# Data-Secure Transfer Learning in Heterogeneous Low-Rank and Sparse Panel VAR Models

Yuchen Xu George Michailidis

University of California, Los Angeles

**Abstract**

Transfer learning (TL) enhances high-dimensional inference by borrowing strength from auxiliary data sets. We extend TL to multivariate time series settings through a panel Vector Autoregression (VAR) model, whose coefficients decompose into heterogeneous low-rank and sparse components. The proposed algorithm treats the two components independently and automatically detects the transferable auxiliary sources for each. Importantly, it operates solely on *pre-trained estimators* from the auxiliary panel, thereby preserving data privacy. We establish that the source detection mechanism achieves high accuracy under mild technical conditions. The proposed TL approach delivers significant performance improvements in data-scarce regimes—precisely the scenarios that motivate transfer learning. Extensive numerical experiments based on synthetic data, conducted for both correctly specified and misspecified time series models, along with applications to macroeconomic time series, confirm the practical effectiveness of our method. The results highlight the potential of TL to enhance predictive tasks in complex multivariate time series models when primary data are limited.

sectionIntroduction

Inspired by the human ability to apply knowledge from familiar areas to new ones, transfer learning (TL) has become a widely adopted approach in statistics and machine learning (Tan et al., 2018; Niu et al., 2020; Iman et al., 2023), with significant applications in medical image classification (Kim et al., 2022), biological signal processing (Bird et al., 2020), and generative language models (Wolf et al., 2019). Recent research has focused on model/parameter transfer in *parametric* statistical models. For example, Li et al. (2022) introduced the Trans-Lasso algorithm that detects informative auxiliary sources and utilizes their data to improve Lasso based sparse estimates of a target regression model. He et al. (2024) proposed a fused regularization approach that leverages both auxiliary and target data robust to covariate shift. Recent advances have extended TL to sparse generalized linear models (Tian and Feng, 2023) and sparse graphical models (Zhao et al., 2024). By carefully borrowing strength from auxiliary data sets, these TL methods improve statistical efficiency in complex models, particularly when working with limited target data—a common challenge in modern high-dimensional applications.

This work represents the first application of TL to high-dimensional time series analysis—a domain that demands substantial computational resources for model selection and estimation. Our method builds upon a heterogeneous panel vector auto-regression (VAR) framework featuring heterogeneous low-rank plus sparse components, originally introduced by Xu and Michailidis (2025). Specifically, it is assumed that multivariate time series data $\{X_t^m \in \mathbb{R}^p;\ m = 1, \cdots, M,\ t = 1, \cdots, T\}$ for $M$ entities can be modeled according to the following PVAR model

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

where $A_m \in \mathbb{R}^{p\times p}$ is the transition matrix of the VAR model for the $m$-th entity, and $\Sigma_m \in \mathbb{R}^{p\times p}$ is the corresponding covariance structure. In particular, the transition matrix $A_m$ comprise of a rescaled version of a low-rank component $\Phi$, while additional heterogeneity is captured by the *sparse* component $S_m$. The lag-1 assumption is made for simplicity of presentation; however, the model can be readily extended to accommodate higher-order lags. Such PVAR formulation (1) presents a well-suited setup for TL, as the shared low-rank component $\Phi$ naturally induces interdependence across entities and promotes knowledge transfer through the structured decomposition of $A_m$. The objective is to estimate the parameters of a target $p$-dimensional VAR(1) process, for which $T_0$ time points are available, and whose structure is specified by:

$$X_t^0 = A_0 X_{t-1}^0 + \epsilon_t^0,\ \epsilon_t^0 \sim N(0, \Sigma_0); \qquad A_0 = W_0\Phi + S_0, \tag{2}$$

where the matrix $A_0 \in \mathbb{R}^{p\times p}$ follows the same low-rank plus sparse decomposition as in (1). In settings where $T_0$ is fairly small, TL should prove useful in leveraging $\Phi$ from the PVAR model and also improve the estimates of $W_0$ and $S_0$. In such model setup, two distinct sets of parameters–namely, the rescaling coefficients $W$ and the sparse component $S$–are amenable to improvement through TL, due to the model's decomposition. Leveraging this structural separation, we design data-secure TL strategies that separately exploit the similarity structures of the target $W_0$ and $S_0$ with those from the auxiliary $M$-entity panel.

The key issues we are going to address are how to efficiently identify the sets $\mathcal{A}_W$ and $\mathcal{A}_S$, and how to improve the estimates of $W_0$ and $S_0$. The main challenges involved in these tasks—and our corresponding contributions—are as follows:

- **Dependent Data:** Unlike existing TL methods for regression that assume independent observations, our approach accommodates time series data with complex temporal dependence structures.

- **Dual Parameter Transfer:** Unlike prior TL work that transfers a single parameter (Li et al., 2022; Tian and Feng, 2023; Zhao et al., 2024), our approach addresses two distinct parameters that may exhibit different similarity patterns with those in the target PVAR model.

- **Privacy Preserving TL:** Unlike existing TL methods that require access to raw auxiliary data (Li et al., 2022; Tian and Feng, 2023), our approach operates solely on pretrained estimates from the auxiliary panel—ensuring data privacy and enabling distributed computation.

- **Generalizability:** The proposed methodology can easily be adapted to many other statistical models (e.g., multivariate regression models, graphical models).

- **Performance Guarantees:** Under fairly mild assumptions, we establish theoretical guarantees showing that the proposed method accurately identifies informative sources and achieves improved estimation rates for the target model parameters. Notably, when the auxiliary panel contains informative members and the target time series is short, our theoretical results are supported by extensive numerical experiments on synthetic data.

- **Robustness to Misspecification:** Through simulations and applications with unknown latent data-generating mechanisms, we demonstrate that our approach consistently delivers competitive results compared to benchmark algorithms, especially when the target series is short.

The remainder of the paper is organized as follows. Section 1 develops the algorithm and discusses consistency when the transfer information is given. Section 2 focus on a more generic setting that detects the transfer structure for TL estimates. The simulation studies with both correctly specified and misspecified settings are included in Section 3. Section 4 illustrates the TL idea in two macroeconomic time series applications. Some concluding remarks are drawn in Section 5.

**Notation** Denote by $e_i$ the $i$-th standard basis vector with an appropriate dimension. The operator $\operatorname{diag}(x)$ returns either a diagonal matrix with the diagonal elements from $x$ if $x$ is a vector, or a vector consisting of the diagonal elements of $x$ if $x$ is a square matrix. For matrix $X$, $X_{i,j}$ is the $(i,j)$-th element, $X'$ denotes the transpose of $X$, $\operatorname{rank}(X)$ returns the rank of $X$, and $\lambda_i(X)$ is the $i$-th singular value of $X$ sorted nonincreasingly. We denote the support of $X$ as $\operatorname{supp}(X) = \{(i,j) : X_{i,j} \neq 0\}$. With $A$ being a collection of values, including matrices and vectors, $Q_L^U(A)$ truncate the values of $A$ by its $L$-quantile at the lower tail and $1-U$-quantile at the upper tail, $\operatorname{mean}(A)$ and $\operatorname{median}(A)$ are respectively the mean and median of the elements in the collection. In addition, we denote $A^{\otimes a}$, the power of collection, as $A^{\otimes a} = \{x^a : x \in A\}$. For norms, $\|X\| = \|X\|_2$ by default represents the spectral norm, $\|X\|_1 = \sum_{i,j} |X_{i,j}|$ is the $\ell_1$ norm that sums up all the absolute values of matrix entries, and $\|X\|_\infty = \max_{i,j} |X_{i,j}|$ is the $\infty$-norm that returns the maximum absolute value. The nuclear norm of matrix $X$ is denoted as $\|X\|_*$. $O(\cdot)$ and $o(\cdot)$ are the usual big and small O notations respectively, and we use $x \gtrsim y$ ($x \lesssim y$) if there exists an absolute positive constant $c$ such that $x \geqslant cy$ ($x \leqslant cy$). We also define $O_{\max}(x_1, \ldots, x_k)$ (and $O_{\min}$) to be the maximum (and minimum) operators among the orders of the quantities $(x_1, \ldots, x_k)$ (up to some constants).

# 1 Estimation with Known Transferrable Auxiliary Models

In this section, we briefly discuss our proposed transfer learning algorithm on the low-rank plus sparse VAR model, in an oracle version that requires the known transferrable auxiliary information.

Let

$$W_m = W_0 + \delta_m^W, \qquad S_m = S_0 + \delta_m^S. \tag{3}$$

Then for positive thresholds $h_W, h_S$, one can identify the indices of informative $W_m$'s and $S_m$'s from the auxiliary panel defined as

$$\mathcal{A}_W := \{m : \|\delta_m^W\| \leqslant h_W\} \subset \{1, \ldots, M\}, \qquad \mathcal{A}_S := \{m : \|\delta_m^S\| \leqslant h_S\} \subset \{1, \ldots, M\}. \tag{4}$$

## 1.1 VAR-TL Algorithm

The following proposition is more algorithmic about our proposed TL roadmap which consists of two steps. Intuitively, the first step (5) exploits the informative auxiliary sources for a good initial estimator in certain neighborhood, and the second step uses the target data for further refinement.

**Proposition 1.1** (Oracle Transfer)**.** *Suppose the auxiliary PVAR model follow* (1) *with accessible pretrained estimates* $\left(\widehat{\Phi}, \{\widehat{W}_m, \widehat{S}_m\}_{m=1}^M\right)$*, and the informative auxiliary models are given by indices sets* $\mathcal{A}_W$ *and* $\mathcal{A}_S$ *as in* (4)*. For the target model* (2)*, the transfer learning procedure takes two steps.*

1. *Pooled Average:*

$$\begin{aligned} \bar{W}_0 &= \text{mean}\{\widehat{W}_m : m \in \mathcal{A}_W\}, \\ \bar{S}_0 &= \Pi_{\mathcal{P}_S}(\text{mean}\{\widehat{S}_m : m \in \mathcal{A}_S\}), \quad \mathcal{P}_S = \text{supp}(\text{median}\{|\widehat{S}_m| : m \in \mathcal{A}_S\}) \end{aligned} \tag{5}$$

*Here the operator $\Pi_{\mathcal{P}}(x)$ projects the matrix $x$ onto the support $\mathcal{P}$. We set $\bar{W}_0$ ($\bar{S}_0$) to be zero if $\mathcal{A}_W$ ($\mathcal{A}_S$) is an empty set.*

2. *Bias correction fit:*

$$(\widehat{\delta}^W, \widehat{\delta}^S) = \underset{\delta^W, \delta^S}{\text{argmin}} \frac{1}{2T_0} \| \left( (\bar{W}_0 + \delta^W)\widehat{\Phi} + (\bar{S}_0 + \delta^S) \right) X_0 - Y_0 \| + p_\lambda(\delta^W, \delta^S), \tag{6}$$

*where $p_\lambda$ is a regularization function with parameters $\lambda = (\lambda^W, \lambda^S)$ to be tuned.*

*Remark* 1.1. 1. Unlike the existing literature, the proposed algorithm here does not require access to the raw auxiliary data. Instead, under mild conditions like Assumption 4, the pretrained auxiliary estimators are sufficient to provide a high-quality initial fit as in (5); refer to ... for theoretical discussions. This initialization derived from the auxiliary models then gets refined in the second step (6) using only the target data.

