B = B_1 \otimes B_2 \otimes B_3$. On the strong-factor event,

$$\kappa(B) = \|B\|\|B^{-1}\| \leq (c_1/c_0)^{3/2}.$$

Under B^{-1} , the tangent blocks become coordinate subspaces supported on (i, k, k) , (k, j, k) , and (k, k, t) , respectively. Their only shared coordinates are (k, k, k) . Assigning these shared coordinates to the first block gives, for every $x \in \mathcal{T}^G$, an orthogonal coordinate decomposition

$$B^{-1}x = z_1 + z_2 + z_3, \quad x_m = Bz_m \in \mathcal{T}_m^G, \quad x = x_1 + x_2 + x_3.$$

Consequently,

$$\sum_{m=1}^3 \|x_m\|^2 \leq \|B\|^2 \sum_{m=1}^3 \|z_m\|^2 = \|B\|^2 \|B^{-1}x\|^2 \leq \kappa(B)^2 \|x\|^2.$$

Since

$$\|x\|^2 = \sum_{m=1}^3 \langle P_m x, x_m \rangle \leq (x' S x)^{1/2} \left(\sum_{m=1}^3 \|x_m\|^2 \right)^{1/2},$$

we obtain

$$\lambda_{\min}^+(S) \geq \kappa(B)^{-2}, \quad \|S^+\| \leq \kappa(B)^2 = O_p(1). \quad (68)$$

The same coordinate construction gives $\dim(\mathcal{T}^G) = \ell(s-2)$. By (64), the estimated factors have full column rank with probability approaching one, so the estimated tangent space has the same dimension. With $\hat{S} = \hat{P}_1 + \hat{P}_2 + \hat{P}_3$, equation (65) implies $\|\hat{S} - S\| = O_p(\rho)$. Therefore,

$$\begin{aligned} \|\Delta^G\| &= \|(I_{d_x} - \hat{P}^G)P^G\| \\ &= \|(I_{d_x} - \hat{P}^G)(S - \hat{S})S^+\| \\ &\leq \|S - \hat{S}\| \|S^+\| = O_p(\rho). \end{aligned}$$

The true and estimated tangent spaces are contained in the space $\mathcal{V}$ generated by $\mathcal{L}_m = \text{col}[A_m, \hat{A}_m]$, whose dimensions are at most 2ℓ . Hence $\Delta^G = \Delta^G P_{\mathcal{V}}$, and Lemma 5 gives

$$\|\Delta^G\| = O_p(\rho), \quad \|\Delta^G e_G\| \leq 2\sqrt{3}\ell \|\Delta^G\| \|E^G\| = O_p(\sqrt{s}\rho). \quad (69)$$

Combining the slope equation with the definition of $\hat{e}_G$ gives the exact residual-score identity

$$\Gamma_n \theta = \mathcal{Z}_n + \mathbf{X}_G P^G \hat{e}_G - \sqrt{n} \mathbf{X}_G M^G h. \quad (70)$$

To use this identity, write $\Delta A_m = \hat{A}_m - A_m$ and expand

$$h = q_1 + q_2 + q_3,$$

where

$$\begin{aligned} q_1 &= \sum_{k=1}^{\ell} [\Delta a_{1,k} \otimes a_{2,k} \otimes a_{3,k} + a_{1,k} \otimes \Delta a_{2,k} \otimes a_{3,k} + a_{1,k} \otimes a_{2,k} \otimes \Delta a_{3,k}], \\ q_2 &= \sum_{k=1}^{\ell} [\Delta a_{1,k} \otimes \Delta a_{2,k} \otimes a_{3,k} + \Delta a_{1,k} \otimes a_{2,k} \otimes \Delta a_{3,k} + a_{1,k} \otimes \Delta a_{2,k} \otimes \Delta a_{3,k}], \\ q_3 &= \sum_{k=1}^{\ell} \Delta a_{1,k} \otimes \Delta a_{2,k} \otimes \Delta a_{3,k}. \end{aligned}$$

Then $q_1 \in \mathcal{T}^G$, and

$$\|q_1\| = O_p(\rho), \quad \|q_2\| = O_p(\rho^2), \quad \|q_3\| = O_p(\rho^3), \quad \|M^G h\| = O_p(\rho^2).$$

Also, $q_1 + 2q_2 + 3q_3$ is the derivative at $t = 1$ of the tensor generated by $A_m + t\Delta A_m$, so it belongs to $\hat{\mathcal{T}}^G$. Direct subtraction consequently gives

$$2M^G h - \Delta^G h = -(I_{d_x} - \hat{P}^G)q_3 + \Delta^G(q_2 + q_3),$$

and therefore

$$M^G h = \frac{1}{2}\Delta^G h + O_p(\rho^3). \quad (71)$$

By (66), $P^G \hat{e}_G = -\Delta^G \hat{e}_G$. Equations (67) and (69) imply

$$\begin{aligned} \|\Delta^G \hat{e}_G\| &\leq \|\Delta^G e_G\| + \|\Delta^G\| \|\mathbf{X}_G\| \|\theta\| + \sqrt{n} \|\Delta^G\| \|h\| \\ &= O_p(\sqrt{n}\rho^2). \end{aligned}$$

Since $\mathcal{Z}_n = O_p(1)$ by Lemma 4 (i), the exact score equation yields

$$\Gamma_n \theta = \mathcal{Z}_n + O_p(s/\sqrt{n}).$$

Define

$$\theta_1 = \Gamma_n^{-1} \mathcal{Z}_n, \quad r_0 = e_G - \mathbf{X}'_G \theta_1, \quad \xi = P^G r_0, \quad q = \frac{\xi}{\sqrt{n}}.$$

Here, θ_1 , ξ , and $M^G r_0$ are the slope change, common-component change, and grouped residual from the model obtained by replacing the nonlinear factor component with its tangent-space approximation, i.e.,

$$\min_{\vartheta \in \mathbb{R}^p, z \in \mathcal{T}^G} \|e - \mathbf{X}'\vartheta - Qz\|^2.$$

Then

$$\theta_1 = O_p(1), \quad \theta - \theta_1 = O_p(s/\sqrt{n}), \quad \|q\| = O_p(\rho).$$

The last bound uses $\|P^G e_G\| = O_p(\sqrt{s})$, which follows from Lemma 5 with the true loading spaces. In particular,

$$\|\gamma\| = O_p(n^{-1/2} + s/n).$$

Let

$$F_0 = Q' - \mathbf{X}'_G \Gamma_n^{-1} \mathbf{X} M_{\mathcal{T}}, \quad r_0 = F_0 e, \quad \xi = P^G F_0 e. \quad (72)$$

The matrices F_0 , $P^G F_0$, and $S^+ P^G F_0$ are $\mathcal{F}$ -measurable and have $O_p(1)$ operator norms. The slope-adjusted linearized error admits the representation $r_0 = F_0 e$. Consequently, the linearized common-component change $q = n^{-1/2} P^G F_0 e$ and the first-order grouped fitted residual $M^G F_0 e$ are also linear in the original errors, conditional on $\mathcal{F}$. This representation is useful because products involving these objects can be written as quadratic forms in e , with $\mathcal{F}$ -measurable coefficient matrices.

The identity $\hat{P}^G \hat{e}_G = 0$ also gives

$$h - q = M^G h - \frac{1}{\sqrt{n}} P^G \mathbf{X}'_G (\theta - \theta_1) + \frac{1}{\sqrt{n}} \Delta^G \hat{e}_G.$$

Each term on the right is $O_p(\rho^2)$. Thus

$$\|h - q\| = O_p(\rho^2), \quad \|\hat{e}_G - M^G r_0\| = O_p(\sqrt{n} \rho^2), \quad (73)$$

where the second bound follows from

$$\hat{e}_G - M^G r_0 = -\mathbf{X}'_G (\theta - \theta_1) - \sqrt{n}(h - q).$$

2. Graph-aligned loading spaces and the exact tangent score. For each mode, define

$$J_m^C = (C'_m C_m)^{-1} C'_m \hat{C}_m, \quad F_m = \hat{C}_m (J_m^C)^{-1} - C_m.$$

The aligned rate gives

$$J_m^C = I_\ell + (C'_m C_m)^{-1} C'_m (\hat{C}_m - C_m) = I_\ell + O_p(\rho).$$

Hence J_m^C is nonsingular with probability approaching one. Since $C_m J_m^C = P_{C_m} \hat{C}_m$,

$$F_m = M_{C_m} (\hat{C}_m - C_m) (J_m^C)^{-1}.$$

It follows that

$$C'_m F_m = 0, \quad \text{col}(C_m + F_m) = \text{col}(\hat{C}_m), \quad \|F_m\|_F = O_p(\rho). \quad (74)$$

The loading normal equations therefore imply

$$\hat{E}_{(m)}^G C_m = -\hat{E}_{(m)}^G F_m. \quad (75)$$

Define

$$\tilde{A}_m = \hat{p}_{(m)} C_m (C'_m C_m)^{-1}.$$

Then

$$\hat{p}_{(m)} = \tilde{A}_m (C_m + F_m)', \quad \tilde{A}_m = A_m + h_{(m)} C_m (C'_m C_m)^{-1},$$

so $\|\tilde{A}_m - A_m\|_F = O_p(\rho)$. Since $C'_m F_m = 0$,

$$F_m = M_{C_m} h'_{(m)} \tilde{A}_m (\tilde{A}'_m \tilde{A}_m)^{-1}.$$

Let

$$F_m^{(1)} = M_{C_m} q'_{(m)} A_m (A'_m A_m)^{-1}.$$

The inverse perturbation identity, the strong-factor bounds, and (73) give

$$\|F_m - F_m^{(1)}\|_F = O_p(\rho^2). \quad (76)$$

Indeed, the difference is bounded by a constant in probability times

$$\|h - q\| + \|h\| \|\tilde{A}_m - A_m\|_F.$$

For slope coordinate $r = 1, \dots, p$, let $\mathbf{e}_r$ denote the corresponding coordinate vector and define

$$x_r = \mathbf{X}'_G \mathbf{e}_r, \quad t_r = S^+ x_r, \quad L_{m,r} = (t_r)_{(m)} C_m (C'_m C_m)^{-1}.$$

These matrices are $\mathcal{F}$ -measurable, and $\|L_{m,r}\|_F = O_p(1)$. Since

$$P^G x_r = \sum_{m=1}^3 P_m t_r, \quad (P_m t_r)_{(m)} = L_{m,r} C'_m,$$

equation (75) gives

$$x'_r P^G \hat{e}_G = - \sum_{m=1}^3 \langle \hat{E}_{(m)}^G, L_{m,r} F'_m \rangle_F.$$

Thus (70) becomes

$$[\Gamma_n \theta]_r = [\mathcal{Z}_n]_r - \sqrt{n} x'_r M^G h - \sum_{m=1}^3 \langle \hat{E}_{(m)}^G, L_{m,r} F'_m \rangle_F. \quad (77)$$

3. The normal common-component contraction. We establish the sharper scalar bound

$$\sqrt{n} x'_r M^G h = O_p \left(\rho + \frac{s^{3/2}}{n} \right). \quad (78)$$

The preliminary bound $\|M^G h\| = O_p(\rho^2)$ alone is insufficient for this conclusion. We begin by obtaining a first-order expansion of the sum of the tangent-block projectors, with an explicit second-order remainder.

Fix $m \in \{1, 2, 3\}$ and write

$$G_m = C'_m C_m.$$

The strong-factor bounds imply that $\|C_m\|$ and $\|G_m^{-1}\|$ are bounded on an event whose probability tends to one. More precisely, if $a < b$ are the modes other than m , then

$$c_0 \|z\| \leq \|C_m z\| = \|A_a \text{diag}(z) A'_b\|_F \leq c_1 \|z\|, \quad z \in \mathbb{R}^\ell.$$

Consequently,

$$\|C_m\| \leq c_1, \quad \lambda_{\min}(G_m) \geq c_0^2, \quad \|G_m^{-1}\| \leq c_0^{-2}.$$

All inverse identities below are evaluated on this event. By (74),

$$C'_m F_m = 0, \quad \text{col}(C_m + F_m) = \text{col}(\hat{C}_m), \quad \|F_m\| \leq \|F_m\|_F = O_p(\rho).$$

In particular,

$$(C_m + F_m)'(C_m + F_m) = G_m + F'_m F_m.$$

Since G_m is positive definite and $F'_m F_m$ is positive semidefinite, $C_m + F_m$ has full column rank. Its column-space projector is therefore

$$P_{\hat{C}_m} = (C_m + F_m)(G_m + F'_m F_m)^{-1}(C_m + F_m)'.$$

To expand the inverse, define

$$R_m^{\text{inv}} = (G_m + F'_m F_m)^{-1} - G_m^{-1}.$$

The exact inverse identity gives

$$R_m^{\text{inv}} = -(G_m + F'_m F_m)^{-1} F'_m F_m G_m^{-1}.$$

Moreover, $G_m + F'_m F_m \geq G_m$ implies

$$\|(G_m + F'_m F_m)^{-1}\| \leq \|G_m^{-1}\|.$$

Hence

$$\begin{aligned} \|R_m^{\text{inv}}\| &\leq \|(G_m + F'_m F_m)^{-1}\| \|F'_m F_m\| \|G_m^{-1}\| \\ &\leq \|G_m^{-1}\|^2 \|F_m\|^2 \\ &= O_p(\rho^2). \end{aligned}$$

Thus the inverse perturbation is second order because the graph orthogonality eliminates the linear cross terms from the Gram matrix.

Substituting $(G_m + F'_m F_m)^{-1} = G_m^{-1} + R_m^{\text{inv}}$ into the projector formula and using $P_{C_m} = C_m G_m^{-1} C'_m$ gives

$$P_{\hat{C}_m} - P_{C_m} = F_m G_m^{-1} C'_m + C_m G_m^{-1} F'_m + R_m^C,$$

where the remainder is exactly

$$R_m^C = F_m G_m^{-1} F'_m + (C_m + F_m) R_m^{\text{inv}} (C_m + F_m)'.$$

Its operator norm satisfies

$$\begin{aligned} \|R_m^C\| &\leq \|G_m^{-1}\| \|F_m\|^2 + \|C_m + F_m\|^2 \|R_m^{\text{inv}}\| \\ &\leq \{\|G_m^{-1}\| + \|C_m + F_m\|^2 \|G_m^{-1}\|^2\} \|F_m\|^2 \\ &= O_p(\rho^2). \end{aligned}$$

We have therefore obtained

$$P_{\hat{C}_m} - P_{C_m} = F_m (C'_m C_m)^{-1} C'_m + C_m (C'_m C_m)^{-1} F'_m + R_m^C, \quad \|R_m^C\| = O_p(\rho^2). \quad (79)$$

We next replace F_m by its first-order approximation. Recall that

$$F_m^{(1)} = M_{C_m} q'_{(m)} A_m (A'_m A_m)^{-1},$$

and, by (76),

$$R_m^F := F_m - F_m^{(1)}, \quad \|R_m^F\|_F = O_p(\rho^2).$$

Substituting $F_m = F_m^{(1)} + R_m^F$ into the two linear terms in (79) gives

$$P_{\hat{C}_m} - P_{C_m} = F_m^{(1)} G_m^{-1} C'_m + C_m G_m^{-1} (F_m^{(1)})' + \tilde{R}_m^C,$$

where

$$\tilde{R}_m^C = R_m^C + R_m^F G_m^{-1} C'_m + C_m G_m^{-1} (R_m^F)'.$$

Because

$$\|C_m G_m^{-1}\| = \frac{1}{\sigma_{\min}(C_m)},$$

the replacement error obeys

$$\begin{aligned} \|\tilde{R}_m^C\| &\leq \|R_m^C\| + 2\|R_m^F\| \|C_m G_m^{-1}\| \\ &\leq \|R_m^C\| + \frac{2\|R_m^F\|_F}{\sigma_{\min}(C_m)} \\ &= O_p(\rho^2). \end{aligned}$$

Thus replacing F_m by $F_m^{(1)}$ in the linear projector terms preserves the second-order remainder.

Recall the normalized common-component pseudoinverse

$$\Psi_m = C_m G_m^{-1} (A'_m A_m)^{-1} A'_m, \quad \Psi'_m = A_m (A'_m A_m)^{-1} G_m^{-1} C'_m.$$

The retained first-order terms can therefore be written as

$$\begin{aligned} F_m^{(1)} G_m^{-1} C'_m &= M_{C_m} q'_{(m)} A_m (A'_m A_m)^{-1} G_m^{-1} C'_m \\ &= M_{C_m} q'_{(m)} \Psi'_m, \end{aligned}$$

and

$$\begin{aligned} C_m G_m^{-1} (F_m^{(1)})' &= C_m G_m^{-1} (A'_m A_m)^{-1} A'_m q_{(m)} M_{C_m} \\ &= \Psi_m q_{(m)} M_{C_m}. \end{aligned}$$

For $v \in \mathbb{R}^{d \times}$, define the symmetric matrix

$$\mathcal{L}_m(v) = M_{C_m} v'_{(m)} \Psi'_m + \Psi_m v_{(m)} M_{C_m}.$$

The preceding calculations show that

$$P_{\hat{C}_m} - P_{C_m} = \mathcal{L}_m(q) + \tilde{R}_m^C, \quad \|\tilde{R}_m^C\| = O_p(\rho^2). \quad (80)$$

Define the associated matrix-valued linear maps

$$\mathcal{K}_m(v) = J'_m (I_{d_m} \otimes \mathcal{L}_m(v)) J_m, \quad \mathcal{K}(v) = \sum_{m=1}^3 \mathcal{K}_m(v).$$

For every v , the two summands defining $\mathcal{L}_m(v)$ are transposes of one another, and

$$\begin{aligned} \|\mathcal{L}_m(v)\| &\leq 2\|\Psi_m\| \|v_{(m)}\| \\ &\leq 2\|\Psi_m\| \|v_{(m)}\|_F \\ &= 2\|\Psi_m\| \|v\|. \end{aligned}$$

Since J_m is orthogonal and $\|I_{d_m} \otimes B\| = \|B\|$ for every matrix B ,

$$\|\mathcal{K}_m(v)\| = \|\mathcal{L}_m(v)\|.$$

It follows that

$$\|\mathcal{K}(v)\| \leq \sum_{m=1}^3 \|\mathcal{K}_m(v)\| \leq 2 \left(\sum_{m=1}^3 \|\Psi_m\| \right) \|v\|.$$

The strong-factor bounds also give

$$\|\Psi_m\| \leq \frac{1}{\sigma_{\min}(C_m)\sigma_{\min}(A_m)} \leq c_0^{-3/2}$$

on the strong-factor event. Hence $\sum_m \|\Psi_m\| = O_p(1)$, and in particular $\|\mathcal{K}(q)\| = O_p(\rho)$.

Finally, recall that

$$P_m = J'_m(I_{d_m} \otimes P_{C_m})J_m, \quad \hat{P}_m = J'_m(I_{d_m} \otimes P_{\hat{C}_m})J_m,$$

and

$$S = \sum_{m=1}^3 P_m, \quad \hat{S} = \sum_{m=1}^3 \hat{P}_m.$$

Equation (80) gives the exact remainder representation

$$\hat{S} - S - \mathcal{K}(q) = \sum_{m=1}^3 J'_m(I_{d_m} \otimes \tilde{R}_m^C)J_m.$$

Consequently,

$$\begin{aligned} \|\hat{S} - S - \mathcal{K}(q)\| &\leq \sum_{m=1}^3 \|J'_m(I_{d_m} \otimes \tilde{R}_m^C)J_m\| \\ &= \sum_{m=1}^3 \|\tilde{R}_m^C\| \\ &= O_p(\rho^2). \end{aligned}$$

Thus

$$\hat{S} - S = \mathcal{K}(q) + O_p(\rho^2)$$

in operator norm. No dimension factor is introduced when passing from the loading-space projectors to the tangent-block projectors, because multiplication by the orthogonal mode-permutation matrices and taking a Kronecker product with an identity matrix preserve the operator norm.

The identities $\hat{S}(I_{d_x} - \hat{P}^G) = 0$ and $(P^G + \Delta^G)^2 = P^G + \Delta^G$ imply

$$\Delta^G = M^G(\hat{S} - S)S^+ + S^+(\hat{S} - S)M^G + R_S,$$

where

$$R_S = (\Delta^G)^2 - S^+(\hat{S} - S)\Delta^G - \Delta^G(\hat{S} - S)S^+.$$

Equations (68) and (69) imply $\|R_S\| = O_p(\rho^2)$. Therefore,

$$\Delta^G = M^G\mathcal{K}(q)S^+ + S^+\mathcal{K}(q)M^G + R_\Delta^G, \quad \|R_\Delta^G\| = O_p(\rho^2).$$

Using (71), $M^G q = 0$, and (73), we obtain

$$\begin{aligned} M^G h &= \frac{1}{2} \Delta^G q + \frac{1}{2} \Delta^G (h - q) + O_p(\rho^3) \\ &= \frac{1}{2} M^G \mathcal{K}(q) S^+ q + O_p(\rho^3). \end{aligned}$$

For fixed m , put

$$Z_m = (M^G x_r)_{(m)}, \quad T_m = (S^+ q)_{(m)}.$$

Orthogonality to the m -th tangent block gives

$$Z_m C_m = 0, \quad Z_m M_{C_m} = Z_m, \quad Z_m \Psi_m = 0.$$

Since $\mathcal{L}_m(q)$ is symmetric, row-major vectorization gives

$$(\mathcal{K}_m(q) S^+ q)_{(m)} = T_m \mathcal{L}_m(q).$$

Consequently,

$$\begin{aligned} \langle M^G x_r, \mathcal{K}_m(q) S^+ q \rangle &= \langle Z_m, T_m M_{C_m} q'_{(m)} \Psi'_m \rangle_F + \langle Z_m, T_m \Psi_m q_{(m)} M_{C_m} \rangle_F \\ &= \langle Z_m \Psi_m, T_m M_{C_m} q'_{(m)} \rangle_F + \langle Z_m M_{C_m}, T_m \Psi_m q_{(m)} \rangle_F \\ &= \langle Z_m, T_m \Psi_m q_{(m)} \rangle_F. \end{aligned}$$

With $\xi = \sqrt{n} q$, this proves

$$\sqrt{n} x'_r M^G h = \frac{1}{2\sqrt{n}} \sum_{m=1}^3 \langle (M^G x_r)_{(m)}, (S^+ \xi)_{(m)} \Psi_m \xi_{(m)} \rangle_F + O_p(\sqrt{n} \rho^3). \quad (81)$$

Fix m and write

$$d_{-m} = \frac{d_\times}{d_m}, \quad Z_m = (M^G x_r)_{(m)} \in \mathbb{R}^{d_m \times d_{-m}}.$$

In addition to the mode-permutation matrix J_m , define the permutation matrix $\tilde{J}_m \in \mathbb{R}^{d_\times \times d_\times}$ by

$$J_m v = \text{vec}(v_{(m)}), \quad \tilde{J}_m v = \text{vec}(v'_{(m)}), \quad v \in \mathbb{R}^{d_\times}.$$

Thus J_m vectorizes the mode- m matricization, whereas $\tilde{J}_m$ vectorizes its transpose. Both matrices are orthogonal:

$$J'_m J_m = \tilde{J}'_m \tilde{J}_m = I_{d_\times}, \quad \|J_m\| = \|\tilde{J}_m\| = 1.$$

Since $\Psi_m \in \mathbb{R}^{d_{-m} \times d_m}$, define

$$\mathcal{Q}_{m,r} = Z_m \otimes \Psi_m \in \mathbb{R}^{d_\times \times d_\times}.$$

The Frobenius inner-product identity $\langle A, BC \rangle_F = \langle AC', B \rangle_F$ and the row-major vectorization identity $\text{vec}(ABC) = (A \otimes C') \text{vec}(B)$ give

$$\begin{aligned}
\langle Z_m, (S^+ \xi)_{(m)} \Psi_m \xi_{(m)} \rangle_F &= \langle Z_m \xi'_{(m)}, (S^+ \xi)_{(m)} \Psi_m \rangle_F \\
&= \text{vec}(Z_m \xi'_{(m)})' \text{vec}((S^+ \xi)_{(m)} \Psi_m) \\
&= \text{vec}(\xi'_{(m)})' (Z'_m \otimes I_{d_m}) (I_{d_m} \otimes \Psi'_m) \text{vec}((S^+ \xi)_{(m)}) \\
&= \text{vec}(\xi'_{(m)})' (Z'_m \otimes \Psi'_m) \text{vec}((S^+ \xi)_{(m)}) \\
&= \text{vec}((S^+ \xi)_{(m)})' \mathcal{Q}_{m,r} \text{vec}(\xi'_{(m)}) \\
&= (J_m S^+ \xi)' \mathcal{Q}_{m,r} (\tilde{J}_m \xi).
\end{aligned}$$

The fourth equality uses the mixed-product identity for Kronecker products, and the fifth follows by transposing the scalar. In particular, $\mathcal{Q}_{m,r}$ is an exact Kronecker product, with the transposed matricization on the right represented by $\tilde{J}_m$.

