

Joint modeling and inference of multiple-subject high-dimensional sparse vector autoregressive models

Younghoon Kim^{*1}, Zachary F. Fisher², and Vlasdas Pipiras²

¹Cornell University

²University of North Carolina at Chapel Hill

October 17, 2025

Abstract

The multiple-subject vector autoregression (multi-VAR) model captures heterogeneous network Granger causality across subjects by decomposing individual sparse VAR transition matrices into commonly shared and subject-unique paths. The model has been applied to characterize hidden shared and unique paths among subjects and has demonstrated performance compared to methods commonly used in psychology and neuroscience. Despite this innovation, the model suffers from using a weighted median for identifying the common effects, leading to statistical inefficiency as the convergence rates of the common and unique paths are determined by the least sparse subject and the smallest sample size across all subjects. We propose a new identifiability condition for the multi-VAR model based on a communication-efficient data integration framework. We show that this approach achieves convergence rates tailored to each subject's sparsity level and sample size. Furthermore, we develop hypothesis tests to assess the nullity and homogeneity of individual paths, using Wald-type test statistics constructed from individual debiased estimators. A test for the significance of the common paths can also be derived through the framework. Simulation studies under various heterogeneity scenarios and a real data application demonstrate the performance of the proposed method compared to existing benchmark across standard evaluation metrics.

Keywords— High-dimensional time series, meta-analysis, heterogeneity, debiased lasso, hypothesis testing, fMRI

^{*}Corresponding author. Email: yk748@cornell.edu

1 Introduction

1.1 Multiple Subject Time Series Model

In recent years, sparse vector autoregressive (VAR) modeling of high-dimensional time series (HDTS) has become a central topic in statistics and machine learning, driven by the increasing availability of high-dimensional, temporally dependent data. These data are collected across diverse scientific domains, including neuroscience, genomics, finance, and social networks (e.g., Song and Bickel; 2011; Han et al.; 2015; Kock and Callot; 2015; Davis et al.; 2016). A key challenge in analyzing these data is characterizing the dynamic dependencies among a large number of variables, often through the framework of network Granger causality (Shojaie and Fox; 2022), which encodes directional relationships, with edges indicating whether past values of one variable improve the prediction of another.

Despite the popularity of sparse VAR modeling, extensions of VAR models to multiple subjects (or datasets) have been relatively uncommon. This scarcity is partly due to the large number of parameters, which grow quadratically with the number of variables, and the difficulty associated with estimation in high-dimensional settings with limited observations per subject. Moreover, representing common and subject-specific (or unique; we use those two terms interchangeably) components through decompositions of VAR transition matrices introduces challenges in interpretation and identifiability. For these reasons, factor models have become a popular alternative in multi-subject time series analysis (e.g., Abdi et al.; 2013; Fan et al.; 2018; O’Connell and Lock; 2019; Kim et al.; 2024). These models capture shared components through common loadings while allowing subject-specific variation. In addition, their computational cost and identifiability requirements are often comparable to those of single-subject models.

Despite their scarcity in the literature, extending VAR models to multiple subjects is interesting; joint VAR modeling can reveal shared mechanisms across individuals and enable comparisons of heterogeneous dynamics. A related idea has been explored in multi-level modeling, which is widely used in individual-differences analyses (e.g., Wright et al.; 2015; Jongerling et al.; 2015; Haslbeck et al.; 2025). A promising extension of this idea to high-dimensional settings is the multiple-subject vector autoregression model (multi-VAR; Fisher et al.; 2022, 2024). Instead of relying on mixed-effects formulations, multi-VAR assumes that the sparsely estimated components of individual VAR transition matrices, called paths, can be decomposed into common or shared, and subject-specific, components. The model encourages similarity across individual parameter vectors while allowing for deviations, thereby capturing both shared temporal dynamics and subject-specific variation. Compared with other VAR-type models commonly used in psychology and neuroscience (e.g., Chen

et al.; 2011; Gates and Molenaar; 2012; Seth et al.; 2015), multi-VAR explicitly borrows information across subjects rather than fitting each subject separately, leading to more efficient estimation of shared paths while still accounting for individual differences.

However, a limitation of the multi-VAR approach is the lack of identifiability of common paths. In practice, common effects are usually determined using a weighted median across individual paths. While this provides a well-defined solution, it can be statistically inefficient (Asiaee et al.; 2019) (see also Maity et al. (2022) for a related discussion), which will be explained. As a result, there is room for improvement in the form of new identification conditions for the multi-VAR framework. In addition, the literature still lacks a formal hypothesis testing framework for multi-subject VAR models. Statistical tests for assessing the nullity, significance, and homogeneity of individual paths across subjects remain underdeveloped, which limits the ability to validate estimated causal structures and to interpret common versus subject-specific effects in scientific applications.

1.2 Original Multi-VAR Model

In this section, we provide an overview of the original multi-VAR model (Fisher et al.; 2022, 2024) and introduce its estimation procedure, which leads to the statistical inefficiency inherent in the original model.

Suppose we have a d -dimensional observation series $\{X_t^{(k)}\}$ from $K > 1$ subjects for T_k time points. Note that the variables across subjects are the same, while their sample length can vary across subjects. We assume that each vector series follows a VAR(p) model,

$$X_t^{(k)} = \Phi_1^{(k)} X_{t-1}^{(k)} + \dots + \Phi_p^{(k)} X_{t-p}^{(k)} + \epsilon_t^{(k)}, \quad \epsilon_t^{(k)} \sim \mathcal{N}(0, \Sigma_\epsilon^{(k)} = \text{diag}(\sigma_{k,1}^2, \dots, \sigma_{k,d}^2)). \quad (1)$$

Note that the lag order p is also the same across all subjects. Each VAR(p) model can be formed into the matrix-valued regression equations,

$$\underbrace{\begin{bmatrix} (X_{p+1}^k)' \\ (X_{p+2}^k)' \\ \vdots \\ (X_T^k)' \end{bmatrix}}_{\mathcal{Y}^{(k)}} = \underbrace{\begin{bmatrix} (X_p^k)' & (X_{p-1}^k)' & \dots & (X_1^k)' \\ (X_{p+1}^k)' & (X_p^k)' & \dots & (X_2^k)' \\ \vdots & \vdots & \ddots & \vdots \\ (X_{T-1}^k)' & (X_{T-2}^k)' & \dots & (X_{T-p}^k)' \end{bmatrix}}_{\mathcal{X}^{(k)}} \underbrace{\begin{bmatrix} (\Phi_1^k)' \\ (\Phi_2^k)' \\ \vdots \\ (\Phi_p^k)' \end{bmatrix}}_{B^{(k)}} + \underbrace{\begin{bmatrix} (\epsilon_{p+1}^k)' \\ (\epsilon_{p+2}^k)' \\ \vdots \\ (\epsilon_T^k)' \end{bmatrix}}_{E^{(k)}},$$

where $\mathcal{Y}^{(k)} \in \mathbb{R}^{N_k \times d}$ and $\mathcal{X}^{(k)} \in \mathbb{R}^{N_k \times dp}$ are response vectors and covariate matrices, respectively, and $N_k = T_k - p$. Note that for the stacked VAR transition matrices $B^{(k)} = (\Phi_1^{(k)'} \dots \Phi_p^{(k)'})' \in \mathbb{R}^{pd \times d}$, consider its vectorization $\beta^{(k)} = \text{vec}(B^{(k)}) \in \mathbb{R}^{d^2 p}$. The multi-VAR assumes that the $d^2 p$ so-called individual paths are decomposed into

$$\beta^{(k)} = \alpha^{(0)} + \alpha^{(k)}, \quad (2)$$

where $\alpha^{(0)}$ are the common paths shared by all K subjects and $\alpha^{(k)}$, $k = 1, \dots, K$ are unique paths of k^{th} subject.

The original multi-VAR model (Fisher et al.; 2022, 2024) employs a joint estimation framework to obtain the decomposed paths (2) in the VAR transition matrices in (1) across all K subjects. Specifically, it builds on the data-shared Lasso (Gross and Tibshirani; 2016) or the stratified Lasso (Ollier and Viallon; 2017);

$$\begin{aligned}
& (\hat{\alpha}^{(0)}, \hat{\beta}^{(1)}, \dots, \hat{\beta}^{(K)}) \\
&= \underset{\alpha^{(0)}, \beta^{(1)}, \dots, \beta^{(K)}}{\operatorname{argmin}} \left\{ \sum_{k=1}^K \frac{1}{2N_k} \|\operatorname{vec}(\mathcal{Y}^{(k)}) - (I_d \otimes \mathcal{X}^{(k)})\beta^{(k)}\|_2^2 + \tilde{\lambda}_0 \|\alpha^{(0)}\|_1 + \sum_{k=1}^K \tilde{\lambda}_k \|\beta^{(k)} - \alpha^{(0)}\|_1 \right\},
\end{aligned} \tag{3}$$

The estimation in the algorithm is performed using the fast iterative shrinkage-thresholding algorithm (FISTA; Beck and Teboulle (2009)) by stacking all individual equations. Specifically, with $\mathbf{Y}^{(k)} = \operatorname{vec}(\mathcal{Y}^{(k)})$ and $\mathbf{Z}^{(k)} = (I_d \otimes \mathcal{X}^{(k)})$, the aggregated equation is

$$\underbrace{\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \\ \vdots \\ \mathbf{Y}^{(K)} \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{Z}^{(1)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{Z}^{(2)} & \mathbf{0} & \mathbf{Z}^{(2)} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}^{(K)} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{Z}^{(K)} \end{bmatrix}}_{\mathbf{Z}} \begin{bmatrix} \alpha^{(0)} \\ \alpha^{(1)} \\ \vdots \\ \alpha^{(K)} \end{bmatrix} + \begin{bmatrix} \mathbf{E}^{(1)} \\ \mathbf{E}^{(2)} \\ \vdots \\ \mathbf{E}^{(K)} \end{bmatrix}. \tag{4}$$

Then for $\boldsymbol{\theta} := (\alpha^{(0)'}, \alpha^{(1)'}, \dots, \alpha^{(K)'})'$, the optimizer solves

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2 + \tilde{\lambda} \|\boldsymbol{\theta}\|_1.$$

Here we use $N = T - k$ so that $N_k = N$ for all k is assumed. Additional computational strategies, such as the backtracking step-size rule, are also included. The multi-VAR modeling and its estimation algorithm are implemented in the R package **multivar** (Fisher et al.; 2021).

Note that the penalty term in (3) is separable. So, each estimated common path is determined by the weighted median of the estimated individual paths,

$$\begin{aligned}
(\hat{\alpha}_i^{(0)})_j &= \underset{(\alpha_i^{(0)})_j \in \mathbb{R}}{\operatorname{argmin}} \left\{ |(\alpha_i^{(0)})_j| + \sum_{k=1}^K \frac{\tilde{\lambda}_k}{\tilde{\lambda}_0} |(\hat{\beta}_i^{(k)})_j - (\alpha_i^{(0)})_j| \right\}, \\
&= \operatorname{median}((\hat{\beta}_i^{(1)})_j, \dots, (\hat{\beta}_i^{(K)})_j; (1, \tilde{\lambda}_1/\tilde{\lambda}_0, \dots, \tilde{\lambda}_K/\tilde{\lambda}_0)),
\end{aligned}$$

and $\hat{\alpha}^{(k)} = \hat{\beta}^{(k)} - \hat{\alpha}^{(0)}$, $k = 1, \dots, K$. However, Asiaee et al. (2019) pointed out that it is statistically inefficient in terms of convergence rate compared to those of individual VAR models. That is,

$$\|\hat{\alpha}^{(0)} - \alpha^{(0)}\|_2 + \sum_{k=1}^K \sqrt{\frac{N_k}{N_0}} \|\hat{\alpha}^{(k)} - \alpha^{(k)}\|_2 \leq \max_{k=1, \dots, K} \frac{N_0}{N_k} \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{\max_k (\|\alpha^{(k)}\|_0 \log d^2 p)}{N_0}} \right),$$

where $N_0 = \sum_k N_k$. Consequently, convergence rates are determined by the least sparse subject and single-individual sample size,

$$\|\hat{\alpha}^{(k)} - \alpha^{(k)}\|_2 \leq \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{\max_k (\|\alpha^{(k)}\|_0) \log d^2 p}{N_k}} \right). \quad (5)$$

As a consequence, the convergence rates of both the common and subject-specific path estimators are determined by the least sparse subject and the smallest sample size, which is problematic when integrating heterogeneous datasets. In addition, since the equations are all stacked as in (4) to solve the single large-scale optimization problem, the computation is consequently slow.

One improvement of the original multi-VAR model is the use of adaptive Lasso penalties (Zou; 2006). Specifically, the estimation problem is formulated as

$$\begin{aligned} & (\hat{\alpha}^{(0)}, \hat{\beta}^{(1)}, \dots, \hat{\beta}^{(K)}) \\ &= \underset{\alpha^{(0)}, \beta^{(1)}, \dots, \beta^{(K)}}{\operatorname{argmin}} \left\{ \sum_{k=1}^K \frac{1}{2N_k} \|\operatorname{vec}(\mathcal{Y}^{(k)}) - (I_d \otimes \mathcal{X}^{(k)})\beta^{(k)}\|_2^2 + \tilde{\lambda}_0 \|\alpha^{(0)}\|_{1,w} + \sum_{k=1}^K \tilde{\lambda}_k \|\beta^{(k)} - \alpha^{(0)}\|_{1,w} \right\}, \end{aligned} \quad (6)$$

where $|\theta|_{1,w} = \sum_i w_i |\theta_i|$ with nonnegative weights $\{w_i\}$. The weights are constructed by initially fitting individual VAR models with Lasso to obtain $\{\hat{\beta}^{(k)}\}$, then setting the weights for the common path as $w_i^{(0)} = 1/|\hat{\beta}_i^{(0)}|$, where $\hat{\beta}_i^{(0)} = \operatorname{median}(\hat{\beta}_i^{(1)}, \dots, \hat{\beta}_i^{(K)})$, and the weights for the unique paths as $w_i^{(k)} = 1/|\hat{\beta}_i^{(0)} - \hat{\beta}_i^{(k)}|$, $k = 1, \dots, K$. A similar equation, stacked across all subjects, is applied in this adaptive weighting scheme, which is the default setting of **multivar**. It is known that the adaptive Lasso penalty reduces the bias of the standard Lasso estimator. However, how these adaptive weights affect the convergence rates of the individual components has not been studied yet. Moreover, the framework still relies on the fully stacked equations in (4).

