JOINT LEARNING OF PANEL VAR MODELS WITH LOW RANK AND SPARSE STRUCTURE

BY YUCHEN XU[1,a] AND GEORGE MICHAILIDIS[1,b]

[1] *University of California, Los Angeles,* [a] *yuchenxu95@g.ucla.edu;* [b] *gmichail@g.ucla.edu*

Panel vector auto-regressive (VAR) models are widely used to capture the dynamics of multivariate time series across different subpopulations, where each subpopulation shares a common set of variables. In this work, we propose a panel VAR model with a shared low-rank structure, modulated by subpopulation-specific weights, and complemented by idiosyncratic sparse components. To ensure parameter identifiability, we impose structural constraints that lead to a nonsmooth, nonconvex optimization problem. We develop a multi-block Alternating Direction Method of Multipliers (ADMM) algorithm for parameter estimation and establish its convergence under mild regularity conditions. Furthermore, we derive consistency guarantees for the proposed estimators under high-dimensional scaling. The effectiveness of the proposed modeling framework and estimators is demonstrated through experiments on both synthetic data and a real-world neuroscience data set.

CONTENTS

Keywords and phrases: ADMM, joint model, low rank matrix, panel time series, sparse matrix, vector auto-regression.

1. Introduction. Vector autoregressions represent a popular modeling framework for capturing dynamic relationships between multivariate time series and have been extensively used in macroeconomics (Kilian and Lütkepohl, 2017), functional genomics (Michailidis and d'Alché Buc, 2013) and neuroscience/neuroimaging (Seth, Barrett and Barnett, 2015; Aslan and Ombao, 2025) applications. Their use in applications wherein the number of time series under consideration is large relative to the number of time points available, led to the introduction of regularized estimators for the VAR model parameters –e.g., assuming sparsity (Basu and Michailidis, 2015; Melnyk and Banerjee, 2016; Kock and Callot, 2015; Kastner and Huber, 2020; Medeiros and Mendes, 2016), group sparsity (Billio, Casarin and Rossini, 2019; Ghosh, Khare and Michailidis, 2019), or low rankness (Basu, Li and Michailidis, 2019; Wang, Zheng and Li, 2024)- and the investigation of their consistency properties under high dimensional scaling, including statistical inference (Krampe, Paparoditis and Trenkler, 2023).

In many of these application domains, it has become increasingly common to consider multivariate time series data from a *collection of entities*. For example, in macroeconomic applications, the entities may represent different countries, while in neuroimaging studies, they may correspond to individual subjects. The resulting data consist of p time series (variables), observed over T time periods, across M entities. A key challenge is to develop modeling strategies that capture both *common effects* shared across entities, as well as *entity-specific* heterogeneities, while simultaneously controlling the total number of parameters to ensure computational and statistical efficiency.

To effectively capture both shared structures and entity-specific variations in panel VAR models, several modeling strategies have been proposed in the literature. One approach involves imposing constraints on the entity-specific VAR parameters through structural assumptions (Canova and Ciccarelli, 2013), penalization techniques (Skripnikov and Michailidis, 2019), or prior distributions (Korobilis, 2016); further details on this line of work are provided in Appendix B in the Supplement. An alternative strategy is to represent the data as a time series of matrix-variate variables (Chen, Xiao and Yang, 2021) or as a three-dimensional tensor (Chen, Yang and Zhang, 2022), enabling the use of factor models or tensor decomposition techniques. A brief overview of these approaches is also provided in Appendix B in the Supplement.

In this paper, we propose a panel VAR-based modeling strategy that captures both shared structures across entity-specific VAR models and idiosyncratic heterogeneities. Specifically, we consider multivariate time series data comprising p variables observed over T time periods for M entities, denoted as $\{X_t^m \in \mathbb{R}^p; t = 1, \cdots, T,\ m = 1, \cdots, M\}$. For ease of presentation, we assume a common number of time periods T across entities, although the proposed model can be readily extended to unbalanced panels where each entity m has its own number of observations T_m. We posit the following model—referred to hereafter as **LSPVAR** (Low-rank & Sparse Panel VAR)—for a single lag for ease of exposition:

$$X_t^m = A_m X_{t-1}^m + \epsilon_t^m,\ \epsilon_t^m \sim N(0, \Sigma_m); \qquad A_m = W_m \Phi + S_m, \tag{1.1}$$

where $A_m \in \mathbb{R}^{p\times p}$ is the transition matrix of the VAR model for the m-th entity, and $\Sigma_m \in \mathbb{R}^{p\times p}$ is the corresponding covariance structure. Models with multiple lags can be handled

similarly by imposing a comparable structure on the coefficient matrices at higher lags; see the discussion in Appendix F in the Supplement. The transition matrices A_m are assumed to exhibit a *low rank plus sparse* structure. Specifically, we model each A_m as a rescaled variant of a *common low rank basis* Φ, modulated by a diagonal matrix of entity-specific weights $W_m \in \mathbb{R}^{p\times p}$, along with an *entity-specific, sparse* component S_m. The shared low-rank component Φ efficiently captures global dependencies across variables, while keeping model complexity low, facilitating scalable estimation in high-dimensional settings. A visual depiction of the transition matrices A_m under the posited model is provided in Figure 1.

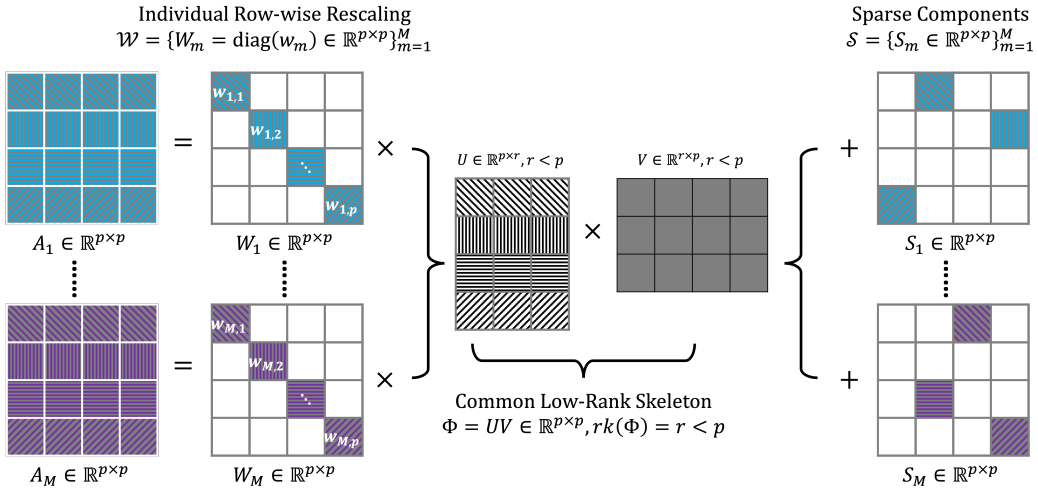

FIG 1. *Illustration of the model setup as specified in* (1.1). *The lead-lag relationships among variables are encoded in the transition matrices A_m. Shared global structures are captured by the common low-rank basis Φ, while entity-specific variations are reflected in the sparse components S_m and the rescaling weights W_m. The different fill patterns across the rows of the matrices visually indicate heterogeneity in the dependency structures across entities.*

Next, we demonstrate the versatility of the proposed model through a synthetic data example. We consider time series data generated from a heterogeneous collection of VAR models as briefly described next.

EXAMPLE 1. Consider $M = 20$ entities, $p = 20$ variables and $T = 1000$ time points. The entities are organized into distinct clusters as follows: (i) two clusters ($m = 1, \cdots, 5$ and $m = 6, \cdots, 10$, respectively) where the corresponding weight matrices W_m share identical diagonal entries; (ii) one cluster ($m = 11, \cdots, 14$) with $W_m = 0$, resulting in a sparse transition matrices A_m; (iii) one cluster ($m = 15, \cdots, 18$) where the weight matrices W_m have identical entries and in addition $S_m = 0$ (no entity specific component); and (iv) two additional VAR models ($m = 19$ and 20) with transition matrices A_m that differ from all previous groups. Additional details on the data generation process and model parameter specifications are provided in Section 4.2.

The parameters of the panel VAR model in Example 1 were estimated using the iterative algorithm described in Algorithm 2. To summarize and visualize the results, we applied principal components analysis (PCA) to the collection of the diagonal entries form the estimated W_m matrices -i.e., $\widehat{\mathbf{W}} = \left(\mathrm{diag}(\widehat{W}_1), \cdots, \mathrm{diag}(\widehat{W}_M)\right)$. The resulting PCA plot, shown in Figure 2, clearly illustrates that the proposed model effectively captures the underlying heterogeneity in the data, correctly reflecting both similarities within clusters and entity-specific

differences. These results highlight the ability of the model in (1.1) to flexibly capture diverse patterns—such as latent cluster structures and purely sparse models—making it highly versatile compared to existing panel VAR approaches; see further discussion in Appendix B in the Supplement.

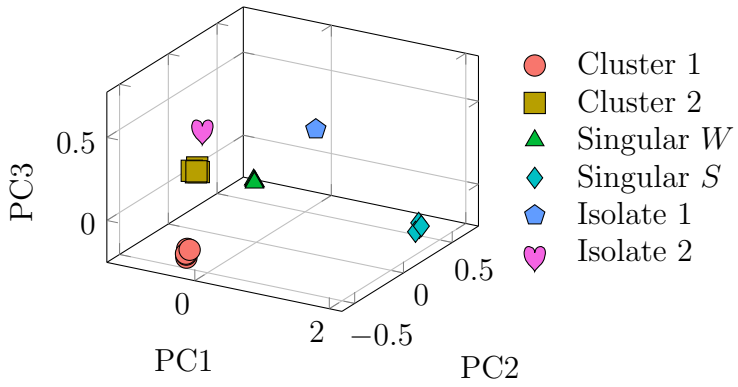

FIG 2. *The top three principal components of the estimated $\mathcal{W}$ of a model consisting of mixed sub-models. Here Singular W (Singular S) represents the special case when the m-th subject has a purely sparse (low-rank) setting with $W_m = 0$ ($S_m = 0$).*

However, the structure of the proposed model presents several technical challenges. The first stems from the presence of both a low-rank component ($W_m\Phi$) and a sparse component (S_m), and the need to reliably distinguish between them—a problem previously studied in multivariate regression (Negahban and Wainwright, 2011) and single VAR models (Basu, Li and Michailidis, 2019). The second challenge concerns the lack of identifiability in the low-rank component $W_m\Phi$ due to the presence of the weight matrices W_m. For instance, rescaling W_m and Φ by a common scalar leaves the product $W_m\Phi$ unchanged, making the parameters non-identifiable without further constraints. For a single VAR model, the first challenge has been effectively addressed in the existing literature by using a nuclear norm and an ℓ_1 regularizer to estimate the low-rank and sparse components, respectively, while imposing an incoherence condition on the entries of the low-rank component to distinguish it from the sparse component. However, these strategies alone are insufficient for resolving the identifiability issues in our model. To overcome this, we impose additional normalization constraints on Φ, which increase both the complexity of the estimation problem and the difficulty of designing a provably convergent optimization algorithm. We resolve these issues satisfactorily, by designing a *multi-block* Alternating Direction Method of Multipliers (ADMM) and establish its convergence properties under mild assumptions, despite the non-convex, non-smooth nature of the underlying objective function. Note that unlike the well-studied convergence analysis of the 2-block ADMM (see Boyd et al., 2011 and references therein), the multi-block version is significantly more challenging, due to requiring careful handling of the interdependencies between the multiple block updates and sometimes additional regularization to guarantee reliable performance. Further, consistency results for the proposed estimators of the various model parameters are provided.

The remainder of the paper is organized as follows. Section 2 addresses parameter identifiability and develops an ADMM algorithm for estimation, along with convergence guarantees. Section 3 establishes consistency of the parameter estimates obtained via this algorithm. Section 4 covers tuning parameter selection and evaluates the algorithm's performance on synthetic data. Section 5 demonstrates the model and algorithm on a neuroscience data set.

Section 6 provides concluding remarks. For appendices, Appendix A provides background on VAR models, while Appendix B expands the literature review on panel VAR models and offers a detailed comparison of our approach with the MAR model. Estimation details are presented in Appendix C. Properties of the proposed fixed nuclear-norm constraint space are studied in Appendix D, and Appendix E discusses the incoherence between the low-rank and sparse components. Appendix F extends our panel VAR setup to the multi-lag case and examines the applicability of our method. The proofs for algorithm convergence and estimator consistency are given in Appendix G and Appendix H, respectively. Finally, Appendix I provides additional simulation details, including data generation, initialization, and supplementary results.

Notation. Denote by $\mathbf{1}_n$ an n-dimensional vector comprising of ones, and by e_i the i-th standard basis vector with an appropriate dimension, I_n as an identity matrix of size $n \times n$. The operator $\mathrm{diag}(x)$ returns either a diagonal matrix with the diagonal elements from x if x is a vector, or a vector consisting of the diagonal elements of x if x is a square matrix. For matrix X, X' denotes the transpose of X, $X[i,j]$ (or $X_{i,j}$ if unambiguous within context) denotes the (i,j)-th element of the matrix X, $\mathrm{rk}(X)$ returns the rank of X, $\mathrm{supp}(X)$ is the support of matrix X with nonzero elements, $\mathrm{range}(X)$ spans the matrix space with column-space contained in that of X, $\det(X)$ and $\mathrm{tr}(X)$ are the determinant and trace function of X, and the sign function $\mathrm{sign}(X)$ maps positive (zero, or negative) matrix entries to 1 (0, or -1) as the signs element-wise. $O(\cdot)$ and $o(\cdot)$ are the usual big and small O notations respectively, and we use $x \gtrsim y$ ($x \lesssim y$) if there exists an absolute positive constant c such that $x \geqslant cy$ ($x \leqslant cy$). For norms, $\|X\| = \|X\|_2$ by default represents the spectral norm, $\|X\|_1 = \sum_{i,j} |X_{i,j}|$ is the ℓ_1 norm that sums up all the absolute values of matrix entries, and $\|X\|_\infty = \max_{i,j} |X_{i,j}|$ is the ∞-norm that returns the maximum absolute value. The norm $\|X\|_{q\to r}$ is defined analogous to matrices' operator norm such as $\|X\|_{q\to r} = \max\{\|Xv\|_r : \|v\|_q \leqslant 1\}$. The nuclear norm of matrix X is denoted as $\|X\|_*$. The ℓ_0-norm $\|X\|_0$ calculates the support size of matrix (or vector) X, i.e., the number of non-zero elements in X. For a square matrix X of dimension n, we write $\lambda_i(X)$ as the i-th eigenvalue of X sorted non-increasingly by magnitudes. Define the indicator function $\mathbb{1}\{A\}$ that outputs 1 if the event A is satisfied, and 0 otherwise.

2. Estimation Procedure for the LSPVAR Model Parameters. Recall that the LSPVAR model under consideration is defined as

$$Y_m = A_m X_m + \epsilon_m; \qquad A_m = W_m \Phi + S_m, \quad m = 1, \cdots, M$$

where $Y_m = (X_1^m, \cdots, X_T^m), X_m = (X_0^m, \cdots, X_{T-1}^m) \in \mathbb{R}^{p\times T}$ and $\epsilon_m = (\epsilon_1^m, \cdots, \epsilon_T^m) \in \mathbb{R}^{p\times T}$. The next assumption ensures that the posited model is stable.

ASSUMPTION 1. LSPVAR model stability: The coefficient matrices $A_m, m = 1, \cdots, M$ satisfy the stability condition $|\lambda_1(A_m)| < 1$; i.e., the eigenvalue of largest magnitude is less than one —an analogous condition to that used in a single VAR model (Basu and Michailidis, 2015).

The following assumption ensures that the decomposition of A_m in (1.1) is nonexplosive and the low-rank component $W_m\Phi$ is non-degenerate.

ASSUMPTION 2. The two summands of A_m, i.e., both $W_m\Phi$ and S_m, are upper bounded in terms of their spectral norms. Further, there exists $w > 0$,

$$\min_j \frac{1}{M} \sum_{m=1}^{M} \|e_j' W_m \Phi\|^2 \geqslant w^2. \tag{2.1}$$

Specifically, (2.1) implies that for each variable $j = 1, \cdots, p$, the set of entities exhibiting singular dynamic patterns (i.e., $W_m = 0$) in the low-rank component must constitute a minority within the panel. Note that the setup in Example 1, where certain $W_m = 0$, does not violate Assumption 2, given the assumed structure of the other clusters.

The model parameters are estimated by minimizing the least squares objective function:

$$f(\mathcal{W}, \mathcal{S}, \Phi; \mathcal{X}) = \sum_{m=1}^{M} \frac{1}{2T} \|Y_m - (W_m \Phi + S_m) X_m\|_F^2, \tag{2.2}$$

where $\mathcal{W} = \{W_1, \ldots, W_M\}$, $\mathcal{S} = \{S_1, \ldots, S_M\}$, and by $\mathcal{X} = \{(X_1, Y_1), \ldots, (X_M, Y_M)\}$. In the unbalanced panel setting, where the number of observations T_m varies across entities, the normalization factor $\frac{1}{2T}$ in (2.2) is replaced by $\frac{1}{2T_m}$ for each entity m. The subsequent analysis follows analogously by replacing T by T_m for entity-specific quantities, while setting $T = \min_m T_m$ when considering the entire panel.

Since the idiosyncratic components S_m are assumed to be sparse, we employ the popular ℓ_1-regularizer, given by

$$\mathcal{P}_S(S_m; \eta) = \eta \|S_m\|_1, \quad m = 1, \cdots, M. \tag{2.3}$$

To address the identifiability issue regarding the low rank component $W_m \Phi$– i.e., distinguishing the weight matrices W_m from the common low-rank basis Φ– we impose the following constraints: (i) all rows of Φ have the same ℓ_2 norm and (ii) its nuclear norm and an upper bound on its rank are fixed; namely,

$$\begin{cases} \Phi \in \mathbb{B}_p := \{\Phi \in \mathbb{R}^{p \times p} : \|e_1' \Phi\| = \cdots = \|e_p' \Phi\| \neq 0\}, \\ \Phi \in \mathbb{L}_p(\hat{r}, \ell) := \{\Phi \in \mathbb{R}^{p \times p} : \|\Phi\|_* = \ell, \ \mathrm{rk}(\Phi) \leqslant \hat{r}\}, \end{cases} \tag{2.4}$$

where $\ell > 0$ is a constant and $\hat{r}$ a positive integer satisfying $r \leqslant \hat{r} \leqslant p$.

REMARK 2.1. Note that each row of the low rank constraint has $p + 1$ unknown parameters and hence a constraint of the type in (2.4) is required to address the identifiability issue. The space $\mathbb{B}_p$ ensures that the rows of Φ are strictly bounded away from singular vectors, thereby preventing any explosive estimation of W_m. Further, the space $\mathbb{L}_p(\hat{r}, \ell)$ corresponds to a simplex for the singular values of Φ. As a result, the minimizer over $\mathbb{L}_p(\hat{r}, \ell)$ usually leads to a solution at the boundary (including vertices) of the space, which naturally results in the low-rankness of Φ. In addition, $\mathbb{L}_p(\hat{r}, \ell)$ imposes an upper bound on the Frobenius norms of the matrices in the space. The intersection of the two constrained spaces has fixed p "degrees of freedom", thus resolving the identifiability issue between $\mathcal{W}$ and Φ. Finally, note that simply imposing a nuclear norm constraint on the product $W_m \Phi$ for each m, does not resolve the identifiability issue, which is critical for model interpretability purposes.

The optimization problem is then formulated as

$$\min_{\mathcal{W}, \mathcal{S}, \Phi} F(\mathcal{W}, \mathcal{S}, \Phi; \mathcal{X}, \eta) = f(\mathcal{W}, \mathcal{S}, \Phi; \mathcal{X}) + \sum_{m=1}^{M} \mathcal{P}_S(S_m; \eta) \ \text{ s.t. } \Phi \in \mathbb{B}_p \cap \mathbb{L}_p(\hat{r}, \ell). \tag{2.5}$$

We develop a multi-block ADMM algorithm[1] based on an auxiliary variable Φ_c that satisfies constraint (2.4). The resulting augmented Lagrangian function is given by

$$G(\mathcal{W}, \mathcal{S}, \Phi_c, \Phi, \Gamma; \mathcal{X}, \eta, \rho) = F(\mathcal{W}, \mathcal{S}, \Phi; \mathcal{X}, \eta) + \frac{\rho}{2} \|\Phi - \Phi_c\|_F^2 + \rho \langle \Gamma, \Phi - \Phi_c \rangle, \tag{2.6}$$

[1] The structure of the constraint in (2.5) makes a block coordinate descent algorithm ineffective in updating the value of Φ.

where $\Phi \in \mathbb{R}^{p\times p}$, $\Phi_c \in \mathbb{B}_p \cap \mathbb{L}_p(\hat{r}, \ell)$, $\Gamma \in \mathbb{R}^{p\times p}$ is the dual variable matrix, and ρ is a scalar coefficient that determines the step size of the parameter updates. Note that the domains $\mathbb{B}_p$ and $\mathbb{L}_p(\hat{r}, \ell)$ of Φ_c, are both nonconvex, but connected compact semi-algebraic sets (as established in the proof of Theorem 2.1).

The ADMM algorithm has three main primal descent blocks and one dual ascent block, summarized in Algorithm 1. The solutions to all primal subproblems (line 4) and the dual update (line 5) are provided in Appendix C.

Algorithm 1: Estimation of $\mathcal{W}, \mathcal{S}, \Phi_c, \Phi$ and Γ

Input: Time series data $\{X_m, Y_m\}_{m=1}^M$, maximum rank $\hat{r}$, fixed nuclear norm ℓ, Lasso parameter η, step size ρ, preset maximum iteration numbers N, convergence tolerance ϵ.

Output: Estimators $\mathcal{W}^{(n)}, \mathcal{S}^{(n)}, \Phi_c^{(n)}, \Phi^{(n)}, \Gamma^{(n)}$.

1 Initialize $\hat{w}_m^{(0)} = \mathbf{1}_p$, $\hat{S}_m^{(0)} = 0$ for $m = 1, \ldots, M$, sample $\Phi^{(0)} = \Phi_c^{(0)}$ from $\mathbb{B}_p \cap \mathbb{L}_p(\hat{r}, \ell)$, and set $\Gamma^{(0)} = 0 \in \mathbb{R}^{p\times p}$.

2 Evaluate the objective function as $G^{(0)}$ at $(\mathcal{W}^{(0)}, \mathcal{S}^{(0)}, \Phi_c^{(0)}, \Phi^{(0)}, \Gamma^{(0)})$.

3 **for** $i = 1 : N$ **do**

4 Update $(\mathcal{W}^{(i)}, \mathcal{S}^{(i)})$, $\Phi_c^{(i)}$, and $\Phi^{(i)}$ sequentially with the optimal solutions of their corresponding subproblems.

5 Update the dual variables $\Gamma^{(i)}$.

6 Evaluate the objective function as $G^{(i)}$ at $(\mathcal{W}^{(i)}, \mathcal{S}^{(i)}, \Phi_c^{(i)}, \Phi^{(i)}, \Gamma^{(i)})$.

7 Terminate and set $n = i$ if some convergence criteria are met, i.e., $|G^{(i-1)} - G^{(i)}| < \epsilon$ or $\frac{\|\Phi^{(i)} - \Phi^{(i-1)}\|_F}{\ell} < \epsilon$.