2. In our implementation, the bias correction problem (6) is solved using Lasso penalty for $p_\lambda$, which induces sparsity in both $\delta^W$ and $\delta^S$. Alternatively, $p_\lambda$ may take different forms, such as ridge-type penalties, whose theoretical guarantees can be established with analogous arguments.

The procedures of the oracle algorithm are briefly summarized in Algorithm 1.

**Algorithm 1:** Oracle Transfer Learning of $W_0$ and $S_0$

**Input:** Pretrained auxiliary model estimates $\{\widehat{\Phi}, (\widehat{W}_m, \widehat{S}_m)_{m=1}^M\}$, Target data $(X_0, Y_0)$, Informative sets $\mathcal{A}_W$ and $\mathcal{A}_S$, regularization parameter $\lambda$.

**Output:** Estimators $(\widehat{W}_0, \widehat{S}_0)$.

1 Solve $(\bar{W}_0, \bar{S}_0)$ from (5).

2 Solve $(\widehat{\delta}^W, \widehat{\delta}^S)$ from (6).

3 Output target estimator $\widehat{W}_0 = \bar{W}_0 + \widehat{\delta}^W$ and $\widehat{S}_0 = \bar{S}_0 + \widehat{\delta}^S$.

## 1.2 Properties of the Oracle Algorithm

We then present the theoretical performance of the oracle algorithm proposed in Proposition 1.1 under some mild conditions discussed below.

As a starting point, and in line with the VAR literature, we impose the following two classical assumptions to ensure that the target VAR model is stationary and that the resulting data matrix of the target data is well-conditioned for convex optimization.

*Assumption* 1 (Stationarity). The target VAR(1) model (2) satisfies $\|A_0\| < 1$.

*Assumption* 2 (Restricted Strong Convexity (RSC)). For arbitrary $\Delta \in \mathbb{R}^{p \times p}$, there exists $\beta = \lambda_p(\frac{X_0 X_0'}{T_0}) > 0$, such that the following inequality holds,

$$\frac{1}{T_0} \|\Delta X_0\|_F^2 \geqslant \beta \|\Delta\|_F^2.$$

To guarantee that the concern of identifiability between $W$ and $\Phi$ is well-addressed, we specify the parameter space of the low rank components.

*Assumption* 3. The low rank matrix $\Phi$ and its pretrained estimator $\widehat{\Phi}$ are from

$$\mathbb{L}_{p,r} := \{\Phi \in \mathbb{R}^{p\times p} : \operatorname{rank}(\Phi) \leqslant r,\ \|e_1'\Phi\| = \cdots = \|e_p'\Phi\| = 1\}. \tag{7}$$

The next necessary condition is imposed to guarantee the consistency of pretraining on the auxiliary PVAR model. In the literature, accurate TL estimation typically requires assumptions on the auxiliary models' noise behaviors (Li et al., 2022). By contrast, our data-secure approach does not need access to the raw data, hence the effectiveness may rely entirely on the quality of the pretrained estimators.

*Assumption* 4. For $\Phi, \widehat{\Phi} \in \mathbb{L}_{p,r}(\varphi)$, the pretraining satisfies

$$\|\Phi - \widehat{\Phi}\|_{2\to\infty}^2 \lesssim \frac{r}{MT}. \tag{8}$$

*Remark* 1.2. Note that (8) is a sufficient condition to imply the classical error rate $\|\Phi-\widehat{\Phi}\|_F^2 \lesssim \frac{rp}{MT}$ in low-rank regression problems (see Basu et al. (2019) for instance), but it imposes slightly stronger restrictions on each row of the estimation error $\Phi - \widehat{\Phi}$. Indeed, given that (1) is identical to the model proposed by Xu and Michailidis (2025), the proposed algorithm therein guarantees well-behaved consistent estimates $(\widehat{\Phi}, \{\widehat{W}_m, \widehat{S}_m\}_{m=1}^M)$ with high-probability statements. Here we set a deterministic error bound on the pretrained low-rank $\widehat{\Phi}$, and consequently derive the high-probability consistency of $(\widehat{W}_m, \widehat{S}_m)$ and their error rates in Proposition A.1.

The final assumption is about the structures of the sparse components $(S_0, \widehat{S}_0)$ and $(S_m, \widehat{S}_m)$ for all $m$. It is a mild condition given model sparsity and the Lasso estimation, and the row-wise sparse structure is essential to guarantee low-rank sparse incoherence; see Hsu et al. (2011).

*Assumption* 5. With $s = o(p^2) > 0$, $\max(\{\|S_m\|_0, \|\widehat{S}_m\|_0\}_{m=0}^M) \leqslant s$. Further, for all $1 \leqslant i \leqslant p$, $\max(\{\|e_i'S_m\|_0, \|e_i'\widehat{S}_m\|_0\}) \lesssim \frac{s}{p}$.

Then we state the following main theorem about high-probability consistency statements of oracle TL estimator from Algorithm 1. See Appendix B for detailed proof.

**Theorem 1.1.** *Suppose Assumptions 1 to 5 are satisfied, the sets $(\mathcal{A}_W, \mathcal{A}_S)$ are known, and suitable irrepresentable conditions hold to ensure that the sparse components are distinguishable. Denote*

$$\rho_W = \sqrt{\frac{rp}{MT}} + \sqrt{\frac{p}{|\mathcal{A}_W|T}}, \qquad \rho_S = \sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{|\mathcal{A}_S|T}}.$$

*We take $\lambda^W \geqslant \frac{1}{2\sqrt{p}}\|\frac{X_0\epsilon_0'}{T_0}\|_2$ and $\lambda^S \geqslant \frac{1}{2}\|\frac{X_0\epsilon_0'}{T_0}\|_\infty$, there exist constants $c_1, c_2, c_3 > 0$, such that with probability at least $1 - c_2\exp(-c_3\log(p))$,*

$$\begin{aligned}&\|\widehat{W}_0 - W_0\|_F^2 + \|\widehat{S}_0 - S_0\|_F^2 \\ &\qquad \leqslant c_1 \max\left((\rho_W + h_W + \rho_S + h_S)^2, \sqrt{\frac{p}{T_0}}(\rho_W + h_W) + \sqrt{\frac{s\log(p)}{T_0}}(\rho_S + h_S)\right)\end{aligned} \tag{9}$$

*Remark* 1.3. Usually when the auxiliary panel is large (large $M$) and the low-rank structure is valid (small $r$), the error term $\sqrt{\frac{rp}{MT}}$ and $\sqrt{\frac{rs}{MT}}$ introduced by the biased covariates $\widehat{\Phi}$, which may

be viewed as a systematic error, can be negligible in $\rho_W$ and $\rho_S$. The leading error terms are essentially from $(\sqrt{\frac{p}{|\mathcal{A}_W|T}}, \sqrt{\frac{s\log(p)}{|\mathcal{A}_S|T}})$ and $(h_W, h_S)$. Observe that for a simple partially penalized Lasso regression only on the target data, the rates of estimation errors typically comprise the orders like $\frac{p}{T_0} + \frac{s\log(p)}{T_0}$ in addition to the systematic error. Compared with (9), when $|\mathcal{A}_W|T \gtrsim T_0$ and $|\mathcal{A}_S|T \gtrsim T_0$, the transfer learning approach can get a sharper high-probability rate if $h_W \ll \sqrt{\frac{p}{T_0}}$ and $h_S \ll \sqrt{\frac{s\log(p)}{T_0}}$. It is equivalent to interpret that, if the parameters of some auxiliary models are in an estimation confidence neighborhood of the target parameters, then TL can be beneficial for improving the accuracy and prediction performance, with supporting evidence from simulation studies in Section 3.

## 2 TL with Unknown Transfer Structure

In practice, the sets of indices $(\mathcal{A}_W, \mathcal{A}_S)$ of informative models are typically unknown. It is a critical task to learn the estimated sets $(\widehat{\mathcal{A}}_W, \widehat{\mathcal{A}}_S)$ from the data. The key idea is to establish some rough estimator of $A_0$ and statistically determine whether a pretrained $\widehat{W}_m$ (or $\widehat{S}_m$) significantly contributes to it accounting for the noises and the source's informative level $h_W$ (or $h_S$).

We first need to guarantee that the informative and noninformative auxiliary models are well separated in the following sense.

*Assumption* 6. In addition to (3) and (4), the parameters of the auxiliary panel satisfy:

1. For any $m \notin \mathcal{A}_W$, $\|\delta_m^W\|_F \gg O_{\max}(\sqrt{\frac{p}{T}}, \sqrt{\frac{rp}{MT}}, \sqrt{\frac{p\log(p)}{T_0}})$.
2. For any $m \notin \mathcal{A}_S$, $\frac{\langle|\delta_m^S|,|S_m|\rangle}{\|S_m\|_F} \gg O_{\max}(\sqrt{\frac{sp}{MT}}, \sqrt{\frac{s\log(p)}{T}}, \sqrt{\frac{s\log(p)}{T_0}})$.

Note that when $m \notin \mathcal{A}_S$, the above assumption on $\delta_m^S$ is a sufficient condition for $\|\delta_m^S\|_F \gg O_{\max}(\sqrt{\frac{sp}{MT}}, \sqrt{\frac{s\log(p)}{T}}, \sqrt{\frac{s\log(p)}{T_0}})$, which is analogous to that of $\|\delta_m^W\|_F$.

Here we express the data version of our target model as $Y_0 = (W_0\Phi + S_0)X_0 + \epsilon_0$, where $Y_0, X_0, \epsilon_0 \in \mathbb{R}^{p\times T_0}$. Assumption 2 implies that $X_0$ is full rank, hence there exists a pseudo inverse $X_0^+ = X_0'(X_0X_0')^{-1}$ such that

$$Y_0X_0^+ = W_0\Phi + S_0 + \epsilon_0X_0^+. \tag{10}$$

We then have the following proposition to detect the informative auxiliary models based on (10).