1.2.1 Related Works

Due to the scarcity of studies, there are few VAR-type models that explicitly identify common and unique paths. Nevertheless, several approaches related to multi-VAR have been proposed. For example, Wilms et al. (2018) developed a multiclass VAR model in which the vectorized VAR transition matrices are assumed to be similar across classes, corresponding to subjects in our setting. Their method employs ℓ_1 and fused Lasso penalties to enforce sparsity and similarity across subjects. However, this approach is closer to differential analysis (Shojaie; 2021); rather than decomposing common and unique components, it focuses on encouraging similarity in individual dynamics across groups.

A model introduced by Skripnikov and Michailidis (2019) is more closely aligned with our setting in that it explicitly distinguishes between common and unique components. In particular, they

impose nonoverlapping supports between the two by adding a penalty, which makes the formulation nonconvex. At the same time, they assume that the supports of the common components are shared across subjects, while their values may differ; that is, $\text{Supp}(\alpha^{(0,k_1)}) = \text{Supp}(\alpha^{(0,k_2)})$ for $k_1 \neq k_2$ but not necessarily $\alpha^{(0,k_1)} = \alpha^{(0,k_2)}$, unlike the decomposition in (2).

Similarly, Manomaisaowapak and Songsiri (2022) proposed three variants of joint VAR models. In their framework, common components are identified using a group Lasso penalty, while individual components are encouraged to be similar across subjects via nonconvex fused penalties. Their approach lies between the two aforementioned models, but distinguishing common from individual components requires fitting multiple models, which cannot be identified simultaneously.

A recent study by Lyu et al. (2024) focuses on covariate-driven population patterns with latent VAR formulations rather than heterogeneous subject-specific effects. In this framework, the observations are generated by a latent VAR model whose transition matrix is decomposed into a low-dimensional individual covariate multiplied by a common sparse matrix, plus a random measurement error component. The primary aim is to identify population-level patterns explained by covariates, whereas our method is explicitly designed to capture both shared and individual dynamics across subjects, with an emphasis on establishing identifiability and providing formal inferential tools. Moreover, their formulation is conceptually closer to low-rank VAR models (e.g., Basu et al.; 2019; Alquier et al.; 2020).

1.3 Contributions

This work makes two main contributions. First, we propose a new identifiability condition for the multi-VAR model, grounded in the communication-efficient data integration framework of Maity et al. (2022). This condition offers a statistically principled alternative to median-based identification, improving estimation efficiency and ensuring robustness in heterogeneous settings. Second, building on this foundation, we develop a hypothesis testing framework specifically designed for multiple-subject high-dimensional VAR models. Our framework enables rigorous assessment of the significance and homogeneity of individual paths as well as the validity of shared paths, thereby addressing a critical methodological gap.

1.4 Organization of Paper

The rest of the paper is organized as follows. Section 2 introduces the new estimation framework and derives the convergence rates of the estimators. Section 3 presents the inference framework associated with the estimation procedure, along with the theory of the hypothesis tests. Section 4

reports numerical experiments comparing our method with existing joint estimation frameworks, as well as hypothesis testing results. Section 5 applies the proposed methods to neuroimaging data. Finally, Section 6 concludes with a discussion.

2 Estimation

In this section, we introduce the proposed estimation framework. We then present the theoretical results on convergence rates, demonstrating that the proposed method achieves improved rates compared to existing approaches.

2.1 Estimation Procedure

The key difference from the original multi-VAR estimation framework is that we first estimate the individual VAR models separately and then aggregate them, rather than jointly estimating all parameters as in (3) and (6).

First, we use equation-by-equation for the VAR models of each k^{th} subject, $k = 1, \dots, K$. Note that the i^{th} equation in the d -dimensional VAR(p) model is written as

$$X_{i,t}^{(k)} = \sum_{\ell=1}^p [\Phi_{\ell}^{(k)}]_{i:} X_{t-\ell}^{(k)} + \epsilon_{i,t}^{(k)} = \begin{pmatrix} X_{t-1}^{(k)'} & \dots & X_{t-p}^{(k)'} \end{pmatrix} \begin{pmatrix} [\Phi_1^{(k)}]_{i:} \\ \vdots \\ [\Phi_p^{(k)}]_{i:} \end{pmatrix} + \epsilon_{i,t}^{(k)} =: \mathcal{X}^{(k)} \beta_i^{(k)} + \epsilon_{i,t}^{(k)},$$

where $\epsilon_{i,t}^{(k)} \sim \mathcal{N}(0, (\sigma_i^{(k)})^2)$ and $\epsilon_{i_1,t}^{(k)} \perp \epsilon_{i_2,t}^{(k)}$ for $i_1 \neq i_2$. The regression equation for i^{th} variable is

$$\underbrace{\begin{pmatrix} X_{i,p+1}^{(k)} \\ \vdots \\ X_{i,T}^{(k)} \end{pmatrix}}_{\mathcal{Y}_i^{(k)}} = \underbrace{\begin{pmatrix} X_p^{(k)'} & \dots & X_1^{(k)'} \\ \vdots & \ddots & \vdots \\ X_{T-1}^{(k)'} & \dots & X_{T-p}^{(k)'} \end{pmatrix}}_{\mathcal{X}^{(k)}} \beta_i^{(k)} + \underbrace{\begin{pmatrix} \epsilon_{i,p+1}^{(k)} \\ \vdots \\ \epsilon_{i,T}^{(k)} \end{pmatrix}}_{E_i^{(k)}}.$$

For $\mathcal{Y}_i^{(k)}$, $i = 1, \dots, d$, we get the estimator $(\hat{\beta}_i^{(k)}) \in \mathbb{R}^{dp}$ by Lasso program,

$$\hat{\beta}_i^{(k)} = \underset{\beta_i^{(k)} \in \mathbb{R}^{dp}}{\operatorname{argmin}} \left\{ \mathcal{L}(\beta_i^{(k)}) = \frac{1}{2N_k} \left\| \mathcal{Y}_i^{(k)} - \mathcal{X}^{(k)} \beta_i^{(k)} \right\|_2^2 + \lambda_i^{(k)} \|\beta_i^{(k)}\|_1 \right\}. \quad (7)$$

By following the debiased Lasso estimator (e.g., Basu et al.; 2024; Adamek et al.; 2023), the each i^{th} equation in (7) has a debiased dp -dimensional estimator

$$\tilde{\beta}_i^{(k)} = \hat{\beta}_i^{(k)} + \frac{1}{N_k} \hat{\Theta}^{(k)} \mathcal{X}^{(k)'} (\mathcal{Y}_i^{(k)} - \mathcal{X}^{(k)} \hat{\beta}_i^{(k)}), \quad i = 1, \dots, d, \quad (8)$$

where $\hat{\Theta}^{(k)}$ is the approximated inverse of the Hessian at k^{th} subject regarding the squared loss function $\frac{1}{2N_k} \|\mathcal{Y}_i^{(k)} - \mathcal{X}^{(k)} \hat{\beta}_i^{(k)}\|_2^2$, which is the common across all $\hat{\beta}_i^{(k)}$ s, $i = 1, \dots, d$. It is computed as $\hat{\Theta}^{(k)} = (\hat{\gamma}^{(k)})^{-2} \hat{\Gamma}^{(k)}$, which consists of

$$\hat{\Gamma}^{(k)} = \begin{pmatrix} 1 & -\hat{\gamma}_{1,2}^{(k)} & \dots & -\hat{\gamma}_{1,dp}^{(k)} \\ -\hat{\gamma}_{2,1}^{(k)} & 1 & \dots & -\hat{\gamma}_{2,dp}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{dp,1}^{(k)} & -\hat{\gamma}_{dp,2}^{(k)} & \dots & 1 \end{pmatrix},$$

where $\hat{\gamma}_j^{(k)} = \{\hat{\gamma}_{j,l}, l \in \{1, \dots, dp\} \setminus \{j\}\}$. Each of the vectors is obtained by the nodewise regression,

$$\hat{\gamma}_j^{(k)} = \underset{\gamma_j^{(k)} \in \mathbb{R}^{dp-1}}{\operatorname{argmin}} \left\{ \frac{1}{2N_k} \|\mathcal{X}_j^{(k)} - \mathcal{X}_{-j}^{(k)} \gamma_j^{(k)}\|_2^2 + \lambda_j \|\gamma_j^{(k)}\|_1 \right\}, \quad (9)$$

where $\mathcal{X}_j^{(k)}$ is j^{th} column in $\hat{\Gamma}^{(k)}$ and $\mathcal{X}_{-j}^{(k)}$ is $\hat{\Gamma}^{(k)}$ with j^{th} column removed. By taking $(\hat{\gamma}_j^{(k)})^2 = \frac{1}{N_k} \|\mathcal{X}_j^{(k)} - \mathcal{X}_{-j}^{(k)} \hat{\gamma}_j^{(k)}\|_2^2 + \lambda_j \|\hat{\gamma}_j^{(k)}\|_1$, we have

$$(\hat{\gamma}^{(k)})^{-2} = \operatorname{diag}(1/(\hat{\gamma}_1^{(k)})^2, \dots, 1/(\hat{\gamma}_{dp}^{(k)})^2).$$

Next, we aggregate the individual estimators to obtain the common paths, then separate the unique paths. By following Maity et al. (2022), one can view the identification of the common effects as finding a robust M -estimator for measurement contaminated by influential errors. That is, for the j^{th} coordinate in the parameter vector of j^{th} variable in k^{th} subject $(\beta_i^{(k)})_j \in \mathbb{R}$, $k = 1, \dots, K$, $j = 1, \dots, dp$, $i = 1, \dots, d$, where d is the number of variables and p is the lag order of the VAR model, it is assumed to follow

$$(\beta_i^{(k)})_j \sim (1 - c)\mathcal{N}((\alpha_i^{(0)})_j, \sigma_0^2) + cG_{ij}, \quad (10)$$

for some unknown distribution G_{ij} , where $(\alpha_i^{(0)})_j$ is the j^{th} coordinate of the common path of i^{th} variable across K subjects.

Inspired by (10), the common path can be obtained by minimizing the sum of the redescending loss function (e.g., Chapter 4.8 in Huber and Ronchetti; 2011)

$$(\tilde{\alpha}_i^{(0)})_j = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \left\{ L_{ij}(x) := \sum_{k=1}^K \min\{((\tilde{\beta}_i^{(k)})_j - x)^2, \eta_j^2\} \right\} \quad (11)$$

where $\tilde{\beta}_i^{(k)}$ is dp -dimensional debiased estimator of β_i in (8). Naturally, the unique paths are defined as $\tilde{\alpha}_i^{(k)} = \tilde{\beta}_i^{(k)} - \tilde{\alpha}_i^{(0)}$. To recover the sparsity, either the hard threshold (HT) or soft threshold (ST) is applied to $(\tilde{\alpha}_i^{(0)})_j$ and $(\tilde{\alpha}_i^{(k)})_j$ to produce sparse estimators $(\hat{\alpha}_i^{(0)})_j$ and $(\hat{\alpha}_i^{(k)})_j$, respectively, $j = 1, \dots, dp$, $i = 1, \dots, d$, and $k = 1, \dots, K$. The thresholds are defined as

$$HT_{\delta_k}(\theta_j) = \theta_j 1_{\{|\theta_j| \geq \delta_k\}}, \quad (12)$$

$$ST_{\delta_k}(\theta_j) = \operatorname{sign}(\theta_j) \max\{|\theta_j| - \delta_k, 0\}, \quad (13)$$

for some univariate parameter θ_j . The desirable scales of the threshold are known as $\delta_k \sim \sqrt{\frac{\log q}{N_k}}$, $k = 1, \dots, K$, and $\delta_0 \sim \sqrt{\frac{\log q}{KN_{\min}}}$, where $q = d^2p$ and $N_{\min} = \min_k N_k$. Throughout this study, we only focus on hard thresholding, as the theoretical results described in Section 2.2 are identical for both choices.

Remark 2.1. In practice, it is not feasible to tune η_{ij} and δ_k individually. Therefore, we use three-layer cross-validation to determine the threshold $\eta = \eta_{ij}$ for the redescending loss function (11), along with two constants, c_0 and c_K , defined as $\delta_0 = \max_k \kappa(\Sigma_\varepsilon^{(k)}) \sqrt{\frac{\log q}{c_0 K N_{\min}}}$ and $\delta_k = c_K \kappa(\hat{\Sigma}_\varepsilon^{(k)}) \sqrt{\frac{\log q}{N_k}}$, $k = 1, \dots, K$, where $\{\kappa(\hat{\Sigma}_\varepsilon^{(k)})\}_{k=1, \dots, K}$ are the condition numbers of the estimated covariance matrices $\hat{\Sigma}_\varepsilon^{(k)}$ of the residuals, defined analogously as $\{\kappa(\Sigma_\varepsilon^{(k)})\}_{k=1, \dots, K}$ in Section 2.2. Note that while the form of δ_0 aligns with the theoretical motivation in Proposition 2.1, its δ_k used in the cross-validation is determined empirically. The mean prediction errors averaged over all K subjects, similar to those commonly used in cross-validation, are highly sensitive to the values of η_{ij} but less sensitive to the grids of the other constants. Despite its robustness, we empirically observed that the level of sparsity depends substantially on the choices of constants c_0 and c_K . Furthermore, in some cases, the thresholding for unique components can be too stringent, even when the cross-validation error does not differ significantly. This often results in eliminating all unique paths if scaling is not properly applied. To our knowledge, there is no established standard for choosing the grids. Based on our empirical experiments, we set the grid for c_0 to range from 0.1 to 1 at equal intervals, the grid for c_K from 0.5 to 1, and the grid for η_{ij} from $\min_{i,j,k} (\tilde{\beta}_i^{(k)})_j$ to $\max_{i,j,k} (\tilde{\beta}_i^{(k)})_j$ at equal intervals.

2.2 Theory on Estimation

In this section, we establish the convergence rates of the proposed estimators. The main result, Proposition 2.1, relies on several technical lemmas, whose proofs are provided in Appendix A. Without loss of generality, we set $p = 1$ and suppress the lag index in $\Phi_\ell^{(k)}$, writing simply $\Phi^{(k)}$. Note that a VAR(p) model with $p > 1$ can be equivalently reformulated as a VAR(1) model (see Basu and Michailidis; 2015 for a related discussion).

To present the result, we define three standard conditions commonly applied in high-dimensional time series modeling, as discussed in Basu et al. (2024).

- (a) Stability regarding VAR transition matrix: for $\Phi^{(k)} = I - \Phi_1^{(k)} z$, $z \in \mathcal{C}$, consider

$$\|\Phi^{(k)}\| = \max_{|z| \leq 1} \|\Phi^{(k)}(z)\|, \quad \|(\Phi^{(k)})^{-1}\| = \max_{|z| \leq 1} \|(\Phi^{(k)})^{-1}(z)\|.$$

Then the condition number is $\kappa(\Phi^{(k)}) := \|\Phi^{(k)}\| \|(\Phi^{(k)})^{-1}\| < \infty$.

