
Transfer Learning for High-dimensional Reduced Rank Time Series Models

Mingliang Ma
University of Florida

Abolfazl Safikhani
George Mason University

Abstract

The objective of transfer learning is to enhance estimation and inference in a target data by leveraging knowledge gained from additional sources. Recent studies have explored transfer learning for independent observations in complex, high-dimensional models assuming sparsity, yet research on time series models remains limited. Our focus is on transfer learning for sequences of observations with temporal dependencies and a more intricate model parameter structure. Specifically, we investigate the vector autoregressive model (VAR), a widely recognized model for time series data, where the transition matrix can be deconstructed into a combination of a sparse matrix and a low-rank one. We propose a new transfer learning algorithm tailored for estimating high-dimensional VAR models characterized by low-rank and sparse structures. Additionally, we present a novel approach for selecting informative observations from auxiliary datasets. Theoretical guarantees are established, encompassing model parameter consistency, informative set selection, and the asymptotic distribution of estimators under mild conditions. The latter facilitates the construction of entry-wise confidence intervals for model parameters. Finally, we demonstrate the empirical efficacy of our methodologies through both simulated and real-world datasets.

1 INTRODUCTION

In many applications, Vector Autoregressive (VAR) model provides a principled framework for a wide

range of tasks, including analyzing speech signal (Juang and Rabiner, 1985; Shannon et al., 2012), investigating causality between economics variables (Granger, 1969), reconstructing gene regulatory interactions (Michailidis and d’Alché Buc, 2013), extracting classifiable features for neuroscience data (Anderson et al., 1998), and finding connectivity between brain regions (Van Den Heuvel and Pol, 2010). A simple form of VAR model is $X_t = BX_{t-1} + \epsilon_t$ where B is a $p \times p$ transition matrix and ϵ_t is the error term at time t . Sparse transition matrices are among popular choices considered in high-dimension regime (the dimension of variables is significantly greater than the number of observations). To get a sparse estimation, there are plenty of studies using penalty for the transition matrix, such as the popular ℓ_1 penalty (lasso), group lasso type penalties employed in Melnyk and Banerjee (2016) and non-convex penalties akin to a square-root lasso (Jiang, 2018). A low-rank transition matrix is assumed in hidden factor model (Bai, 2003). In this scenario, B can be written as the product of two rank- r ($r \ll p$) matrices U, V , i.e. $B = UV'$ so that the resulting model specification of the original p time series is expressed as linear combinations $Z_t = V'X_t$ of the original ones and U specifies the dependence between X_t and Z_t ; namely $X_t = UZ_{t-1} + \epsilon_t$. Z_t is in fact the hidden factor with dimension r , driving the evolution of the process (as a simple example, X_t can be the GDP of each country, and the hidden factors can reflect the general state of the economics at a continent scale). Recent works have generalized the mentioned model by assuming a low-rank plus sparse structure for the transition matrix, i.e. $B = L + S$ where L is the low-rank part and S is the sparse component (Basu et al., 2019; Bai et al., 2020). This decomposition is a natural assumption for dynamic imaging since the L and S components can represent the background and the dynamic foreground, respectively (Otao et al., 2015). To include the low-rank structure, the estimation is achieved by imposing a nuclear penalty. However the algorithm requires relatively large amount of data for training and testing purposes to guarantee a consistent estimation, i.e. number of observations N

should have at least the same order as the dimension p . An important question then arises: when insufficient data is provided, is there a way to estimate L and S with high accuracy? A nature idea of solving the data shortage is leveraging knowledge from additional sources whose observations have similar behavior as original ones. This is where transfer learning comes in. There has been a growing body of literature on transfer learning under high-dimension regime with sparse transition matrices. For example, Cai and Wei (2021) studies transfer learning in the context of nonparametric classification, while Tian and Feng (2022) provides theoretical analysis of transfer learning algorithm, a transferable source detection approach, as well as constructing confidence intervals for model parameters for generalized linear models. Also, a low-rank transfer learning algorithm is proposed by Tian et al. (2023), while they mainly focus on low-rank component instead of a decomposed structure and/or the model parameters are p -dimensional vectors as opposed to squared matrices of dimension p in our paper. Other related works include investigating transfer learning in large-scale Gaussian graphical models with false discovery rate control (Li et al., 2022b), and leveraging big data information via weighted estimator for logistic regression (Zheng et al., 2019); see also related works on hypothesis transfer and meta-learning (Kuzborskij and Orabona, 2013; Wang et al., 2016; Kuzborskij and Orabona, 2017; Aghbalou and Staerman, 2023; Lin and Reimherr, 2024; Tripuraneni et al., 2021). To the best of our knowledge, theoretical analysis of transfer learning for VAR models has not been investigated in the literature. The main goal of this paper is to bridge this gap. To that end, in this work, we focus on VAR models with low-rank plus sparse structure for the transition matrix and propose a transfer learning algorithm to improve the estimation and inference for target model.

Suppose that we have $K + 1$ groups of observations, $X_t^{(i)} = B_i' X_{t-1}^{(i-1)} + \epsilon_t^{(i)}$, $0 \leq i \leq K$ while the first one is the target group. Hence, B_0 is the transition matrix of the target model while $B_i, i \geq 1$ is the transition matrix of the i -th auxiliary model. We assume that B_i can be decomposed as $L + S_i$. In other words, all models have a common low-rank component, which can be interpreted as a background information, while S_i varies across different models. For example, if the i -th group is collected in chronological order, S_i captures a dynamic evolution across time and our model can be applied in a dynamic imaging problem as mentioned before (see more details in Section 5).

The main goal of this paper is estimating sparse component of the target model S_0 with high accuracy while also estimating the shared low-rank structure L . To

address this problem, we propose a transfer learning algorithm by using observations from those informative models whose S_i is close to S_0 . To be more specific, our algorithm comprises of two main steps. First, since L is a common part over all groups, we merge all observations from target and auxiliary sets to estimate the low-rank component denoted by \hat{L} . With more observations, our theory verifies that the first step derives a more accurate estimation of L compared with the result of estimating L with only the target data (Theorem 1). In the second step, we remove the effect of the low-rank component by defining $Y_t^{(i)} := X_t^{(i)} - \hat{L}' X_{t-1}^{(i)}$ and then apply a transfer learning algorithm for the remaining sparse part. It should be noticed that different from merging all collected data, observations from informative models are merged in the transfer learning step. Taking into account observations from non-informative models whose sparse component S_i deviates far away from S_0 could damage the estimation performance, which is a phenomenon often called as negative transfer (Torrey and Shavlik, 2010). Since it is crucial to have a correct informative set, we also propose a novel algorithm to select informative groups from all auxiliary groups. Under some mild condition, it is shown that informative groups and non-informative groups can be separated perfectly with high probability using the proposed algorithm (Theorem 2). Finally, inference for model parameters is performed by adding an additional debiasing step.

In summary, the main contributions of this work include: (1) propose a new algorithm to perform transfer learning for VAR models with low-rank plus sparse structure with theoretical guarantees; (2) develop a novel algorithm for selecting informative sets from all auxiliary groups; (3) constructing confidence intervals for model parameters. These developments came with addressing important challenges due to the complex model structure and temporal dependence. More specifically, (a) due to existence of a shared low rank component, an additional step to the transfer learning algorithm had to be added to estimate that shared piece consistently; (b) inclusion of temporal dependence makes the theoretical development more complicated. We need to design new (A1) Restricted Strong Convexity and (A2) Deviation Bound Conditions. The original version of such conditions are written for single group data while we deal with multiple groups in transfer learning scenarios. As such, these conditions had to be adjusted to the multi-group cases where models from different groups are not same but only similar. Further, these two conditions had to be verified for the proposed model. This is done successfully in the paper by listing appropriate sufficient conditions under which they are satisfied. Certain type of Hanson-Wright inequality is applied to prove that (A1)

and (A2) hold for VAR models; (c) certain parts of the algorithm need to be adapted to respect the temporal dependence. For example, the last step on debiasing the estimate to get confidence intervals had to be adjusted appropriately using an “online debiasing” method so that Martingale-type CLT can be used to derive the asymptotic distribution. A simple debiasing step is not appropriate here since the usual CLT will not work for such models. It is worth noting that developing inferential framework is not only for theoretical purposes, but can also be utilized in practice, i.e. in real data analysis. For example, in Figure 2, we illustrate how the inferential framework helps in finding pixels with potential changes in the surveillance video data, i.e. locating potential root cause of changes in the video data by checking significantly different from zero parameter estimates. Using invalid confidence intervals can potentially ruin this analysis and either select additional unchanged pixels by mistake or fail to select important/changed pixels (see more details in Section 5).

The remainder of the paper is organized as follows. In Section 2, we present the modeling framework, introduce some prior knowledge for VAR models and background for transfer learning algorithm. In Section 3, we introduce the two-step transfer learning algorithm and establish its theoretical properties, introduce an strategy of selecting informative set as well as making inference for sparse component. Section 4 presents our simulation results while a real data is analyzed using our algorithm in Section 5. Finally, some concluding remarks are summarized in Section 6.

Notation: For a $p \times p$ matrix A , $\Lambda_{\min}(A)$, $\Lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A , respectively. $\|A\|_2$ denotes the operator norm for matrix A , i.e. $\|A\|_2 = \sqrt{\Lambda_{\max}(A'A)}$. $\|A\|_\infty$ denotes the infinity norm $\|A\|_\infty = \max_{i,j} |A_{ij}|$. $\|A\|_F$ denotes the Frobenius norm, i.e. $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$. $\|A\|_*$ denote the nuclear norm, i.e. $\sum_{j=1}^p \sigma_j(A)$, the sum of the singular values of a matrix. A^* denotes the conjugate transpose of a matrix A . $\|A\|_0$ denotes the number of non-zero entry in A . $\|A\|_1$ denotes the ℓ_1 norm of A , i.e. $\|A\|_1 = \sum_{i,j} |A_{ij}|$. For a vector $u \in \mathbb{R}^p$, $\|u\|_2 = \sqrt{\sum u_i^2}$ and $\|u\|_1 = \sum |u_i|$. For two sequences $(a_t)_{t \geq 1}$ and $(b_t)_{t \geq 1}$, we write $a_t \lesssim b_t$ if there exists a constant $c \geq 1$ such that $a_t \leq cb_t$ for all t . If both $a_t \lesssim b_t$ and $b_t \lesssim a_t$, we write $a_t \asymp b_t$. Also, $a_t = o(b_t)$ implies $a_t/b_t \rightarrow 0$ as $t \rightarrow \infty$; $a_t = O(b_t)$ implies $a_t/b_t < \infty$; $a_t = \Omega(b_t)$ implies $b_t/a_t \rightarrow 0$. For two real numbers a and b , $a \vee b$ denotes $\max\{a, b\}$ and $a \wedge b$ denotes $\min\{a, b\}$. Finally, let e_j be a vector such that its j -th element is 1 and all other elements are zero.

2 MODEL FORMULATION

As mentioned, we focus on transfer learning for the VAR model. To that end, assume that we observe samples from the target model and K other auxiliary models. Each model has the following expression

$$X_t^{(i)} = B_i' X_{t-1}^{(i)} + \epsilon_t^{(i)}, \quad B_i = L + S_i, \quad \text{rank}(L) = r, \quad (1)$$

where $X_t^{(i)}$ is the p dimensional vector of observed time series at time t for the i -th group. Observations from different models are assumed to be independent with each other. L represents the low-rank matrix while S_i represents the sparse matrix and the transition matrix B_i is low-rank plus sparse. We assume that the number of none-zero entries in S_i is s with $s \ll p^2$. We further assume that the rank of low-rank component L is far less than p , $r \ll p$. The length of i -th group is defined by n_i . Let $N := \sum_{i=0}^K n_i$ be the overall sample size. We can rewrite (1) as

$$\underbrace{\begin{pmatrix} (X_{n_i}^{(i)})' \\ \vdots \\ (X_1^{(i)})' \end{pmatrix}}_{\mathcal{Y}_i} = \underbrace{\begin{pmatrix} (X_{n_i-1}^{(i)})' \\ \vdots \\ (X_0^{(i)})' \end{pmatrix}}_{\mathcal{X}_i} B_i + \underbrace{\begin{pmatrix} (\epsilon_{n_i}^{(i)})' \\ \vdots \\ (\epsilon_1^{(i)})' \end{pmatrix}}_{\mathcal{E}_i}. \quad (2)$$

This model is a generalization of standard sparse VAR model and is able to deal with the setting where there is an invariant cross-autocorrelation structure L across target groups and auxiliary groups. S_i captures the additional cross-sectional autocorrelation structure for each group. Basu et al. (2019) consider the low-rank plus structured sparse model in one-group case and propose an algorithm to estimate L and S accurately. Our goal is to improve the estimation accuracy for sparse component of the target group S_0 and the shared low-rank L provided that more information from auxiliary groups are available.

In the context of transfer learning, to improve the estimation accuracy, we need to select useful observations that have similar behavior as target data. Those observations from auxiliary models that have similar behavior as target data are named as informative observations and their corresponding models are named as informative groups. In this work, similarity is characterized by the difference between the sparse component of transition matrices S_i , i.e. $\delta^k := S_k - S_0$. A small δ^k implies a high level of similarity. When δ^k is relatively small, taking into account observations from k -th group via transfer learning could improve the estimation accuracy of transition matrices. On the contrary, incorporating information from non-informative groups will damage the transfer learning performance, which is known as negative transfer (Zhang et al.,

2022). Informative groups are mathematically defined as $\mathcal{A} = \{k \in \{1, 2, \dots, K\} : \|\delta^{(k)}\|_1 \leq h\}$ where h is some positive number. We also define $\mathcal{A}_0 := \{0\} \cup \mathcal{A}$ to simplify the notation. We use $n_{\mathcal{A}}$ to denote the sample size of informative sets, i.e. $n_{\mathcal{A}} := \sum_{i \in \mathcal{A}} n_i$, and similarly $n_{\mathcal{A}_0} := \sum_{i \in \mathcal{A}_0} n_i$. We also define $\mathcal{X}_{\mathcal{A}_0}$ as the design matrix constructed by \mathcal{X}_i , $i \in \mathcal{A}_0$.

3 ESTIMATION PROCEDURE AND THEORETICAL RESULT

In this section, we introduce the proposed transfer learning algorithm and present theoretical results. As an overview, our estimation procedure comprises of two steps. First, we estimate low-rank matrix given all observations. Since all models share one low-rank matrix, considering all observations could improve estimation accuracy. Then we focus on informative set and apply transfer learning to estimate sparse matrix of the target model. Our algorithm is shown in Algorithm 1.

Algorithm 1 : Transfer learning for sparse component

Input : observations from target model and auxiliary model $\{X_t^{(i)}\}, i = 0, 1, \dots, K$; penalty parameters $\lambda_\beta, \lambda_\delta$; informative set \mathcal{A} and some $\theta > 0$.

Output : coefficient estimator for the target model $\hat{\beta}$.

Step 1 Let $\Omega := \{L \in \mathbb{R}^{p \times p} : \|L\|_\infty \leq \theta\}$

$$\begin{aligned} \hat{L}, \hat{S}_1, \dots, \hat{S}_K = \underset{\substack{L, S_1, \dots, S_K \\ L \in \Omega}}{\operatorname{argmax}} \sum_i \frac{1}{N} \|\mathcal{Y}_i - \mathcal{X}_i(L + S_i)\|_F^2 \\ + \lambda \|L\|_* + \sum_i \frac{1}{\sqrt{N}} \mu_i \|S_i\|_1 \end{aligned} \quad (3)$$

Step 2

$$\tilde{S} = \underset{S \in \mathbb{R}^{p^2}}{\operatorname{argmin}} \sum_{i \in \mathcal{A}} \frac{1}{2n_{\mathcal{A}_0}} \|\mathcal{Y}_i - \mathcal{X}_i(\hat{L} + S)\|_F^2 + \lambda_\beta \|S\|_1 \quad (4)$$

$$\hat{S}_{tran} := \tilde{S} - \tilde{\delta}, \text{ where}$$

$$\tilde{\delta} := \underset{\delta \in \mathbb{R}^{p^2}}{\operatorname{argmin}} \left\{ \frac{1}{2n_0} \|\mathcal{Y}_0 - (\hat{L} + \tilde{S} + \delta)\mathcal{X}_0\|_F^2 + \lambda_\delta \|\delta\|_1 \right\} \quad (5)$$

3.1 Step 1: Estimating The Low-rank Component

The first step is a low-rank plus sparse decomposition problem takes the form of (3), where λ and μ_i are

non-negative tuning parameters controlling the regularizations of low-rank and sparse parts. The parameters θ controls the degree of non-identifiability of decomposition of the low-rank and sparse matrices. For example, if the sparse component $\{S_i\}_{1 \leq i \leq K}$ is also low-rank and low-rank component L is sparse itself, there will be multiple choices of decomposition $L + S_i$ without imposing any further constraints. Larger values of θ provide sparser estimates of sparse component and allow both sparse and low-rank components to be absorbed in \hat{L} . A smaller value of θ , on the other hand, tends to produce a matrix L with smaller rank and pushes both low-rank and sparse components to be absorbed in $\{\hat{S}_i\}_{1 \leq i \leq K}$. We refer to Agarwal et al. (2012) for more details about this identifiability issue. In the low-rank plus sparse regime, consistent estimation relies on the following assumption:

- (A1) Restricted Strong Convexity (RSC): There exist $\alpha > 0$ and $\tau \geq \tau' > 0$ such that for all $\Delta \in \mathbb{R}^{p \times p}$.

$$\begin{aligned} \frac{1}{2N} \sum_i \|\mathcal{X}_i \Delta\|_F^2 &\geq \alpha \|\Delta\|_F^2 - \tau' \Phi^2(\Delta), \\ \frac{1}{2n_i} \|\mathcal{X}_i \Delta\|_F^2 &\geq \alpha \|\Delta\|_F^2 - \tau \|\Delta\|_1^2 \end{aligned}$$

where $\Phi(\Delta) = \inf_{L+S=\Delta} \{\|L\|_* + \frac{\mu}{\lambda} \|S\|_1\}$, $\mu = \max\{\mu_0, \dots, \mu_K\}$ and $\tau = O(\frac{\log p}{\max_i n_i})$.