**Proposition 2.1.** *Suppose that the pretrained auxiliary estimates are available, and Assumptions 1 to 6 are satisfied. Denote $c_i > 0$ for $i = 1, 2$ as scale-free constants that may depend on the model parameters $(A_0, \Sigma_0)$, and two conditions $C_W$ and $C_S$ as follow.*

$$C_W := \{T_0 \gtrsim p, \qquad T \gtrsim p, \qquad h_W \lesssim O_{\max}(\sqrt{\frac{rp}{MT}}, \sqrt{\frac{p}{T}}, \sqrt{\frac{p\log(p)}{T_0}})\},$$

$$C_S := \{T_0 \gtrsim \log(p), \quad T \gtrsim \log(p), \quad h_S \lesssim O_{\max}(\sqrt{\frac{rs}{MT}}, \sqrt{\frac{s\log(p)}{T}}, \sqrt{\frac{s\log(p)}{T_0}})\}.$$

1. *Define $\gamma = \frac{\log(p^2)}{p^2}$ and $\mathbf{S}_m := Y_0X_0^+ - \widehat{W}_m\widehat{\Phi}$. If the condition $C_W$ is satisfied, then with probability at least $1 - c_1\exp(-c_2\log(p))$,*

$$\mu_m := \|\operatorname{diag}(\frac{Q_\gamma^\gamma(\mathbf{S}_m)X_0X_0'\widehat{\Phi}'}{T_0})\|_2 \begin{cases} \lesssim \sqrt{\frac{rp}{MT}} + \sqrt{\frac{p}{T}} + \sqrt{\frac{p\log(p)}{T_0}}, & m \in \mathcal{A}_W \\ \gtrsim \|\delta_m^W\|_F, & m \notin \mathcal{A}_W \end{cases}. \tag{11}$$

*And under the condition $C_W$ with $p \to \infty$, $P(\max_{m\in\mathcal{A}_W} \mu_m < \min_{m\notin\mathcal{A}_W} \mu_m) \to 1$.*

2. *Define* $\mathbf{W}_m := Y_0 X_0^+ - \widehat{S}_m - \widetilde{W}_m\widehat{\Phi}$, *where* $\widetilde{W}_m = \underset{W=\text{diag}(w)}{\text{argmin}} \|Y_0 X_0^+ - \widehat{S}_m - W\widehat{\Phi}\|_1$. *If the condition $C_S$ is satisfied, then with probability at least* $1 - c_1\exp(-c_2\log(p))$,

$$\kappa_m := \frac{\langle|\widehat{S}_m|, |\frac{\mathbf{W}_m X_0 X_0'}{T_0}|\rangle}{\|\widehat{S}_m\|_F} \begin{cases} \lesssim \sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{T}} + \sqrt{\frac{s\log(p)}{T_0}}, & m \in \mathcal{A}_S \\ \gtrsim \frac{\langle|\delta_m^S|, |S_m|\rangle}{\|S_m\|_F}, & m \notin \mathcal{A}_S \end{cases}. \tag{12}$$

*And under the condition $C_S$ with $p \to \infty$, $P(\max_{m\in\mathcal{A}_S} \kappa_m < \min_{m\notin\mathcal{A}_S} \kappa_m) \to 1$*

*Remark* 2.1. The rates of $h_W$ and $h_S$ in conditions $C_W$ and $C_S$, to a large extent, coincide with our discussions in Remark 1.3. It requires that an auxiliary model is identified informative only if it is (in terms of the order) within an estimation confidence distance from the target parameters. This proposition essentially designs proxy quantities that evaluate such distances asymptotically, and provides the guidelines for selecting the estimated informative sets:

$$\widehat{\mathcal{A}}_W = \{m : \mu_m \lesssim \sqrt{\frac{rp}{MT}} + \sqrt{\frac{p}{T}} + \sqrt{\frac{p\log(p)}{T_0}}\},$$
$$\widehat{\mathcal{A}}_S = \{m : \kappa_m \lesssim \sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{T}} + \sqrt{\frac{s\log(p)}{T_0}}\}.$$

# 3 Performance Evaluation based on Synthetic Data

In this section, we design simulation studies to present the discussions about the advantages of the proposed TL approach under various conditions.

## 3.1 Correctly Specified Setups

For correctly specified settings, we simulate the auxiliary PVAR model for $(p, M, T) = (30, 10, 800)$ and vary the target $T_0 \in \{100, 200, 400, 800, 1600, 6400\}$. Further, we set $(\omega_W, \omega_S) = (0.4, 0.3)$, where $\omega_W = \frac{|\mathcal{A}_W|}{M}$ and $\omega_S = \frac{|\mathcal{A}_S|}{M}$. We compare the performance of the VAR-TL algorithm with transfer structure detection (Full TL) against Local Low-rank and Sparse VAR Basu et al. (2019), which relies solely on target data; Nominal TL, which uses only the pretrained $\widehat{\Phi}$; and PVAR Refit, which fits a combined PVAR model over the $M + 1$ datasets by incorporating the target data into the original auxiliary data (assuming full data access). Figure 1 depicts the relative performance of the competing methods against the Local Low-rank and Sparse VAR across three metrics, based on 100 data replicates. It is evident that for small sample sizes ($T_0 \leqslant 400$), the Full TL method demonstrates significant improvements across all metrics compared to all competitors, while its gains become smaller for larger $T_0$. This is expected, because the target problem standalone gradually becomes well-posed with the increasing amount of data.

We also highlight that even nominally transferring the background structure $\Phi$–under the correctly specified model (1) and (2)–can introduce descent performance gains under the data-limited settings.

## 3.2 Robustness on Misspecified Setups

We also evaluated the Full TL method under misspecified scenarios, where both the auxiliary panel data and the target data deviate from the assumed low-rank plus sparse structure. Specifically, the

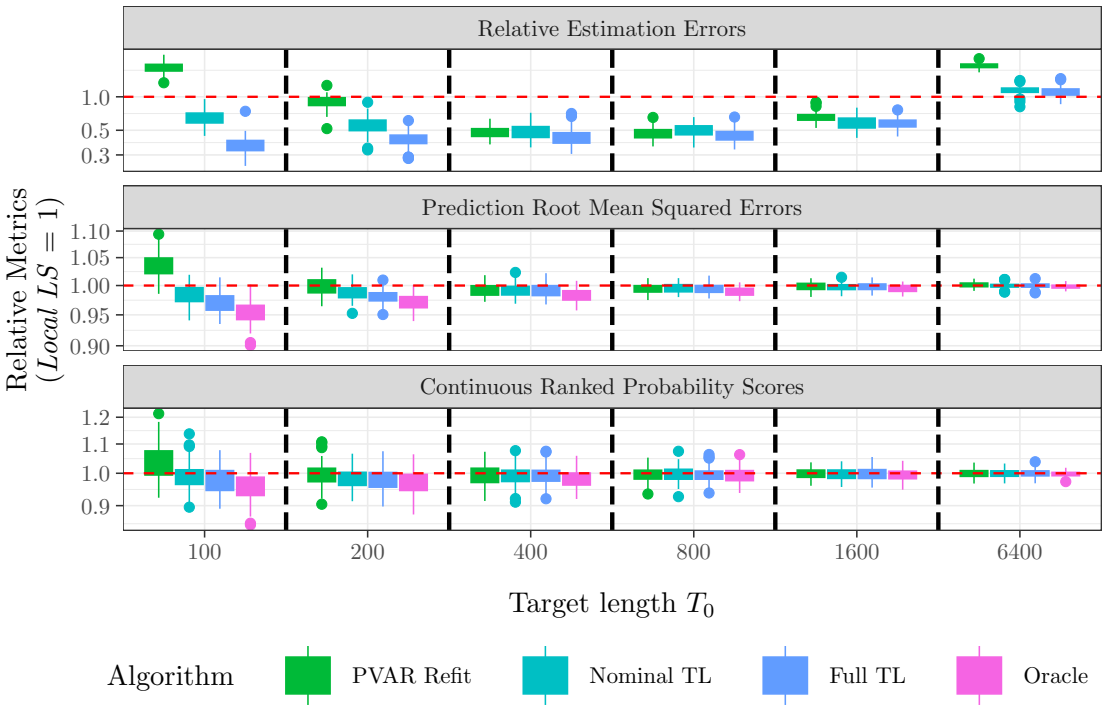


Figure 1: The summary metrics of comparison on benchmark algorithms.

auxiliary data are generated either from a panel of dense VAR models or from a panel of dynamic factor models (DFM) driven by common low-dimensional factor series.

For more details, the first dense VAR(1) panel has coefficient matrices $A_m$ consisting of both a dense and a sparse component. Specifically,

$$X_t^m = A_m X_{t-1}^m + \epsilon_t^m,\ \epsilon_t^m \sim N(0, \Sigma_m); \qquad A_m = B_m + S_m, \tag{13}$$

where $B_m \in \mathbb{R}^{p\times p}$ is dense and $S_m \in \mathbb{R}^{p\times p}$ is sparse.

The common lower-dimensional ($r < p$) factor series of the second example follows a VAR(1) structure, and the innovations to the factor series exhibits some sparse contemporaneous correlations. The observations are generated by loading the common factors and adding noises. Formally,

$$\begin{aligned} F_t &= A_f F_{t-1} + (I + S_f) f_t, \quad & f_t &\sim N(0, I), \\ X_t^m &= B_m F_t + \epsilon_t^m, & \epsilon_t^m &\sim N(0, \Sigma_m), \end{aligned} \tag{14}$$

where $A_f \in \mathbb{R}^{r\times r}$ is the coefficient matrix of the factor VAR model, $S_f \in \mathbb{R}^{r\times r}$ introduces some sparse contemporaneous dependence among factor innovations, and $B_m \in \mathbb{R}^{p\times r}$ is the loading matrix for the $m$-th entity.

For both models, we embed TL-type structures analogous to our main problem setting, by respectively assigning cluster patterns to the collections of $\{B_m\}_{m=1}^M$ and, when applicable, $\{S_m\}_{m=1}^M$.

Despite the mismatch, we treat the PVAR model in (1) and the VAR model in (2) as the *working* models for the two cases and fit them accordingly. In the first setting a correctly specified Ridge VAR model is used as a typical benchmark (Basu et al., 2015; Nicholson et al., 2020), while in the second setting a correctly specified DFM model is used as a benchmark.

Results are shown in Figure 2 (dense VAR setting) and Figure 3 (DFM setting). Figure 2 presents the simulation results of the first dense-plus-sparse specification (13), under the setting $(M, p, \omega_W, \omega_S) = (20, 40, 0.2, 0.2)$. We find that Full TL is strongly beneficial across all three performance metrics for smaller sample size for the target data ($T_0 \leqslant 400$).