(b) Error covariance matrix of VAR model: for $\sigma_{k,\max}^2 = \max_j \sigma_{k,j}^2$ and $\sigma_{k,\min}^2 = \min_j \sigma_{k,j}^2$, the condition number is $\kappa(\Sigma_\epsilon^{(k)}) := \sigma_{k,\max}^2 / \sigma_{k,\min}^2 < \infty$.

(c) Sparsity level: $s_{0,k} = \|\Phi^{(k)}\|_0$ so that $s_{0,\max} = \max_k s_{0,k}$. Also, for $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$, define

$$s_{j,k} = \|\Theta_{j\cdot}^{(k)}\|_0, \quad s_{\max,k} = \max_j s_{j,k}. \quad (14)$$

Note that Van de Geer et al. (2014), on which our paper is founded, requires sparsity of $\Theta^{(k)}$, which the theory in Maity et al. (2022) also relies on (see Sections 2.4 and 2.5 therein). They defined

$$s_k = \sum_{1 \leq i,j \leq d} |[\Theta^{(k)}]_{ij}| = \text{vec}(\Theta^{(k)}).$$

In contrast, Basu et al. (2024) adopt the weak sparsity assumption for $\Theta^{(k)}$ as proposed by Javanmard and Montanari (2014). Regardless of the specific sparsity assumptions employed, the resulting convergence rates remain unchanged, consistent with the findings of Zhang and Zhang (2014). Consequently, the primary purpose of our proof is to bridge the gap between these differing assumptions.

Assumption 2.1. For each k , we consider an asymptotic regime where $d, N_k \rightarrow \infty$,

$$\kappa^2(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 \max\{s_{\max,k}, s_{0,k}\} \frac{\log d}{\sqrt{N_k}} \rightarrow 0. \quad (15)$$

This allows d to grow with N_k as long as $\max\{s_{\max,k}, s_{0,k}\}$ grow as $\mathcal{O}(\sqrt{N_k})$. In particular, according to Basu et al. (2024), if the eigenvalues of $\Sigma_\epsilon^{(k)}$ and the modulus of eigenvalues of $\Phi^{(k)}(z)$ are both bounded away from zero and infinity, the terms involving $\Sigma_\epsilon^{(k)}$ and $\Phi^{(k)}$ will not appear in the convergence analysis and match the known error bounds in the high-dimensional regression with i.i.d. data.

In addition, we introduce a set of conditions that directly correspond to Assumption 4 in Maity et al. (2022).

Assumption 2.2. The following conditions are assumed to be held.

- (a) Let I_j be the set of indices for $(\beta_i^{(k)})_j$'s which are considered as inliers. We assume $|I_j|/K \geq 4/7$.
- (b) Let $\mu_j = \frac{1}{|I_j|} \sum_{k \in I_j} (\beta_i^{(k)})_j$. Let δ be the smallest positive real number such that $(\beta_i^{(k)})_j \in [\mu_j - \delta, \mu_j + \delta]$ for all $k \in I_j$. We assume that none of the $(\beta_i^{(k)})_j$'s are in the intervals $[\mu_j - 5\delta, \mu_j - \delta]$ or $[\mu_j + \delta, \mu_j + 5\delta]$.
- (c) Let $\delta_2 = \min_{k_1 \in I_j, k_2 \notin I_j} |(\beta_i^{(k_1)})_j - (\beta_i^{(k_2)})_j|$ is the minimum separation between inliers and outliers. Clearly, $4\delta < \delta_2$. We choose η_j such that $2\delta < \eta_j < \delta_2/2$.

Note that conditions (a) and (c) are technical conditions required to complete Result 5 in Maity et al. (2022). Condition (b) is crucial, as it distinguishes between inliers and outliers. The number factors appearing in conditions (a) through (c) are symbolic rather than carrying actual meaning.

Proposition 2.1. *Suppose that Assumptions 2.1 and 2.2 hold. Define $\delta_0 = \max_k \kappa(\Sigma_\epsilon^{(k)}) \sqrt{\frac{\log d^2}{(1-e)KN_{\min}}}$, $0 < e < 1$, and $\delta_k = \kappa(\Sigma_\epsilon^{(k)}) \sqrt{\frac{\log d^2}{N_k}}$. For sufficiently large N_k , $k = 1, \dots, K$, we have the following:*

- (a) $\|\hat{\alpha}^{(0)} - \alpha^{(0)}\|_\infty \leq \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{\log d^2}{KN_{\min}}} \right),$
- (b) $\|\hat{\alpha}^{(0)} - \alpha^{(0)}\|_1 \leq \mathcal{O}_{\mathbb{P}} \left(s_{0,\max} \sqrt{\frac{\log d^2}{KN_{\min}}} \right),$
- (c) $\|\hat{\alpha}^{(0)} - \alpha^{(0)}\|_2 \leq \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{s_{0,\max} \log d^2}{KN_{\min}}} \right),$
- (d) $\|\hat{\alpha}^{(k)} - \alpha^{(k)}\|_\infty \leq \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{\log d^2}{N_k}} \right),$
- (e) $\|\hat{\alpha}^{(k)} - \alpha^{(k)}\|_1 \leq \mathcal{O}_{\mathbb{P}} \left(s_{k,\max} \sqrt{\frac{\log d^2}{N_k}} \right),$
- (f) $\|\hat{\alpha}^{(k)} - \alpha^{(k)}\|_2 \leq \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{s_{k,\max} \log d^2}{N_k}} \right).$

Note that the convergence rates of both the common and unique components (5) change, as they now depend on their respective sparsity levels.

Proof. Recall the Lemma 11 in Lee et al. (2017): If $\|\tilde{\theta} - \theta^*\|_\infty < \delta$, then for $\hat{\theta} = HT_\delta(\tilde{\theta})$, where $HT_\delta(\cdot)$ is defined in (12), the following holds:

- (a) $\|\hat{\theta} - \theta^*\|_\infty < 2\delta,$
- (b) $\|\hat{\theta} - \theta^*\|_2 < 2\sqrt{2s}\delta,$
- (c) $\|\hat{\theta} - \theta^*\|_1 < 2\sqrt{2s}\delta,$

where s is the sparsity level of θ^* . The analogous results hold for $\hat{\theta} = ST_\delta(\tilde{\theta})$, where $ST_\delta(\cdot)$ is defined in (13). It remains to show that values of δ_k , $k = 0, 1, \dots, K$, satisfy the condition in Proposition 2.1, which is directly from Lemma A.4. $\square$

3 Hypothesis Tests

In this section, we introduce the hypothesis testing framework for the proposed approach. Within this framework, we describe three representative and practically relevant tests. First, we present

the test of nullity and the test of homogeneity across all subjects; these two tests share a common foundation, as they represent special cases of a more general setting. In addition, we introduce the test of significance for common paths, which is also derived from the framework.

3.1 Inference Procedure

We begin by formulating the most general setting for the hypothesis tests. Define $\hat{V}_{ij}^{(k)} = \hat{\sigma}_{k,i}^2 [\hat{\Theta}^{(k)} \hat{\Sigma}^{(k)} \hat{\Theta}^{(k)'}]_{jj}$ where $\hat{\Sigma}^{(k)} = \frac{1}{N_k} \mathcal{X}^{(k)'} \mathcal{X}^{(k)}$, $\hat{\Theta}^{(k)}$ is defined in (8), and

$$\hat{\sigma}_{k,i}^2 = \frac{1}{N_k} \sum_{t=1}^{N_k} \left((\mathcal{Y}_i^{(k)})_t - (\mathcal{X}^{(k)})_{t,:} \hat{\beta}_i^{(k)} \right)^2, \quad i = 1, \dots, d, \quad k = 1, \dots, K.$$

Let $\tilde{\beta}_{(i,j)} = ((\tilde{\beta}_i^{(1)})_j, \dots, (\tilde{\beta}_i^{(K)})_j)' \in \mathbb{R}^K$ and $\beta_{(i,j)} = ((\beta_i^{(1)})_j, \dots, (\beta_i^{(K)})_j)' \in \mathbb{R}^K$ be K -dimensional estimators of j^{th} entries in i^{th} variable across K subjects and its population analog. We also denote K -dimensional diagonal matrix from $\hat{V}_{ij}^{(k)}$, $k = 1, \dots, K$, by

$$\hat{V}_{(i,j)} = [\hat{V}_{ij}^{(k)}]_{k=1}^K \in \mathbb{R}^{K \times K}. \quad (16)$$

In addition, we have a contrast $D \in \mathbb{R}^{a \times K}$ and a scalar matrix M ,

$$M = \text{diag}(\sqrt{N_1}, \dots, \sqrt{N_K}) \in \mathbb{R}^{K \times K}. \quad (17)$$

Suppose that we are interested in the hypothesis $D\beta_{(i,j)} = c$ for some contrast D whose rank is $\text{rank}(D) = a$ (that is, every hypothesis is tested separately). The test statistic

$$\chi_{ij}^2(c) = [D\tilde{\beta}_{(i,j)} - c]' [DM\hat{V}_{(i,j)}MD']^{-1} [D\tilde{\beta}_{(i,j)} - c], \quad (18)$$

will follow χ^2 -distribution with degree of freedom a . Note that the result corresponds to inference on a single entry in a single-subject VAR model in Section 2.7 in Basu et al. (2024) by taking $D = \text{diag}(1, 0, \dots, 0) \in \mathbb{R}^K$ and $M = \text{diag}(\sqrt{N_1}, 0, \dots, 0)$.

We use (18) to introduce hypothesis tests that are practically relevant to our setting. We begin with the test of nullity, which assesses whether the paths are null across all subjects. For $i = 1, \dots, d$ and $j = 1, \dots, dp$, we are interested in

$$H_0 : (\beta_i^{(1)})_j = \dots (\beta_i^{(K)})_j = 0 \quad \text{vs} \quad H_1 : \text{not } H_0. \quad (19)$$

Write the hypothesis (19) into

$$H_0 : D\beta_{(i,j)} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} (\beta_i^{(1)})_j \\ (\beta_i^{(2)})_j \\ \vdots \\ (\beta_i^{(K)})_j \end{pmatrix} = 0 \quad \text{vs} \quad H_1 : \text{not } H_0.$$

Under the null hypotheses in (19) is true, the test statistic $\chi_{ij}^2(0)$ follows χ^2 distribution with degree of freedom K . We reject the null hypothesis if

$$\mathbb{P}\left(\chi^2(K) > \chi_{ij}^2(0)\right) \leq \alpha. \quad (20)$$

for significance level α . We refer to this as the test of nullity.

Next, we define the test of homogeneity, which assesses whether the paths are consistent across all subjects. For $i = 1, \dots, d$ and $j = 1, \dots, dp$,

$$H_0 : (\beta_i^{(1)})_j = \dots = (\beta_i^{(K)})_j \quad vs \quad H_1 : \text{not } H_0. \quad (21)$$

Write the hypothesis (21) into

$$H_0 : D\beta_{(i,j)} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix} \begin{pmatrix} (\beta_i^{(1)})_j \\ (\beta_i^{(2)})_j \\ (\beta_i^{(3)})_j \\ \vdots \\ (\beta_i^{(K-1)})_j \\ (\beta_i^{(K)})_j \end{pmatrix} = 0 \quad vs \quad H_1 : \text{not } H_0.$$

Under the null hypotheses in (21) is true, the test statistic $\chi_{ij}^2(0)$ follows χ^2 distribution with degree of freedom $K - 1$. Then we can compute p -values, similar to (20).

Finally, we propose the test of significance, which evaluates the importance of a common path $(\alpha_i^{(0)})_j$ across subjects. Note that we use the standard Z-test while the hypothesis tests described above are based on Wald statistics. Suppose that we are interested in

$$H_0 : (\alpha_i^{(0)})_j = 0, \quad H_1 : (\alpha_i^{(0)})_j \neq 0, \quad (22)$$

by taking into account the average of the variances of the subjects that contribute to the common path. That is, let $\mathcal{J}_{ij} \subseteq \{1, \dots, K\}$ be the set of indices of k whose value of $(\tilde{\beta}_i^{(k)})_j$ is considered as inliers by (11):

$$|(\tilde{\beta}_i^{(k)})_j - (\tilde{\alpha}_i^{(0)})_j| \leq \eta_j \quad (23)$$

Then for $N_{ij} = \frac{1}{|\mathcal{J}_{ij}|} \sum_{k \in \mathcal{J}_{ij}} N_k$ we have

$$Z_{ij}((\alpha_i^{(0)})_j) = \frac{\sqrt{N_{ij}}((\tilde{\alpha}_i^{(0)})_j - (\alpha_i^{(0)})_j)}{\sqrt{\frac{1}{|\mathcal{J}_{ij}|^2} \sum_{k \in \mathcal{J}_{ij}} \hat{V}_{ij}}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (24)$$

Therefore, the hypothesis (22) is considered as a standard normal test: We reject the null hypothesis in (22) if $\mathbb{P}(Z > |Z_{ij}(0)|) < \alpha/2$ for a standard normal random variable Z with the significance level α .

3.2 Theory on Inference

In this section, we establish the asymptotic distributions of the test statistics. Specifically, Proposition 3.1 shows that the Wald-type test statistic (18) converges in distribution to a chi-squared random variable. Corollary 3.1 then presents the three practical hypothesis tests derived from this result.

Proposition 3.1. *Suppose that Assumptions 2.1 and 2.2 hold. Consider the hypothesis test written in the form $D\beta = c$ for some contrast D with rank a and constant c . Suppose that $\hat{V}_{(i,j)}, D, M$ are defined in (16) and (17). Under the null hypothesis,*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left([D\tilde{\beta}_{(i,j)}]' [DM\hat{V}_{(i,j)}MD']^{-1} [D\tilde{\beta}_{(i,j)}] \leq x \right) - \mathbb{P}(\chi^2(a) \leq x) \right| = \mathcal{O}_{\mathbb{P}}(1). \quad (25)$$

where $\chi^2(a)$ is the Wald-type test statistic with degree of freedom a .