We next bound the nuclear norm of $\mathcal{Q}_{m,r}$. Since M^G is an orthogonal projector,

$$\|Z_m\|_F = \|M^G x_r\| \leq \|x_r\| = O_p(1).$$

The strong-factor bounds imply $\|\Psi_m\| = O_p(1)$ and $\text{rank}(\Psi_m) = \ell$ with probability approaching one. Since ℓ is fixed,

$$\|\Psi_m\|_F \leq \sqrt{\ell} \|\Psi_m\| = O_p(1).$$

The Frobenius norm and rank identities for Kronecker products therefore give

$$\|\mathcal{Q}_{m,r}\|_F = \|Z_m\|_F \|\Psi_m\|_F = O_p(1)$$

and

$$\text{rank}(\mathcal{Q}_{m,r}) = \text{rank}(Z_m) \text{rank}(\Psi_m) \leq \ell d_m.$$

Applying Cauchy-Schwarz to the nonzero singular values yields

$$\|\mathcal{Q}_{m,r}\|_* \leq \sqrt{\text{rank}(\mathcal{Q}_{m,r})} \|\mathcal{Q}_{m,r}\|_F \leq \sqrt{\ell d_m} \|\mathcal{Q}_{m,r}\|_F = O_p(\sqrt{d_m}).$$

By (72), $\xi = P^G F_0 e$, so

$$\text{vec}((S^+ \xi)_{(m)}) = J_m S^+ P^G F_0 e, \quad \text{vec}(\xi'_{(m)}) = \tilde{J}_m P^G F_0 e.$$

Consequently,

$$\langle Z_m, (S^+ \xi)_{(m)} \Psi_m \xi_{(m)} \rangle_F = e' \mathcal{H}_{m,r} e,$$

where

$$\mathcal{H}_{m,r} = (J_m S^+ P^G F_0)' \mathcal{Q}_{m,r} (\tilde{J}_m P^G F_0) \in \mathbb{R}^{n \times n}.$$

All matrices defining $\mathcal{H}_{m,r}$ are $\mathcal{F}$ -measurable. Moreover, J_m and $\tilde{J}_m$ are orthogonal, while $S^+ P^G F_0$ and $P^G F_0$ have $O_p(1)$ operator norms. The nuclear-norm inequality $\|ABC\|_* \leq \|A\| \|B\|_* \|C\|$ therefore gives

$$\begin{aligned}
\|\mathcal{H}_{m,r}\|_* &\leq \|J_m S^+ P^G F_0\| \|\mathcal{Q}_{m,r}\|_* \|\tilde{J}_m P^G F_0\| \\
&= \|S^+ P^G F_0\| \|\mathcal{Q}_{m,r}\|_* \|P^G F_0\| \\
&= O_p(\sqrt{d_m}).
\end{aligned}$$

On the conditional covariance-summability event, Lemma 4 (iii) implies

$$\mathbb{E}[|e' \mathcal{H}_{m,r} e| \mid \mathcal{F}] \leq c_\varepsilon \|\mathcal{H}_{m,r}\|_*.$$

Conditional Markov's inequality, localized to events on which $\|\mathcal{H}_{m,r}\|_*/\sqrt{d_m}$ is bounded, yields

$$\langle Z_m, (S^+ \xi)_{(m)} \Psi_m \xi_{(m)} \rangle_F = e' \mathcal{H}_{m,r} e = O_p(\sqrt{d_m}).$$

Substituting these bounds into (81) gives

$$\begin{aligned} \sqrt{n} x'_r M^G h &= \frac{1}{2\sqrt{n}} \sum_{m=1}^3 e' \mathcal{H}_{m,r} e + O_p(\sqrt{n} \rho^3) \\ &= O_p\left(\frac{\sqrt{d_1} + \sqrt{d_2} + \sqrt{d_3}}{\sqrt{n}} + \sqrt{n} \rho^3\right). \end{aligned}$$

Finally,

$$\sqrt{d_1} + \sqrt{d_2} + \sqrt{d_3} \leq \sqrt{3(d_1 + d_2 + d_3)} = \sqrt{3s}, \quad \sqrt{n} \rho^3 = \frac{s^{3/2}}{n}.$$

Therefore,

$$\sqrt{n} x'_r M^G h = O_p\left(\rho + \frac{s^{3/2}}{n}\right),$$

which proves (78).

4. First-order expansion of the graph score. Define

$$D_{m,r} = L_{m,r} (A'_m A_m)^{-1} A'_m = (t_r)_{(m)} \Psi_m.$$

These matrices are $\mathcal{F}$ -measurable and have $O_p(1)$ operator norms. By (73) and (74),

$$\begin{aligned} \left| \langle \hat{E}_{(m)}^G - (M^G r_0)_{(m)}, L_{m,r} F'_m \rangle_F \right| &\leq \|\hat{e}_G - M^G r_0\| \|L_{m,r}\| \|F_m\|_F \\ &= O_p(\sqrt{n} \rho^3). \end{aligned}$$

Next, let

$$z_m = \mathbf{v}_m \left(L_{m,r} (F_m - F_m^{(1)})' \right).$$

Then $\|z_m\| = O_p(\rho^2)$ due to (76) and $\|L_{m,r}\| = O_p(1)$. To obtain an explicit tangent-block decomposition of q , define

$$B_j^q = (S^+ q)_{(j)} C_j (C'_j C_j)^{-1} \in \mathbb{R}^{d_j \times \ell}, \quad j = 1, 2, 3.$$

The inverses are well defined on the strong-factor event. Multiplying by C'_j gives

$$B_j^q C'_j = (S^+ q)_{(j)} C_j (C'_j C_j)^{-1} C'_j = (S^+ q)_{(j)} P_{C_j}.$$

This matrix is the mode- j matricization of the projection of $S^+ q$ onto $\mathcal{T}_j^G$. Indeed, recall that

$$P_j = J'_j (I_{d_j} \otimes P_{C_j}) J_j, \quad J_j v = \text{vec}(v_{(j)}).$$

For any $v \in \mathbb{R}^{d_\times}$, the row-major vectorization identity and the symmetry of P_{C_j} imply

$$\begin{aligned} J_j P_j v &= (I_{d_j} \otimes P_{C_j}) J_j v \\ &= (I_{d_j} \otimes P_{C_j}) \text{vec}(v_{(j)}) \\ &= \text{vec}(v_{(j)} P_{C_j}). \end{aligned}$$

Consequently,

$$(P_j v)_{(j)} = v_{(j)} P_{C_j}.$$

Applying this identity to $v = S^+ q$ yields

$$B_j^q C_j' = (P_j S^+ q)_{(j)},$$

and hence, because $\mathbf{v}_j$ reverses mode- j matricization,

$$\mathbf{v}_j(B_j^q C_j') = P_j S^+ q \in \mathcal{T}_j^G.$$

Since $S = P_1 + P_2 + P_3$ has range $\mathcal{T}^G$, its Moore–Penrose pseudoinverse satisfies $SS^+ = P^G$. Moreover, $q \in \mathcal{T}^G$, so $P^G q = q$. Summing the preceding identities therefore gives

$$\begin{aligned} \sum_{j=1}^3 \mathbf{v}_j(B_j^q C_j') &= \sum_{j=1}^3 P_j S^+ q \\ &= (P_1 + P_2 + P_3) S^+ q \\ &= SS^+ q \\ &= P^G q = q. \end{aligned}$$

Thus

$$q = \sum_{j=1}^3 \mathbf{v}_j(B_j^q C_j').$$

This is an exact tangent-block decomposition, but its three summands need not be mutually orthogonal. The use of $S^+ q$, rather than q , accounts for the overlap among the tangent blocks: in general, $\sum_j P_j q = Sq \neq q$, whereas $\sum_j P_j S^+ q = q$. Let

$$\mathcal{L}_j = \text{col}[A_j, \hat{A}_j, B_j^q], \quad r_j = \dim(\mathcal{L}_j) \leq 3\ell, \quad j = 1, 2, 3.$$

The dimension bound holds because each of the three matrices has ℓ columns. These spaces contain the true loading directions, the estimated loading directions, and the directions appearing in the tangent decomposition of q .

Fix m , let $a < b$ be the other two modes, and define

$$\mathcal{U}_{-m} = \mathcal{L}_a \otimes \mathcal{L}_b \subseteq \mathbb{R}^{d_a d_b}.$$

We first verify that the columns of both F_m and $F_m^{(1)}$ belong to $\mathcal{U}_{-m}$. Recall that

$$C_m = A_a * A_b, \quad \hat{C}_m = \hat{A}_a * \hat{A}_b.$$

Every column of C_m is a tensor product of a column of A_a and a column of A_b , and the same statement holds for the estimated matrices. Since $\text{col}(A_j), \text{col}(\hat{A}_j) \subseteq \mathcal{L}_j$,

$$\text{col}(C_m) \subseteq \mathcal{U}_{-m}, \quad \text{col}(\hat{C}_m) \subseteq \mathcal{U}_{-m}.$$

The graph representation

$$F_m = \hat{C}_m (J_m^C)^{-1} - C_m$$

then implies

$$\text{col}(F_m) \subseteq \text{col}(\hat{C}_m) + \text{col}(C_m) \subseteq \mathcal{U}_{-m},$$

because right multiplication only forms linear combinations of columns.

Next, the tangent decomposition

$$q = \sum_{j=1}^3 \mathbf{v}_j (B_j^q C_j')$$

has the following mode- m representation:

$$q_{(m)} = B_m^q C_m' + A_m (B_a^q * A_b + A_a * B_b^q)'$$

The first term is the contribution from the m -th tangent block, while the two terms inside the parentheses arise from the other two tangent blocks. Transposing gives

$$q'_{(m)} = C_m (B_m^q)' + (B_a^q * A_b + A_a * B_b^q) A_m'.$$

Since $\text{col}(B_j^q) \subseteq \mathcal{L}_j$, each of C_m , $B_a^q * A_b$, and $A_a * B_b^q$ has columns in $\mathcal{U}_{-m}$. Consequently,

$$\text{col}(q'_{(m)}) \subseteq \text{col}[C_m, B_a^q * A_b, A_a * B_b^q] \subseteq \mathcal{U}_{-m}.$$

The space $\mathcal{U}_{-m}$ is also preserved by M_{C_m} . Indeed, for any $u \in \mathcal{U}_{-m}$,

$$P_{C_m} u \in \text{col}(C_m) \subseteq \mathcal{U}_{-m},$$

and therefore

$$M_{C_m} u = u - P_{C_m} u \in \mathcal{U}_{-m}.$$

Using

$$F_m^{(1)} = M_{C_m} q'_{(m)} A_m (A_m' A_m)^{-1},$$

we conclude that

$$\text{col}(F_m^{(1)}) \subseteq \mathcal{U}_{-m}.$$

Thus

$$\text{col}(F_m - F_m^{(1)}) \subseteq \mathcal{U}_{-m}.$$

Recall that

$$z_m = \mathbf{v}_m \left(L_{m,r} (F_m - F_m^{(1)})' \right).$$

Its transposed mode- m matricization satisfies

$$(z_m)'_{(m)} = (F_m - F_m^{(1)}) L'_{m,r},$$

so

$$\text{col}((z_m)'_{(m)}) \subseteq \mathcal{U}_{-m}.$$

Equivalently, the mode- m direction of z_m is unrestricted, while its other two modes are restricted to $\mathcal{L}_a$ and $\mathcal{L}_b$. In the original tensor ordering, this gives

$$\begin{aligned} z_1 &\in \mathbb{R}^{d_1} \otimes \mathcal{L}_2 \otimes \mathcal{L}_3, \\ z_2 &\in \mathcal{L}_1 \otimes \mathbb{R}^{d_2} \otimes \mathcal{L}_3, \\ z_3 &\in \mathcal{L}_1 \otimes \mathcal{L}_2 \otimes \mathbb{R}^{d_3}. \end{aligned}$$

Hence every z_m belongs to the structured space

$$\mathcal{V} = (\mathbb{R}^{d_1} \otimes \mathcal{L}_2 \otimes \mathcal{L}_3) + (\mathcal{L}_1 \otimes \mathbb{R}^{d_2} \otimes \mathcal{L}_3) + (\mathcal{L}_1 \otimes \mathcal{L}_2 \otimes \mathbb{R}^{d_3}).$$

Because $\text{col}(A_j) \subseteq \mathcal{L}_j$ for every j , each true tangent block is contained in the corresponding summand of $\mathcal{V}$. Therefore,

$$\mathcal{T}^G \subseteq \mathcal{V}.$$

In particular, $P^G z_m \in \mathcal{T}^G \subseteq \mathcal{V}$. Since $z_m \in \mathcal{V}$ and $\mathcal{V}$ is a linear space,

$$M^G z_m = z_m - P^G z_m \in \mathcal{V}.$$

Let $P_{\mathcal{V}}$ denote the orthogonal projector onto $\mathcal{V}$. Lemma 5 with $r_j = \dim(\mathcal{L}_j) \leq 3\ell$ gives the deterministic bound

$$\begin{aligned} \|P_{\mathcal{V}} e_G\| &\leq \sqrt{r_2 r_3 + r_1 r_3 + r_1 r_2} \|E^G\| \\ &\leq 3\sqrt{3}\ell \|E^G\|. \end{aligned}$$

Here $\|E^G\|$ is the tensor spectral norm. Since $M^G z_m \in \mathcal{V}$,

$$\langle e_G, M^G z_m \rangle = \langle P_{\mathcal{V}} e_G, M^G z_m \rangle.$$

Using $r_0 = e_G - \mathbf{X}'_G \theta_1$, the symmetry of M^G , and Cauchy–Schwarz, we therefore obtain

$$\begin{aligned} |\langle M^G r_0, z_m \rangle| &= |\langle r_0, M^G z_m \rangle| \\ &\leq |\langle e_G, M^G z_m \rangle| + |\langle \mathbf{X}'_G \theta_1, M^G z_m \rangle| \\ &= |\langle P_{\mathcal{V}} e_G, M^G z_m \rangle| + |\langle \mathbf{X}'_G \theta_1, M^G z_m \rangle| \\ &\leq (\|P_{\mathcal{V}} e_G\| + \|\mathbf{X}_G\| \|\theta_1\|) \|M^G z_m\| \\ &\leq \left(3\sqrt{3}\ell \|E^G\| + \|\mathbf{X}_G\| \|\theta_1\| \right) \|z_m\|. \end{aligned}$$

Since ℓ is fixed and

$$\|E^G\| = O_p(\sqrt{s}), \quad \|\mathbf{X}_G\| = O_p(1), \quad \|\theta_1\| = O_p(1), \quad \|z_m\| = O_p(\rho^2),$$

it follows that

$$|\langle M^G r_0, z_m \rangle| = O_p(\sqrt{s} \rho^2) = O_p(\sqrt{n} \rho^3).$$

The subspace inclusions and the structured contraction inequality hold for every realized choice of the factors, graph errors, and mode spaces. Consequently, the argument remains valid when F_m , q , z_m , and $\mathcal{L}_j$ depend on the errors. It does not require these objects to be $\mathcal{F}$ -measurable, and no conditional independence between them and e_G is used.

Finally, $(M^G r_0)_{(m)} C_m = 0$ and $q = \xi/\sqrt{n}$ imply

$$\langle (M^G r_0)_{(m)}, L_{m,r}(F_m^{(1)})' \rangle_F = \frac{1}{\sqrt{n}} \langle (M^G r_0)_{(m)}, D_{m,r} \xi_{(m)} \rangle_F.$$

Substitution into (77), together with (78), gives

$$[\Gamma_n \theta]_r = [\mathcal{Z}_n]_r - \frac{1}{\sqrt{n}} \sum_{m=1}^3 \langle (M^G r_0)_{(m)}, D_{m,r}(P^G r_0)_{(m)} \rangle_F + O_p\left(\rho + \frac{s^{3/2}}{n}\right). \quad (82)$$

5. Reduction to error Gram matrices. Note that we have

$$\begin{aligned} \langle (M^G r_0)_{(m)}, D_{m,r}(P^G r_0)_{(m)} \rangle_F &= \text{vec}((M^G r_0)_{(m)})' \text{vec}(D_{m,r}(P^G r_0)_{(m)}) \\ &= \text{vec}((M^G r_0)_{(m)})' (D_{m,r} \otimes I_{d-m}) \text{vec}((P^G r_0)_{(m)}) \\ &= r_0' M^G J_m' (D_{m,r} \otimes I_{d-m}) J_m P^G r_0. \end{aligned}$$

Let

$$\Lambda_{m,r} = J_m' (D_{m,r} \otimes I_{d-m}) J_m, \quad \Lambda_r = \sum_{m=1}^3 \Lambda_{m,r}.$$

Then $\|\Lambda_r\| = O_p(1)$, and the quadratic term in (82) is

$$\frac{1}{\sqrt{n}} r_0' M^G \Lambda_r P^G r_0.$$

We first remove the slope adjustment. By (72), $F_0 - Q' = -X_G' \Gamma_n^{-1} X_M \mathcal{T}$ has rank at most p and operator norm $O_p(1)$. Therefore,

$$F_0' M^G \Lambda_r P^G F_0 - Q M^G \Lambda_r P^G Q'$$

has rank at most $3p$ and operator norm $O_p(1)$, and hence has nuclear norm $O_p(1)$. Lemma 4 (iii) gives

$$\frac{1}{\sqrt{n}} r_0' M^G \Lambda_r P^G r_0 = \frac{1}{\sqrt{n}} e_G' M^G \Lambda_r P^G e_G + O_p(n^{-1/2}). \quad (83)$$

Decompose the tangent space into its component inside the product loading space and its three orthogonal exterior blocks:

$$P^G = P_0 + \sum_{h=1}^3 P_h^\circ, \quad P_h^\circ = J_h' (M_{A_h} \otimes P_{C_h}) J_h.$$

Here P_0 projects onto

$$\mathcal{T}^G \cap \{\text{col}(A_1) \otimes \text{col}(A_2) \otimes \text{col}(A_3)\},$$

so $\text{rank}(P_0) \leq \ell^3$. This decomposition follows by splitting each unrestricted loading perturbation into its projection onto the corresponding loading space and its orthogonal complement. The exterior blocks are mutually orthogonal and orthogonal to the product loading space.

Fix a mode h , and let $a < b$ be the other two modes. Define

$$D_{h,r}^{(-)} = D_{a,r} \otimes I_{d_b} + I_{d_a} \otimes D_{b,r}, \quad B_{h,r} = M_{C_h} D_{h,r}^{(-)} P_{C_h}.$$

For every matrix Y satisfying $A'_h Y = 0$,

$$(M^G \mathbf{v}_h(Y))_{(h)} = Y M_{C_h}.$$

Indeed, $Y P_{C_h}$ belongs to the h -th tangent block, while $Y M_{C_h}$ is orthogonal to that block. It is also orthogonal to the other tangent blocks, whose mode- h matricizations have columns in $\text{col}(A_h)$.

For an arbitrary $v \in \mathbb{R}^{d \times}$, define, for each $h = 1, 2, 3$,

$$V_h = v_{(h)}, \quad z_h^\circ = P_h^\circ v, \quad Z_h^\circ = (z_h^\circ)_{(h)} = M_{A_h} V_h P_{C_h}.$$

Thus V_h is the mode- h matricization of the common input v , whereas Z_h° is the mode- h matricization of its h -th exterior tangent component. In particular,

$$A'_h Z_h^\circ = 0, \quad Z_h^\circ P_{C_h} = Z_h^\circ.$$

Fix h , and let $a < b$ be the other two modes. The transformation in mode h satisfies

$$\begin{aligned} (\Lambda_{h,r} z_h^\circ)_{(h)} &= D_{h,r} Z_h^\circ \\ &= D_{h,r} Z_h^\circ P_{C_h} \\ &= [D_{h,r} Z_h^\circ C_h (C'_h C_h)^{-1}] C'_h. \end{aligned}$$

Hence $\Lambda_{h,r} z_h^\circ \in \mathcal{T}_h^G$, and therefore

$$M^G \Lambda_{h,r} z_h^\circ = 0.$$

The transformations in the other two modes act on the same input component z_h° and satisfy

$$(\Lambda_{a,r} z_h^\circ)_{(h)} = Z_h^\circ (D_{a,r} \otimes I_{d_b})', \quad (\Lambda_{b,r} z_h^\circ)_{(h)} = Z_h^\circ (I_{d_a} \otimes D_{b,r})'.$$

Consequently, we have

$$((\Lambda_{a,r} + \Lambda_{b,r}) z_h^\circ)_{(h)} = Z_h^\circ (D_{h,r}^{(-)})'.$$

Since

$$A'_h Z_h^\circ (D_{h,r}^{(-)})' = 0,$$

the previously established residual-projection identity applies to $Y = Z_h^\circ (D_{h,r}^{(-)})'$. Using $\Lambda_r = \Lambda_{h,r} + \Lambda_{a,r} + \Lambda_{b,r}$, we obtain

$$\begin{aligned} (M^G \Lambda_r P_h^\circ v)_{(h)} &= (M^G \Lambda_r z_h^\circ)_{(h)} \\ &= (M^G (\Lambda_{a,r} + \Lambda_{b,r}) z_h^\circ)_{(h)} \\ &= Z_h^\circ (D_{h,r}^{(-)})' M_{C_h} \\ &= M_{A_h} V_h P_{C_h} (D_{h,r}^{(-)})' M_{C_h} \\ &= M_{A_h} V_h B'_{h,r}. \end{aligned}$$

By row-major vectorization,

$$\begin{aligned} J_h M^G \Lambda_r P_h^\circ v &= \text{vec}((M^G \Lambda_r P_h^\circ v)_{(h)}) \\ &= \text{vec}(M_{A_h} V_h B'_{h,r}) \\ &= (M_{A_h} \otimes B_{h,r}) \text{vec}(V_h) \\ &= (M_{A_h} \otimes B_{h,r}) J_h v. \end{aligned}$$

Since v was arbitrary and J_h is orthogonal, this proves

$$M^G \Lambda_r P_h^\circ = J_h'(M_{A_h} \otimes B_{h,r}) J_h, \quad h = 1, 2, 3.$$

Finally, the decomposition $P^G = P_0 + \sum_{h=1}^3 P_h^\circ$ gives

$$\begin{aligned} M^G \Lambda_r P^G &= M^G \Lambda_r P_0 + \sum_{h=1}^3 M^G \Lambda_r P_h^\circ \\ &= M^G \Lambda_r P_0 + \sum_{h=1}^3 J_h'(M_{A_h} \otimes B_{h,r}) J_h. \end{aligned}$$

Each $B_{h,r}$ has rank at most ℓ , and

$$\|B_{h,r}\| \leq \|D_{h,r}^{(-)}\| \leq \|D_{a,r}\| + \|D_{b,r}\| = O_p(1).$$

Replacing M_{A_h} by $I_{d_h} - P_{A_h}$ gives

$$M^G \Lambda_r P^G = \sum_{h=1}^3 J_h'(I_{d_h} \otimes B_{h,r}) J_h + \mathcal{R}_r,$$

where

$$\mathcal{R}_r = M^G \Lambda_r P_0 - \sum_{h=1}^3 J_h'(P_{A_h} \otimes B_{h,r}) J_h.$$

Since $\text{rank}(P_{A_h}) = \ell$,

$$\text{rank}(\mathcal{R}_r) \leq \ell^3 + 3\ell^2, \quad \|\mathcal{R}_r\| \leq \|\Lambda_r\| + \sum_{h=1}^3 \|B_{h,r}\| = O_p(1).$$

Thus $\|\mathcal{R}_r\|_* = O_p(1)$. All these coefficients are $\mathcal{F}$ -measurable, so Lemma 4 (iii) implies

$$e_G' \mathcal{R}_r e_G = O_p(1).$$

If $f_{h,i}'$ is row i of $E_{(h)}^G$, then

$$\begin{aligned} e_G' J_h'(I_{d_h} \otimes B_{h,r}) J_h e_G &= \sum_{i=1}^{d_h} f_{h,i}' B_{h,r} f_{h,i} \\ &= \text{tr} \left(B_{h,r} (E_{(h)}^G)' E_{(h)}^G \right). \end{aligned}$$

Therefore,

$$\frac{1}{\sqrt{n}} e_G' M^G \Lambda_r P^G e_G = \frac{1}{\sqrt{n}} \sum_{h=1}^3 \text{tr} \left(B_{h,r} (E_{(h)}^G)' E_{(h)}^G \right) + O_p(n^{-1/2}). \quad (84)$$

The coefficients admit an expression that does not involve S^+ . Since $\Psi_m Y_m = P_{C_m}$ and $D_{m,r} = (t_r)_{(m)} \Psi_m$,

$$(\Lambda_{m,r} p_0)_{(m)} = D_{m,r} Y_m = (t_r)_{(m)} P_{C_m},$$

and hence

$$\Lambda_r p_0 = \sum_{m=1}^3 \Lambda_{m,r} p_0 = \sum_{m=1}^3 P_m t_r = S t_r = P^G x_r.$$

The mode- m version of this identity is

$$(P^G x_r)_{(m)} = D_{m,r} Y_m + Y_m (D_{m,r}^{(-)})'.$$

Premultiplying its transpose by M_{C_m} and postmultiplying by Ψ'_m gives

$$B_{m,r} = M_{C_m} (P^G x_r)_{(m)}' \Psi'_m. \quad (85)$$

Here we used $M_{C_m} Y_m' = 0$ and $Y_m' \Psi'_m = P_{C_m}$. In particular,

$$\|B_{m,r}\|_F = O_p(1), \quad \|B_{m,r}\|_* = O_p(1).$$

Define the sample bias vectors

$$\mathcal{B}_m^s = \frac{1}{\sqrt{n}} \mathbf{X}_G P^G \mathbf{v}_m \left(\Psi'_m (E_{(m)}^G)' E_{(m)}^G M_{C_m} \right).$$

Equation (85) implies

$$[\mathcal{B}_m^s]_r = \frac{1}{\sqrt{n}} \text{tr} \left(B_{m,r} (E_{(m)}^G)' E_{(m)}^G \right).$$

Combining (82), (83), and (84), using fixed p and $n^{-1/2} \leq \rho$, yields

$$\Gamma_n \theta = \mathcal{Z}_n - \sum_{m=1}^3 \mathcal{B}_m^s + O_p \left(\rho + \frac{s^{3/2}}{n} \right). \quad (86)$$

6. Conditional centering and bias orders. For each coordinate r ,

$$[\mathcal{B}_m^s - \mathcal{B}_m]_r = \frac{1}{\sqrt{n}} \text{tr} \left[B_{m,r} \left\{ (E_{(m)}^G)' E_{(m)}^G - \Omega_m^G \right\} \right].$$

The corresponding quadratic coefficient in the original error coordinates is

$$K_{m,r} = Q J_m' (I_{d_m} \otimes B_{m,r}) J_m Q'.$$

Because $Q'Q = I_{d_x}$ and J_m is orthogonal,

$$\|K_{m,r}\|_F^2 = d_m \|B_{m,r}\|_F^2 = O_p(d_m). \quad (87)$$

The nonzero coefficients of $K_{m,r}$ are supported on $\mathcal{R}^D$. For $m = 1$,

$$(E_{(1)}^G)' E_{(1)}^G = E_{(1)}' P_D E_{(1)},$$

and $(P_D)_{ii'}$ vanishes unless $g(i) = g(i')$. For the other two modes,

$$E_{(2)}^G = E_{(2)}(Q_D \otimes I_T), \quad E_{(3)}^G = E_{(3)}(Q_D \otimes I_M),$$

so the Gram products retain a common teacher index and a common time index, respectively. Consequently,

$$[\mathcal{B}_m^s - \mathcal{B}_m]_r = \frac{1}{\sqrt{n}} \sum_{(a,b) \in \mathcal{R}^D} [K_{m,r}]_{ab} z_{ab}.$$

The fourth-order covariance row-sum condition in Assumption 3 therefore gives

$$\mathbb{E}\left[|[\mathcal{B}_m^s - \mathcal{B}_m]_r|^2 \middle| \mathcal{F}\right] \leq \frac{c_4}{n} \|K_{m,r}\|_F^2$$

on an event whose probability tends to one. By (87) and conditional Markov's inequality,

$$\mathcal{B}_m^s - \mathcal{B}_m = O_p\left(\sqrt{\frac{d_m}{n}}\right).$$

Summing over the three modes produces an $O_p(\rho)$ difference. Substitution into (86) gives

$$\Gamma_n \sqrt{n} \gamma = \mathcal{Z}_n - \mathcal{B}_1 - \mathcal{B}_2 - \mathcal{B}_3 + O_p\left(\rho + \frac{s^{3/2}}{n}\right).$$

Since $\gamma = \hat{\delta} - \delta$, this proves (10).