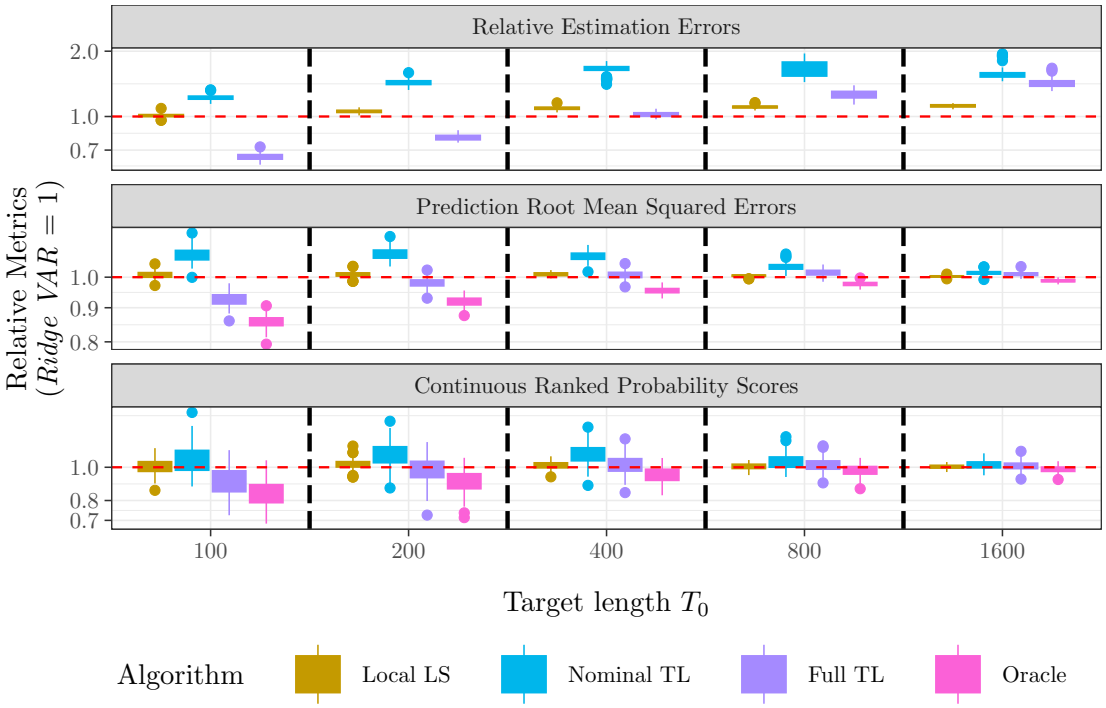


Figure 2: The metrics on a misspecified panel of dense VAR models (13) with $\omega_W = \omega_S = 0.2$.

For the factor model, the simulations from Figure 3 show that TL can be a competitive model in various settings. Especially for shorter target series, Full TL approach has better prediction errors and distributional forecasts when compared against the correctly specified dynamic factor model.

## 4 Application

We apply the proposed methodology to two economic time series data sets. For both data sets, a low-rank VAR plus sparse structure is a good candidate model.

The first example focuses on $p = 35$ key variables from countries within the Euro-zone. We build the PVAR model (1) based on data from $M = 8$ countries (Belgium, Finland, France, Germany, Netherlands, Norway, Spain, and Sweden) and assess the performance of Full TL on three target countries (Austria, Denmark, and Italy). The results in Table 1 indicate that Full TL demonstrates competitive performance in both point prediction error (PRMSE) and distributional forecasts (mean CRPS).

Our second example analyzes $p = 18$ labor-market time series (unemployment rates, hourly earnings, weekly working hours etc.) from 35 US states, using $M = 30$ states for the auxiliary PVAR model and reserving 5 states for evaluation–Arkansas (AR), Colorado (CO), Connecticut (CT), Idaho (ID), and West Virginia (WV). Results in Table 2 show Full TL delivers competitive

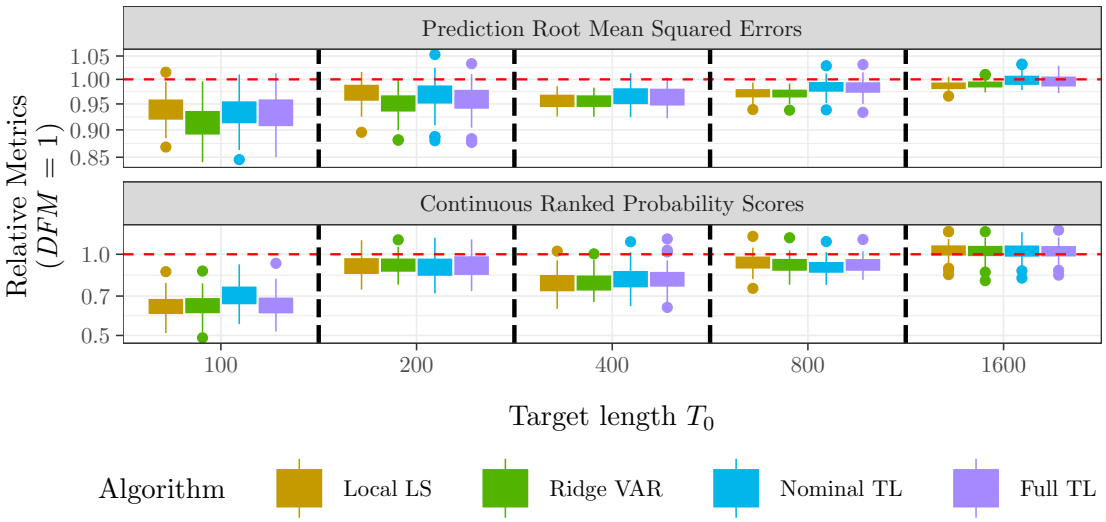


Figure 3: The metrics on a misspecified DFM (14) with $\omega_W = \omega_S = 0.2$.

Table 1: Performance summary of TL versus benchmarks (Ridge VAR as baseline) for the EU macroeconomic application.

| | Algorithm \ Country | Austria | Finland | Italy |
|---|---|---|---|---|
| Relative | Local LS | 1.004 | 0.980 | 1.003 |
| Prediction RMSE | Nominal TL | 0.969 | 0.985 | 0.947 |
| (Ridge VAR = 1) | Full TL | 0.964 | 0.979 | 0.981 |
| Relative | Local LS | 0.995 | 0.994 | 1.006 |
| Prediction CRPS | Nominal TL | 0.890 | 0.942 | 0.838 |
| (Ridge VAR = 1) | Full TL | 0.888 | 0.881 | 0.943 |

forecast accuracy despite data quality issues affecting predictive performance.

## 5 Conclusions

Our transfer learning methodology for time series data offers a data-secure framework for panel VARs, selectively leveraging low-rank and sparse structures. The method provides significant empirical advantages over competing methods, as well as over using only the target data, particularly in the case of short time series. It maintains robustness against model misspecification and demonstrates strong performance in economic data applications. Additionally, the method is supported by theoretical guarantees.

## References

Basu, S., X. Li, and G. Michailidis (2019, March). Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions. *IEEE Transactions on Signal Processing 67*(5), 1207–1222.

Table 2: Performance summary of TL versus benchmarks (Ridge VAR as baseline) for the US labor market application.

| | Algorithm \ State | AR | CO | CT | ID | WV |
|---|---|---|---|---|---|---|
| Relative | Local LS | 1.002 | 0.989 | 0.994 | 0.984 | 0.997 |
| Prediction RMSE | Nominal TL | 0.992 | 0.981 | 0.967 | 0.956 | 0.985 |
| (Ridge VAR = 1) | Full TL | 0.992 | 0.975 | 0.957 | 0.948 | 0.986 |
| Relative | Local LS | 0.978 | 1.016 | 1.079 | 1.079 | 1.064 |
| Prediction CRPS | Nominal TL | 1.189 | 1.034 | 1.082 | 1.072 | 1.073 |
| (Ridge VAR = 1) | Full TL | 1.189 | 1.026 | 1.075 | 1.017 | 1.063 |
| Top Informative State from Detected $\mathcal{A}_W$ | | MN | AZ | WI | OH | LA |


Basu, S. and G. Michailidis (2015, August). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*(4), 1535–1567.

Basu, S., A. Shojaie, and G. Michailidis (2015). Network Granger Causality with Inherent Grouping Structure. *Journal of Machine Learning Research 16*(13), 417–453.

Bird, J. J., J. Kobylarz, D. R. Faria, A. Ekárt, and E. P. Ribeiro (2020). Cross-Domain MLP and CNN Transfer Learning for Biological Signal Processing: EEG and EMG. *IEEE Access 8*, 54789–54801.

Dai, D., P. Rigollet, and T. Zhang (2012, June). Deviation optimal learning using greedy $Q$-aggregation. *The Annals of Statistics 40*(3), 1878–1905.

He, Z., Y. Sun, and R. Li (2024, April). TransFusion: Covariate-Shift Robust Transfer Learning for High-Dimensional Regression. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 703–711. PMLR.

Hsu, D., S. M. Kakade, and T. Zhang (2011, November). Robust Matrix Decomposition With Sparse Corruptions. *IEEE Transactions on Information Theory 57*(11), 7221–7234.

Iman, M., H. R. Arabnia, and K. Rasheed (2023, April). A Review of Deep Transfer Learning and Recent Advancements. *Technologies 11*(2), 40.

Kim, H. E., A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt (2022, April). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging 22*(1), 69.

Li, S., T. T. Cai, and H. Li (2022, February). 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.

Nicholson, W. B., I. Wilms, J. Bien, and D. S. Matteson (2020). High Dimensional Forecasting via Interpretable Vector Autoregression. *Journal of Machine Learning Research 21*(166), 1–52.

Niu, S., Y. Liu, J. Wang, and H. Song (2020, October). A Decade Survey of Transfer Learning (2010–2020). *IEEE Transactions on Artificial Intelligence 1*(2), 151–166.

Rudelson, M. and R. Vershynin (2013, January). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability 18*(none), 1–9.

Tan, C., F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu (2018). A Survey on Deep Transfer Learning. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pp. 270–279. Springer, Cham.

Tian, Y. and Y. Feng (2023, October). Transfer Learning Under High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association 118*(544), 2684–2697.

Wolf, T., V. Sanh, J. Chaumond, and C. Delangue (2019, February). TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents.

Xu, Y. and G. Michailidis (2025, May). Joint Learning of Panel VAR models with Low Rank and Sparse Structure.

Zhao, B., C. Ma, and M. Kolar (2024, November). Trans-Glasso: A Transfer Learning Approach to Precision Matrix Estimation.

Zhao, P. and B. Yu (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research 7*(90), 2541–2563.

# A Rate of Pretraining

Note that for the auxiliary PVAR model, when conditioning on a fixed low-rank $\widehat{\Phi}$, the joint estimation algorithm with respect to $(W_m, S_m)$ is equivalent to individual estimations on the sub-models independently. In this section, we show that the average of the estimation errors in the panel can have an enhanced finite-sample rate under mild conditions.