Proof. Recall that from (30),

$$\sqrt{N_k} \left((\tilde{\beta}_i^{(k)})_j - (\beta_i^{(k)})_j \right) = \frac{1}{\sqrt{N_k}} \hat{\Theta}_{j \cdot}^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} + \sqrt{N_k} (\Delta_i^{(k)})_j.$$

From Lemma A.2, $\sqrt{N_k} \|\Delta_i^{(k)}\|_{\infty} = \mathcal{O}_{\mathbb{P}}(1)$. By using (36) and Lemma A.6, the first term converges to $\mathcal{N}(0, \sigma_{k,i}^2 \Theta_{jj}^{(k)})$. Hence, by Slutsky's theorem,

$$\frac{\sqrt{N_k} \left((\tilde{\beta}_i^{(k)})_j - (\beta_i^{(k)})_j \right)}{\hat{\sigma}_{k,i} \sqrt{[\hat{\Theta}^{(k)} \hat{\Sigma}^{(k)} \hat{\Theta}^{(k)'}]_{jj}}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (26)$$

Recall that $\hat{V}_{(i,j)} = \text{diag}(\hat{V}_{ij}^{(1)}, \dots, \hat{V}_{ij}^{(K)})$, $\tilde{\beta}_{(i,j)} = ((\tilde{\beta}_i^{(1)})_j, \dots, (\tilde{\beta}_i^{(K)})_j)'$ and $\beta_{(i,j)} = ((\beta_i^{(1)})_j, \dots, (\beta_i^{(K)})_j)'$. For a given contrast D with rank a and $M = \text{diag}(\sqrt{N_1}, \dots, \sqrt{N_K})$, by Cramér-Wold theorem,

$$DM(\tilde{\beta}_{(i,j)} - \beta_{(i,j)}) \xrightarrow{d} \mathcal{N}(0, DMV_{(i,j)}MD').$$

Since each diagonal entry in $\hat{V}_{(i,j)}$ converges in probability to the corresponding diagonal entry in $V_{(i,j)}$, and the multiplication $D\hat{V}_{(i,j)}D'$ is continuous, we have $D\hat{V}_{(i,j)}D' \xrightarrow{p} DV_{(i,j)}D'$ by using continuous mapping theorem with a sufficiently large $N_{\min}$. Then, we have

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left(\frac{DM(\tilde{\beta}_{(i,j)} - \beta_{(i,j)})}{\sqrt{D\hat{V}_{(i,j)}D'}} \leq x \right) - \Phi(x) \right| = \mathcal{O}_{\mathbb{P}}(1).$$

where $\Phi(\cdot)$ is the standard normal CDF. Hence, for $D\beta_{(i,j)} = c$ with the contrast D of rank a ,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left([D\tilde{\beta}_{(i,j)} - c]' [DM\hat{V}_{(i,j)}MD']^{-1} [D\tilde{\beta}_{(i,j)} - c] \leq x \right) - \mathbb{P}(\chi^2(a) \leq x) \right| = \mathcal{O}_{\mathbb{P}}(1),$$

holds by Cochran's theorem (Cochran; 1934). $\square$

Corollary 3.1. *For the three specific hypothesis tests,*

- (a) *Under H_0 of the hypothesis (19), the test statistic $\chi_{ij}^2(0)$ follows χ^2 distribution with degree of freedom K .*
- (b) *Under H_0 of the hypothesis (21), the test statistic $\chi_{ij}^2(0)$ follows χ^2 distribution with degree of freedom $K - 1$.*
- (c) *Under H_0 of the hypothesis (22), the test statistic $Z_{ij}(0)$ follows the standard normal distribution.*

Proof. The first two statements are immediate from Proposition 3.1. For the third statement, it is sufficient to show that (24) holds. Define the indicator that the element of k^{th} subject contributes to the common path,

$$I_{ij}^{(k)} := 1_{\{|((\beta)_i^{(k)})_j - ((\alpha)_i^{(0)})_j| \leq \eta_j\}}.$$

From the conditions (a) – (c) in Assumption 2.2, the minimizer in (11) is unique, and none of $|((\beta)_i^{(k)})_j - ((\alpha)_i^{(0)})_j| = \eta_j$ holds for all k . Then the derivative of the objective function in (11) is defined by

$$L'_{ij}(x) = -2 \sum_{k=1}^K ((\tilde{\beta}_i^{(k)})_j - x) 1_{\{|(\tilde{\beta}_i^{(k)})_j - x| \leq \eta_j\}},$$

and it satisfies $L'_{ij}((\tilde{\alpha}_i^{(0)})_j) = 0$. So, by Taylor expansion around $(\alpha_i^{(0)})_j$,

$$0 = L'_{ij}((\alpha_i^{(0)})_j) + L''_{ij}((\alpha_i^{(0)})_j)((\tilde{\alpha}_i^{(0)})_j - (\alpha_i^{(0)})_j).$$

Note that $1_{\{|(\tilde{\beta}_i^{(k)})_j - ((\alpha)_i^{(0)})_j| \leq \eta_j\}} \xrightarrow{p} I_{ij}^{(k)}$. With $N_{ij} = (\sum_{k=1}^K I_{ij}^{(k)} N_k) / (\sum_{k=1}^K I_{ij}^{(k)})$,

$$L'_{ij}((\alpha_i^{(0)})_j) = -2 \sum_{k=1}^K I_{ij}^{(k)} ((\tilde{\beta}_i^{(k)})_j - ((\alpha)_i^{(0)})_j) + \mathcal{O}_{\mathbb{P}}(N_{ij}^{-1/2}),$$

$$L''_{ij}((\alpha_i^{(0)})_j) = 2 \sum_{k=1}^K I_{ij}^{(k)} + \mathcal{O}_{\mathbb{P}}(1).$$

Therefore,

$$(\tilde{\alpha}_i^{(0)})_j - (\alpha_i^{(0)})_j = \frac{\sum_{k=1}^K I_{ij}^{(k)} ((\tilde{\beta}_i^{(k)})_j - (\alpha_i^{(0)})_j)}{\sum_{k=1}^K I_k} + \mathcal{O}_p(N_{ij}^{-1/2}) = \frac{\sum_{k=1}^K I_{ij}^{(k)} ((\tilde{\beta}_i^{(k)})_j - (\beta_i^{(k)})_j)}{\sum_{k=1}^K I_k} + \mathcal{O}_p(N_{ij}^{-1/2}),$$

since $\sum_{k=1}^K I_{ij}^{(k)} ((\beta_i^{(k)})_j - (\alpha_i^{(0)})_j) = 0$. Therefore, by Slutsky's theorem,

$$\sqrt{N_{ij}}((\tilde{\alpha}_i^{(0)})_j - (\alpha_i^{(0)})_j) \xrightarrow{d} \mathcal{N}(0, W_{ij}^{(0)}),$$

where $\hat{W}_{ij}^{(0)} := \sum_{k=1}^K I_{ij}^{(k)} \hat{V}_{ij} / (\sum_{k=1}^K I_{ij}^{(k)})^2 \xrightarrow{p} W_{ij}^{(0)}$. Hence, the test statistic in (24) also follows the standard normal distribution under the null hypothesis. $\square$

4 Numerical Experiments

In this section, we present numerical experiments evaluating the proposed method in comparison with the benchmark approach and report its performance according to the defined metrics.

4.1 Simulation Setups

We focus on the case $p = 1$ with independent Gaussian errors $\epsilon_{i,t}^{(k)} \sim \mathcal{N}(0, 1)$. Among the d^2 possible paths, $s_0 d^2$ are designated as common, and $(\sum_{k=1}^K s_k) d^2$ unique paths are selected so that no overlaps occur across subjects. For estimation, we set $K = 10, 15$ and vary $d = 10, 20$ and the average sample lengths $T = 50, 200$, where the ranges are between 45-55 for $T = 50$ and 190-210 for $T = 200$. Three relative heterogeneity levels are considered, given by $(s_0, s_k) = (0.02, 0.04), (0.03, 0.03), (0.04, 0.02)$, denoted as high, medium, and low, respectively, while the overall sparsity is fixed at 6%. For each combination of settings, we repeat the simulations 50 times.

The proposed estimation framework in Section 2.1 is compared with the multi-VAR model in (3) (multi-VAR) and its adaptive Lasso variant in (6) (multi-VAR (A)). The benchmark methods in this simulation study are implemented using the R package `multivar` (Fisher et al.; 2022). As discussed in Section 1.2.1, existing methods outside the multi-VAR framework either do not estimate identically defined common paths, cannot jointly estimate common and subject-specific paths, or rely on low-rank modeling with individualized covariates. Moreover, the implementations of the second and third approaches described in that section cannot be modified to fit our simulation setup. In addition to the estimation results, we conduct three hypothesis tests based on the estimated models, tests of nullity, homogeneity, and significance, as described in Section 3.1.

To evaluate estimation performance, we compute the root mean square error (RMSE), sensitivity (Sens), and specificity (Spec) for $\alpha^{(0)}$ and averaged $\alpha^{(k)}$ across the K subjects. Specifically, for the true $(\alpha_i^{(0)})_j$ and its estimator $(\hat{\alpha}_i^{(0)})_j$,

$$\begin{aligned} \text{RMSE}(\alpha^{(0)}) &= \frac{\|\hat{\alpha}^{(0)} - \alpha^{(0)}\|_2}{\|\alpha^{(0)}\|_2}, & \text{RMSE}(\alpha^{(K)}) &= \frac{1}{K} \sum_{k=1}^K \frac{\|\hat{\alpha}^{(k)} - \alpha^{(k)}\|_2}{\|\alpha^{(k)}\|_2}, \\ \text{Sens}(\alpha^{(0)}) &= \frac{\sum_{i,j} 1_{\{(\hat{\alpha}_i^{(0)})_j \neq 0 \text{ \& } (\alpha_i^{(0)})_j \neq 0\}}}{\sum_{i,j} 1_{\{(\alpha_i^{(0)})_j \neq 0\}}}, & \text{Sens}(\alpha^{(K)}) &= \frac{1}{K} \sum_{k=1}^K \frac{\sum_{i,j} 1_{\{(\hat{\alpha}_i^{(k)})_j \neq 0 \text{ \& } (\alpha_i^{(k)})_j \neq 0\}}}{\sum_{i,j} 1_{\{(\alpha_i^{(k)})_j \neq 0\}}}, \\ \text{Spec}(\alpha^{(0)}) &= \frac{\sum_{i,j} 1_{\{(\hat{\alpha}_i^{(0)})_j = 0 \text{ \& } (\alpha_i^{(0)})_j = 0\}}}{\sum_{i,j} 1_{\{(\alpha_i^{(0)})_j = 0\}}}, & \text{Spec}(\alpha^{(K)}) &= \frac{1}{K} \sum_{k=1}^K \frac{\sum_{i,j} 1_{\{(\hat{\alpha}_i^{(k)})_j = 0 \text{ \& } (\alpha_i^{(k)})_j = 0\}}}{\sum_{i,j} 1_{\{(\alpha_i^{(k)})_j = 0\}}}. \end{aligned}$$

For the inference study, we compute the false discovery rate (FDR) and statistical power of the three tests at significance level $\alpha = 0.05$. To compute these metrics, let $\mathcal{S}$ and $\mathcal{S}^c$ denote the sets of index pairs (i, j) , $i, j = 1, \dots, d$, for which the null and alternative hypotheses are true,

respectively. Let $\hat{\mathcal{S}}$ and $\hat{\mathcal{S}}^c$ denote the sets of indices (i, j) for which the corresponding decisions are non-rejection and rejection, respectively. The FDR and power are then computed as

$$\text{FDR} = \frac{\sum_{i,j} 1_{\{(i,j) \in \hat{\mathcal{S}}^c \text{ \& } (i,j) \in \mathcal{S}\}}}{\sum_{i,j} 1_{\{(i,j) \in \hat{\mathcal{S}}^c\}}},$$

$$\text{Power} = \frac{\sum_{i,j} 1_{\{(i,j) \in \hat{\mathcal{S}}^c \text{ \& } (i,j) \in \mathcal{S}^c\}}}{\sum_{i,j} 1_{\{(i,j) \in \mathcal{S}^c\}}}.$$

Here, define the index sets $\mathcal{T}_{0K}, \mathcal{T}_{0^cK}, \mathcal{T}_{0K^c}, \mathcal{T}_{0^cK^c}$ as the sets of index pairs (i, j) , $i, j = 1, \dots, d$, corresponding to the following cases, respectively: (i) both the common and unique paths are zero; (ii) the common path is nonzero but the unique path is zero; (iii) the common path is zero but the unique path is nonzero; and (iv) both the common and unique paths are nonzero. These sets are disjoint from each other and

$$\mathcal{T}_{0K} \cup \mathcal{T}_{0^cK} \cup \mathcal{T}_{0K^c} \cup \mathcal{T}_{0^cK^c} = \{(i, j) : i, j = 1, \dots, d\}.$$

Then the null and alternative sets for each test are given by

- (a) Test of nullity: $\mathcal{S} = \mathcal{T}_{0K}$, $\mathcal{S}^c = \mathcal{T}_{0K^c} \cup \mathcal{T}_{0^cK} \cup \mathcal{T}_{0^cK^c}$.
- (b) Test of homogeneity: $\mathcal{S} = \mathcal{T}_{0K} \cup \mathcal{T}_{0^cK}$, $\mathcal{S}^c = \mathcal{T}_{0K^c} \cup \mathcal{T}_{0^cK^c}$.
- (c) Test of significance: $\mathcal{S} = \mathcal{T}_{0K} \cup \mathcal{T}_{0K^c}$, $\mathcal{S}^c = \mathcal{T}_{0^cK} \cup \mathcal{T}_{0^cK^c}$.

4.2 Estimation Results

The simulation results for estimation are presented in Figure (1). In the lower-dimensional setting ($d = 10$), both the identified common and unique paths from the proposed methods tend to yield smaller RMSEs compared with the original approaches. Although the trend reverses in the higher-dimensional setting ($d = 20$), the gap quickly narrows as the sample size increases (on average, as T grows from 50 to 200). Regarding other performance metrics, such as sensitivity and specificity, the behavior is similar to that observed in multi-VAR modeling with an adaptive scheme. We conjecture that the individually adjusted thresholding applied during sparsity recovery plays a role similar to that of adaptive weights in the adaptive Lasso approach. There are no significant differences across different numbers of subjects (K) for all performance metrics. This result is not surprising, as similar observations have been reported in previous studies of multi-subject time series modeling (e.g., Fisher et al.; 2022; Kim et al.; 2024). While the improvement in estimation accuracy may not be dramatic in higher dimensions, Figure 4 shows that a substantial amount of computational time is saved in achieving comparable results.