Finally, each row-block conditional covariance of $E_{(m)}^G$ has operator norm at most c_ε on the covariance-summability event. Therefore,

$$\|\Omega_m^G\| \leq c_\varepsilon d_m.$$

Using (85),

$$\begin{aligned} |[\mathcal{B}_m]_r| &= \frac{1}{\sqrt{n}} |\text{tr}(B_{m,r} \Omega_m^G)| \\ &\leq \frac{1}{\sqrt{n}} \|B_{m,r}\|_* \|\Omega_m^G\| \\ &= O_p(d_m/\sqrt{n}). \end{aligned}$$

Since p is fixed and $(d_1, d_2, d_3) = (G, M, T)$, this proves the three asserted bias orders. The stated expansion uses only $\rho \rightarrow 0$; its displayed remainder is $o_p(1)$ if, additionally, $s^{3/2}/n \rightarrow 0$. $\square$

Proposition 2 (Fixed-index expansions with mode-specific remainders). *Suppose that Assumptions 1, 2, 3, 4, and 5 hold. Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be the global least-squares estimator defined in Assumption 1, and let H_u, H_v, H_w be the alignment matrices defined as*

$$\begin{aligned} H'_u &= (V * W)'(\hat{V} * \hat{W}) \left[(\hat{V} * \hat{W})'(\hat{V} * \hat{W}) \right]^{-1}, \\ H'_v &= (DU * W)'(D\hat{U} * \hat{W}) \left[(D\hat{U} * \hat{W})'(D\hat{U} * \hat{W}) \right]^{-1}, \\ H'_w &= (DU * V)'(D\hat{U} * \hat{V}) \left[(D\hat{U} * \hat{V})'(D\hat{U} * \hat{V}) \right]^{-1} \end{aligned}$$

and let $\Sigma_{VW} = (\frac{1}{M}V'V) \odot (\frac{1}{T}W'W)$, $\Sigma_{UW} = (\frac{1}{N}U'D'DU) \odot (\frac{1}{T}W'W)$, and $\Sigma_{UV} = (\frac{1}{N}U'D'DU) \odot (\frac{1}{M}V'V)$.

For each fixed admissible group g , teacher j , and period t , define

$$\begin{aligned} S_{u,g} &= \frac{1}{\sqrt{N_g MT}} \sum_{\substack{1 \leq i \leq N \\ g(i)=g}} \sum_{j'=1}^M \sum_{t'=1}^T \varepsilon_{ij't'} (v_{j'} \odot w_{t'}), \\ S_{v,j} &= \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t'=1}^T \varepsilon_{ijt'} (u_{g(i)} \odot w_{t'}), \\ S_{w,t} &= \frac{1}{\sqrt{NM}} \sum_{i=1}^N \sum_{j'=1}^M \varepsilon_{ij't} (u_{g(i)} \odot v_{j'}), \end{aligned}$$

and the $\ell \times p$ cross-moment matrices

$$\begin{aligned} Q_{u,g} &= \frac{1}{N_g MT} \sum_{\substack{1 \leq i \leq N \\ g(i)=g}} \sum_{j'=1}^M \sum_{t'=1}^T (v_{j'} \odot w_{t'}) x'_{ij't'}, \\ Q_{v,j} &= \frac{1}{NT} \sum_{i=1}^N \sum_{t'=1}^T (u_{g(i)} \odot w_{t'}) x'_{ij't'}, \\ Q_{w,t} &= \frac{1}{NM} \sum_{i=1}^N \sum_{j'=1}^M (u_{g(i)} \odot v_{j'}) x'_{ij't}. \end{aligned}$$

Let $\theta_{NMT} = \sqrt{NMT}(\hat{\delta} - \delta)$. Then

$$\begin{aligned} \sqrt{NMT} H_w^{-1} (\hat{w}_t - H_w w_t) &= \Sigma_{UV}^{-1} \left[S_{w,t} - \frac{1}{\sqrt{T}} Q_{w,t} \theta_{NMT} \right] + R_{w,t,NMT}, \\ \sqrt{NT} H_v^{-1} (\hat{v}_j - H_v v_j) &= \Sigma_{UW}^{-1} \left[S_{v,j} - \frac{1}{\sqrt{M}} Q_{v,j} \theta_{NMT} \right] + R_{v,j,NMT}, \\ \sqrt{N_g MT} H_u^{-1} (\hat{u}_g - H_u u_g) &= \Sigma_{VW}^{-1} \left[S_{u,g} - \sqrt{\frac{N_g}{N}} Q_{u,g} \theta_{NMT} \right] + R_{u,g,NMT}. \end{aligned}$$

The remainder vectors satisfy

$$\begin{aligned} \|R_{w,t,NMT}\| &= O_p \left(\rho + \frac{\sqrt{N}}{T\sqrt{M}} + \frac{\sqrt{M}}{T\sqrt{N}} \right), \\ \|R_{v,j,NMT}\| &= O_p \left(\rho + \frac{\sqrt{N}}{M\sqrt{T}} + \frac{\sqrt{T}}{M\sqrt{N}} \right), \\ \|R_{u,g,NMT}\| &= O_p \left(\rho + \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}} + \sqrt{N_g MT} \rho^4 \right), \end{aligned}$$

where $\rho = \sqrt{\frac{G+M+T}{NMT}}$. If $N_g G/N = O_p(1)$, we have

$$\|R_{u,g,NMT}\| = O_p \left(\rho + \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}} \right).$$

The group-size condition holds, in particular, when G remains fixed or when $N_g = O_p(N/G)$ for the group under consideration.

Proof. Write

$$n = NMT, \quad s = G + M + T, \quad \rho = \sqrt{\frac{s}{n}}, \quad \gamma = \hat{\delta} - \delta, \quad \theta = \theta_{NMT} = \sqrt{n} \gamma.$$

Assumption 1 implies $\rho \rightarrow 0$, while p and ℓ remain fixed. All slice arguments below are pointwise in the fixed admissible index under consideration.

1. Grouped coordinates and an exact remainder identity. By Lemma 1, the estimated factor and Khatri–Rao loading matrices have full column rank and well-conditioned normalized Gram matrices with probability approaching one, the relevant loading spaces converge, and the unrestricted block normal equations hold. All inverses and normal equations below are evaluated on these events. Inverse expressions may be assigned arbitrary values wherever they are undefined without affecting the asserted stochastic orders.

Define

$$Q_D = D(D'D)^{-1/2}, \quad Q = Q_D \otimes I_M \otimes I_T, \quad e = \text{vec}(E), \quad E^G = E \times_1 Q'_D, \quad e_G = Q' e.$$

Let

$$A_1 = \frac{(D'D)^{1/2}U}{\sqrt{N}}, \quad A_2 = \frac{V}{\sqrt{M}}, \quad A_3 = \frac{W}{\sqrt{T}}, \quad (d_1, d_2, d_3) = (G, M, T).$$

Set $d_\times = GMT$, $d_{-m} = d_\times/d_m$, and

$$C_1 = A_2 * A_3, \quad C_2 = A_1 * A_3, \quad C_3 = A_1 * A_2, \quad \Sigma_m = C'_m C_m.$$

Thus

$$(\Sigma_1, \Sigma_2, \Sigma_3) = (\Sigma_{VW}, \Sigma_{UW}, \Sigma_{UV}).$$

These are the normalized Gram matrices evaluated at the true factors. Use hats for the estimated counterparts.

For $v \in \mathbb{R}^{d_\times}$, let $v_{(m)}$ denote the mode- m matricization of its tensor reshape, with the remaining modes ordered increasingly, and let $\mathbf{v}_m$ denote the inverse operation. Define the mode-permutation matrices by

$$J_m = K_{m,1}^G, \quad J_m v = \text{vec}(v_{(m)}), \quad \mathbf{v}_m(B) = J'_m \text{vec}(B).$$

Writing $a_{m,\lambda}$ for column λ of A_m , define

$$p_0 = \sum_{\lambda=1}^{\ell} a_{1,\lambda} \otimes a_{2,\lambda} \otimes a_{3,\lambda}, \quad (p_0)_{(m)} = A_m C'_m,$$

and define $\hat{p}$ analogously. Put

$$h = \hat{p} - p_0.$$

Write $(H_1, H_2, H_3) = (H_u, H_v, H_w)$. The scalar normalizations and the group-coordinate isometries preserve the alignment matrices in the proposition, so

$$H_m = (\hat{C}'_m \hat{C}_m)^{-1} \hat{C}'_m C_m.$$

The true and estimated Gram matrices are well conditioned in probability, and the corresponding loading spaces converge. Consequently, $\hat{C}_m' C_m$ and H_m are nonsingular with probability approaching one, and

$$\|H_m^{-1}\| = O_p(1).$$

Define the graph-alignment matrix and graph error by

$$J_m^C = (C_m' C_m)^{-1} C_m' \hat{C}_m, \quad F_m = \hat{C}_m (J_m^C)^{-1} - C_m.$$

Then

$$C_m' F_m = 0, \quad \hat{C}_m = (C_m + F_m) J_m^C, \quad \hat{C}_m' C_m = (J_m^C)' \Sigma_m.$$

The graph error is invariant to changes of basis in the estimated loading space: replacing $\hat{C}_m$ by $\hat{C}_m R$, for nonsingular R , replaces J_m^C by $J_m^C R$ and leaves F_m unchanged. Hence the graph-error bounds established using the signed-permutation alignment in Proposition 1 apply to the present representatives as well.

Define $X^G = X \times_1 Q_D'$. Let $e_{m,k}^G \in \mathbb{R}^{d-m}$ be the transpose of row k of $E_{(m)}^G$, and let $X_{m,k}^G \in \mathbb{R}^{d-m \times p}$ stack the corresponding regressor vectors from X^G . To distinguish factor rows from the columns $a_{m,\lambda}$, write

$$\alpha_{m,k} = A_m' \mathbf{e}_k^{(d_m)}, \quad \hat{\alpha}_{m,k} = \hat{A}_m' \mathbf{e}_k^{(d_m)},$$

where $\mathbf{e}_k^{(d_m)}$ is the k -th coordinate vector in $\mathbb{R}^{d_m}$. Thus $\alpha_{m,k}$ and $\hat{\alpha}_{m,k}$ are the transposes of row k of the corresponding normalized factor matrices.

The block normal equations give

$$\sqrt{n}(\hat{\alpha}_{m,k} - H_m \alpha_{m,k}) = (\hat{C}_m' \hat{C}_m)^{-1} \hat{C}_m' \{e_{m,k}^G - X_{m,k}^G \gamma\}.$$

Premultiplying by H_m^{-1} and using the graph representation yields the exact identity

$$\begin{aligned} \sqrt{n} H_m^{-1}(\hat{\alpha}_{m,k} - H_m \alpha_{m,k}) &= (\hat{C}_m' C_m)^{-1} \hat{C}_m' \{e_{m,k}^G - X_{m,k}^G \gamma\} \\ &= \Sigma_m^{-1} (C_m + F_m)' \{e_{m,k}^G - X_{m,k}^G \gamma\}. \end{aligned} \quad (88)$$

Indeed,

$$(\hat{C}_m' C_m)^{-1} \hat{C}_m' = ((J_m^C)' \Sigma_m)^{-1} (J_m^C)' (C_m + F_m)' = \Sigma_m^{-1} (C_m + F_m)'.$$

The normalized rows satisfy

$$\alpha_{1,g} = \sqrt{\frac{N_g}{N}} u_g, \quad \alpha_{2,j} = \frac{v_j}{\sqrt{M}}, \quad \alpha_{3,t} = \frac{w_t}{\sqrt{T}},$$

with analogous identities for their estimators. Therefore, the left-hand side of (88) has exactly the group, teacher, and period normalizations in the proposition. Moreover,

$$C_1' e_{1,g}^G = S_{u,g}, \quad C_2' e_{2,j}^G = S_{v,j}, \quad C_3' e_{3,t}^G = S_{w,t},$$

and

$$\begin{aligned} C_1' X_{1,g}^G \gamma &= \sqrt{\frac{N_g}{N}} Q_{u,g} \theta, \\ C_2' X_{2,j}^G \gamma &= \frac{1}{\sqrt{M}} Q_{v,j} \theta, \\ C_3' X_{3,t}^G \gamma &= \frac{1}{\sqrt{T}} Q_{w,t} \theta. \end{aligned}$$

It therefore remains to bound

$$R_{m,k} = \Sigma_m^{-1} F'_m \{e_{m,k}^G - X_{m,k}^G \gamma\}, \quad (89)$$

where

$$(R_{1,g}, R_{2,j}, R_{3,t}) = (R_{u,g,NMT}, R_{v,j,NMT}, R_{w,t,NMT}).$$

2. First-order graph expansion and slice moments. Retain the tangent-space notation from the proof of Proposition 1. For $m = 1, 2, 3$, define

$$\mathcal{T}_m^G = \{v \in \mathbb{R}^{d_\times} : v_{(m)} = BC'_m, \ B \in \mathbb{R}^{d_m \times \ell}\}, \quad P_m = J'_m(I_{d_m} \otimes P_{C_m})J_m.$$

Let P^G be the orthogonal projector onto

$$\mathcal{T}^G = \mathcal{T}_1^G + \mathcal{T}_2^G + \mathcal{T}_3^G,$$

and write

$$S = P_1 + P_2 + P_3, \quad P^G = SS^+, \quad M^G = I_{d_\times} - P^G, \quad M_{\mathcal{T}} = I_n - QP^GQ'.$$

The bound $\|S^+\| = O_p(1)$ is established in the proof of Proposition 1.

Define

$$\mathbf{X} = \frac{X_{(4)}}{\sqrt{n}}, \quad \mathbf{X}_G = \mathbf{X}Q, \quad \Gamma_n = \mathbf{X}M_{\mathcal{T}}\mathbf{X}', \quad \mathcal{Z}_n = \mathbf{X}M_{\mathcal{T}}e,$$

and

$$\theta_1 = \Gamma_n^{-1}\mathcal{Z}_n, \quad F_0 = Q' - \mathbf{X}'_G\Gamma_n^{-1}\mathbf{X}M_{\mathcal{T}}.$$

Then

$$r_0 = F_0e = e_G - \mathbf{X}'_G\theta_1, \quad \xi = P^G r_0, \quad q = \frac{\xi}{\sqrt{n}}.$$

For convenience, put

$$\mathcal{A} = P^G F_0, \quad q = \frac{1}{\sqrt{n}}\mathcal{A}e.$$

The conditional regressor-moment bound gives $\|\mathbf{X}\|_F = O_p(1)$, while identification gives $\|\Gamma_n^{-1}\| = O_p(1)$. Hence $\mathcal{A}$ is $\mathcal{F}$ -measurable and

$$\|\mathcal{A}\| = O_p(1).$$

The preliminary linearization bounds established in the proof of Proposition 1 give

$$\|\gamma\| = O_p(n^{-1/2} + \rho^2), \quad \theta_1 = O_p(1), \quad \|h\| + \|q\| = O_p(\rho), \quad \|h - q\| = O_p(\rho^2). \quad (90)$$

In particular, no bound $\sqrt{n}\gamma = O_p(1)$ is used.

Define the linear map

$$\mathcal{D}_m(v) = M_{C_m}v'_{(m)}A_m(A'_m A_m)^{-1}.$$

Its operator norm between the relevant Euclidean and Frobenius spaces is $O_p(1)$. The graph expansion is

$$F_m = F_m^{(1)} + \mathcal{E}_m, \quad F_m^{(1)} = \mathcal{D}_m(q), \quad \|\mathcal{E}_m\|_F = O_p(\rho^2). \quad (91)$$

To recall its algebraic basis, set

$$\tilde{A}_m = \hat{p}_{(m)} C_m \Sigma_m^{-1} = A_m + h_{(m)} C_m \Sigma_m^{-1}.$$

The graph representation implies

$$\hat{p}_{(m)} = \tilde{A}_m (C_m + F_m)', \quad F_m = M_{C_m} h'_{(m)} \tilde{A}_m (\tilde{A}_m' \tilde{A}_m)^{-1}.$$

The strong-factor bounds, the inverse perturbation identity, and (90) therefore yield (91).

Write

$$\kappa_{1,g} = \sqrt{N_g M T}, \quad \kappa_{2,j} = \sqrt{N T}, \quad \kappa_{3,t} = \sqrt{N M}.$$

The conditional fourth-moment bound on the regressors implies

$$\|X_{m,k}^G\|_F = O_p(\kappa_{m,k}). \quad (92)$$

Indeed, conditional on D , each regressor second moment is at most $c_x^{1/2}$. For the group slice,

$$\|X_{1,g}^G\|_F^2 \leq \sum_{i:g(i)=g} \sum_{j=1}^M \sum_{t=1}^T \|x_{ijt}\|^2,$$

whose conditional expectation is at most $c_x^{1/2} N_g M T$. For teacher and period slices, multiplication by Q'_D is a Frobenius contraction, giving the corresponding conditional expectation bounds $c_x^{1/2} N T$ and $c_x^{1/2} N M$.

On the conditional covariance-summability event, Assumption 3 also gives

$$\mathbb{E}[\|e_{m,k}^G\|^2 \mid \mathcal{F}] \leq c_\varepsilon d_{-m}, \quad \mathbb{E}[\|B e\|^2 \mid \mathcal{F}] \leq c_\varepsilon \|B\|_F^2$$

for every conformable $\mathcal{F}$ -measurable matrix B . We additionally use

$$\mathbb{E}[|e' K e| \mid \mathcal{F}] \leq c_\varepsilon \|K\|_* \quad (93)$$

for every $\mathcal{F}$ -measurable square matrix K , where $\|\cdot\|_*$ denotes the nuclear norm. This follows from a singular value decomposition and conditional Cauchy–Schwarz, as in Lemma 4 (iii). The same inequality holds for e_G , since

$$\|\mathbb{E}[e_G e'_G \mid \mathcal{F}]\| = \|Q' \mathbb{E}[e e' \mid \mathcal{F}] Q\| \leq c_\varepsilon.$$

All conditional bounds below are used on these high-probability events, with localization for random coefficient norms.

3. The graph–regressor contribution. Because $q = \mathcal{A}e/\sqrt{n}$, each entry of $(F_m^{(1)})' X_{m,k}^G$ is a linear contraction of e with an $\mathcal{F}$ -measurable coefficient vector. More precisely, the matrix mapping e to the vectorization of this product has Frobenius norm at most

$$\frac{\|\mathcal{A}\| \|A_m (A_m' A_m)^{-1}\|_F \|X_{m,k}^G\|_F}{\sqrt{n}}.$$

Lemma 4(i), (92), and fixed p, ℓ therefore give

$$\|(F_m^{(1)})' X_{m,k}^G\|_F = O_p\left(\frac{\kappa_{m,k}}{\sqrt{n}}\right).$$

Combining this bound with (91) and the slope rate in (90),

$$\begin{aligned}\|F'_m X_{m,k}^G \gamma\| &\leq \|(F_m^{(1)})' X_{m,k}^G\|_F \|\gamma\| + \|\mathcal{E}_m\|_F \|X_{m,k}^G\|_F \|\gamma\| \\ &= O_p \left(\kappa_{m,k} (n^{-1/2} + \rho^2)^2 \right) \\ &= O_p \left(\rho + \kappa_{m,k} \rho^4 \right).\end{aligned}\tag{94}$$

For the last equality, use $\kappa_{m,k} \leq \sqrt{n}$, $n^{-1/2} \leq \rho$, and $\rho \rightarrow 0$:

$$\kappa_{m,k} (n^{-1/2} + \rho^2)^2 = \frac{\kappa_{m,k}}{n} + \frac{2\kappa_{m,k}\rho^2}{\sqrt{n}} + \kappa_{m,k}\rho^4 \leq n^{-1/2} + 2\rho^2 + \kappa_{m,k}\rho^4.$$

The conditional linear-contraction argument is what avoids the larger bound obtained by replacing $\|(F_m^{(1)})' X_{m,k}^G\|_F$ with $\|F_m^{(1)}\|_F \|X_{m,k}^G\|_F$.

4. The higher-order graph-error contribution. Use the tangent decomposition from the proof of Proposition 1. Define

$$B_a^q = (S^+ q)_{(a)} C_a (C_a' C_a)^{-1}, \quad a = 1, 2, 3.$$

Then

$$\mathbf{v}_a(B_a^q C_a') = P_a S^+ q.$$

Since $q \in \mathcal{T}^G$ and $SS^+ = P^G$,

$$q = \sum_{a=1}^3 \mathbf{v}_a(B_a^q C_a').$$

For each mode a , let

$$\mathcal{L}_a = \text{col}[A_a, \hat{A}_a, B_a^q], \quad \dim(\mathcal{L}_a) \leq 3\ell.$$

Fix m , and let $a < b$ be the other two modes. Every column of F_m belongs to $\mathcal{L}_a \otimes \mathcal{L}_b$, because

$$F_m = \hat{C}_m (J_m^C)^{-1} - C_m, \quad C_m = A_a * A_b, \quad \hat{C}_m = \hat{A}_a * \hat{A}_b.$$

The tangent representation of q gives

$$q'_{(m)} = C_m (B_m^q)' + (B_a^q * A_b + A_a * B_b^q) A'_m,$$

so every column of $q'_{(m)}$ also lies in $\mathcal{L}_a \otimes \mathcal{L}_b$. Since $\text{col}(C_m)$ is contained in that product space, applying M_{C_m} preserves it. Thus the columns of

$$F_m^{(1)} = M_{C_m} q'_{(m)} A_m (A'_m A_m)^{-1}$$

belong to the same product space. Consequently, each column of $\mathcal{E}_m = F_m - F_m^{(1)}$, reshaped as a $d_a \times d_b$ matrix, has rank at most 3ℓ .