8 **end**

9 Use ordinary least squares to refine and update the estimators $(\mathcal{W}^{(n)}, \mathcal{S}^{(n)})$ with fixed $\Phi^{(n)}$ and support of $\mathcal{S}^{(n)}$.

10 Output the final estimators $\mathcal{W}^{(n)}, \mathcal{S}^{(n)}, \Phi_c^{(n)}, \Phi^{(n)}, \Gamma^{(n)}$.

REMARK 2.2. Note that the estimate of Φ_c is based on Dykstra's algorithm (Boyle and Dykstra, 1986). The convergence of the corresponding subproblem's iterates can be verified using a similar proof strategy as outlined in Section 2.1; for more details, see Section C.2 in the Supplement.

REMARK 2.3. The selection of the step size ρ and the nuclear norm paramter ℓ is discussed in Theorems 2.1 and 3.1, while the choice of the input rank $\hat{r}$ is addressed in Section 3. These selections demonstrate strong empirical performance based on the synthetic data experiments presented in Section 4.1. Finally, the sparsity tuning parameter η is selected based on BIC (Wang, Li and Jiang, 2007; Zou, Hastie and Tibshirani, 2007), with additional implementation details provided in Section 4.

2.1. *Convergence Guarantees for Algorithm 1.* As previously noted, establishing convergence guarantees for multi-block ADMM algorithms in the context of non-smooth and non-convex problems remains a challenging task. We start by introducing assumptions and some additional notation.

ASSUMPTION 3 (Restricted Strong Convexity (RSC)). For an arbitrary matrix $\Delta \in \mathbb{R}^{p\times p}$, the following relationships holds

$$\sum_{m=1}^{M} \frac{1}{2T} \|\Delta X_m\|_F^2 \geqslant \frac{\beta}{2} \sum_{m=1}^{M} \|\Delta\|_F^2,$$

with $\beta = \min_m\{\beta_m\}$, and $\beta_m := \lambda_p(\frac{X_m X_m'}{T})$,

REMARK 2.4. This is a commonly used assumption in the high-dimensional statistics literature (Wainwright, 2019) and also in VAR models (Basu and Michailidis, 2015). It is leveraged both in the convergence analysis of Algorithm 1 and the consistency of the model parameters in Theorem 3.1.

The following assumptions are imposed on the objective function F.

ASSUMPTION 4. Given a data realization $\mathcal{X}$, the function F:

1. is β_W-convex with respect to W_m for every m, and β_Φ-convex with respect to Φ.
2. Its gradient is α_W-Lipschitz continuous with respect to W_m (for every m), α_S-Lipschitz continuous with respect to S_m (for every m), and α_Φ-Lipschitz continuous with respect to Φ.

REMARK 2.5. Recall that the function F in (2.5) consists of a sum of least squares terms and ℓ_1 penalties. Therefore, the above convexity and continuity assumptions are satisfied provided the maximum and minimum eigenvalues of the corresponding data Gram matrices $\frac{X_m X_m'}{T}$ satisfy $0 < \lambda_p(\frac{X_m X_m'}{T}) \leqslant \lambda_1(\frac{X_m X_m'}{T}) < \infty$. Indeed, this condition is partially subsumed in Assumption 3 regarding the lower bound of $\beta_m = \lambda_p(\frac{X_m X_m'}{T})$.

The next result characterizes the behavior of the iterates of the parameter estimates generated by Algorithm 1.

THEOREM 2.1. *Suppose that Assumptions 1 to 4 hold, and based on a data realization $\mathcal{X}$ from model (1.1) and for step size satisfying*

$$\rho \gtrsim \max\left\{\frac{M\alpha_W^2}{\beta_W}, \frac{M\alpha_S^2}{\beta}, \alpha_\Phi\right\}, \tag{2.7}$$

the sequence of iterates of the model parameters $(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)})$ from Algorithm 1 converges globally to a stationary point $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi}_c, \widehat{\Phi}, \widehat{\Gamma})$ of the augmented Lagrangian function G.

The first step is to establish a "sufficient descent property," provided in Proposition 2.1. The second step is to show that the augmented Lagrangian function is lower bounded along the sequence of iterates, as shown in Proposition 2.2. When these two conditions hold, the set of accumulation points of Algorithm 1 is guaranteed to be non-empty, compact, and connected (see Remark 5 in Bolte, Sabach and Teboulle, 2014). The final requirement for proving global convergence to a critical point of G is to verify that G satisfies the Kurdyka-Łojasiewicz (KL) property (Attouch, Bolte and Svaiter, 2013), which ensures that the sequence of iterates generated by Algorithm 1 forms a Cauchy sequence. Together, these three components establish the claims in Theorem 2.1. All technical proofs are deferred to Appendix G.

PROPOSITION 2.1 (Sufficient Descent). *Under the Assumptions of Theorem 2.1, the objective function evaluated at the sequence $\{(\mathcal{W}^{(i)}, \mathcal{S}^{(i)}, \Phi_c^{(i)}, \Phi^{(i)}, \Gamma^{(i)})\}$ is monotonically decreasing. Moreoever, there exists a constant $\mathcal{C} > 0$, such that*

$$G(\mathcal{W}^{(i)}, \mathcal{S}^{(i)}, \Phi_c^{(i)}, \Phi^{(i)}, \Gamma^{(i)}) - G(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)})$$

$$\geqslant \mathcal{C}\Big(\sum_{m=1}^{M}\|W_m^{(i)}-W_m^{(i+1)}\|_F^2+\sum_{m=1}^{M}\|S_m^{(i)}-S_m^{(i+1)}\|_F^2+\|\Phi^{(i)}-\Phi^{(i+1)}\|_F^2+\|\Phi_c^{(i)}-\Phi_c^{(i+1)}\|_F^2\Big).$$

PROPOSITION 2.2 (Lower bound of the augmented Lagrangian function). *Under the Assumptions of Theorem 2.1, the evaluation of function G at the estimated sequence from Algorithm 1, $G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i+1)},\Gamma^{(i+1)})$, is lower bounded for all i and converges as $i\to\infty$.*

The next result is based on Proposition 2.2, and the fact that the sequence $G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i+1)},\Gamma^{(i+1)})$ forms a Cauchy sequence.

PROPOSITION 2.3. *Under the Assumptions of Theorem 2.1, $(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i+1)},\Gamma^{(i+1)})$ is convergent to some limiting point $(\widehat{\mathcal{W}},\widehat{\mathcal{S}},\widehat{\Phi}_c,\widehat{\Phi},\widehat{\Gamma})$ with $i\to\infty$.*

Moreover, the limiting point $(\widehat{\mathcal{W}},\widehat{\mathcal{S}},\widehat{\Phi}_c,\widehat{\Phi},\widehat{\Gamma})$ satisfies the local first-order optimality of (2.6) within the corresponding support domain.

REMARK 2.6. The arguments used to prove Proposition 2.3 also imply that the sequence of ADMM iterates achieves an $o(\frac{1}{i})$ convergence rate (Deng et al., 2017, Lemma 1.1). Consequently, for a given error tolerance $\epsilon>0$ (measuring the difference between consecutive iterates), the number of iterations required for convergence is of order $O(1/\epsilon)$.

3. Consistency of the Model Parameter Estimates. In this section, we discuss the consistency results for the estimators obtained from Algorithm 1, focusing on the case where the model structure in (1.1) is correctly specified. Notably, the derivation shows that the effect of a potentially overspecified $\hat{r}$ can be well controlled under mild conditions.

Based on the convergence guarantees in Theorem 2.1, the final output of Algorithm 1 corresponds to a limiting stationary point $(\widehat{\mathcal{W}},\widehat{\mathcal{S}},\widehat{\Phi}_c,\widehat{\Phi},\widehat{\Gamma})$. Given the nonconvex nature of the optimization problem, Theorem 2.1 does not ensure convergence to a global minimum of the objective function (2.5). Moreover, not all stationary points of (2.5) possess equally desirable statistical properties. Therefore, in the following analysis, we focus on stationary points obtained by initializing Algorithm 1 within a suitable neighborhood of the true model parameters $(\mathcal{W}^*,\mathcal{S}^*,\Phi^*)$ with $\Phi^*\in\mathbb{B}_p\cap\mathbb{L}_p(r,\ell)$. In practice, however, our simulation studies in Section 4 indicate that Algorithm 1 consistently achieves strong performance, regardless of the initialization. These empirical results suggest that, even under broader initialization settings, the favorable statistical properties we establish below may be implicitly supported by the structure of the optimization problem.

Before stating the theoretical results, we introduce some additional notation. Define $W=\max_m\|W_m^*\|$, $\widehat{W}=\max(\max_m\|\widehat{W}_m\|,W)$, and $W_m^\dagger=\frac{W_m^*}{W}$ for every m. The proposition below describes the estimation power of Algorithm 1 with $(\mathcal{W},\mathcal{S},\Phi,\Phi_c)$ initialized in a neighborhood of $(\mathcal{W}^*,\mathcal{S}^*,\Phi^*,\Phi^*)$, given the data realization $\mathcal{X}$. Its proof is presented in Appendix H.

PROPOSITION 3.1. *Suppose Assumptions 1 to 4 hold, and that Algorithm 1 is initialized in a neighborhood of the true parameter $(\mathcal{W}^*,\mathcal{S}^*,\Phi^*)$. Further, assume the following:*

(B1) *there exists a constant $\phi>0$ such that*
- *the true parameter $\Phi^*\in\mathbb{B}_p\cap\mathbb{L}_p(r,\ell)$ satisfies $\|\Phi^*\|_\infty\leqslant\frac{\phi\ell}{\sqrt{rp}}$;*
- *the estimator $\widehat{\Phi}\in\mathbb{B}_p\cap\mathbb{L}_p(\hat{r},\ell)$ satisfies $\|\widehat{\Phi}\|_\infty\leqslant\frac{\phi\ell}{\sqrt{rp}}$;*

(B2) *for every m, the sparse matrix S_m^* has at most s non-zero entries.*

Then, for tuning parameter $\eta \geqslant \frac{4\beta\phi\widehat{W}\ell}{\sqrt{rp}} + \max_m \left|\frac{\varepsilon_m X_m'}{T}\right|_\infty$, *and selecting* $\zeta = \min\{\frac{\beta_\Phi}{M\beta}, \frac{\beta_W}{\beta}, 1\} \gtrsim \min\left\{\frac{rp}{\ell^2}, \frac{\ell^2}{\hat{r}p}, 1\right\}$, *the solution* $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi})$ *satisfies*

$$\|\widehat{\Phi} - \Phi^*\|_F^2 + \frac{1}{M}\sum_{m=1}^{M}(\|\widehat{S}_m - S_m^*\|_F^2 + \|\widehat{W}_m - W_m^*\|_F^2) \tag{3.1}$$

$$\lesssim \frac{1}{\zeta^2}\left(\frac{\hat{r}^2}{\ell^2} + \frac{\hat{r}s}{rp} + \frac{s}{\beta^2}\max_m\|\frac{\varepsilon_m X_m'}{T}\|_\infty^2 + \frac{\ell^2}{\beta^2 Mrp}\sum_{m=1}^{M}\|\frac{\varepsilon_m X_m'}{T}\|_2^2 + \frac{r\hat{r}p}{\beta^2\ell^2}\|\sum_{m=1}^{M}\frac{W_m^\dagger \varepsilon_m X_m'}{MT}\|_2^2\right).$$

REMARK 3.1. The assumptions on the ∞-norm bounds of Φ^* and $\widehat{\Phi}$ are mild, especially given that p is assumed to be large and growing with the sample size T; see Appendix H in the Supplement for further explanation and discussion. Moreover, since the constants in Assumption 4 have become tighter at the estimate $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi})$, the feasible choice for the step size coefficient ρ can be refined from (2.7) to $\rho \gtrsim \frac{M\hat{r}p}{\beta\ell^2}\max_m(\|\frac{X_m X_m'}{T}\|^2 + \|\frac{\epsilon_m X_m'}{T}\|^2)$.

Then, for the deviation bounds appearing in (3.1), we adapt analogous results from Basu and Michailidis (2015); Basu, Li and Michailidis (2019) to the LSPVAR model under consideration, as formalized in Proposition 3.2.

PROPOSITION 3.2. *Given the model setup in* (1.1), *suppose that Assumptions 1 and 4 are satisfied. Consider a realization of the data* $\{X_m\}_{m=1}^M$ *and* $\{\varepsilon_m\}_{m=1}^M$, *and define*[2]

$$\xi = \max_m \lambda_1(\Sigma_m)\left[1 + \frac{1 + \tau_{\max}(\mathcal{A}_m)}{\tau_{\min}(\mathcal{A}_m)}\right],$$

$$\xi_\dagger = \max_m \lambda_1(\Sigma_m) + \max_m \frac{\lambda_1(\Sigma_m)}{\tau_{\min}(\mathcal{A}_m)} + \max_m \frac{\lambda_1(\Sigma_m)\tau_{\max}(\mathcal{A}_m)}{\tau_{\min}(\mathcal{A}_m)}.$$

Assume $\log(p) \gtrsim \log(M)$. *Then, there exist constants* $c_i > 0$ $(i = 1, 2, 3)$ *such that:*

1. *for* $T \gtrsim p$ *and* $T \gtrsim \log(p)$, *with probability at least* $1 - c_2\exp(-c_3\log(p))$,

$$\max_m\|\frac{\varepsilon_m X_m'}{T}\|_2 \leqslant c_1\xi\sqrt{\frac{p}{T}}, \qquad \max_m\|\frac{\varepsilon_m X_m'}{T}\|_\infty \leqslant c_1\xi\sqrt{\frac{\log(p)}{T}};$$

2. *for* $T \gtrsim p$, *with probability at least* $1 - c_2\exp(-c_3\log(p))$,

$$\|\sum_{m=1}^{M}\frac{W_m^\dagger\varepsilon_m X_m'}{MT}\|_2 \leqslant c_1\xi_\dagger\sqrt{\frac{p}{MT}};$$

3. *for* $T \gtrsim p\Psi^2(h_{X_m})/\psi^2(h_{X_m})$, *with probability at least* $1 - c_2\exp(-c_3\log(p))$,

$$\min_m \lambda_p(\frac{X_m X_m'}{T}) \geqslant \min_m \frac{1}{4\pi}\cdot\frac{\lambda_p(\Sigma_m)}{\tau_{\max}(\mathcal{A}_m)}.$$

Combining Proposition 3.1 and Proposition 3.2, and noting that $\xi \leqslant \xi_\dagger$, we obtain the following theorem, which establishes the high-probability consistency of the proposed estimator.

[2]The quantities $\tau_{\max}$, $\tau_{\min}$, $\Psi(h_{X_m})$, $\psi(h_{X_m})$ and $\mathcal{A}_m$ are related to the characteristic polynomials and spectral densities of the individual VAR models. Their rigorous definitions are given in Appendix A.

THEOREM 3.1. *Under the setting of Propositions 3.1 and 3.2, assume there exists a constant ι such that $\hat{r} \leqslant \iota r$, and set $\ell = \sqrt{\hat{r}p}$. Then, there exist universal positive constants C_i for $i = 1, 2, 3, 4$, such that the solution obtained from Algorithm 1 satisfies, with probability at least $1 - C_3 \exp(-C_4 \log(p))$,*

$$\|\widehat{\Phi} - \Phi^*\|_F^2 + \frac{1}{M}\sum_{m=1}^{M}(\|\widehat{S}_m - S_m^*\|_F^2 + \|\widehat{W}_m - W_m^*\|_F^2) \leqslant C_1\iota \cdot \frac{r+s}{p} + C_2\xi_\dagger^2 \max_m \frac{\tau_{\max}^2(\mathcal{A}_m)}{\lambda_p^2(\Sigma_m)}\Big(\frac{s\log(p)+\iota p}{T} + \frac{rp}{MT}\Big). \tag{3.2}$$

The roadmap to establish Theorem 3.1 is as follows. First, Proposition 3.1 establishes an upper bound on the estimation errors, which is dependent on the deviation bounds of the time series data. Next, the deviation bounds are controlled with high-probability based on Proposition 3.2. Then, the consistency rate is obtained in a straightforward manner, by plugging the obtained bounds back in (3.1) and selecting appropriate values for ℓ and $\hat{r}$.

REMARK 3.2. Note that ι is treated more like a pseudo-constant and serves as a guiding reference for selecting $\hat{r}$. The derived error bound consists of two distinct parts. The first part arises from the inherent non-identifiability of the model in separating the low-rank and sparse components. It depends only on the rank r, the sparsity s and the overspecification ι. It will not vanish even with increasing sample size, but remains small under a reasonably chosen ι in correctly specified settings, where $r \ll p$ and $s \ll p$. The second term reflects the randomness of the data and vanishes as the sample size (time series length T) increases. In particular, the term $\frac{s\log(p)+\iota p}{T}$ can be interpreted as the proxy parametric convergence rate of estimating (W_m, S_m), whose effective degrees of freedom are approximately $DOF \approx s\log(p) + p$, based on time series of length T for each entity. Similarly, the term $\frac{rp}{MT}$ corresponds to the rate of estimating the shared low-rank component Φ, with $DOF \approx rp$, using the entire panel $\{X_m\}_{m=1}^M$ of size MT. The signal-to-noise ratio of the model further influences the second term through factors such as $\max_m \frac{\tau_{\max}^2(\mathcal{A}_m)}{\lambda_p^2(\Sigma_m)}$ and $\xi_\dagger$.

As previously noted, there exists a source of indeterminacy between the diagonal matrices $\mathcal{W}$ and the low-rank structure Φ, as scaling both factors in the product $W_m\Phi$ by the same constant leaves the transition matrices A_m unchanged. While the constraint set $\mathbb{B}_p \cap \mathbb{L}_p(\hat{r}, \ell)$ serves to fix the overall magnitude of $\widehat{\Phi}$, there remains a possibility that part of $\widehat{\Phi}$'s nuclear norm may "leak" outside the primary rank-r subspace of the true low-rank component Φ^*. To account for this, we present the following corollary, which provides classical consistency guarantees for the combination of $\widehat{\mathcal{W}}$ and $\widehat{\Phi}$. A proof sketch is included in Appendix H.

COROLLARY 3.1. *Under the Assumptions of Theorem 3.1 and in addition assuming $\eta \geqslant \frac{2\beta\phi\widehat{W}\ell}{\sqrt{rp}} + \max_m \|\frac{\varepsilon_m X_m'}{T}\|_\infty$, there exist universal positive constants C_i for $i = 1, 2, 3, 4$, such that the solution obtained from Algorithm 1 satisfies, with probability at least $1 - C_3\exp(-C_4\log(p))$,*

$$\frac{1}{M}\sum_{m=1}^{M}(\|\widehat{S}_m - S_m^*\|_F^2 + \|\widehat{W}_m\widehat{\Phi} - W_m^*\Phi^*\|_F^2) \leqslant C_1 \cdot \frac{\iota s}{p} + C_2\xi^2 \cdot \max_m \frac{\tau_{\max}^2(\mathcal{A}_m)}{\lambda_p^2(\Sigma_m)} \cdot \frac{s\log(p)+\iota rp}{T}. \tag{3.3}$$

REMARK 3.3. The error bound in (3.3) aligns with analogous results in the literature for a single VAR model; see, e.g., Basu, Li and Michailidis (2019). The main difference between

the right-hand side bound in Corollary 3.1 and that in Theorem 3.1 lies in the presence of the factor $\frac{1}{M}$ in the rates $\frac{\hat{r}p}{T}$ versus $\frac{\hat{r}p}{MT}$. This is due to the Φ-relevant error bounds being influenced by idiosyncratic rescaling factors $\mathcal{W}$. Hence, the bound (3.3) is less sharp, requiring $T \gg rp$, compared to $T \gg \frac{rp}{M}$ in (3.2).

4. Performance Evaluation. The proposed LSPVAR parameter estimation strategy based on Algorithm 1 is assessed through numerical experiments with synthetic data. he simulation studies in Section 4.1 primarily focus on practical strategies for selecting the input rank $\hat{r}$. Additionally, Section 4.2 explores variants of the scenario described in Example 1, highlighting the model's capability to uncover heterogeneous patterns across a panel of VAR models.

To select the optimal penalty coefficient η, we perform a grid search over a geometric sequence covering an appropriate range, aiming to minimize the Bayesian Information Criterion (BIC). For each estimate $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi})$ corresponding to a given η, we calculate $RSS_\eta = \sum_{m=1}^{M} \|Y_m - (\widehat{W}_m\widehat{\Phi} + \widehat{S}_m)X_m\|_F^2$, and define the model degrees of freedom $\hat{d}$ as

$$\hat{d} = (2p - \mathrm{rk}(\widehat{\Phi})) \cdot \mathrm{rk}(\widehat{\Phi}) + p(M-1) + \sum_{m=1}^{M} \|\widehat{S}_m\|_0.$$

The BIC is then computed as

$$BIC_\eta = MTp \cdot \log(\frac{RSS_\eta}{MTp}) + \hat{d} \cdot \log(MT).$$

Besides information criteria, we also consider several classical metrics to evaluate the performance of our approach. Specifically, we assess sensitivity, specificity, and overall accuracy to study the sparse recovery of $\mathcal{S}$. Additionally, we compute the relative errors of the coefficient matrices and the Frobenius norms of the normalized components to evaluate the overall quality of the algorithm's estimates. These metrics are provided in Table 4 in Section I.3.

4.1. *Choice of Input Rank $\hat{r}$.* Next, the performance of Algorithm 1 is evaluated based on different choices for $\hat{r}$, while also detailing tuning parameter selection guidelines for η and ρ. The setting considered is $(M, p, r, s) = (20, 40, 5, 30)$ with $T = 2rp = 400$. Candidate ranks are $\hat{r} \in \{3, 5, 10, 15, 20, 25, 30, 35, 40\}$, and $\ell = \sqrt{\hat{r}p}$ based on Theorem 3.1. The step size is selected as $\rho \in \{\frac{M}{20}, \frac{M}{5}, M\}$ based on (2.7) and Remark 3.1. The grid search interval for η is set as $[4 \times 10^{-2}, 2.5 \times 10^{-1}]$, selected from a coarser pilot run.

Figure 3 illustrates how key performance metrics vary with the tuning parameter η, input rank $\hat{r}$, and step ρ. Focusing on η, with other parameters fixed, we observe that the minimal BIC is consistently attained around $\eta \approx 10^{-2}$, which coincides with the minimum relative Frobenius error of the coefficient matrices. Additionally, Figure 3 demonstrates that the sparse component's recovery accuracy is also maximized near this η value. These results validate the grid search procedure as an effective strategy for tuning η in practice, even without prior knowledge of the sparsity level. The other two parameters $\hat{r}$ and ρ, based on Figure 3 exhibit minimal influence on performance, provided $\hat{r} \geqslant r$ and $\rho = O(M)$. This observation is consistent with our theoretical discussion of $\hat{r}$ and ρ in Theorem 3.1 and Remarks 3.1 and 3.2.

It is worth emphasizing that the setting with $T = 2rp$ under consideration, is challenging for either a single low-rank plus sparse model, or a general VAR model, due to lack of adequate sample size. Nevertheless, our empirical results demonstrate that Algorithm 1 accurately recovers the LSPVAR model parameters.

Based on the above findings and discussion, a choice of $\hat{r} \geqslant r$ and $\rho = O(M)$, is recommended in practice.