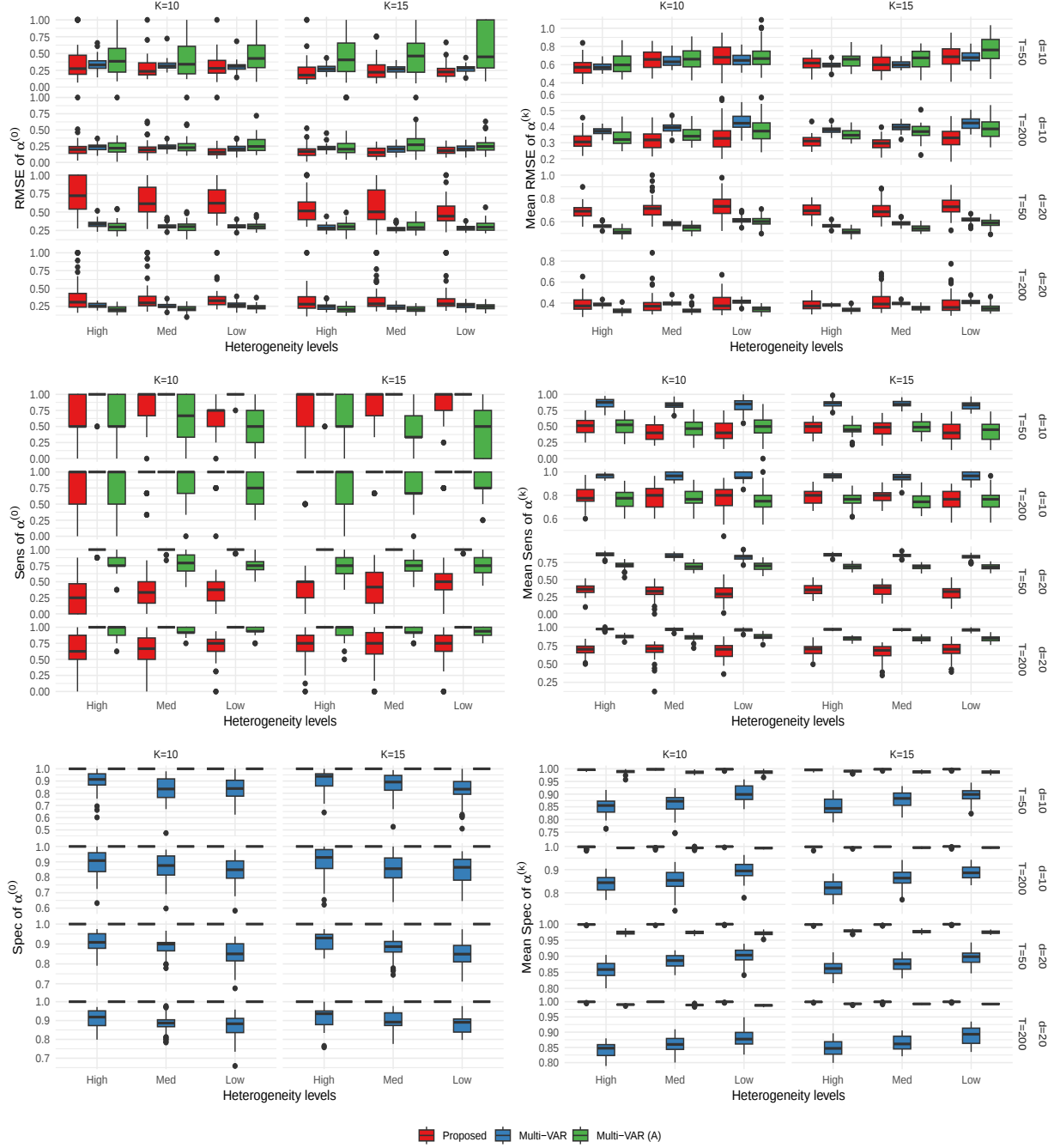


Figure 1: Boxplots of the root mean square error (RMSE) of $\alpha^{(0)}$ (top left), the average RMSE of $\alpha^{(k)}$ (top right), the sensitivity (Sens) of $\alpha^{(0)}$ (middle left), the average sensitivity of $\alpha^{(k)}$ (middle right), the specificity (Spec) of $\alpha^{(0)}$ (bottom left), and the average specificity of $\alpha^{(k)}$ (bottom right) under different combinations of d and average T (combinations indicated on the right tabs), K (each column), and heterogeneity levels (each axis). Red indicates the proposed method, while blue and green represent the benchmark methods.

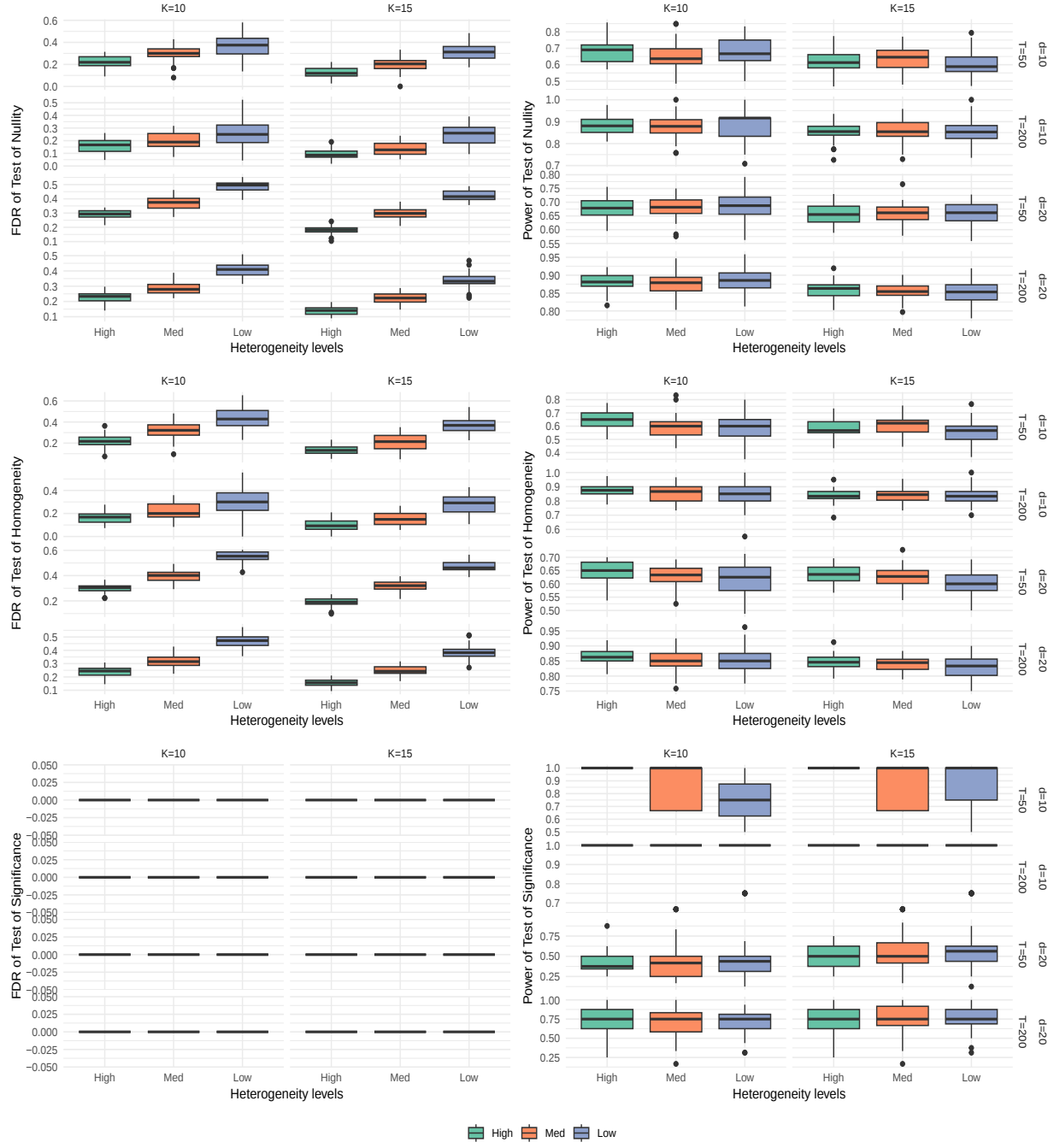


Figure 2: Boxplots of the FDRs (left columns) and powers (right columns) for the three hypothesis tests: test of nullity (top), test of homogeneity (middle), and test of significance (bottom), under different combinations of d and average T (combinations indicated on the right tabs), K (each column), and heterogeneity levels (each axis), presented through three colors.

4.3 Hypothesis Tests Results

The simulation results for the hypothesis tests are presented in Figure (2). After obtaining the estimation results, hypothesis testing was performed. Interestingly, while the FDR varies with the level of heterogeneity, being more favorable at higher levels, the power remains relatively stable across settings. For both measures, performance improves as the sample size increases, while remaining robust with respect to dimensionality. This robustness may arise because the hypothesis tests are conducted entrywise. As expected, the number of subjects does not substantially affect test performance. Across different types of tests, the results for the test of nullity are generally similar to those for the test of homogeneity, although the test of nullity tends to perform slightly better. Notably, for the test of significance, the FDR remains zero across all simulation settings, while the corresponding power behaves as expected. This finding suggests that the proposed testing procedure is both accurate and sensitive, which demonstrates the reliability of the hypothesis tests under the given scenarios.

5 Data Application

5.1 Data Description

We use task fMRI (tfMRI) data from the WU-Minn Human Connectome Project (HCP) (Van Essen et al.; 2013). The data have been preprocessed through the minimal pipeline described in Glasser et al. (2013). The HCP emotion processing task probes brain circuits involved in affective perception, particularly the amygdala. Participants complete two short fMRI runs (up to three minutes each) that alternate between emotion blocks and control blocks. In the emotion blocks, they match faces displaying fearful or angry expressions; in the control blocks, they match simple geometric shapes. Each block lasts approximately 18 seconds and includes several trials, isolating neural responses to emotional faces from general visual or matching processes. This task is widely used to study emotion reactivity, regulation, and individual differences across the large HCP sample (Barch et al.; 2013).

For our analysis, we select subjects whose behavioral and imaging data were both acquired and released in Quarter 1 (Q1) and who completed the full HCP 3T MRI protocol, ensuring that all scans are available across all time points. Our final sample consists of 12 females ($K = 12$), with ages ranging from 22 to 30 years. For cortical parcellation, we adopt the Schaefer2018 local-global atlas (Schaefer et al.; 2018), using the 400-parcel solution aligned with Yeo’s 17-network functional organization (Yeo et al.; 2011).

The atlas provides a predefined map that divides the brain into regions of interest (parcels), allowing researchers to summarize neural activity at the regional level rather than at individual voxels or vertices. The Schaefer atlas is derived from resting-state functional connectivity and combines fine local gradients with global clustering, producing parcellations at multiple resolutions. In the 400-parcel version, each parcel is assigned to one of $d = 17$ cortical networks, including the Default Mode, Salience/Ventral Attention, Dorsal Attention, Somatomotor, and Visual networks. Following this preprocessing step, we excluded abnormally high spikes observed at the beginning and end of the scans, yielding an average sample length of $T_k = 165$ for all subjects.

5.2 Application Results

The results are presented in Figure (3). The first row shows the estimated common paths across four approaches: multi-VAR without the adaptive scheme, multi-VAR with the adaptive Lasso penalty, the proposed estimation framework, and the proposed framework after hypothesis-based filtering. In terms of sparsity, the proposed framework (third column) yields sparser results than the adaptive multi-VAR model, and the subsequent hypothesis tests confirm this finding (fourth column). All nonzero paths identified by the proposed method are contained within the set of nonzero paths from the multi-VAR model, and approximately 88.9% of the nonzero paths overlap with those identified by the adaptive multi-VAR model.

The second row summarizes the number of unique nonzero paths across the 12 subjects. For each path, green indicates that more than six subjects exhibit a nonzero path, while orange indicates otherwise. Both the non-adaptive and adaptive multi-VAR models produce an excessively large number of nonzero paths. Specifically, 86.9% and 72.7% of paths are identified as nonzero by the two benchmarks, respectively, whereas only 19.7% and 12.5% of paths are identified as nonzero by the proposed method and hypothesis test results. Regarding the frequency with which each path is identified as nonzero, the non-adaptive and adaptive multi-VAR models yield medians of 2 and 1, third quartiles of 3 and 2, and maximum values of 8 for both. In contrast, the proposed method and hypothesis test results yield both medians and third quartiles of 0, with maximum values of only 2. This indicates that most of the unique nonzero paths occur only in single individuals.

The third row reports the number of nonzero individual paths. The non-adaptive multi-VAR model suggests that nearly all paths are present across subjects, resulting in uniformly dark green cells. Although somewhat less dense, the adaptive multi-VAR model still identifies 74.7% of paths as nonzero, yielding similar results. By contrast, the two proposed frameworks identify only 25.6% and 18% of paths as nonzero. In terms of frequency, while all four approaches reach a maximum of 12, their median values are 12, 3.46, 1.29, and 0.96, respectively. Under limited sample lengths,

the proposed methods facilitate easier interpretation by producing sparser models while preserving heterogeneous patterns.

Three paths are particularly relevant for emotion processing. At the individual level, hypothesis tests show that all participants have nonzero paths between ventral attention A and default mode subdivision B in both directions. Using the proposed method reveals additional connections among multiple default mode subdivisions (not only B but also C and D). This corresponds to the idea that salience detection systems influence self-referential and internally oriented processes (e.g., Seeley et al.; 2007; Andrews-Hanna et al.; 2010; Menon; 2011) and confirms that externally salient emotional stimuli are integrated with internal evaluations, linking perception of emotion to autobiographical and self-related representations. Paths from frontoparietal network A to limbic B are consistently observed across all participants. This corresponds to the integration of affective appraisal with cognitive control systems (e.g., Ochsner and Gross; 2005; Buhle et al.; 2014; Etkin et al.; 2015) and shows that executive networks help regulate responses to emotional stimuli, consistent with prior findings on top-down control of emotion.

Focusing on the result from the estimation only, one path shared by all participants is a directional connection from Limbic A to Limbic B. This corresponds to strong coordination within the limbic system, where interactions between the amygdala, hippocampus, and orbitofrontal cortex support emotion evaluation and integrate emotional experiences with memory (e.g., Critchley et al.; 2004; Ochsner and Gross; 2005; Buhle et al.; 2014). This corresponds to the central role of limbic networks in coordinating emotion processing.