The matrix spectral norm of a fixed mode- m slice of E^G is at most the tensor spectral norm $\|E^G\|$. Applying the nuclear/Frobenius inequality to each rank-at-most- 3ℓ column reshape, and also using ordinary Cauchy-Schwarz, gives the deterministic bound

$$\|\mathcal{E}'_m e_{m,k}^G\| \leq \min\{\sqrt{3\ell} \|E^G\|, \|e_{m,k}^G\|\} \|\mathcal{E}_m\|_F.$$

The spectral-error assumption, the slice second-moment bound, and (91) therefore imply

$$\|\mathcal{E}'_m e_{m,k}^G\| = O_p \left(\rho^2 \min\{\sqrt{s}, \sqrt{d-m}\} \right).\tag{95}$$

This argument holds for every realized choice of the estimated loading spaces; it does not treat these spaces or $\mathcal{E}_m$ as $\mathcal{F}$ -measurable.

5. The first-order graph-error contribution. We first separate the overlap among the tangent blocks. Let P_0 denote the orthogonal projector onto

$$\mathcal{T}^G \cap \{\text{col}(A_1) \otimes \text{col}(A_2) \otimes \text{col}(A_3)\},$$

and let

$$P_a^\circ = J'_a(M_{A_a} \otimes P_{C_a})J_a, \quad a = 1, 2, 3.$$

Splitting each tangent direction into its component in $\text{col}(A_a)$ and its orthogonal complement gives

$$P^G = P_0 + \sum_{a=1}^3 P_a^\circ, \quad P_a = P_a^\circ + J'_a(P_{A_a} \otimes P_{C_a})J_a.$$

Consequently,

$$P^G = \sum_{a=1}^3 P_a + \mathcal{R}_P, \quad \mathcal{R}_P := P^G - S = P_0 - \sum_{a=1}^3 J'_a(P_{A_a} \otimes P_{C_a})J_a.$$

Every term in the last expression has range contained in the product loading space. Hence

$$\text{rank}(\mathcal{R}_P) \leq \ell^3, \quad \|\mathcal{R}_P\| \leq \|P^G\| + \sum_{a=1}^3 \|P_a\| \leq 4.$$

The mode- m block makes no contribution to the graph map. Indeed,

$$(P_m v)_{(m)} = v_{(m)} P_{C_m}$$

implies

$$\mathcal{D}_m(P_m v) = M_{C_m} P_{C_m} v'_{(m)} A_m (A'_m A_m)^{-1} = 0.$$

The term generated by $\mathcal{R}_P$ contributes $O_p(n^{-1/2})$ to $(F_m^{(1)})' e_{m,k}^G$. For each coordinate, its quadratic-form coefficient has rank at most ℓ^3 and operator norm $O_p(1)$ before multiplication by $n^{-1/2}$, so (93) applies.

The part of q equal to

$$-\frac{1}{\sqrt{n}} P^G \chi'_G \theta_1$$

also contributes $O_p(n^{-1/2})$. To see this, expand over the fixed number p of coordinates of θ_1 . Each remaining coefficient vector is $\mathcal{F}$ -measurable and has $O_p(1)$ norm, so its contraction with $e_{m,k}^G$ is $O_p(1)$, while $\theta_1 = O_p(1)$.

For every $a \neq m$, replacing $M_{C_m} e_{m,k}^G$ by $e_{m,k}^G$ in the remaining quadratic term introduces an additional $O_p(n^{-1/2})$ error. The difference contains $P_{C_m} e_{m,k}^G$, and its quadratic-form coefficient has rank at most ℓ and operator norm $O_p(1)$. Therefore,

$$(F_m^{(1)})' e_{m,k}^G = \frac{1}{\sqrt{n}} \sum_{a \neq m} (A'_m A_m)^{-1} A'_m (P_a e_G)_{(m)} e_{m,k}^G + O_p(n^{-1/2}). \quad (96)$$

Fix distinct m, a , and let b be the remaining mode. Write $\mathcal{E}^{mab}$ for E^G with its modes ordered as (m, a, b) . To distinguish factor-component indices from the slope-coordinate index used in Proposition 1, use $\mu, \nu \in \{1, \dots, \ell\}$. Since

$$(A'_m A_m)^{-1} A'_m A_m = I_\ell,$$

and P_a projects the other two modes onto $\text{col}(C_a)$, the μ -th coordinate of the mode- a summand satisfies

$$\left[\frac{1}{\sqrt{n}} (A'_m A_m)^{-1} A'_m (P_a e_G)_{(m)} e_{m,k}^G \right]_\mu = \sum_{\nu=1}^{\ell} [\Sigma_a^{-1}]_{\mu\nu} T_{ma,k}^{\mu\nu},$$

where

$$T_{ma,k}^{\mu\nu} = \frac{1}{\sqrt{n}} \sum_{u=1}^{d_m} \sum_{v=1}^{d_a} \sum_{w,w'=1}^{d_b} \mathcal{E}_{kvw}^{mab} \mathcal{E}_{uvw'}^{mab} [A_m]_{u\nu} [A_b]_{w\mu} [A_b]_{w'\nu}. \quad (97)$$

The strong-factor bounds give $\|\Sigma_a^{-1}\| = O_p(1)$.

We first bound the centered contraction. Let $K_{ma,k}^{G,\mu\nu}$ be the coefficient matrix satisfying

$$T_{ma,k}^{\mu\nu} = e'_G K_{ma,k}^{G,\mu\nu} e_G.$$

Its squared Frobenius norm is

$$\|K_{ma,k}^{G,\mu\nu}\|_F^2 = \frac{d_a}{n} \| [A_m]_{\cdot\nu} \|^2 \| [A_b]_{\cdot\mu} \|^2 \| [A_b]_{\cdot\nu} \|^2 = O_p(d_a/n).$$

The coefficient in the original error coordinates is

$$K_{ma,k}^{\mu\nu} = Q K_{ma,k}^{G,\mu\nu} Q'.$$

Since $Q'Q = I_{d_\times}$,

$$\|K_{ma,k}^{\mu\nu}\|_F = \|K_{ma,k}^{G,\mu\nu}\|_F.$$

Its nonzero ordered error pairs share the mode- a index in grouped coordinates. In the original coordinates, they therefore have a common group, teacher, or period, according as $a = 1, 2, 3$. Thus the coefficient is supported on $\mathcal{R}^D$. Lemma 4(ii) yields

$$T_{ma,k}^{\mu\nu} - \mathbb{E}[T_{ma,k}^{\mu\nu} \mid \mathcal{F}] = O_p \left(\sqrt{\frac{d_a}{n}} \right).$$

For the conditional means, let

$$\mathcal{I}_{1,g} = \mathcal{I}_g^{(g)}, \quad \mathcal{I}_{2,j} = \mathcal{I}_j^{(j)}, \quad \mathcal{I}_{3,t} = \mathcal{I}_t^{(t)}$$

be the source slices in Assumption 5. Their cardinalities are $\kappa_{m,k}^2$. For an original error index α , abbreviate

$$u_\alpha = u_{g(i_\alpha)}, \quad v_\alpha = v_{j_\alpha}, \quad w_\alpha = w_{t_\alpha}.$$

Also define

$$\chi_J(\alpha, \beta) = \mathbf{1}\{j_\alpha = j_\beta\}, \quad \chi_T(\alpha, \beta) = \mathbf{1}\{t_\alpha = t_\beta\}, \quad p_{\alpha\beta} = (P_D)_{i_\alpha i_\beta},$$

where

$$p_{\alpha\beta} = \frac{\mathbf{1}\{g(i_\alpha) = g(i_\beta)\}}{N_{g(i_\alpha)}}.$$

Expanding the normalized grouped errors and factors in (97) gives

$$T_{ma,k}^{\mu\nu} = \frac{1}{\kappa_{m,k} c_{ma}} \sum_{\alpha \in \mathcal{I}_{m,k}} \sum_{\beta \in \mathcal{I}} \chi_{ma}(\alpha, \beta) \phi_\alpha \psi_\beta \varepsilon_\alpha \varepsilon_\beta, \quad (98)$$

with the following coefficients, where the arguments of χ_J, χ_T, p are suppressed:

(m, a)	ϕ_α	ψ_β	χ_{ma}	c_{ma}
$(1, 2)$	$w_{\alpha, \mu}$	$u_{\beta, \nu} w_{\beta, \nu}$	χ_J	NT
$(1, 3)$	$v_{\alpha, \mu}$	$u_{\beta, \nu} v_{\beta, \nu}$	χ_T	NM
$(2, 1)$	$w_{\alpha, \mu}$	$v_{\beta, \nu} w_{\beta, \nu}$	p	MT
$(2, 3)$	$u_{\alpha, \mu}$	$u_{\beta, \nu} v_{\beta, \nu}$	χ_T	NM
$(3, 1)$	$v_{\alpha, \mu}$	$v_{\beta, \nu} w_{\beta, \nu}$	p	MT
$(3, 2)$	$u_{\alpha, \mu}$	$u_{\beta, \nu} w_{\beta, \nu}$	χ_J	NT

For example,

$$T_{12, g}^{\mu\nu} = \frac{1}{NT\sqrt{N_g MT}} \sum_{i: g(i)=g} \sum_{i'=1}^N \sum_{j=1}^M \sum_{t, t'=1}^T \varepsilon_{ijt} \varepsilon_{i'jt'} u_{g(i'), \nu} w_{t, \mu} w_{t', \nu}.$$

The source grouped error supplies the factor $N_g^{-1/2}$, while the destination grouped-error normalization cancels the corresponding group-size normalization in A_1 .

In every row of the table,

$$0 \leq \chi_{ma}(\alpha, \beta) \leq 1, \quad |\psi_\beta|^2 \leq \omega_\beta,$$

where ω_β is defined in Assumption 5. Let

$$\Gamma_{\alpha\beta} = \mathbb{E}[\varepsilon_\alpha \varepsilon_\beta \mid \mathcal{F}].$$

The conditional covariance row-sum bound and the factor Gram bounds imply

$$\begin{aligned} \sum_{\alpha \in \mathcal{I}_{m, k}} \sum_{\beta \in \mathcal{I}} |\Gamma_{\alpha\beta}| |\phi_\alpha|^2 &\leq c_\varepsilon \sum_{\alpha \in \mathcal{I}_{m, k}} |\phi_\alpha|^2 \\ &= O_p(\kappa_{m, k}^2). \end{aligned}$$

Assumption 5 also gives

$$\begin{aligned} \sum_{\alpha \in \mathcal{I}_{m, k}} \sum_{\beta \in \mathcal{I}} |\Gamma_{\alpha\beta}| |\psi_\beta|^2 &\leq \sum_{\alpha \in \mathcal{I}_{m, k}} \sum_{\beta \in \mathcal{I}} |\Gamma_{\alpha\beta}| \omega_\beta \\ &= O_p(\kappa_{m, k}^2). \end{aligned}$$

Weighted Cauchy–Schwarz in (98) therefore yields

$$\begin{aligned} \left| \mathbb{E}[T_{ma, k}^{\mu\nu} \mid \mathcal{F}] \right| &\leq \frac{1}{\kappa_{m, k} c_{ma}} \left(\sum_{\alpha \in \mathcal{I}_{m, k}} \sum_{\beta \in \mathcal{I}} |\Gamma_{\alpha\beta}| |\phi_\alpha|^2 \right)^{1/2} \\ &\quad \times \left(\sum_{\alpha \in \mathcal{I}_{m, k}} \sum_{\beta \in \mathcal{I}} |\Gamma_{\alpha\beta}| |\psi_\beta|^2 \right)^{1/2} \\ &= O_p \left(\frac{\kappa_{m, k}}{c_{ma}} \right). \end{aligned} \tag{99}$$

Define

$$\begin{aligned} b_{1,g} &= \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}}, \\ b_{2,j} &= \frac{\sqrt{N}}{M\sqrt{T}} + \frac{\sqrt{T}}{M\sqrt{N}}, \\ b_{3,t} &= \frac{\sqrt{N}}{T\sqrt{M}} + \frac{\sqrt{M}}{T\sqrt{N}}. \end{aligned}$$

These are exactly

$$b_{m,k} = \sum_{a \neq m} \frac{\kappa_{m,k}}{c_{ma}}.$$

Combining (96)–(99), using fixed ℓ , $\|\Sigma_a^{-1}\| = O_p(1)$, and $\sqrt{d_a/n} \leq \rho$, gives

$$\|(F_m^{(1)})' e_{m,k}^G\| = O_p(\rho + b_{m,k}). \quad (100)$$

6. Rate comparisons and conclusion. The higher-order error term in (95) is absorbed by the same rate:

$$\rho^2 \min\{\sqrt{s}, \sqrt{d_{-m}}\} \lesssim \rho + b_{m,k}. \quad (101)$$

To verify this, first suppose $d_m = \max_{a=1,2,3} d_a$. Since $s \leq 3d_m$ and $n = (N/G)d_m d_{-m}$,

$$\rho^2 \sqrt{d_{-m}} \leq \frac{3d_m \sqrt{d_{-m}}}{n} \leq 3\sqrt{\frac{G}{N}} \rho \leq 3\rho.$$

Otherwise, one of the other two modes attains the largest dimension, so

$$\rho^2 \sqrt{s} = \frac{s^{3/2}}{n} \leq \frac{3\sqrt{3}}{n} \sum_{a \neq m} d_a^{3/2} \leq 3\sqrt{3} b_{m,k}.$$

The last inequality follows from the displayed expressions for $b_{m,k}$, using $G \leq N$, $N_g \geq 1$, and $N, M, T \geq 1$.

Equations (91), (95), (100), and (101) imply

$$\|F_m' e_{m,k}^G\| \leq \|(F_m^{(1)})' e_{m,k}^G\| + \|\mathcal{E}_m' e_{m,k}^G\| = O_p(\rho + b_{m,k}).$$

Combining this with (94), (89), and $\|\Sigma_m^{-1}\| = O_p(1)$ proves the unified remainder bound

$$\|R_{m,k}\| = O_p(\rho + b_{m,k} + \kappa_{m,k} \rho^4). \quad (102)$$

For the period block, $G \leq N$ gives

$$\begin{aligned} \kappa_{3,t} \rho^4 &= \rho^2 \frac{G + M + T}{T\sqrt{NM}} \\ &\leq \rho^2 \left(\frac{\sqrt{N}}{T\sqrt{M}} + \frac{\sqrt{M}}{T\sqrt{N}} + \frac{1}{\sqrt{NM}} \right). \end{aligned}$$

Since $1/\sqrt{NM} \leq \rho$ and $\rho \rightarrow 0$, this term is absorbed into $\rho + b_{3,t}$. Similarly,

$$\kappa_{2,j} \rho^4 \leq \rho^2 \left(\frac{\sqrt{N}}{M\sqrt{T}} + \frac{\sqrt{T}}{M\sqrt{N}} + \frac{1}{\sqrt{NT}} \right)$$

is absorbed into $\rho + b_{2,j}$. Thus (102) yields

$$\|R_{w,t, NMT}\| = O_p \left(\rho + \frac{\sqrt{N}}{T\sqrt{M}} + \frac{\sqrt{M}}{T\sqrt{N}} \right)$$

and

$$\|R_{v,j, NMT}\| = O_p \left(\rho + \frac{\sqrt{N}}{M\sqrt{T}} + \frac{\sqrt{T}}{M\sqrt{N}} \right).$$

For the group block, retaining $\kappa_{1,g}\rho^4$ gives

$$\|R_{u,g, NMT}\| = O_p \left(\rho + \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}} + \sqrt{N_g M T} \rho^4 \right).$$

Finally, suppose $N_g G/N = O_p(1)$. Then

$$\begin{aligned} \sqrt{N_g M T} \rho^4 &= \rho^2 \left[\sqrt{\frac{N_g G}{N}} \sqrt{\frac{G}{n}} + \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}} \right] \\ &= \rho^2 [O_p(\rho) + b_{1,g}] \\ &= o_p(\rho + b_{1,g}). \end{aligned}$$

Hence the group remainder simplifies to

$$\|R_{u,g, NMT}\| = O_p \left(\rho + \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}} \right).$$

The group-size condition holds when G remains fixed, since $N_g \leq N$, and when $N_g = O_p(N/G)$ for the group under consideration. Together with the exact identity (88) and the score and regressor identities in Step 1, these bounds establish all three expansions and the stated remainder rates. The remainder vectors are $o_p(1)$ whenever their corresponding displayed rates tend to zero. $\square$

Theorem 1 (Joint asymptotic normality of the slope and aligned factor errors). *Suppose that Assumptions 1, 2, 3, 4, 5, and 6 hold. Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be the global least-squares estimator defined in Assumption 1. Suppose that*

$$\frac{(G + M + T)^{3/2}}{NMT} \rightarrow 0. \quad (11)$$

Let $b_\delta = b_1 + b_2 + b_3$, and use the limiting matrices and vectors defined in Assumption 6.

The estimated factors are expressed in the true-factor column labeling and sign orientation. Specifically, among the equivalent least-squares triples

$$(\hat{U}\Pi_\ell S_v S_w, \hat{V}\Pi_\ell S_v, \hat{W}\Pi_\ell S_w),$$

where Π_ℓ ranges over $\ell \times \ell$ permutation matrices and S_v, S_w range over diagonal matrices with entries in $\{-1, 1\}$, choose a representative minimizing

$$\frac{1}{M} \|\hat{V}\Pi_\ell S_v - V\|_F^2 + \frac{1}{T} \|\hat{W}\Pi_\ell S_w - W\|_F^2.$$

Retain the hat notation for the selected representative, and define H_u, H_v, H_w from it as in Proposition 2. The followings hold.

1. (Identity limits of the alignment matrices.)

The matrices H_u, H_v, H_w are nonsingular with probability approaching one and satisfy

$$H_u \xrightarrow{p} I_\ell, \quad H_v \xrightarrow{p} I_\ell, \quad H_w \xrightarrow{p} I_\ell.$$

2. (Slope central limit theorem.)

The slope estimator satisfies

$$\sqrt{NMT}(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(-\Gamma_\delta^{-1}b_\delta, \Gamma_\delta^{-1}\Omega_Z\Gamma_\delta^{-1}).$$

3. (Joint fixed-index central limit theorem.)

Fix an admissible triple (g, j, t) and suppose additionally that

$$\frac{N+M}{T\sqrt{NM}} \rightarrow 0, \quad \frac{N+T}{M\sqrt{NT}} \rightarrow 0, \quad \frac{\sqrt{N_g}(M+T)}{N\sqrt{MT}} \xrightarrow{p} 0. \quad (12)$$

Let

$$C_{u,g} = \sqrt{\pi_g} \bar{Q}_{u,g}.$$

Then the aligned factor errors satisfy

$$\begin{pmatrix} \sqrt{N_g MT}(\hat{u}_g - H_u u_g) \\ \sqrt{NT}(\hat{v}_j - H_v v_j) \\ \sqrt{NM}(\hat{w}_t - H_w w_t) \\ \sqrt{NMT}(\hat{\delta} - \delta) \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mu_{gjt}, \mathcal{A}_{gjt} \Sigma_{gjt} \mathcal{A}'_{gjt}), \quad (13)$$

where

$$\mathcal{A}_{gjt} = \begin{pmatrix} \bar{\Sigma}_{VW}^{-1} & 0 & 0 & -\bar{\Sigma}_{VW}^{-1} C_{u,g} \Gamma_\delta^{-1} \\ 0 & \bar{\Sigma}_{UW}^{-1} & 0 & 0 \\ 0 & 0 & \bar{\Sigma}_{UV}^{-1} & 0 \\ 0 & 0 & 0 & \Gamma_\delta^{-1} \end{pmatrix}, \quad \mu_{gjt} = \begin{pmatrix} \bar{\Sigma}_{VW}^{-1} C_{u,g} \Gamma_\delta^{-1} b_\delta \\ 0_\ell \\ 0_\ell \\ -\Gamma_\delta^{-1} b_\delta \end{pmatrix}.$$

For completeness, define

$$\begin{aligned} V_{u,g} &= \bar{\Sigma}_{VW}^{-1} \left[\Gamma_g^{vw} + C_{u,g} \Gamma_\delta^{-1} \Omega_Z \Gamma_\delta^{-1} C'_{u,g} - \Xi_{gZ} \Gamma_\delta^{-1} C'_{u,g} - C_{u,g} \Gamma_\delta^{-1} \Xi'_{gZ} \right] \bar{\Sigma}_{VW}^{-1}, \\ V_{v,j} &= \bar{\Sigma}_{UW}^{-1} \Gamma_j^{uw} \bar{\Sigma}_{UW}^{-1}, \\ V_{w,t} &= \bar{\Sigma}_{UV}^{-1} \Gamma_t^{uv} \bar{\Sigma}_{UV}^{-1}. \end{aligned}$$

The marginal fixed-index limits are

$$\begin{aligned} \sqrt{N_g MT}(\hat{u}_g - H_u u_g) &\xrightarrow{d} \mathcal{N}(\bar{\Sigma}_{VW}^{-1} C_{u,g} \Gamma_\delta^{-1} b_\delta, V_{u,g}), \\ \sqrt{NT}(\hat{v}_j - H_v v_j) &\xrightarrow{d} \mathcal{N}(0, V_{v,j}), \\ \sqrt{NM}(\hat{w}_t - H_w w_t) &\xrightarrow{d} \mathcal{N}(0, V_{w,t}). \end{aligned} \quad (14)$$

Proof. Write

$$n = NMT, \quad s = G + M + T, \quad \rho = \sqrt{\frac{s}{n}}, \quad \gamma = \hat{\delta} - \delta, \quad \theta = \theta_{NMT} = \sqrt{n} \gamma.$$

Assumption 1 implies $\rho \rightarrow 0$. Throughout, the estimated factors are represented using the column matching and compensating signs specified in the theorem. All inverse expressions are evaluated on events whose probabilities tend to one and may be extended arbitrarily elsewhere.

1. Identity limits of the alignment matrices. Introduce the normalized grouped factors

$$Q_D = D(D'D)^{-1/2}, \quad A_1 = \frac{(D'D)^{1/2}U}{\sqrt{N}}, \quad A_2 = \frac{V}{\sqrt{M}}, \quad A_3 = \frac{W}{\sqrt{T}},$$

with estimated counterparts $\hat{A}_m$.

Lemma 1 supplies an equivalent signed-permuted representative whose sum of squared normalized V - and W -errors is $O_p(\rho^2)$. The representative specified in the theorem minimizes this sum over the same finite collection. Therefore,

$$\|\hat{A}_2 - A_2\|_F^2 + \|\hat{A}_3 - A_3\|_F^2 = O_p(\rho^2),$$

and hence

$$\|\hat{A}_2 - A_2\|_F + \|\hat{A}_3 - A_3\|_F = O_p(\rho).$$

The minimum is well defined because the collection is finite, and the stated tie-breaking rule defines a particular representative.

Let Π and $\hat{\Pi}$ denote the true and estimated common-component tensors. Define their normalized grouped vectors by

$$p_0 = \frac{1}{\sqrt{n}} \text{vec}(\Pi \times_1 Q'_D), \quad \hat{p} = \frac{1}{\sqrt{n}} \text{vec}(\hat{\Pi} \times_1 Q'_D), \quad h = \hat{p} - p_0.$$

Also, put

$$C_1 = A_2 * A_3, \quad C_2 = A_1 * A_3, \quad C_3 = A_1 * A_2,$$

and define $\hat{C}_m$ analogously. The factor-alignment rate in Lemma 1, the bounded normalized factor norms, and trilinearity imply the common-component rate

$$\|h\| = \frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{n}} = O_p(\rho).$$

The equality follows because $Q'_D Q_D = I_G$ and the first-mode fibers of both common components belong to $\text{col}(D)$. The common-component difference is unchanged by the choice of signed-permuted representative.