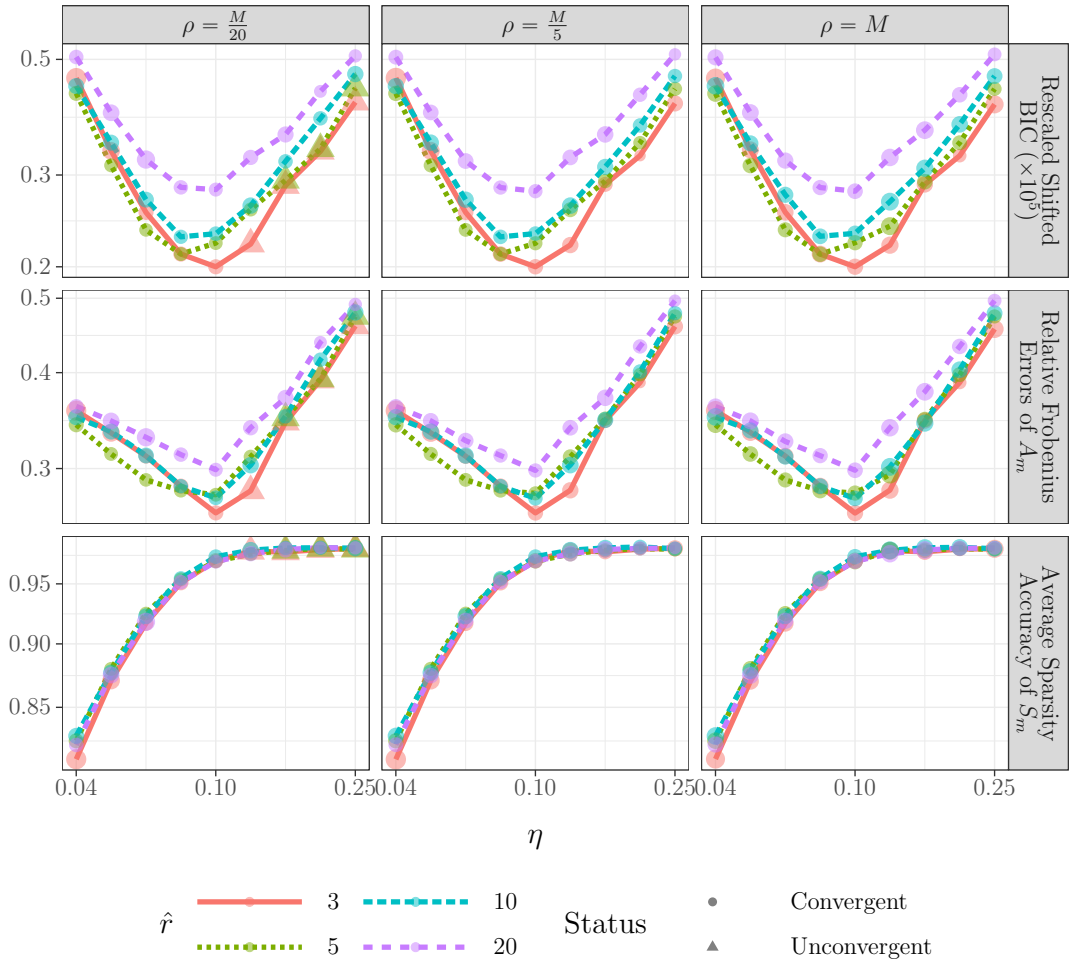

FIG 3. *The trajectories of BIC, relative errors of A_m and sparse recovery accuracy of S_m as functions of η are depicted. For better visualization, the BIC values are shifted and rescaled. The size of each scatter point reflects the number of iterations required for convergence, while triangular shaped points indicate cases where Algorithm 1 fails to meet the convergence tolerance $\epsilon = 5 \times 10^{-6}$ within the maximum allowed $N = 4 \times 10^5$ iterations.*

4.2. *Assessing Panel Heterogeneity.* As stated in Section 1, LSPVAR can effectively infer latent heterogeneity in the panel based on the obtained estimates from Algorithm 1, even in the presence of "degenerate configurations", where some of the constituent models in the panel exhibit pure sparse or pure low-rank structure.

Revisiting the model in Example 1: recall the setup with $(M, p, r, s, T) = (20, 40, 5, 30, 400)$. The structure of the VAR models within the panel is as follows:

1. Two clusters with nonsingular diagonal matrices $W_1 = \cdots = W_5 \neq 0$, $W_6 = \cdots = W_{10} \neq 0$, and non-restrictive sparse matrices $\{S_m\}_{m=1}^{10}$. We refer to these as *Cluster 1* and *Cluster 2*, respectively.
2. One cluster of purely sparse structure, i.e., the diagonal matrices $W_{11} = \cdots = W_{14} = 0$. This is labeled as *Singular W*.
3. One cluster with identical diagonal matrices $W_{15} = \cdots = W_{18} \neq 0$ and singular sparse matrices $S_{15} = \cdots = S_{18} = 0$. This is labeled as *Singular S*.

4. Two isolated entities, *Isolate 1* and *Isolate 2*, whose diagonal matrices and sparse matrices are different than all the previous ones.

The estimators are obtained by minimizing the BIC through grid search over η, with $(\hat{r}, \rho) = (\frac{p}{2}, \frac{M}{20})$. For visualization, Figure 2 presents a 3-dimensional scatter plot of the leading principal components of $\widehat{\mathbf{W}} = \left(\mathrm{diag}(\widehat{W}_1), \ldots, \mathrm{diag}(\widehat{W}_M)\right)$. It is noticeable from Figure 2 that the cluster patterns and/or isolated outliers are well separated and captured by the LSPVAR estimates.

Additionally, Table 1 reports evaluation metrics based on twenty five replicates of the same underlying model. Notably, Notably, the estimation errors for W_m in the *Singular W* group and for S_m in the *Singular S* group are significantly smaller, demonstrating the model's ability to uncover latent singular structures in the panel. Overall, both the error metrics and sparsity recovery results confirm the effectiveness of the proposed model and algorithm in handling heterogeneous and idiosyncratic settings.

Cluster	Average Relative Frobenius Error of A_m	Average Absolute Frobenius Error of W_m	Average Absolute Frobenius Error of S_m	Sparsity Recovery Accuracy	Sparsity Recovery Sensitivity	Sparsity Recovery Specificity
Cluster 1	0.099	0.303	0.274	0.992	0.877	0.994
Cluster 2	0.099	0.409	0.390	0.990	0.822	0.995
Singular W	0.081	**0.107**	0.264	0.991	0.875	0.995
Singular S	0.134	0.953	**0.070**	0.999	NA	0.999
Isolate 1	0.119	0.271	0.340	0.988	0.822	0.993
Isolate 2	0.090	0.774	1.300	0.986	0.792	0.993

TABLE 1

Summary statistics by cluster based on 25 simulation replicates. The absolute errors of the rescaling effects W_m's are computed under the normalization $\|\Phi\|_* = \sqrt{\hat{r}p}$, consistent with the setup in Theorem 3.1.

A larger size heterogeneous panel: we consider a setting with $(M, p, r, s, T) = (50, 80, 5, 100, 2000)$, wherein the structure of the VAR models in the panel is as follows: two 19-entity clusters, *Cluster 1* and *Cluster 2*, with identical weight matrices W_m within each cluster; one 5-entity cluster with *Singular W* ($W_m = 0$); one 5-entity cluster with *Singular S* ($S_m = 0$), and two isolated entities, *Isolate 1* and *Isolate 2*.

The summary of the estimated weight matrices W_m based on PCA is depicted in Figure 4. Analogously to the smaller heterogeneous panel analyzed above, the estimated weight matrices accurately capture the latent structure of the panel. Additional performance metrics are reported in Table 2, demonstrating that Algorithm 1 yields highly accurate estimates of the LSPVAR model paramters.

Cluster	Average Relative Frobenius Error of A_m	Average Absolute Frobenius Error of W_m	Average Absolute Frobenius Error of S_m	Sparsity Recovery Accuracy	Sparsity Recovery Sensitivity	Sparsity Recovery Specificity
Cluster 1	0.104	0.160	0.326	0.996	0.809	0.999
Cluster 2	0.103	0.159	0.313	0.997	0.796	1.000
Singular W	0.187	**0.130**	0.372	0.994	0.694	1.000
Singular S	0.102	1.670	**0.018**	1.000	NA	1.000
Isolate 1	0.097	0.357	0.982	0.995	0.754	0.998
Isolate 2	0.053	0.201	0.317	0.996	0.852	0.999

TABLE 2

The cluster-based summary statistics from 25 simulation replicates of the large size heterogeneous panel, with parameters $(M, p, r, s, T) = (50, 80, 5, 100, 2000)$.

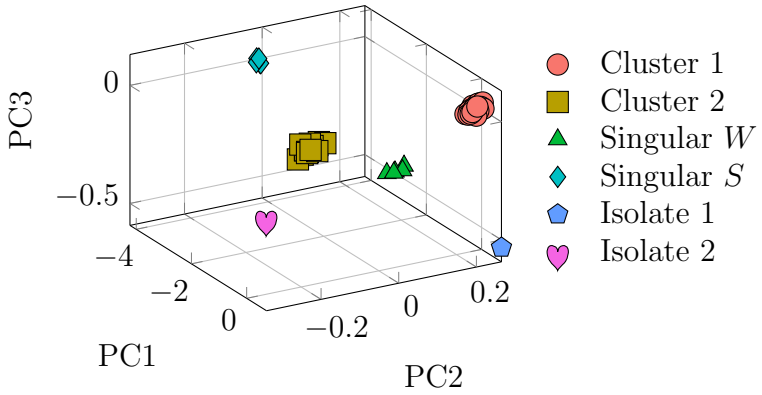

FIG 4. *Top three principal components of the estimated $\mathcal{W}$ for the larger heterogeneous panel comprising mixed sub-models, with parameters $(M, p, r, s, T) = (50, 80, 5, 100, 2000)$.*

5. Application to a Neuroscience Data Set. To illustrate the usefulness in real life applications of LSPVAR, we apply it to a data set comprising EEG signals (Chen et al., 2008) obtained from 22 subjects. Specifically, data on 11 female and 11 male undergraduate students at Texas State University were collected (age range 18-26 years, mean age=21.1 years), while they performed the following two consecutive tasks: alternating one-minute intervals of eyes open (EO) and eyes closed (EC). The EEG signals, sampled at 256 Hz, were recorded from 71 scalp channels as illustrated in Figure 5, with channel locations summarized in Table 3.

For each subject, we separately extracted the EO and EC segments and filtered the alpha band signals (8–13 Hz), which are known to be relevant for visual processing tasks. We followed the preprocessing pipeline described in Bai, Safikhani and Michailidis (2022) and related references. The resulting dataset comprises $M = 44 = 22 \times 2$ panels, with dimension $p = 71$ and time length $T = 2000$.

The key objective is to identify potential clustering patterns corresponding to the EO and EC conditions, as well as to capture additional subject-level heterogeneity. Guided by the parameter selection discussion in Section 4.1, we set $(\hat{r}, \rho) = (18, 4.4)$, approximately $(\frac{p}{4}, \frac{M}{10})$, and estimate the parameters W_m, Φ, and S_m accordingly. Similar to the setup in Example 1, we summarize the results using PCA of the estimated W_m matrices in Figure 6 and Figure 7. The findings reveal considerable heterogeneity across subjects, particularly reflected in the W_m scaling components. Notably, male subjects exhibit a somewhat greater degree of heterogeneity than females, an effect that is even more pronounced in the alpha band data.

Eye-Movement:	LHEOG, LVEOG, RVEOG, RHEOG;
Front:	NFpz, Fp1 - Fp2, AF7 - AF8, F7 - F8;
Central:	FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2;
Central-Left:	FT7, FC5, FC3, T7, C5, C3, M1, TP7, CP5, CP3;
Central-Right:	FT8, FC6, FC4, T8, C6, C4, M2, TP8, CP6, CP4;
Posterior:	P9 - P10, PO7 - PO8, O1 - O2, Iz.

TABLE 3

The 71 EEG channels are classified into the 5 classes for collective analysis and inferences.

To delineate further differences, selected channels from Table 3 are examined in greater detail. Specifically, boxplots of the rescaling factors $\{\widehat{W}_m[i] : m = 1, \ldots, M\}$ for these channels are depicted in Figure 7. It can be seen that the signs of the estimated rescaling factors

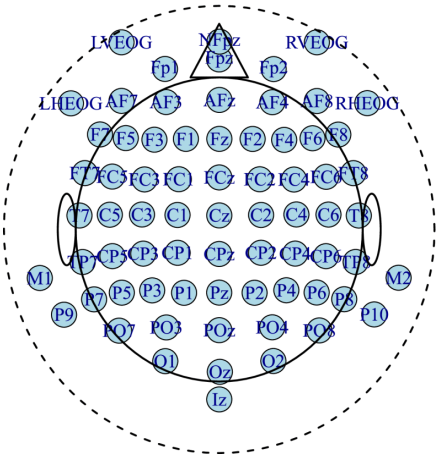

FIG 5. *The abbreviations and scalp locations of the 71 EEG channels in the data set are summarized below. The illustration figure is reproduced from Bai, Safikhani and Michailidis (2022).*

are overall stable amongst subjects. Further, in the alpha band, greater variability is exhibited in the EC condition compared to the EO, but also smaller magnitudes. These observations align with findings reported in the literature (Chen et al., 2008).

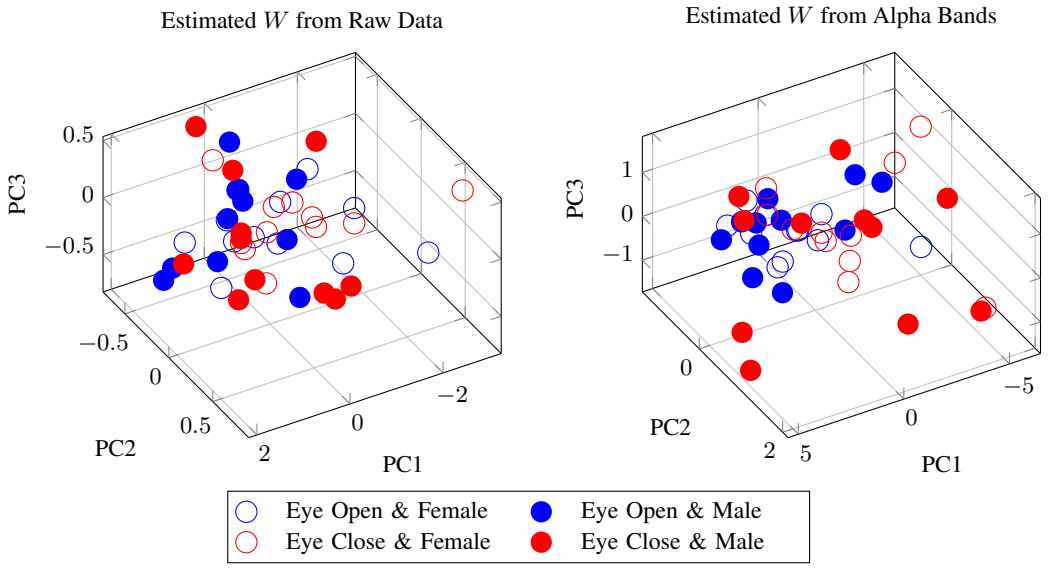

FIG 6. *The 3-dimensional scatter plot for the principal components of the rescaling effects estimated from the raw data and alpha bands, respectively.*

Next, we examine the sparse components $\widehat{\mathcal{S}}_m$. For every (i, j)-th entry of the $p \times p$ sparse matrices, we calculate the frequency of non-zero values separately for the EO and EC conditions, for different groups of channels based on their scalp locations. A heatmap illustrating these frequencies is shown in Table 3. The results reveal which groups of channels exhibit additional activity, as captured by Granger causal effects from the sparse components $\hat{\mathcal{S}}_m$ on top of the low-rank basis $\hat{\Phi}$. It can be seen from Figure 8 that channels associated with eye movement-related alpha bands help filter out noise, especially for dynamics originating from the frontal channels under the EC condition. The channels in the Eye-Movement

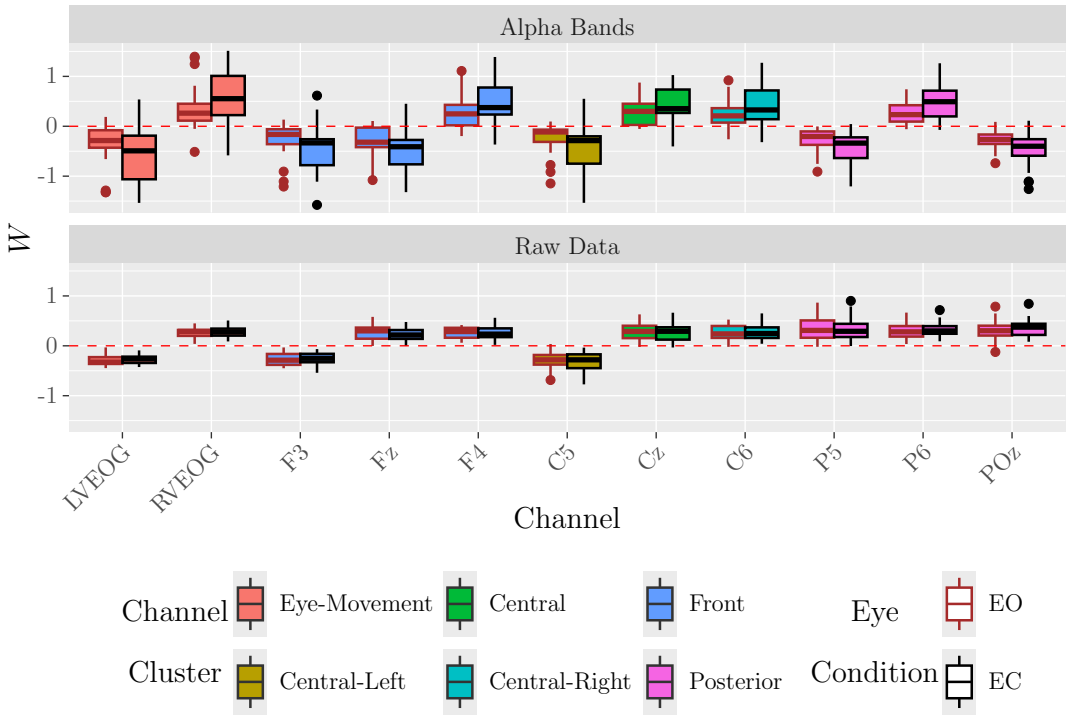

FIG 7. *The boxplot of the estimated rescaling effects for selective channels.*

cluster, which are related to physical eyeball movements (Issa and Juhasz, 2019), are more active during the EO condition, especially in the alpha band data. channels involved in visual processing—primarily located in the posterior scalp region—show more frequent outward connections. These findings are consistent with existing literature (see Bai, Safikhani and Michailidis, 2022, for instance).

6. Conclusion. There has been growing interest in modeling time series data across multiple entities. One approach organizes the data into a three-dimensional array and applies matrix-variate or tensor-based models, which naturally scale to large numbers of entities (M), but often sacrifice interpretability of the estimated parameters. In contrast, our proposed panel VAR framework preserves the classical VAR structure for each entity, capturing inter-entity relationships through structured constraints while allowing substantial heterogeneity across entities. Such flexibility is difficult to achieve with matrix variate or tensor models.

Building on this motivation, this work develop a panel VAR model–LSPVAR–designed to effectively capture both similarities and differences in Granger causal relationships of different entities in a meaningful and easily interpretable manner. The identification conditions for the model parameters lead to a nonsmooth, nonconvex optimization problem, which we address by developing a multi-block ADMM algorithm with established convergence guarantees. Additionally, we prove consistency of the estimators under mild assumptions commonly adopted in single high-dimensional VAR models. Simulation studies with synthetic data and an application to EEG signals demonstrate the effectiveness of the proposed approach.

APPENDIX A: VECTOR AUTO-REGRESSION (VAR) MODEL

Vector Auto-Regression (VAR) is a classical and widely used model for modeling multivariate time series data. Consider the time series observations $X_t^m : t = 0, \ldots, T$ for the m-th

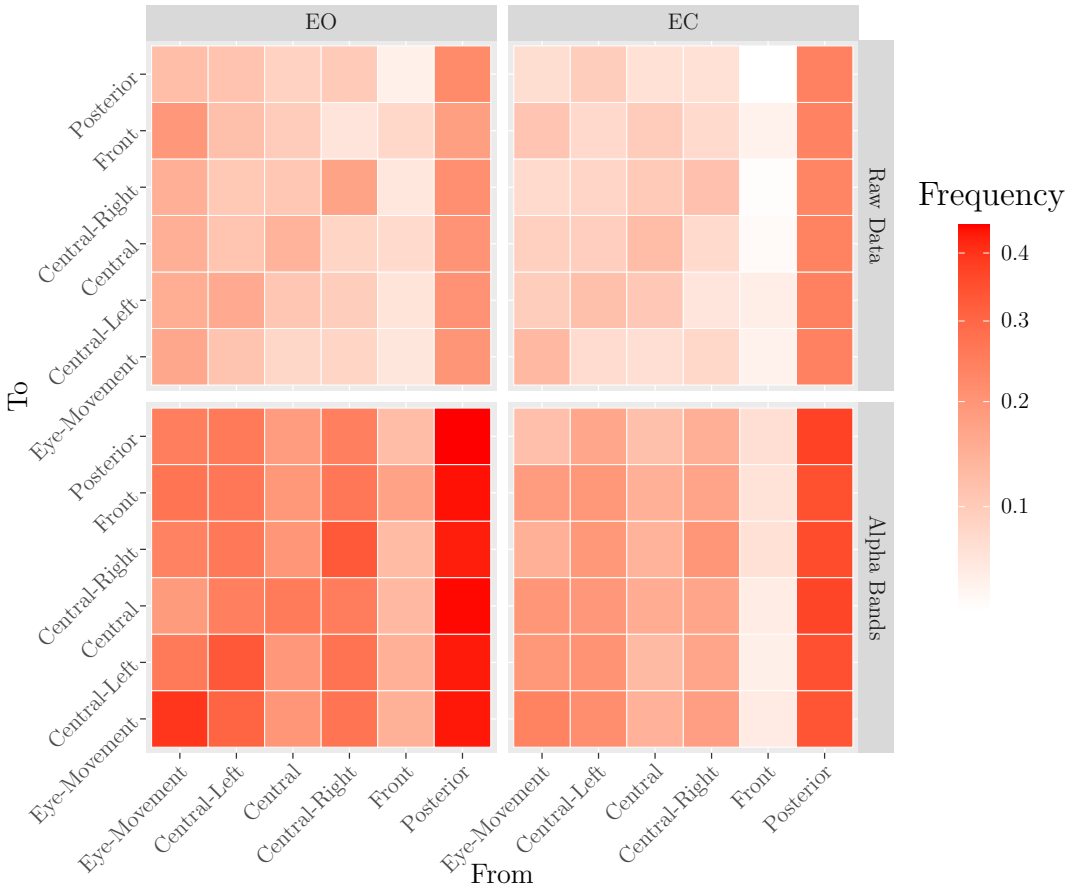

FIG 8. *The frequency of nonzero entries in the estimated sparse components $\widehat{\mathcal{S}}$ is summarized according to the channel clusters classified in Table 3.*

subject. The VAR model is expressed as in (1.1):

$$X_t^m = A_m X_{t-1}^m + \epsilon_t^m,\ \epsilon_t^m \sim N(0, \Sigma_m), \qquad A_m = W_m \Phi + S_m \tag{1.1}$$

where $A_m \in \mathbb{R}^{p\times p}$ is the coefficient matrix, and $\Sigma_m \in \mathbb{R}^{p\times p}$ is the covariance matrix for the model innovations.