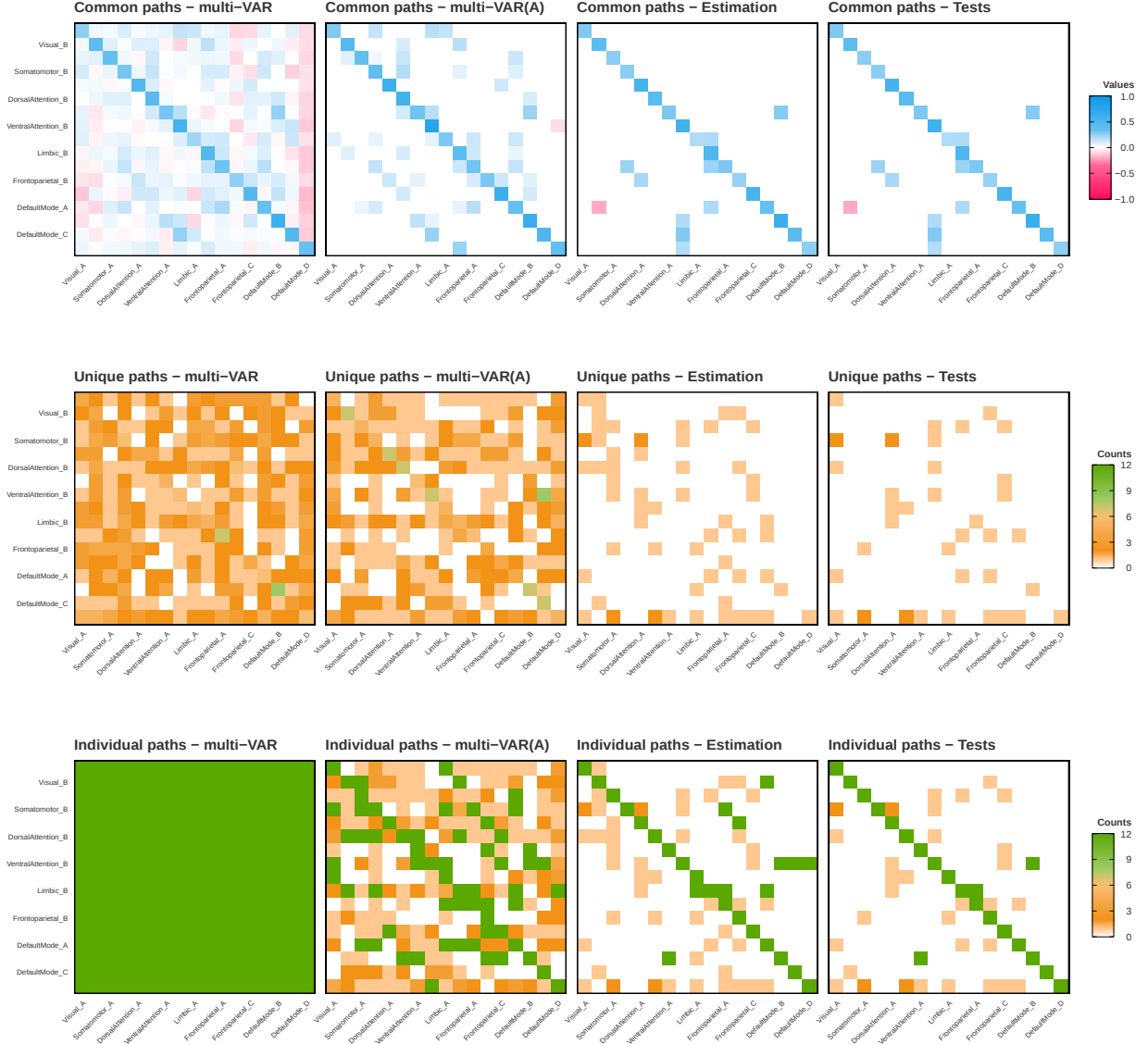


Figure 3: Path identification from tfMRI on emotion processing. Each row presents results for common paths (top), unique paths (middle), and individual paths (bottom). The common paths are shown with their estimated values, while the unique and individual paths are summarized by the counts of nonzero entries across subjects. The first two columns report results from the stratified Lasso and its adaptive analogue (multi-VAR and multi-VAR(A), respectively), the third column shows results from the estimators of the proposed framework, and the last column presents the estimators after filtering through the hypothesis testing procedure. For each node in the common, unique, and individual path matrices, only entries for which the null hypotheses of significance, homogeneity, or nullity are rejected are colored.

6 Discussion

Adopting the new identifiability restriction for the common path enables convergence rates that are tailored to each subject’s sparsity level and sample size. To the best of our knowledge, this work provides the first systematic framework for conducting inference on both commonality and heterogeneity in multiple-subject HDTS models (multi-VAR). Across the simulation studies, the proposed algorithm performs reliably and is sufficiently accurate and fast to replace existing methods in terms of estimator quality, as measured by various metrics. The performance of the hypothesis testing framework, evaluated using standard criteria, aligns closely with the asymptotic theory. The data application offers new insights into heterogeneous tfMRI dynamics across multiple subjects by identifying both shared and individual-specific paths.

The framework can be extended in several directions. First, it naturally accommodates sub-Gaussian (or strongly bounded) innovations (e.g., Section 2.3.4 in Van de Geer et al.; 2014), but extending it to heavier-tailed distributions, such as sub-exponential innovations, remains an open challenge. While additional mixing conditions may ensure consistent estimation for individual subjects (Wong et al.; 2020), debiasing time series models under such distributions is nontrivial, even in single-subject settings. Second, Crawford et al. (2024) considered the decomposition of partially shared paths via clustering; however, establishing the consistency of these estimators within a communication-efficient data integration framework is still unresolved, representing an important direction for future work. Finally, the framework can be extended to higher-lag VAR models. Imposing simple sparsity across lags or using a standard group Lasso uniformly may fail to capture the natural decay of higher-lag effects. Approaches such as overlapping group sparsity with increasing penalties (Nicholson et al.; 2020) offer promising alternatives, yet debiasing these structurally penalized models remains an open problem, providing another avenue for methodological development.

Acknowledgements

Vladas Pipiras’s research was partially supported by the grants NSF DMS 1712966, DMS 2113662, and DMS 2134107.

Data Availability Statement

Human Connectome Project (HCP) data used in the data application of Section 6 is publicly accessible. The dataset can be downloaded at <https://humanconnectome.org/>. The R code used in the simulations of Sections 4 and in the data analysis of Section 5 is available on GitHub at <https://github.com/yk748/multiVARSE>.

A Proofs of Lemmas

In this section, we provide several technical Lemmas and their proofs.

Lemma A.1. *With the conditions (a) – (c) in Assumption 2.1, the following holds.*

$$\|\hat{\beta}_i^{(k)} - \beta_i^{(k)}\|_1 = \|\hat{\Phi}_{i:}^{(k)'} - \Phi_{i:}^{(k)'}\|_1 = \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_{\epsilon}^{(k)}) \kappa^2(\Phi^{(k)}) \|\Phi^{(k)}\|^2 s_{0,k} \sqrt{\frac{\log d}{N_k}} \right). \quad (27)$$

Proof. From the deterministic function of the VAR model (e.g., Proposition 4.3 in Basu and Michailidis; 2015), we have

$$\begin{aligned} \mathbb{Q}(\beta_k, \Sigma_{\epsilon}^{(k)}) &= c_0 \left(\Lambda_{\max}(\Sigma_{\epsilon}^{(k)}) + \frac{\Lambda_{\max}(\Sigma_{\epsilon}^{(k)})}{\mu_{\min}(\Phi^{(k)})} + \frac{\Lambda_{\max}(\Sigma_{\epsilon}^{(k)}) \mu_{\max}(\Phi^{(k)})}{\mu_{\min}(\Phi^{(k)})} \right) \\ &= c_0 \sigma_{k,\max}^2 \left(1 + \|(\Phi^{(k)})^{-1}\| (1 + \|\Phi^{(k)}\|) \right) \\ &\leq c_0 \sigma_{k,\max}^2 (1 + 2\kappa^2(\Phi^{(k)})) \\ &= \mathcal{O}_{\mathbb{P}}(\sigma_{k,\max}^2 \kappa^2(\Phi^{(k)})). \end{aligned} \quad (28)$$

For (28), by using Proposition 2.2 in Basu et al. (2024) with $\lambda^{(k)} = \mathcal{O}_{\mathbb{P}}(\sigma_{k,\max}^2 \kappa^2(\Phi^{(k)}) \sqrt{\log d / N_k})$, it completes the proof. $\square$

Lemma A.2. *Consider that*

$$\Delta_i^{(k)} := (I - \hat{\Theta}^{(k)} \hat{\Sigma}^{(k)}) (\hat{\beta}_i^{(k)} - \beta_i^{(k)}).$$

Then

$$\|\Delta_i^{(k)}\|_{\infty} \leq \mathcal{O}_{\mathbb{P}} \left(\kappa^2(\Sigma_{\epsilon}^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 s_{0,k} \frac{\log d}{N_k} \right). \quad (29)$$

Proof. Recall the debiased equation (8). For each i and k , one has

$$\tilde{\beta}_i^{(k)} - \beta_i^{(k)} = \frac{1}{N_k} \hat{\Theta}^{(k)} \mathcal{X}^{(k)'} \underbrace{(\mathcal{Y}_i^{(k)} - \mathcal{X}^{(k)} \beta_i^{(k)})}_{=E_i^{(k)}} + \underbrace{(I - \hat{\Theta}^{(k)} \hat{\Sigma}^{(k)}) (\hat{\beta}_i^{(k)} - \beta_i^{(k)})}_{=\Delta_i^{(k)}}. \quad (30)$$

Note that

$$\|\Delta_i^{(k)}\|_\infty \leq \|I - \hat{\Theta}^{(k)} \hat{\Sigma}^{(k)}\|_\infty \|\hat{\beta}_i^{(k)} - \beta_i^{(k)}\|_1 = \max_j |e_j - \hat{\Theta}_{j:}^{(k)} \hat{\Sigma}^{(k)}| \|\hat{\beta}_i^{(k)} - \beta_i^{(k)}\|_1.$$

From KKT condition for nodewise regression in (9),

$$-\frac{1}{N_k} \mathcal{X}_{-j}^{(k)'} (\mathcal{X}_j^{(k)} - \mathcal{X}_{-j}^{(k)} \hat{\gamma}_j^{(k)}) + \lambda_j^{(k)} \hat{z}_j^{(k)},$$

where $\|\hat{z}_j^{(k)}\|_1 \leq 1$. This implies

$$-\frac{1}{N_k} \hat{\gamma}_j^{(k)'} \mathcal{X}_{-j}^{(k)'} \mathcal{X}^{(k)} \hat{\Theta}_{j:}^{(k)'} (\hat{\tau}_j^{(k)})^2 + \lambda_j^{(k)} \|\hat{\gamma}_j^{(k)}\|_1 = 0.$$

This gives us

$$\begin{aligned} (\hat{\tau}_j^{(k)})^2 &= \frac{1}{N_k} \|\mathcal{X}_j^{(k)} - \mathcal{X}_{-j}^{(k)} \hat{\gamma}_j^{(k)}\|_2^2 + \lambda_j^{(k)} \|\hat{\tau}_j^{(k)}\|_1 \\ &= \frac{1}{N_k} \mathcal{X}_j^{(k)'} \mathcal{X}^{(k)} \hat{\Theta}_{j:}^{(k)'} (\hat{\tau}_j^{(k)})^2 - \frac{1}{N_k} \hat{\gamma}_j^{(k)'} \mathcal{X}_{-j}^{(k)'} \mathcal{X}^{(k)} \hat{\Theta}_{j:}^{(k)'} (\hat{\tau}_j^{(k)})^2 + \lambda_j^{(k)} \|\hat{\gamma}_j^{(k)}\|_1 \\ &= \frac{1}{N_k} \mathcal{X}_j^{(k)'} \mathcal{X}^{(k)} \hat{\Theta}_{j:}^{(k)'} (\hat{\tau}_j^{(k)})^2. \end{aligned} \tag{31}$$

Hence, $\frac{1}{N_k} \mathcal{X}_j^{(k)'} \mathcal{X}^{(k)} \hat{\Theta}_{j:}^{(k)} = 1$. This implies

$$|e_j - \hat{\Theta}_{j:}^{(k)} \hat{\Sigma}^{(k)}| \leq \frac{\lambda_j^{(k)}}{(\hat{\tau}_j^{(k)})^2} \leq \frac{\sigma_{k,\max}^2 \kappa^2(\Phi^{(k)})}{(\hat{\tau}_j^{(k)})^2} \sqrt{\frac{\log d}{N_k}}$$

with a suitable choice of $\lambda_j^{(k)} = O_{\mathbb{P}}(\sigma_{k,\max}^2 \kappa^2(\Phi^{(k)}) \sqrt{\log d / N_k})$. To complete the proof, we reconstruct Lemma 5.3 in Van de Geer et al. (2014): the population error variance $(\tau_j^{(k)})^2 = \mathbb{E}[(\mathcal{X}_{1,j} - \sum_{\ell \neq j} \gamma_{j,\ell} X_{1,\ell})^2]$ satisfies $(\tau_j^{(k)})^2 = \mathbb{E}\epsilon_{1,j}^2 = \frac{1}{\Theta_{jj}^{(k)}} \geq \sigma_{k,\min}^2 > 0$ and $(\tau_j^{(k)})^2 \leq \mathcal{M}(f_X) \leq \mathcal{O}(\sigma_{k,\max}^2 \|(\Phi^{(k)})^{-1}\|^2) < \infty$. This successfully replaces Assumption (A2) in Van de Geer et al. (2014). From (31),

$$\begin{aligned} \frac{1}{N_k} \|\mathcal{X}_j^{(k)} - \mathcal{X}_{-j}^{(k)} \hat{\gamma}_j^{(k)}\|_2^2 &= \frac{1}{N_k} \|\mathcal{X}_j^{(k)} - \mathcal{X}_{-j}^{(k)} \gamma_j^{(k)}\|_2^2 + \frac{1}{N_k} \|\mathcal{X}_{-j}^{(k)} (\hat{\gamma}_j^{(k)} - \gamma_j^{(k)})\|_2^2 \\ &\quad + \frac{2}{N_k} (\mathcal{X}_j^{(k)} - \mathcal{X}_{-j}^{(k)} \gamma_j^{(k)})' \mathcal{X}_{-j}^{(k)} (\hat{\gamma}_j^{(k)} - \gamma_j^{(k)}). \end{aligned} \tag{32}$$

The first term in (32) is $(\tau_j^{(k)})^2$. The second term in (32) is, by Proposition 3.3 in Basu and Michailidis (2015), bounded above by

$$\mathcal{O}_{\mathbb{P}} \left(\frac{s_{j,k}(\lambda_j^{(k)}(\mathcal{M}(f_X) + \mathcal{M}(f_\epsilon))^2)}{\alpha_{\text{RE}}} \right) = \mathcal{O}_{\mathbb{P}} \left(\kappa^2(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 \frac{s_{j,k} \log d}{N_k} \right) = \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{N_k}} \right),$$

where the last equality holds by Assumption 2.1. Note that it is equivalent to assume explicitly that $s_{\max,k} = \mathcal{O}(N_k / \log d)$ in Van de Geer et al. (2014). The third term in (32) is, by Propositions