The true and estimated columns of A_2 and A_3 have unit norm. Expanding the Khatri–Rao difference gives

$$\hat{C}_1 - C_1 = (\hat{A}_2 - A_2) * \hat{A}_3 + A_2 * (\hat{A}_3 - A_3),$$

so

$$\|\hat{C}_1 - C_1\|_F \leq \|\hat{A}_2 - A_2\|_F + \|\hat{A}_3 - A_3\|_F = O_p(\rho).$$

On the strong-factor event,

$$\sigma_{\min}(C_1) \geq c_0.$$

Consequently, $\sigma_{\min}(\hat{C}_1)$ is bounded away from zero with probability approaching one, and

$$\left\| \hat{C}_1(\hat{C}_1' \hat{C}_1)^{-1} \right\| = \frac{1}{\sigma_{\min}(\hat{C}_1)} = O_p(1).$$

As in the preceding proofs, $h_{(1)}$ denotes the mode-1 matricization of the tensor reshape of h . Since

$$h_{(1)} = \hat{A}_1 \hat{C}_1' - A_1 C_1',$$

we have the exact identity

$$\hat{A}_1 - A_1 = \left[h_{(1)} + A_1(C_1 - \hat{C}_1)' \right] \hat{C}_1(\hat{C}_1' \hat{C}_1)^{-1}.$$

The preceding bounds and $\|A_1\| = O_p(1)$ imply

$$\|\hat{A}_1 - A_1\|_F = O_p(\rho).$$

Expanding the other two Khatri–Rao differences in the same way therefore gives

$$\sum_{m=1}^3 \|\hat{A}_m - A_m\|_F = O_p(\rho), \quad \sum_{m=1}^3 \|\hat{C}_m - C_m\|_F = O_p(\rho). \quad (103)$$

The strong-factor bounds imply that the singular values of each C_m are bounded above and away from zero. Hence

$$\max_{m=1,2,3} \left\{ \|\hat{C}_m\| + \|(\hat{C}_m' \hat{C}_m)^{-1}\| \right\} = O_p(1).$$

Writing $(H_1, H_2, H_3) = (H_u, H_v, H_w)$, the definitions in Proposition 2 become

$$H_m = (\hat{C}_m' \hat{C}_m)^{-1} \hat{C}_m' C_m, \quad m = 1, 2, 3.$$

The scalar normalizations and isometric group embeddings cancel from these projection coefficients. Thus

$$H_m - I_\ell = (\hat{C}_m' \hat{C}_m)^{-1} \hat{C}_m' (C_m - \hat{C}_m).$$

Equation (103) implies

$$\sum_{m=1}^3 \|H_m - I_\ell\| = O_p(\rho).$$

Because $\rho \rightarrow 0$, the event

$$\max_{m=1,2,3} \|H_m - I_\ell\| < \frac{1}{2}$$

has probability tending to one. On this event, every H_m is nonsingular, $\|H_m^{-1}\| \leq 2$, and

$$H_m^{-1} - I_\ell = -H_m^{-1}(H_m - I_\ell).$$

Therefore,

$$\sum_{m=1}^3 \left\{ \|H_m - I_\ell\| + \|H_m^{-1} - I_\ell\| \right\} = O_p(\rho). \quad (104)$$

This proves part (i) without imposing any alignment-limit condition in Assumption 6.

2. Slope linearization and marginal limit. Proposition 1 gives

$$\Gamma_{NMT}\theta = \mathcal{Z}_{NMT} - \mathcal{B}_1 - \mathcal{B}_2 - \mathcal{B}_3 + O_p\left(\rho + \frac{s^{3/2}}{n}\right).$$

The remainder is $o_p(1)$ by (11) and $\rho \rightarrow 0$. Assumption 6 implies

$$\mathcal{Z}_{NMT} = O_p(1), \quad \sum_{m=1}^3 \mathcal{B}_m \xrightarrow{p} b_\delta, \quad \Gamma_{NMT}^{-1} \xrightarrow{p} \Gamma_\delta^{-1}.$$

Consequently,

$$\theta = \Gamma_\delta^{-1}(\mathcal{Z}_{NMT} - b_\delta) + o_p(1) = O_p(1). \quad (105)$$

The marginal score CLT in Assumption 6 gives

$$\mathcal{Z}_{NMT} \xrightarrow{d} \mathcal{N}(0, \Omega_Z).$$

The continuous mapping theorem and Slutsky's theorem therefore imply

$$\theta \xrightarrow{d} \mathcal{N}(-\Gamma_\delta^{-1}b_\delta, \Gamma_\delta^{-1}\Omega_Z\Gamma_\delta^{-1}).$$

This proves part (ii), without using (12).

3. Fixed-index linearizations and removal of the inverse alignments. Fix an admissible triple (g, j, t) and impose (12). Define

$$\begin{aligned} a_{u,g,n} &= \frac{\sqrt{N_g}(M+T)}{N\sqrt{MT}} = \frac{\sqrt{N_g M}}{N\sqrt{T}} + \frac{\sqrt{N_g T}}{N\sqrt{M}}, \\ a_{v,n} &= \frac{N+T}{M\sqrt{NT}} = \frac{\sqrt{N}}{M\sqrt{T}} + \frac{\sqrt{T}}{M\sqrt{N}}, \\ a_{w,n} &= \frac{N+M}{T\sqrt{NM}} = \frac{\sqrt{N}}{T\sqrt{M}} + \frac{\sqrt{M}}{T\sqrt{N}}. \end{aligned}$$

Proposition 2 gives

$$\begin{aligned} \|R_{u,g,NMT}\| &= O_p\left(\rho + a_{u,g,n} + \sqrt{N_g MT} \rho^4\right), \\ \|R_{v,j,NMT}\| &= O_p(\rho + a_{v,n}), \\ \|R_{w,t,NMT}\| &= O_p(\rho + a_{w,n}). \end{aligned}$$

The three quantities $a_{u,g,n}, a_{v,n}, a_{w,n}$ tend to zero under (12). Moreover, $N_g \leq N$ implies

$$\sqrt{N_g MT} \rho^4 \leq \sqrt{n} \rho^4 = \rho \frac{s^{3/2}}{n} = o(\rho).$$

Hence

$$\|R_{u,g,NMT}\| + \|R_{v,j,NMT}\| + \|R_{w,t,NMT}\| = o_p(1). \quad (106)$$

In particular, no additional condition on $N_g G/N$ is needed.

The first two conditions in (12) imply $T \rightarrow \infty$ and $M \rightarrow \infty$, respectively, because

$$\frac{N+M}{T\sqrt{NM}} \geq \frac{2}{T}, \quad \frac{N+T}{M\sqrt{NT}} \geq \frac{2}{M}.$$

Define the effective slope coefficients

$$C_{u,g,n} = \sqrt{\frac{N_g}{N}} Q_{u,g}, \quad C_{v,j,n} = \frac{1}{\sqrt{M}} Q_{v,j}, \quad C_{w,t,n} = \frac{1}{\sqrt{T}} Q_{w,t}.$$

The group-share and group cross-moment limits in Assumption 6, together with continuity of the square-root function on $[0, 1]$, give

$$C_{u,g,n} \xrightarrow{p} \sqrt{\pi_g} \bar{Q}_{u,g} = C_{u,g}.$$

For the other two modes, the boundedness assumptions $Q_{v,j} = O_p(1)$ and $Q_{w,t} = O_p(1)$, together with $M, T \rightarrow \infty$, imply

$$C_{u,g,n} \xrightarrow{p} C_{u,g}, \quad C_{v,j,n} = o_p(1), \quad C_{w,t,n} = o_p(1). \quad (107)$$

Since $\theta = O_p(1)$ by (105), the teacher and period slope-estimation contributions are $o_p(1)$, while the group contribution need not vanish.

Put

$$D_u = \bar{\Sigma}_{VW}^{-1}, \quad D_v = \bar{\Sigma}_{UW}^{-1}, \quad D_w = \bar{\Sigma}_{UV}^{-1}, \quad D_\delta = \Gamma_\delta^{-1},$$

and define the inverse-alignment-adjusted errors

$$\begin{aligned} \zeta_{u,g,n} &= \sqrt{N_g M T} H_u^{-1} (\hat{u}_g - H_u u_g), \\ \zeta_{v,j,n} &= \sqrt{N T} H_v^{-1} (\hat{v}_j - H_v v_j), \\ \zeta_{w,t,n} &= \sqrt{N M} H_w^{-1} (\hat{w}_t - H_w w_t). \end{aligned}$$

Proposition 2 gives

$$\begin{aligned} \zeta_{u,g,n} &= \Sigma_{VW}^{-1} [S_{u,g} - C_{u,g,n} \theta] + R_{u,g,NMT}, \\ \zeta_{v,j,n} &= \Sigma_{UW}^{-1} [S_{v,j} - C_{v,j,n} \theta] + R_{v,j,NMT}, \\ \zeta_{w,t,n} &= \Sigma_{UV}^{-1} [S_{w,t} - C_{w,t,n} \theta] + R_{w,t,NMT}. \end{aligned}$$

The joint score CLT implies that the three loading scores and Z_{NMT} are jointly bounded in probability. The loading Gram limits, (106), and (107) therefore imply

$$\begin{aligned} \zeta_{u,g,n} &= D_u [S_{u,g} - C_{u,g} \theta] + o_p(1), \\ \zeta_{v,j,n} &= D_v S_{v,j} + o_p(1), \\ \zeta_{w,t,n} &= D_w S_{w,t} + o_p(1). \end{aligned}$$

For the group block, replacing $C_{u,g,n}$ by $C_{u,g}$ is valid because

$$(C_{u,g,n} - C_{u,g}) \theta = o_p(1).$$

Substituting (105) into the group expansion gives

$$\begin{aligned} \zeta_{u,g,n} &= D_u [S_{u,g} - C_{u,g} D_\delta (Z_{NMT} - b_\delta)] + o_p(1), \\ \zeta_{v,j,n} &= D_v S_{v,j} + o_p(1), \\ \zeta_{w,t,n} &= D_w S_{w,t} + o_p(1). \end{aligned} \quad (108)$$

These three vectors are jointly $O_p(1)$. Thus (104) gives

$$\begin{aligned}\sqrt{N_g MT}(\hat{u}_g - H_u u_g) - \zeta_{u,g,n} &= (H_u - I_\ell)\zeta_{u,g,n} = o_p(1), \\ \sqrt{NT}(\hat{v}_j - H_v v_j) - \zeta_{v,j,n} &= (H_v - I_\ell)\zeta_{v,j,n} = o_p(1), \\ \sqrt{NM}(\hat{w}_t - H_w w_t) - \zeta_{w,t,n} &= (H_w - I_\ell)\zeta_{w,t,n} = o_p(1).\end{aligned}$$

Consequently, (108) remains valid without the premultiplying inverse alignment matrices. This proves

$$\begin{aligned}\sqrt{N_g MT}(\hat{u}_g - H_u u_g) &= \bar{\Sigma}_{VW}^{-1} [S_{u,g} - C_{u,g} \Gamma_\delta^{-1} (\mathcal{Z}_{NMT} - b_\delta)] + o_p(1), \\ \sqrt{NT}(\hat{v}_j - H_v v_j) &= \bar{\Sigma}_{UV}^{-1} S_{v,j} + o_p(1), \\ \sqrt{NM}(\hat{w}_t - H_w w_t) &= \bar{\Sigma}_{UV}^{-1} S_{w,t} + o_p(1).\end{aligned}$$

The finite-sample alignment matrices remain inside the centers throughout this argument. For example, replacing $H_u u_g$ by u_g would additionally require

$$\sqrt{N_g MT}(H_u - I_\ell)u_g = o_p(1),$$

which does not follow from (104) alone. No such replacement is made here.

4. Joint Gaussian limit and marginal covariance matrices. Define the score and estimator stacks

$$S_n = \begin{pmatrix} S_{u,g} \\ S_{v,j} \\ S_{w,t} \\ \mathcal{Z}_{NMT} \end{pmatrix}, \quad L_n = \begin{pmatrix} \sqrt{N_g MT}(\hat{u}_g - H_u u_g) \\ \sqrt{NT}(\hat{v}_j - H_v v_j) \\ \sqrt{NM}(\hat{w}_t - H_w w_t) \\ \theta \end{pmatrix}.$$

Equations (105) and (108), together with the preceding removal of the inverse alignment matrices, give

$$L_n = \mathcal{A}_{gjt} S_n + \mu_{gjt} + o_p(1),$$

where $\mathcal{A}_{gjt}$ and μ_{gjt} are the deterministic matrix and vector specified in the theorem. In particular,

$$D_u [S_{u,g} - C_{u,g} D_\delta (\mathcal{Z}_{NMT} - b_\delta)] = D_u S_{u,g} - D_u C_{u,g} D_\delta \mathcal{Z}_{NMT} + D_u C_{u,g} D_\delta b_\delta.$$

Thus the group-loading mean has the displayed plus sign, while the slope mean is $-D_\delta b_\delta$. Neither mean vector is asserted to have componentwise positive or negative entries.

Assumption 6 gives

$$S_n \xrightarrow{d} \mathcal{N}(0, \Sigma_{gjt}).$$

The continuous mapping theorem and Slutsky's theorem therefore imply

$$L_n \xrightarrow{d} \mathcal{N}(\mu_{gjt}, \mathcal{A}_{gjt} \Sigma_{gjt} \mathcal{A}_{gjt}').$$

This proves (13). The stack with the first three blocks premultiplied by $H_u^{-1}, H_v^{-1}, H_w^{-1}$ differs from L_n by $o_p(1)$ and therefore has the same joint limit.

To obtain the marginal covariance formulas, let

$$\begin{pmatrix} S_u^\infty \\ S_v^\infty \\ S_w^\infty \\ Z^\infty \end{pmatrix} \sim \mathcal{N}(0, \Sigma_{gjt}).$$

The random part of the group-loading limit is

$$D_u(S_u^\infty - C_{u,g}D_\delta Z^\infty).$$

Since

$$\text{Var}(S_u^\infty) = \Gamma_g^{vw}, \quad \text{Var}(Z^\infty) = \Omega_Z, \quad \text{Cov}(S_u^\infty, Z^\infty) = \Xi_{gZ},$$

and D_u, D_δ are symmetric, its covariance is

$$\begin{aligned} \text{Var}[D_u(S_u^\infty - C_{u,g}D_\delta Z^\infty)] &= D_u \left[\Gamma_g^{vw} + C_{u,g}D_\delta \Omega_Z D_\delta C'_{u,g} - \Xi_{gZ} D_\delta C'_{u,g} - C_{u,g}D_\delta \Xi'_{gZ} \right] D_u \\ &= V_{u,g}. \end{aligned}$$

Its mean is $D_u C_{u,g} D_\delta b_\delta$. The teacher and period limits are $D_v S_v^\infty$ and $D_w S_w^\infty$, respectively, so their means are zero and their covariance matrices are

$$D_v \Gamma_j^{uw} D_v = V_{v,j}, \quad D_w \Gamma_t^{uv} D_w = V_{w,t}.$$

These calculations establish (14). All cross-covariance blocks in Assumption 6 remain represented in the full covariance matrix $\mathcal{A}_{gjt} \Sigma_{gjt} \mathcal{A}'_{gjt}$; none is assumed to vanish.

Finally, $\mathcal{A}_{gjt}$ is invertible because it is block upper triangular with invertible diagonal blocks. Therefore,

$$\text{rank}(\mathcal{A}_{gjt} \Sigma_{gjt} \mathcal{A}'_{gjt}) = \text{rank}(\Sigma_{gjt}).$$

The joint limiting distribution may thus be degenerate when Σ_{gjt} is singular. This does not affect the asserted convergence in distribution. All applications of the score CLT and the fixed-index expansions concern the selected triple (g, j, t) , so the conclusions are pointwise, as stated. $\square$

Proposition 3. *Suppose that Assumptions 1, 2, 3, and 4 hold. Suppose also that $\mathbb{E}[\varepsilon_a \varepsilon_b \mid \mathcal{F}] = 0$ for all distinct $a, b \in \mathcal{I}$. The true factor rows satisfy*

$$\frac{1}{N} \sum_{g=1}^G N_g \|u_g\|^4 + \frac{1}{M} \sum_{j=1}^M \|v_j\|^4 + \frac{1}{T} \sum_{t=1}^T \|w_t\|^4 = O_p(1)$$

and the group sizes satisfies

$$\sum_{g=1}^G \frac{1}{N_g} = O_p\left(\frac{G^2}{N}\right). \quad (16)$$

Let $\mathcal{T}$ be the tangent space defined in (6) evaluated at the true parameters (U, V, W) and let Γ_{NMT} and $\mathcal{Z}_{NMT}$ be as defined in Proposition 1. Set $\hat{\hat{U}} = (D'D)^{1/2} \hat{U}$ and

$$(\hat{\hat{F}}_1, \hat{\hat{F}}_2, \hat{\hat{F}}_3) = (\hat{\hat{U}}, \hat{V}, \hat{W}), \quad (\hat{\hat{C}}_1, \hat{\hat{C}}_2, \hat{\hat{C}}_3) = (\hat{V} * \hat{W}, \hat{\hat{U}} * \hat{W}, \hat{\hat{U}} * \hat{V}).$$

Let $\hat{\Omega}_m^G$, $m = 1, 2, 3$, be data-based estimators of the conditional covariance defined in (15), of dimensions $MT \times MT$, $GT \times GT$, and $GM \times GM$, respectively. Define

$$\hat{\mathcal{B}}_m = \frac{1}{\sqrt{NMT}} X_{(4)} P_{\hat{\tau}} Q K_{1,m}^G \text{vec} \left(\hat{F}_m (\hat{F}_m' \hat{F}_m)^{-1} (\hat{C}_m' \hat{C}_m)^{-1} \hat{C}_m' \hat{\Omega}_m^G M_{\hat{C}_m} \right), \quad m = 1, 2, 3.$$

The bias-corrected estimator $\hat{\delta}_{bc}$ is defined as

$$\hat{\delta}_{bc} = \hat{\delta} + \frac{1}{\sqrt{NMT}} \hat{\Gamma}_{NMT}^{-1} (\hat{\mathcal{B}}_1 + \hat{\mathcal{B}}_2 + \hat{\mathcal{B}}_3), \quad \hat{\Gamma}_{NMT} = \frac{1}{NMT} X_{(4)} M_{\hat{\tau}} X_{(4)}'$$

Then, we have

$$\Gamma_{NMT} \sqrt{NMT} (\hat{\delta}_{bc} - \delta) = \mathcal{Z}_{NMT} + O_p \left(\sqrt{\frac{G+M+T}{NMT}} + \frac{(G+M+T)^{3/2}}{NMT} \right). \quad (17)$$

Proof. Write $n = NMT$, $s = G + M + T$, $\rho = \sqrt{s/n}$, and $\gamma = \hat{\delta} - \delta$. Assumption 1 implies $\rho \rightarrow 0$. Let $e = \text{vec}(E)$ and $\hat{e} = \text{vec}(\hat{E})$, and put

$$B_{NMT} = \sum_{m=1}^3 \mathcal{B}_m, \quad \hat{B}_{NMT} = \sum_{m=1}^3 \hat{\mathcal{B}}_m.$$

We use the signed-permutation-aligned estimates supplied by Lemma 1. These relabelings leave the fitted common component, the slope estimate, and the displayed bias formulas unchanged. Unless a weighted norm or operator norm is specified, matrix norms below are Frobenius norms. The conditional inequalities are applied on the covariance-bound events from Assumption 3, whose probabilities approach one.

Step 1: Normalized coordinates and preliminary bounds. Put $(d_1, d_2, d_3) = (G, M, T)$ and define

$$\mu_1(g) = \frac{N_g}{N}, \quad \mu_2(j) = \frac{1}{M}, \quad \mu_3(t) = \frac{1}{T}, \quad \mu = \mu_1 \otimes \mu_2 \otimes \mu_3.$$

For a probability-weight vector ν , write $D_\nu = \text{diag}(\nu)$. For a matrix B with rows b'_q , define

$$\|B\|_{\alpha, \nu} = \left(\sum_q \nu(q) \|b_q\|_2^\alpha \right)^{1/\alpha}, \quad \alpha \geq 1.$$

For a square matrix P , write

$$\|P\|_{4, \nu \rightarrow 4, \nu} = \sup_{z \neq 0} \frac{\|Pz\|_{4, \nu}}{\|z\|_{4, \nu}}.$$

All weights are positive because D has full column rank.

Let

$$(B_1, B_2, B_3) = (U, V, W), \quad A_m = D_{\mu_m}^{1/2} B_m.$$

Thus

$$A_1 = \frac{(D'D)^{1/2} U}{\sqrt{N}}, \quad A_2 = \frac{V}{\sqrt{M}}, \quad A_3 = \frac{W}{\sqrt{T}}.$$

For complementary modes $a < b$, put

$$B_{-m} = B_a * B_b, \quad \mu_{-m} = \mu_a \otimes \mu_b, \quad L_m = A_a * A_b = D_{\mu_{-m}}^{1/2} B_{-m}.$$

Here L_m denotes a normalized pair matrix, whereas the matrices C_m in the proposition are unnormalized. Define

$$G_m = A'_m A_m = B'_m D_{\mu_m} B_m, \quad J_m = L'_m L_m = G_a \odot G_b,$$

and use hats for the estimated counterparts.

Lemma 1 gives

$$\sum_{m=1}^3 \|\hat{A}_m - A_m\|_F = O_p(\rho).$$

Since ℓ is fixed and the normalized factor norms are bounded, expanding the Khatri–Rao products gives

$$\sum_{m=1}^3 \|\hat{L}_m - L_m\|_F = O_p(\rho).$$

Consequently,

$$\|\hat{G}_m - G_m\|_{\text{op}} + \|\hat{J}_m - J_m\|_{\text{op}} = O_p(\rho).$$

Assumption 2 gives $c_0 I_\ell \leq G_m \leq c_1 I_\ell$. Moreover, J_m is a principal submatrix of $G_a \otimes G_b$, so

$$c_0^2 I_\ell \leq J_m \leq c_1^2 I_\ell.$$

The true and estimated Gram inverses are therefore $O_p(1)$ in operator norm, and the inverse identity gives

$$\|\hat{G}_m^{-1} - G_m^{-1}\|_{\text{op}} + \|\hat{J}_m^{-1} - J_m^{-1}\|_{\text{op}} = O_p(\rho). \quad (109)$$

Write $(H_1, H_2, H_3) = (H_u, H_v, H_w)$ and define

$$H'_m = L'_m \hat{L}_m \hat{J}_m^{-1}, \quad \Delta_m = \hat{A}_m - A_m H'_m.$$

Since

$$H'_m - I_\ell = (L_m - \hat{L}_m)' \hat{L}_m \hat{J}_m^{-1},$$

we obtain

$$\sum_m \|H_m - I_\ell\|_{\text{op}} + \sum_m \|\Delta_m\|_F = O_p(\rho).$$

The conditional regressor fourth-moment bound implies

$$\frac{1}{n} \sum_{i,j,t} \|x_{ijt}\|_2^4 = O_p(1), \quad \frac{\|X_{(4)}\|_F}{\sqrt{n}} = O_p(1).$$

Assumption 3 implies $\|\mathbb{E}[ee' \mid \mathcal{F}]\|_{\text{op}} \leq c_\varepsilon$ on its covariance-bound event. Hence

$$\mathbb{E}[\|\mathcal{Z}_{NMT}\|_2^2 \mid \mathcal{F}] \leq \frac{c_\varepsilon}{n} \|X_{(4)} M_{\mathcal{T}}\|_F^2 \leq \frac{c_\varepsilon}{n} \|X_{(4)}\|_F^2,$$

and conditional Markov's inequality gives $\mathcal{Z}_{NMT} = O_p(1)$.

Every CP tangent variation is annihilated by $M_{DU*V} \otimes M_W$: variations in U or V are annihilated by M_W , and variations in W are annihilated by M_{DU*V} . Thus

$$M_{\mathcal{T}} \geq M_{DU*V} \otimes M_W.$$

Assumption 4, evaluated at the true factors, consequently gives $\lambda_{\min}(\Gamma_{NMT}) \geq c_\delta$ with probability approaching one. Proposition 1 and $B_{NMT} = O_p(s/\sqrt{n})$ now imply

$$\begin{aligned} \|\gamma\|_2 &= O_p\left(n^{-1/2} + \frac{s}{n} + \frac{\rho}{\sqrt{n}} + \frac{s^{3/2}}{n^{3/2}}\right) \\ &= O_p(n^{-1/2} + \rho^2) = O_p(\rho). \end{aligned}$$

Here $\rho/\sqrt{n} = \sqrt{s}/n \leq s/n$ and $s^{3/2}/n^{3/2} = \rho^3 \leq \rho^2$ eventually. No additional condition on $s^{3/2}/n$ is needed for this preliminary bound.