- (A2) Deviation Bound Condition (DBC): There exists a constant ϕ depending on the model parameters B_0, \dots, B_K and $\Sigma_0, \dots, \Sigma_K$ such that

$$\begin{aligned} \left\| \frac{1}{N} \sum_{i=0}^K \mathcal{X}_i' \mathcal{E}_i \right\|_2 &\leq \phi \sqrt{\frac{p}{N}} \\ \max_{0 \leq i \leq K} \frac{1}{N} \|\mathcal{X}_i' \mathcal{E}_i\|_\infty &\leq \phi \sqrt{\frac{\log p}{N}} \end{aligned}$$

RSC and DBC are basic assumptions for low-rank plus sparse models (Basu et al., 2019). We show that all stable VAR models satisfy these assumptions with high probability in Proposition 1 in the Appendix. Applying the above deviation bounds, we obtain the consistency result for both sparse and low-rank parts.

Theorem 1. Suppose that the low-rank matrix L has rank at most r , while the sparse matrix S_i has at most s nonzero entries for $i \in \{0, 1, \dots, K\}$. Assume that $p = O(N)$ and $\log p = O(n_i)$. Let $\mu_i = 2c_0 \phi \sqrt{\frac{\log p}{N}} + \theta$, $\lambda = 2c_0 \phi \sqrt{\frac{p}{N}}$ and $\theta = o(\sqrt{\frac{p}{N}})$. Under Conditions (A1) and (A2), the estimator of (3) satisfies $\|L - \hat{L}\|_F^2 + \sum_{i=0}^K \frac{n_i}{N} \|S_i - \hat{S}_i\|_F^2 \lesssim s \frac{\log p}{N} + r \frac{p}{N}$.

Remark 1. The convergence rate is a combination of low-rank component and sparse component. The result implies that the upper bound on low-rank component $\|L - \hat{L}\|_F^2$ is $\frac{s \log p + r p}{N}$ and the upper bound on

sparse component $\|S_i - \hat{S}_i\|_F^2$ is $\frac{s \log p + rp}{n_i}$. When there is no auxiliary observations, the upper bound becomes $\|L - \hat{L}\|_F^2 + \|S_0 - \hat{S}_0\|_F^2 \lesssim s \frac{\log p}{n_0} + r \frac{p}{n_0}$. In this case, the upper bound on $\|L - \hat{L}\|_F^2$ is $\frac{s \log p + rp}{n_0}$ and the upper bound on $\|S_0 - \hat{S}_0\|_F^2$ is $\frac{s \log p + rp}{n_0}$. Comparing the conclusions of these two scenarios, we can see that auxiliary observations help improve the estimation accuracy of L but not for S_0 . An additional step is required to improve the estimation for the sparse components, see Theorem 2 and the discussion after the theorem for more details.

Theorem 1 provides estimation consistency for the first step. Since all auxiliary models have a common low-rank component, merging all observations is helpful for estimating L . We next show that a better estimator of S_0 could be obtained from a transfer learning algorithm utilizing the better estimator \hat{L} we found in the first step.

3.2 Step 2: Transfer Learning for Sparse Component

In the second step, we estimate sparse component of the target model S_0 via the transfer learning method summarized in equations (4) and (5). \tilde{S} is an intermediate estimator in transfer learning method calculated by merging target observations and informative observations as source data. This estimator will slightly deviates from S_0 due to the usage of informative observations. We show that \tilde{S} converge to $\bar{S} := (\sum_{i \in \mathcal{A}_0} \Gamma_i)^{-1} (\sum_{i \in \mathcal{A}_0} \Gamma_i S_i)$, $\Gamma_i := \text{Cov}(X_1^{(i)}, X_1^{(i)})$ in the Appendix. To get a consist estimator for S_0 , we need to debias \tilde{S} further, as shown in (5). Next, we introduce the form of Restricted Eigenvalue and Deviation Bound Condition we need in the analysis of this high-dimensional transfer learning problem.

(B1) Restricted Eigenvalue(RE):

$$\begin{aligned} \alpha'_2 \|\Delta\|_F^2 + \tau_{n_{\mathcal{A}_0}} \|\Delta\|_1^2 &\geq \frac{1}{n_{\mathcal{A}_0}} \|\mathcal{X}_{\mathcal{A}_0} \Delta\|_F^2 \\ &\geq \alpha_2 \|\Delta\|_F^2 - \tau_{n_{\mathcal{A}_0}} \|\Delta\|_1^2 \end{aligned}$$

, where $\alpha > 0$, $\alpha' > 0$ and $\tau_{n_{\mathcal{A}_0}} = O(\frac{\log p}{n_{\mathcal{A}_0}})$.

(B2) Deviation Bound Condition:

$$\frac{1}{n_{\mathcal{A}_0}} \left\| \sum_{i \in \mathcal{A}_0} \mathcal{X}'_i \mathcal{E}_i \right\|_{\max} \leq \phi_{\mathcal{A}_0} \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}}$$

, where $\phi_{\mathcal{A}_0}$ is a constant depending on $\{B_i\}_{i \in \mathcal{A}_0}$ and $\{\Sigma_i\}_{i \in \mathcal{A}_0}$.

Proposition 2 in the Appendix shows that (B1) and (B2) are satisfied with high probability in the high-dimensional regime.

Theorem 2. Assume that $\|S_0\|_0 \leq s$ and $\|S_i - S_0\|_1 \leq h$ for all $i \in \mathcal{A}$. We take $\mu = 2(c_3 + c_\Sigma)(1 \vee h^2) \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}}$, $\lambda_\delta = c \sqrt{\frac{\log p}{n_0}}$. Assume that $\frac{n_{\mathcal{A}_0}(pr + s \log p)}{N n_0} = o(1)$ and $\frac{n_0(pr + s \log p)}{N \log p} = o(1)$. Under the condition (B1) and (B2), the estimator of (5) satisfies

$$\begin{aligned} \|\hat{S}_{\text{tran}} - S_0\|_F^2 &\lesssim h \sqrt{\frac{\log p}{n_0}} \wedge h^2 + (1 \vee h^4) \frac{s \log p}{n_{\mathcal{A}_0}} \\ &\quad + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{n_0 N^2} \end{aligned}$$

with high probability.

Theorem 2 provides the convergence rate of S_0 . This consistency rate underscores the non-trivial nature of our method and theoretical developments, as it deviates from existing rates in the literature (Li et al., 2022b; Tian and Feng, 2022; Li et al., 2022a). This distinctive consistency rate offers valuable insights into how the similarity between target and informative groups —quantified by h — affects the overall estimation error. Specifically, it elucidates the interplay among the similarity metric h , dimensionality p , sample sizes of the target and informative groups (n_0 and $n_{\mathcal{A}_0}$), rank r , and sparsity level s . Further, the upper bound on estimation error consists of two parts. First part, $h \sqrt{\frac{\log p}{n_0}} \wedge h^2 + (1 \vee h^4) \frac{s \log p}{n_{\mathcal{A}_0}}$, which we call the transfer learning error, is coming from transfer learning steps. This rate is the same as the rate of traditional transfer learning algorithm (Li et al., 2022a) when no low-rank component is present in the model. Second part is the last term representing the error due to the estimation error of L in our first step. When we estimate L with high accuracy given enough observations (i.e $N \gtrsim \frac{n_{\mathcal{A}_0}(pr + s \log p)}{\sqrt{s n_0 \log p}}$), the third term will be dominated by the transfer learning error. When the informative set \mathcal{A} is empty ($h = 0$ and $n_{\mathcal{A}_0} = n_0$), transfer learning error becomes $\frac{s \log p}{n_0}$, which is the same as the rate of traditional lasso method (Basu and Michailidis, 2015). As we can see, using extra information received from informative groups improves the estimation accuracy when $h = o(s \sqrt{\frac{\log p}{n_0}})$.

3.3 Selecting Informative Set

Algorithm 1 is based on a known informative set, while informative set is typically unknown in practice. Misclassifying non-informative observations as informative observations does harm to the performance of transfer learning. Therefore, we need to select useful observations before applying Algorithm 1. The goal of this section is to determine informative models from all auxiliary models. This algorithm is inspired by Tian and Feng (2022).

The basic idea of selecting informative set comes from cross validation. We evenly split target data into two groups $X_{\mathcal{I}}^{(0)}$ and $X_{\mathcal{I}^c}^{(0)}$ where \mathcal{I} plays the training set role and \mathcal{I}^c as testing set. For each k , we estimate transition matrices for the k -th auxiliary model based on observations from both k -th group and \mathcal{I} . Then, we compute squared residual on the testing set, i.e. $R^{(k)} = \|Y_{\mathcal{I}^c}^{(0)} - X_{\mathcal{I}^c}^{(0)} \hat{\beta}^{(k)}\|_2^2$. A lower test error $R^{(k)}$ implies a closer transition matrix S_k to S_0 , and thus the ones with lower $R^{(k)}$ is selected as informative set. The proposed algorithm is summarized in Algorithm 2 in the Appendix. Next, we make some additional assumptions before presenting the theoretical properties of $R^{(k)}$.

Assumption 1. *There exists some constant $M > 0$, such that, $\sup_k \|S_k - S_0\|_1 \leq M$.*

Assumption 2. *For $k \in \mathcal{A}$, $\|\delta^{(k)}\|_2^2 = O(\sqrt{\frac{\log(p)}{n_0/2}})$; For $k \in \mathcal{A}^c$, $\|\delta^{(k)}\|_2^2 = \Omega(\sqrt{\frac{\log(p)}{n_0/2}})$.*

Theorem 3. *Suppose $\|S_0\|_0 \leq s$, $\frac{s \log(p)}{n_0} = o(1)$ and $K = o(p^2)$. Taking $\lambda_k = C_1(1 \vee M^2) \sqrt{\frac{\log(p)}{n_k + n_0/2}}$, where C_1 depends on \mathcal{M} and \mathbf{m} . Under Assumption 1, We have*

$$\|\delta^{(k)}\|_2^2 - \sqrt{\frac{\log(p)}{n_0/2}} \lesssim R^{(k)} - R_1^{(0)} \lesssim \|\delta^{(k)}\|_2^2 + \sqrt{\frac{\log(p)}{n_0/2}}, \quad (6)$$

with high probability. $R_1^{(0)}$ is defined in Algorithm 2. Further, suppose that Assumption 2 holds. Then, we have $\mathbb{P}\{\sup_{k \in \mathcal{A}} R^{(k)} < \inf_{k \in \mathcal{A}^c} R^{(k)}\} \rightarrow 1$.

Theorem 3 implies that informative groups and non-informative groups can be perfectly separated based on $R^{(k)} - R_1^{(0)}$. Since $R^{(k)} - R_1^{(0)}$ can be treated as testing error, models with lower $R^{(k)} - R_1^{(0)}$ will be preferred for transfer learning. Basically, we can set a threshold and select models with the value of $R^{(k)} - R_1^{(0)}$ below the threshold as informative set. In Algorithm 2, we use $|R_1^{(0)} - R_2^{(0)}|$ as a threshold to select models. As we can see in (6), $|R_1^{(0)} - R_2^{(0)}|$ is lower than $\sqrt{\frac{\log p}{n_0}}$ with high probability, which implies any model in \mathcal{A}^c will be excluded from the estimated informative set by our proposed algorithm.

3.4 Inference for Sparse Component

We propose an additional debiasing step to help with inference. The explicit form of debiased estimator is $\hat{S}^{on} = \hat{S}_{tran} + \frac{1}{n_0} \sum_{i=1}^{n_0} M_i X_i^{(0)} (X_{i+1}^{(0)} - X_i^{(0)} (\hat{L} + \hat{S}_{tran}))'$, where M_i is called the debiasing matrix and needs to be estimated by target model. If observations are i.i.d, setting $M_1 = M_2 \cdots = M_{n_0}$ is an effective way to debias \hat{L} (Javanmard and Montanari,

2014). However, for VAR models, the existence of dependency destroys the asymptotic normality. To fix this problem, we follow the online procedure in Deshpande et al. (2021) in estimating M_i by past observations, $\{X_t\}_{t < i}$, which makes M_i predictable. This online algorithm works for any estimator \hat{S} as long as $\|\hat{S} - S\|_1 = o(s\sqrt{\frac{\log p}{n_0}})$. For the proposed transfer learning estimator, such a rate holds when $h = o(s\sqrt{\frac{\log p}{n_0}})$. Details are deferred to the Appendix due to page limits.

4 SIMULATION RESULTS

In this section, we compare the performance of the proposed transfer learning algorithms with the lasso method (see also additional simulation studies on improvement for recovering the low-rank part utilizing the proposed transfer learning algorithm as well as reporting on computation time in the appendix). Transfer learning algorithms include the Oracle Trans-Lasso (Algorithm 1 with known \mathcal{A}), Trans-Lasso (Algorithm 1 with \mathcal{A} selected by Algorithm 2), and naive Trans-lasso (Algorithm 1 with $\mathcal{A} = \{1, 2, \dots, K\}$). The lasso method is applied only for the target data. Discuss on hyperparameter selection and their sensitivity analysis are summarized in the Appendix due to space consideration.

We focus on both estimation and inference performances of oracle transfer learning algorithms with the lasso method. All simulations are repeated 200 times. In this simulation setting, the entries of S_0 are generated independently from a Bernoulli distribution with success probability $q = 0.02$, multiplied by $b \cdot \text{uniform}(\{+1, -1\})$ with $b = 0.25$, i.e. $b \cdot \text{Bernoulli}(q) \cdot \text{uniform}(\{+1, -1\})$. L is generated by $L = UDV'$, where $D := \text{diag}(0.2, r)$ is a diagonal matrix. Rank of L , r is set to be 8, We set $p = 100$, $n_0 = 200$, $n_1 = n_2 = \dots = n_K = 100$, and $K = 10$. Note that in this setting, the total number of parameters in the target model is $p^2 = 10,000$ while there are only 200 time points. Thus, this can be regarded as the high-dimensional case. Let \mathcal{A} denote the informative set. We define \mathcal{J} as the set of non-zero entries in S_0 , and \mathcal{J}^c as the set of zero entries. For the transition matrices of auxiliary models S_k , we construct them by modifying entries of S_0 in \mathcal{J} and \mathcal{J}^c separately. For a given k , let H_k be a random subset of \mathcal{J} such that $|H_k| = \gamma|\mathcal{J}|$, and G_k be a random subset of \mathcal{J}^c such that $|G_k| = \gamma|\mathcal{J}^c|$, where $\gamma = \gamma_1$ if $k \in \mathcal{A}$, and $\gamma = \gamma_2$ otherwise, and it ranges from 0 to 1. If $(i, j) \in H_k$, we set $S_{ij}^{(k)} = -S_{ij}^{(0)}$. If $(i, j) \in G_k$, we set $S_{ij}^{(k)} = S_{ij}^{(0)} + \eta_{ij}$, where $\eta_{ij} \sim \text{uniform}(-0.1, 0.1)$. The two terms γ_1 and γ_2 are the percentage of changes we make for entries of A_0 . We set $\gamma_1 = 0.04$, $\gamma_2 = 0.4$ and $|\mathcal{A}| = \{0, 2, 4, \dots, 10\}$.

The estimation error is shown in Figure 3 in the Appendix. We present the absolute estimation error, i.e. $\|S_0 - \hat{S}\|_1$, for lasso, Oracle Trans-lasso, naive Trans-lasso and Trans-lasso with an increasing samples in informative set. As expected, Trans-lasso has better estimation error than lasso method when we consider enough informative samples. The similar behavior of Oracle Trans-lasso and lasso implies that Algorithm 2 selects informative set accurately. As for naive Trans-lasso, it outputs the worst estimation when the size of informative set $|\mathcal{A}|$ is small and reach the same accuracy level when $|\mathcal{A}|$ increases. This is because naive Trans-lasso takes non-informative set into consideration, which damages the estimation performance.