We denote that, for any $m$ (including $m = 0$),

$$\begin{aligned} z_m &= \begin{pmatrix} \mathrm{diag}(W_m) \\ \mathrm{vec}(S_m') \end{pmatrix} = \begin{pmatrix} z_m^W \\ z_m^S \end{pmatrix}, \quad \widehat{z}_m = \begin{pmatrix} \mathrm{diag}(\widehat{W}_m) \\ \mathrm{vec}(\widehat{S}_m') \end{pmatrix} = \begin{pmatrix} \widehat{z}_m^W \\ \widehat{z}_m^S \end{pmatrix} \in \mathbb{R}^{p^2+p}, \\ \mathbf{X}_m &= \left(\mathrm{blkdiag}(\{X_m' \Phi' e_i\}_{i=1}^p), I_p \otimes X_m'\right) = (\mathbf{X}_m^W, \mathbf{X}_m^S) \in \mathbb{R}^{pT_0 \times (p^2+p)}, \\ \widehat{\mathbf{X}}_m &= \left(\mathrm{blkdiag}(\{X_m' \widehat{\Phi}' e_i\}_{i=1}^p), I_p \otimes X_m'\right) = (\widehat{\mathbf{X}}_m^W, \mathbf{X}_m^S) \in \mathbb{R}^{pT_0 \times (p^2+p)}. \end{aligned} \tag{15}$$

In addition, if we denote $\Phi = \mathrm{blkdiag}(\{e_i' \Phi\}_{i=1}^p)$ and $\widehat{\Phi} = \mathrm{blkdiag}(\{e_i' \widehat{\Phi}\}_{i=1}^p)$, then

$$\mathbf{X}_m = \mathbf{X}_m^S(\Phi', I), \qquad \widehat{\mathbf{X}}_m = \mathbf{X}_m^S(\widehat{\Phi}', I).$$

Similar to Assumption 2, we assume the convex conditions also hold in the panel of auxiliary models with the same constant $\beta$. We further assume that all these data gram matrices are upper-bounded by $\alpha$. Therefore,

$$\beta \leqslant \lambda_p(\frac{X_m X_m'}{T_m}) \leqslant \lambda_1(\frac{X_m X_m'}{T_m}) \leqslant \alpha, \ \forall m = 0, \ldots, M. \tag{16}$$

Before proceeding to the main result, we also need the following essential assumption on Irrepresentable Condition (IRC) which guarantees that the sparse component cannot be well-approximated by any dense combination of the predictors.

*Assumption* 7 (Irrepresentable Condition). For any $m$ (including the target $m = 0$) and any subspace $G \subset \mathbb{R}^{p^2}$ with $|G| \leqslant 2s$, there exists a constant $\eta$,

$$\| [\mathbf{X}_m^S]_{G^\perp}' [\mathbf{X}_m^S]_G \left( [\mathbf{X}_m^S]_G' [\mathbf{X}_m^S]_G \right)^{-1} \| \leqslant \eta,$$

where $[*]_G$ and $[*]_{G^\perp}$ represent the columns of the matrix $*$ with indices from $G$ and $G^\perp$ respectively.

In the literature like Zhao and Yu (2006); Basu et al. (2015), the IRC is usually defined with a constant order for $\| [\mathbf{X}_m^S]_{G^\perp}' [\mathbf{X}_m^S]_G \left( [\mathbf{X}_m^S]_G' [\mathbf{X}_m^S]_G \right)^{-1} \|_{\infty \to \infty}$. Based on the equivalence between the two induced matrix norms, we may expect $\eta$ to satisfy $\eta \lesssim \frac{1}{\sqrt{s}}$. With Assumption 7, it is then possible to show restricted strong convexity (RSC) of the design with respect to sparse vector spaces.

**Lemma A.1.** *Suppose Assumptions 3 and 7 are satisfied, and (16) holds as an extended version of Assumption 2. Let $\eta \leqslant \frac{\beta}{3\alpha}$. Denote the data Gram matrix*

$$\mathbf{G}_m = \frac{\widehat{\mathbf{X}}' \widehat{\mathbf{X}}_m}{T_m} = \frac{1}{T_m} \begin{pmatrix} \widehat{\Phi} (\mathbf{X}_m^S)' \mathbf{X}_m^S \widehat{\Phi}' & \widehat{\Phi} (\mathbf{X}_m^S)' \mathbf{X}_m^S \\ (\mathbf{X}_m^S)' \mathbf{X}_m^S \widehat{\Phi}' & (\mathbf{X}_m^S)' \mathbf{X}_m^S \end{pmatrix} = \begin{pmatrix} \mathbf{G}_m^{WW} & \mathbf{G}_m^{WS} \\ \mathbf{G}_m^{SW} & \mathbf{G}_m^{SS} \end{pmatrix},$$

1. *Then for any vector $v^S \in \mathbb{R}^{p^2}$ with $\|v^S\|_0 \leqslant 2s$, denote the Schur complement of the top-left block in $\mathbf{G}_m$ as*

$$\mathbf{C}_m = \mathbf{G}_m^{SS} - \mathbf{G}_m^{SW} (\mathbf{G}_m^{WW})^{-1} \mathbf{G}_m^{WS},$$

*and it satisfies that*

$$(v^S)' \mathbf{C}_m v^S \geqslant \frac{3\beta}{4} \|v^S\|^2.$$

2. *Then for any vector* $v = \begin{pmatrix} v^W \\ v^S \end{pmatrix} \in \mathbb{R}^{p^2+p}$ *with* $\|v^S\| \leqslant 2s$, $\mathbf{G}_m$ *satisfies that*

$$v'\mathbf{G}_m v \geqslant \frac{\beta}{2}\|v\|^2.$$

*Proof.* Here we slightly abuse the notation $\varphi$ from Assumption 3 such that $\varphi = \|e_i'\widehat{\Phi}\|$ for all $i$ from 1 to $p$, then automatically $\varphi \geqslant 1$.

1. The $\beta$-convexity of $\mathbf{G}_m^{SS}$ implies that

$$g_1 = (v^S)'\mathbf{C}_m v^S \geqslant \beta\|v^S\|^2 - \frac{1}{\beta\varphi^2}\|\frac{\widehat{\Phi}(\mathbf{X}_m^S)'\mathbf{X}_m^S v^S}{T_m}\|^2.$$

Utilizing Assumption 7 with $G$ being the support of $v^S$, we can bound that

$$\begin{aligned}\|\frac{\widehat{\Phi}(\mathbf{X}_m^S)'\mathbf{X}_m^S v^S}{T_m}\| &\leqslant \|\frac{[\widehat{\Phi}]_{G^\perp}[\mathbf{X}_m^S]'_{G^\perp}\mathbf{X}_m^S v^S}{T_m}\| + \|\frac{[\widehat{\Phi}]_G[\mathbf{X}_m^S]'_G\mathbf{X}_m^S v^S}{T_m}\| \\ &\leqslant \alpha\left(\varphi\eta + \|[\widehat{\Phi}]_G\|\right)\|v^S\|.\end{aligned}$$

The incoherence between $\widehat{\Phi}$ and the sparse component indicates that $\frac{\|[\widehat{\Phi}]_G\|}{\varphi} \lesssim \frac{\sqrt{s}}{p}$, then under valid sparse structures with $s = o(p^2)$, we have $\frac{\|[\widehat{\Phi}]_G\|}{\varphi} = o(\eta)$, and

$$g_1 \geqslant \left(\beta - \frac{\alpha^2\left(\varphi\eta + \|[\widehat{\Phi}]_G\|\right)^2}{\beta\varphi^2}\right)\|v^S\|^2 \geqslant \frac{3\beta}{4}\|v^S\|^2.$$

2. Observe that

$$\begin{aligned}g_2 = v'\mathbf{G}_m v &= (v^W)'\mathbf{G}_m^{WW}v^W + (v^S)'\mathbf{G}_m^{SS}v^S + 2(v^W)'\mathbf{G}_m^{WS}v^S \\ &\geqslant \beta\varphi^2\|v^W\|^2 + \beta\|v^S\|^2 - \frac{\beta\varphi^2}{2}\|v^W\|^2 - \frac{2}{\beta\varphi^2}\|\frac{\widehat{\Phi}(\mathbf{X}_m^S)'\mathbf{X}_m^S v^S}{T_m}\|^2 \\ &\geqslant \frac{\beta}{2}\|v^W\|^2 + \left(\beta - \frac{2\alpha^2(\varphi\eta + \|[\widehat{\Phi}]_G\|)^2}{\beta\varphi^2}\right)\|v^S\|^2 \\ &\geqslant \frac{\beta}{2}\|v\|^2\end{aligned}$$

The first inequality comes from Cauchy-Schwarz, and the second last inequality uses the conclusion from the above derivation about $\mathbf{C}_m$.

□

In addition, we also prepare the following lemma to give analogous high-probability deviation bounds for the norms of $(X_m X_m')^{-1}X_m E_m'$. It can also be used as the baseline for bounding the sum $\sum_{m\in\mathcal{A}}(X_m X_m')^{-1}X_m E_m'$; refer to Xu and Michailidis, 2025, Proposition 2 for more details.

**Lemma A.2.** *Assume $X, E \in \mathbb{R}^{p\times T}$ are respectively the data matrices for 1-lagged observations and innovations from a length-$T$ realization of a centered stationary Gaussian VAR(1) model. Condition on $\frac{XX'}{T}$ being $\beta$-convex with $\beta > 0$, there exist constants $c_1, c_2, c_3 > 0$ such that for any $u, v \in \mathbb{R}^p$ with $\|u\| \leqslant 1$, $\|v\| \leqslant 1$, we have*

$$P(|u'(XX')^{-1}XE'v| \geqslant c_1\eta) \leqslant c_2 \exp(-c_3 T \min(\eta, \eta^2)).$$

*Proof.* The classical results, in Basu and Michailidis (2015) for instance, show that

$$P(|u'\frac{XE'}{T}v| \geqslant c_1\eta) \leqslant c_2 \exp(-c_3 T \min(\eta, \eta^2)). \tag{17}$$

Then consider arbitrary $u \in \mathbb{R}^p$ with $\|u\| \leqslant 1$, there exists a corresponding vector $\tilde{u} = \beta(\frac{XX'}{T})^{-1}u \in \mathbb{R}^p$ with $\|\tilde{u}\| \leqslant 1$ such that

$$u'(XX')^{-1}XE'v = \left((\frac{XX'}{T})^{-1}u\right)'\frac{XE'}{T}v \leqslant \frac{1}{\beta}\tilde{u}'\frac{XE'}{T}v.$$

And therefore for such arbitrary unit vectors $u$ and $v$,

$$P(|u'(XX')^{-1}XE'v| \geqslant c_1\eta) \leqslant P(|\tilde{u}'\frac{XE'}{T}v| \geqslant c_1\beta\eta).$$

Then with slight manipulation of the constants in (17), we get the result. □

Below we supplement Assumption 4 with the refined rates for the other pretrained estimators under mild conditions, and provide key evidence to support the claim of good initialization for Lemma B.1.