Step 2: Fourth moments of factor errors and residual squares. Define the source slices

$$\mathcal{I}_g^{(1)} = \{(i, j, t) \in \mathcal{I} : g(i) = g\}, \quad \mathcal{I}_j^{(2)} = \{(i, j', t) \in \mathcal{I} : j' = j\}, \quad \mathcal{I}_t^{(3)} = \{(i, j, t') \in \mathcal{I} : t' = t\}.$$

Their cardinalities are $N_g MT$, NT , and NM , respectively. Let S be any member of $\{\mathcal{I}_r^{(m)} : m = 1, 2, 3, 1 \leq r \leq d_m\}$, and let $L = \sum_{a \in S} b_a \varepsilon_a$ for $\mathcal{F}$ -measurable scalar coefficients. Since $S^2 \subseteq \mathcal{R}^D$, Assumption 3 yields

$$\text{Var}(L^2 \mid \mathcal{F}) \leq c_4 \sum_{a, b \in S} b_a^2 b_b^2 = c_4 \left(\sum_{a \in S} b_a^2 \right)^2,$$

while its second-order bound gives $\mathbb{E}[L^2 \mid \mathcal{F}] \leq c_\varepsilon \sum_{a \in S} b_a^2$. Therefore,

$$\mathbb{E}[L^4 \mid \mathcal{F}] \leq (c_4 + c_\varepsilon^2) \left(\sum_{a \in S} b_a^2 \right)^2. \quad (110)$$

In particular, taking a single nonzero coefficient gives a uniform conditional bound on $\mathbb{E}[\varepsilon_a^4 \mid \mathcal{F}]$. Conditional Markov's inequality implies

$$\frac{1}{n} \sum_a \varepsilon_a^4 = O_p(1), \quad \|e\|_4 = O_p(n^{1/4}), \quad \|e \odot e\|_2 = O_p(\sqrt{n}). \quad (111)$$

Let $Q_D = D(D'D)^{-1/2}$, $E^G = E \times_1 Q'_D$, and $X_\gamma^G = (X \times_4 \gamma') \times_1 Q'_D$. Constraint inactivity gives the unrestricted block equations

$$\Delta_m = \frac{1}{\sqrt{n}} (E^G - X_\gamma^G)_{(m)} \hat{L}_m \hat{J}_m^{-1}.$$

For complementary modes $a < b$, expand

$$\hat{L}_m = (A_a \otimes A_b)(H'_a * H'_b) + R_m,$$

where

$$R_m = \Delta_a * (A_b H'_b) + (A_a H'_a) * \Delta_b + \Delta_a * \Delta_b, \quad \|R_m\|_F = O_p(\rho).$$

Define the true-factor score vectors

$$\begin{aligned} z_{1,g}^\varepsilon &= \frac{1}{N_g MT} \sum_{i:g(i)=g} \sum_{j',t'} \varepsilon_{ij't'} (v_{j'} \otimes w_{t'}), \\ z_{2,j}^\varepsilon &= \frac{1}{NT} \sum_{i,t'} \varepsilon_{ij't'} (u_{g(i)} \otimes w_{t'}), \\ z_{3,t}^\varepsilon &= \frac{1}{NM} \sum_{i,j'} \varepsilon_{ij't} (u_{g(i)} \otimes v_{j'}). \end{aligned}$$

Let Z_m^ε have rows $(z_{m,r}^\varepsilon)'$. The normalized Gram bounds imply that the squared coefficient norms of each scalar score coordinate are bounded by constants times $(N_g MT)^{-1}$, $(NT)^{-1}$, and $(NM)^{-1}$, respectively. Equation (110) therefore gives

$$\mathbb{E}[\|z_{1,g}^\varepsilon\|_2^4 \mid \mathcal{F}] \leq \frac{C}{(N_g MT)^2}, \quad \mathbb{E}[\|z_{2,j}^\varepsilon\|_2^4 \mid \mathcal{F}] \leq \frac{C}{(NT)^2}, \quad \mathbb{E}[\|z_{3,t}^\varepsilon\|_2^4 \mid \mathcal{F}] \leq \frac{C}{(NM)^2}.$$

Writing $H_N = \sum_g N_g^{-1}$, the corresponding weighted averages satisfy

$$\begin{aligned} \mathbb{E}[\|Z_1^\varepsilon\|_{4,\mu_1}^4 \mid \mathcal{F}] &\leq \frac{CNH_N}{n^2}, \\ \mathbb{E}[\|Z_2^\varepsilon\|_{4,\mu_2}^4 \mid \mathcal{F}] &\leq \frac{CM^2}{n^2}, \\ \mathbb{E}[\|Z_3^\varepsilon\|_{4,\mu_3}^4 \mid \mathcal{F}] &\leq \frac{CT^2}{n^2}. \end{aligned}$$

By (16) and conditional Markov's inequality,

$$\|Z_m^\varepsilon\|_{4,\mu_m} = O_p\left(\sqrt{\frac{d_m}{n}}\right).$$

Undoing the normalizations in the block equations gives the exact decomposition

$$D_{\mu_m}^{-1/2} \Delta_m = Z_m^\varepsilon (H'_a * H'_b) \hat{J}_m^{-1} + E_m^R - X_m^\gamma,$$

where

$$E_m^R = \frac{1}{\sqrt{n}} D_{\mu_m}^{-1/2} E_{(m)}^G R_m \hat{J}_m^{-1}, \quad X_m^\gamma = \frac{1}{\sqrt{n}} D_{\mu_m}^{-1/2} (X_\gamma^G)_{(m)} \hat{L}_m \hat{J}_m^{-1}.$$

The first contribution has weighted fourth norm $O_p(\sqrt{d_m/n})$, since its right multiplier is bounded in operator norm.

For the adaptive error contribution, define

$$\eta_{1,g} = \frac{\|E_{g,:}^G\|_F}{\sqrt{N_g MT}}, \quad \eta_{2,j} = \frac{\|E_{:,j}^G\|_F}{\sqrt{NT}}, \quad \eta_{3,t} = \frac{\|E_{:,t}^G\|_F}{\sqrt{NM}}.$$

Rowwise Cauchy–Schwarz gives

$$\|(E_m^R)_{r,:}\|_2 \leq \|\hat{J}_m^{-1}\|_{\text{op}} \|R_m\|_F \eta_{m,r}.$$

Equation (110), applied to $E_{gjt}^G = N_g^{-1/2} \sum_{i:g(i)=g} \varepsilon_{ijt}$, gives $\mathbb{E}[|E_{gjt}^G|^4 \mid \mathcal{F}] \leq C$. Using $(\sum_{k=1}^K e_k^2)^2 \leq K \sum_{k=1}^K e_k^4$ within each grouped-error slice therefore yields

$$\begin{aligned} \mathbb{E} \left[\sum_g \mu_1(g) \eta_{1,g}^4 \mid \mathcal{F} \right] &\leq \frac{CH_N}{N} \leq C, \\ \mathbb{E} \left[\frac{1}{M} \sum_j \eta_{2,j}^4 \mid \mathcal{F} \right] &\leq C(G/N)^2 \leq C, \\ \mathbb{E} \left[\frac{1}{T} \sum_t \eta_{3,t}^4 \mid \mathcal{F} \right] &\leq C(G/N)^2 \leq C. \end{aligned}$$

Here $N_g \geq 1$ implies $H_N \leq G \leq N$. Consequently,

$$\|E_m^R\|_{4,\mu_m} \leq \|\hat{J}_m^{-1}\|_{\text{op}} \|R_m\|_F \left(\sum_r \mu_m(r) \eta_{m,r}^4 \right)^{1/4} = O_p(\rho).$$

This deterministic bound permits R_m to depend on the errors.

For the regressor contribution, source-slice Cauchy–Schwarz and $\|\hat{L}_m \hat{J}_m^{-1}\|_{\text{op}} = O_p(1)$ give a common $K_{NMT} = O_p(1)$ such that

$$\|(X_m^\gamma)_{r,:}\|_2 \leq K_{NMT} \|\gamma\|_2 \left(\frac{1}{|\mathcal{I}_r^{(m)}|} \sum_{a \in \mathcal{I}_r^{(m)}} \|x_a\|_2^2 \right)^{1/2}.$$

For modes 2 and 3, the grouped contraction cannot increase the source-slice Euclidean norm; for mode 1, the same statement follows directly from Cauchy–Schwarz over the N_g observations. Raising the bound to the fourth power and applying Jensen’s inequality gives

$$\|X_m^\gamma\|_{4,\mu_m}^4 \leq K_{NMT}^4 \|\gamma\|_2^4 \sum_r \frac{\mu_m(r)}{|\mathcal{I}_r^{(m)}|} \sum_{a \in \mathcal{I}_r^{(m)}} \|x_a\|_2^4 = K_{NMT}^4 \|\gamma\|_2^4 \frac{1}{n} \sum_a \|x_a\|_2^4.$$

Thus $\|X_m^\gamma\|_{4,\mu_m} = O_p(\|\gamma\|_2) = O_p(\rho)$.

Combining the three contributions gives

$$\|D_{\mu_m}^{-1/2} \Delta_m\|_{4,\mu_m} = O_p(\rho).$$

Since

$$\hat{B}_m - B_m = D_{\mu_m}^{-1/2} \Delta_m + B_m(H'_m - I_\ell),$$

the assumed true-factor fourth moments and the alignment bounds imply

$$\sum_{m=1}^3 \|\hat{B}_m - B_m\|_{4,\mu_m} = O_p(\rho), \quad \sum_{m=1}^3 (\|B_m\|_{4,\mu_m} + \|\hat{B}_m\|_{4,\mu_m}) = O_p(1). \quad (112)$$

Now let $h_{ijt} = x'_{ijt} \gamma + \hat{\pi}_{ijt} - \pi_{ijt}$, so $\hat{\varepsilon}_{ijt} = \varepsilon_{ijt} - h_{ijt}$. For scalar column vectors on separate modes, their weighted tensor fourth norm is the product of their mode-specific fourth norms. For each k , use the exact expansion

$$\begin{aligned} \hat{u}_{gk} \hat{v}_{jk} \hat{w}_{tk} - u_{gk} v_{jk} w_{tk} &= (\hat{u}_{gk} - u_{gk}) \hat{v}_{jk} \hat{w}_{tk} \\ &\quad + u_{gk} (\hat{v}_{jk} - v_{jk}) \hat{w}_{tk} \\ &\quad + u_{gk} v_{jk} (\hat{w}_{tk} - w_{tk}). \end{aligned}$$

Equation (112), fixed ℓ , and the triangle inequality imply

$$\left(\frac{1}{n} \sum_{i,j,t} |\hat{\pi}_{ijt} - \pi_{ijt}|^4 \right)^{1/4} = O_p(\rho).$$

The raw observation average equals the product-weighted grouped average because the common-component error depends on i only through $g(i)$. Also,

$$\left(\frac{1}{n} \sum_{i,j,t} |x'_{ijt} \gamma|^4 \right)^{1/4} \leq \|\gamma\|_2 \left(\frac{1}{n} \sum_{i,j,t} \|x_{ijt}\|_2^4 \right)^{1/4} = O_p(\rho).$$

Consequently, identifying h with its vectorization,

$$\|h\|_4 = O_p(n^{1/4}\rho), \quad \|e \odot h\|_2 \leq \|e\|_4 \|h\|_4 = O_p(\sqrt{s}), \quad \|h \odot h\|_2 = \|h\|_4^2 = O_p(s/\sqrt{n}).$$

Using $\hat{e} \odot \hat{e} - e \odot e = -2e \odot h + h \odot h$ gives

$$\|\hat{e} \odot \hat{e} - e \odot e\|_2 = O_p\left(\sqrt{s} + \frac{s}{\sqrt{n}}\right). \quad (113)$$

Step 3: Tangent projectors and projected regressors. For a full-column-rank matrix B and probability weights ν , put $G_B = B' D_\nu B$ and

$$P_B^\nu = B G_B^{-1} B' D_\nu.$$

This is the column-space projector for the weighted inner product induced by D_ν . Hölder's inequality gives

$$\|P_B^\nu\|_{4,\nu \rightarrow 4,\nu} \leq \|B\|_{4,\nu} \|G_B^{-1}\|_{\text{op}} \|B\|_{4/3,\nu}. \quad (114)$$

If $\|\hat{B} - B\|_{4,\nu} = O_p(\rho)$, the true and estimated fourth norms are $O_p(1)$, and the Gram inverses are bounded, then $\|\hat{G}_B - G_B\|_{\text{op}} = O_p(\rho)$ and $\|\hat{G}_B^{-1} - G_B^{-1}\|_{\text{op}} = O_p(\rho)$. The exact expansion

$$\begin{aligned} P_B^\nu - P_{\hat{B}}^\nu &= (\hat{B} - B) \hat{G}_B^{-1} \hat{B}' D_\nu \\ &\quad + B (\hat{G}_B^{-1} - G_B^{-1}) \hat{B}' D_\nu \\ &\quad + B G_B^{-1} (\hat{B} - B)' D_\nu \end{aligned}$$

and $\|\hat{B} - B\|_{4/3,\nu} \leq \|\hat{B} - B\|_{4,\nu}$ imply

$$\|P_B^\nu - P_{\hat{B}}^\nu\|_{4,\nu \rightarrow 4,\nu} = O_p(\rho). \quad (115)$$

Define

$$P_m = B_m G_m^{-1} B_m' D_{\mu_m}, \quad P_{-m} = B_{-m} J_m^{-1} B_{-m}' D_{\mu_{-m}},$$

with analogous estimated projectors. The rowwise inequality $\|b_a \odot b_b\|_2 \leq \|b_a\|_2 \|b_b\|_2$ implies

$$\|B_{-m}\|_{4,\mu_{-m}} + \|\hat{B}_{-m}\|_{4,\mu_{-m}} = O_p(1).$$

Expanding $\hat{B}_{-m} - B_{-m} = (\hat{B}_a - B_a) * \hat{B}_b + B_a * (\hat{B}_b - B_b)$ also gives

$$\|\hat{B}_{-m} - B_{-m}\|_{4,\mu_{-m}} = O_p(\rho).$$

Thus the mode and pair projectors have bounded weighted fourth-norm operator norms, and their estimated-minus-true differences have order $O_p(\rho)$.

Let R_ℓ select the canonical coordinates (r, k, k) , (k, r, k) , and (k, k, r) in $\mathbb{R}^{\ell^3}$, with duplicates removed. It has $r_0 = 3\ell^2 - 2\ell$ orthonormal columns. Define

$$\Phi = (B_1 \otimes B_2 \otimes B_3)R_\ell, \quad G_{\text{core}} = \Phi' D_\mu \Phi = R'_\ell (G_1 \otimes G_2 \otimes G_3) R_\ell.$$

The normalized Gram bounds imply

$$c_0^3 I_{r_0} \leq G_{\text{core}} \leq c_1^3 I_{r_0}.$$

The product weights and the telescoping expansion of the full Kronecker product give

$$\|\Phi\|_{4,\mu} + \|\hat{\Phi}\|_{4,\mu} = O_p(1), \quad \|\hat{\Phi} - \Phi\|_{4,\mu} = O_p(\rho).$$

Equations (114) and (115) therefore apply to the core projectors P_Φ^μ and $P_{\hat{\Phi}}^\mu$.

The weighted projector onto the grouped physical tangent space has the orthogonal decomposition

$$P_{\mathcal{T}}^\mu = \sum_{m=1}^3 K_{1,m}^G [(I_{d_m} - P_m) \otimes P_{-m}] (K_{1,m}^G)' + P_\Phi^\mu. \quad (116)$$

Indeed, split each mode variation into its weighted projection onto $\text{col}(B_m)$ and its orthogonal complement. The complementary variations generate $\text{col}(B_m)^{\perp_{\mu_m}} \otimes \text{col}(B_{-m})$. These three spaces are mutually orthogonal and orthogonal to the core space. The remaining variations generate $\text{col}(\Phi)$, since their coefficient coordinates are exactly those selected by R_ℓ .

Applying a projector to each complementary slice and summing the fourth powers shows that tensor extensions preserve the weighted fourth-norm bounds. Expanding each tensor-product projector difference in (116) consequently gives

$$\|P_{\mathcal{T}}^\mu\|_{4,\mu \rightarrow 4,\mu} + \|P_{\hat{\mathcal{T}}}^\mu\|_{4,\mu \rightarrow 4,\mu} = O_p(1), \quad \|P_{\hat{\mathcal{T}}}^\mu - P_{\mathcal{T}}^\mu\|_{4,\mu \rightarrow 4,\mu} = O_p(\rho). \quad (117)$$

We also need the Euclidean projector bound. In normalized grouped coordinates, the same tangent decomposition uses A_m , L_m , and the core matrix $\Psi = (A_1 \otimes A_2 \otimes A_3)R_\ell$. All these matrices have bounded norms, Gram matrices bounded away from singularity, and Frobenius perturbations $O_p(\rho)$. The ordinary column-projector expansion therefore gives an $O_p(\rho)$ operator-norm perturbation for each component and hence for their sum. Writing $Q = Q_D \otimes I_M \otimes I_T$, the original tangent projectors are obtained by the isometric embedding Q . Consequently,

$$\|P_{\hat{\mathcal{T}}} - P_{\mathcal{T}}\|_{\text{op}} = O_p(\rho). \quad (118)$$

This argument uses normalized coordinates directly.

Let $\bar{X}^G \in \mathbb{R}^{GMT \times p}$ have rows $\bar{x}'_{gjt}$, where $\bar{x}_{gjt} = N_g^{-1} \sum_{i:g(i)=g} x_{ijt}$, and put

$$\mathcal{X} = P_{\mathcal{T}}^\mu \bar{X}^G, \quad \hat{\mathcal{X}} = P_{\hat{\mathcal{T}}}^\mu \bar{X}^G.$$

Jensen's inequality gives

$$\|\bar{X}^G\|_{4,\mu}^4 \leq \frac{1}{n} \sum_{i,j,t} \|x_{ijt}\|_2^4 = O_p(1).$$

Since the projections act columnwise and p is fixed, (117) implies

$$\|\mathcal{X}\|_{4,\mu} + \|\hat{\mathcal{X}}\|_{4,\mu} = O_p(1), \quad \|\hat{\mathcal{X}} - \mathcal{X}\|_{4,\mu} = O_p(\rho). \quad (119)$$

The relation between physical and normalized grouped coordinates is

$$Q' P_T X'_{(4)} = \sqrt{n} D_\mu^{1/2} \mathcal{X}, \quad Q' P_{\hat{T}} X'_{(4)} = \sqrt{n} D_\mu^{1/2} \hat{\mathcal{X}}. \quad (120)$$

Step 4: Exact bias coefficients and coefficient rates. Write $b'_{m,r}$ for row r of B_m and $c'_{m,q}$ for row q of B_{-m} . Let $\xi'_{r,q}$ be the row of $\mathcal{X}$ at source index r and complementary index q . Define

$$Z_{m,q} = \sum_r \mu_m(r) b_{m,r} \xi'_{r,q} \in \mathbb{R}^{\ell \times p}.$$

Let $Z_m \in \mathbb{R}^{d_a d_b \times \ell p}$ have row q equal to $\text{vec}(Z_{m,q})'$, put $T_m = (I_{d_a d_b} - P_{-m})Z_m$, and let $T_{m,q}$ denote the $\ell \times p$ reshaping of row q . Define

$$k_{m,q} = T'_{m,q} G_m^{-1} J_m^{-1} c_{m,q},$$

and let K_m have rows $k'_{m,q}$. Define estimated quantities by replacing the factors, projected regressors, Grams, and projectors with their estimated counterparts.

Hölder's inequality in the source weights gives

$$\|Z_m\|_{4,\mu_{-m}} \leq \|B_m\|_{4/3,\mu_m} \|\mathcal{X}\|_{4,\mu} = O_p(1).$$

The blockwise identity

$$\hat{Z}_{m,q} - Z_{m,q} = \sum_r \mu_m(r) (\hat{b}_{m,r} - b_{m,r}) \hat{\xi}'_{r,q} + \sum_r \mu_m(r) b_{m,r} (\hat{\xi}_{r,q} - \xi_{r,q})'$$

and (112) and (119) give $\|\hat{Z}_m - Z_m\|_{4,\mu_{-m}} = O_p(\rho)$. The pair-projector bounds and

$$\hat{T}_m - T_m = (I - \hat{P}_{-m})(\hat{Z}_m - Z_m) - (\hat{P}_{-m} - P_{-m})Z_m$$

then imply

$$\|T_m\|_{4,\mu_{-m}} + \|\hat{T}_m\|_{4,\mu_{-m}} = O_p(1), \quad \|\hat{T}_m - T_m\|_{4,\mu_{-m}} = O_p(\rho).$$

Hölder's inequality for the product of two weighted fourth-norm matrices gives

$$\|K_m\|_{2,\mu_{-m}} \leq \|G_m^{-1} J_m^{-1}\|_{\text{op}} \|T_m\|_{4,\mu_{-m}} \|B_{-m}\|_{4,\mu_{-m}} = O_p(1).$$

Expanding $\hat{k}_{m,q} - k_{m,q}$ one factor at a time, keeping the inverse-Gram order $G_m^{-1} J_m^{-1}$, and using (109) gives

$$\|K_m\|_{2,\mu_{-m}} + \|\hat{K}_m\|_{2,\mu_{-m}} = O_p(1), \quad \|\hat{K}_m - K_m\|_{2,\mu_{-m}} = O_p(\rho). \quad (121)$$

To identify the bias coefficients exactly, note the unnormalized scalings

$$(F_1, F_2, F_3) = (\bar{U}, V, W) = (\sqrt{N}A_1, \sqrt{M}A_2, \sqrt{T}A_3),$$

and

$$(C_1, C_2, C_3) = (V * W, \bar{U} * W, \bar{U} * V) = (\sqrt{MT}L_1, \sqrt{NT}L_2, \sqrt{NM}L_3).$$

Thus the matrix inside the mode- m bias formula is

$$\frac{1}{\sqrt{n}} A_m G_m^{-1} J_m^{-1} L'_m \Omega_m^G M_{L_m}.$$

Fix a regressor coordinate r , and let $e_r^{(p)}$ be the r th canonical basis vector of $\mathbb{R}^p$. Let $Y_m^{(r)}$ unfold the grouped tensor with entries $\sqrt{\mu(g, j, t)} \xi_{gjt}^{(r)}$. Let $Z_m^{(r)}$ have row q equal to $(Z_{m,q} e_r^{(p)})'$, and define $T_m^{(r)}$ analogously. Then

$$(Y_m^{(r)})' A_m = D_{\mu-m}^{1/2} Z_m^{(r)}, \quad M_{L_m} D_{\mu-m}^{1/2} = D_{\mu-m}^{1/2} (I - P_{-m}).$$

Equation (120) and cyclicity of the trace give

$$\begin{aligned} \mathcal{B}_m^{(r)} &= \frac{1}{\sqrt{n}} \text{tr} \left((Y_m^{(r)})' A_m G_m^{-1} J_m^{-1} L'_m \Omega_m^G M_{L_m} \right) \\ &= \frac{1}{\sqrt{n}} \text{tr} \left(D_{\mu-m}^{1/2} T_m^{(r)} G_m^{-1} J_m^{-1} B'_{-m} D_{\mu-m}^{1/2} \Omega_m^G \right) \\ &= \frac{1}{\sqrt{n}} \sum_q \mu_{-m}(q) (\Omega_m^G)_{qq} k_{m,q}^{(r)}. \end{aligned}$$

The last equality uses conditional uncorrelatedness, which makes Ω_m^G diagonal. Stacking coordinates, and repeating the calculation for estimated quantities, gives

$$\mathcal{B}_m = \frac{1}{\sqrt{n}} \sum_q \mu_{-m}(q) (\Omega_m^G)_{qq} k_{m,q}, \quad \hat{\mathcal{B}}_m = \frac{1}{\sqrt{n}} \sum_q \mu_{-m}(q) (\hat{\Omega}_m^G)_{qq} \hat{k}_{m,q}.$$

Put $\sigma_a^2 = \mathbb{E}[\varepsilon_a^2 \mid \mathcal{F}]$ and $\sigma_{gjt}^{G,2} = N_g^{-1} \sum_{i:g(i)=g} \sigma_{ijt}^2$. Under conditional uncorrelatedness,

$$(\Omega_1^G)_{(j,t),(j,t)} = \sum_g \sigma_{gjt}^{G,2}, \quad (\Omega_2^G)_{(g,t),(g,t)} = \sum_j \sigma_{gjt}^{G,2}, \quad (\Omega_3^G)_{(g,j),(g,j)} = \sum_t \sigma_{gjt}^{G,2}.$$

Substituting these formulas and the residual-based estimators in (15) yields the exact representations

$$\mathcal{B}_m = \sum_a \lambda_{m,a} \sigma_a^2, \quad \hat{\mathcal{B}}_m = \sum_a \hat{\lambda}_{m,a} \hat{\varepsilon}_a^2,$$

where

$$\lambda_{1,ijt} = \frac{N}{N_{g(i)} n^{3/2}} k_{1,(j,t)}, \quad \lambda_{2,ijt} = \frac{M}{n^{3/2}} k_{2,(g(i),t)}, \quad \lambda_{3,ijt} = \frac{T}{n^{3/2}} k_{3,(g(i),j)}.$$

The estimated coefficients replace $k_{m,q}$ by $\hat{k}_{m,q}$.