In addition, we construct entry-wise confidence interval for sparse matrix based on Trans-lasso and lasso separately. To make comparison for inference performance, we consider four metrics: True Positive Rate (TPR), False Positive Rate (FPR), coverage rate of confidence intervals and average confidence interval length (Avg CI length). Figure 1 summarizes the results for all methods at significance level $\alpha = 0.05$. As we can see, Trans-lasso shows a comparable result with lasso method when $|\mathcal{A}| = 0$ (all auxiliary sets are non-informative). This is because transfer learning with no informative set is equivalent to lasso method. As more informative sets are provided, Trans-lasso performs better in terms of FPR, TPR and coverage rate. The coverage rate gradually goes up to 0.95 as $|\mathcal{A}|$ increases, which is consistent with our significant level $\alpha = 0.05$. Since both lasso and transfer learning use the same method to generate conditional variance V_n , the length of confidence intervals for lasso and transfer learning are at the same level all the time.

5 A REAL DATA EXAMPLE

The proposed transfer learning algorithm is applied to a surveillance video data set obtained from the CAVIAR project¹. A number of video clips record different actions by people in diverse settings, including walking alone, meeting with others, entering and exiting a room, etc. The data set we analyze is “Two other people meet and walk together”. The original data set has 837 images in total and the resolution of each image is half-resolution PAL standard (384×288 pixels, 25 frames per second). Before applying our proposed algorithm, we re-sized the original images from 384×288 pixels to 32×24 pixels and used a gray-scaled scheme instead of the original colored image to accelerate computations. We then digitalize and vectorize 32×24 image to be a 768 dimension vector. Therefore, the resulting time series process has $n = 837$ time points and $p = 768$ features.

¹<http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/>

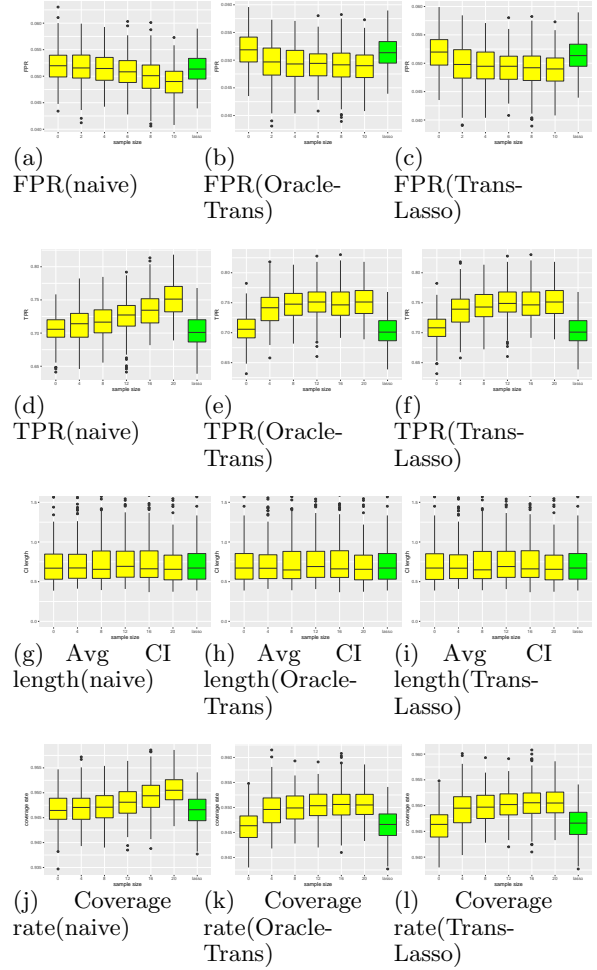


Figure 1: FPR (False Positive Rate), TPR (True Positive Rate), Coverage Rate and Average Length of Confidence Intervals (Avg CI Length) at significance level $\alpha = 0.05$. The x axis is the size of informative set $|\mathcal{A}|$.

The whole video data can be divided into 12 segments depending on human activities in the video (see more details in Bai et al. (2020)). The background (mainly the lobby) seems fixed during the time while certain human movements occur during the video. When our model is applied to this data, the low-rank part will capture the background (lobby) while the sparse components will capture the additional human movements during the video clip. To ensure the proposed model is a good fit, we compared it with several additional parameterizations (including low-rank only, sparse only, etc.) and concluded that the proposed algorithm coupled with low rank plus sparse model parameters outperforms all other competing models. More specifically, we consider five scenarios in total: Trans-lasso(L+S), Trans-lasso(S), lasso(L+S), lasso(S) and Low-rank. Trans-lasso(L+S) and lasso(L+S) refer to methods that we model VAR with low-rank plus sparse

structure, with and without transfer learning, respectively. Trans-lasso(S) and lasso(S) refer to methods that we model VAR with sparse components only, with and without transfer learning, respectively. Low-rank method implies that we model VAR with only low-rank component for each segment. Results of this comparison are summarized in Table 4 in the Appendix. As seen from this table, the proposed modeling framework with the help from the proposed transfer learning algorithm achieves the best prediction error overall.

First row of Figure 2 shows the start time point for four of these segments/movements: 1st segment, 4th segment, 6th segment and 9th segment. We apply Algorithm 1 to estimate the low-rank component (dimension is 768×768) for each segment. Since non-changing low-rank component corresponding to the stationary background of the space surveyed and the changing sparse component corresponds to movement of people in and out of the space in the evolving foreground, the sparse component can imply the position of people in the lobby. To visualize the information contained in sparse component, we (1) construct entry-wise 95% confidence interval of the sparse estimator \hat{S} , (2) count the number of significant entries in each row, i.e $V := (v_1, \dots, v_{768})$, $v_i = \#\{j : \hat{S}_{ij} \text{ is significant}\}$, (3) map vector V back to a 32×24 matrix M . Figure 2(e)-(h) shows the heatmap of M . As we can see, the dark region of the heatmap perfectly matches the position of people in original image. Results for other segments are summarized in the Appendix. Further, Table 4 summarizes out-of-sample mean squared prediction error (each segment is split such that its first 2/3 observations are used as training and the remaining parts as testing data) obtained from lasso and Trans-lasso (T-lasso) algorithms which clearly illustrates the great reduction of prediction error when similar images are used in the estimation procedure for each segment.

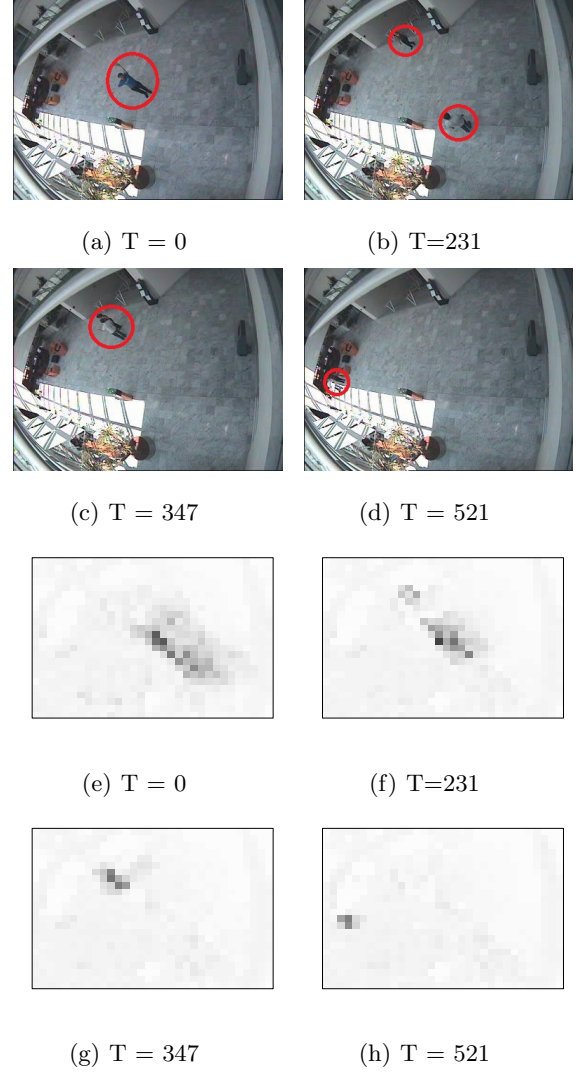


Figure 2: 2(a)-2(d) are the view of footage for 1st segment, 4th segment, 6th segment and 9th segment respectively. 2(e)-2(h) are corresponding heatmaps of sparse estimators.

Table 1: Mean Squared Prediction Error for Each Segment; Standard Errors Are Shown in Parentheses.

	seg1	seg2	seg3
T-lasso	7.205(0.015)	0.133(0.003)	2.177(0.012)
lasso	8.660(0.017)	0.241(0.004)	4.484(0.017)
	seg4	seg5	seg6
T-lasso	0.468(0.005)	0.092(0.002)	0.916(0.008)
lasso	1.928(0.011)	0.450(0.005)	2.686(0.013)
	seg7	seg8	seg9
T-lasso	0.372(0.005)	0.233(0.004)	0.146(0.003)
lasso	1.653(0.010)	1.235(0.009)	0.523(0.006)
	seg10	seg11	seg12
T-lasso	0.094(0.002)	0.071(0.002)	0.139(0.002)
lasso	0.189(0.004)	0.102(0.004)	0.153(0.002)

6 DISCUSSION

In this paper, we propose an step-wise algorithm for implementing transfer learning for VAR models with low-rank plus sparse structure. Theoretical results confirm that our transfer learning algorithm can improve the estimation accuracy for the low-rank and sparse components given known informative set. We also provide an approach based on prediction error for properly selecting the informative sets when it is unknown. Numerical experiments and real data applications support the theoretical findings. In our model, all auxiliary models have a common low-rank component. How to relax to the case where auxiliary models have different but similar low-rank components is an

interesting future research direction. Measuring the similarity according to the column space of low-rank components (Tian et al., 2023) can be a feasible approach. Another limitation of our work is considering a VAR model with single lag. Extensions to VAR model of general lag, i.e. VAR(d) models (utilizing techniques in Lütkepohl (2005)) is of interest.

References

- Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions.
- Aghbalou, A. and Staerman, G. (2023). Hypothesis transfer learning with surrogate classification losses: Generalization bounds through algorithmic stability. In *International Conference on Machine Learning*, pages 280–303. PMLR.
- Anderson, C. W., Stolz, E. A., and Shamsunder, S. (1998). Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3):277–286.
- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.
- Bai, P., Safikhani, A., and Michailidis, G. (2020). Multiple change points detection in low rank and sparse high dimensional vector autoregressive models. *IEEE Transactions on Signal Processing*, 68:3074–3089.
- 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.
- Cai, T. T. and Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128.
- Deshpande, Y., Javanmard, A., and Mehrabi, M. (2021). Online debiasing for adaptively collected high-dimensional data with applications to time series analysis. *Journal of the American Statistical Association*, pages 1–14.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.
- Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.
- 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.
- Jiang, H. (2018). Sparse estimation based on square root nonconvex optimization in high-dimensional data. *Neurocomputing*, 282:122–135.
- Juang, B.-H. and Rabiner, L. (1985). Mixture autoregressive hidden markov models for speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1404–1413.
- Kuzborskij, I. and Orabona, F. (2013). Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR.
- Kuzborskij, I. and Orabona, F. (2017). Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106:171–195.
- Li, S., Cai, T. T., and Li, H. (2022a). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173.
- Li, S., Cai, T. T., and Li, H. (2022b). Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, pages 1–13.
- Lin, H. and Reimherr, M. (2024). Smoothness adaptive hypothesis transfer learning. *International Conference on Machine Learning*.
- Loh, P.-L. and Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24.
- Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.
- Melnyk, I. and Banerjee, A. (2016). Estimating structured vector autoregressive models. In *International Conference on Machine Learning*, pages 830–839. PMLR.
- Michailidis, G. and d’Alché Buc, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246(2):326–334.
- Otazo, R., Candes, E., and Sodickson, D. K. (2015). Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components. *Magnetic resonance in medicine*, 73(3):1125–1136.
- Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix

equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

Rudelson, M. and Vershynin, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9.

Shannon, M., Zen, H., and Byrne, W. (2012). Autoregressive models for statistical parametric speech synthesis. *IEEE transactions on audio, speech, and language processing*, 21(3):587–597.

Tian, Y. and Feng, Y. (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, (just-accepted):1–30.

Tian, Y., Gu, Y., and Feng, Y. (2023). Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

Tripuraneni, N., Jin, C., and Jordan, M. (2021). Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR.

Van Den Heuvel, M. P. and Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534.

Wang, X., Oliva, J. B., Schneider, J. G., and Póczos, B. (2016). Nonparametric risk and stability analysis for multi-task learning problems. In *IJCAI*, pages 2146–2152.

Zhang, W., Deng, L., Zhang, L., and Wu, D. (2022). A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*.

Zheng, C., Dasgupta, S., Xie, Y., Haris, A., and Chen, Y. Q. (2019). On data enriched logistic regression. *arXiv preprint arXiv:1911.06380*.

result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:

- (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
- (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
- (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:

- (a) Statements of the full set of assumptions of all theoretical results. [Yes]
- (b) Complete proofs of all theoretical results. [Yes]
- (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

- (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
- (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
- (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
- (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

- (a) Citations of the creator If your work uses existing assets. [Yes]
- (b) The license information of the assets, if applicable. [Not Applicable]
- (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't

- (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
- (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

In this section, some prior information about VAR models are summarized in Section A, some useful lemmas with their proofs are provided in Section B while some propositions with their proofs are summarized in Section C. Proof of main theorems are stated in Section D while additional details on the proposed algorithms are described in Section E. Further, some additional details on numerical studies and a new simulation study are explained in Section F. Finally, computer information is summarized in Section G.

A Prior Knowledge for VAR Model in High-dimensions

For a p -dimensional centered, covariance-stationary process $\{X_t\}_{t \in \mathbb{Z}}$ with autocovariance function $\Gamma_X(h) = \text{Cov}(X_t, X_{t+h})$, its spectral density is defined as $f_X(\theta) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} \Gamma_X(h) e^{-ih\theta}$. For the VAR model (1) in the main file, the spectral density has the closed form $f_{X^{(k)}}(\theta) = \frac{1}{2\pi} (\mathcal{B}_k^{-1}(e^{i\theta})) \Sigma_k (\mathcal{B}_k^{-1}(e^{i\theta}))^*$ where $\mathcal{B}_k(z) = I_p - B_k' z$ is the characteristic polynomial and Σ_k is the covariance matrix of the error term. To introduce some useful properties for VAR model, we need the following quantities

$$\begin{aligned} \mathcal{M}(f_{X^{(k)}}) &:= \sup_{\theta \in [-\pi, \pi]} \Lambda_{\max}(f_{X^{(k)}}(\theta)) \\ \mathbf{m}(f_{X^{(k)}}) &:= \sup_{\theta \in [-\pi, \pi]} \Lambda_{\min}(f_{X^{(k)}}(\theta)) \\ \mu_{\max}(\mathcal{B}_k) &:= \max_{|z|=1} \Lambda_{\max}(\mathcal{B}_k^*(z) \mathcal{B}_k(z)) \\ \mu_{\min}(\mathcal{B}_k) &:= \min_{|z|=1} \Lambda_{\max}(\mathcal{B}_k^*(z) \mathcal{B}_k(z)). \end{aligned}$$

Stability is always a basic assumption in time series model to ensure consistent estimation. Basu and Michailidis (2015) provide a new measure of stability described by $\mathcal{M}(f_X)$ and shows that a larger $\mathcal{M}(f_X)$ implies a less stable process. $\mathcal{M}(f_X)$ and $\mathbf{m}(f_X)$ capture the dependence among the univariate components of the vector-valued time series and help quantify dependence among the columns of the design matrix in our analysis. For VAR model, the boundness of $\mathcal{M}(f_X)$ and $\mathbf{m}(f_X)$ are related to $\mu_{\max}(\mathcal{B})$ and $\mu_{\min}(\mathcal{B})$: $\mathcal{M}(f_{X^{(k)}}) \leq \frac{1}{2\pi} \frac{\Lambda_{\max}(\Sigma_k)}{\mu_{\min}(\mathcal{B}_k)}$, $\mathbf{m}(f_{X^{(k)}}) \geq \frac{1}{2\pi} \frac{\Lambda_{\min}(\Sigma_k)}{\mu_{\max}(\mathcal{B}_k)}$.