3.2 and 3.3 in Basu and Michailidis (2015), bounded above by

$$\begin{aligned}
& 2 \left\| \frac{1}{N_k} E_{-j}^{(k)'} \mathcal{X}_{-j}^{(k)} \right\|_{\infty} \|\hat{\gamma}_j^{(k)} - \gamma_j^{(k)}\|_1 \\
& \leq \mathcal{O}_{\mathbb{P}} \left(\lambda_j^{(k)} (\mathcal{M}(f_X) + \mathcal{M}(f_{\epsilon})) \right) \mathcal{O}_{\mathbb{P}} \left(\frac{s_{j,k} (\mathcal{M}(f_X) + \mathcal{M}(f_{\epsilon}))}{\alpha_{\text{RE}}} \sqrt{\frac{\log d}{N_k}} \right) \\
& \leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_{\epsilon}^{(k)}) \kappa^2(\Phi^{(k)}) \sqrt{\frac{\log d}{N_k}} \right) \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_{\epsilon}^{(k)}) \kappa^2(\Phi^{(k)}) \|\Phi^{(k)}\|^2 s_{j,k} \sqrt{\frac{\log d}{N_k}} \right) \\
& = \mathcal{O}_{\mathbb{P}} \left(\kappa^2(\Sigma_{\epsilon}^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 \frac{s_{j,k} \log d}{N_k} \right) \\
& = \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{N_k}} \right).
\end{aligned}$$

In addition,

$$\lambda_j^{(k)} \|\hat{\gamma}_j^{(k)}\|_1 \leq \lambda_j^{(k)} \|\gamma_j^{(k)}\|_1 + \lambda_j^{(k)} \|\hat{\gamma}_j^{(k)} - \gamma_j^{(k)}\|_1 = \lambda_j^{(k)} \mathcal{O}_{\mathbb{P}}(\sqrt{s_{j,k}}) + \lambda_j^{(k)} \mathcal{O}_{\mathbb{P}}(\lambda_j^{(k)} s_{j,k}) = \mathcal{O}_{\mathbb{P}}(1). \quad (33)$$

Combining (32), (33), and $(\tau_j^{(k)})^2 \geq \sigma_{k,\min}^2$ yields

$$\max_j \frac{1}{(\hat{\tau}_j^{(k)})^2} = \mathcal{O}_{\mathbb{P}}(1/\sigma_{k,\min}^2).$$

This implies

$$\|I - \hat{\Theta}^{(k)} \hat{\Sigma}^{(k)}\|_{\infty} \leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_{\epsilon}^{(k)}) \kappa^2(\Phi^{(k)}) \sqrt{\frac{\log d}{N_k}} \right)$$

so that combining with (27) in Lemma A.1 yields the desired result (29). $\square$

Lemma A.3. *The bound on $\tilde{\beta}_i^{(k)} - \beta_i^{(k)}$ in (30) is*

$$\|\tilde{\beta}_i^{(k)} - \beta_i^{(k)}\|_{\infty} \leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma^{(k)}) \sqrt{\frac{\log d}{N_k}} \right). \quad (34)$$

Proof. Note that

$$\|\tilde{\beta}_i^{(k)} - \beta_i^{(k)}\|_{\infty} \leq \left\| \frac{1}{N_k} \Theta^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} \right\|_{\infty} + \|(\hat{\Theta}^{(k)} - \Theta^{(k)}) \frac{1}{N_k} \mathcal{X}^{(k)'} E_i^{(k)}\|_{\infty} + \|\Delta_i^{(k)}\|_{\infty}. \quad (35)$$

Note that the first term is

$$\max_j \frac{1}{\sqrt{N_k}} \left| \frac{1}{\sqrt{N_k}} e_j \Theta^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} \right|.$$

In the proof of Proposition 2.3 in Basu et al. (2024), they showed that under Assumption 2.1 holds,

$$\frac{1}{\sqrt{N_k}} e_j \Theta^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} = \frac{1}{\sqrt{N_k}} \sum_{t=p+1}^{T_k} (\Theta_{j:}^{(k)'} X_{t-1}) \epsilon_{i,t} \xrightarrow{d} \mathcal{N}(0, \sigma_{k,i}^2 \Theta_{jj}^{(k)}). \quad (36)$$

Therefore, by using Borell–TIS inequality with $u = \log d$ (e.g., Theorem 2.1.1 in Adler and Taylor; 2007) and $\Theta_{jj}^{(k)} \leq 1/\sigma_{k,\min}^2$, the first term in (35) is bounded above by

$$\mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma^{(k)}) \sqrt{\frac{\log d}{N_k}} \right). \quad (37)$$

Note that the second term is

$$\begin{aligned}
\left\| (\hat{\Theta}^{(k)} - \Theta^{(k)}) \frac{1}{N_k} \mathcal{X}^{(k)'} E_i^{(k)} \right\|_\infty &= \max_j \left| (\hat{\Theta}_{j\cdot}^{(k)} - \Theta_{j\cdot}^{(k)}) \frac{1}{N_k} \mathcal{X}^{(k)'} E_i^{(k)} \right| \\
&\leq \max_j \|\hat{\Theta}_{j\cdot}^{(k)} - \Theta_{j\cdot}^{(k)}\|_1 \left\| \frac{1}{N_k} \mathcal{X}^{(k)'} E_i^{(k)} \right\|_\infty \\
&\leq \max_j \|\hat{\Theta}_{j\cdot}^{(k)} - \Theta_{j\cdot}^{(k)}\|_1 \mathcal{O}_{\mathbb{P}} \left(\sigma_{k,\max}^2 \kappa^2(\Phi^{(k)}) \sqrt{\frac{\log d}{N_k}} \right) \\
&\leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 s_{\max,k} \frac{\log d}{N_k} \right) \tag{38}
\end{aligned}$$

where the second last inequality holds by (28) in Lemma A.1. To show the last inequality, note that from $\hat{\Theta}^{(k)} = (\hat{\gamma}^{(k)})^{-2} \hat{\Gamma}^{(k)}$, so $\hat{\Theta}_{j\cdot}^{(k)} = \hat{\gamma}_j^{(k)} / (\hat{\tau}_j^{(k)})^2$. This implies

$$\begin{aligned}
&\|\hat{\Theta}_{j\cdot}^{(k)} - \Theta_{j\cdot}^{(k)}\|_1 \\
&= \|\hat{\gamma}_j^{(k)} / (\hat{\tau}_j^{(k)})^2 - \gamma_j^{(k)} / (\tau_j^{(k)})^2\|_1 \\
&\leq \|\hat{\gamma}_j^{(k)} - \gamma_j^{(k)}\|_1 / (\hat{\tau}_j^{(k)})^2 + \|\gamma_j^{(k)}\|_1 (1 / (\hat{\tau}_j^{(k)})^2 - 1 / (\tau_j^{(k)})^2) \\
&\leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_\epsilon^{(k)}) \kappa^2(\Phi^{(k)}) \|\Phi^{(k)}\|^2 s_{j,k} \sqrt{\frac{\log d}{N_k}} \right) \mathcal{O}_{\mathbb{P}}(1 / \sigma_{k,\min}^2) + \mathcal{O}_{\mathbb{P}}(\sqrt{s_{j,k}}) \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{s_{j,k} \log d}{N_k}} \right) \\
&= \mathcal{O}_{\mathbb{P}} \left(\frac{\kappa(\Sigma_\epsilon^{(k)}) \kappa^2(\Phi^{(k)})}{\sigma_{k,\min}^2} \|\Phi^{(k)}\|^2 s_{j,k} \sqrt{\frac{\log d}{N_k}} \right).
\end{aligned}$$

Combining (29) in Lemma A.2, (37), and (38) yields the upper bound of (35),

$$\|\tilde{\beta}_i^{(k)} - \beta_i^{(k)}\|_\infty \leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma^{(k)}) \sqrt{\frac{\log d}{N_k}} + \kappa(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 \max\{s_{\max,k}, s_{0,k}\} \frac{\log d}{N_k} \right).$$

For a sufficiently large N_k , by Assumptions 2.1, we have the desired result (34). $\square$

Lemma A.4. *For a sufficiently large N_k , by Assumptions 1 and 2, we have for all $k = 0, 1, \dots, K$,*

$$\|\tilde{\alpha}^{(k)} - \alpha^{(k)}\|_\infty \leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_\epsilon^{(k)}) \sqrt{\frac{\log d^2}{N_k}} \right).$$

Proof. Recall Result 5 in Maity et al. (2022): Under Assumption 2.2, the objective function (11) has a unique minimizer $(\alpha_i^{(0)})_j = \mu_{ij}$. Note that if $\kappa(\Sigma^{(k)}) \sqrt{\log d / N_k} \leq \frac{1}{4}(\eta_j - 2\delta) \wedge (\eta_j - \delta_2/2)$ for all k , there exists $\hat{\delta}, \hat{\delta}_2$ such that

$$2\delta < 2\hat{\delta} < \eta_j < \hat{\delta}_2/2 \leq \delta_2/2,$$

and Assumption 2.2 holds with $\hat{\delta}$ and $\hat{\delta}_2$. Hence, by Result 5 in Maity et al. (2022), we have $(\tilde{\alpha}_i^{(0)})_j = \frac{1}{|I_j|} \sum_{k \in I_j} (\tilde{\beta}_i^{(k)})_j$. Let $w_j^{(k)} = \frac{1}{|I_j|} 1_{\{I_j\}}(k)$, the indicator function of the events $k \in I_j$.

Then

$$\begin{aligned}
(\tilde{\alpha}_i^{(0)})_j &= \sum_{k=1}^K w_j^{(k)} (\tilde{\beta}^{(k)})_j \\
&= \sum_{k=1}^K w_j^{(k)} \left((\beta^{(k)})_j - \frac{1}{N_k} \hat{\Theta}_{j\cdot}^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} + (\Delta_i^{(k)})_j \right) \\
&= (\alpha_i^{(0)})_j - \sum_{k=1}^K \frac{w_j^{(k)}}{N_k} \hat{\Theta}_{j\cdot}^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} + \sum_{k=1}^K w_j^{(k)} (\Delta_i^{(k)})_j,
\end{aligned}$$

Then

$$\begin{aligned}
& |(\tilde{\alpha}^{(0)})_j - (\alpha^{(0)})_j| \tag{39} \\
& \leq \left| \sum_{k=1}^K \frac{w_j^{(k)}}{N_k} \hat{\Theta}_{j\cdot}^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} \right| + \left| \sum_{k=1}^K w_j^{(k)} (\Delta_i^{(k)})_j \right| \\
& \leq \left| \sum_{k=1}^K \frac{w_j^{(k)}}{N_k} \Theta_{j\cdot}^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} \right| + \left| \sum_{k=1}^K \frac{w_j^{(k)}}{N_k} (\hat{\Theta}_{j\cdot}^{(k)} - \Theta_{j\cdot}^{(k)}) \mathcal{X}^{(k)'} E_i^{(k)} \right| + \left| \sum_{k=1}^K w_j^{(k)} (\Delta_i^{(k)})_j \right| \\
& \leq \left| \sum_{k=1}^K \frac{w_j^{(k)}}{N_k} \Theta_{j\cdot}^{(k)} \mathcal{X}^{(k)'} E_i^{(k)} \right| + \sum_{k=1}^K w_j^{(k)} \left(\|\hat{\Theta}_{j\cdot}^{(k)} - \Theta_{j\cdot}^{(k)}\|_1 \frac{1}{N_k} \mathcal{X}^{(k)'} E_i^{(k)}\|_\infty + |(\Delta_i^{(k)})_j| \right). \tag{40}
\end{aligned}$$

The second term in (40) is, from (29) and (38) in Lemmas A.2 and 34, bounded above by

$$\mathcal{O}_{\mathbb{P}} \left(\sum_{k=1}^K w_j^{(k)} \left(\kappa(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 \max\{s_{\max}, s_0\} \frac{\log d}{N_{\min}} \right) \right)$$

for all j and k , which converges to 0 fast by Assumption 2.1. Note that the bound does not depend on i . Hence, we take the union bound across $j = 1, \dots, d^2$. For the first term in (40), one can show that for each k , the sum of $(\Theta_{j\cdot}^{(k)})' X_{t-1} \epsilon_{i,t} = Z_t^{(k)}$, $t = p+1, \dots, T_k$ converges to

$$\frac{1}{\sqrt{N_k}} \sum_{t=1}^{N_k} Z_t^{(k)} := S_{N_k}^{(k)} \xrightarrow{d} \mathcal{N}(0, \sigma_{k,i}^2 \Theta_{jj}^{(k)}).$$

Therefore, by using generalized Hoeffding inequality (e.g., Theorem 2.6.3 in Vershynin; 2018), with $a = (w_j^{(1)}/\sqrt{N_1}, \dots, w_j^{(K)}/\sqrt{N_K})$,

$$\mathbb{P} \left(\left| \sum_{k=1}^K \frac{1}{\sqrt{N_k}} S_{N_k}^{(k)} \right| \geq \eta \right) \leq 2 \exp \left(- \frac{c_1 \eta^2}{-c_2 \max_k (\sigma_{k,i}^2 \Theta_{jj}^{(k)}) \|a\|_2^2} \right).$$

for some constants c_1 and c_2 . By taking $\eta = \mathcal{O}(\max_k \kappa(\Sigma_\epsilon^{(k)}) \sqrt{\log d^2} \|a\|_2)$ and using $\Theta_{jj}^{(k)} \leq 1/\sigma_{k,\min}^2$, the first term in (40) is bounded above by

$$\mathcal{O}_{\mathbb{P}} \left(\max_k \kappa(\Sigma_\epsilon^{(k)}) \sqrt{\frac{\log d^2}{|I_j|^2} \sum_{k \in I_j} \frac{1}{N_k}} \right) \leq \mathcal{O}_{\mathbb{P}} \left(\max_k \kappa(\Sigma_\epsilon^{(k)}) \sqrt{\frac{\log d^2}{K N_{\min}}} \right)$$

and by Assumption 2.1, the bound on the first term dominates the bound of $|(\hat{\alpha}^{(0)})_j - (\alpha^{(0)})_j|$. For the second inequality, by using the triangular inequality,

$$\begin{aligned} \|\tilde{\alpha}^{(k)} - \alpha^{(k)}\|_\infty &\leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma^{(k)}) \sqrt{\frac{\log d^2}{N_k}} + \kappa(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \max\{s_{\max,k}, s_{0,k}\} \frac{\log d^2}{N_k} \right. \\ &\quad \left. + \max_k \kappa(\Sigma_\epsilon^{(k)}) \sqrt{\frac{\log d^2}{KN_{\min}}} + \max_k \kappa(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \max\{s_{\max}, s_0\} \frac{\log d^2}{N_{\min}} \right). \end{aligned} \quad (41)$$

Note that the first term in (41) dominates for a sufficiently large N_k with the condition (a) in Assumption 2.2 so that

$$\|\tilde{\alpha}^{(k)} - \alpha^{(k)}\|_\infty \leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_\epsilon^{(k)}) \sqrt{\frac{\log d^2}{N_k}} \right).$$

This corresponds to the Lemma 6 in Maity et al. (2022) by replacing σ with $\max_k \kappa(\Sigma^{(k)})$. $\square$

Lemma A.5. For each k , $\hat{\sigma}_{k,i}^2 \xrightarrow{P} \sigma_{k,i}^2$, $i = 1, \dots, d$.