Let $\mathcal{L}_m \in \mathbb{R}^{n \times p}$ have rows $\lambda'_{m,a}$, and define $\hat{\mathcal{L}}_m$ similarly. Direct summation gives

$$\|\mathcal{L}_1\|_F^2 = \frac{NH_N}{n^2} \|K_1\|_{2,\mu_2 \otimes \mu_3}^2, \quad \|\mathcal{L}_2\|_F^2 = \frac{M^2}{n^2} \|K_2\|_{2,\mu_1 \otimes \mu_3}^2, \quad \|\mathcal{L}_3\|_F^2 = \frac{T^2}{n^2} \|K_3\|_{2,\mu_1 \otimes \mu_2}^2.$$

The same identities apply to estimated coefficients and their differences. Equations (16) and (121) consequently imply

$$\|\mathcal{L}_m\|_F + \|\hat{\mathcal{L}}_m\|_F = O_p(d_m/n), \quad \|\hat{\mathcal{L}}_m - \mathcal{L}_m\|_F = O_p(d_m \rho/n). \quad (122)$$

Step 5: The bias-estimation rate. Let q have entries $q_a = \varepsilon_a^2 - \sigma_a^2$. The exact coefficient representations give

$$\hat{\mathcal{B}}_m - \mathcal{B}_m = \mathcal{L}'_m q + (\hat{\mathcal{L}}_m - \mathcal{L}_m)'(e \odot e) + \hat{\mathcal{L}}'_m(\hat{e} \odot \hat{e} - e \odot e).$$

The true coefficients are $\mathcal{F}$ -measurable. Since all diagonal pairs (a, a) belong to $\mathcal{R}^D$, Assumption 3 implies

$$\text{Var} \left(\sum_a \lambda_{m,a}^{(r)} q_a \mid \mathcal{F} \right) \leq c_4 \sum_a |\lambda_{m,a}^{(r)}|^2.$$

Conditional Chebyshev's inequality and fixed p therefore give

$$\|\mathcal{L}'_m q\|_2 = O_p(d_m/n).$$

For the other terms, deterministic Cauchy–Schwarz inequalities, (111), (113), and (122) give

$$\|(\hat{\mathcal{L}}_m - \mathcal{L}_m)'(e \odot e)\|_2 \leq \|\hat{\mathcal{L}}_m - \mathcal{L}_m\|_F \|e \odot e\|_2 = O_p(d_m \rho / \sqrt{n})$$

and

$$\|\hat{\mathcal{L}}'_m(\hat{e} \odot \hat{e} - e \odot e)\|_2 \leq \|\hat{\mathcal{L}}_m\|_F \|\hat{e} \odot \hat{e} - e \odot e\|_2 = O_p \left(\frac{d_m \sqrt{s}}{n} + \frac{d_m s}{n \sqrt{n}} \right).$$

No independence involving the estimated coefficients is used. Summing over modes and using $\sum_m d_m = s$ gives

$$\hat{B}_{NMT} - B_{NMT} = O_p \left(\frac{s}{n} + \frac{s^{3/2}}{n} + \frac{s^2}{n^{3/2}} \right).$$

Because $s/n = \rho^2$, $s^2/n^{3/2} = \rho s^{3/2}/n$, and $\rho \rightarrow 0$, this proves

$$\hat{B}_{NMT} - B_{NMT} = O_p \left(\rho^2 + \frac{s^{3/2}}{n} \right) = O_p \left(\rho + \frac{s^{3/2}}{n} \right). \quad (123)$$

Step 6: The bias-corrected expansion. Equation (118) and $\|X_{(4)}\|_{\text{op}}^2/n = O_p(1)$ imply

$$\|\hat{\Gamma}_{NMT} - \Gamma_{NMT}\|_{\text{op}} \leq \frac{\|X_{(4)}\|_{\text{op}}^2}{n} \|P_{\hat{\mathcal{T}}} - P_{\mathcal{T}}\|_{\text{op}} = O_p(\rho).$$

The lower eigenvalue bound established in Step 1 therefore gives

$$\|\hat{\Gamma}_{NMT}^{-1}\|_{\text{op}} = O_p(1), \quad \|\Gamma_{NMT} \hat{\Gamma}_{NMT}^{-1} - I_p\|_{\text{op}} = O_p(\rho).$$

All inverses are well defined with probability approaching one; the estimator may be extended arbitrarily outside that event.

Let $R_{NMT} = O_p(\rho + s^{3/2}/n)$ denote the remainder in Proposition 1. Substituting the definition of $\hat{\delta}_{\text{bc}}$ into that expansion gives

$$\begin{aligned} \Gamma_{NMT} \sqrt{n}(\hat{\delta}_{\text{bc}} - \delta) &= \mathcal{Z}_{NMT} + R_{NMT} \\ &\quad + (\Gamma_{NMT} \hat{\Gamma}_{NMT}^{-1} - I_p) B_{NMT} \\ &\quad + \Gamma_{NMT} \hat{\Gamma}_{NMT}^{-1} (\hat{B}_{NMT} - B_{NMT}). \end{aligned}$$

Since $B_{NMT} = O_p(s/\sqrt{n})$,

$$(\Gamma_{NMT} \hat{\Gamma}_{NMT}^{-1} - I_p) B_{NMT} = O_p(\rho s/\sqrt{n}) = O_p(s^{3/2}/n).$$

The last term is $O_p(\rho + s^{3/2}/n)$ by (123). Consequently,

$$\Gamma_{NMT} \sqrt{NMT} (\hat{\delta}_{bc} - \delta) = \mathcal{Z}_{NMT} + O_p \left(\sqrt{\frac{G+M+T}{NMT}} + \frac{(G+M+T)^{3/2}}{NMT} \right),$$

which proves (17). $\square$

B Proofs for Section 6 (Selecting the Number of Interactive Components)

Proposition 4 (Consistency of the information criterion). *Suppose that Assumptions 1, 2, 3, and 7 hold, with the true-factor restrictions evaluated at rank ℓ_0 . Let $\ell_{\max}$ be fixed, with $1 \leq \ell_0 \leq \ell_{\max}$, and define $\hat{\ell}$ by (18). If the positive deterministic penalty satisfies*

$$g(G, N, M, T) \rightarrow 0, \quad \frac{NMT}{G+M+T} g(G, N, M, T) \rightarrow \infty, \quad (19)$$

then, along the asymptotic sequence in Assumption 1, we have

$$\mathbb{P}\{\hat{\ell} = \ell_0\} \rightarrow 1.$$

Proof. Let $g_n = g(G, N, M, T)$. Retain the notation

$$n = NMT, \quad s = G + M + T, \quad \rho = \sqrt{\frac{s}{n}}, \quad \mathbf{X} = \frac{X_{(4)}}{\sqrt{n}},$$

and define

$$e = \text{vec}(E), \quad V_E = \frac{\|E\|_F^2}{n}, \quad Q_D = D(D'D)^{-1/2}, \quad E^G = E \times_1 Q_D'.$$

Vector norms are Euclidean norms, matrix norms without subscripts are operator norms, and $\|E^G\|$ denotes the tensor spectral norm. The asymptotic framework gives

$$\rho^2 = \frac{G}{NMT} + \frac{1}{NT} + \frac{1}{NM} \rightarrow 0.$$

Assumption 2 and Lemma 4(i) imply

$$\|\mathbf{X}\|_F = \frac{\|X\|_F}{\sqrt{n}} = O_p(1), \quad \|\mathbf{X}e\| = \frac{\|X_{(4)}e\|}{\sqrt{n}} = O_p(1), \quad \|E^G\| = O_p(\sqrt{s}). \quad (124)$$

Only the conditional mean-zero and second-order covariance restrictions in Assumption 3 are used in this proof.

1. A uniform criterion bound. For an arbitrary candidate rank $0 \leq \ell \leq \ell_{\max}$ and an admissible pair $(\tilde{\delta}, \Pi)$, define

$$\gamma = \tilde{\delta} - \delta_0, \quad \Delta_{\Pi} = \Pi - \Pi_0, \quad R = X \times_4 \gamma' + \Delta_{\Pi}, \quad r = \frac{\|R\|_F}{\sqrt{n}}.$$

Here R is the displacement of the fitted conditional mean from the true conditional mean. Consequently,

$$Y - X \times_4 \tilde{\delta}' - \Pi = E - R.$$

For $\ell \geq 1$, choose any admissible representation of Π with candidate factors $(\tilde{U}, \tilde{V}, \tilde{W})$, and put

$$\mathcal{M}_{\tilde{U}, \tilde{V}} = M_{D\tilde{U} * \tilde{V}} \otimes M_{W_0}.$$

For $\ell = 0$, use $\mathcal{M}_{\tilde{U}, \tilde{V}} = I_{NM} \otimes M_{W_0}$. This is an orthogonal projector. It annihilates $\text{vec}(\Pi)$ because the candidate mode-3 matricization satisfies

$$\Pi_{(3)} = \tilde{W}(D\tilde{U} * \tilde{V})',$$

and it annihilates $\text{vec}(\Pi_0)$ because

$$\Pi_{0(3)} = W_0(DU_0 * V_0)'.$$

Therefore,

$$\mathcal{M}_{\tilde{U}, \tilde{V}} \text{vec}(R) = \mathcal{M}_{\tilde{U}, \tilde{V}} X'_{(4)} \gamma.$$

On the identification event in Assumption 7,

$$\begin{aligned} r^2 &\geq \frac{1}{n} \left\| \mathcal{M}_{\tilde{U}, \tilde{V}} \text{vec}(R) \right\|^2 \\ &= \gamma' \left[\mathbf{X} \mathcal{M}_{\tilde{U}, \tilde{V}} \mathbf{X}' \right] \gamma \\ &\geq c_{\text{rank}} \|\gamma\|^2. \end{aligned}$$

This inequality is uniform over all candidate ranks and admissible representations. Hence

$$\|\gamma\| \leq c_{\text{rank}}^{-1/2} r.$$

Also,

$$\begin{aligned} \frac{\|\Delta_{\Pi}\|_F}{\sqrt{n}} &\leq \frac{\|R\|_F}{\sqrt{n}} + \frac{\|X \times_4 \gamma'\|_F}{\sqrt{n}} \\ &= r + \|\mathbf{X}' \gamma\| \\ &\leq r + \|\mathbf{X}\|_F \|\gamma\|. \end{aligned}$$

Consequently,

$$\|\gamma\| \leq c_{\text{rank}}^{-1/2} r, \quad \frac{\|\Delta_{\Pi}\|_F}{\sqrt{n}} \leq a_n r, \quad a_n := 1 + c_{\text{rank}}^{-1/2} \|\mathbf{X}\|_F = O_p(1). \quad (125)$$

The same random bound a_n applies to every candidate.

We next establish a deterministic error-contraction bound. Let

$$L = \ell_{\max} + \ell_0.$$

The difference Δ_Π admits a grouped CP representation with at most $\ell + \ell_0 \leq L$ terms. We claim that

$$|\langle E, \Delta_\Pi \rangle_F| \leq L \|E^G\| \|\Delta_\Pi\|_F. \quad (126)$$

To verify this, define the grouped tensor difference

$$\Delta_\Pi^G = \Delta_\Pi \times_1 Q'_D.$$

Because the first-mode fibers of Δ_Π belong to $\text{col}(D) = \text{col}(Q_D)$ and $Q'_D Q_D = I_G$,

$$\langle E, \Delta_\Pi \rangle_F = \langle E^G, \Delta_\Pi^G \rangle_F, \quad \|\Delta_\Pi^G\|_F = \|\Delta_\Pi\|_F.$$

For $\Delta_\Pi \neq 0$, choose a representation of Δ_Π^G with at most L terms. Let

$$O_1 \in \mathbb{R}^{G \times r_1}, \quad O_2 \in \mathbb{R}^{M \times r_2}$$

be orthonormal bases of the spans of its first-mode and second-mode factor vectors. Then

$$O_1' O_1 = I_{r_1}, \quad O_2' O_2 = I_{r_2}, \quad r_1, r_2 \leq L,$$

and

$$(\Delta_\Pi^G)_{(3)} = (\Delta_\Pi^G)_{(3)} (O_1 \otimes O_2) (O_1 \otimes O_2)'. \quad (127)$$

If $o_{1,a}$ and $o_{2,b}$ are columns of O_1 and O_2 , respectively, then

$$\begin{aligned} \left\| E_{(3)}^G(o_{1,a} \otimes o_{2,b}) \right\| &= \sup_{\substack{w \in \mathbb{R}^T \\ \|w\|=1}} |\langle E^G, o_{1,a} \circ o_{2,b} \circ w \rangle_F| \\ &\leq \|E^G\|. \end{aligned}$$

Since there are $r_1 r_2$ such columns,

$$\|E_{(3)}^G(O_1 \otimes O_2)\|_F \leq \sqrt{r_1 r_2} \|E^G\|.$$

Using the displayed support identity and Cauchy-Schwarz,

$$\begin{aligned} |\langle E, \Delta_\Pi \rangle_F| &= \left| \left\langle E_{(3)}^G(O_1 \otimes O_2), (\Delta_\Pi^G)_{(3)}(O_1 \otimes O_2) \right\rangle_F \right| \\ &\leq \|E_{(3)}^G(O_1 \otimes O_2)\|_F \|(\Delta_\Pi^G)_{(3)}(O_1 \otimes O_2)\|_F \\ &\leq \sqrt{r_1 r_2} \|E^G\| \|\Delta_\Pi^G\|_F \\ &\leq L \|E^G\| \|\Delta_\Pi\|_F. \end{aligned}$$

The bound is immediate when $\Delta_\Pi = 0$. Thus (126) holds for every realization and every admissible candidate. In particular, it applies to data-dependent candidate factors without conditioning on them, and it does not require candidate factor matrices to satisfy strong-factor lower bounds.

Since $Y = X \times_4 \delta'_0 + \Pi_0 + E$, equations (125) and (126) give

$$\begin{aligned}
\frac{1}{n} \|Y - X \times_4 \tilde{\delta}' - \Pi\|_F^2 - V_E &= r^2 - \frac{2}{n} \langle E, R \rangle_F \\
&\geq r^2 - \frac{2}{\sqrt{n}} \|\gamma\| \|Xe\| - \frac{2L}{n} \|E^G\| \|\Delta_\Pi\|_F \\
&\geq r^2 - b_n r \\
&= \left(r - \frac{b_n}{2}\right)^2 - \frac{b_n^2}{4} \\
&\geq -\frac{1}{4} b_n^2,
\end{aligned} \tag{127}$$

where

$$b_n = \frac{2c_{\text{rank}}^{-1/2}}{\sqrt{n}} \|Xe\| + \frac{2La_n}{\sqrt{n}} \|E^G\|.$$

By (124), fixed L , and $n^{-1/2} \leq \rho$,

$$b_n = O_p(\rho), \quad b_n \geq 0, \quad b_n = o_p(1).$$

The criterion inequality holds simultaneously over all candidate ranks and admissible parameter values on the identification event.

2. Overfitted ranks. For every $\ell \geq \ell_0$, the true common component Π_0 belongs to $\mathcal{C}_{\ell, c_U}^D$ by zero padding. Evaluating the criterion at (δ_0, Π_0) therefore gives $\mathcal{V}(\ell) \leq V_E$. Together with (127),

$$V_E - \frac{1}{4} b_n^2 \leq \mathcal{V}(\ell) \leq V_E, \quad \ell \geq \ell_0,$$

simultaneously on an event whose probability tends to one.

The candidate classes are nested, so $\mathcal{V}(\ell) \leq \mathcal{V}(\ell_0)$ when $\ell > \ell_0$. Consequently,

$$0 \leq \mathcal{V}(\ell_0) - \mathcal{V}(\ell) \leq \frac{1}{4} b_n^2 = O_p(s/n), \quad \ell > \ell_0.$$

It follows that

$$PC(\ell) - PC(\ell_0) \geq (\ell - \ell_0)g_n - \frac{1}{4} b_n^2.$$

The penalty condition gives

$$\frac{b_n^2}{g_n} = O_p\left(\frac{s/n}{g_n}\right) = o_p(1).$$

Therefore,

$$PC(\ell) > PC(\ell_0)$$

with probability approaching one, simultaneously over all overfitted candidate ranks.

3. Underfitted ranks. On the strong-factor event,

$$\lambda_{\min} \left(\frac{1}{N} U_0' D' D U_0 \right) \geq c_0, \quad \lambda_{\min} \left(\frac{1}{M} V_0' V_0 \right) \geq c_0, \quad \lambda_{\min} \left(\frac{1}{T} W_0' W_0 \right) \geq c_0.$$

For every $z \in \mathbb{R}^{\ell_0}$,

$$\begin{aligned} \frac{1}{NM} \|(DU_0 * V_0)z\|^2 &= \left\| \frac{DU_0}{\sqrt{N}} \text{diag}(z) \left(\frac{V_0}{\sqrt{M}} \right)' \right\|_F^2 \\ &\geq \lambda_{\min} \left(\frac{1}{N} U_0' D' D U_0 \right) \lambda_{\min} \left(\frac{1}{M} V_0' V_0 \right) \|\text{diag}(z)\|_F^2 \\ &\geq c_0^2 \|z\|^2. \end{aligned}$$

Hence

$$\sigma_{\min}(DU_0 * V_0) \geq c_0 \sqrt{NM}, \quad \sigma_{\min}(W_0) \geq \sqrt{c_0 T}.$$

Since

$$\Pi_{0,(3)} = W_0(DU_0 * V_0)',$$

the smallest nonzero singular value satisfies

$$\sigma_{\ell_0}(\Pi_{0,(3)}) \geq \sigma_{\min}(W_0) \sigma_{\min}(DU_0 * V_0) \geq c_0^{3/2} \sqrt{n}.$$

Every candidate $\Pi \in \mathcal{C}_{\ell, c_V}^D$ has $\text{rank}(\Pi_{(3)}) \leq \ell$. The matrix rank-approximation inequality therefore gives, for $\ell < \ell_0$,

$$\begin{aligned} \frac{1}{n} \|\Pi - \Pi_0\|_F^2 &= \frac{1}{n} \|\Pi_{(3)} - \Pi_{0,(3)}\|_F^2 \\ &\geq \frac{1}{n} \sum_{k=\ell+1}^{\ell_0} \sigma_k(\Pi_{0,(3)})^2 \\ &\geq (\ell_0 - \ell) c_0^3. \end{aligned} \tag{128}$$

Combining (125) and (128), every admissible candidate at rank $\ell < \ell_0$ satisfies

$$r \geq \underline{r}_{\ell, n}, \quad \underline{r}_{\ell, n} := \frac{c_0^{3/2} \sqrt{\ell_0 - \ell}}{a_n}.$$

Because $a_n = O_p(1)$ and $b_n = o_p(1)$,

$$a_n b_n = o_p(1).$$

In particular, with probability approaching one,

$$b_n \leq \frac{1}{2} \underline{r}_{\ell, n}$$

simultaneously for every $\ell < \ell_0$. On this event, $r \geq \underline{r}_{\ell, n}$ implies $b_n \leq r/2$, and (127) gives

$$\begin{aligned} \frac{1}{n} \|Y - X \times_4 \tilde{\delta}' - \Pi\|_F^2 - V_E &\geq r^2 - b_n r \\ &\geq \frac{1}{2} r^2 \\ &\geq \frac{(\ell_0 - \ell) c_0^3}{2 a_n^2}. \end{aligned}$$

Taking the minimum over all admissible candidates at rank ℓ , and using $\mathcal{V}(\ell_0) \leq V_E$, yields

$$\mathcal{V}(\ell) - \mathcal{V}(\ell_0) \geq \frac{(\ell_0 - \ell)c_0^3}{2a_n^2}, \quad \ell < \ell_0,$$

simultaneously with probability approaching one.

Therefore,

$$\begin{aligned} PC(\ell) - PC(\ell_0) &= \mathcal{V}(\ell) - \mathcal{V}(\ell_0) - (\ell_0 - \ell)g_n \\ &\geq (\ell_0 - \ell) \left(\frac{c_0^3}{2a_n^2} - g_n \right). \end{aligned}$$

The lower bound may be random, but its positivity follows from

$$a_n^2 g_n = o_p(1),$$

because $a_n = O_p(1)$ and $g_n \rightarrow 0$. Thus

$$PC(\ell) > PC(\ell_0)$$

with probability approaching one, simultaneously over all underfitted candidate ranks.

Combining the underfitting and overfitting comparisons over the finite candidate set gives

$$\mathbb{P}\{PC(\ell) > PC(\ell_0) \text{ for every } \ell \neq \ell_0, 0 \leq \ell \leq \ell_{\max}\} \rightarrow 1.$$

On this event, ℓ_0 is the unique minimizer of the penalized criterion. Hence

$$\mathbb{P}\{\hat{\ell} = \ell_0\} \rightarrow 1.$$

□

C Proofs for Section 7 (Inclusion of Additive Fixed Effects)

Lemma 2 (Consistency with group, teacher, and period additive effects). *Suppose Assumptions 1, 2, 3, and 8 hold. Both the true and candidate additive effects satisfy the observation-weighted normalizations in (20). Define the global least-squares estimator by (21). Then the slope and the interactive component satisfy*

$$\|\hat{\delta} - \delta\| + \frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{n}} = O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right). \quad (24)$$

The uncentered joint loading spaces satisfy

$$\begin{aligned} \frac{1}{\sqrt{MT}} \|M_{\hat{V} * \hat{W}}(V * W)\|_F &= O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right), \\ \frac{1}{\sqrt{NT}} \|M_{D\hat{U} * \hat{W}}(DU * W)\|_F &= O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right), \\ \frac{1}{\sqrt{NM}} \|M_{D\hat{U} * \hat{V}}(DU * V)\|_F &= O_p \left(\sqrt{\frac{G + M + T}{NMT}} \right). \end{aligned} \quad (25)$$

Moreover, the additive-effect estimators satisfy

$$\begin{aligned}
|\hat{\kappa} - \kappa| &= O_p \left(\frac{1}{\sqrt{NMT}} + \sqrt{\frac{G+M+T}{NMT}} \right), \\
\frac{1}{\sqrt{N}} \|D(\hat{\mu} - \mu)\| &= \left[\frac{1}{N} \sum_{g=1}^G N_g (\hat{\mu}_g - \mu_g)^2 \right]^{1/2} = O_p \left(\sqrt{\frac{G}{NMT}} + \sqrt{\frac{G+M+T}{NMT}} \right), \\
\frac{1}{\sqrt{M}} \|\hat{\nu} - \nu\| &= O_p \left(\frac{1}{\sqrt{NT}} + \sqrt{\frac{G+M+T}{NMT}} \right), \\
\frac{1}{\sqrt{T}} \|\hat{\xi} - \xi\| &= O_p \left(\frac{1}{\sqrt{NM}} + \sqrt{\frac{G+M+T}{NMT}} \right).
\end{aligned} \tag{26}$$

Proof. Write $n = NMT$, $s = G + M + T$, and $\rho = \sqrt{s/n}$. Assumption 1 gives $\rho \rightarrow 0$. Let $\mathcal{A}$ and $\hat{\mathcal{A}}$ denote the true and estimated additive-effect tensors. All stochastic bounds involving conditional covariances are understood on the events in Assumption 3, whose probabilities tend to one, with localization for random coefficient norms. Only the conditional mean-zero and second-order covariance restrictions in that assumption are needed below.

The conditional regressor-moment bound implies $\|X\|_F = O_p(\sqrt{n})$. Let $e = \text{vec}(E)$. Assumption 3 gives $\|\mathbb{E}[ee' \mid \mathcal{F}]\| \leq C_\varepsilon$ with probability approaching one. Therefore,

$$\mathbb{E}[\|X_{(4)} M_{\mathcal{A}_D} e\|^2 \mid \mathcal{F}] \leq C_\varepsilon \|X_{(4)} M_{\mathcal{A}_D}\|_F^2 \leq C_\varepsilon \|X\|_F^2.$$

Conditional Markov's inequality and Assumption 2 yield

$$\frac{\|X\|_F}{\sqrt{n}} = O_p(1), \quad \|X_{(4)} M_{\mathcal{A}_D} e\| = O_p(\sqrt{n}), \quad \|E^G\| = O_p(\sqrt{s}). \tag{129}$$

1. The slope and additive-within common component. For any factor triple,

$$\begin{aligned}
\Pi(U, V, W)^{\mathcal{A}_D} &= \sum_{k=1}^{\ell} \left[Du_k^c \circ v_k^c \circ w_k^c + \bar{u}_k \iota_N \circ v_k^c \circ w_k^c \right. \\
&\quad \left. + \bar{v}_k Du_k^c \circ \iota_M \circ w_k^c + \bar{w}_k Du_k^c \circ v_k^c \circ \iota_T \right].
\end{aligned} \tag{130}$$

Each first-mode vector belongs to $\text{col}(D)$. Hence every transformed common component has a grouped CP representation with at most 4ℓ terms, and the difference of two such components has at most 8ℓ terms.

For any grouped tensor Z with a representation of at most r terms,

$$|\langle E, Z \rangle| \leq r \|E^G\| \|Z\|_F. \tag{131}$$

Indeed, let $Z^G = Z \times_1 Q'_D$, where $Q_D = D(D'D)^{-1/2}$. Then $\langle E, Z \rangle = \langle E^G, Z^G \rangle$ and $\|Z^G\|_F = \|Z\|_F$. Choose orthonormal bases B_1, B_2 for the spans of the first- and second-mode vectors in a representation of Z^G with at most r terms. Their column counts r_1, r_2 are at most r , and

$$Z_{(3)}^G = Z_{(3)}^G (B_1 \otimes B_2) (B_1 \otimes B_2)'.$$

Each column of $E_{(3)}^G(B_1 \otimes B_2)$ has Euclidean norm at most $\|E^G\|$. Frobenius Cauchy–Schwarz therefore proves (131), with $\sqrt{r_1 r_2} \leq r$. The bound is pathwise and applies to data-dependent tensors.