**Proposition A.1.** *Suppose Assumptions 3 and 7 are satisfied, and (16) holds as an extended version of Assumption 2. In addition, let $\eta \gg \frac{\sqrt{s}}{p}$. Then for an arbitrary index set $\mathcal{A} \subset \{1, \ldots, M\}$ that satisfies $\log(p) \gtrsim \log(|\mathcal{A}|)$, there exist constants $c_1, c_2, c_3$, such that*

1. *For $T \gtrsim p$ and $\eta \lesssim \sqrt{\frac{p}{s\log(p)|\mathcal{A}|}}$, with probability at least $1 - c_2\exp(-c_3\log(p))$,*

$$\frac{1}{|\mathcal{A}|}\|\sum_{m\in\mathcal{A}}(\widehat{W}_m - W_m)\|_F \leqslant c_1\left(\sqrt{\frac{rp}{MT}} + \sqrt{\frac{p}{|\mathcal{A}|T}}\right).$$

2. *For $T \gtrsim \log(p)$, $\eta \lesssim |\mathcal{A}|^{-1/4}$, and $\mathcal{P}$ a $3s$-sparse index set ($|\mathcal{P}| \leqslant 3s$), with probability at least $1 - c_2\exp(-c_3\log(p))$,*

$$\frac{1}{|\mathcal{A}|}\|\Pi_{\mathcal{P}}(\sum_{m\in\mathcal{A}}\widehat{S}_m - S_m)\|_F \leqslant c_1\left(\sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{|\mathcal{A}|T}}\right).$$

*Proof.* Recall the actual model is of the form $\text{vec}(Y_m') = \mathbf{X}_m z_m + \text{vec}(\epsilon_m')$, the first-order optimality of $\hat{z}_m^W$ gives

$$\begin{aligned}
\frac{(\hat{\mathbf{X}}_m^W)'\hat{\mathbf{X}}_m^W}{T_m}\hat{z}_m^W &= -\frac{(\hat{\mathbf{X}}_m^W)'\mathbf{X}_m^S}{T_m}\hat{z}_m^S + \frac{(\hat{\mathbf{X}}_m^W)'\,\text{vec}(Y_m')}{T_m}, \\
\frac{(\mathbf{X}_m^S)'(I - \hat{P}_m)\mathbf{X}_m^S}{T_m}\hat{z}_m^S &= \frac{(\mathbf{X}_m^S)'(I_{p^2} - \hat{P}_m)\,\text{vec}(Y_m')}{T_m} - \lambda\hat{\zeta}_m,
\end{aligned} \tag{18}$$

where $\widehat{P}_m = \widehat{\mathbf{X}}_m^W \left((\widehat{\mathbf{X}}_m^W)'\widehat{\mathbf{X}}_m^W\right)^{-1} (\widehat{\mathbf{X}}_m^W)'$, and $\widehat{\zeta}_m$ is the subgradient of $\|\widehat{z}_m^S\|_1$. Denote

$$\Delta_m = \widehat{z}_m - z_m = \begin{pmatrix} \Delta_m^W \\ \Delta_m^S \end{pmatrix}.$$

We first show the high-probability bound for $\max_{m\in\mathcal{A}} \|\Delta_m^S\|_F$.

Plug in $\mathrm{vec}(Y_m')$ to the second equation of (18),

$$\frac{(\mathbf{X}_m^S)'(I_{p^2} - \widehat{P}_m)\mathbf{X}_m^S}{T_m}\Delta_m^S = \frac{(\mathbf{X}_m^S)'(I - \widehat{P}_m)(\mathbf{X}_m^W - \widehat{\mathbf{X}}_m^W)z_m^W}{T_m} + \frac{(\mathbf{X}_m^S)'(I - \widehat{P}_m)\,\mathrm{vec}(\epsilon_m')}{T_m} - \lambda\widehat{\zeta}_m.$$

We observe that $\frac{(\mathbf{X}_m^S)'(I-\widehat{P}_m)\mathbf{X}_m^S}{T_m}$ is exactly the Schur complement $\mathbf{C}_m$, hence is restricted strong convex with respect to the $2s$-sparse vectors according to Lemma A.1, i.e.,

$$(\Delta_m^S)'\frac{(\mathbf{X}_m^S)'(I - \widehat{P}_m)\mathbf{X}_m^S}{T_m}\Delta_m^S \geqslant \frac{3\beta}{4}\|\Delta_m^S\|^2.$$

Then, with appropriate choice of $\lambda$ (for instance $\|\frac{(\mathbf{X}_m^S)'(I-\widehat{P}_m)\,\mathrm{vec}(\epsilon_m')}{T_m}\|_\infty \lesssim \lambda$), there exists a universal constant $C_1 > 0$, such that

$$\|\Delta_m^S\| \leqslant C_1\left(\sqrt{\frac{rs}{MT}} + \sqrt{s}\left\|\frac{(\mathbf{X}_m^S)'(I - \widehat{P}_m)\,\mathrm{vec}(\epsilon_m')}{T_m}\right\|_\infty\right).$$

Then classical deviation bounds argument implies that, with constants $C_2, C_3 > 0$,

$$P\left(\|\Delta_m^S\| \geqslant C_1(\sqrt{\frac{rs}{MT}} + \sqrt{s}\xi)\right) \leqslant C_2 p^2 \exp(-C_3 T_m \min(\xi, \xi^2)),$$

and taking the union bound for $m \in \mathcal{A}$, we have

$$P\left(\max_{m\in\mathcal{A}}\|\Delta_m^S\| \geqslant C_1(\sqrt{\frac{rs}{MT}} + \sqrt{s}\xi)\right) \leqslant C_2|\mathcal{A}|p^2 \exp(-C_3 T \min(\xi, \xi^2)).$$

Therefore, if $T \gtrsim \log(p)$ and $\log(p) \gtrsim \log(|\mathcal{A}|)$, there exist constants $c_1, c_2, c_3 > 0$, with probability at least $1 - c_2\exp(-c_3\log(p))$,

$$P\left(\max_{m\in\mathcal{A}}\|\Delta_m^S\| \leqslant c_1(\sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{T}})\right) \geqslant 1 - c_2\exp(-c_3\log(p)). \tag{19}$$

1. For the first part,

$$\bar{\Delta}^W = -\frac{1}{|\mathcal{A}|}\sum_{m\in\mathcal{A}}\left((\widehat{\mathbf{X}}_m^W)'\widehat{\mathbf{X}}_m^W\right)^{-1}(\widehat{\mathbf{X}}_m^W)'\left(\mathbf{X}_m^S\Delta_m^S + (\mathbf{X}_m^W - \widehat{\mathbf{X}}_m^W)z_m^W + \mathrm{vec}(\epsilon_m')\right). \tag{20}$$

Then take the norm at both sides and utilize (19), for updated constants $c_1, c_2, c_3 > 0$, with probability at least $1 - c_2\exp(-c_3\log(p))$,

$$\begin{aligned}\|\bar{\Delta}^W\| &\lesssim \frac{\alpha\eta}{\beta}\max_{m\in\mathcal{A}}\|\Delta_m^S\| + \sqrt{\frac{rp}{MT}} + \|\frac{1}{|\mathcal{A}|}\sum_{m\in\mathcal{A}}\left((\widehat{\mathbf{X}}_m^W)'\widehat{\mathbf{X}}_m^W\right)^{-1}(\widehat{\mathbf{X}}_m^W)'\,\mathrm{vec}(\epsilon_m')\| \\ &\lesssim \frac{\alpha\eta}{\beta}(\sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{T}}) + \sqrt{\frac{rp}{MT}} + \sqrt{\frac{p}{|\mathcal{A}|T}}.\end{aligned}$$

Given $\frac{\alpha\eta}{\beta}\sqrt{s\log(p)} \lesssim \sqrt{\frac{p}{|\mathcal{A}|}}$ and $\frac{\alpha\eta}{\beta}\sqrt{s} \lesssim \sqrt{p}$, this high probability upper bound becomes

$$\|\bar{\Delta}^W\| \leqslant c_1 \left(\sqrt{\frac{rp}{MT}} + \sqrt{\frac{p}{|\mathcal{A}|T}}\right).$$

2. For the second part, we can reorganize the second equation of (18) as

$$\Delta_m^S = \left((\mathbf{X}_m^S)'\mathbf{X}_m^S\right)^{-1}\left((\mathbf{X}_m^S)'\hat{P}_m\mathbf{X}_m^S\Delta_m^S + (\mathbf{X}_m^S)'(I-\hat{P}_m)(\mathbf{X}_m^W - \hat{\mathbf{X}}_m^W)z_m^W\right) + \left((\mathbf{X}_m^S)'\mathbf{X}_m^S\right)^{-1}\left((\mathbf{X}_m^S)'(I-\hat{P}_m)\operatorname{vec}(\epsilon_m') - \lambda T\widehat{\zeta}_m\right). \quad (21)$$

Then taking the average over $\mathcal{A}$, we have

$$\bar{\Delta}^S = \frac{1}{|\mathcal{A}|}\sum_{m\in\mathcal{A}}\left((\mathbf{X}_m^S)'\mathbf{X}_m^S\right)^{-1}(\mathbf{X}_m^S)'\hat{P}_m\mathbf{X}_m^S\Delta_m^S + \left((\mathbf{X}_m^S)'\mathbf{X}_m^S\right)^{-1}(\mathbf{X}_m^S)'(I-\hat{P}_m)(\mathbf{X}_m^W - \hat{\mathbf{X}}_m^W)z_m^W + \frac{1}{|\mathcal{A}|}\sum_{m\in\mathcal{A}}\left((\mathbf{X}_m^S)'\mathbf{X}_m^S\right)^{-1}(\mathbf{X}_m^S)'(I-\hat{P}_m)\operatorname{vec}(\epsilon_m') - \lambda T\left((\mathbf{X}_m^S)'\mathbf{X}_m^S\right)^{-1}\hat{\zeta}_m.$$

Taking the inner product with $\Pi_{\mathcal{P}}(\bar{\Delta}^S)$, for updated constants $c_1, c_2, c_3 > 0$, with probability at least $1 - c_2\exp(-c_3\log(p))$, we have

$$\begin{aligned}\|\Pi_{\mathcal{P}}(\bar{\Delta}^S)\|_F &\lesssim \frac{\alpha^2\eta^2}{\beta^2}\max_{m\in\mathcal{A}}\|\Delta_m^S\| + \sqrt{\frac{rs}{MT}} \\ &\quad + \sqrt{3s}\|\frac{1}{|\mathcal{A}|}\sum_{m\in\mathcal{A}}(\mathbf{X}_S'\mathbf{X}_S)^{-1}\mathbf{X}_S'(I-\hat{P})\operatorname{vec}(\epsilon')\|_\infty + \lambda T(\mathbf{X}_S'\mathbf{X}_S)^{-1}\|\widehat{\zeta}_m\|_\infty \\ &\lesssim \frac{\alpha^2\eta^2}{\beta^2}\left(\sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{T}}\right) + \sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{|\mathcal{A}|T}}\end{aligned}$$

Given $\frac{\alpha^2\eta^2}{\beta^2} \lesssim \frac{1}{\sqrt{|\mathcal{A}|}}$, this high probability upper bound becomes

$$\|\Pi_{\mathcal{P}}(\bar{\Delta}^S)\| \leqslant c_1\left(\sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{|\mathcal{A}|T}}\right).$$

□

# B Proof for Oracle Consistency

We first show with the proposition below that the pool average step (5) is a good initialization.