REMARK A.1. For analytical simplicity, one may assume the noise covariance matrix is isotropic:

$$\Sigma_m = \sigma_m^2 I_p,$$

which aligns the log-likelihood function closely with the least squares objective used in estimation. Nonetheless, Σ_m can take more general forms with complex structure to model dependencies in high-dimensional settings. Our least squares estimation approach remains robust and applicable even under such general covariance structures.

We next introduce some standard definitions and key properties of VAR models, which will be useful in our subsequent analysis.

Recall that Assumption 1,

$$|\lambda_1(A_m)| < 1, \tag{A.1}$$

ensures a VAR model to be stationary (Lütkepohl, 2005). The inequality (A.1) can also be expressed with regard to the VAR model's characteristic polynomial $\mathcal{A}_m(z) = I_p - A_m z$ that $\det(\mathcal{A}_m(z)) \neq 0$ for $|z| \leqslant 1$. As a consequence, the spectral density of the VAR model (1.1), defined as

$$h_m(\vartheta) = \frac{1}{2\pi}(\mathcal{A}_m^{-1}(e^{i\vartheta}))\Sigma_m(\mathcal{A}_m^{-1}(e^{i\vartheta}))^\dagger, \quad \vartheta \in [-\pi, \pi], \tag{A.2}$$

is also bounded above.

The following quantities play an essential role in the estimation analysis and will be used to control the consistency rates established in Section 3:

$$\Psi(h_{X^m}) = \sup_{\vartheta\in[-\pi,\pi]} \lambda_1(h_{X^m}(\vartheta)),$$

$$\psi(h_{X^m}) = \sup_{\vartheta\in[-\pi,\pi]} \lambda_p(h_{X^m}(\vartheta)),$$

$$\tau_{\max}(\mathcal{A}_m) = \max_{|z|=1} \lambda_1(\mathcal{A}_m^\dagger(z)\mathcal{A}_m(z)),$$

$$\tau_{\min}(\mathcal{A}_m) = \min_{|z|=1} \lambda_p(\mathcal{A}_m^\dagger(z)\mathcal{A}_m(z)).$$

These quantities are also related in VAR models in terms of the following inequalities

$$\psi(h_{X^m}) \geqslant \frac{1}{2\pi}\frac{\lambda_p(\Sigma_m)}{\tau_{\max}(\mathcal{A}_m)}, \qquad \Psi(h_{X^m}) \leqslant \frac{1}{2\pi}\frac{\lambda_1(\Sigma_m)}{\tau_{\min}(\mathcal{A}_m)}.$$

APPENDIX B: PANEL VAR LITERATURE OVERVIEW & COMPARISONS

In this section, we provide a supplementary review of existing panel VAR (PVAR) models and compare them with our proposed approach.

A highly restrictive class of models assumes a common transition matrix shared across all entities, as studied extensively by Canova and Ciccarelli (2013, Section 3.1), Breitung (2015), Sigmund and Ferstl (2021), among others. These models leverage the entire data pool to estimate a single $p \times p$ matrix (for VAR(1)), typically using classical least squares. While simple and computationally efficient, the effectiveness of such models hinges on the assumption of homogenous dynamics across entities. When this assumption is violated, the model's validity deteriorates, often producing unreliable results.

At the other extreme, one can fully vectorize the panel data and model it with a large $Mp \times Mp$ transition matrix, allowing for maximal flexibility. However, for practical estimation, this flexibility is often curtailed by structural constraints, such as low-rank assumptions (Canova and Ciccarelli, 2013, Section 3.2), or sparsity and group sparsity, as in the Bayesian framework of Korobilis (2016). Despite their flexibility, these approaches sacrifice interpretability at the entity level. For instance, a low-rank structure imposed on the large transition matrix does not necessarily induce meaningful or identifiable patterns in the individual $p \times p$ blocks corresponding to each entity. Moreover, unstructured sparsity can inadvertently produce degenerate cases, such as exactly zero transition matrices for certain entities, leading to unintended random walk behavior.

An intermediate line of work retains individual VAR models for each entity while introducing structured inter-entity dependencies. For example, Skripnikov and Michailidis (2019) model the entity-specific matrices $A_1, \ldots, A_M$ as sparse perturbations of a shared baseline matrix, while Xu, Düker and Matteson (2024) explore commonalities in eigen-structures through post-estimation hypothesis testing. While these models relax strict homogeneity, they remain sensitive to structural assumptions, making them difficult to apply robustly to complex real-world data where such assumptions may not hold.

Our proposed model (1.1) follows this intermediate philosophy but introduces a more flexible and interpretable structure. Specifically, we link the M transition matrices through a shared low-rank basis, entity-specific rescaling effects, and sparse deviations. This formulation accommodates substantial heterogeneity while preserving clear interpretability at both the entity and panel level. As demonstrated in the main paper and through simulations, this setup strikes a balance between flexibility, interpretability, and robustness for analyzing high-dimensional panel VAR data.

B.1. Matrix Variate & Tensor Models. An alternative line of research for analyzing panel time series treats the data as a three-dimensional array and applies matrix-variate or tensor decomposition methods. In this section, we compare our proposed model with these approaches, focusing particularly on the matrix auto-regression (MAR) model proposed by Chen, Xiao and Yang (2021), as well as selected tensor factor models; see, for example, Chen, Yang and Zhang (2022) and Babii, Ghysels and Pan (2023).

The MAR model extends the classical VAR structure (1.1) to accommodate panel data in matrix format. Specifically, the panel time series are arranged as a matrix-valued process $Z_t = (X_t^1, \ldots, X_t^M) \in \mathbb{R}^{p\times M}$. To illustrate, consider the following MAR(1) example, which is conceptually comparable to the setup proposed in our work:

$$Z_t = \sum_{i=1}^{r} B_i^f Z_{t-1} B_i^b + \epsilon_t, \tag{B.1}$$

where $B_i^f \in \mathbb{R}^{p\times p}$, $B_i^b \in \mathbb{R}^{M\times M}$, and $\epsilon_t \in \mathbb{R}^{p\times M}$ are noise terms.

The MAR model can be interpreted as a combination of two VAR models: one operating along the columns and the other along the rows of the matrix-valued process. In particular, when the coefficient matrices satisfy $B_i^b = I_M$ (identity matrices), the MAR model in (B.1) simplifies to a panel of VAR models where every entity shares a common coefficient matrix $B^f = \sum_{i=1}^r B_i^f$, i.e., $Z_t[\cdot, m] = B^f Z_{t-1}[\cdot, m] + \epsilon_t[\cdot, m]$, which corresponds to the classical VAR formulation in (1.1) with $A_m \equiv B^f$) for all m. Alternatively if $B_i^f = I_p$, the MAR model (B.1) describes a panel of VAR models of dimension M, where the p variables are treated as entities. In this case, the shared coefficient matrix is $B^b = \sum_{i=1}^r B_i^b$ leading to $Z_t[i, \cdot] = Z_{t-1}[i, \cdot]B^b + \epsilon_t[i, \cdot]$.

Estimation of the coefficient matrices B_i^f and B_i^b typically follows an alternating regression procedure, where left and right pseudo-inverses are applied iteratively.

Analogous to the MAR setting, tensor factor models have also been explored in the context of panel time series analysis (Chen, Yang and Zhang, 2022; Babii, Ghysels and Pan, 2023). These models generalize (B.1) by introducing latent factors $F_t \in \mathbb{R}^{q\times N}$ to capture low-dimensional structures, resulting in the formulation:

$$Z_t = B^f F_t B^b + \epsilon_t, \tag{B.2}$$

where $B^f \in \mathbb{R}^{p\times q}$ and $B^b \in \mathbb{R}^{N\times M}$ are loading matrices, with $q < p$ and $N < M$ to capture the underlying lower-dimensional structure. Mathematically, this formulation is equivalent to concatenating the matrix-variate data matrices Z_t into a multi-way array $\mathcal{Z} \in \mathbb{R}^{p\times M\times T}$ and then performing inference from a tensor point of view. In this framework,classical tensor decomposition techniques, such as Tucker decomposition or canonical polyadic (CP) decomposition, can be employed to characterize cross-sectional dependencies. For instance, one may write (B.2) as the tensor decomposition with the outer product algebra

$$\mathcal{Z} = \mathcal{F} \times_1 B^f \times_2 (B^b)' \times_3 I_T + \epsilon,$$

where $\times_i$ denotes the usual tensor product of the the mode i. Relevant estimation procedures for such tensor models have been discussed in Han et al. (2024a,b).

Notably, the tensor factor model (B.2) does not, by itself, incorporate explicit temporal dynamics. To address this limitation, it is common to impose an MAR structure (B.1) on the latent factors $\mathcal{F}$; see Surana, Patterson and Rajapakse (2016). This leads to a tensor model with auto-regressive latent factors:

$$Z_t = B^f F_t B^b + \epsilon_t, \qquad F_t = \sum_{i=1}^{r} A_i^f F_{t-1} A_i^b + \varepsilon_t, \tag{B.3}$$

where A_i^f and A_i^b are autoregressive coefficient matrices of appropriate dimensions. This formulation effectively embeds temporal dependence into the model but does so under a strong assumption: the latent temporal dynamics governing all entities are driven by a shared set of core processes.

Compared to our LSPVAR model (1.1), a key distinction is that the MAR and tensor factor models explicitly mix panel data through post-coefficient (loading) matrices B_i^b. In contrast, the dependence across entities in our panel VAR framework is implicit, induced by algebraic constraints imposed on the set of transition matrices A_m, as illustrated in Figure 1.

In practice, both the MAR and tensor factor models can be overly restrictive for heterogeneous panels. If the entities exhibit significant heterogeneity, the bilinear structures in (B.1) and (B.2) may fail to capture the true underlying dynamics, especially for the tensor MAR factor model (B.3), which enforces that all entities share identical auto-regressive latent processes. In contrast, the proposed LSPVAR model flexibly accommodates heterogeneous structures while preserving interpretability at both the entity and panel levels.

On the other hand, if the panel is truly homogeneous, models like (B.1) and (B.2) may achieve more efficient recovery by leveraging their stronger mixture structure through pre- and post-loading matrices. However, as the number of entities M grows large relative to p or T, estimating the $M \times M$ loading matrices B_i^b in these models can become computationally expensive, particularly under classical least squares approaches. Moreover, when sparsity or low-rank constraints are introduced on B_i^b, these models begin to resemble the structure of the LSPVAR model

In summary, the posited model strikes a favorable balance by providing enhanced flexibility and interpretability in heterogeneous settings, making it particularly appealing for complex real-world applications.

APPENDIX C: ESTIMATION PROCEDURE

Next, we provide detailed descriptions of the subproblem updates within Algorithm 1, which iteratively optimizes the proposed objective function.

$$G(\mathcal{W}, \mathcal{S}, \Phi_c, \Phi, \Gamma; \mathcal{X}, \eta, \rho) = F(\mathcal{W}, \mathcal{S}, \Phi; \mathcal{X}, \eta) + \frac{\rho}{2}\|\Phi - \Phi_c\|_F^2 + \rho\langle\Gamma, \Phi - \Phi_c\rangle. \tag{2.6}$$

With slight abuse of notation, we continue to use G to denote the the objective functions for the subproblems, where the precise form and inputs of G may vary depending on the context. The following subsections detail the update routines for each estimator at the $(i+1)$-th iteration.

C.1. Subproblem of $(\mathcal{W}, \mathcal{S})$. The first block update involves optimizing (2.6) with respect to $(\mathcal{W}, \mathcal{S})$. This subproblem can be viewed as a classical LASSO-type regression with blockwise penalties. Thanks to the blockwise convexity of the objective, the optimization is well-posed and can be efficiently solved using standard algorithms developed for high-dimensional penalized regression.

Moreover, this subproblem naturally decomposes into independent optimizations across entities. Specifically, for the m-th entity, the objective function simplifies to:

$$G^{(i+1)}(W_m, S_m) \propto \frac{1}{2T}\|Y_m - W_m\Phi^{(i)}X_m - S_mX_m\|_F^2 + \eta\|S_m\|_1, \tag{C.1}$$

and the optimization is separable row-wise with LASSO-type solutions.

For the purpose of proving the sufficient descent property (Proposition 2.1) of an update by optimizing (C.1), we may further split and optimize the two arguments W_m and S_m sequentially, as (C.1) is strongly convex with respect to W_m and S_m separately when Assumption 4 is satisfied.

C.2. Subproblem for Φ_c. Next, we consider the subproblem of optimizing Φ_c. This task essentially involves finding a feasible point in the intersection of two constraint sets, $\mathbb{L}_p(\hat{r}, \ell)$ and $\mathbb{B}_p$. To address this, we draw inspiration from Dykstra's algorithm (Boyle and Dykstra, 1986), which computes the nearest projection onto the intersection by iteratively projecting onto each set.

We implement this via an internal ADMM routine, formulated as a proximal problem with an additional step size parameter κ, which will be specified later in Appendix G (see the proof of Proposition 2.1). Specifically,

$$G^{(i+1)}(\Phi_c) = \frac{\rho}{2}\|\Phi^{(i)} + \Gamma^{(i)} - \Phi_c\|_F^2 + \frac{\kappa\rho}{2}\|\Phi_c - \Phi_c^{(i)}\|_F^2,$$

and it can be reformulated under the ADMM framework by introducing auxiliary and dual variables as

$$G_c(\Phi_B, \Phi_L, \Gamma_{BL}) = \frac{1}{2}\|\Phi_0^{(i)} - \Phi_L\|_F^2 + \frac{1}{2}\|\Phi_B - \Phi_L\|_F^2 + \langle\Gamma_{BL}, \Phi_B - \Phi_L\rangle, \tag{C.2}$$

where $\Phi_0^{(i)} = \frac{\Phi^{(i)} + \Gamma^{(i)} + \kappa\Phi_c^{(i)}}{1+\kappa}$. Analogously we solve it by blockwise updates. For instance, at the k-th iteration,

1. Solve $\Phi_L^{(k+1)} = \operatorname{argmin}_{\Phi_L \in \mathbb{L}_p(\hat{r},\ell)} \|\Phi_L - \frac{1}{2}\left(\Phi_0^{(i)} + \Phi_B^{(k)} + \Gamma_{BL}^{(k)}\right)\|_F^2$. The solution is indeed the projection of $\frac{1}{2}\left(\Phi_0^{(i)} + \Phi_B^{(k)} + \Gamma_{BL}^{(k)}\right)$ onto the set $\mathbb{L}_p(\hat{r}, \ell)$. Explicit form can be found below in Section C.2.1.
2. Solve $\Phi_B^{(k+1)} = \operatorname{argmin}_{\Phi_B \in \mathbb{B}_p} \|\Phi_B - \left(\Phi_L^{(k+1)} - \Gamma_{BL}^{(k)}\right)\|_F^2$, which is the projection of $\Phi_L^{(k+1)} - \Gamma_{BL}^{(k)}$ onto the set $\mathbb{B}_p$. Details are covered in Section C.2.2.
3. Update $\Gamma_{BL}^{(k+1)} = \Gamma_{BL}^{(k)} + \Phi_B^{(k+1)} - \Phi_L^{(k+1)}$.

For notational simplicity, we denote by Π_B and Π_L the projection operators onto the sets $\mathbb{B}_p$ and $\mathbb{L}_p(r, \ell)$, respectively.

C.2.1. *Subproblem for Φ_L.* Note that any $\Phi_L \in \mathbb{L}_p(\hat{r}, \ell)$ can be expressed via its singular value decomposition (SVD) as

$$\Phi_L = UDV' \tag{C.3}$$

where U, V are all orthogonal matrices in $\mathbb{R}^{p\times p}$, $D = \operatorname{diag}(d) \in \mathbb{R}^{p\times p}$, $d = (d_1, \ldots, d_p) \in \mathbb{R}_+^p$ and $\|d\|_1 = \ell$. Note that here we restrict the parameterization (C.3) to be at most rank-$\hat{r}$, then at most $\hat{r}$ elements of d are potentially non-zero, and hence only the corresponding $\hat{r}$ columns of U and V matter in our optimization.

With the expression (C.3) and condition on $\Phi_{L,0}^{(k)} := \frac{1}{2}\left(\Phi_0^{(i)} + \Phi_B^{(k)} + \Gamma_{BL}^{(k)}\right)$, the objective function of $\Phi_L = UDV'$ can be viewed as

$$\begin{aligned} G_c^{(k+1)}(\Phi_L) &\propto \operatorname{tr}(\Phi_L\Phi_L' - 2\Phi_{L,0}^{(k)}\Phi_L') \\ &\Leftrightarrow G_c^{(k+1)}(U,V,D) \propto \operatorname{tr}(D^2 - DU'\Phi_{L,0}^{(k)}V). \end{aligned} \tag{C.4}$$

The optimization can be sequentially decomposed into two main steps, as formalized in the following lemmas, which separately address the updates of the singular vector matrices and the singular values, respectively.

LEMMA C.1. *Conditioned on a given matrix $\Phi_{L,0}^{(k)}$, the objective function (C.4) attains its minimum at $(U,V) = (U_0,V_0)$, if and only if U_0 and V_0 are the left and right singular vector matrices of $\Phi_{L,0}^{(k)}$, respectively.*

PROOF. This result corresponds to a well-known inequality frequently encountered in matrix inner product analysis. Nonetheless, we provide here a more intricate proof that offers a different perspective.

The following parameterization for orthogonal matrices facilitates deriving a closed-form solution for this subproblem. Specifically, we consider perturbations of the pair (U,V) around a candidate optimizer (U_0,V_0) using the matrix exponential of skew-symmetric matrices, a popular tool for exploring neighborhoods of orthogonal matrices. The parameterization is given by:

$$U(k_u;K_U,U_0) = U_0e^{k_uK_U}, \quad V(k_v;K_V,V_0) = V_0e^{k_vK_V},$$

where $(k_u,k_v) \in \mathbb{R}^2$, and $K_U, K_V \in \mathbb{R}^{p\times p}$ are skew-symmetric matrices. Then, the objective function as a variant of (C.4) is introduced with $\Phi_L = U(k_u;K_U,U_0)DV(k_v;K_V,V_0)'$ plugged in,

$$G_{U_0,V_0}^{(k+1)}(k_u,k_v,D;K_U,K_V) = \operatorname{tr}(D^2 - 2De^{-k_uK_U}U_0'\Phi_{L,0}^{(k)}V_0e^{k_vK_V}).$$

This reformulation allows us to analyze the minimizer by examining the behavior of $G_{U_0,V_0}^{(k+1)}$ with respect to small perturbations around (U_0,V_0).

The following claim, which follows naturally from critical point analysis, provides a necessary and sufficient condition for when the optimizer Φ_L of the subproblem admits the singular vector matrices U_0 and V_0. Specifically, the objective function

$$\begin{aligned} G_c^{(k+1)}(U_0,V_0,D) &= G_{U_0,V_0}^{(k+1)}(0,0,D;K_U,K_V) \\ &= \operatorname{tr}(D^2 - DV_0'(\Phi_{L,0}^{(k)})'U_0 - DU_0'\Phi_{L,0}^{(k)}V_0) \end{aligned} \tag{C.5}$$

attains a (local) minimum of (C.4) with respect to all orthogonal pairs (U,V), given a fixed $\Phi_{L,0}^{(k)}$, if and only if (U_0,V_0) correspond to the left and right singular vector matrices of $\Phi_{L,0}^{(k)}$.

CLAIM C.1. *Conditioned on a given $\Phi_{L,0}^{(k)}$, a pair of matrices (U_0,V_0) corresponds to the left and right singular vectors of Φ_L as defined in (C.3), and Φ_L optimizes the objective function (C.4), if and only if, for any arbitrary skew-symmetric matrix pair (K_U,K_V), the following holds:*

$$\begin{cases} \frac{\partial}{\partial k_u} G_{U_0,V_0}^{(i+1)}(0,0,D;K_U,K_V) = 0, \\ \frac{\partial}{\partial k_v} G_{U_0,V_0}^{(i+1)}(0,0,D;K_U,K_V) = 0. \end{cases} \tag{C.6}$$

Indeed, solving the system (C.6), with $\Theta = U_0' \Phi_{L,0}^{(k)} V_0$, we have

$$\begin{cases} D\Theta - \Theta' D = 0, \\ D\Theta' - \Theta D = 0. \end{cases}$$

The above system implies that the only valid solution requires the matrix Θ to be diagonal. Hence, the matrices U_0 and V_0 are the left and right singular vectors of $\Phi_{L,0}^{(k)}$, respectively. □

Based on Lemma C.1, set $\Theta = \mathrm{diag}(\theta) = U_0' \Phi_{L,0}^{(k)} V_0$ with $\theta = (\theta_1, \ldots, \theta_p)$ being the singular values of $\Phi_{L,0}^{(k)}$, recall that $D = \mathrm{diag}(d)$, then (C.4) can be derived as

$$G_c^{(k+1)}(U_0, V_0, D) \propto \sum_{i=1}^{r} (d_i^2 - 2\theta_i d_i), \quad \text{subject to } \|d\|_1 = \ell,\ \|d\|_0 \leqslant \hat{r},\ d_i \geqslant 0. \tag{C.7}$$

The optimization can be efficiently handled using classical quadratic programming techniques over a simplex (Frank and Wolfe, 1956). In fact, the following lemma provides a closed-form solution to (C.7).

LEMMA C.2. *Define the cumulative sum-type function as*

$$CS_\theta(j) = \frac{\sum_{k=1}^{j} \theta_k - \ell}{j}, \quad j = 1, \ldots, \hat{r}.$$

Then, the optimizer to (C.7) *at the* $(k+1)$*-th iteration is given by*

$$\hat{d} = \mathbb{P}_\gamma(\theta) - CS_\theta(\gamma) \cdot (\mathbf{1}_\gamma', 0)', \quad \textit{where } \gamma = \max\{j : CS_\theta(j) \leqslant \theta_j, j = 1, \ldots, \hat{r}\}.$$

Here $\mathbb{P}_\gamma(\theta)$ *denotes the projection of the vector* θ *onto the subspace spanned by its first* γ *elements* .

REMARK C.1. Note that when we input the extreme case $\hat{r} = p$, the above solution indicates that the quadratic programming may automatically induce sparsity on $\hat{d}$, since it is likely that $\gamma < p$. Hence, instead of tuning the hyper-parameter $\hat{r}$, the true rank r may be estimated as γ by our algorithm adaptively from the data when an over-specified $\hat{r}$ is used. Further discussion on this adaptive behavior is provided in our consistency analysis; see Remark 3.2 and corresponding simulation results in Section 4.1.

In summary, the results from the preceding lemmas lead to the following proposition, which characterizes the solution to the subproblem for Φ_L, that is, the projection $\Pi_L(\Phi_{L,0}^{(k)})$.

PROPOSITION C.1. *Given* $\Phi_{L,0}^{(k)}$*, the subproblem* (C.4) *admits a unique minimizer* $\Phi_L^{(k+1)}$ *of the form*

$$\Phi_L^{(k+1)} = U_0 \hat{D} V_0', \tag{C.8}$$

where U_0 *and* V_0 *are the matrices containing the left and the right singular vectors of* $\Phi_{L,0}^{(k)}$*, respectively, and* $\hat{D} = \mathrm{diag}(\hat{d})$ *is given by the solution in Lemma C.2.*

PROOF. It follows directly from Lemma C.1 and Lemma C.2 that Proposition C.1 holds. Therefore, it remains to prove Lemma C.2.