Notations. For a matrix A , its transpose is denoted by A' while $\text{vec}(A)$ is the vectorized version of matrix A . $\Lambda_{\min}(A)$, $\Lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A , respectively. Define $\Gamma_{X^{(k)}}(i-j) := \text{Cov}(X_i^{(k)}, X_j^{(k)})$. $\Upsilon_{n_k}^{X^{(k)}} = \text{Cov}(\text{vec}((\mathcal{X}^{(k)})'), \text{vec}((\mathcal{X}^{(k)})'))$. From Proposition 2.3 in Basu and Michailidis (2015), we know that

$$\begin{aligned} 2\pi \mathbf{m}(f_{X^{(k)}}) &\leq \Lambda_{\min}(\Upsilon_{n_k}^{X^{(k)}}) \leq \Lambda_{\max}(\Upsilon_{n_k}^{X^{(k)}}) \leq 2\pi \mathcal{M}(f_{X^{(k)}}) \\ 2\pi \mathbf{m}(f_{X^{(k)}}) &\leq \Lambda_{\min}(\Gamma_k) \leq \Lambda_{\max}(\Gamma_k) \leq 2\pi \mathcal{M}(f_{X^{(k)}}). \end{aligned}$$

Similarly, for $\{\epsilon_0^{(k)}, \dots, \epsilon_{n_k}^{(k)}\}$, we have

$$\begin{aligned} 2\pi \mathbf{m}(f_{\epsilon^{(k)}}) &\leq \Lambda_{\min}(\Upsilon_{n_k}^{\epsilon^{(k)}}) \leq \Lambda_{\max}(\Upsilon_{n_k}^{\epsilon^{(k)}}) \leq 2\pi \mathcal{M}(f_{\epsilon^{(k)}}) \\ 2\pi \mathbf{m}(f_{\epsilon^{(k)}}) &\leq \Lambda_{\min}(\Sigma_k) \leq \Lambda_{\max}(\Sigma_k) \leq 2\pi \mathcal{M}(f_{\epsilon^{(k)}}). \end{aligned}$$

We define $\mathcal{M}_{\epsilon} := \max_k \mathcal{M}(f_{\epsilon^{(k)}})$ and $\mathbf{m}_{\epsilon} := \min_k \mathbf{m}(f_{\epsilon^{(k)}})$. For two matrices A and B , the inner product of A and B is defined as $\langle A, B \rangle := \sum_{i,j} (AB')_{ij}$.

B Useful Lemmas with Proofs

lemma 1. Consider model (1). Recall the following notation $2\pi \mathcal{M} = \max_k \Lambda_{\max}(\Upsilon_{n_k}^{X^{(k)}})$, $\Gamma_k = \text{Cov}(X_0^{(k)}, X_0^{(k)})$. Suppose $v_0, \dots, v_K \in \mathbb{R}^p$. We have that,

$$\mathbb{P} \left(\frac{1}{N} \left| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} v'_k [(X_i^{(k)})(X_i^{(k)})' - \Gamma_k] v_k \right| \geq 2\pi \mathcal{M} \max_k (\|v_k\|_2^2) t \right) \leq 2 \exp[-cN \min\{t, t^2\}], \quad (7)$$

$$\begin{aligned} \mathbb{P} \left(\frac{1}{N} \left| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} u'_k [(X_i^{(k)})(X_i^{(k)})' - \Gamma_k] v_k \right| \geq 6\pi \mathcal{M} (\max_k (\|v_k\|_2^2) + \max_k (\|u_k\|_2^2)) t \right) \\ \leq 6 \exp[-cN \min\{t, t^2\}]. \end{aligned} \quad (8)$$

Proof. Let $V_i^{(k)} = (X_i^{(k)})' v_k$ and $Q = \text{Var}(V_1^{(0)}, \dots, V_{n_0}^{(0)}, V_1^{(1)}, \dots, V_{n_2}^{(1)}, \dots, V_1^{(K)}, \dots, V_{n_K}^{(K)})$. The entry of Q is shown as follow:

$$\text{Cov}(V_i^{(k)}, V_j^{(k)}) = v'_k \Gamma_{\bar{X}^{(k)}}(i-j) v_k; \quad \text{Cov}(V_i^{(k_1)}, V_j^{(k_2)}) = 0.$$

Define $Q^{(k)} = \text{Var}(V_1^{(k)}, \dots, V_{n_k}^{(k)})$. We can see that Q is a block diagonal matrix,

$$Q = \begin{pmatrix} Q^{(0)} & 0 & \dots & 0 \\ 0 & Q^{(1)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & Q^{(K)} \end{pmatrix}.$$

For $Q^{(k)}$ and any $\|w\|_2 = 1$, we have

$$\begin{aligned} w' Q^{(k)} w &= \sum_{r=1}^{n_k} \sum_{s=1}^{n_k} w_r w_s Q_{rs}^{(k)} = \sum_{r=1}^{n_k} \sum_{s=1}^{n_k} w_r w_s (v^{(k)})' \Gamma_{\bar{X}^{(k)}}(r-s) v^{(k)} \\ &= (w \otimes v^{(k)})' \Upsilon_{n_k}^{X^{(k)}}(w \otimes v^{(k)}) \\ &\leq \Lambda_{\max}(\Upsilon_{n_k}^{X^{(k)}}) \|v_k\|_2^2 \leq 2\pi \mathcal{M}(f_{X^{(k)}}) \|v_k\|_2^2. \end{aligned} \quad (9)$$

Since $\|Q\|_2 \leq \max_k (\|Q^{(k)}\|_2)$, we obtain $\|Q\|_2 \leq 2\pi \mathcal{M} \max_k (\|v_k\|_2^2)$. By the Hanson–Wright inequality (Rudelson and Vershynin, 2013)

$$\mathbb{P} \left(\frac{1}{N} \left| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} v'_k [(X_i^{(k)})(X_i^{(k)})' - \Gamma_k] v_k \right| \geq \eta \right) \leq 2 \exp[-c \min\{ \frac{N^2 \eta^2}{\|Q\|_F^2}, \frac{N \eta}{\|Q\|_2} \}]. \quad (10)$$

Setting $\eta = \|Q\|_2 t$ and using $\|Q\|_F^2 \leq N \|Q\|_2$, we get (7).

To prove (8), notice that

$$\begin{aligned} \frac{1}{N} \left| \sum_{\substack{k=1, \dots, K \\ i=1, \dots, n_k}} u'_k [(X_i^{(k)})(X_i^{(k)})' - \Gamma_k] v_k \right| &\leq \frac{1}{N} \left| \sum_{\substack{k=1, \dots, K \\ i=1, \dots, n_k}} u'_k [(X_i^{(k)})(X_i^{(k)})' - \Gamma_k] u_k \right| \\ &\quad + \frac{1}{N} \left| \sum_{\substack{k=1, \dots, K \\ i=1, \dots, n_k}} v'_k [(X_i^{(k)})(X_i^{(k)})' - \Gamma_k] v_k \right| \\ &\quad + \frac{1}{N} \left| \sum_{\substack{k=1, \dots, K \\ i=1, \dots, n_k}} (u'_k + v'_k) [(X_i^{(k)})(X_i^{(k)})' - \Gamma_k] (u_k + v_k) \right|. \end{aligned}$$

By applying (7) on each of three terms separately, we get (8).

lemma 2. Consider model (1), we have

$$\mathbb{P} \left(\frac{1}{N} \left\| \sum_{k=0, \dots, K} (\mathcal{X}^{(k)})' \mathcal{E}^{(k)} \right\|_{\infty} \geq 6\pi(\mathcal{M} + \mathcal{M}_{\epsilon})t \right) \leq 6p^2 \exp(-cN \min\{t, t^2\}).$$

Proof. Note that

$$\frac{1}{N} \left\| \sum_{k=0, \dots, K} (\mathcal{X}^{(k)})' \mathcal{E}^{(K)} \right\|_{\infty} = \frac{1}{N} \left\| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} (X_i^{(k)}) (\epsilon_i^{(k)})' \right\|_{\infty}.$$

Next we get the upper bound for

$$\mathbb{P} \left(\frac{1}{N} \left| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} u' X_i^{(k)} (\epsilon_i^{(k)})' v \right| \geq t \right),$$

where $u, v \in \mathbb{R}^p$ and $\|u\|_2, \|v\|_2 = 1$. Since $\text{Cov}((X_i^{(k)})' u, (\epsilon_i^{(k)})' v) = 0$, we have the following decomposition:

$$\begin{aligned} \frac{1}{N} \left(\sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} u' X_i^{(k)} (\epsilon_i^{(k)})' v \right) &= \underbrace{\left[\frac{1}{N} \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} ((X_i^{(k)})' u + (\epsilon_i^{(k)})' v)^2 - \text{Var}((X_i^{(k)})' u + (\epsilon_i^{(k)})' v) \right]}_{(a)} \\ &\quad - \underbrace{\left[\frac{1}{N} \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} ((X_i^{(k)})' u)^2 - \text{Var}((X_i^{(k)})' u) \right]}_{(b)} - \underbrace{\left[\frac{1}{N} \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} ((\epsilon_i^{(k)})' v)^2 - \text{Var}((\epsilon_i^{(k)})' v) \right]}_{(c)}. \end{aligned}$$

We can apply (7) to obtain that

$$\begin{aligned} \mathbb{P} \left(\frac{1}{N} \left| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} u' [(X_i^{(k)}) (X_i^{(k)})' - \Gamma_k] u \right| \geq 2\pi \mathcal{M} t \right) &\leq 2 \exp[-cN \min\{t, t^2\}], \\ \mathbb{P} \left(\frac{1}{N} \left| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} v' [(\epsilon_i^{(k)}) (\epsilon_i^{(k)})' - \Gamma_k] v \right| \geq 2\pi \mathcal{M}_{\epsilon} t \right) &\leq 2 \exp[-cN \min\{t, t^2\}]. \end{aligned}$$

For (a), similar to the proof of Lemma 1, we let

$$\begin{aligned} Q &= \text{Var}((X_1^{(0)})' u + (\epsilon_0^{(0)})' v, \dots, (X_{n_K}^{(K)})' u + (\epsilon_{n_K}^{(K)})' v), \\ Q_k &= \text{Var}((X_1^{(k)})' u + (\epsilon_0^{(k)})' v, \dots, (X_{n_k}^{(k)})' u + (\epsilon_{n_k}^{(k)})' v), \quad k = 1, \dots, K. \end{aligned}$$

Since Q is a block diagonal matrix

$$Q = \begin{pmatrix} Q_0 & 0 & \dots & 0 \\ 0 & Q_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & Q_K \end{pmatrix}.$$

We have that $\|Q\|_2 \leq \max_k \{\|Q_k\|_2\}$. For every k , we define

$$\begin{aligned} Q_r' &= \text{Var}((X_1^{(k)})'u - (\epsilon_0^{(k)})'v, \dots, (X_{n_r}^{(k)})'u - (\epsilon_{n_k}^{(k)})'v) \\ Q_{k,1} &= \text{Var}((X_1^{(r)})'u, \dots, (X_{n_k-h}^{(k)})'u) \\ Q_{k,2} &= \text{Var}((\epsilon_0^{(k)})'v, \dots, (\epsilon_{n_k}^{(k)})'v). \end{aligned}$$

Note that $Q_r, Q_r', Q_{r,1}, Q_{r,2}$ are positive definite. Since $Q_r + Q_r' = 2Q_{r,1} + 2Q_{r,2}$, we have that $\|Q_r\|_2 \leq 2\|Q_{r,1}\|_2 + 2\|Q_{r,2}\|_2$. Using the same method as (9), we get $\|Q_{r,1}\|_2 \leq \Lambda_{\max}(\Upsilon_{n_k}^{X^{(r)}})$, $\|Q_{r,2}\|_2 \leq \Lambda_{\max}(\Sigma_{\epsilon^{(r)}})$. Therefore, $\|Q\|_2 \leq \max_r \{2\|Q_{r,1}\|_2 + 2\|Q_{r,2}\|_2\} \leq 4\pi\mathcal{M}^* + 4\pi\mathcal{M}_\epsilon \leq 4\pi\mathcal{M} + 4\pi\mathcal{M}_\epsilon$. Applying Hanson–Wright inequality (Rudelson and Vershynin, 2013) again, we obtain that

$$\mathbb{P}(|(a)| \geq 4\pi(\mathcal{M} + \mathcal{M}_\epsilon)t) \leq 2\exp[-cN \min\{t, t^2\}].$$

By the probability inequalities derived for (a), (b), (c), we have that

$$\mathbb{P}\left(\frac{1}{N} \left| \sum_{\substack{k=0, \dots, K \\ i=1, \dots, n_k}} u' X_i^{(k)} (\epsilon_i^{(k)})' v \right| \geq 6\pi(\mathcal{M} + \mathcal{M}_\epsilon)t \right) \leq 6\exp[-cN \min\{t, t^2\}].$$

Let e_j be a vector such that its j -th element is 1 and all other elements are zero. Observe that

$$\frac{1}{N} \left\| \sum_{k=0, \dots, K} (\mathcal{X}^{(k)})' \mathcal{E}^{(K)} \right\|_\infty = \max_{1 \leq r, s \leq p} \frac{1}{N} \left| \sum_{\substack{k=1, \dots, K \\ i=1, \dots, n_k}} e_r' (X_i^{(k)}) (\epsilon_i^{(k)})' e_s \right|.$$

Taking a union bound over r, s yields the final result.

lemma 3. Consider model (1), we have

$$\mathbb{P}\left(\frac{1}{n_i} \|(\mathcal{X}^{(k)})' \mathcal{X}^{(k)} - \Gamma_k\|_\infty \geq 2\pi\mathcal{M}t\right) \leq 6p^2 \exp(-cn_i \min\{t, t^2\}).$$

Proof. Similar to lemma 2, we can prove that for $u, v \in \mathbb{R}^p$

$$\mathbb{P}\left(\frac{1}{n_i} \left| \sum_{i=1, \dots, n_k} u' (X_i^{(k)} (X_i^{(k)})' - \Gamma_k) v \right| \geq 2\pi\mathcal{M}t \|u\|_2 \|v\|_2 \right) \leq 2\exp[-cn_i \min\{t, t^2\}].$$

Observe that

$$\frac{1}{n_i} \|(\mathcal{X}^{(k)})' \mathcal{X}^{(k)} - \Gamma_k\|_\infty = \max_{1 \leq r, s \leq p} \frac{1}{n_i} |e_r' (\sum_{i=1, \dots, n_k} X_i^{(k)} (X_i^{(k)})' - \Gamma_k) e_s|.$$

Taking a union bound over r, s yields the final result.

lemma 4. Consider model (1). Under the conditions of Theorem 1, we have

$$\left\| \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}_0} (S_i - \bar{S}) \mathcal{X}_i' \mathcal{X}_i \right\|_\infty \leq c_{\mathcal{M}} \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}} (1 + h^2) \quad (11)$$

with high probability

Proof. Define $\delta_i := \text{vec}(S_i - \bar{S})$. Notice that

$$\begin{aligned} \text{vec} \left(\frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}_0} (S_i - \bar{S}) \mathcal{X}_i' \mathcal{X}_i \right) &= \text{vec} \left(\frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}_0} (S_i - \bar{S}) (\mathcal{X}_i' \mathcal{X}_i - \Gamma_i) \right) \\ &= \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}_0} \delta_i' (I_p \otimes \mathcal{X}_i' \mathcal{X}_i - I_p \otimes \Gamma_i). \end{aligned}$$

Therefore, we have

$$\left\| \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}_0} (S_i - \bar{S}) \mathcal{X}'_i \mathcal{X}_i \right\|_{\infty} = \max_{s \in \{1, \dots, p^2\}} \left\| \sum_{i \in \mathcal{A}_0} \alpha_k \delta'_k \left[\frac{1}{n_k} (I_p \otimes \mathcal{X}'_i \mathcal{X}_i) - I_p \otimes \Gamma_k \right] e_s \right\|_{\infty}. \quad (12)$$

Similar to lemma 2, we can prove that

$$\begin{aligned} & \mathbb{P} \left(\left| \sum_{k \in \mathcal{A}_0} \alpha_k u'_k \left[\frac{1}{n_k} (I_p \otimes \mathcal{X}'_i \mathcal{X}_i) - I_p \otimes \Gamma_k \right] v_k \right| \geq 6\pi \mathcal{M} (\max_k \|v_k\|_2^2 + \max_k \|v_k\|_2^2) t \right) \\ & \leq 6p \exp[-cn_{\mathcal{A}_0} \min\{t, t^2\}], \end{aligned} \quad (13)$$

where $\alpha_k := \frac{n_k}{n_{\mathcal{A}_0}}$ and $u_k, v_k \in \mathbb{R}^{p^2}$. Recall that $\|\delta_k\|_1 \leq h$, and thus $\|\delta_k\|_2 \leq h$. Applying (13) to (12), we arrive at

$$\begin{aligned} & \mathbb{P} \left(\left| \sum_{i \in \mathcal{A}_0} \alpha_k \delta'_k \left[\frac{1}{n_k} (I_p \otimes \mathcal{X}'_i \mathcal{X}_i) - I_p \otimes \Gamma_k \right] e_s \right| \geq 6\pi \mathcal{M} (1 + h^2) t \right) \\ & \leq 6p \exp[-cn_{\mathcal{A}_0} \min\{t, t^2\}]. \end{aligned}$$

Setting $t = \sqrt{\frac{6 \log p}{cn_{\mathcal{A}_0}}}$ and taking the union bound over s yield the final result.