**Lemma B.1.** *Assume the conditions of Proposition A.1, and suppose Assumptions 4 and 5 are satisfied. Given sets $\mathcal{A}_W$ and $\mathcal{A}_S$ as in (4), there exist constants $c, C > 0$, such that with probability at least $1 - c\exp(-C\log(p))$,*

$$\|\bar{W}_0 - W_0\|_F \lesssim \rho_W + h_W, \qquad \|\bar{S}_0 - S_0\|_F \lesssim \rho_S + h_S.$$

*Remark* B.1. This proposition is equivalent to state that, the pool average step (5) of our estimation procedure, under mild conditions, has estimation rate comparable to the initiation approach in other state-of-the-art TL algorithms that require accessibility to the raw auxiliary data for joint fits.

*Proof.* Denote $\nu_m^W = \widehat{W}_m - W_m$ and $\nu_m^S = \widehat{S}_m - S_m$. The first inequality can be immediately derived from the definition of (5) and Proposition A.1, as

$$\bar{W}_0 - W_0 = \frac{1}{|\mathcal{A}_W|}\sum_{m\in\mathcal{A}_W}\nu_m^W + \frac{1}{|\mathcal{A}_W|}\sum_{m\in\mathcal{A}_W}\delta_m^W.$$

For the second part on the sparse components, we denote $\mathcal{P} = \mathcal{P}_S \cup \mathrm{supp}(S_0)$, then Assumption 5 indicates $|\mathcal{P}| \leqslant 3s$ and

$$\begin{aligned}\bar{S}_0 - S_0 &= \frac{1}{|\mathcal{A}_S|}\sum_{m\in\mathcal{A}_S}(\Pi_{\mathcal{P}_S}(\widehat{S}_m) - \Pi_{\mathcal{P}}(\widehat{S}_m)) + \frac{1}{|\mathcal{A}_S|}\sum_{m\in\mathcal{A}_S}\Pi_{\mathcal{P}}(\nu_m^S) + \frac{1}{|\mathcal{A}_S|}\sum_{m\in\mathcal{A}_S}\Pi_{\mathcal{P}}(\delta_m^S)\\ &= \frac{1}{|\mathcal{A}_S|}\Pi_{\mathcal{P}_S^\perp\cap\mathcal{P}}(\sum_{m\in\mathcal{A}_S}\widehat{S}_m) + \frac{1}{|\mathcal{A}_S|}\Pi_{\mathcal{P}}(\sum_{m\in\mathcal{A}_S}\nu_m^S) + \frac{1}{|\mathcal{A}_S|}\sum_{m\in\mathcal{A}_S}\Pi_{\mathcal{P}}(\delta_m^S)\end{aligned}$$

Therefore, denote $\mathcal{A}_S(i,j) = \{m \in \mathcal{A}_S : e_i'\widehat{S}_m e_j \neq 0\}$, then we have $|\mathcal{A}_S(i,j)| < \frac{|\mathcal{A}_S|}{2}$, and

$$\begin{aligned}\|\bar{S}_0 - S_0\|_F &\leqslant \frac{1}{|\mathcal{A}_S|}\|\Pi_{\mathcal{P}_S^\perp\cap\mathcal{P}}(\sum_{m\in\mathcal{A}_S}\widehat{S}_m)\|_F + \frac{1}{|\mathcal{A}_S|}\|\Pi_{\mathcal{P}}(\sum_{m\in\mathcal{A}_S}\nu_m^S)\|_F + \sum_{m\in\mathcal{A}_S}\frac{\|\Pi_{\mathcal{P}}(\delta_m^S)\|_F}{|\mathcal{A}_S|}\\ &\leqslant \frac{1}{|\mathcal{A}_S|}\sqrt{\sum_{(i,j)\in\mathcal{P}_S^\perp\cap\mathcal{P}}\left(\sum_{m\in\mathcal{A}_S(i,j)}e_i'\widehat{S}_m e_j\right)^2} + \frac{1}{|\mathcal{A}_S|}\|\Pi_{\mathcal{P}}(\sum_{m\in\mathcal{A}_S}\nu_m^S)\|_F + h_S.\end{aligned} \tag{22}$$

Observe that for those indices $(i,j) \in \mathcal{P}_S^\perp \cap \mathcal{P}$, we have

$$e_i'\widehat{S}_m e_j = \begin{cases} e_i'(\nu_m^S + \delta_m^S - \delta_n^S - \nu_n^S)e_j, & m \in \mathcal{A}_S(i,j), \forall n \notin \mathcal{A}_S(i,j)\\ 0, & m \notin \mathcal{A}_S(i,j)\end{cases},$$

and we can formulate the following,

$$\sqrt{\sum_{(i,j)\in\mathcal{P}_S^\perp\cap\mathcal{P}}\left(\sum_{m\in\mathcal{A}_S(i,j)}e_i'\widehat{S}_m e_j\right)^2} = \|\sum_{m\in\mathcal{A}_S}(\widetilde{\nu}_m^S + \widetilde{\delta}_m^S)\|_F \leqslant \|\sum_{m\in\mathcal{A}_S}\widetilde{\nu}_m^S\|_F + \sum_{m\in\mathcal{A}_S}\|\widetilde{\delta}_m^S\|_F,$$

where $\widetilde{\nu}_m^S$ satisfies that $\widetilde{\mathcal{P}}_m = \mathrm{supp}(\widetilde{\nu}_m^S) \subset \mathcal{P}$ and $|\widetilde{\nu}_m^S| = |\Pi_{\widetilde{\mathcal{P}}_m}(\nu_m^S)|$. The other terms $\widetilde{\delta}_m^S$ are analogously defined with respect to $\delta_m^S$. Such conditions guarantee that $\|\widetilde{\delta}_m^S\|_F \leqslant \|\delta_m^S\|_F \leqslant h_S$, then

$$\|\bar{S}_0 - S_0\|_F \leqslant \frac{1}{|\mathcal{A}_S|}\|\sum_{m\in\mathcal{A}_S}\widetilde{\nu}_m^S\|_F + \frac{1}{|\mathcal{A}_S|}\|\sum_{m\in\mathcal{A}_S}\nu_m^S\|_F + 2h_S.$$

The arguments of Proposition A.1 are applicable here for both the first and second summands above, hence with probability at least $1 - c\exp(-C\log(p))$,

$$\|\bar{S}_0 - S_0\|_F \lesssim \rho_S + h_S.$$

□

*Proof of Theorem 1.1.* Recall the notations $z_0, \widehat{z}_0, \mathbf{X}_0, \widehat{\mathbf{X}}_0$ as in (15), and similarly denote

$$\delta = z_0 - \begin{pmatrix} \mathrm{diag}(\bar{W}_0) \\ \mathrm{vec}(\bar{S}_0') \end{pmatrix}, \quad \widehat{\delta} = \widehat{z}_0 - \begin{pmatrix} \mathrm{diag}(\bar{W}_0) \\ \mathrm{vec}(\bar{S}_0') \end{pmatrix}.$$

Then we have $\Delta_0 = \widehat{z}_0 - z_0 = \widehat{\delta} - \delta$. According the optimality of $\widehat{\delta}$ in (6), we have the following inequality,

$$\frac{1}{2T_0}\|\widehat{\mathbf{X}}_0\Delta_0\|^2 \leqslant \frac{1}{T_0}\langle \widehat{\mathbf{X}}_0\Delta_0, \mathrm{vec}(\epsilon_0') + \mathbf{X}_0\delta + (\mathbf{X}_0^W - \widehat{\mathbf{X}}_0^W)z_0^W\rangle \\ + \lambda^W(\|\delta^W\|_1 - \|\widehat{\delta}^W\|_1) + \lambda^S(\|\delta^S\|_1 - \|\widehat{\delta}^S\|_1).$$

Then if $\frac{\lambda^W\sqrt{p}}{2} \geqslant \|\frac{(\widehat{\mathbf{X}}_0^W)'\,\mathrm{vec}(\epsilon_0')}{T_0}\|_2$ and $\frac{\lambda^S}{2} \geqslant \|\frac{(\widehat{\mathbf{X}}_0^S)'\,\mathrm{vec}(\epsilon_0')}{T_0}\|_\infty$

$$\frac{\beta}{8}\|\Delta_0\|^2 \leqslant \frac{1}{T_0}\|(\mathbf{X}_0^W - \widehat{\mathbf{X}}_0^W)z_0^W + \mathbf{X}_0\delta\|^2 + \lambda^W\sqrt{p}\left(2\|\delta^W\|_2 - \frac{\|\Delta_0^W\|_2}{2}\right) + \lambda^S\left(2\|\delta^S\|_1 - \frac{\|\Delta_0^S\|_1}{2}\right)$$

1. If $2\lambda^W\sqrt{p}\|\delta^W\|_2 + 2\lambda^S\|\delta^S\|_1 \leqslant \frac{1}{T_0}\|(\mathbf{X}_0^W - \widehat{\mathbf{X}}_0^W)z_0^W + \mathbf{X}_0\delta\|^2$, then with probability at least $1 - c\exp(-C\log(p))$,

$$\|\Delta_0\|^2 \lesssim \frac{1}{T_0}\|(\mathbf{X}_0^W - \widehat{\mathbf{X}}_0^W)z_0^W + \mathbf{X}_0\delta\|^2 \lesssim (\rho_W + \rho_S + h_W + h_S)^2.$$

2. If instead $2\lambda^W\sqrt{p}\|\delta^W\|_2 + 2\lambda^S\|\delta^S\|_1 \geqslant \frac{1}{T_0}\|(\mathbf{X}_0^W - \widehat{\mathbf{X}}_0^W)z_0^W + \mathbf{X}_0\delta\|^2$, then with probability at least $1 - c\exp(-C\log(p))$,

$$\lambda^W\|\Delta_0^W\|_1 + \lambda^S\|\Delta_0^S\|_1 \leqslant 8\lambda^W\sqrt{p}\|\delta^W\|_2 + 8\lambda^S\|\delta^S\|_1 \lesssim \lambda^W\sqrt{p}(\rho_W + h_W) + \sqrt{s}(\rho_S + h_S),$$

and

$$\|\Delta_0\|^2 \lesssim \lambda^W\sqrt{p}\|\delta^W\|_2 + \lambda^S\|\delta^S\|_1 \lesssim \lambda^W\sqrt{p}(\rho_W + h_W) + \lambda^S\sqrt{s}(\rho_S + h_S).$$

Note that $\|\frac{(\widehat{\mathbf{X}}_0^S)'\,\mathrm{vec}(\epsilon_0')}{T_0}\|_\infty \lesssim \sqrt{\frac{\log(p)}{T_0}}$ and $\|\frac{(\widehat{\mathbf{X}}_0^W)'\,\mathrm{vec}(\epsilon_0')}{T_0}\|_2 \lesssim \sqrt{\frac{p}{T_0}}$ with probability at least $1 - c\exp(-C\log(p))$, we can summarize from above that, with probability at least $1 - c\exp(-C\log(p))$,

$$\|\Delta_0\|^2 \lesssim \max\left((\rho_W + h_W + \rho_S + h_S)^2, \sqrt{\frac{p}{T_0}}(\rho_W + h_W) + \sqrt{\frac{s\log(p)}{T_0}}(\rho_S + h_S)\right).$$

□

# C Proof for Detection Guarantees

Before proving Proposition 2.1, we first diverge to state the following lemma of some high-probability bounds that are of essential significance. The proof can be referred to Basu et al. (2019).