To this end, assume that U_0 and V_0 are full $p \times p$ matrices containing the singular vector of $\Phi_{L,0}^{(k)}$, and the low-rank structure of Φ_L is induced by sparsifying the diagonal matrix of singular values D. The following claim characterizes how this low-rank structure influences the optimization in (C.7).

CLAIM C.2. *The rank-$\hat{r}$ minimizer D of (C.7) (with $\hat{r} \leqslant p$) has nonzero entries only on the diagonal of its leading principal $\hat{r} \times \hat{r}$ submatrix, denoted by D_0.*

Claim C.2 can be verified by contradiction: if there exists a rank-r minimizer with a nonzero diagonal element outside D_0, one can construct an alternative D by moving this nonzero element into D_0, without increasing the objective in (C.7), leveraging the non-increasing property of the diagonal entries of Θ.

Given Claim C.2, Lemma C.2 follows by sequentially allocating the total nuclear norm budget ℓ to the diagonal entries of $\hat{d}$ in descending order, until the threshold specified in Lemma C.2 is reached. □

C.2.2. *Subproblem of Φ_B.* The corresponding objective function has the form

$$G_c^{(k+1)}(\Phi_B) \propto \operatorname{tr}(\Phi_B \Phi_B' - 2\Phi_{B,0}^{(k)} \Phi_B'),$$

where $\Phi_{B,0}^{(k)} = \Phi_L^{(k+1)} - \Gamma_{BL}^{(k)}$.

Next, if we assume that the rows of $\Phi_B \in \mathbb{B}_p$ have norms equal to K, then

$$\begin{aligned} G_c^{(k+1)}(\Phi_B) &\propto pK^2 - 2\sum_{i=1}^{p} e_i' \Phi_{B,0}^{(k)} \Phi_B' e_i \\ &\geqslant pK^2 - 2K \sum_{i=1}^{p} \|e_i' \Phi_{B,0}^{(k)}\|, \end{aligned}$$

with the equality obtained when there exist positive constants $k_i > 0$ such that $e_i' \Phi_B = k_i e_i' \Phi_{B,0}^{(k)}$ for $i = 1, \ldots, p$. The optimal K can then be derived as $K = \frac{1}{p} \sum_{i=1}^{p} \|e_i' \Phi_{B,0}^{(k)}\|$, the average norm of the rows in the matrix $\Phi_{B,0}^{(k)}$. Note that the derivation requires that $e_i' \Phi_{B,0}^{(k)} \neq 0$ for all i. The solution leads to the following proposition introducing the property of the projection operator Π_B onto $\mathbb{B}_p$.

PROPOSITION C.2. *For arbitrary $x \in \mathbb{R}^{p \times p}$ that $e_i x \neq 0$ for all i, there exists a positive definite diagonal matrix $A = \operatorname{diag}(a_1, \ldots, a_p)$ with $\operatorname{tr}(A) = p$, such that $y_0 = \Pi_B(x)$ and $x = A y_0$.*

C.2.3. *Convergence Analysis.* For the convergence of this subproblem, we provide an analogous roadmap for proving Theorem 2.1 below.

1. The objective function G_c exhibits sufficient descent in every iteration (Proposition C.3).
2. The evaluations of the objective function G_c are bounded below (Proposition C.4).
3. The norm of the subgradients are convergent to zero.
4. The objective function G_c satisfies the KL-property.

PROPOSITION C.3. *Let the subproblem for Φ_L in Section C.2.1 adopt the proximal setup, and with the notations from Proposition C.2, assume that a_i (obtained from the projection of $\Phi_{B,0}^{(k)}$) are uniformly lower bounded. Then, there exists a constant $\mathcal{C} > 0$ such that*

$$G_c(\Phi_B^{(k)}, \Phi_L^{(k)}, \Gamma_{BL}^{(k)}) - G_c(\Phi_B^{(k+1)}, \Phi_L^{(k+1)}, \Gamma_{BL}^{(k+1)}) \geqslant \mathcal{C} \left(\|\Phi_B^{(k)} - \Phi_B^{(k+1)}\|_F^2 + \|\Phi_L^{(k)} - \Phi_L^{(k+1)}\|_F^2 \right).$$

Indeed, Proposition C.3 is a straightforward corollary of the proximal setup of the Φ_L's subproblem and the three-point property of $\mathbb{B}_p$ (as stated in Proposition D.1 in Appendix D), particularly given that the a_i's of the matrices $\Phi_{B,0}^{(k)}$ are bounded below according to the assumption in Proposition D.1.

PROPOSITION C.4. *The evaluations of the objective function G_c (C.2) at the sequence of estimators $(\Phi_B^{(k)}, \Phi_L^{(k)}, \Gamma_{BL}^{(k)})$ are lower bounded.*

PROOF. We note that the objective function satisfies

$$\begin{aligned} G_c^{(k)} &:= \frac{1}{2}\|\Phi_0^{(i)} - \Phi_L^{(k)}\|_F^2 + \frac{1}{2}\|\Phi_B^{(k)} - \Phi_L^{(k)}\|_F^2 + \langle \Gamma_{BL}^{(k)}, \Phi_B^{(k)} - \Phi_L^{(k)} \rangle \\ &\geqslant \frac{1}{2}\|\Phi_0^{(i)} - \Phi_L^{(k)}\|_F^2 + \frac{1}{2}\|\Phi_B^{(k)} - \Phi_L^{(k)}\|_F^2 - \frac{1}{2}\|\Gamma_{BL}^{(k)}\|_F^2 - \frac{1}{2}\|\Phi_B^{(k)} - \Phi_L^{(k)}\|_F^2 \\ &\geqslant \frac{1}{2}\|\Phi_0^{(i)} - \Phi_L^{(k)}\|_F^2 - \frac{1}{2}\|\Gamma_{BL}^{(k)}\|_F^2. \end{aligned}$$

Then it suffices to show that $\Gamma_{BL}^{(k)}$ is upper bounded. We prove it by induction. Assume there is a constant $\Upsilon > 0$ that $\|\Phi_L^{(k+1)} - \Gamma_{BL}^{(k)}\|_F \leqslant \Upsilon$, then $\|\Gamma_{BL}^{(k+1)}\|_F \leqslant \sqrt{\frac{p-1}{p}}\Upsilon$, and hence noting that $\|\Phi_L^{(k+2)}\|_F \leqslant \ell$, we have

$$\|\Phi_L^{(k+2)} - \Gamma_{BL}^{(k+1)}\|_F \leqslant \ell + \sqrt{\frac{p-1}{p}}\Upsilon \leqslant \Upsilon,$$

given any $\Upsilon \geqslant 2p\ell$. Note that when $k = 0$, $\|\Phi_L^{(k+1)} - \Gamma_{BL}^{(k)}\|_F = \|\Phi_L^1\|_F \leqslant \ell$, we can simply choose $\Upsilon = 2p\ell$, and hence the induction is valid. □

The KL-property of the two sets will be established later in Appendix G in the proof of Theorem 2.1. Hence, as a result, the subproblem is convergent irrespective of the initialization. Indeed, with the sufficient descent property and our choice of initialization $\Phi_L^{(0)} = \Phi_B^{(0)} = \Phi_c^{(i)}$, use the convergence result to set $\Phi_L^{(\infty)} = \Phi_B^{(\infty)} = \Phi_c^{(i+1)}$. We then obtain

$$\frac{\rho}{2}\|\Phi^{(i)} + \Gamma^{(i)} - \Phi_c^{(i)}\|_F^2 \geqslant \frac{\rho}{2}\|\Phi^{(i)} + \Gamma^{(i)} - \Phi_c^{(i+1)}\|_F^2 + \frac{\kappa\rho}{2}\|\Phi_c^{(i)} - \Phi_c^{(i+1)}\|_F^2,$$

which will be used in the proof of Proposition 2.1.

C.3. Subproblem of Φ. The next primal subproblem focuses on updating the matrix Φ. The corresponding objective function with respect to Φ is

$$\begin{aligned} G^{(i+1)}(\Phi) \propto \operatorname{tr}(\rho\Phi\Phi' + \sum_{m=1}^{M} \frac{1}{T} W_m^{(i+1)} \Phi X_m X_m' \Phi' W_m^{(i+1)}) \\ - 2\operatorname{tr}(\Phi'(\rho(\Phi_c^{(i+1)} - \Gamma^{(i)}) + \sum_{m=1}^{M} \frac{1}{T} W_m^{(i+1)}(Y_m - S_m^{(i+1)} X_m) X_m')). \end{aligned}$$

The first order optimality condition yields

$$\rho\Phi + \sum_{m=1}^{M} (W_m^{(i+1)})^2 \Phi \frac{X_m X_m'}{T} = \rho(\Phi_c^{(i+1)} - \Gamma) + \sum_{m=1}^{M} \frac{1}{T} W_m^{(i+1)}(Y_m - S_m^{(i+1)} X_m) X_m'.$$

This equation is row-wise separable and admits explicit solutions for each row, given the other variables $(\mathcal{W}, \mathcal{S}, \Phi_c, \Gamma)$ fixed.

C.4. Dual Ascent Update of Γ. Finally, the update of the dual variable Γ is given by

$$\Gamma^{(i+1)} \mapsto \Gamma^{(i)} + (\Phi^{(i+1)} - \Phi_c^{(i+1)}). \tag{C.9}$$

APPENDIX D: CONSTRAINT SPACE PROPERTIES

In this section, we discuss properties of the constraint spaces, with a particular focus on the intersection $\mathbb{B}_p \cap \mathbb{L}_p(r,\ell)$. Specifically, Section D.1 highlights properties that make $\mathbb{B}_p$ behave similarly to a convex space, while Section D.2 examines the tangent spaces and normal cones of both $\mathbb{B}_p$ and $\mathbb{L}_p(r,\ell)$ characterizing their intersection.

D.1. Convex-like Property of $\mathbb{B}_p$. We begin by presenting the three-point property of the set $\mathbb{B}_p$ (see Zhu and Li, 2019, Assumption 1). The following proposition states this property, and its proof follows straightforwardly. This three-point property exhibits characteristics similar to those of (strongly) convex sets in certain respects.

PROPOSITION D.1 (Three-point Property). *Under the conditions of Proposition C.2, if there exists a constant $0 < k_B \leqslant 1$ such that $a_i \geqslant k_B > 0$, then the set $\mathbb{B}_p$ satisfies the three-point property at x. Specifically, for arbitrary $y \in \mathbb{B}_p$,*

$$\|x - y\|_F^2 - \|x - y_0\|_F^2 \geqslant k_B \|y - y_0\|_F^2.$$

PROOF. For an arbitrary $y \in \mathbb{B}_p$,

$$\begin{aligned}
\|x - y\|_F^2 - \|x - y_0\|_F^2 &= \|y\|_F^2 - \|y_0\|_F^2 - 2\langle x, y - y_0\rangle \\
&= \|y\|_F^2 + \langle y_0, (2A - I_p) y_0\rangle - 2\langle A y_0, y\rangle \\
&= \|y\|_F^2 + \|y_0\|_F^2 - 2\langle A y_0, y\rangle \\
&= \|y\|_F^2 + \|y_0\|_F^2 - \frac{2}{p}\sum_{i=1}^{p} a_i \|y\|_F \|y_0\|_F \cos(\omega_i),
\end{aligned}$$

where ω_i is the angle of the i-th rows of the two matrices y_0 and y. In order to prove the property with some constant k (to be determined), it suffices to show that with $0 < k \leqslant 1$,

$$(1-k)\left(\|y\|_F^2 + \|y_0\|_F^2\right) - \frac{2}{p}\|y\|_F \|y_0\|_F \sum_{i=1}^{p} (a_i - k)\cos(\omega_i) \geqslant 0$$

for all y and y_0 satisfying our assumptions. Some algebra then implies that

$$k \leqslant \min\left(\frac{p - \sum_{i=1}^p a_i \cos(\omega_i)}{p - \sum_{i=1}^p \cos(\omega_i)}, \frac{p + \sum_{i=1}^p a_i \cos(\omega_i)}{p + \sum_{i=1}^p \cos(\omega_i)} \right).$$

Now given that $\sum_{i=1}^p a_i = p$, $a_i \geqslant k_B$ and $\cos(\omega_i) \in [-1, 1]$, we can then derive that the right hand side of the above inequality is lower bounded by k_B, and hence the three point property holds with the constant $k = k_B$.

□

REMARK D.1. Note that Proposition D.1 plays a crucial role in proving Proposition C.3 by setting $z = \Phi_B^{(k)}$ and $y = \Phi_B^{(k+1)}$. Further, Proposition D.1 leads to the following inequality,

$$\langle x - y_0, y - y_0\rangle \leqslant \frac{1 - k_B}{2}\|y - y_0\|_F^2.$$

Although this condition is still weaker than full convexity, it effectively controls the dual ascent steps in the algorithm described in Section C.2.

D.2. Tangent Spaces & Normal Cones of $\mathbb{B}_p$ and $\mathbb{L}_p(r,\ell)$. Assume there exists a matrix $\Phi \in \mathbb{B}_p \cap \mathbb{L}_p(r,\ell)$, where

$$\Phi = UDV' = (\Phi_1', \dots, \Phi_p')'$$

is of rank-$\tilde{r}$ ($\tilde{r} \leqslant r$) with $U, V \in \mathbb{R}^{p\times\tilde{r}}$ being singular vector matrices and $D = \operatorname{diag}(d_1, \dots, d_{\tilde{r}})$ storing the singular values that $\sum_{i=1}^{\tilde{r}} d_i = \ell$. Below we analyze the tangent spaces and the normal cones of the two spaces at the intersection Φ.

Regarding the space $\mathbb{B}_p$, we can derive from its definition that the tangent space and normal cone can be expressed as

$$\mathcal{T}_{\mathbb{B}_p}(\Phi) = \{\Delta\Phi = (\Delta\Phi_1', \dots, \Delta\Phi_p')' : \Phi_i \Delta\Phi_i' = \Phi_j \Delta\Phi_j',\ \forall i \neq j\}.$$

$$\mathcal{N}_{\mathbb{B}_p}(\Phi) = \{H = \operatorname{diag}(b_1, \dots, b_p)\Phi : \sum_{i=1}^{p} b_i = 0\}.$$

For the low-rank component, we draw on the works of Schneider and Uschmajew (2015); Hosseini, Luke and Uschmajew (2019); Li and Luo (2023) and note that the stratified space

$$\mathbb{L}_p^{\tilde{r}}(\ell) = \{\Phi : \operatorname{rk}(\Phi) = \tilde{r},\ \|\Phi\|_* = \ell,\ \lambda_i(\Phi\Phi') \neq \lambda_j(\Phi\Phi'),\ i \neq j\}$$

is a smooth manifold as it restricts all the singular values to be away from zero and of order 1. In addition, define

$$\mathbb{L}_p^{\tilde{r}}(\ell, U, V) = \mathbb{L}_p^{\tilde{r}}(\ell) \cap \{\Phi : U'\Phi V \text{ diagonal}\}.$$

Then, the tangent space and normal cone can be derived with the following form,

$$\mathcal{T}_{\mathbb{L}_p^{\tilde{r}}(\ell)}(\Phi) = \{\Delta\Phi : U_\perp \Delta\Phi V_\perp' = 0,\ \operatorname{tr}(U'\Delta\Phi V) = 0\} := \mathcal{T}(U,V),$$

$$\mathcal{N}_{\mathbb{L}_p^{\tilde{r}}(\ell)}(\Phi) = \{H = U_\perp E V_\perp' : E \in \mathbb{R}^{(p-\tilde{r})\times(p-\tilde{r})}\} \oplus UV' := \mathcal{N}(U,V) \oplus UV',$$

where the orthogonal complement $U_\perp$ and $V_\perp$ can be arbitrary as long as $U'U_\perp = V'V_\perp = 0$. Finally, note that $\mathbb{L}_p^{\tilde{r}}(\ell) \subset \mathbb{L}_p(r,\ell)$, the tangent space and normal cone of $\mathbb{L}_p(r,\ell)$ at Φ can be derived as

$$\mathcal{T}_{\mathbb{L}_p(r,\ell)}(\Phi) = \{\Delta\Phi + H - kUV' : \Delta\Phi \in \mathcal{T}(U,V),\ H \in \mathcal{N}(U,V),\ \operatorname{rk}(H) \leqslant r - \tilde{r},\ \|H\|_* = \tilde{r}k\}.$$

$$\mathcal{N}_{\mathbb{L}_p(r,\ell)}(\Phi) = \{H \in \mathcal{N}(U,V) : \operatorname{rk}(H) \leqslant p - r\} \oplus UV'.$$

Note that the closest rank-$\tilde{r}$ approximation to any rank-r ($r > \tilde{r}$) matrix preserves the singular vector structure. Therefore, any sequence satisfies $\Phi_n \to \Phi$ ($n \to \infty$) with $\Phi_n \in \mathbb{L}_p^{r_n}(\ell)$ ($r_n > \tilde{r}$), one can design $\delta\ell_n > 0$ and $\delta\ell_n \to 0$, such that

$$\Phi_n - (\underline{\Phi}_n + H_n) \to 0, \text{ where } \begin{cases} \underline{\Phi}_n \in \mathbb{L}_p^{\tilde{r}}(l - \delta\ell_n, U, V), \\ H_n \in \mathcal{N}(U,V),\ \operatorname{rk}(H_n) = r_n - \tilde{r},\ \|H_n\|_* = \delta\ell_n. \end{cases}$$

Further, note that for $\underline{\Phi}_n \in \mathbb{L}_p^{\tilde{r}}(l - \delta\ell_n, U, V)$,

$$\delta\Phi_n := \Phi - \underline{\Phi}_n = U\delta D_n V' = k_n UV' + U(\delta D_n - k_n I_{\tilde{r}})V' \subset UV' \oplus \mathcal{T}_{\tilde{r}}(U,V)$$

with $k_n = \frac{\delta\ell_n}{\tilde{r}}$ and $\tilde{r}k_n = \|H_n\|_*$, then

$$\Delta\Phi_n := H_n - \delta\Phi_n \in \mathcal{T}_{\mathbb{L}_p(r,\ell)}(\Phi),\ (\Phi_n - \Phi) - \Delta\Phi_n \to 0.$$

The other direction is rather straightforward.

In summary, if the left singular vectors of Φ in U are not standard basis vectors, we can then verify that

$$\mathcal{N}_{\mathbb{L}_p(r,\ell)}(\Phi) \cap \mathcal{N}_{\mathbb{B}_p}(\Phi) = \{0\},$$

which implies that Φ is a linearly regular intersection of the two sets. This result suggests that any alternating projection or averaged projection algorithms exhibit local convergence.

APPENDIX E: RANK-SPARSITY INCOHERENCE

To ensure identifiability between the basis Φ and the sparse components $\mathcal{S}$ in (1.1), certain incoherence conditions are required. Following Hsu, Kakade and Zhang (2011), we consider the following ones.

Assume the setup (1.1), where each coefficient matrix decomposes as $A_m = W_m\Phi + S_m$. Corresponding to this decomposition, define the following matrix spaces:

$$\Omega_m = \{S \in \mathbb{R}^{p\times p} : \operatorname{supp}(S) \subset \operatorname{supp}(S_m) = \operatorname{supp}(W_m^{-1}S_m)\}, \quad m = 1,\dots,M$$
$$\Phi = \{\Phi \in \mathbb{R}^{p\times p} : \Phi = \Phi_1 + \Phi_2, \Phi_1 \in \operatorname{range}(\Phi), \Phi_2' \in \operatorname{range}(\Phi')\}.$$

The projections of an arbitrary matrix $A \in \mathbb{R}^{p\times p}$ onto the two spaces (Ω_m, Φ) are given by

$$\Pi_{\Omega_m}(A) = (A[i,j]\cdot \mathbb{1}\{(i,j)\in \operatorname{supp}(S_m)\})_{i,j=1}^p,$$
$$\Pi_{\Phi}(A) = UU'A + AVV' - UU'AVV',$$

respectively, where $U, V \in \mathbb{R}^{p\times r}$ are the left and right orthonormal singular vector matrices of Φ respectively, and r is the rank of Φ.

DEFINITION E.1. For the coefficient matrices $A_m = W_m\Phi + S_m$ for $m = 1,\dots,M$, define the following two quantities:

$$\mu_m(\varsigma) = \max\{\varsigma\|\operatorname{sign}(S_m)\|_{1\to 1}, \varsigma^{-1}\|\operatorname{sign}(S_m)\|_{\infty\to\infty}\}, \tag{E.1}$$
$$\nu(\varsigma) = \varsigma^{-1}\|UU'\|_\infty + \varsigma\|VV'\|_\infty + \|U\|_{2\to\infty}\|V\|_{2\to\infty}, \tag{E.2}$$

where $\varsigma > 0$ is a balancing parameter to adjust for disparity between the number of rows and columns.

Note that the coefficient matrices A_m are all square matrices of dimension $p \times p$, hence ς can typically be set to 1. The first quantity μ_m measures the number of nonzero entries in the sparse matrix S_m, while the second quantity ν quantifies the sparsity of the low-rank basis Φ. In addition, based on Definition E.1, we impose the following assumption for estimation guarantee, which will be detailed in the sequel.

ASSUMPTION 5. For arbitrary m, $\inf_\varsigma \nu(\varsigma)\mu_m(\varsigma) < 1$.

PROPOSITION E.1. *Given the model* (1.1) *and under Assumption 5, with $\mathcal{W}$ fixed, each coefficient matrix A_m admits a unique decomposition into the pair (Φ, S_m) for $m = 1,\dots,M$.*

REMARK E.1. We do not impose a specific random sparsity model on the matrices in $\mathcal{S}$, nor do we assume particular properties on the singular vectors of the low-rank matrix Φ. Therefore, it is challenging to derive explicit estimation constraints directly from Assumption 5. To better understand how Algorithm 1's performance relates to the intrinsic low-rank plus sparse structure, we empirically examine violations of Assumption 5 in cases where the estimation results are suboptimal.

The key idea behind the proof of Proposition E.1 is to show that, under Assumption 5, the intersection set satisfies $\Omega_m \cap \Phi = \{0\}$ for every m. In particular, assume there is $M \in \Omega_m \cap \Phi$, then

$$\Pi_{\Omega_m}\big(\Pi_\Phi(M)\big) = M,$$

and using the triangular inequality, we obtain

$$\|M\| \leqslant \mu_m(\varsigma)\nu(\varsigma)\|M\|.$$

By Assumption 5, the product $\mu_m(\varsigma)\nu(\varsigma) < 1$, which forces $\|M\| = 0$.

REMARK E.2. Despite the discussion in Remark E.1, we emphasize that Proposition E.1 provides only a sufficient condition for unique decomposition. Therefore, the manual verification approach suggested in Remark E.1 should be viewed as a heuristic or supplementary diagnostic tool, rather than a definitive criterion, especially when the algorithm exhibits poor performance.

APPENDIX F: MULTI-LAG LSPVAR EXTENSION

A natural extension of the LSPVAR model (1.1) to higher-order lags can be handled using analogous methods. For instance, consider the lag-q model

$$X_t^m = \sum_{i=1}^{q} A_{m,i} X_{t-i}^m + \epsilon_t^m, \qquad A_{m,i} = W_{m,i}\Phi_i + S_{m,i},\ \forall\ i = 1, \ldots, q,$$

where the transition matrices $A_{m,i}$, satisfy the same type of constraints as in (1.1) for each lag i.