C Propositions with Their Proofs

proposition 1. Consider model (1). Define

$$\phi((B_0, \Sigma_0) \cdots, (B_K, \Sigma_K)) = \max_{0 \leq i \leq K} \Lambda_{\max}(\Sigma^{(i)}) \left[1 + \frac{1 + \mu_{\max}(B_i)}{\mu_{\min}(B_i)} \right].$$

We will write ϕ instead of $\phi((B_0, \Sigma_0) \cdots, (B_K, \Sigma_K))$ when the meaning of the deterministic constant is clear from the context. Then, there exist universal positive constants $c_i > 0$ such that

(1) for $N \gtrsim p$,

$$\mathbb{P} \left[\left\| \frac{1}{N} \sum_i \mathcal{X}'_i \mathcal{E}_i \right\|_2 > c_0 \phi \sqrt{\frac{p}{N}} \right] \leq c_1 \exp[-c_2 \log p] \quad (14)$$

and for $N \gtrsim \log p$

$$\mathbb{P} \left[\max_i \frac{1}{N} \|\mathcal{X}'_i \mathcal{E}_i\|_{\infty} \geq c_0 \phi \sqrt{\frac{\log p}{N}} \right] \leq c_1 \exp[-c_2 \log p] \quad (15)$$

(2) for $N \gtrsim p$,

$$\mathbb{P} \left[\Lambda_{\min} \left(\frac{1}{N} \sum_i \mathcal{X}'_i \mathcal{X}_i \right) > \frac{\min_i \Lambda_{\min}(\Sigma_i)}{2 \max_i \mu_{\max}(B_i)} \right] \leq c_1 \exp[-c_2 \log p] \quad (16)$$

and for $n_i \gtrsim \max\{1, t^{-2}\} \log p$

$$\mathbb{P} \left[\frac{1}{n_i} \|\mathcal{X}_{\mathcal{A}_0} \Delta\|_F^2 \leq \alpha \|\Delta\|_F^2 - \tau_{n_i} \|\Delta\|_1^2 \right] \leq \exp\left\{-\frac{c}{2} n_i \min\{1, t^2\}\right\} \quad (17)$$

where $\alpha = \pi \mathbf{m}(f_{X^{(i)}})$, $t = \frac{\mathbf{m}(f_{X^{(i)}})}{54 \mathcal{M}(f_{X^{(i)}})}$, $\tau_{n_i} = \frac{3 \mathbf{m}(f_{X^{(i)}}) \log p^2}{cn_i \min\{1, t^2\}}$ and $c > 0$

proposition 2. Let $\mathcal{M} := \max_{i \in \mathcal{A}_0} \mathcal{M}(f_{X^{(i)}})$ and $\mathbf{m} := \min_{i \in \mathcal{A}_0} \mathbf{m}(f_{X^{(i)}})$. If $\mathbf{m} > 0$ and $\mathcal{M} < \infty$, then there exists universal constants $c_3, c_4, c_5 > 0$ such that

(1) For $n_{\mathcal{A}_0} \gtrsim \log p$

$$\mathbb{P} \left[\|\mathcal{X}'_i \mathcal{E}_i\|_\infty \geq \phi_{\mathcal{A}_0} c_3 \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}} \right] \leq c_4 \exp[-c_5 \log p]$$

(2) For $n_{\mathcal{A}_0} \gtrsim \max\{1, t^{-2}\} \log p$

$$\mathbb{P} \left[\frac{1}{n_{\mathcal{A}_0}} \|\mathcal{X}_{\mathcal{A}_0} \Delta\|_F^2 \leq \alpha_2 \|\Delta\|_F^2 - \tau_{n_{\mathcal{A}_0}} \|\Delta\|_1^2 \right] \leq \exp\left\{-\frac{c}{2} n_{\mathcal{A}_0} \min\{1, t^2\}\right\} \quad (18)$$

$$\mathbb{P} \left[\frac{1}{n_{\mathcal{A}_0}} \|\mathcal{X}_{\mathcal{A}_0} \Delta\|_F^2 \geq \alpha'_2 \|\Delta\|_F^2 + \tau_{n_{\mathcal{A}_0}} \|\Delta\|_1^2 \right] \leq \exp\left\{-\frac{c}{2} n_{\mathcal{A}_0} \min\{1, t^2\}\right\} \quad (19)$$

where $\alpha_2 = \pi \mathbf{m}$, $\alpha'_2 = 3\pi \mathcal{M}$, $t = \frac{\mathbf{m}}{54\mathcal{M}}$, $\tau_{n_{\mathcal{A}_0}} = \frac{3\mathbf{m} \log p^2}{cn_{\mathcal{A}_0} \min\{1, t^2\}}$ and $c > 0$.

Remark 2. The assumption $\mathbf{m}(f_{X^{(i)}}) > 0$ and $\mathcal{M}(f_{X^{(i)}}) < \infty$ are fairly mild and hold for stable, invertible ARMA processes. In our case, all auxiliary models are VAR model. Therefore, it is reasonable to assume that these models are uniformly bounded by some constant \mathbf{m} and \mathcal{M} .

C.1 Proof of Proposition 1

Proof of Proposition 1 (14) and (16) are simple modifications of Proposition 3 in Basu et al. (2019). Also, (17) is a special case of (18). So we omit their proofs. We next prove (15). Using Proposition 3.2 in Basu and Michailidis (2015), we know that

$$\mathbb{P} \left[\frac{1}{n_i} \|\mathcal{X}_i \mathcal{E}_i\|_\infty > c_0 \phi \eta \right] \leq 6p \exp[-cn_i \min\{\eta^2, \eta\}].$$

With the choice of $\eta = \sqrt{\frac{(1+c_1) \log p}{cN}}$, we arrive at

$$\mathbb{P} \left[\frac{1}{N} \|\mathcal{X}_i \mathcal{E}_i\|_\infty > c_0 \phi \frac{n_i}{N} \sqrt{\frac{\log p}{N}} \right] \leq 6p \exp[-(c_1 + 1) \frac{n_i \log p}{N}] = 6 \exp[-c_1 \frac{n_i \log p}{N}].$$

Taking the union set over i , we have

$$\begin{aligned} \mathbb{P} \left[\max_i \frac{1}{N} \|\mathcal{X}_i \mathcal{E}_i\|_\infty > c_0 \phi \sqrt{\frac{\log p}{N}} \right] &\leq \sum_i 6 \exp[-c_1 \frac{n_i \log p}{N}] \\ &\leq \sum_i 6 \frac{n_i}{N} \exp[-c_1 \log p] = 6 \exp[-c_1 \log p]. \end{aligned}$$

This implies that

$$\mathbb{P} \left[\max_i \frac{1}{N} \|\mathcal{X}'_i \mathcal{E}_i\|_\infty \geq c_0 \phi \sqrt{\frac{\log p}{N}} \right] \leq c_1 \exp[-c_2 \log p] \quad (20)$$

for some universal constants $c_i > 0$.

C.2 Proof of Proposition 2

Proof of Proposition 2. Let $\alpha_k = \frac{n_k}{n_{\mathcal{A}_0}}$. According to the definition of \mathcal{M} and \mathbf{m} , we have that

$$2\pi \mathcal{M} \geq \Lambda_{\max} \left(\sum_{k \in \mathcal{A}_0} \frac{n_k}{n_{\mathcal{A}_0}} \Gamma_k \right) \geq \Lambda_{\min} \left(\sum_{k \in \mathcal{A}_0} \frac{n_k}{n_{\mathcal{A}_0}} \Gamma_k \right) \geq 2\pi \mathbf{m}.$$

For every $u \in \mathbb{R}^{p^2}$, $\|u\|_2 \leq 1$, we obtain the following inequality from Lemma 2,

$$\mathbb{P} \left(\left| u' \left(\sum_{k \in \mathcal{A}_0} \alpha_k \left[\frac{1}{n_k} (I_p \otimes \mathcal{X}'_i \mathcal{X}_i) - I_p \otimes \Gamma_k \right] \right) u \right| \geq 2\pi \mathcal{M} t \right) \leq 2p \exp(-cn_{\mathcal{A}_0} \min\{t, t^2\}).$$

To simplify the notation, we define $\hat{\Gamma}^{\mathcal{A}_0} := \sum_{k \in \mathcal{A}_0} \alpha_k [\frac{1}{n_k} (I_p \otimes \mathcal{X}_i' \mathcal{X}_i)]$. Then the inequality can be simplified as

$$\mathbb{P} \left(\left| u' \left(\hat{\Gamma}^{\mathcal{A}_0} - I_p \otimes \sum_{k \in \mathcal{A}_0} \alpha_k' \Gamma_k \right) u \right| \geq 2\pi \mathcal{M} t \right) \leq 2p \exp(-cn_{\mathcal{A}_0} \min\{t, t^2\}).$$

Applying Supplementary Lemma F.2 in Basu and Michailidis (2015), we have

$$\begin{aligned} \mathbb{P} \left(\frac{1}{n_{\mathcal{A}_0}} \sup_{u \in \mathcal{K}(2s)} \left| u' \left(\hat{\Gamma}^{\mathcal{A}_0} - I_p \otimes \sum_{k \in \mathcal{A}_0} \alpha_k \Gamma_k \right) u \right| \geq 2\pi \mathcal{M} t \right) \\ \leq 2p \exp[-cn_{\mathcal{A}_0} \{t, t^2\} + 2s \min\{\log p^2, \log(\frac{21ep^2}{s})\}] \\ \leq 2 \exp(-cn_{\mathcal{A}_0} \min\{1, t^2\} + 3s \log p^2). \end{aligned} \quad (21)$$

Setting $t = \frac{\mathbf{m}}{54\mathcal{M}}$, we have

$$\sup_{u \in \mathcal{K}(2s)} \left| u' \left(\hat{\Gamma}^{\mathcal{A}_0} - I_p \otimes \sum_{k \in \mathcal{A}_0} \alpha_k \Gamma_k \right) u \right| \leq \frac{2\pi \mathbf{m}}{54}, \quad (22)$$

with probability greater than $1 - 2 \exp(-cn_{\mathcal{A}_0} \min\{1, t^2\} + 3s \log p^2)$. Given (22), applying supplementary Lemma 12 in Loh and Wainwright (2011), we have that for all $u \in \mathbb{R}^{p^2}$,

$$\left| u' \left(\hat{\Gamma}^{\mathcal{A}_0} - I_p \otimes \sum_{k \in \mathcal{A}_0} \alpha_k \Gamma_k \right) u \right| \leq \frac{2\pi \mathbf{m}}{2} \|u\|_2^2 + \frac{2\pi \mathbf{m}}{2s} \|u\|_1^2. \quad (23)$$

Note that

$$2\pi \mathcal{M} \|u\|_2^2 \geq u' \left(I_p \otimes \sum_{k \in \mathcal{A}_0} \alpha_k \Gamma_k \right) u \geq 2\pi \mathbf{m} \|u\|_2^2. \quad (24)$$

By (23) and (24), we have that

$$3\pi \mathcal{M} \|u\|_2^2 - \frac{\pi \mathbf{m}}{s} \|u\|_1^2 \geq u' \hat{\Gamma}^{\mathcal{A}_0} u \geq \pi \mathbf{m} \|u\|_2^2 - \frac{\pi \mathbf{m}}{s} \|u\|_1^2, \quad \text{for all } u \in \mathbb{R}^{p^2},$$

with probability greater than $1 - 2 \exp(-cn_{\mathcal{A}_0} \min\{1, t^2\} + 3s \log p^2)$. Setting $s = \frac{cn_{\mathcal{A}_0} \min\{1, t^2\}}{6 \log(p^2)}$ yields the final result.

lemma 5. Recall that $\bar{S} := (\sum_{i \in \mathcal{A}_0} \Gamma_i)^{-1} (\sum_{i \in \mathcal{A}_0} \Gamma_i S_i)$ and \tilde{S} defined in Algorithm 1. Under the assumptions of Theorem 2, we have

$$\begin{aligned} \|\tilde{S} - \bar{S}\|_F^2 &\lesssim \frac{s \log p}{n_{\mathcal{A}_0}} (1 \vee h^4) + \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}} h + \frac{s \log p + rp}{N} \\ \|\tilde{S} - \bar{S}\|_1 &\lesssim s \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}} (1 \vee h^2) + 2h + \frac{pr + s \log p}{N} \sqrt{\frac{n_{\mathcal{A}_0}}{\log p}} \end{aligned}$$

proof According to (4), we have

$$\frac{1}{2n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{Y}_i - \mathcal{X}_i(\hat{L} + \tilde{S})\|_F^2 + \mu \|\tilde{S}\|_1 \leq \frac{1}{2n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{Y}_i - \mathcal{X}_i(\hat{L} + \bar{S})\|_F^2 + \mu \|\bar{S}\|_1 \quad (25)$$

After some algebra

$$\begin{aligned}
 \frac{1}{2n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{X}_i(\tilde{S} - \bar{S})\|_F^2 &\leq \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \langle \tilde{S} - \bar{S}, \mathcal{X}'_i \mathcal{E}_i \rangle + \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \langle \mathcal{X}_i(\tilde{S} - \bar{S}), \mathcal{X}_i(L - \hat{L}) \rangle \\
 &\quad + \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \langle \mathcal{X}_i(\tilde{S} - \bar{S}), \mathcal{X}_i(S_i - \bar{S}) \rangle + \mu \|\bar{S}\|_1 - \mu \|\tilde{S}\|_1 \\
 &\leq \|\tilde{S} - \bar{S}\|_1 \left\| \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \mathcal{X}'_i \mathcal{E}_i \right\|_\infty + \frac{1}{4n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{X}_i(\tilde{S} - \hat{S})\|_F^2 + \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{X}_i(L - \hat{L})\|_F^2 \\
 &\quad + \|\tilde{S} - \bar{S}\|_1 \left\| \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} (S_i - \bar{S}) \mathcal{X}'_i \mathcal{X}_i \right\|_\infty + \mu \|\bar{S}\|_1 - \mu \|\tilde{S}\|_1
 \end{aligned} \tag{26}$$

According to lemma 4, we know that

$$\left\| \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}_0} (S_i - \bar{S}) \mathcal{X}'_i \mathcal{X}_i \right\|_\infty \leq c_{\mathcal{M}} \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}} (1 + h^2) \tag{27}$$

From proposition 1, we know that

$$\left\| \frac{1}{n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}_0} \mathcal{X}_i \mathcal{E}_i \right\|_\infty \leq c_0 \phi \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}}. \tag{28}$$

Inserting (27) and (28) into (26) and setting $\mu = 2(c_0 + c_{\Sigma})(1 \vee h^2) \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}}$, we have

$$\begin{aligned}
 \frac{1}{4n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{X}_i(\tilde{S} - \bar{S})\|_F^2 &\leq \frac{1}{2} \mu \|\tilde{S} - \bar{S}\|_1 + \mu \|\bar{S}\|_1 - \mu \|\tilde{S}\|_1 + \frac{pr + s \log p}{N} \\
 &\leq \frac{3\mu}{2} \|\tilde{S} - \bar{S}\|_{1, M_1} - \frac{\mu}{2} \|\tilde{S} - \bar{S}\|_{1, M_1^c} + 2\mu \|\bar{S}\|_{1, M_1^c} + \frac{pr + s \log p}{N} \\
 &\leq \frac{3\mu}{2} \|\tilde{S} - \bar{S}\|_{1, M_1} - \frac{\mu}{2} \|\tilde{S} - \bar{S}\|_{1, M_1^c} + 2\mu h + \frac{pr + s \log p}{N}
 \end{aligned} \tag{29}$$

(i) If $\|\tilde{S} - \bar{S}\|_{1, M_1} \geq 2h + \frac{pr + s \log p}{N\mu}$, we arrive at,

$$\frac{1}{4n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{X}_i(\tilde{S} - \bar{S})\|_F^2 \leq \frac{5\mu}{2} \|\tilde{S} - \bar{S}\|_{1, M_1} - \frac{\mu}{2} \|\tilde{S} - \bar{S}\|_{1, M_1^c} \tag{30}$$

This implies $\|\tilde{S} - \bar{S}\|_1 \leq 6\|\tilde{S} - \bar{S}\|_{1, M_1} \leq 6\sqrt{s} \|\tilde{S} - \bar{S}\|_F$. Using RE condition again, we have $\frac{1}{4n_{\mathcal{A}_0}} \sum_{i \in \mathcal{A}} \|\mathcal{X}_i(\tilde{S} - \bar{S})\|_F^2 \geq \frac{\alpha}{8} \|\tilde{S} - \bar{S}\|_F^2$. Using (30), we arrive at $\frac{\alpha}{8} \|\tilde{S} - \bar{S}\|_F^2 \leq \frac{5\mu}{2} \sqrt{s} \|\tilde{S} - \bar{S}\|_F$, which implies

$$\|\tilde{S} - \bar{S}\|_F^2 \lesssim \frac{s \log p}{n_{\mathcal{A}_0}} (1 \vee h^4), \quad \|\tilde{S} - \bar{S}\|_1 \lesssim s \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}} (1 \vee h^2) \tag{31}$$