Set

$$\begin{aligned} \gamma &= \hat{\delta} - \delta_0, & \Delta &= \hat{\Pi} - \Pi_0, & \Delta^{\mathcal{A}_D} &= \hat{\Pi}^{\mathcal{A}_D} - \Pi_0^{\mathcal{A}_D}, \\ R_{\mathcal{A}_D} &= M_{\mathcal{A}_D} \{X'_{(4)} \gamma + \text{vec}(\Delta)\}, & r_n &= \frac{\|R_{\mathcal{A}_D}\|}{\sqrt{n}}. \end{aligned}$$

Global optimality of the profiled estimator

$$(\hat{\delta}, \hat{U}, \hat{V}, \hat{W}) \in \underset{\substack{\delta \in \mathbb{R}^P \\ (U, V, W) \in \Theta_\ell}}{\text{argmin}} \frac{1}{n} \left\| M_{\mathcal{A}_D} \{y - X'_{(4)} \delta - \pi(U, V, W)\} \right\|^2.$$

implies $\|M_{\mathcal{A}_D} e - R_{\mathcal{A}_D}\|^2 \leq \|M_{\mathcal{A}_D} e\|^2$ so that

$$\begin{aligned} \|R_{\mathcal{A}_D}\|^2 &\leq 2 \langle M_{\mathcal{A}_D} e, R_{\mathcal{A}_D} \rangle \\ &\leq 2 \|\gamma\| \|X_{(4)} M_{\mathcal{A}_D} e\| + 2 |\langle E, \Delta^{\mathcal{A}_D} \rangle| \\ &\leq 2 \|\gamma\| \|X_{(4)} M_{\mathcal{A}_D} e\| + 16 \ell \|E^G\| \|\Delta^{\mathcal{A}_D}\|_F. \end{aligned} \tag{132}$$

The last line follows from (130)–(131).

Let

$$\mathcal{L}_{\hat{U}, \hat{V}} = M_{[\iota_N \otimes \iota_M, D\hat{U} * \hat{V}]} \otimes M_{[\iota_T, W_0]}.$$

This orthogonal projector annihilates $\mathcal{A}_D$, $\text{vec}(\Pi_0)$, and $\text{vec}(\hat{\Pi})$. To see this, note that the column $\iota_N \otimes \iota_M$ in C_A removes arbitrary year effects and the column ι_T in $W_A = [\iota_T, W]$ removes group and teacher effects, which are constant over time.

Thus $\mathcal{L}_{\hat{U}, \hat{V}} M_{\mathcal{A}_D} = \mathcal{L}_{\hat{U}, \hat{V}}$, and the uniform identification condition gives

$$r_n^2 \geq \frac{1}{n} \|\mathcal{L}_{\hat{U}, \hat{V}} R_A\|^2 = \gamma' \left[\frac{1}{n} X_{(4)} \mathcal{L}_{\hat{U}, \hat{V}} X'_{(4)} \right] \gamma \geq c_{\delta, A} \|\gamma\|^2$$

with probability approaching one. Consequently,

$$\|\gamma\| \leq c_{\delta, A}^{-1/2} r_n, \quad \frac{\|\Delta^{\mathcal{A}_D}\|_F}{\sqrt{n}} \leq a_n r_n, \quad a_n = 1 + c_{\delta, A}^{-1/2} \frac{\|X\|_F}{\sqrt{n}} = O_p(1).$$

Dividing (132) by n gives $r_n^2 \leq b_n r_n$, where

$$b_n = \frac{2c_{\delta, A}^{-1/2}}{n} \|X_{(4)} M_{\mathcal{A}_D} e\| + 16 \ell a_n \frac{\|E^G\|}{\sqrt{n}} = O_p(\rho).$$

Since $r_n, b_n \geq 0$, this implies $r_n \leq b_n$ on the identification event. Therefore,

$$\|\hat{\delta} - \delta_0\| + \frac{\|\hat{\Pi}^{\mathcal{A}_D} - \Pi_0^{\mathcal{A}_D}\|_F}{\sqrt{n}} = O_p(\rho). \tag{133}$$

2. Recovering the untransformed interactive component. We recover the full component from its fully centered tensor and its three pairwise contractions.

Let $(d_1, d_2, d_3) = (G, M, T)$ and define the unit constant directions in grouped coordinates by

$$q_1 = \frac{(D'D)^{1/2} \iota_G}{\sqrt{N}}, \quad q_2 = \frac{\iota_M}{\sqrt{M}}, \quad q_3 = \frac{\iota_T}{\sqrt{T}}, \quad M_{q_m} = I_{d_m} - q_m q_m'.$$

Introduce normalized true factors and their centered parts,

$$A_1 = \frac{(D'D)^{1/2} U_0}{\sqrt{N}}, \quad A_2 = \frac{V_0}{\sqrt{M}}, \quad A_3 = \frac{W_0}{\sqrt{T}}, \quad B_m = M_{q_m} A_m, \quad \alpha_m = A_m' q_m.$$

Then $A_m = B_m + q_m \alpha_m'$. Use hats for the estimated counterparts. In particular,

$$B_1 = \frac{(D'D)^{1/2} U_0^c}{\sqrt{N}}, \quad B_2 = \frac{V_0^c}{\sqrt{M}}, \quad B_3 = \frac{W_0^c}{\sqrt{T}},$$

so the centered-strength condition gives $c_c I_\ell \leq B_m' B_m \leq c_1 I_\ell$ with probability approaching one.

Since $A_m' A_m = B_m' B_m + \alpha_m \alpha_m'$, Assumptions 2 and 8(i) imply

$$c_c I_\ell \leq B_m' B_m \leq A_m' A_m \leq c_1 I_\ell, \quad m = 1, 2, 3, \quad (134)$$

with probability approaching one. Below, $\lesssim$ suppresses constants depending only on ℓ, c_c, c_1, c_U .

Define the normalized grouped tensors and their difference by

$$\mathbf{C} = \frac{1}{\sqrt{n}} \Pi_0 \times_1 Q_D', \quad \hat{\mathbf{C}} = \frac{1}{\sqrt{n}} \hat{\Pi} \times_1 Q_D', \quad z = \text{vec}(\hat{\mathbf{C}} - \mathbf{C}).$$

Thus $\mathbf{C} = \sum_{k=1}^\ell a_{1,k} \circ a_{2,k} \circ a_{3,k}$, with the analogous representation for $\hat{\mathbf{C}}$. Let

$$Q = Q_D \otimes I_M \otimes I_T, \quad M_{\mathcal{A}_D}^G = Q' M_{\mathcal{A}_D} Q.$$

Since $Q'Q = I_{GMT}$ and the additive space is contained in $\text{col}(Q)$, we have $M_{\mathcal{A}_D} Q = Q M_{\mathcal{A}_D}^G$, and $M_{\mathcal{A}_D}^G$ is the additive residual-maker in grouped coordinates. Consequently,

$$\|z\| = \frac{\|\Delta\|_F}{\sqrt{n}}, \quad \eta_n := \|M_{\mathcal{A}_D}^G z\| = \frac{\|\Delta^{\mathcal{A}_D}\|_F}{\sqrt{n}} = O_p(\rho) = o_p(1).$$

We first recover the raw loading spaces. Define

$$\mathbf{S}_1 = \mathbf{C} \times_2 M_{q_2} \times_3 M_{q_3}, \quad \hat{\mathbf{S}}_1 = \hat{\mathbf{C}} \times_2 M_{q_2} \times_3 M_{q_3}.$$

Every grouped additive term has at least two modes proportional to the corresponding q_m . Hence centering in modes two and three annihilates the grouped additive space, giving

$$(I_G \otimes M_{q_2} \otimes M_{q_3}) M_{\mathcal{A}_D}^G = I_G \otimes M_{q_2} \otimes M_{q_3}.$$

It follows that

$$\|\hat{\mathbf{S}}_1 - \mathbf{S}_1\|_F \leq \eta_n, \quad (\mathbf{S}_1)_{(1)} = A_1 (B_2 * B_3)'.$$

The Schur-product identity and (134) imply

$$(B_2 * B_3)'(B_2 * B_3) = (B_2' B_2) \odot (B_3' B_3) \geq c_c^2 I_\ell,$$

and therefore $\sigma_\ell((S_1)_{(1)}) \geq c_c^{3/2}$. Singular-value perturbation gives

$$\sigma_\ell((\hat{S}_1)_{(1)}) \geq c_c^{3/2} - \eta_n > 0$$

with probability approaching one. Since $(\hat{S}_1)_{(1)} = \hat{A}_1(\hat{B}_2 * \hat{B}_3)'$ has rank at most ℓ , it has rank exactly ℓ , as does $\hat{A}_1$. Moreover,

$$c_c \|M_{\hat{A}_1} A_1\|_F \leq \|M_{\hat{A}_1} (S_1)_{(1)}\|_F = \|M_{\hat{A}_1} \{(S_1)_{(1)} - (\hat{S}_1)_{(1)}\}\|_F \leq \eta_n.$$

Repeating this argument cyclically yields

$$\text{rank}(\hat{A}_m) = \ell, \quad \|M_{\hat{A}_m} A_m\|_F \leq c_c^{-1} \eta_n, \quad m = 1, 2, 3,$$

with probability approaching one.

Let $O_m = A_m(A_m' A_m)^{-1/2}$, so that $O_m' O_m = I_\ell$ and $P_{A_m} = O_m O_m'$. Equality of the true and estimated loading-space dimensions gives

$$\|P_{\hat{A}_m} - P_{A_m}\|_F^2 = 2\|M_{\hat{A}_m} O_m\|_F^2.$$

Thus, by (134),

$$\|P_{\hat{A}_m} - P_{A_m}\| \leq \sqrt{2} \|M_{\hat{A}_m} O_m\|_F \leq \sqrt{2} c_c^{-3/2} \eta_n, \quad m = 1, 2, 3.$$

Next, define

$$\mathcal{P} = P_{A_1} \otimes P_{A_2} \otimes P_{A_3}, \quad \hat{\mathcal{P}} = P_{\hat{A}_1} \otimes P_{\hat{A}_2} \otimes P_{\hat{A}_3}, \quad \mathcal{J} = M_{q_1} \otimes M_{q_2} \otimes M_{q_3}.$$

A telescoping expansion of the Kronecker products gives

$$\|\hat{\mathcal{P}} - \mathcal{P}\| \leq \sum_{m=1}^3 \|P_{\hat{A}_m} - P_{A_m}\| \lesssim \eta_n.$$

The parameter-space restrictions imply

$$\|\hat{\mathbf{C}}\|_F \leq \sum_{k=1}^{\ell} \|\hat{a}_{1,k}\| \|\hat{a}_{2,k}\| \|\hat{a}_{3,k}\| \leq \ell \sqrt{C_U}.$$

Since $\mathcal{P} \text{vec}(\mathbf{C}) = \text{vec}(\mathbf{C})$ and $\hat{\mathcal{P}} \text{vec}(\hat{\mathbf{C}}) = \text{vec}(\hat{\mathbf{C}})$,

$$\begin{aligned} \|(I_{GMT} - \mathcal{P})z\| &= \|(I_{GMT} - \mathcal{P})(\hat{\mathcal{P}} - \mathcal{P}) \text{vec}(\hat{\mathbf{C}})\| \\ &\leq \|\hat{\mathcal{P}} - \mathcal{P}\| \|\hat{\mathbf{C}}\|_F \lesssim \eta_n. \end{aligned} \tag{135}$$

To control the remaining component $\|\mathcal{P}z\|$, observe that

$$O_m' M_{q_m} O_m = (A_m' A_m)^{-1/2} B_m' B_m (A_m' A_m)^{-1/2} \geq \frac{c_c}{c_1} I_\ell.$$

Let $O = O_1 \otimes O_2 \otimes O_3$. Then $O'O = I_{\ell^3}$, $\mathcal{P} = OO'$, and

$$(\mathcal{J}O)'(\mathcal{J}O) = \bigotimes_{m=1}^3 (O'_m M_{q_m} O_m) \geq \left(\frac{c_c}{c_1}\right)^3 I_{\ell^3}.$$

Consequently,

$$\|\mathcal{J}v\| \geq \left(\frac{c_c}{c_1}\right)^{3/2} \|v\|, \quad v \in \text{range}(\mathcal{P}).$$

Full centering annihilates the grouped additive space, so $\mathcal{J}M_{\mathcal{A}_D}^G = \mathcal{J}$ and $\|\mathcal{J}z\| \leq \|M_{\mathcal{A}_D}^G z\| = \eta_n$. Therefore, using (135),

$$\|\mathcal{P}z\| \leq \left(\frac{c_1}{c_c}\right)^{3/2} \|\mathcal{J}\mathcal{P}z\| \leq \left(\frac{c_1}{c_c}\right)^{3/2} \{\eta_n + \|(I_{GMT} - \mathcal{P})z\|\} \lesssim \eta_n.$$

where we've used

$$\|\mathcal{J}\mathcal{P}z\| = \|\mathcal{J}(z - (I_{GMT} - \mathcal{P})z)\| = \|\mathcal{J}z - \mathcal{J}(I_{GMT} - \mathcal{P})z\| \leq \|\mathcal{J}z\| + \|\mathcal{J}(I_{GMT} - \mathcal{P})z\|.$$

Together with (135), this proves

$$\frac{\|\hat{\Pi} - \Pi_0\|_F}{\sqrt{n}} = \|z\| \leq \|\mathcal{P}z\| + \|(I_{GMT} - \mathcal{P})z\| \lesssim \eta_n = O_p(\rho). \quad (136)$$

Combining (133) and (136) establishes (24).

3. The uncentered joint loading spaces. The first-mode matricizations satisfy

$$(\Pi_{0,(1)} - \hat{\Pi}_{(1)})M_{\hat{V}*\hat{W}} = DU_0(V_0 * W_0)'M_{\hat{V}*\hat{W}}.$$

Since $U_0'D'DU_0 \geq Nc_cI_\ell$, it follows that

$$\sqrt{Nc_c}\|M_{\hat{V}*\hat{W}}(V_0 * W_0)\|_F \leq \|(\Pi_{0,(1)} - \hat{\Pi}_{(1)})M_{\hat{V}*\hat{W}}\|_F \leq \|\Delta\|_F.$$

Hence

$$\frac{\|M_{\hat{V}*\hat{W}}(V_0 * W_0)\|_F}{\sqrt{MT}} \leq c_c^{-1/2} \frac{\|\Delta\|_F}{\sqrt{n}} = O_p(\rho).$$

The other two bounds in (25) follow by cyclic permutation.

4. The additive effects. Let $\Delta_a = \text{vec}(\hat{\mathcal{A}} - \mathcal{A})$. The additive profiling identity gives

$$\Delta_a = P_{\mathcal{A}_D}\{e - X'_{(4)}\gamma - \text{vec}(\Delta)\}.$$

Denote the four mutually orthogonal summands of $P_{\mathcal{A}_D}$ by $P_\kappa, P_\mu, P_\nu, P_\xi$, with respective ranks

$$r_\kappa = 1, \quad r_\mu = G - 1, \quad r_\nu = M - 1, \quad r_\xi = T - 1.$$

These projectors are $\mathcal{F}$ -measurable. The conditional covariance bound implies

$$\mathbb{E}[\|P_b e\|^2 \mid \mathcal{F}] \leq C_\varepsilon r_b, \quad b \in \{\kappa, \mu, \nu, \xi\}.$$

By conditional Markov's inequality, (129), (133), and (136),

$$\begin{aligned}\frac{\|P_b \Delta_a\|}{\sqrt{n}} &\leq \frac{\|P_b e\|}{\sqrt{n}} + \frac{\|X\|_F}{\sqrt{n}} \|\gamma\| + \frac{\|\Delta\|_F}{\sqrt{n}} \\ &= O_p\left(\sqrt{\frac{r_b}{n}} + \rho\right), \quad b \in \{\kappa, \mu, \nu, \xi\}.\end{aligned}$$

The observation-weighted additive normalizations give

$$\begin{aligned}\frac{\|P_\kappa \Delta_a\|}{\sqrt{n}} &= |\hat{\kappa} - \kappa|, & \frac{\|P_\mu \Delta_a\|}{\sqrt{n}} &= \frac{\|D(\hat{\mu} - \mu)\|}{\sqrt{N}}, \\ \frac{\|P_\nu \Delta_a\|}{\sqrt{n}} &= \frac{\|\hat{\nu} - \nu\|}{\sqrt{M}}, & \frac{\|P_\xi \Delta_a\|}{\sqrt{n}} &= \frac{\|\hat{\xi} - \xi\|}{\sqrt{T}}.\end{aligned}$$

Substitution of the four ranks proves (26). $\square$

D Proofs for Section 11 (Appendix)

Lemma 3 (Convergence with an unbalanced sample). *Maintain the model, membership conditions, normalized parameter space, and true-parameter feasibility specified above. Let p, ℓ be fixed, and suppose that*

$$\rho_\Omega = \sqrt{\frac{G + M + T}{n_\Omega}}, \quad \rho_\Omega \rightarrow_p 0. \quad (41)$$

Let $(\hat{\delta}, \hat{U}, \hat{V}, \hat{W})$ be a global minimizer in (37), and suppose that Assumptions 13 and 14 hold.

(i) *Observed-sample convergence. The slope and observed interactive component satisfy*

$$\|\hat{\delta} - \delta\| + \frac{\|R_\Omega(\hat{\pi} - \pi)\|}{\sqrt{n_\Omega}} = O_p(\rho_\Omega). \quad (42)$$

(ii) *Full-component convergence. If Assumption 15 also holds, then*

$$\frac{\|\hat{\Pi} - \Pi\|_F}{\sqrt{NMT}} = O_p(\rho_\Omega). \quad (43)$$

(iii) *Full joint loading-space convergence. Suppose Assumption 15 holds and, for a constant $c_f > 0$ independent of the panel dimensions, with probability approaching one,*

$$\frac{1}{N} U' D' D U \geq c_f I_\ell, \quad \frac{1}{M} V' V \geq c_f I_\ell, \quad \frac{1}{T} W' W \geq c_f I_\ell. \quad (44)$$

Then

$$\begin{aligned}\frac{1}{\sqrt{MT}} \|M_{\hat{V} * \hat{W}}(V * W)\|_F &= O_p(\rho_\Omega), \\ \frac{1}{\sqrt{NT}} \|M_{D\hat{U} * \hat{W}}(DU * W)\|_F &= O_p(\rho_\Omega), \\ \frac{1}{\sqrt{NM}} \|M_{D\hat{U} * \hat{V}}(DU * V)\|_F &= O_p(\rho_\Omega).\end{aligned} \quad (45)$$

The observed-sample conclusion requires neither Assumption 15 nor the strong-factor bounds in (44). All conclusions are aggregate-norm statements, not uniform fixed-index consistency results.

Proof. In the proof, write $(\delta_0, U_0, V_0, W_0, \Pi_0)$ for the true parameters and common component denoted by (δ, U, V, W, Π) in the statement. Write $m = n_\Omega$, $n = NMT$, $s = G + M + T$, and $\rho_\Omega = \sqrt{s/m}$. All stochastic orders involving the random sample size are relative to its realized positive value.

1. The observed regressor score. On the conditional covariance event in Assumption 14,

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{X'_\Omega e_\Omega}{\sqrt{m}} \right\|^2 \middle| \mathcal{F}_\Omega \right] &= \frac{1}{m} \text{tr} (X'_\Omega \mathbb{E}[e_\Omega e'_\Omega | \mathcal{F}_\Omega] X_\Omega) \\ &\leq \frac{c_\varepsilon}{m} \|X_\Omega\|_F^2. \end{aligned}$$

Conditional Markov's inequality, the regressor bound in Assumption 13, and the probability of the covariance event imply

$$\left\| \frac{X'_\Omega e_\Omega}{\sqrt{m}} \right\| = O_p(1), \quad \frac{\|X_\Omega\|_F}{\sqrt{m}} = O_p(1).$$

For the first assertion, one may first restrict to $\{\|X_\Omega\|_F^2/m \leq A\}$ for a fixed A , apply the conditional probability bound there, and then make the complementary probability arbitrarily small by choosing A sufficiently large.

2. Identification and the basic inequality. Set

$$\gamma = \hat{\delta} - \delta_0, \quad \Delta = \hat{\Pi} - \Pi_0, \quad r_\Omega = X_\Omega \gamma + R_\Omega \text{vec}(\Delta), \quad t_\Omega = \frac{\|r_\Omega\|}{\sqrt{m}}, \quad \hat{B}_\Omega = B_\Omega(\hat{U}, \hat{V}).$$

The candidate component lies in the first block of $\hat{B}_\Omega$, because

$$\hat{\pi}_\Omega = R_\Omega \{(D\hat{U} * \hat{V}) \otimes I_T\} \text{vec}(\hat{W}').$$

The true component lies in its second block. Specifically, define $b_0 \in \mathbb{R}^{GM\ell}$ by stacking the vectors $u_{0j} \odot v_{0j}$ in increasing group-teacher order. Then

$$\pi_{0,\Omega} = R_\Omega(D \otimes I_M \otimes W_0)b_0.$$

Hence $M_{\hat{B}_\Omega} R_\Omega \text{vec}(\Delta) = 0$. On the uniform identification event,

$$t_\Omega^2 \geq \frac{1}{m} \|M_{\hat{B}_\Omega} r_\Omega\|^2 = \gamma' \left(\frac{1}{m} X'_\Omega M_{\hat{B}_\Omega} X_\Omega \right) \gamma \geq c_{\delta,\Omega} \|\gamma\|^2.$$

Although $\hat{B}_\Omega$ depends on the errors, the inequality applies because the identification bound is uniform over all candidate matrices. It follows that

$$\|\gamma\| \leq c_{\delta,\Omega}^{-1/2} t_\Omega, \quad \frac{\|R_\Omega \text{vec}(\Delta)\|}{\sqrt{m}} \leq a_\Omega t_\Omega, \quad a_\Omega = 1 + c_{\delta,\Omega}^{-1/2} \frac{\|X_\Omega\|_F}{\sqrt{m}} = O_p(1). \quad (137)$$

The true parameters are feasible, so global least-squares optimality gives $\|e_\Omega - r_\Omega\|^2 \leq \|e_\Omega\|^2$. Moreover, $\Delta \in \mathcal{L}_{2\ell}^D$, and therefore

$$|e'_\Omega R_\Omega \text{vec}(\Delta)| \leq \|E\|_{\Omega, 2\ell} \|R_\Omega \text{vec}(\Delta)\|.$$

This inequality is also valid when $R_\Omega \text{vec}(\Delta) = 0$, in which case both sides are zero. Consequently,

$$\begin{aligned} mt_\Omega^2 &\leq 2e'_\Omega r_\Omega \\ &\leq 2\|\gamma\| \|X'_\Omega e_\Omega\| + 2\|E\|_{\Omega, 2\ell} \|R_\Omega \text{vec}(\Delta)\|. \end{aligned}$$

Using (137), divide by m to obtain $t_\Omega^2 \leq b_\Omega t_\Omega$, where

$$b_\Omega = \frac{2c_{\delta, \Omega}^{-1/2}}{m} \|X'_\Omega e_\Omega\| + \frac{2a_\Omega}{\sqrt{m}} \|E\|_{\Omega, 2\ell} = O_p\left(m^{-1/2} + \sqrt{\frac{s}{m}}\right) = O_p(\rho_\Omega).$$

Both t_Ω and b_Ω are nonnegative, so $t_\Omega \leq b_\Omega$ on the identification event. Thus $t_\Omega = O_p(\rho_\Omega)$, and (137) proves (42). The convergence to zero follows from (41).

3. Recovery of the full component. On the sampling-stability event, uniformity of (40) over the feasible parameter space permits evaluation at $(\hat{U}, \hat{V}, \hat{W})$ and gives

$$\frac{\|\Delta\|_F}{\sqrt{n}} \leq c_\Pi^{-1/2} \frac{\|R_\Omega \text{vec}(\Delta)\|}{\sqrt{m}} = O_p(\rho_\Omega).$$

This proves (43). No true-factor lower eigenvalue bound is used in this step.

4. The three joint loading spaces. For $h = 1, 2, 3$, write $\Pi_{0, (h)} = A_{h0} C'_{h0}$, where

$$(A_{h0}, C_{h0}, d_h) = \begin{cases} (DU_0, V_0 * W_0, N), & h = 1, \\ (V_0, DU_0 * W_0, M), & h = 2, \\ (W_0, DU_0 * V_0, T), & h = 3. \end{cases}$$

Let $\hat{A}_h, \hat{C}_h$ denote the corresponding estimated matrices. Since $\hat{\Pi}_{(h)} = \hat{A}_h \hat{C}'_h$, we have

$$\Delta_{(h)} M_{\hat{C}_h} = -A_{h0} C'_{h0} M_{\hat{C}_h}.$$

The Frobenius norm is unchanged by matricization and contracts under orthogonal projection. Hence

$$\|\Delta\|_F \geq \|A_{h0} C'_{h0} M_{\hat{C}_h}\|_F \geq \{\lambda_{\min}(A'_{h0} A_{h0})\}^{1/2} \|M_{\hat{C}_h} C_{h0}\|_F.$$

On the true-factor strength event, it follows that

$$\sqrt{\frac{d_h}{n}} \|M_{\hat{C}_h} C_{h0}\|_F \leq \left\{ \lambda_{\min} \left(\frac{A'_{h0} A_{h0}}{d_h} \right) \right\}^{-1/2} \frac{\|\Delta\|_F}{\sqrt{n}} \leq c_f^{-1/2} \frac{\|\Delta\|_F}{\sqrt{n}} = O_p(\rho_\Omega).$$

The factors $\sqrt{d_h/n}$ equal $1/\sqrt{MT}$, $1/\sqrt{NT}$, and $1/\sqrt{NM}$, respectively. This proves (45). The argument uses Moore–Penrose projectors and does not assume that the estimated loading matrices or observed slice Gram matrices are nonsingular. $\square$