In the simplest case where $\Phi_1 = \cdots = \Phi_q$, the objective function and estimation procedure remain unchanged. If the low-rank basis matrices Φ_i differ across lags, the optimization framework extends naturally by augmenting the objective with terms for each lag:

$$G = F + \rho \sum_{i=1}^{q} \left(\frac{1}{2}\|\Phi_i - \bar{\Phi}_i\|_F^2 + \langle \Gamma_i, \Phi_i - \bar{\Phi}_i \rangle \right),$$

where F is the classical least squares objective combined with LASSO penalties on the sparse components $\mathcal{S}$.

For estimation, the algorithmic steps described in Appendix C remain applicable, as the augmented variables $\bar{\Phi}_i$ and dual variables Γ_i are separable across lags in the augmented objective. Similarly, the convergence guarantees established in Section 2.1 extend directly to this higher-lag setting.

APPENDIX G: PROOF OF CONVERGENCE RESULT FOR ALGORITHM 1

PROOF OF PROPOSITION 2.1. For the primal update of $\mathcal{W}$ and $\mathcal{S}$, Assumption 4 implies that the relevant objective function for the m-th subproblem,

$$G(W_m, S_m) = f_m(W_m, S_m) + \eta\|S_m\|_1 = \frac{1}{2T}\|W_m\Phi X_m + S_m X_m - Y_m\|_F^2 + \eta\|S_m\|_1,$$

is strongly convex with respect to its two arguments, respectively. Since Assumptions 3 and 4 require that the objective function G function β_W-convex with respect to W_m and β-convex with respect to S_m, it follows that the corresponding updates satisfy

$$\begin{aligned} G(W_m^{(i)}, S_m^{(i)}) - G(W_m^{(i+1)}, S_m^{(i)}) &\geqslant \frac{\beta_W}{2}\|W_m^{(i)} - W_m^{(i+1)}\|_F^2, \\ G(W_m^{(i+1)}, S_m^{(i)}) - G(W_m^{(i+1)}, S_m^{(i+1)}) &\geqslant \frac{\beta}{2}\|S_m^{(i)} - S_m^{(i+1)}\|_F^2. \end{aligned}$$

Therefore,

$$\text{(G.1)}\quad G(\mathcal{W}^{(i)},\mathcal{S}^{(i)},\Phi_r^{(i)},\Phi^{(i)},\Gamma_\Phi^{(i)}) - G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_r^{(i)},\Phi^{(i)},\Gamma_\Phi^{(i)}) \geqslant \sum_{m=1}^{M}\frac{\beta_W}{2}\|W_m^{(i)}-W_m^{(i+1)}\|_2^2 + \sum_{m=1}^{M}\frac{\beta}{2}\|S_m^{(i)}-S_m^{(i+1)}\|_F^2.$$

By optimality and convergence of the subproblem's algorithm in Section C.2, and noting that the objective function is monotonically decreasing and the algorithm is initialized at $\Phi_c^{(i)}$, we have that

$$\text{(G.2)}\quad G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i)},\Phi^{(i)},\Gamma^{(i)}) - G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i)},\Gamma^{(i)}) \geqslant \frac{\kappa\rho}{2}\|\Phi_c^{(i)}-\Phi_c^{(i+1)}\|_F^2.$$

For Φ, since Assumption 4 states that the objective function F is β_Φ-convex with respect to Φ, we consequently have

$$\text{(G.3)}\quad G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i)},\Gamma^{(i)}) - G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i+1)},\Gamma^{(i)}) \geqslant \frac{\beta_\Phi+\rho}{2}\|\Phi^{(i)}-\Phi^{(i+1)}\|_F^2,$$

given $\Phi^{(i+1)}$ is the subpoblem minimizer that satisfies the first order optimality condition. In addition, the first order condition also suggests $\rho\Gamma^{(i+1)} = -F_\Phi(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi^{(i+1)})$, where F_Φ is the gradient of F with respect to Φ defined as

$$F_\Phi(\mathcal{W},\mathcal{S},\Phi) = \nabla_\Phi F(\mathcal{W},\mathcal{S},\Phi;X,\eta) = \sum_{m=1}^{M} W_m\frac{Y_mX_m'}{T} - \sum_{m=1}^{M} W_m(W_m\Phi+S_m)\frac{X_mX_m'}{T}.$$

Denote $\nabla_\Phi F^{(i)} = F_\Phi(\mathcal{W}^{(i)},\mathcal{S}^{(i)},\Phi^{(i)})$. Since Assumption 4 ensures that F_Φ is α_W-Lipschitz continuous with respect to W_m, α_S-Lipschitz continuous with respect to S_m and α_Φ-Lipschitz continuous with respect to Φ, we get that

$$\begin{aligned}\text{(G.4)}\quad &\|\nabla_\Phi F^{(i)} - \nabla_\Phi F^{(i+1)}\|_F^2 \\ &\leqslant \left(\alpha_\Phi\|\Phi^{(i)}-\Phi^{(i+1)}\|_F + \sum_{m=1}^{M}\alpha_W\|W_m^{(i)}-W_m^{(i+1)}\|_F + \sum_{m=1}^{M}\alpha_S\|S_m^{(i)}-S_m^{(i+1)}\|_F\right)^2 \\ &\leqslant 3\alpha_\Phi^2\|\Phi^{(i)}-\Phi^{(i+1)}\|_F^2 + \sum_{m=1}^{M}3M\alpha_W^2\|W_m^{(i)}-W_m^{(i+1)}\|_F^2 + \sum_{m=1}^{M}3M\alpha_S^2\|S_m^{(i)}-S_m^{(i+1)}\|_F^2\end{aligned}$$

Hence, for the dual ascent step,

$$\begin{aligned}\text{(G.5)}\quad &G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i+1)},\Gamma^{(i)}) - G(\mathcal{W}^{(i+1)},\mathcal{S}^{(i+1)},\Phi_c^{(i+1)},\Phi^{(i+1)},\Gamma^{(i+1)}) \\ &= -\frac{1}{\rho}\|\nabla_\Phi F^{(i)} - \nabla_\Phi F^{(i+1)}\|_F^2 \\ &\geqslant -\frac{3\alpha_\Phi^2}{\rho}\|\Phi^{(i)}-\Phi^{(i+1)}\|_F^2 - \sum_{m=1}^{M}\frac{3M\alpha_W^2}{\rho}\|W_m^{(i)}-W_m^{(i+1)}\|_F^2 - \sum_{m=1}^{M}\frac{3M\alpha_S^2}{\rho}\|S_m^{(i)}-S_m^{(i+1)}\|_F^2\end{aligned}$$

where the last line uses (G.4). The final result then combines Equations (G.1) to (G.3) and (G.5) and yields

$$
\begin{aligned}
&G(\mathcal{W}^{(i)}, \mathcal{S}^{(i)}, \Phi_c^{(i)}, \Phi^{(i)}, \Gamma^{(i)}) - G(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) \\
&\quad \geqslant \sum_{m=1}^{M} \|W_m^{(i)} - W_m^{(i+1)}\|_F^2 \left(\frac{\beta_W}{2} - \frac{3M\alpha_W^2}{\rho}\right) + \sum_{m=1}^{M} \|S_m^{(i)} - S_m^{(i+1)}\|_F^2 \left(\frac{\beta}{2} - \frac{3M\alpha_S^2}{\rho}\right) \\
&\qquad + \|\Phi^{(i)} - \Phi^{(i+1)}\|_F^2 \left(\frac{\beta_\Phi + \rho}{2} - \frac{3\alpha_\Phi^2}{\rho}\right) + \frac{\kappa\rho}{2} \|\Phi_c^{(i)} - \Phi_c^{(i+1)}\|_F^2.
\end{aligned}
$$

Given a fixed data realization, a sufficiently large ρ leads to the sufficient descent property.

Note that in terms of orders, $\frac{\beta}{2} - \frac{3M\alpha_S^2}{\rho}$ is essentially of order $O(1)$. Hence, for balanced coefficients, one can select κ proportionally as $\kappa \propto \frac{M}{\rho}$. Similarly, imposing $\frac{\beta_W}{2} - \frac{3M\alpha_W^2}{\rho} = O(1)$ and $\frac{\beta_\Phi + \rho}{2} - \frac{3\alpha_\Phi^2}{\rho} = O(M)$ provides practical guidance for selecting ℓ accordingly. □

PROOF OF PROPOSITION 2.2. For bounding $G(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)})$ below, note that

$$
\begin{aligned}
&G(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) \\
&\quad = F(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi^{(i+1)}) + \frac{\rho}{2} \|\Phi^{(i+1)} - \Phi_c^{(i+1)}\|_F^2 + \rho \langle \Gamma^{(i+1)}, \Phi^{(i+1)} - \Phi_c^{(i+1)} \rangle \\
&\quad = F(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi^{(i+1)}) + \langle \nabla_\Phi F^{(i+1)}, \Phi_c^{(i+1)} - \Phi^{(i+1)} \rangle + \frac{\rho}{2} \|\Phi^{(i+1)} - \Phi_c^{(i+1)}\|_F^2 \\
&\quad \geqslant F(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}) + \frac{\rho - \alpha_\Phi}{2} \|\Phi^{(i+1)} - \Phi_r^{(i+1)}\|_F^2 \\
&\quad \geqslant 0,
\end{aligned}
$$

where the first inequality uses Assumption 4, namely that the partial derivative F_Φ is α_Φ-Lipschitz continuous with respect to Φ, and the last inequality is due to the selection criteria of ρ in (2.7). Therefore, the sequence $G(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)})$ is lower bounded.

Combined with the sufficient descent property established in Proposition 2.1, the sequence is monotonically decreasing and lower bounded, and is therefore guaranteed to converge. □

PROOF OF PROPOSITION 2.3. According to Propositions 2.1 and 2.2, the sequence of estimates, $(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)})$, forms a Cauchy sequence, and hence as $i \to \infty$, there exists a limiting point $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi}_c, \widehat{\Phi}, \widehat{\Gamma})$ such that

$$
\lim_{s \to \infty} (\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) = (\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi}_c, \widehat{\Phi}, \widehat{\Gamma}).
$$

For the second part, concerning first-order optimality, we will show that the sub-gradient sets of the function $G(\mathcal{W}, \mathcal{S}, \Phi_c, \Phi, \Gamma)$ at $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi}, \widehat{\Gamma}_\Phi)$ include the element 0, so that $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi}, \widehat{\Gamma}_\Phi)$ is a critical point. Indeed, given that the function G and its sub-gradient sets are continuous, it suffices to establish the following lemma.

LEMMA G.1. *There exist sequences of sub-gradients of G with respect to the variables $\{\mathcal{W}, \mathcal{S}, \Phi_c, \Phi, \Gamma\}$, evaluated at $(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)})$, that converge to 0,*

i.e., for $i \to \infty$,

$$\frac{\partial G}{\partial w_m}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) \ni \mathbf{d}_{w_m}^{(i+1)} \to 0, \ m = 1, \dots, M,$$
$$\frac{\partial G}{\partial S_m}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) \ni \mathbf{d}_{S_m}^{(i+1)} \to 0, \ m = 1, \dots, M,$$
$$\frac{\partial G}{\partial \Phi_c}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) \ni \mathbf{d}_{\Phi_c}^{(i+1)} \to 0,$$
$$\frac{\partial G}{\partial \Phi}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) \ni \mathbf{d}_{\Phi}^{(i+1)} \to 0,$$
$$\frac{\partial G}{\partial \Gamma_\Phi}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i+1)}) \ni \mathbf{d}_{\Gamma}^{(i+1)} \to 0.$$

□

PROOF OF LEMMA G.1. We prove these limits by upper bounding selective sequences of sub-gradients that converge to 0, and utilize the fact that the sets of subgradients are continuous.

For $\mathbf{d}_{w_m}^{(i+1)}$, we know that $\tilde{\mathbf{d}}_{w_m}^{(i+1)} = 0 \in \frac{\partial G}{\partial w_m}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i)}, \Phi_c^{(i)}, \Phi^{(i)}, \Gamma^{(i)})$, hence

$$\mathbf{d}_{w_m}^{(i+1)} = \tilde{\mathbf{d}}_{w_m}^{(i+1)} + \sum_{i=1}^{p} E_i(\Phi^{(i+1)} \frac{X_m X_m'}{T} (\Phi^{(i+1)})' - \Phi^{(i)} \frac{X_m X_m'}{T} (\Phi^{(i)})') E_i w_m^{(i+1)}$$
$$+ \sum_{i=1}^{p} E_i(\Phi^{(i+1)} - \Phi^{(i)}) \frac{X_m (S_m^{(i+1)} X_m - Y_m)'}{T} e_i + \sum_{i=1}^{p} E_i \Phi^{(i)} \frac{X_m X_m'}{T} (S_m^{(i+1)} - S_m^{(i)})' e_i.$$

Regarding S_m, optimality implies that $\tilde{\mathbf{d}}_{S_m}^{(i+1)} = 0 \in \frac{\partial G}{\partial S_m}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i)}, \Phi^{(i)}, \Gamma_\Phi^{(i)})$, therefore one of the sub-gradients $\mathbf{d}_{S_m}^{(i+1)}$ is

$$\mathbf{d}_{S_m}^{(i+1)} = \tilde{\mathbf{d}}_{S_m}^{(i+1)} + \frac{X_m X_m'}{T} (\Phi^{(i+1)} - \Phi^{(i)})' W_m^{(i+1)}.$$

The fact that the subproblem of Φ_c converges implies that there exists $H^{(i+1)}$ in the normal cone of $\mathbb{B}_p \cap \mathbb{L}_p(\hat{r}, \ell)$ at $\Phi_c^{(i+1)}$, such that

$$\tilde{\mathbf{d}}_{\Phi_c}^{(i+1)} = H^{(i+1)} + \rho(\Phi_c^{(i+1)} - \Phi^{(i)} - \Gamma^{(i)}) + \kappa\rho(\Phi_c^{(i+1)} - \Phi_c^{(i)}) = 0,$$

and $\tilde{\mathbf{d}}_{\Phi_c}^{(i+1)} \in \frac{\partial G}{\partial \Phi_c}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i)}, \Gamma^{(i)})$. Therefore, we have

$$\begin{aligned}\mathbf{d}_{\Phi_c}^{(i+1)} &= \tilde{\mathbf{d}}_{\Phi_c}^{(i+1)} - \rho(\Phi^{(i+1)} - \Phi^{(i)} + \Gamma^{(i+1)} - \Gamma^{(i)}) - \kappa\rho(\Phi_c^{(i+1)} - \Phi_c^{(i)}) \\ &= -\rho(\Phi^{(i+1)} - \Phi^{(i)} + \Gamma^{(i+1)} - \Gamma^{(i)}) - \kappa\rho(\Phi_c^{(i+1)} - \Phi_c^{(i)}).\end{aligned}$$

The optimality of Φ implies that $\tilde{\mathbf{d}}_{\Phi}^{(i+1)} = 0 \in \frac{\partial G}{\partial \Phi}(\mathcal{W}^{(i+1)}, \mathcal{S}^{(i+1)}, \Phi_c^{(i+1)}, \Phi^{(i+1)}, \Gamma^{(i)})$, and

$$\mathbf{d}_{\Phi}^{(i+1)} = \tilde{\mathbf{d}}_{\Phi}^{(i+1)} + \rho(\Gamma^{(i+1)} - \Gamma^{(i)}) = \rho_\Phi(\Gamma^{(i+1)} - \Gamma^{(i)}).$$

Finally, for the dual variable, we get

$$\mathbf{d}_{\Gamma}^{(i+1)} = \rho(\Phi^{(i+1)} - \Phi_c^{(i+1)}) = \rho(\Gamma^{(i+1)} - \Gamma^{(i)}).$$

Then, the limiting behaviors of the gradients' norms are all controlled by norm quantities that converge to 0. □

PROOF OF THEOREM 2.1. Note that the convergence and first-order optimality of Algorithm 1 follow directly from the sequence of results established in Propositions 2.1 to 2.3. It remains to verify that the objective function (2.6) satisfies the Kurdyka–Łojasiewicz (KL) property, ensuring that the convergence of Algorithm 1 holds regardless of initialization.

If we temporarily disregard the domain constraints $\Phi_c \in \mathbb{B}_p \cap \mathbb{L}_p(\hat{r}, \ell)$, the objective function (2.6) is evidently semi-algebraic. Therefore, it suffices to establish that both $\mathbb{B}_p$ and $\mathbb{L}_p(\hat{r}, \ell)$ are semi-algebraic sets. Under this condition, the optimization problem satisfies the KL property, guaranteeing global convergence.

One can easily verify the semi-algebraic property of the space $\mathbb{B}_p$, as its definition is given by polynomial constraints.

The space $\mathbb{L}_p(\hat{r}, \ell)$ can be considered as the intersection of two spaces: a matrix space with rank at most $\hat{r}$, and a matrix space with fixed nuclear norm ℓ. Since the former is known to be semi-algebraic, we establish next the semi-algebraic property of the latter.

The nuclear norm can be expressed as

$$\|\Phi\|_* = \operatorname{tr}(\sqrt{\Phi\Phi'}).$$

Here the square root of a symmetric (semi-)positive definite matrix $A = QDQ'$ with diagonal non-negative D is defined as $\sqrt{A} = QD^{1/2}Q'$ where $D^{1/2}$ takes element-wise square roots of the diagonal entries. This square root operation is a semi-algebraic operator. Thus, since the nuclear norm is a composition of multiple semi-algebraic functions, the set $\mathbb{L}_p(\hat{r}, \ell)$ is semi-algebraic. □

APPENDIX H: PROOFS OF CONSISTENCY RESULTS

PROOF OF PROPOSITION 3.1. We first briefly digress to present the following lemma, which suggests that the assumptions on the ∞-norm bounds of Φ^* and $\widehat{\Phi}$ in Proposition 3.1 are not restrictive.

LEMMA H.1. *Given an arbitrary rank-r matrix Φ ($r \leqslant p$), whose matrices contain the singular vectors and are sampled uniformly from the set of rank-r partial isometries in $\mathbb{R}^{p\times r}$, there exist positive constants φ and c, such that*

$$\|\Phi\|_\infty \leqslant \varphi \|\Phi\|_2 \frac{\sqrt{r_p}}{p}$$

holds with probability at least $1 - cp^{-3}\log(p)$, where $r_p = \log^2(p) \cdot \max(r, \log(p))$.

Lemma H.1 can be derived by adapting results from the literature (for example, see Candès and Recht, 2009, Lemma 2.2). In particular, we justify the bound on $\widehat{\Phi}$ by arguing that $\widehat{\Phi}$ primarily lies in a subspace of rank at least r.

Then, referring to the model setup discussed in Assumption 2, and given that $\Phi^* \in \mathbb{B}_p \cap \mathbb{L}_p(r, \ell)$, we can derive from $\|W_m^*\Phi^*\| = O(1)$ that $W \precsim \frac{\sqrt{rp}}{\ell}$.

The following claim is helpful to understand the constants given in Assumption 4 and the function F itself evaluated at the output $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi})$.

CLAIM H.1. *Given a data realization $\mathcal{X}$, Assumptions 1 and 2, together with the least squares structure of the subproblems, imply that the output $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi})$ from Algorithm 1 satisfies the following:*

- $\frac{\ell^2}{\hat{r}} \precsim \|\widehat{\Phi}\|_F^2 \precsim \frac{\ell^2}{r}$;
- $\|\widehat{W}_m\widehat{\Phi}\|^2 = O(1)$, $\|\widehat{S}_m\|^2 = O(1)$;

- $\min_j \frac{1}{M}\sum_{m=1}^{M}\|e_j'\widehat{W}_m\widehat{\Phi}\|^2 = O(1)$.

Consequently, when evaluated at the true parameters $(\mathcal{W}^*, \mathcal{S}^*, \Phi^*)$ *and the algorithm output* $(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi})$*, the constants in Assumption 4 can be tightened to satisfy:*

- $\frac{\beta_W}{\beta} = \min\left\{\min_j \|e_j'\widehat{\Phi}\|^2, \min_j \|e_j'\Phi^*\|^2\right\} \gtrsim \frac{\ell^2}{\hat{r}p}$;
- $\frac{\beta_\Phi}{\beta} = \min\left\{\min_j \sum_{m=1}^{M}\|e_j'\widehat{W}_m\|^2, \min_j \sum_{m=1}^{M}\|e_j'W_m^*\|^2\right\} \gtrsim \frac{Mrp}{\ell^2}$;
- $\widehat{W} \lesssim \frac{\sqrt{\hat{r}p}}{\ell}$.

We provide some further justification for Claim H.1. The first three bullet points follow directly from the model structure and the least squares formulation of the associated regression subproblems. The latter three can be verified by explicitly expanding the corresponding second-order partial derivatives of the objective function F.