(ii) If $\|\tilde{S} - \bar{S}\|_{1, M_1} \leq 2h + \frac{pr + s \log p}{N\mu}$, from (29) we know that $\|\tilde{S} - \bar{S}\|_{1, M_1^c} \lesssim 2h + \frac{pr + s \log p}{N\mu}$. This implies $\|\tilde{S} - \bar{S}\|_1 \lesssim h + \frac{pr + s \log p}{N\mu}$. Applying RSC condition again for (29), we arrive at,

$$\begin{aligned}
 \frac{\alpha}{4} \|\tilde{S} - \bar{S}\|_F^2 &\lesssim \frac{\log p}{n_{\mathcal{A}_0}} \left(h + \frac{pr + s \log p}{N\mu} \right)^2 + h\mu + \frac{pr + s \log p}{N} \\
 &\lesssim \frac{\log p}{n_{\mathcal{A}_0}} h^2 + h\mu + \frac{pr + s \log p}{N}
 \end{aligned} \tag{32}$$

Combining (i) and (ii), we have

$$\begin{aligned}
 \|\tilde{S} - \bar{S}\|_F^2 &\lesssim \frac{s \log p}{n_{\mathcal{A}_0}} (1 \vee h^4) + h\mu + \frac{pr + s \log p}{N} \\
 \|\tilde{S} - \bar{S}\|_1 &\lesssim s \sqrt{\frac{\log p}{n_{\mathcal{A}_0}}} (1 \vee h^2) + h + \frac{pr + s \log p}{N} \sqrt{\frac{n_{\mathcal{A}_0}}{\log p}}.
 \end{aligned} \tag{33}$$

lemma 6. Define $\delta = S_0 - \bar{S}$. Under the assumptions of Theorem 2, we have

$$\begin{aligned}\|\tilde{\delta} - \delta\|_F^2 &\lesssim h\sqrt{\frac{\log p}{n_0}} \wedge h^2 + (1 \vee h^4) \frac{s \log p}{n_{\mathcal{A}_0}} + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{n_0 N^2} \\ \|\tilde{\delta} - \delta\|_1 &\lesssim h + (1 \vee h^4) \frac{s\sqrt{n_0 \log p}}{n_{\mathcal{A}_0}} + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{\sqrt{n_0 \log p} N^2}.\end{aligned}$$

Remark 3. As shown in Theorem 1, the upper bound of estimation error for S_0 is $\frac{s \log p + rp}{n_0}$ without transfer learning algorithm. Under the above assumption, considering other informative set improves estimation result when $h = o(\frac{s \log p + rp}{\sqrt{n_0 \log p}} \vee (\frac{n_{\mathcal{A}_0}}{n_0})^{\frac{1}{4}})$. For traditional lasso method, the estimation rate is $O_p(\frac{s \log p}{n_0})$. Transfer learning has a faster convergence rate when $h \lesssim s\sqrt{\frac{\log(p)}{n_0}}$ and $N \gtrsim \sqrt{\frac{n_{\mathcal{A}_0}}{s \log p}}(pr + s \log p)$.

Proof According to (5), we have

$$\frac{1}{2n_0} \|\mathcal{Y}_i - \mathcal{X}_i(\hat{L} + \tilde{S} + \tilde{\delta})\|_F^2 + \lambda_\delta \|\tilde{\delta}\|_1 \leq \frac{1}{2n_0} \|\mathcal{Y}_i - \mathcal{X}_i(\hat{L} + \tilde{S} + \delta)\|_F^2 + \lambda_\delta \|\delta\|_1 \quad (34)$$

Similar to (26) and (28), we have

$$\frac{1}{4n_0} \|\mathcal{X}_1(\tilde{\delta} - \delta)\|_F^2 \leq 2\lambda_\delta \|\delta\|_1 - \frac{\lambda_\delta}{2} \|\tilde{\delta} - \delta\|_1 + \frac{pr + s \log p}{N} + \frac{2}{n_0} \|\mathcal{X}(\tilde{S} - \bar{S})\|_F^2 \quad (35)$$

If (i) $\|\tilde{S} - \bar{S}\|_{1, M_1} \geq 2h + \frac{pr + s \log p}{N\mu}$, we know that $\|\tilde{S} - \bar{S}\|_1 \leq 6\sqrt{s} \|\tilde{S} - \bar{S}\|_F$ from lemma 5. Using RSC condition, we have

$$\begin{aligned}\frac{2}{n_0} \|\mathcal{X}_1(\tilde{S} - \bar{S})\|_F^2 &\lesssim \|\tilde{S} - \bar{S}\|_F^2 + \frac{\log p}{n_0} \|\tilde{S} - \bar{S}\|_1^2 \\ &\lesssim (1 + 36s \frac{\log p}{n_{\mathcal{A}_0}}) \|\tilde{S} - \bar{S}\|_F^2\end{aligned} \quad (36)$$

If (ii) $\|\tilde{S} - \bar{S}\|_{1, M_1} \leq 2h + \frac{pr + s \log p}{N\mu}$, using RE condition again we have

$$\begin{aligned}\frac{2}{n_0} \|\mathcal{X}_1(\tilde{S} - \bar{S})\|_F^2 &\lesssim \|\tilde{S} - \bar{S}\|_F^2 + \frac{\log p}{n_0} (2h + \frac{pr + s \log p}{N\mu})^2 \\ &\lesssim \|\tilde{S} - \bar{S}\|_F^2 + \frac{\log p}{n_0} h^2 + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{N^2 n_0}\end{aligned} \quad (37)$$

Inserting (i) and (ii) into (35), we arrive at

$$\begin{aligned}\frac{1}{4n_0} \|\mathcal{X}_1(\tilde{\delta} - \delta)\|_F^2 \vee \frac{\lambda_\delta}{2} \|\tilde{\delta} - \delta\|_1 &\lesssim 2h\sqrt{\frac{\log p}{n_0}} + \|\tilde{S} - \bar{S}\|_F^2 + \frac{\log p}{n_0} h^2 + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{N^2 n_0} \\ &\lesssim 2h\sqrt{\frac{\log p}{n_0}} + \frac{s \log p}{n_{\mathcal{A}_0}} (1 \vee h^4) + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{N^2 n_0}\end{aligned} \quad (38)$$

Using RE condition again for $\mathcal{X}_1(\tilde{\delta} - \delta)$, we have

$$\|\tilde{\delta} - \delta\|_F^2 \lesssim h\sqrt{\frac{\log p}{n_0}} + \frac{s \log p}{n_{\mathcal{A}_0}} (1 \vee h^4) + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{N^2 n_0} \quad (39)$$

Inserting $\|\tilde{\delta} - \delta\|_F \leq \|\tilde{\delta} - \delta\|_1$ into (38), we have

$$\begin{aligned}\|\tilde{\delta} - \delta\|_F^2 &\lesssim h^2 + (\frac{s \log p}{n_{\mathcal{A}_0}} (1 \vee h^4))^2 / \lambda_\delta + (\frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{N^2 n_0})^2 / \lambda_\delta \\ &< h^2 + \frac{s \log p}{n_{\mathcal{A}_0}} (1 \vee h^4) + \frac{n_{\mathcal{A}_0}(pr + s \log p)^2}{N^2 n_0}\end{aligned} \quad (40)$$

Using (39) and (40) yields the final result.

D Proof of Theorems

D.1 Proof of Theorem 1

Proof According to (3), we have

$$\sum_i \frac{1}{N} \|\mathcal{Y}_i - \mathcal{X}_i B_i\|_F^2 + \lambda \|\widehat{L}\|_* + \sum_i \mu_i \sqrt{\frac{n_i}{N}} \|\widehat{S}_i\|_1 \leq \sum_i \frac{1}{N} \|\mathcal{Y}_i - \mathcal{X}_i B_i\|_F^2 + \lambda \|L\|_* + \sum_i \mu_i \sqrt{\frac{n_i}{N}} \|S_i\|_1 \quad (41)$$

Let $\Delta^L = \widehat{L} - L$ and $\Delta_i^S = \widehat{S}_i - S_i$. Using $\mathcal{Y}_i = \mathcal{X}_i(L + S_i) + \mathcal{E}_i$ and simple algebra, we have

$$\begin{aligned} \frac{1}{N} \sum_i \|\mathcal{X}_i(\Delta^L + \Delta_i^S)\|_F^2 &\leq \frac{2}{N} \sum_i \langle \widehat{\Delta}^L + \widehat{\Delta}_i^S, \mathcal{X}_i \mathcal{E}_i \rangle + \lambda \|L\|_* + \sum_i \sqrt{\frac{n_i}{N}} \mu_i \|S_i\|_1 - \lambda \|\widehat{L}\|_* - \sum_i \sqrt{\frac{n_i}{N}} \mu_i \|\widehat{S}_i\|_1 \\ &\leq \sum_i \frac{2}{N} \|\Delta_i^S\|_1 \cdot \|\mathcal{X}_i' \mathcal{E}_i\|_\infty + \|\widehat{\Delta}^L\|_* \left\| \frac{2}{N} \sum_i \mathcal{X}_i' \mathcal{E}_i \right\|_2 + \lambda (\|\widehat{\Delta}_A^L\|_* - \|\widehat{\Delta}_B^L\|_*) \\ &\quad + \sum_i \mu_i \sqrt{\frac{n_i}{N}} (\|\Delta_i^S\|_{1, M_i} - \|\Delta_i^S\|_{1, M_i^c}) \\ &\leq \sum_i \frac{\mu_i}{2} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_1 + \frac{\lambda}{2} \|\Delta^L\|_* + \lambda (\|\Delta_A^L\|_* - \|\Delta_B^L\|_*) \\ &\quad + \sum_i \mu_i \sqrt{\frac{n_i}{N}} (\|\Delta_i^S\|_{1, M_i} - \|\Delta_i^S\|_{1, M_i^c}) \\ &= \frac{3\lambda}{2} \|\Delta_A^L\|_* - \frac{\lambda}{2} \|\Delta_B^L\|_* + \sum_i \frac{3\mu_i}{2} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1, M_i} - \sum_i \frac{\mu_i}{2} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1, M_i^c} \end{aligned} \quad (42)$$

where the matrices $(A, B) \in \{(A, B) : AB' = 0 \text{ and } A'B = 0\}$, M_i and M_i^c corresponds to non-zero entries and zero entries of matrices L_i separately. The second inequality derives from Lemma 1 in Agarwal et al. (2012) and lemma 2.3 in Recht et al. (2010). The third inequality derives from Proposition 1, $\mu_i = 2c_0\phi\sqrt{\frac{\log p}{N}} + \theta$ and $\lambda = 2c_0\phi\sqrt{\frac{p}{N}}$. Now, (42) implies

$$\lambda \|\Delta_B^L\|_* + \sum_i \mu_i \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1, M_i} \leq 3\lambda \|\Delta_A^L\|_* + \sum_i \mu_i \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1, M_i^c} \quad (43)$$

Using RSC conditions and $\tau' \leq \tau$, we have

$$\begin{aligned}
 \sum_i \frac{1}{N} \|\mathcal{X}_i(\Delta^L + \Delta_i^S)\|_F^2 &= \sum_i \frac{n_i}{N} \frac{1}{n_i} \|\mathcal{X}_i(\Delta^L + \Delta_i^S)\|_F^2 \\
 &= \sum_i \frac{1}{N} \|\mathcal{X}_i \Delta^L\|_F^2 + \sum_i \frac{n_i}{N} \frac{1}{n_i} \|\mathcal{X}_i \Delta_i^S\|_F^2 + \sum_i \frac{n_i}{N} \langle \Delta^L, \frac{1}{n_i} \mathcal{X}_i \mathcal{X}_i' \Delta_i^S \rangle \\
 &\geq \alpha \|\Delta^L\|_F^2 - \tau' \|\Delta^L\|_*^2 + \sum_i \frac{n_i}{N} (\|\Delta_i^S\|_F^2 - \tau \frac{\mu_i^2}{\lambda^2} \|\Delta_i^S\|_1^2) \\
 &\quad - \|\Delta^L\|_\infty (\max_i \|\Gamma_i\|_\infty + \phi \sqrt{\frac{\log p}{n_i}}) (\sum_i \frac{n_i}{N} \|\Delta_i^S\|_1) \\
 &\geq \alpha \|\Delta^L\|_F^2 + \alpha \sum_i \frac{n_i}{N} \|\Delta_i^S\|_F^2 - 2 \|\Delta^L\|_\infty (\sum_i \frac{n_i}{N} \|\Delta_i^S\|_1) \\
 &\quad - 2\tau \|\Delta_A^L\|_*^2 - 2\tau \|\Delta_B^L\|_*^2 - 2\tau \sum_i \frac{n_i}{N} \frac{\mu_i^2}{\lambda^2} \|\Delta_i^S\|_{1,M_i}^2 - 2\tau \sum_i \frac{n_i}{N} \frac{\mu_i^2}{\lambda^2} \|\Delta_i^S\|_{1,M_i^c}^2 \\
 &\geq \alpha \|\Delta^L\|_F^2 + \alpha \sum_i \frac{n_i}{N} \|\Delta_i^S\|_F^2 - 2 \|\Delta^L\|_\infty (\sum_i \frac{n_i}{N} \|\Delta_i^S\|_1) \\
 &\quad - 2\tau (\|\Delta_A^L\|_* + \sum_i \frac{\mu_i}{\lambda} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1,M_i})^2 - 2\tau (\|\Delta_B^L\|_* + \sum_i \frac{\mu_i}{\lambda} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1,M_i^c})^2,
 \end{aligned} \tag{44}$$

where the first inequality comes from lemma 3 with the choice of $t = \sqrt{\frac{\log p}{n_i}}$, $\|\frac{1}{n_i} \mathcal{X}_i \mathcal{X}_i'\|_\infty \leq \|\Gamma_i\|_\infty + \phi \sqrt{\frac{\log p}{n_i}}$. Combining (43) and (44), we arrive at

$$\begin{aligned}
 \sum_i \frac{1}{N} \|\mathcal{X}_i(\Delta^L + \Delta_i^S)\|_F^2 &\geq \alpha \|\Delta^L\|_F^2 + \alpha \sum_i \frac{n_i}{N} \|\Delta_i^S\|_F^2 - 2 \|\Delta^L\|_\infty (\sum_i \frac{n_i}{N} \|\Delta_i^S\|_1) \\
 &\quad - 20\tau (\|\Delta_A^L\|_* + \sum_i \frac{\mu_i}{\lambda} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1,M_i})^2
 \end{aligned} \tag{45}$$

Since Δ_A^L has rank at most $2r$ and $M_i \leq s$, we have

$$\begin{aligned}
 \|\Delta_A^L\|_* + \sum_i \frac{\mu_i}{\lambda} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1,M_i} &\leq \sqrt{2r} \|\Delta^L\|_F + \sum_i \frac{\mu_i}{\lambda} \sqrt{s} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_F \\
 &\leq (\|\Delta\|_F^2 + \sum_i \frac{n_i}{N} \|\Delta_i^S\|_F^2)^{\frac{1}{2}} (2r + s \sum_i \frac{\mu_i}{\lambda})^{\frac{1}{2}}
 \end{aligned} \tag{46}$$

Recall that $\mu_i = 2c_0\phi\sqrt{\frac{\log p}{N}} + \theta$, $\lambda = 2c_0\phi\sqrt{\frac{p}{N}}$ and $\theta = o(\sqrt{\frac{p}{N}})$. Thus $\sum_i \frac{\mu_i}{\lambda} = o(1)$. With (45) and (46), we arrive at

$$\sum_i \frac{1}{N} \|\mathcal{X}_i(\Delta^L + \Delta_i^S)\|_F^2 \geq \frac{\alpha}{2} \|\Delta^L\|_F^2 + \frac{\alpha}{2} \sum_i \frac{n_i}{N} \|\Delta_i^S\|_F^2 - 2 \|\Delta^L\|_\infty (\sum_i \frac{n_i}{N} \|\Delta_i^S\|_1) \tag{47}$$

Inserting (47) into (42), we have

$$\begin{aligned}
 \frac{\alpha}{2} \|\Delta^L\|_F^2 + \frac{\alpha}{2} \sum_i \frac{n_i}{N} \|\Delta_i^S\|_F^2 &\leq \frac{3\lambda}{2} \|\Delta_A^L\|_* + \sum_i \frac{5\mu_i}{2} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_{1,M_i} \\
 &\leq \frac{3\lambda}{2} \sqrt{2r} \|\Delta^L\|_F + \sum_i \frac{5\mu_i}{2} \sqrt{s} \sqrt{\frac{n_i}{N}} \|\Delta_i^S\|_F \\
 &\leq \sqrt{(3\lambda\sqrt{2r})^2 + (5\mu\sqrt{s})^2} \cdot \sqrt{\frac{1}{2} \|\Delta^L\|_F^2 + \frac{1}{2} \sum_i \frac{n_i}{N} \|\Delta_i^S\|_F^2}
 \end{aligned} \tag{48}$$

Using $\theta = o(\sqrt{\frac{p}{N}})$ yields the final result.

D.2 Proof of Theorem 2

Proof Using lemma 5 and lemma 6 yields Theorem 2 directly.