Proof. Note that for fixed i ,

$$\begin{aligned} \hat{\sigma}_{k,i}^2 &= \frac{1}{N_k} \sum_{t=1}^{N_k} \left((\mathcal{Y}_i^{(k)})_t - (\mathcal{X}^{(k)})_t \hat{\beta}_i^{(k)} \right)^2 = \frac{1}{N_k} \|\mathcal{Y}_i^{(k)} - \mathcal{X}^{(k)} \hat{\beta}_i^{(k)}\|_2^2 \\ &= \frac{1}{N_k} \|\mathcal{X}^{(k)} \beta_i^{(k)} - \mathcal{X}^{(k)} \hat{\beta}_i^{(k)} + E_i^{(k)}\|_2^2 \\ &= \frac{1}{N_k} \|\mathcal{X}^{(k)} (\hat{\beta}_i^{(k)} - \beta_i^{(k)})\|_2^2 - \frac{2}{N_k} \langle \mathcal{X}^{(k)} (\hat{\beta}_i^{(k)} - \beta_i^{(k)}), E_i^{(k)} \rangle + \frac{1}{N_k} \|E_i^{(k)}\|_2^2. \end{aligned} \quad (42)$$

Note that by the law of large numbers, the last term in (42) converges to $\sigma_{k,i}^2$. For the first term in (42), by using Proposition 3.3 in Basu and Michailidis (2015), it is bounded above by

$$\mathcal{O}_{\mathbb{P}} \left(\frac{s_{0,k} (\lambda^{(k)} (\mathcal{M}(f_X) + \mathcal{M}(f_\epsilon)))^2}{\alpha_{\text{RE}}} \right) = \mathcal{O}_{\mathbb{P}} \left(\kappa^2(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 \frac{s_{0,k} \log d}{N_k} \right) = \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{N_k}} \right).$$

Note that the second term in (42), from (28) and (27) in Lemma A.1, is bounded above by

$$\begin{aligned} 2 \left\| \frac{\mathcal{X}^{(k)'} E_i^{(k)}}{N_k} \right\|_\infty \|\hat{\beta}_i^{(k)} - \beta_i^{(k)}\|_1 &\leq \mathcal{O}_{\mathbb{P}} \left(\sigma_{k,\max}^2 \kappa^2(\Phi^{(k)}) \sqrt{\frac{\log d}{N_k}} \right) \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_\epsilon^{(k)}) \kappa^2(\Phi^{(k)}) \|\Phi^{(k)}\|^2 s_{0,k} \sqrt{\frac{\log d}{N_k}} \right) \\ &\leq \mathcal{O}_{\mathbb{P}} \left(\kappa^2(\Sigma_\epsilon^{(k)}) \kappa^4(\Phi^{(k)}) \|\Phi^{(k)}\|^2 \frac{s_{0,k} \log d}{N_k} \right) = \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{N_k}} \right) \\ &= \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{N_k}} \right). \end{aligned}$$

Hence, we have the desired result. $\square$

Lemma A.6. $\hat{\sigma}_{i,k} \sqrt{\hat{\Omega}_{jj}^{(k)}} \xrightarrow{P} \sigma_{i,k} \sqrt{\Theta_{jj}^{(k)}}$ for all i, j , and k .

Proof. Define $\hat{\Omega}^{(k)} = \hat{\Theta}^{(k)} \hat{\Sigma}^{(k)} \hat{\Theta}^{(k)'}.$ Note that

$$\|\hat{\Omega}^{(k)} - \Theta^{(k)}\|_{\infty} = \|(\hat{\Theta}^{(k)} \hat{\Sigma}^{(k)} - I) \hat{\Theta}^{(k)'}\|_{\infty} + \|\hat{\Theta}^{(k)} - \Theta^{(k)}\|_{\infty}. \quad (43)$$

The first term in (43) is bounded above by

$$\begin{aligned} \max_j |e_j - \hat{\Theta}_{j:}^{(k)} \Sigma^{(k)}| \|\hat{\Theta}_{j:}^{(k)}\|_1 &\leq \max_j |e_j - \hat{\Theta}_{j:}^{(k)} \Sigma^{(k)}| \|\hat{\gamma}_j^{(k)} / (\hat{\tau}_j^{(k)})^2\|_1 \\ &\leq \mathcal{O}_{\mathbb{P}} \left(\kappa(\Sigma_{\epsilon}^{(k)}) \kappa^2(\Phi^{(k)}) \sqrt{\frac{\log d}{N_k}} \right) \mathcal{O}(\max_j \sqrt{s_{j,k}}) \mathcal{O}_{\mathbb{P}}(1/\sigma_{k,\min}^2) \\ &\leq \mathcal{O}_{\mathbb{P}} \left(\frac{\kappa(\Sigma_{\epsilon}^{(k)}) \kappa^2(\Phi^{(k)})}{\sigma_{k,\min}^2} \sqrt{\frac{s_{\max,k} \log d}{N_k}} \right). \end{aligned}$$

The second term in (43) is bounded above by

$$\begin{aligned} \max_j \|\hat{\Theta}_{j:}^{(k)} - \Theta_{j:}^{(k)}\| &\leq \max_j \left(\|\hat{\gamma}_j^{(k)} - \gamma_j^{(k)}\|_2 / (\hat{\tau}_j^{(k)})^2 + \|\gamma_j^{(k)}\|_2 (1/(\hat{\tau}_j^{(k)})^2 - 1/(\tau_j^{(k)})^2) \right) \\ &\leq \max_j \left\{ \mathcal{O}_{\mathbb{P}} \left(\frac{\sqrt{s_{j,k}} \lambda^{(k)}(\mathcal{M}(f_X) + \mathcal{M}(f_{\epsilon}))}{\alpha_{\text{RE}}} \right) \mathcal{O}_{\mathbb{P}}(1/\sigma_{k,\min}^2) + \mathcal{O}_{\mathbb{P}} \left(\sqrt{\frac{s_{j,k} \log d}{N_k}} \right) \right\} \\ &\leq \mathcal{O}_{\mathbb{P}} \left(\frac{\kappa(\Sigma_{\epsilon}^{(k)}) \kappa^2(\Phi^{(k)}) \|\Phi^{(k)}\|^2}{\sigma_{k,\min}^2} \sqrt{\frac{s_{\max,k} \log d}{N_k}} \right) \end{aligned}$$

Hence, $\|\hat{\Omega}^{(k)} - \Theta^{(k)}\|_{\infty} = \mathcal{O}_{\mathbb{P}}(1).$ By combining with Lemma A.5, it completes the proof. $\square$

B Additional Figure

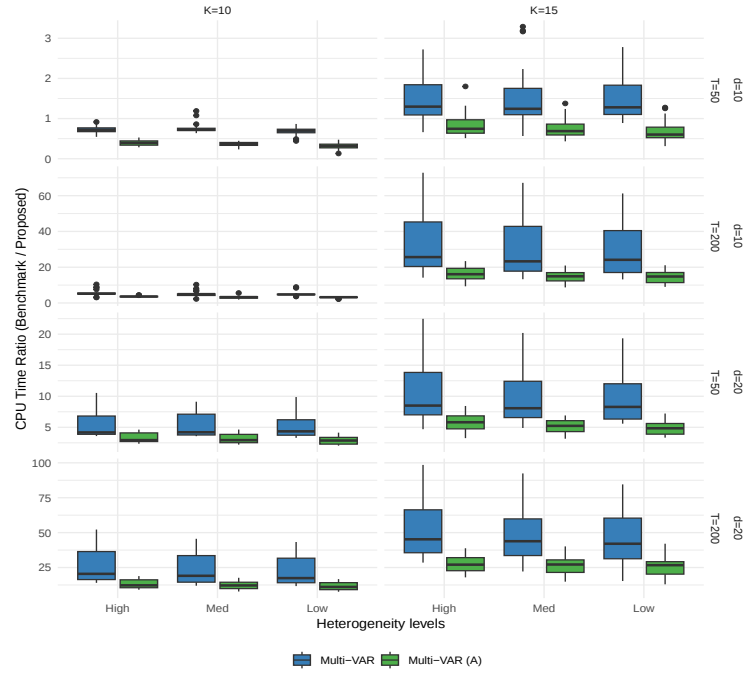


Figure 4: Boxplots of the CPU time ratio (Benchmark / Proposed) under different combinations of d and average T (combinations indicated on the right tabs), K (each column), and heterogeneity levels (each axis).

References

- Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179.
- Adamek, R., Smeekes, S., and Wilms, I. (2023). Lasso inference for high-dimensional time series. *Journal of Econometrics*, 235(2):1114–1143.
- Adler, R. J. and Taylor, J. E. (2007). Gaussian inequalities. *Random Fields and Geometry*, pages 49–64.
- Alquier, P., Bertin, K., Doukhan, P., and Garnier, R. (2020). High-dimensional VAR with low-rank transition. *Statistics and Computing*, 30(4):1139–1153.
- Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain’s default network. *Neuron*, 65(4):550–562.
- Asiaee, A., Oymak, S., Coombes, K. R., and Banerjee, A. (2019). Data enrichment: Multi-task learning in high dimension with theoretical guarantees. In *Adaptive and Multitask Learning Workshop at the ICML. IMLS, Long Beach, CA*.
- Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169–189.
- Basu, S., Das, S., Michailidis, G., and Purnanandam, A. (2024). A high-dimensional approach to measure connectivity in the financial sector. *The Annals of Applied Statistics*, 18(2):922–945.
- Basu, S., Li, X., and Michailidis, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, 67(5):1207–1222.
- Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.
- Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.
- Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., Weber, J., and Ochsner, K. N. (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cerebral Cortex*, 24(11):2981–2990.

- Chen, G., Glen, D. R., Saad, Z. S., Hamilton, J. P., Thomason, M. E., Gotlib, I. H., and Cox, R. W. (2011). Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis. *Computers in Biology and Medicine*, 41(12):1142–1155.
- Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 30, pages 178–191. Cambridge University Press.
- Crawford, C. M., Park, J. J., Chow, S.-M., Ernst, A. F., Pipiras, V., and Fisher, Z. F. (2024). Penalized subgrouping of heterogeneous time series. *arXiv preprint arXiv:2409.03085*.
- Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., and Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7(2):189–195.
- Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.
- Etkin, A., Büchel, C., and Gross, J. J. (2015). The neural bases of emotion regulation. *Nature Reviews Neuroscience*, 16(11):693–700.
- Fan, J., Liu, H., Wang, W., and Zhu, Z. (2018). Heterogeneity adjustment with applications to graphical model inference. *Electronic Journal of Statistics*, 12(2):3908.
- Fisher, Z., Kim, Y., and Pipiras, V. (2021). Multivar: Penalized estimation of multiple-subject vector autoregressive (multi-var) models. R package version 1.1.0. Available at <https://CRAN.R-project.org/package=multivar>.
- Fisher, Z. F., Kim, Y., Fredrickson, B. L., and Pipiras, V. (2022). Penalized estimation and forecasting of multiple subject intensive longitudinal data. *Psychometrika*, 87(2):403–431.
- Fisher, Z. F., Kim, Y., Pipiras, V., Crawford, C., Petrie, D. J., Hunter, M. D., and Geier, C. F. (2024). Structured estimation of heterogeneous time series. *Multivariate Behavioral Research*, 59(6):1270–1289.
- Gates, K. M. and Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, 63(1):310–319.
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124.
- Gross, S. M. and Tibshirani, R. (2016). Data shared Lasso: A novel tool to discover uplift. *Computational statistics & Data Analysis*, 101:226–235.

- Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150.
- Haslbeck, J. M., Epskamp, S., and Waldorp, L. J. (2025). Testing for group differences in multilevel vector autoregressive models. *Behavior Research Methods*, 57(3):100.
- Huber, P. and Ronchetti, E. (2011). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley.
- Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.
- Jongerling, J., Laurenceau, J.-P., and Hamaker, E. L. (2015). A multilevel AR (1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, 50(3):334–349.
- Kim, Y., Fisher, Z. F., and Pipiras, V. (2024). Group integrative dynamic factor models with application to multiple subject brain connectivity. *Biometrical Journal*, 66(8):e202300370.
- Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.
- Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30.
- Lyu, X., Kang, J., and Li, L. (2024). High-dimensional multisubject time series transition matrix inference with application to brain connectivity analysis. *Biometrics*, 80(2):ujae021.
- Maity, S., Sun, Y., and Banerjee, M. (2022). Meta-analysis of heterogeneous data: Integrative sparse regression in high-dimensions. *Journal of Machine Learning Research*, 23(198):1–50.
- Manomaisaowapak, P. and Songsiri, J. (2022). Joint learning of multiple Granger causal networks via non-convex regularizations: Inference of group-level brain connectivity. *Neural Networks*, 149:157–171.
- Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in Cognitive Sciences*, 15(10):483–506.
- Nicholson, W. B., Wilms, I., Bien, J., and Matteson, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21(166):1–52.
- Ochsner, K. N. and Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, 9(5):242–249.
- O’Connell, M. J. and Lock, E. F. (2019). Linked matrix factorization. *Biometrics*, 75(2):582–592.

- Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the Lasso. *Biometrika*, 104(1):83–96.
- Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral Cortex*, 28(9):3095–3114.
- Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., and Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9):2349–2356.
- Seth, A. K., Barrett, A. B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297.
- Shojaie, A. (2021). Differential network analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(2):e1508.
- Shojaie, A. and Fox, E. B. (2022). Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319.
- Skipnikov, A. and Michailidis, G. (2019). Joint estimation of multiple network Granger causal models. *Econometrics and Statistics*, 10:120–133.
- Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*.
- Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79.
- Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press.
- Wilms, I., Barbaglia, L., and Croux, C. (2018). Multiclass vector auto-regressive models for multistore sales data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(2):435–452.
- Wong, K. C., Li, Z., Tewari, A., et al. (2020). Lasso guarantees for β -mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142.
- Wright, A. G., Beltz, A. M., Gates, K. M., Molenaar, P. C., and Simms, L. J. (2015). Examining the dynamic structure of daily internalizing and externalizing behavior at multiple levels of analysis. *Frontiers in Psychology*, 6:1914.

- Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*.
- Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.