Since Algorithm 1 is initialized within a neighborhood of the true parameters $(\mathcal{W}^*, \mathcal{S}^*, \Phi^*)$, the sufficient descent argument from Proposition 2.1, combined with the local smoothness conditions in Assumption 4, ensures that the final output$(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi})$ satisfies

$$F(\widehat{\mathcal{W}}, \widehat{\mathcal{S}}, \widehat{\Phi}) \leqslant F(\mathcal{W}^*, \mathcal{S}^*, \Phi^*).$$

Therefore, we obtain

$$\sum_{m=1}^{M}\frac{1}{2T}\|Y_m - (\widehat{W}_m\widehat{\Phi} + \widehat{S}_m)X_m\|_F^2 + \eta\|\widehat{S}_m\|_1 \leqslant \sum_{m=1}^{M}\frac{1}{2T}\|Y_m - (W_m^*\Phi^* + S_m^*)X_m\|_F^2 + \eta\|S_m^*\|_1,$$

The model implies that $Y_m - (W_m^*\Phi^* + S_m^*)X_m = \varepsilon_m$. Introducing the notation $\Delta_\Phi = \widehat{\Phi} - \Phi^*$, $\Delta_{W_m} = \widehat{W}_m - W_m^*$, and $\Delta_{S_m} = \widehat{S}_m - S_m^*$, we consider pairs of decomposable subspaces $(\mathcal{Q}_m, \mathcal{Q}_m^\perp)$ satisfying$\|S_m\|_1 = \|S_m^{\mathcal{Q}_m}\|_1 + \|S_m^{\mathcal{Q}_m^\perp}\|_1$ and $\mathcal{Q}_m \bigcup \mathcal{Q}_m^\perp = \mathbb{R}^{p\times p}$ for all $m = 1, \ldots, M$. Then

$$\begin{aligned}
&\sum_{m=1}^{M}\frac{1}{2T}\|(\Delta_{W_m}\widehat{\Phi} + W_m^*\Delta_\Phi + \Delta_{S_m})X_m\|_F^2 \\
&\leqslant \sum_{m=1}^{M}\langle\Delta_{W_m}\widehat{\Phi} + W_m^*\Delta_\Phi + \Delta_{S_m}, -\frac{\varepsilon_m X_m'}{T}\rangle + \sum_{m=1}^{M}\eta(\|S_m^*\|_1 - \|\widehat{S}_m\|_1) \\
&\leqslant \sum_{m=1}^{M}\langle\Delta_{W_m}\widehat{\Phi} + W_m^*\Delta_\Phi + \Delta_{S_m}, -\frac{\varepsilon_m X_m'}{T}\rangle + \sum_{m=1}^{M}\left(\eta\|\Delta_{S_m}^{\mathcal{Q}_m}\|_1 - \eta\|\Delta_{S_m}^{\mathcal{Q}_m^\perp}\|_1 + 2\eta\|(S_m^*)^{\mathcal{Q}_m^\perp}\|_1\right)
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{m=1}^{M}\langle\Delta_{W_m}\widehat{\Phi} + W_m^*\Delta_\Phi + \Delta_{S_m}, -\frac{\varepsilon_m X_m'}{T}\rangle \\
&\leqslant \sum_{m=1}^{M}\|\Delta_{W_m}\widehat{\Phi}\|_*\|\frac{\varepsilon_m X_m'}{T}\|_2 + \|\Delta_\Phi\|_*\|\sum_{m=1}^{M}\frac{W_m^*\varepsilon_m X_m'}{T}\|_2 + \sum_{m=1}^{M}\|\Delta_{S_m}\|_1\|\frac{\varepsilon_m X_m'}{T}\|_\infty \\
&\leqslant W\|\Delta_\Phi\|_*\|\sum_{m=1}^{M}\frac{W_m^\dagger\varepsilon_m X_m'}{T}\|_2 + \sum_{m=1}^{M}\frac{\phi\ell}{\sqrt{rp}}\|\Delta_{W_m}\|_*\|\frac{\varepsilon_m X_m'}{T}\|_2 + \sum_{m=1}^{M}(\|\Delta_{S_m}^{\mathcal{Q}}\|_1 + \|\Delta_{S_m}^{\mathcal{Q}^\perp}\|_1)\|\frac{\varepsilon_m X_m'}{T}\|_\infty
\end{aligned}$$

Then, combining the above two inequalities yields

$$
\begin{aligned}
&\sum_{m=1}^{M} \frac{1}{2T} \|(\Delta_{W_m} \widehat{\Phi} + W_m^* \Delta_\Phi + \Delta_{S_m}) X_m\|_F^2 \\
&\quad \leqslant W \|\Delta_\Phi\|_* \| \sum_{m=1}^{M} \frac{W_m^\dagger \varepsilon_m X_m'}{T} \|_2 + \frac{\phi \ell}{\sqrt{rp}} \sum_{m=1}^{M} \|\Delta_{W_m}\|_F \| \frac{\varepsilon_m X_m'}{T} \|_2 \\
&\qquad + \sum_{m=1}^{M} (\|\Delta_{S_m}^{\mathcal{Q}_m}\|_1 + \|\Delta_{S_m}^{\mathcal{Q}_m^\perp}\|_1) \| \frac{\varepsilon_m X_m'}{T} \|_\infty + \sum_{m=1}^{M} \left(\eta \|\Delta_{S_m}^{\mathcal{Q}_m}\|_1 - \eta \|\Delta_{S_m}^{\mathcal{Q}_m^\perp}\|_1 + 2\eta \|(S_m^*)^{\mathcal{Q}_m^\perp}\|_1 \right) \\
&\quad \leqslant W \|\Delta_\Phi\|_* \| \sum_{m=1}^{M} \frac{W_m^\dagger \varepsilon_m X_m'}{T} \|_2 + \frac{\phi \ell}{\sqrt{rp}} \sum_{m=1}^{M} \|\Delta_{W_m}\|_F \| \frac{\varepsilon_m X_m'}{T} \|_2 \\
&\qquad + \sum_{m=1}^{M} \left[(\| \frac{\varepsilon_m X_m'}{T} \|_\infty + \eta) \|\Delta_{S_m}^{\mathcal{Q}_m}\|_1 + (\| \frac{\varepsilon_m X_m'}{T} \|_\infty - \eta) \|\Delta_{S_m}^{\mathcal{Q}_m^\perp}\|_1 \right]
\end{aligned}
$$

The last inequality follows from choosing $\mathcal{Q}_m$ as the support of the true sparse S_m^*. Under this choice, and by applying the restricted strong convexity condition stated in Assumption 3, we obtain

$$
\begin{aligned}
&\sum_{m=1}^{M} \frac{1}{2T} \|(\Delta_{W_m} \widehat{\Phi} + W_m^* \Delta_\Phi + \Delta_{S_m}) X_m\|_F^2 \\
&\quad \geqslant \frac{\beta}{2} \sum_{m=1}^{M} \|\Delta_{W_m} \widehat{\Phi} + W_m^* \Delta_\Phi + \Delta_{S_m}\|_F^2 \\
&\quad \geqslant \frac{\beta_\Phi}{2} \|\Delta_\Phi\|_F^2 + \sum_{m=1}^{M} (\frac{\beta_W}{2} \|\Delta_{W_m}\|_F^2 + \frac{\beta}{2} \|\Delta_{S_m}\|_F^2 - \beta \|W_m^* \Delta_{W_m}\|_2 \|\widehat{\Phi}\|_\infty \|\Delta_\Phi\|_* \\
&\qquad\qquad - \beta W \|\Delta_\Phi\|_\infty \|\Delta_{S_m}\|_1 - 2\beta \widehat{W} \|\widehat{\Phi}\|_\infty \|\Delta_{S_m}\|_1) \\
&\quad \geqslant \frac{\beta_\Phi}{2} \|\Delta_\Phi\|_F^2 - \frac{2\beta\phi W \widehat{W} M \ell}{\sqrt{rp}} \|\Delta_\Phi\|_* + \sum_{m=1}^{M} (\frac{\beta_W}{2} \|\Delta_{W_m}\|_F^2 + \frac{\beta}{2} \|\Delta_{S_m}\|_F^2 - \frac{4\beta\phi \widehat{W} \ell}{\sqrt{rp}} \|\Delta_{S_m}\|_1)
\end{aligned}
$$

Hence, we can find $\zeta = \min\{\frac{\beta_\Phi}{M\beta}, \frac{\beta_W}{\beta}, 1\} \gtrsim \min\Big\{\frac{rp}{\ell^2}, \frac{\ell^2}{\hat{r}p}, 1\Big\}$, and $\eta \geqslant \frac{4\beta\phi\widehat{W}\ell}{\sqrt{rp}} + \|\frac{\varepsilon_m X_m'}{T}\|_\infty$ for all m,

$$
\begin{aligned}
&\frac{\beta\zeta}{2}\big(\|\Delta_\Phi\|_F^2 + \frac{1}{M}\sum_{m=1}^M(\|\Delta_{W_m}\|_F^2 + \|\Delta_{S_m}\|_F^2)\big) \\
&\quad\leqslant \frac{\beta_\Phi}{2M}\|\Delta_\Phi\|_F^2 + \frac{1}{2M}\sum_{m=1}^M(\beta_W\|\Delta_{W_m}\|_F^2 + \beta\|\Delta_{S_m}\|_F^2) \\
&\quad\leqslant \|\Delta_\Phi\|_*\big(W\|\sum_{m=1}^M\frac{W_m^\dagger\varepsilon_m X_m'}{MT}\|_2 + \frac{2\beta\phi W\widehat{W}\ell}{\sqrt{rp}}\big) + \frac{1}{M}\sum_{m=1}^M\frac{\phi\ell}{\sqrt{rp}}\|\Delta_{W_m}\|_F\|\frac{\varepsilon_m X_m'}{T}\|_2 \\
&\qquad + \frac{1}{M}\sum_{m=1}^M\Big((\frac{4\beta\phi\widehat{W}\ell}{\sqrt{rp}} + \|\frac{\varepsilon_m X_m'}{T}\|_\infty + \eta)\|\Delta_{S_m}^{\mathcal{Q}_m}\|_1 \\
&\qquad\qquad + (\frac{4\beta\phi\widehat{W}\ell}{\sqrt{rp}} + \|\frac{\varepsilon_m X_m'}{T}\|_\infty - \eta)\|\Delta_{S_m}^{\mathcal{Q}_m^\perp}\|_1\Big) \\
&\quad\leqslant \frac{\phi\ell}{M\sqrt{rp}}\sum_{m=1}^M\|\Delta_{W_m}\|_F\|\frac{\varepsilon_m X_m'}{T}\|_2 + \frac{2\eta\sqrt{s}}{M}\sum_{m=1}^M\|\Delta_{S_m}\|_F \\
&\qquad + \|\Delta_\Phi\|_F\big(W\sqrt{2\hat{r}}\|\sum_{m=1}^M\frac{W_m^\dagger\varepsilon_m X_m'}{MT}\|_2 + \frac{2\sqrt{2}\beta\phi W\widehat{W}\ell\sqrt{\hat{r}}}{\sqrt{rp}}\big) \\
&\quad\leqslant \sqrt{\frac{1}{M}\sum_{m=1}^M\Big(\frac{\phi^2\ell^2}{rp}\|\frac{\varepsilon_m X_m'}{T}\|_2^2 + 4\eta^2 s + 4W^2\hat{r}\|\sum_{m=1}^M\frac{W_m^\dagger\varepsilon_m X_m'}{MT}\|_2^2 + \frac{16\beta^2\phi^2W^2\widehat{W}^2\ell^2\hat{r}}{rp^2}\Big)} \\
&\qquad\times\sqrt{\|\Delta_\Phi\|_F^2 + \frac{1}{M}\sum_{m=1}^M(\|\Delta_{S_m}\|_F^2 + \|\Delta_{W_m}\|_F^2)}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{\beta^2\zeta^2}{4}\big(\|\Delta_\Phi\|_F^2 + \frac{1}{M}\sum_{m=1}^M(\|\Delta_{S_m}\|_F^2 + \|\Delta_{W_m}\|_F^2)\big) \\
&\leqslant \frac{1}{M}\sum_{m=1}^M\Big(\frac{\phi^2\ell^2}{rp}\|\frac{\varepsilon_m X_m'}{T}\|_2^2 + 4\eta^2 s + \frac{16\beta^2\phi^2W^2\widehat{W}^2\ell^2\hat{r}}{rp^2}\Big) + 4W^2\hat{r}\|\sum_{m=1}^M\frac{W_m^\dagger\varepsilon_m X_m'}{MT}\|_2^2.
\end{aligned}
$$

Hence, when we select $\eta = \frac{4\beta\phi\widehat{W}\ell}{\sqrt{rp}} + \max_m \|\frac{\varepsilon_m X_m'}{T}\|_\infty$, we obtain

$$\begin{aligned}
&\|\Delta_\Phi\|_F^2 + \frac{1}{M}\sum_{m=1}^M (\|\Delta_{S_m}\|_F^2 + \|\Delta_{W_m}\|_F^2) \\
&\leqslant \frac{4}{\zeta^2}\Bigg(16\phi^2\widehat{W}^2\ell^2\frac{W^2\hat{r}+8s}{rp^2} + \frac{8s}{\beta^2}\max_m\|\frac{\varepsilon_m X_m'}{T}\|_\infty^2 \\
&\qquad + \frac{\phi^2\ell^2}{\beta^2 Mrp}\sum_{m=1}^M \|\frac{\varepsilon_m X_m'}{T}\|_2^2 + \frac{4W^2\hat{r}}{\beta^2}\|\sum_{m=1}^M \frac{W_m^\dagger\varepsilon_m X_m'}{MT}\|_2^2\Bigg) \\
&\precsim \frac{1}{\zeta^2}\left(\frac{\hat{r}^2}{\ell^2} + \frac{\hat{r}s}{rp} + \frac{s}{\beta^2}\max_m\|\frac{\varepsilon_m X_m'}{T}\|_\infty^2 + \frac{\ell^2}{\beta^2 Mrp}\sum_{m=1}^M\|\frac{\varepsilon_m X_m'}{T}\|_2^2 + \frac{r\hat{r}p}{\beta^2\ell^2}\|\sum_{m=1}^M\frac{W_m^\dagger\varepsilon_m X_m'}{MT}\|_2^2\right).
\end{aligned}$$

Note that the last inequality uses the bounds of W and $\widehat{W}$ from Claim H.1. □

PROOF OF PROPOSITION 3.2. The proofs for the first and last statements closely follow the arguments in Basu and Michailidis (2015); Basu, Li and Michailidis (2019) and are therefore omitted. We focus here on proving the second statement, for which we begin with the following lemma.

LEMMA H.2. *For M arbitrary stationary centered Gaussian time series $\{H_m \in \mathbb{R}^{p\times T}\}_{m=1}^M$ that are mutually independent, and an arbitrary unit vector $v \in \mathbb{R}^p$ that $\|v\| = 1$, there exists constant $c > 0$, such that for any $k > 0$,*

$$P(|v'(\sum_{m=1}^M \frac{H_m H_m'}{MT} - \sum_{m=1}^M \frac{\Gamma_{H_m}(0)}{M})v| > 2\pi k\max_m \Psi(h_{H_m})) \leqslant 2\exp(-cMT\min\{k^2,k\}).$$

The proof of the lemma also follows based on arguments in Basu and Michailidis (2015). Specifically, note that $(v'H_1,\cdots,v'H_M)' \sim N(0,Q_H)$ with

$$Q_H = \text{blkdiag}(\Upsilon_1,\ldots,\Upsilon_M), \qquad \Upsilon_m[r,s] = v'\Gamma_{H_m}(r-s)v,$$

and $\|Q_H\| = \max_m\|\Upsilon_m\|$ due to its block-diagonal structure.

Next, by denoting $\bar{\varepsilon}_m = W_m^\dagger\varepsilon_m$, we get

$$\begin{aligned}
\sum_{m=1}^M \frac{2v'\bar{\varepsilon}_m X_m' v}{MT} &= [\sum_{m=1}^M \frac{(X_m'v+\bar{\varepsilon}_m'v)'(X_m'v+\bar{\varepsilon}_m'v)}{MT} - \sum_{m=1}^M \frac{v'(\Gamma_{X_m}(0)+W_m^\dagger\Sigma_m W_m^\dagger)v}{M}] \\
&\quad - [\sum_{m=1}^M \frac{v'X_m X_m' v}{MT} - \sum_{m=1}^M \frac{v'\Gamma_{X_m}(0)v}{M}] - [\sum_{m=1}^M \frac{v'\bar{\varepsilon}_m\bar{\varepsilon}_m' v}{MT} - \sum_{m=1}^M \frac{v'W_m^\dagger\Sigma_m W_m^\dagger v}{M}]
\end{aligned}$$

Applying Lemma H.2 to the collections $\{X_m\}_{m=1}^M$, $\{\bar{\varepsilon}_m\}_{m=1}^M$ and $\{X_m+\bar{\varepsilon}_m\}_{m=1}^M$, and using the fact that

$$\Psi(h_{X_m}) \leqslant \frac{\lambda_1(\Sigma_m)}{\tau_{\min}(\mathcal{A}_m)}, \quad \Psi(h_{\bar{\varepsilon}_m}) \leqslant \lambda_1(\Sigma_m), \quad \Psi(h_{X_m+\bar{\varepsilon}_m}) \leqslant \frac{\lambda_1(\Sigma_m)\tau_{\max}(\mathcal{A}_m)}{\tau_{\min}(\mathcal{A}_m)},$$

the conclusion follows by arguments analogous to those in Basu, Li and Michailidis (2019). □

PROOF OF COROLLARY 3.1. With similar arguments to those used in the proof of Proposition 3.1, define $\Delta_m = \widehat{W}_m\hat{\Phi} - W_m^*\Phi^*$. Then, we have

$$\frac{1}{2T}\sum_{m=1}^{M}\|(\Delta_m+\Delta_{S_m})X_m\|_F^2$$
$$\leqslant \sum_{m=1}^{M}\|\Delta_m\|_*\|\frac{\varepsilon_m X_m'}{T}\|_2 + \sum_{m=1}^{M}\left[(\|\frac{\varepsilon_m X_m'}{T}\|_\infty+\eta)\|\Delta_{S_m}^{\mathcal{Q}_m}\|_1 + (\|\frac{\varepsilon_m X_m'}{T}\|_\infty-\eta)\|\Delta_{S_m}^{\mathcal{Q}_m^\perp}\|_1\right].$$

and with the assumption implying that $\|\Delta_m\|_\infty \leqslant \frac{2\phi\widehat{W}\ell}{\sqrt{rp}}$,

$$\frac{1}{2T}\sum_{m=1}^{M}\|(\Delta_m+\Delta_{S_m})X_m\|_F^2 \geqslant \frac{\beta}{2}\sum_{m=1}^{M}(\|\Delta_m\|_F^2+\|\Delta_{S_m}\|_F^2 - \frac{4\phi\widehat{W}\ell}{\sqrt{rp}}\|\Delta_{S_m}\|_1).$$

Hence, selecting η as $\eta \geqslant \frac{2\beta\phi\widehat{W}\ell}{\sqrt{rp}} + \max_m\|\frac{\varepsilon_m X_m'}{T}\|_\infty$, we get

$$\frac{1}{M}\sum_{m=1}^{M}(\|\Delta_m\|_F^2+\|\Delta_{S_m}\|_F^2) \leqslant \frac{2}{\beta M}\sum_{m=1}^{M}(\|\Delta_m\|_*\|\frac{\varepsilon_m X_m'}{T}\|_2 + 2\eta\|\Delta_{S_m}^{\mathcal{Q}_m}\|_1),$$

and therefore

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}(\|\Delta_m\|_F^2+\|\Delta_{S_m}\|_F^2) &\leqslant \frac{4}{\beta^2 M}\sum_{m=1}^{M}(2\hat{r}\|\frac{\varepsilon_m X_m'}{T}\|_2^2 + 4\eta^2 s)\\ &\leqslant \frac{4}{\beta^2 M}\sum_{m=1}^{M}(\frac{4c_1^2\xi^2\hat{r}p}{T} + \frac{8c_1^2\xi^2 s\log(p)}{T}) + \frac{128\phi^2\widehat{W}^2\ell^2 s}{rp^2}\\ &\lesssim \xi^2\cdot\max_m\frac{\tau_{\max}^2(\mathcal{A}_m)}{\lambda_p^2(\Sigma_m)}\cdot\frac{s\log(p)+\hat{r}p}{T} + \frac{\iota s}{p}.\end{aligned}$$

□

APPENDIX I: IMPLEMENTATION DETAILS

I.1. Data Generating Process (DGP). Next, we introduce the data generating process (DGP) corresponding to the LSPVAR model in (1.1). The parameters (M, p, r, s) are specified upfront, and the time series length T is chosen accordingly, guided by the consistency results discussed in Section 3.

For the simulation in Example 1, we generate only 6 distinct diagonal matrices in $\mathcal{W}$, each representing one of the 6 clusters. In contrast, the 20 sparse matrices $\mathcal{S}$ are sampled independently following the procedure described in Algorithm 2.

I.2. Initialization. We also provide an initialization algorithm intended to produce estimates close to the true parameters. Algorithm 3 is motivated by the observation that, when temporarily ignoring the sparse components, corresponding rows from individual fits tend to align consistently in the same direction, differing primarily by a scaling factor.

I.3. Supplementary Figures and Tables. In this section, we supplement the simulation and neuroscience application results with additional visualizations, including figures and tables.

Algorithm 2: DGP for $\{\mathcal{X}_n(M,p,r,s,T)\}_{n=1}^N$

Input: Number of models M, time series dimension p, time series length T, rank r of matrix Φ, expected percentage of non-zero elements $\frac{s}{p^2}$ in sparse matrices $\mathcal{S}$, time series length T, number of replicates N per setup.

Output: Model parameters $(\Phi, \mathcal{W}, \mathcal{S}, \{A_m, \Sigma_m\}_{m=1}^M)$, and the replicates of Panel time series data $\{\mathcal{X}_n(M,T,p,r,s)\}_{n=1}^N$.

Randomly generate two singular vector matrices $U, V \in \mathbb{R}^{p\times r}$ $(r \leqslant p)$, sample the entries of the diagonal matrix D from an r-dimensional Dirichlet distribution, and generate $\Phi = UDV'$.

for $m = 1, \ldots, M$ **do**

  Sample s_m from a Poisson distribution with mean s, generate a sparse matrix S_m with support $\|S_m\|_0 = s_m$, and elements from centered normal distribution with standard deviation $\sqrt{p}\|\Phi\|_\infty$.

  Compute the eigenvalues of $\Phi + S_m$, and sample the diagonal entries of W_m from the uniform distribution on $[\frac{1}{2}|\lambda_1(\Phi + S_m)|^{-1}, |\lambda_1(\Phi + S_m)|^{-1}]$.

  Update the sparse matrix $S_m \mapsto W_m S_m$, and define $A_m = W_m\Phi + S_m$.

  Randomly sample $\Sigma_m = \sigma_m^2 I_p$ as the innovation covariance matrix, where σ_m^2 are sampled from an inverse Gamma distribution.

end

for $n = 1, \ldots, N$ **do**

  Sample the time series of length T as $\mathcal{X}_n(M,p,r,s,T) = \{\{X_t^m\}_{t=1}^T\}_{m=1}^M$ according to our PVAR model setup specified by (1.1) and parameters $\{A_m, \Sigma_m\}_{m=1}^M$.

end

Algorithm 3: Initialization for $\Phi^{(0)} = \Phi_c^{(0)}$

Input: Time series data $\{(X_m, Y_m) \in \mathbb{R}^{p\times T} \times \mathbb{R}^{p\times T}\}_{m=1}^M$, maximum rank $\hat{r}$, fixed nuclear norm ℓ.

Output: Initialization $(\Phi^{(0)}, \Phi_c^{(0)})$.

For each entity with data pair (X_m, Y_m), individually fit the VAR coefficient matrix $\Phi_m \in \mathbb{R}^{p\times p}$.

for $i = 1, \ldots, p$ **do**

  Take the i-th rows of Φ_m $(m = 1, \ldots, M)$ and stack them as $\Phi_i \in \mathbb{R}^{M\times p}$.

  Find the top right singular vector of Φ_i and set it as the i-th row of $\Phi^{(0)}$.

end

Update $\Phi^{(0)} \mapsto \frac{\ell}{\|\Phi^{(0)}\|_*}\Phi^{(0)}$.

Calculate $\Phi_c^{(0)}$ by projecting $\Phi^{(0)}$ onto $\mathbb{B}_p \cap \mathbb{L}_p(\hat{r}, \ell)$.

We begin with Table 4, which summarizes the key statistics from our simulation study in Section 4.1, focusing on the effects of the input rank $\hat{r}$ and the step size ρ.

Figure 9 depicts the histograms of the estimated $\|\widehat{W}_m\|^2$ under the setting specified in Example 1. They illustrate that the entities within the singular low-rank cluster have significantly smaller magnitude norms $\|\widehat{W}_m\|_F^2$. This observation closely aligns with the model's ground truth, which states that their Frobenius norms should be zero.

Finally, for the EEG application presented in Section 5, we display boxplots of all rescaling effects W, estimated from both the raw data and the alpha band-filtered data. Similar conclusions can be drawn from Figures 10 and 11: the estimates based on the alpha band data exhibit greater variability across entities (students). In addition, aside from some potential outliers, the effects under the EO condition condition appear more stable, as indicated by shorter box lengths, and less pronounced in magnitude—especially evident in the alpha band estimates.