D.3 Proof of Theorem 3

Proof Assumption 1 implies $h = O(1)$. Using lemma 5, we have

$$\|\widehat{S}_k - (S_0 + \tilde{\delta}_k)\|_2^2 \lesssim \frac{s \log(p)}{n_0/2 + n_k} + \sqrt{\frac{\log(p)}{n_0/2 + n_k}} + \frac{s \log p + rp}{N} \quad (49)$$

$$\|\widehat{S}_k - (S_0 + \tilde{\delta}_k)\|_1 \lesssim s \sqrt{\frac{\log(p)}{n_0/2 + n_k}} + h + \frac{pr + s \log p}{N} \sqrt{\frac{n_0/2 + n_k}{\log p}} \quad (50)$$

where

$$\begin{aligned} \tilde{\delta}_k &= [(\alpha_0 \Gamma_0 + \alpha_k \Gamma_k)]^{-1} [\alpha_k \Gamma_k \delta^{(k)}], \\ \alpha_0 &= \frac{n_0/2}{n_0/2 + n_k}, \alpha_k = \frac{n_k}{n_0/2 + n_k} \end{aligned}$$

We can see that $\|\tilde{\delta}^{(k)}\|_2^2 \asymp \|\delta^{(k)}\|_2^2$ and $\|\tilde{\delta}^{(k)}\|_1 \leq C_1 h$, where C depends on Σ_k and $\Sigma^{(0)}$. From assumption 1 and equation (50), we know that $\|\widehat{S}_k - S_0\|_1$ is bounded by C_2 , C_2 depends on Σ_k and Σ_1 . We can prove $\|\widehat{S}_{0,\mathcal{I}} - S_0\|_1$ is bounded by $s \sqrt{\frac{\log(p^2)}{n_0/2}} + \frac{pr + s \log p}{N} \sqrt{\frac{n_0/2 + n_k}{\log p}}$ for the same reason as (50). With the boundness of $\|\widehat{S}_{0,\mathcal{I}} - S_0\|_1$ and $\|\widehat{S}_k - S_0\|_1$, we have that $\|\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k\|_1 \leq CM$.

According to the definition of $R^{(k)}$ and $R_1^{(0)}$,

$$\begin{aligned} R^{(k)} - R_1^{(0)} &= \|\mathcal{Y}_{0,\mathcal{I}^c} - \mathcal{X}_{0,\mathcal{I}^c}(\widehat{L} + \widehat{S}_k)\|_F^2 - \|\mathcal{Y}_{0,\mathcal{I}^c} - \mathcal{X}_{0,\mathcal{I}^c}(\widehat{L} + \widehat{S}_{0,\mathcal{I}})\|_F^2 \\ &\leq 2 \langle \mathcal{E}'_{0,\mathcal{I}^c} \mathcal{X}_{0,\mathcal{I}^c}, \widehat{S}_{0,\mathcal{I}} - \widehat{S}_k \rangle + 2 \langle (\mathcal{X}_{0,\mathcal{I}^c})(S_0 - \widehat{S}_{0,\mathcal{I}}), (\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k) \rangle \\ &\quad + 2 \langle (\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k), (\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k) \rangle + 2 \|\mathcal{X}_{0,\mathcal{I}^c}(\widehat{L} - L)\|_F^2 \end{aligned}$$

Using Proposition 1 for the first term, we know that

$$\begin{aligned} \langle \mathcal{E}'_{0,\mathcal{I}^c} \mathcal{X}_{0,\mathcal{I}^c}, \widehat{S}_{0,\mathcal{I}} - \widehat{S}_k \rangle &\leq \|(\mathcal{E}_{0,\mathcal{I}^c})' \mathcal{X}_{0,\mathcal{I}^c}\|_\infty \|\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k\|_1 \\ &\lesssim CM \sqrt{\frac{\log(p^2)}{n_0/2}} \end{aligned}$$

with probability greater than $1 - O(p^{-2})$.

For the second term, we have

$$\begin{aligned} \left| \langle (\mathcal{X}_{0,\mathcal{I}^c})(S_0 - \widehat{S}_{0,\mathcal{I}}), (\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k) \rangle \right| &\leq 2 \langle (\mathcal{X}_{0,\mathcal{I}^c})(S_0 - \widehat{S}_{0,\mathcal{I}}), (\mathcal{X}_{0,\mathcal{I}^c})(S_0 - \widehat{S}_{0,\mathcal{I}}) \rangle \\ &\quad + \frac{1}{2} \langle (\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k), (\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k) \rangle \end{aligned}$$

For the last term, $\|\mathcal{X}_{0,\mathcal{I}^c}(\widehat{L} - L)\|_F^2 \leq \frac{s \log p + rp}{N}$ Therefore,

$$R^{(k)} - R_1^{(0)} \lesssim \|(\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k)\|_F^2 + \|(\mathcal{X}_{0,\mathcal{I}^c})(S_0 - \widehat{S}_{0,\mathcal{I}})\|_F^2 + \sqrt{\frac{\log p}{n_0/2}} + \frac{s \log p + rp}{N} \quad (51)$$

Similarly, we have

$$R^{(k)} - R_1^{(0)} \gtrsim \|(\mathcal{X}_{0,\mathcal{I}^c})(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k)\|_F^2 + \|(\mathcal{X}_{0,\mathcal{I}^c})(S_0 - \widehat{S}_{0,\mathcal{I}})\|_F^2 + \sqrt{\frac{\log p}{n_0/2}} + \frac{s \log p + rp}{N} \quad (52)$$

Using Proposition 1 and the boundedness of $\|S_0 - \widehat{S}_{0,\mathcal{I}}\|_1$, $\|\widehat{S}_k - \widehat{S}_{0,\mathcal{I}}\|_1$, we have with probpability greater than $1 - O(\exp(-n_0))$

$$\begin{aligned}
 \|\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k\|_F^2 - M^2 \frac{\log p}{n_0/2} - \frac{(s \log p + rp)^2}{N^2} &\lesssim \|\mathcal{X}_{0,\mathcal{I}^c}(\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k)\|_F^2 \\
 &\lesssim \|\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k\|_F^2 + M^2 \frac{\log p}{n_0/2} + \frac{(s \log p + rp)^2}{N^2} \\
 \|S_0 - \widehat{S}_{0,\mathcal{I}}\|_F^2 - M^2 \frac{\log p}{n_0/2} - \frac{(s \log p + rp)^2}{N^2} &\lesssim \|\mathcal{X}_{0,\mathcal{I}^c}(S_0 - \widehat{S}_{0,\mathcal{I}})\|_F^2 \\
 &\lesssim \|S_0 - \widehat{S}_{0,\mathcal{I}}\|_F^2 + M^2 \frac{\log p}{n_0/2} + \frac{(s \log p + rp)^2}{N^2}.
 \end{aligned} \tag{53}$$

Plugging (53) in (51) and (52), we arrive at

$$\begin{aligned}
 &\frac{1}{2}C_0\|\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k\|_2^2 - 2C_0\|S_0 - \widehat{S}_{0,\mathcal{I}}\|_2^2 - \sqrt{\frac{\log(p)}{n_0/2}} - \frac{s \log p + rp}{N} \\
 &\leq R^{(k)} - R_1^{(0)} \leq \\
 &\frac{3}{2}C_1\|\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k\|_2^2 + 2C_1\|S_0 - \widehat{S}_{0,\mathcal{I}}\|_2^2 + \sqrt{\frac{\log(p)}{n_0/2}} + \frac{s \log p + rp}{N}
 \end{aligned}$$

with probability greater than $1 - O(p^{-2})$. From $\frac{n_0(pr+s \log p)}{N \log p} = o(1)$, we know that $\sqrt{\frac{\log p}{n_0/2}} \gtrsim \frac{s \log p + rp}{N}$. Note that $\|S_0 - \widehat{S}_{0,\mathcal{I}}\|_F \leq s\sqrt{\frac{\log p}{n_0}}$ and $\|\widehat{S}_{0,\mathcal{I}} - \widehat{S}_k\|_F \leq \|\tilde{\delta}^{(k)}\|_F + \|S_0 - \widehat{S}_{0,\mathcal{I}}\|_F$. The upper bound for $R^{(k)} - R_1^{(0)}$ is

$$R^{(k)} - R_1^{(0)} \lesssim \|\tilde{\delta}^{(k)}\|_F^2 + \sqrt{\frac{\log(p^2)}{n_0/2}} \asymp \|\delta^{(k)}\|_F^2 + \sqrt{\frac{\log p}{n_0/2}}$$

The lower bound for $R^{(k)} - R_1^{(0)}$ is

$$R^{(k)} - R_1^{(0)} \gtrsim \|\tilde{\delta}^{(k)}\|_F^2 + \sqrt{\frac{\log(p^2)}{n_0/2}} \asymp \|\delta^{(k)}\|_F^2 + \sqrt{\frac{\log p}{n_0/2}}$$

Taking the union bound over $1 \leq k \leq K$ yields the final result.

E Additional Details For The Algorithms

E.1 Informative Set Selection Algorithm

Shown in Algorithm 2

E.2 Inference for Model Parameter Algorithm

The first step is debiasing \widehat{S}_{tran} . The explicit form of debiased estimator is as below,

$$\widehat{S}^{on} = \widehat{S}_{tran} + \frac{1}{n_0} \sum_{i=1}^{n_0} M_i X_i^{(0)} (X_{i+1}^{(0)} - X_i^{(0)} (\widehat{L} + \widehat{S}_{tran}))'.$$

M_i is called debiasing matrix and needs to be estimated by target model. If observations are i.i.d, setting $M_1 = M_2 \cdots = M_{n_0}$ is an effective way to debias \widehat{L} (Javanmard and Montanari, 2014). However, for VAR model, the existence of dependency destroy the asymptotic normality. To fix this problem, Deshpande et al. (2021) estimate M_i by the past observations, $\{X_t\}_{t < i}$, which makes M_i predictable. The term ‘‘online’’ comes from the imposed predictability.

Algorithm 2 : Selecting Informative Set

Input : observations from target model and auxiliary model $\{X_t^{(k)}\}, k = 0, \dots, K$.
penalty parameters $\lambda_k, k = 1, \dots, K$. Some constant $c > 0$.
low-rank matrix estimator \hat{L} from algorithm 1

Output : Informative set \hat{A} .

Step 1 Split the target data into two parts $X_{0,\mathcal{I}}$ and X_{0,\mathcal{I}^c} , where $\mathcal{I} = \{1, 2, \dots, n_0/2\}$, $\mathcal{I}^c = \{n_0/2 + 1, \dots, n_0\}$.

Step 2 For each $k \in \{1, \dots, K\}$, compute

$$\hat{S}_k = \underset{S}{\operatorname{argmin}} \frac{1}{n_k + n_0/2} \left(\|\mathcal{Y}_{0,\mathcal{I}} - \mathcal{X}_{0,\mathcal{I}}(\hat{L} + S)\|_2^2 + \|\mathcal{Y}_k - \mathcal{X}_k(\hat{L} + S)\|_2^2 \right) + \lambda_k \|S\|_1$$

Step 3 For $k = 0$, compute

$$\begin{aligned} \hat{S}_{0,\mathcal{I}} &= \underset{S}{\operatorname{argmin}} \frac{1}{n_0/2} \|\mathcal{Y}_{0,\mathcal{I}} - \mathcal{X}_{0,\mathcal{I}}(\hat{L} + S)\|_2^2 + \lambda_0 \|S\|_1 \\ \hat{S}_{0,\mathcal{I}^c} &= \underset{S}{\operatorname{argmin}} \frac{1}{n_0/2} \|\mathcal{Y}_{0,\mathcal{I}^c} - \mathcal{X}_{0,\mathcal{I}^c}(\hat{L} + S)\|_2^2 + \lambda_0 \|S\|_1 \end{aligned}$$

Step 4 For $k \in \{1, \dots, K\}$, $R^{(k)} = \|\mathcal{Y}_{0,\mathcal{I}^c} - \mathcal{X}_{0,\mathcal{I}^c}(\hat{L} + \hat{S}_k)\|_2^2$.

For $k = 0$, $R_1^{(0)} = \|\mathcal{Y}_{0,\mathcal{I}^c} - \mathcal{X}_{0,\mathcal{I}^c}(\hat{L} + \hat{S}_{0,\mathcal{I}})\|_2^2$ and $R_2^{(0)} = \|\mathcal{Y}_{0,\mathcal{I}^c} - \mathcal{X}_{0,\mathcal{I}^c}(\hat{L} + \hat{S}_{0,\mathcal{I}^c})\|_2^2$

Step 5 $\hat{\mathcal{A}} = \{k \in \{1, 2, \dots, K\} : R^{(k)} - R_1^{(0)} \leq c|R_1^{(0)} - R_2^{(0)}|\}$

We next introduce the first step specifically. We split processed target data $\{X_t\}_{t=1}^{n_0}$ into ℓ segments

$$\{X_t^{(0)}\}_{t \in E_0}, \{X_t^{(0)}\}_{t \in E_1}, \dots, \{X_t^{(0)}\}_{t \in E_{\ell-1}}$$

where $E_i := \{m_i + 1, m_i + 2, \dots, m_{i+1}\}$ and $0 = m_0 < m_1 < \dots < m_\ell = n_0$. Define the length of E_i as r_i , $r_i := m_{i+1} - m_i$. Define the sample covariance of the observations in the first j segments,

$$\hat{\Sigma}^{(j)} := \frac{1}{m_j} \sum_{t \in E_0 \cup \dots \cup E_{j-1}} X_t^{(0)} (X_t^{(0)})', \quad j = 1, \dots, \ell$$

The matrix $M^{(j)} = M_j$ is a $p \times p$ matrix. The a -th row of $M^{(j)}$ denoted as $\mathbf{m}_a^{(j)}$, $a = 1, \dots, p$, is the solution of the following optimization:

$$\begin{aligned} &\text{minimize} \quad \mathbf{m}_a^{(j)} \hat{\Sigma}^{(j)} (\mathbf{m}_a^{(j)})' \\ &\text{subject to} \quad \|\hat{\Sigma}^{(j)} (\mathbf{m}_a^{(j)})' - e_a\|_\infty \leq \mu_j, \quad \|\mathbf{m}_a^{(j)}\|_1 \leq C \end{aligned}$$

for appropriate values of $\mu_j, C > 0$. Then constructing \hat{S}^{on} as below,

$$\hat{S}^{on} = \hat{S}_{tran} + \frac{1}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} M^{(j)} X_t^{(0)} (X_{t+1}^{(0)} - X_t^{(0)} (\hat{L} + \hat{S}_{tran}))' \quad (54)$$

r_0 is selected to be $\sqrt{n_0}$ and $r_j \asymp \alpha^j$ for some constant $\alpha > 1$.

The second step is constructing the variance of \hat{S}^{on} . Corollary 3.7 in Deshpande et al. (2021) shows that the (a, i) -th entry of \hat{S}^{on} has the following variance

$$V_{a,i} = \frac{(\Sigma_\epsilon^{(0)})_{i,i}}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} (\mathbf{m}_a^{(j)} X_t^{(0)})^2$$

where $a \in \{1, 2, \dots, p\}$ and $i \in \{1, 2, \dots, p\}$. Therefore, we can use the scaled residual $\sqrt{n_0}(\hat{S}_{a,i}^{on} - S_{a,i}^{(0)})/\sqrt{V_{a,i}}$ as the test statistics and construct entry-wise confidence intervals accordingly. Theorem is given as below.

Corollary 1. Let \hat{S}^{on} be the debiased estimator (54) with $\mu_j = C_1 \omega \sqrt{\log p / m_j}$ and \hat{S} derived from Algorithm 1. Let $C = C_0 \|\Omega\|_1$ for an arbitrary constant $C_0 \geq 1$, where $\Omega = (\mathbb{E}[X_t^{(0)}(X_t^{(0)})'])^{-1}$. For any fixed sequence of integers $a(n) \in \{1, \dots, p\}$, define the conditional variance V_n as

$$V_{a,i} = \frac{(\Sigma_\epsilon^{(0)})_{i,i}}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} (\mathbf{m}_a^{(j)} X_t^{(1)})^2 \quad (55)$$

Assume that $\|\Omega\|_1 = o(\sqrt{n_0}/\log p)$, $h = o(s\sqrt{\log(p)/n_0})$, $\frac{n_0(s \log p + rp)}{N} = o(1)$. For any fixed coordinate $a \in \{1, \dots, p\}$, $i \in \{1, 2, \dots, p\}$ and for all $x \in \mathbb{R}$, we have

$$\lim_{n \rightarrow \infty} \left| \mathbb{P}\left\{ \frac{\sqrt{n_0}(\hat{S}_{a,i}^{on} - S_{a,i}^{(0)})}{\sqrt{V_{a,i}}} \leq x \right\} - \Phi(x) \right| = 0,$$

where Φ is the standard Gaussian cdf.

Remark 4. For getting the conditional variance V_n , Corollary requires the knowledge of $\Sigma_\epsilon^{(0)}$. Typically, $\Sigma_\epsilon^{(0)}$ is estimated by the training error. Since transfer learning improves the estimation accuracy and lowers the training error, confidence intervals constructed with transfer learning are always narrower than confidence intervals constructed with typical lasso method without losing any confidence. More detailed comparison is provided in simulation experiments section.