We also include a brief visualization of the estimated low-rank component Φ in Figure 12. We observe that the outward connections from the posterior channels are relatively stronger

	ρ \ $\hat{r}$	3	5	10	15	20	30	40
Average Rescaled Shifted BIC ($\times 10^5$)	$\frac{M}{20}$	0.200	0.212	0.228	0.260	0.281	0.315	0.322
	$\frac{M}{5}$	0.200	0.212	0.228	0.260	0.279	0.315	0.322
	M	0.200	0.211	0.229	0.260	0.279	0.315	0.322
Average Relative Error of A_m	$\frac{M}{20}$	0.263	0.281	0.285	0.290	0.299	0.322	0.322
	$\frac{M}{5}$	0.263	0.281	0.285	0.289	0.298	0.322	0.322
	M	0.262	0.281	0.285	0.289	0.298	0.321	0.322
Average Relative Error of S_m	$\frac{M}{20}$	0.245	0.285	0.350	0.406	0.451	0.453	0.455
	$\frac{M}{5}$	0.245	0.285	0.350	0.406	0.449	0.453	0.455
	M	0.245	0.285	0.350	0.405	0.449	0.452	0.455
Average Sparsity Accuracy	$\frac{M}{20}$	0.970	0.954	0.955	0.971	0.969	0.951	0.951
	$\frac{M}{5}$	0.970	0.954	0.955	0.971	0.970	0.951	0.951
	M	0.970	0.954	0.955	0.971	0.970	0.951	0.951
Average Sparsity Sensitivity	$\frac{M}{20}$	0.745	0.767	0.745	0.687	0.670	0.699	0.702
	$\frac{M}{5}$	0.745	0.767	0.745	0.684	0.671	0.699	0.702
	M	0.745	0.767	0.745	0.685	0.671	0.700	0.702
Average Sparsity Specificity	$\frac{M}{20}$	0.976	0.959	0.960	0.979	0.977	0.958	0.958
	$\frac{M}{5}$	0.976	0.959	0.960	0.979	0.977	0.958	0.958
	M	0.976	0.959	0.960	0.979	0.977	0.958	0.958

TABLE 4

Supplement to Section 4.1. Summary metrics from simulations examining the choice of input rank $\hat{r}$ and ADMM step size ρ

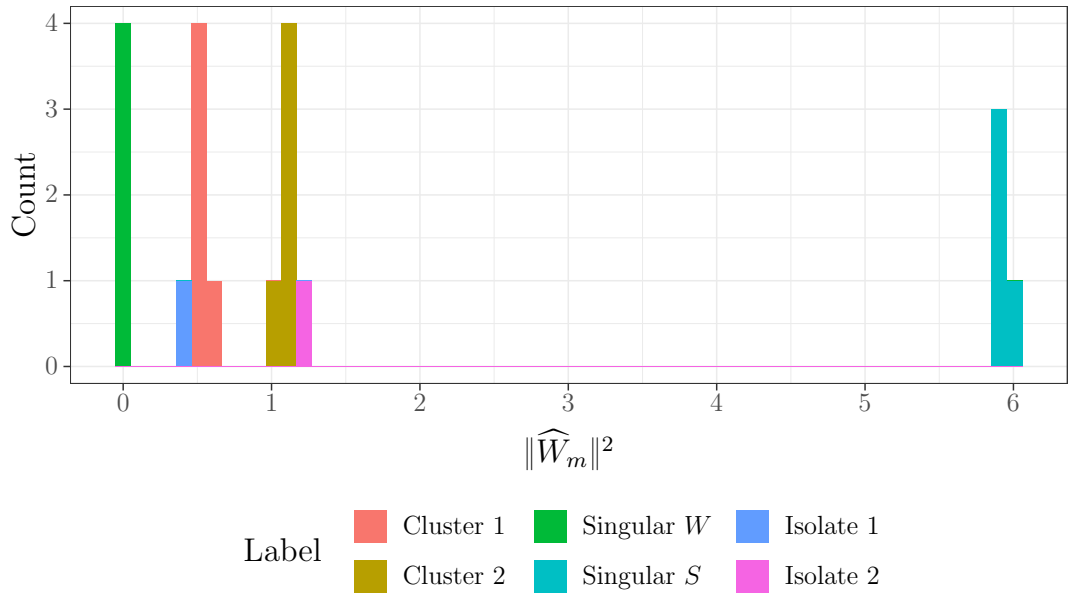

FIG 9. *Supplement to Section 4.2. The histogram above displays the frequencies of the estimated Frobenius norms $|\widehat{W}_m|_F^2$ from a single replicate. It demonstrates that the clustering structure can be consistently recovered in accordance with our simulation setup.*

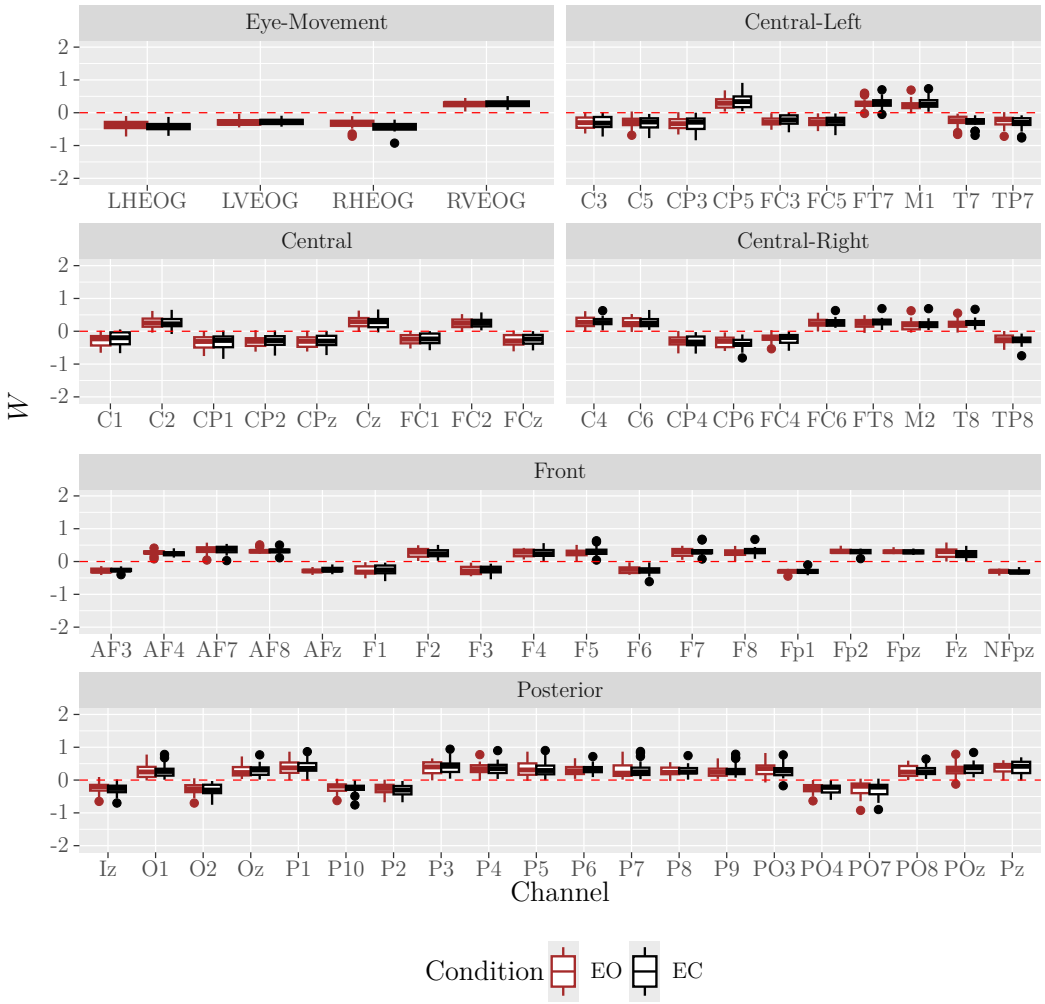

FIG 10. *Boxplots of the rescaling effects for all channels estimated from the raw data.*

on the human scalp, and these connections become even more pronounced after filtering out the underlying noise signals.

REFERENCES

ASLAN, S. and OMBAO, H. (2025). Granger Causality in High-Dimensional Networks of Time Series. arXiv:2406.02360. https://doi.org/10.48550/arXiv.2406.02360

ATTOUCH, H., BOLTE, J. and SVAITER, B. F. (2013). Convergence of Descent Methods for Semi-Algebraic and Tame Problems: Proximal Algorithms, Forward–Backward Splitting, and Regularized Gauss–Seidel Methods. *Mathematical Programming* **137** 91–129. https://doi.org/10.1007/s10107-011-0484-9

BABII, A., GHYSELS, E. and PAN, J. (2023). Tensor PCA for Factor Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4312303

BAI, P., SAFIKHANI, A. and MICHAILIDIS, G. (2022). A Fast Detection Method of Break Points in Effective Connectivity Networks. *IEEE Transactions on Medical Imaging* **41** 1017–1030. https://doi.org/10.1109/TMI.2021.3131142

BASU, S., LI, X. and MICHAILIDIS, G. (2019). Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions. *IEEE Transactions on Signal Processing* **67** 1207–1222. https://doi.org/10.1109/TSP.2018.2887401

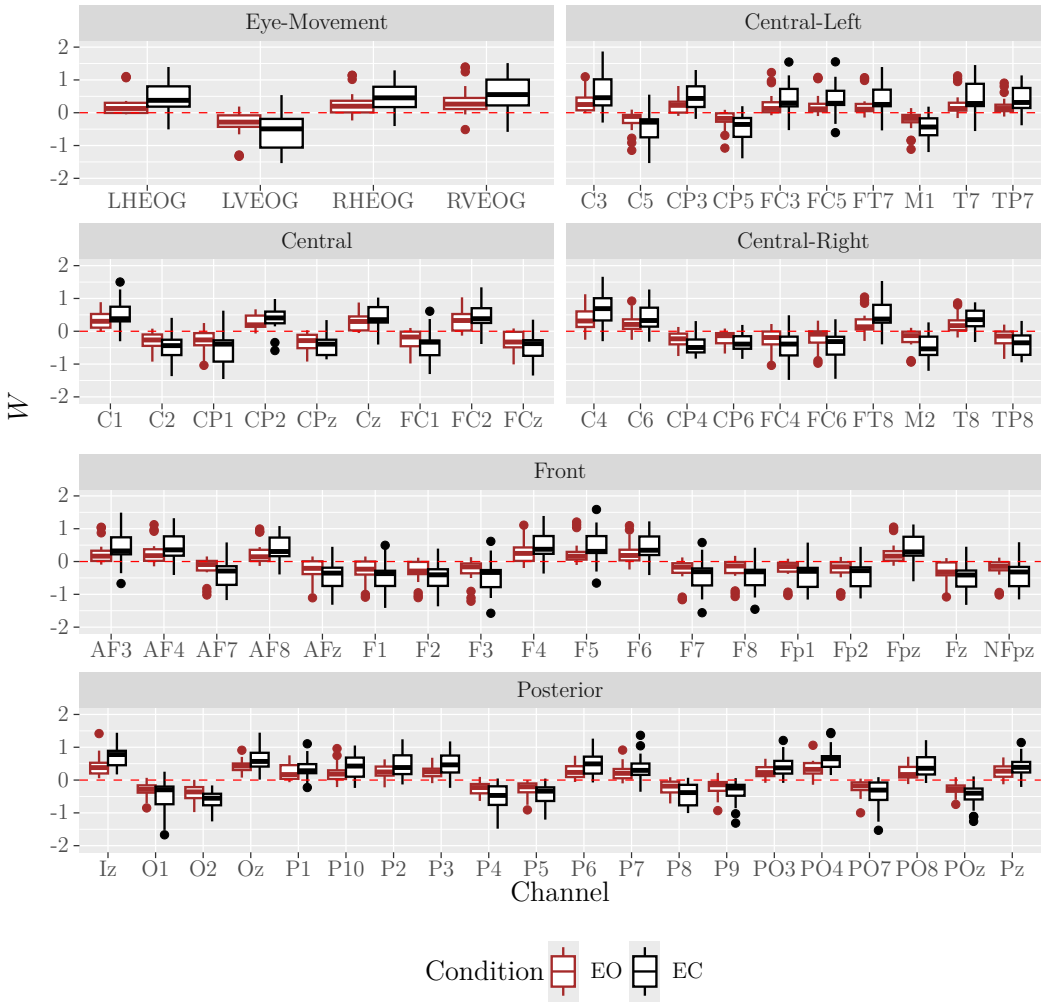

FIG 11. *Boxplots of the rescaling effects for all channels estimated from the alpha bands.*

BASU, S. and MICHAILIDIS, G. (2015). Regularized Estimation in Sparse High-Dimensional Time Series Models. *The Annals of Statistics* **43** 1535–1567. https://doi.org/10.1214/15-AOS1315

BILLIO, M., CASARIN, R. and ROSSINI, L. (2019). Bayesian nonparametric sparse VAR models. *Journal of Econometrics* **212** 97–115.

BOLTE, J., SABACH, S. and TEBOULLE, M. (2014). Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems. *Mathematical Programming* **146** 459–494. https://doi.org/10.1007/s10107-013-0701-9

BOYD, S., PARIKH, N., CHU, E., PELEATO, B., ECKSTEIN, J. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* **3** 1–122.

BOYLE, J. P. and DYKSTRA, R. L. (1986). A Method for Finding Projections onto the Intersection of Convex Sets in Hilbert Spaces. In *Advances in Order Restricted Statistical Inference* 28–47. Springer, New York, NY. https://doi.org/10.1007/978-1-4613-9940-7_3

BREITUNG, J. (2015). The Analysis of Macroeconomic Panel Data. In *The Oxford Handbook of Panel Data* 0. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199940042.013.0015

CANDÈS, E. J. and RECHT, B. (2009). Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics* **9** 717–772. https://doi.org/10.1007/s10208-009-9045-5

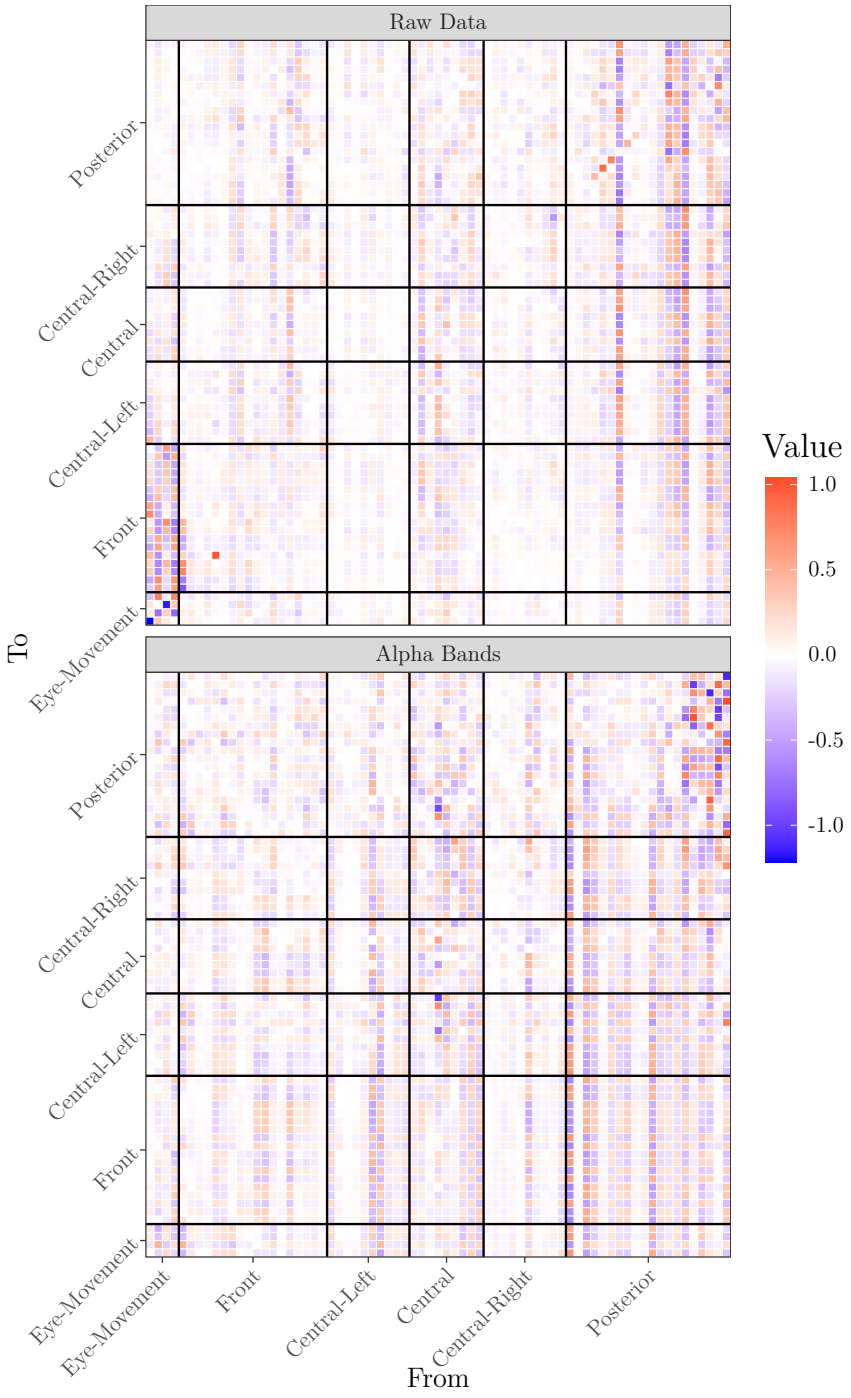

FIG 12. *The heatmap for the estimated* Φ*, with blocks illustrating the dynamics at the cluster level.*

CANOVA, F. and CICCARELLI, M. (2013). Panel Vector Autoregressive Models: A Survey. In *VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A. Sims*, **32** Emerald Group Publishing Limited. https://doi.org/10.1108/S0731-9053(2013)0000031006

CHEN, R., XIAO, H. and YANG, D. (2021). Autoregressive Models for Matrix-Valued Time Series. *Journal of Econometrics* **222** 539–560. https://doi.org/10.1016/j.jeconom.2020.07.015

CHEN, R., YANG, D. and ZHANG, C.-H. (2022). Factor Models for High-Dimensional Tensor Time Series. *Journal of the American Statistical Association* **117** 94–116. https://doi.org/10.1080/01621459.2021.1912757

CHEN, A. C. N., FENG, W., ZHAO, H., YIN, Y. and WANG, P. (2008). EEG Default Mode Network in the Human Brain: Spectral Regional Field Powers. *NeuroImage* **41** 561–574. https://doi.org/10.1016/j.neuroimage.2007.12.064

DENG, W., LAI, M.-J., PENG, Z. and YIN, W. (2017). Parallel Multi-Block ADMM with o(1 / k) Convergence. *Journal of Scientific Computing* **71** 712–736. https://doi.org/10.1007/s10915-016-0318-2

FRANK, M. and WOLFE, P. (1956). An Algorithm for Quadratic Programming. *Naval Research Logistics Quarterly* **3** 95–110. https://doi.org/10.1002/nav.3800030109

GHOSH, S., KHARE, K. and MICHAILIDIS, G. (2019). High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models. *Journal of the American Statistical Association* **114** 735–748. https://doi.org/10.1080/01621459.2018.1437043

HAN, Y., YANG, D., ZHANG, C.-H. and CHEN, R. (2024a). CP Factor Model for Dynamic Tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86** 1383–1413. https://doi.org/10.1093/jrsssb/qkae036

HAN, Y., CHEN, R., YANG, D. and ZHANG, C.-H. (2024b). Tensor Factor Model Estimation by Iterative Projection. *The Annals of Statistics* **52** 2641–2667. https://doi.org/10.1214/24-AOS2412

HOSSEINI, S., LUKE, D. R. and USCHMAJEW, A. (2019). Tangent and Normal Cones for Low-Rank Matrices. In *Nonsmooth Optimization and Its Applications* (S. Hosseini, B. S. Mordukhovich and A. Uschmajew, eds.) 45–53. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-11370-4_3

HSU, D., KAKADE, S. M. and ZHANG, T. (2011). Robust Matrix Decomposition With Sparse Corruptions. *IEEE Transactions on Information Theory* **57** 7221–7234. https://doi.org/10.1109/TIT.2011.2158250

ISSA, M. F. and JUHASZ, Z. (2019). Improved EOG Artifact Removal Using Wavelet Enhanced Independent Component Analysis. *Brain Sciences* **9** 355. https://doi.org/10.3390/brainsci9120355

KASTNER, G. and HUBER, F. (2020). Sparse Bayesian Vector Autoregressions in Huge Dimensions. *Journal of Forecasting* **39** 1142–1165. https://doi.org/10.1002/for.2680

KILIAN, L. and LÜTKEPOHL, H. (2017). *Structural vector autoregressive analysis*. Cambridge University Press.

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

KOROBILIS, D. (2016). Prior Selection for Panel Vector Autoregressions. *Computational Statistics & Data Analysis* **101** 110–120. https://doi.org/10.1016/j.csda.2016.02.011

KRAMPE, J., PAPARODITIS, E. and TRENKLER, C. (2023). Structural inference in sparse high-dimensional vector autoregressions. *Journal of Econometrics* **234** 276–300.

LI, X. and LUO, Z. (2023). Normal Cones Intersection Rule and Optimality Analysis for Low-Rank Matrix Optimization with Affine Manifolds. *SIAM Journal on Optimization* **33** 1333–1360. https://doi.org/10.1137/22M147863X

LÜTKEPOHL, H. (2005). Stable Vector Autoregressive Processes. In *New Introduction to Multiple Time Series Analysis* (H. Lütkepohl, ed.) 13–68. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-27752-1_2

MEDEIROS, M. C. and MENDES, E. F. (2016). ℓ1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics* **191** 255–271.

MELNYK, I. and BANERJEE, A. (2016). Estimating structured vector autoregressive models. In *International Conference on Machine Learning* 830–839. PMLR.

MICHAILIDIS, G. and D'ALCHÉ BUC, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences* **246** 326–334.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39** 1069 – 1097. https://doi.org/10.1214/10-AOS850

SCHNEIDER, R. and USCHMAJEW, A. (2015). Convergence Results for Projected Line-Search Methods on Varieties of Low-Rank Matrices Via Łojasiewicz Inequality. *SIAM Journal on Optimization* **25** 622–646. https://doi.org/10.1137/140957822

SETH, A. K., BARRETT, A. B. and BARNETT, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience* **35** 3293–3297.

SIGMUND, M. and FERSTL, R. (2021). Panel Vector Autoregression in R with the Package Panelvar. *The Quarterly Review of Economics and Finance* **80** 693–720. https://doi.org/10.1016/j.qref.2019.01.001

SKRIPNIKOV, A. and MICHAILIDIS, G. (2019). Regularized Joint Estimation of Related Vector Autoregressive Models. *Computational statistics & data analysis* **139** 164–177. https://doi.org/10.1016/j.csda.2019.05.007

SURANA, A., PATTERSON, G. and RAJAPAKSE, I. (2016). Dynamic Tensor Time Series Modeling and Analysis. In *2016 IEEE 55th Conference on Decision and Control (CDC)* 1637–1642. https://doi.org/10.1109/CDC.2016.7798500

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge university press.

WANG, H., LI, G. and JIANG, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *Journal of Business & Economic Statistics* **25** 347–355. https://doi.org/10.1198/073500106000000251

WANG, D., ZHENG, Y. and LI, G. (2024). High-Dimensional Low-Rank Tensor Autoregressive Time Series Modeling. *Journal of Econometrics* **238** 105544. https://doi.org/10.1016/j.jeconom.2023.105544

XU, Y., DÜKER, M.-C. and MATTESON, D. S. (2024). Testing Simultaneous Diagonalizability. *Journal of the American Statistical Association* **119** 1513–1525. https://doi.org/10.1080/01621459.2023.2202435

ZHU, Z. and LI, X. (2019). Convergence Analysis of Alternating Projection Method for Nonconvex Sets. https://doi.org/10.48550/arXiv.1802.03889

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "Degrees of Freedom" of the Lasso. *The Annals of Statistics* **35** 2173–2192. https://doi.org/10.1214/009053607000000127