Proof of Corollary 1. The proof is similar to Corollary 3.7 in Deshpande et al. (2021), so we omit some details. We rewrite (54) as

$$\begin{aligned} \hat{S}^{on} - S_0 &= \underbrace{\left(I - \frac{1}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} M^{(j)} X_t^{(0)} (X_t^{(0)})' \right)}_{\Delta} (\hat{S} - S_0) + \underbrace{\frac{1}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} M^{(j)} X_t^{(0)} (\epsilon_t^{(0)})'}_W \\ &\quad + \underbrace{\frac{1}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} M^{(j)} X_t^{(0)} (X_t^{(0)})' (\hat{L} - L)}_U \end{aligned}$$

From Theorem 3.4 in Deshpande et al. (2021), we have that $\mathbb{P}(\|\Delta\|_\infty \geq C s \frac{\log(p)}{n_0}) \leq (p)^{-4}$ which implies that $\sqrt{n_0} \Delta$ is negligible with high probability. Lemma 3.6 in Deshpande et al. (2021) shows that $\sqrt{n_0} W$ is a martingale with variance equal to $V_{a,i}$. U can be decomposed as

$$\underbrace{\frac{1}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} M^{(j)} X_t^{(0)} (X_t^{(0)})' (\hat{L} - L)}_U = \underbrace{\left(\frac{1}{n_0} \sum_{j=1}^{\ell} \sum_{t \in E_j} M^{(j)} X_t^{(0)} (X_t^{(0)})' - I \right)}_{U_1} (\hat{L} - L) + \underbrace{(\hat{L} - L)}_{U_2}$$

Similar to Δ , we have that $U_1 \leq \sqrt{\frac{\log p}{n_0}} \sqrt{\frac{s \log p + rp}{N}}$ with high probability. We also have $\|U_2\|_\infty \leq \sqrt{\frac{s \log p + rp}{N}}$. Thus $\|\sqrt{n_0} U\|_\infty \leq \sqrt{\frac{n_0(s \log p + rp)}{N}} = o(1)$. The final result derives from Martingale central limit Theorem in Hall and Heyde (2014), i.e. Corollary 3.2. \square

F Additional Numerical Results and Considerations

F.1 Figure for Section 4

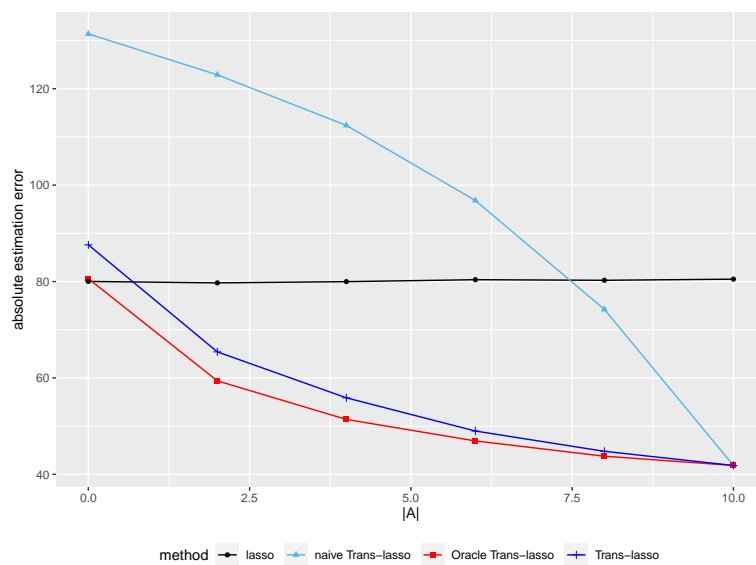


Figure 3: Absolute Estimation Error for Sparse Matrix

F.2 Figure for Section 5

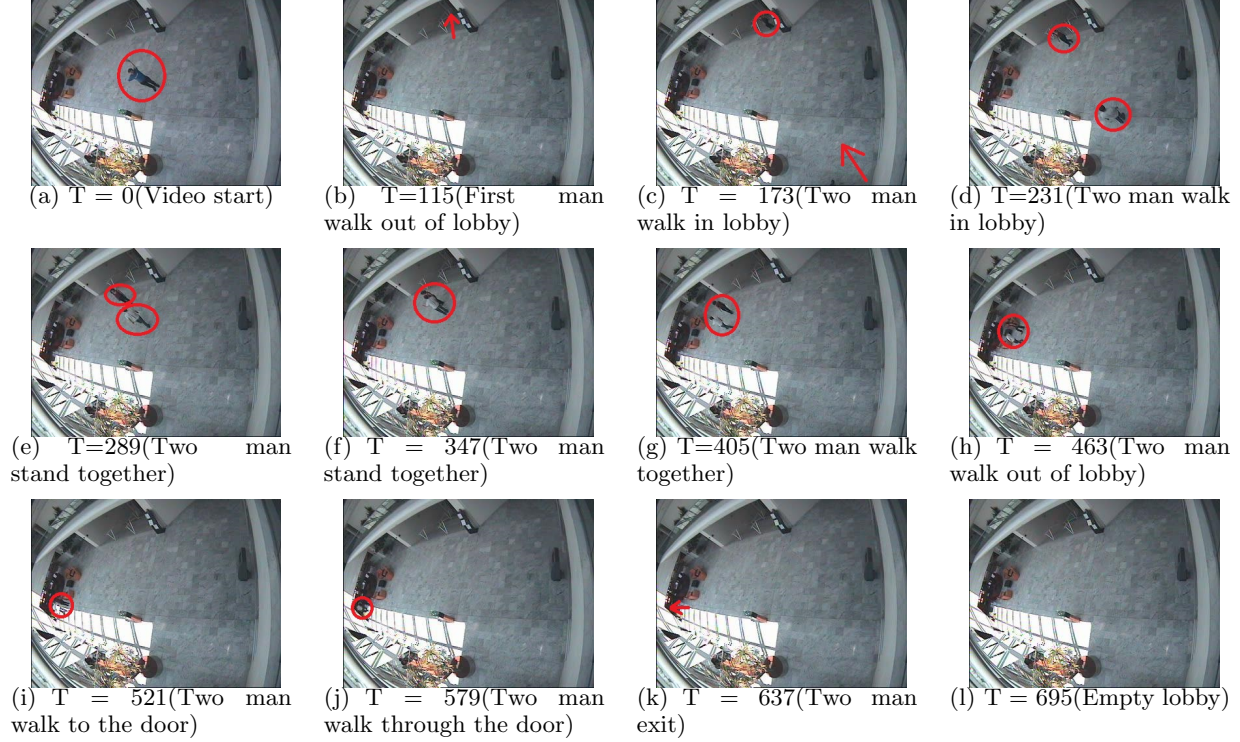


Figure 4: View of Footage

F.3 Additional Simulation Study

In this simulation, we show the accuracy of recovering low rank matrix in the first step of Algorithm 1 with increasing dimension. Consider the VAR model

$$\text{Target model: } X_t^{(0)} = (L + S_0)X_{t-1}^{(0)} + \epsilon_t^{(0)}; \text{ Auxiliary model: } X_t^{(k)} = (L + S_k)X_{t-1}^{(k)} + \epsilon_t^{(k)}, \quad 1 \leq k \leq K,$$

where S_0 is one-off diagonal matrix having the following structure

$$S_0 = \begin{bmatrix} 0 & 0.5 & 0 & \cdots & 0 \\ 0 & 0 & 0.5 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0.5 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{p \times p}.$$

L is generated by $L = UDV'$, where $D := \text{diag}(0.2, r)$ is a diagonal matrix. The dimension p is set to 20, 50 and 100, respectively, while the rank r is fixed to be 4 for all p . $S_k, k \geq 1$ is constructed by randomly replacing four entries of S_0 . The sample size for each group is set to be 200, $n_0 = n_1 = \cdots = n_K = 200$. Define \mathcal{S} as the set of non-zero entries in S_0 . Let H be a random subset of $\{(i, j) : 1 \leq i \leq p, 1 \leq j \leq p\}$ such that $|H| = 4$. If $(i, j) \in H \cap \mathcal{S}$, we set $(S_k)_{ij} = 0$; if $(i, j) \in H \cap \mathcal{S}^c$, $(S_k)_{ij} = \eta_{ij} + 0.5$, where $\eta_{ij} \sim b \cdot \text{uniform}(\{+1, -1\})$. Size of auxiliary set is set to 0, 1, \dots , 9.

we utilize a grid search to select the optimal values of λ and μ_i . To simplify the selection procedure, we use the same μ_i for all groups. We compare estimation error of low-rank matrix for different values of p . The squared Frobenius norm error of estimation given by $\|L - \hat{L}\|_F^2$ is shown in Table 2. As we can see, our proposed algorithm gets better low-rank matrix estimator when more observations are provided. This result is consistent with Theorem 1.

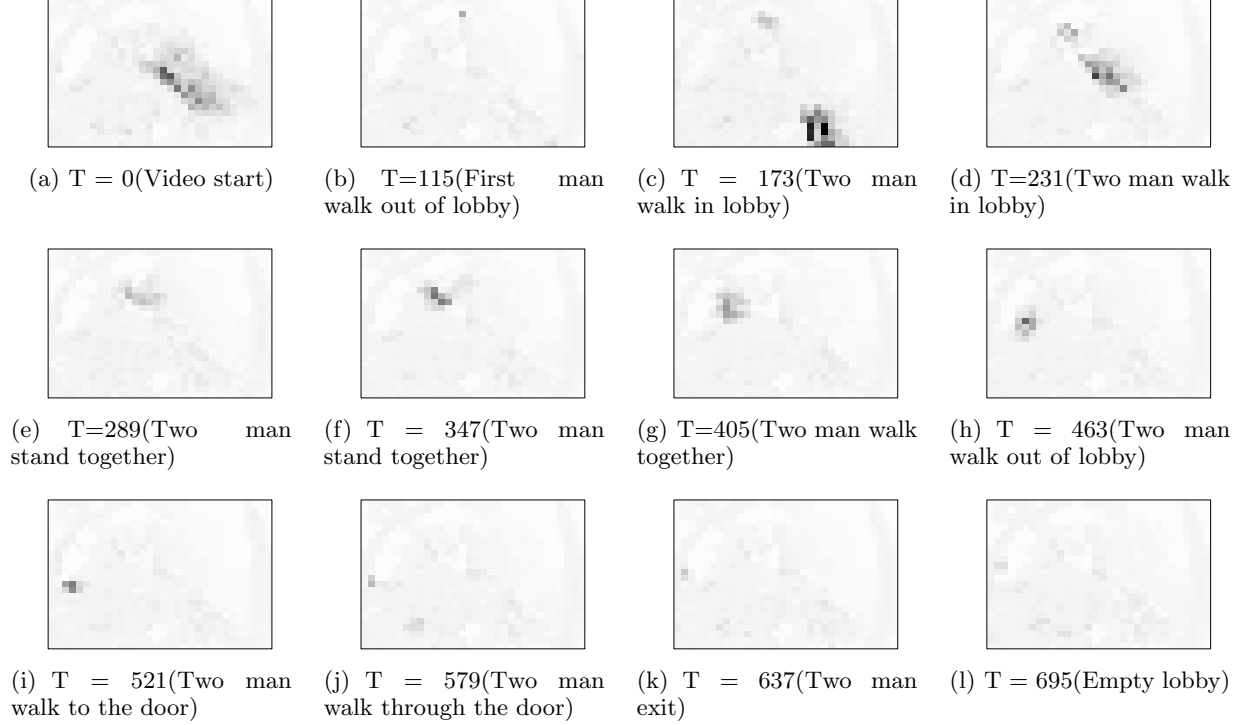


Figure 5: View of Footage

F.4 Computation Time

We provide average computation time for the proposed algorithm in Table 3.

F.5 Alternative Models for the Real Data

In this part, for the real data analysis, we make a comparison with other parameterizations (results are provided in Table 4). As we can see from these results, the proposed algorithm coupled with low rank plus sparse model parameters outperforms all other competing methods.

F.6 Hyperparameter selection

Using cross-validation to select hyperparameters (tuning parameters) in our model is difficult due to the time series nature of the data. Also, it would be computationally demanding since each private sparse component corresponds to one hyperparameter. It seems difficult to compute all cases when dealing with large groups. There are three types of hyperparameters in the proposed algorithms. The first group is related to lasso penalty terms. For those, we use suggestions in the literature for tuning parameter selection (Li et al., 2022a). Specifically, we set $\lambda_\beta = \sqrt{2 \log p / (n_0 + n_{\mathcal{A}_0})}$ and $\lambda_\delta = \sqrt{2 \log p / (n_0)}$. Second, there will be a tuning parameter for the low rank penalty. For that, we set $\mu = \tau * \sqrt{np}$. We performed several simulations with different τ . Empirically speaking, our algorithm reaches a good result when $\tau \in (0.1, 2)$. We set $\tau = 0.1$ in all numerical analyses. Further, there will be an additional tuning parameter for the selection algorithm which is the constant c . We performed sensitivity analysis and given $c \in (0.01, 0.5)$, the algorithm always selects helpful groups for the next transfer learning step. We set $c = 0.01$ in all numerical analyses. Finally, for the lasso method, we set $\lambda = \sqrt{2 \log p / (n_0)}$ while for the inference part we use $r_0 = \sqrt{n_0}$, $r_i = 2^i$ and $\mu_j = \sqrt{\frac{\log p}{2m_j}}$ for $i, j \geq 1$ (Deshpande et al., 2021).

Table 2: Squared Frobenius Norm Error of L

	$p = 20$	$p = 50$	$p = 100$
$K = 0$	0.639	0.639	0.639
$K = 1$	0.616	0.631	0.636
$K = 2$	0.546	0.612	0.628
$K = 3$	0.468	0.579	0.611
$K = 4$	0.414	0.529	0.597
$K = 5$	0.368	0.484	0.578
$K = 6$	0.320	0.451	0.555
$K = 7$	0.294	0.429	0.533
$K = 8$	0.274	0.407	0.508
$K = 9$	0.260	0.389	0.486

N	200	400	600	800	1000
$p = 20$	0.1869	0.5492	1.115	2.040	3.351
$p = 40$	0.2011	0.5986	1.206	2.213	3.607
$p = 60$	0.2184	0.6418	1.314	2.378	3.848
$p = 80$	0.2284	0.6890	1.410	2.553	4.109
$p = 100$	0.2448	0.7413	1.499	2.722	4.381

Table 3: Computation time of algorithm 1. Each entry is the total time of 200 replicates. The unit of computation time is second. We use the same target model as simulation 1 in our paper and consider only one auxiliary model. The sample size of the target model is set to be 100 and sample size of auxiliary model take the value from $\{100, 300, 500, 700, 900\}$.

G Additional Details on Numerical Experiments

Computer Information:

Processor: Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz

Installed RAM: 16.0 GB (15.9 GB usable)

System type: Windows 10 Home 64-bit operating system, x64-based processor

Time of execution: 3h(one computer)

	seg1	seg2	seg3	seg4
Trans-lasso(L+S)	7.205(0.015)	0.133(0.003)	2.177(0.012)	0.468(0.005)
Trans-lasso(S)	7.415(0.015)	0.289(0.004)	3.398(0.014)	0.673(0.006)
lasso(L+S)	8.660(0.017)	0.241(0.004)	4.484(0.017)	1.928(0.011)
lasso(S)	8.686(0.017)	0.257(0.004)	4.579(0.017)	1.954(0.011)
Low-rank	14.090(0.021)	1.955(0.005)	6.854(0.018)	5.008(0.014)
	seg5	seg6	seg7	seg8
Trans-lasso(L+S)	0.092(0.002)	0.916(0.008)	0.372(0.005)	0.233(0.004)
Trans-lasso(S)	0.251(0.004)	1.164(0.009)	0.920(0.008)	0.503(0.006)
lasso(L+S)	0.450(0.005)	2.686(0.013)	1.653(0.010)	1.235(0.009)
lasso(S)	0.476(0.005)	2.682(0.013)	1.683(0.010)	1.273(0.009)
Low-rank	2.388(0.007)	5.682(0.016)	3.938(0.012)	3.509(0.011)
	seg9	seg10	seg11	seg12
Trans-lasso(L+S)	0.146(0.003)	0.094(0.002)	0.071(0.002)	0.139(0.002)
Trans-lasso(S)	0.297(0.004)	0.209(0.003)	0.163(0.003)	0.212(0.002)
lasso(L+S)	0.523(0.006)	0.189(0.004)	0.102(0.004)	0.153(0.002)
lasso(S)	0.529(0.006)	0.211(0.004)	0.117(0.003)	0.219(0.003)
Low-rank	2.377(0.007)	1.910(0.005)	1.815(0.004)	0.876(0.003)

Table 4: Mean squared prediction error for each segment. Standard errors are shown in parentheses. Trans-lasso(L+S) and lasso(L+S) refer to methods that we model VAR(1) with low-rank plus sparse structure. Trans-lasso(S) and lasso(S) refer to methods that we model VAR(1) without low-rank component. Low-rank methods in the last row implies that we model VAR(1) with only low-rank component for each segment.