**Lemma C.1.** *Based on the model setup (2), suppose Assumption 1 is satisfied. Then there exist constants $c_i > 0$ $(i = 1, 2, 3)$ which may depend on the model parameters $(A_0, \Sigma_0)$, such that:*

1. *for $T_0 \gtrsim p$, with probability at least $1 - c_1\exp(-c_2\log(p))$,*

$$\|\frac{\epsilon_0 X_0'}{T_0}\|_2 \leqslant c_3\sqrt{\frac{p}{T_0}}.$$

*2. for $T_0 \gtrsim \log(p)$, then with probability at least $1 - c_1 \exp(-c_2 \log(p))$,*

$$\|\frac{\epsilon_0 X_0'}{T_0}\|_\infty \leqslant c_3 \sqrt{\frac{\log(p)}{T_0}}.$$

*3. for $T_0 \gtrsim p\Psi^2(h_{X_0})/\psi^2(h_{X_0})$, then with probability at least $1 - c_1 \exp(-c_2 \log(p))$,*

$$\lambda_p(\frac{X_0 X_0'}{T_0}) \geqslant c_3.$$

**Lemma C.2.** *For a linear model $y_i = w^* x_i + \epsilon_i + \varsigma_i$ for $i = 1, \ldots, n$ with univariate response $y_i$, univariate covariate $x_i$, unknown but deterministic deviation terms $\epsilon_i$, and $\varsigma_i$ that are nonzero for only a small portion among the $n$ values. We assume that*

- *The index set $\mathcal{I} = \{i : \varsigma_i \neq 0\}$ for the nonzero $\varsigma_i$ satisfies that $\frac{|\mathcal{I}|}{n} \ll 1$;*
- *The covariates $x_i$ are normalized that $\sum_{i=1}^n x_i^2 = 1$, and bounded that $\max_i |x_i| \lesssim \frac{1}{\sqrt{n}}$.*

*Then the LAD estimator of the regression problem*

$$\hat{w} = \operatorname*{argmin}_w \sum_{i=1}^n |y_i - w x_i|$$

*satisfies that, with $\mathcal{I}^\perp$ denoting the complement of $\mathcal{I}$,*

$$|w^* - \hat{w}| \lesssim \frac{|\sum_{i \in \mathcal{I}^\perp} \epsilon_i \operatorname{sign}(x_i)|}{\sqrt{n}}.$$

*Proof.* Note that the LAD solution $\hat{w}$, by definition, satisfies that

$$\hat{w} = \operatorname*{argmin}_w \sum_{i=1}^n |y_i - w x_i| = \operatorname*{argmin}_w \sum_{i=1}^n |(w - w^*) x_i - \epsilon_i - \varsigma_i|,$$

or equivalently the estimation error $\hat{\delta} = \hat{w} - w^*$ satisfies

$$\hat{\delta} = \operatorname*{argmin}_\delta \sum_{i=1}^n |x_i| \cdot |\delta - \frac{\epsilon_i + \varsigma_i}{x_i}|.$$

The minimizer above is known to be the weighted median of $\frac{\epsilon_i + \varsigma_i}{x_i}$ with weights $|x_i|$. Without loss of generality, assume $x_i \neq 0$, otherwise the zero-weight indicates the observation has zero contribution and can be ignored.

Construct a discrete random variable $Z$ such that $P(Z = \frac{\epsilon_i + \varsigma_i}{x_i}) \propto |x_i|$, then

$$\mathbb{E}[Z|\mathcal{I}^\perp] = \frac{\sum_{i \in \mathcal{I}^\perp} \epsilon_i \operatorname{sign}(x_i)}{\sum_{i \in \mathcal{I}^\perp} |x_i|} \lesssim \frac{|\sum_{i \in \mathcal{I}^\perp} \epsilon_i \operatorname{sign}(x_i)|}{\sqrt{n}(1 - \frac{|\mathcal{I}|}{n})} \lesssim \frac{|\sum_{i \in \mathcal{I}^\perp} \epsilon_i \operatorname{sign}(x_i)|}{\sqrt{n}}.$$

The last step uses the inequality $\frac{1}{1 - \frac{|\mathcal{I}|}{n}} \lesssim 1$ given $\frac{|\mathcal{I}|}{n} \ll 1$. Markov inequality analogously gives that,

$$|\operatorname{median}(Z)| \leqslant \frac{|\operatorname{median}(Z|\mathcal{I}^\perp)|}{1 - \frac{2|\mathcal{I}|}{n}} \lesssim \frac{|\sum_{i \in \mathcal{I}^\perp} \epsilon_i \operatorname{sign}(x_i)|}{\sqrt{n}}.$$

□

*Proof of Proposition 2.1.* The analysis is basically about controlling the norms of the term $\frac{\epsilon_0 X_0'}{T_0}$. In addition to Lemma C.1, we will also need that, for $T_0 \gtrsim p$, and denote $\mathbf{1}_\pm$ an arbitrary vector consisting of only $\pm 1$ and 0, with probability at least $1 - c_1 \exp(-c_2 \log(p))$, then

$$\| \operatorname{diag}(\frac{\epsilon_0 X_0' \hat{\Phi}'}{T_0}) \|_2 \leqslant \sqrt{\varphi^2 \sum_{i=1}^{p} \left( e_i' \frac{\epsilon_0 X_0'}{T_0} \left( \frac{\hat{\Phi}_i}{\|\hat{\Phi}_i\|} \right) \right)'^2} \leqslant c_3 \sqrt{\frac{p \log(p)}{T_0}}. \tag{23}$$

$$\max_i |e_i' \frac{\epsilon_0 X_0'}{T_0} \left( \left( \frac{X_0 X_0'}{T_0} \right)^{-1} \frac{\mathbf{1}_\pm}{\sqrt{p}} \right)| \leqslant c_3 \sqrt{\frac{\log(p)}{T_0}}. \tag{24}$$

For the last steps of the two inequalities, observing that $\|e_i\| \leqslant 1$, $\| \frac{\hat{\Phi}_i}{\|\hat{\Phi}_i\|} \| \leqslant 1$, and $\| \left( \left( \frac{X_0 X_0'}{T_0} \right)^{-1} \frac{\mathbf{1}_\pm}{\sqrt{p}} \right) \| \leqslant 1$, we refer to the Hanson-Wright inequality (Rudelson and Vershynin, 2013) and its extensions as in Basu and Michailidis (2015).

1. Consider the pretrained estimator $\widehat{W}_m$, we define

$$\mathbf{S}_m = S_0 + (W_m - \widehat{W}_m - \delta_m^W)\hat{\Phi} + W_m(\Phi - \hat{\Phi}) + \epsilon_0 X_0^+. \tag{25}$$

Note that the sparsity of $S_0$ implies that the majority of entries in $\mathbf{S}_m$ are identical to those of

$$\mathbf{S}_m - S_0 = (W_m - \widehat{W}_m)\hat{\Phi} - \delta_m^W \hat{\Phi} + W_m(\Phi - \hat{\Phi}) + \epsilon_0 X_0^+.$$

Then, with high probability, the maximum absolute values of the entries in $\mathbf{S}_m - S_0$ are of order $h_\gamma = O_{\max}(\frac{1}{\sqrt{T}} + \sqrt{\frac{r}{MT}} + \sqrt{\frac{\log(p)}{T_0}}) = o(\frac{1}{\sqrt{p}})$ given $T_0 \gtrsim p$ and $T \gtrsim p$. The nonzero elements in $S_0$ are of order at least $\frac{1}{\sqrt{p}}$, the tails of $\mathbf{S}_m$ are primarily contributed by the sparse $S_0$. By trimming an appropriate proportion $\gamma$, for example $\gamma = \frac{\log(p^2)}{p^2}$, at both tails of $\mathbf{S}_m$, we obtain a truncated version

$$Q_\gamma^\gamma(\mathbf{S}_m) = (W_m - \widehat{W}_m)\hat{\Phi} - \delta_m^W \hat{\Phi} + W_m(\Phi - \hat{\Phi}) + \epsilon_0 X_0^+ + (S_0 - \mathbf{S}_m + Q_\gamma^\gamma(\mathbf{S}_m)).$$

Note that the last term is sparse and bounded in the sense that $\|S_0 - \mathbf{S}_m + Q_\gamma^\gamma(\mathbf{S}_m)\|_\infty \lesssim h_\gamma$. Post-multiply $Q_\gamma^\gamma(\mathbf{S}_m)$ by $\frac{X_0 X_0' \hat{\Phi}'}{T_0}$, we get

$$\begin{aligned} \frac{Q_\gamma^\gamma(\mathbf{S}_m) X_0 X_0' \hat{\Phi}'}{T_0} &= \frac{\epsilon_0 X_0' \hat{\Phi}'}{T_0} \\ &\quad + \left( (W_m - \widehat{W}_m)\hat{\Phi} - \delta_m^W \hat{\Phi} + W_m(\Phi - \hat{\Phi}) + (S_0 - \mathbf{S}_m + Q_\gamma^\gamma(\mathbf{S}_m)) \right) \frac{X_0 X_0' \hat{\Phi}'}{T_0}. \end{aligned}$$

With probability at least $1 - c_1' \exp(-c_2 \log(p))$, utilizing (23), the diagonal entries form a vector with the $\ell_2$ norm upper bounded that

$$\begin{aligned} \| \operatorname{diag}(\frac{Q_\gamma^\gamma(\mathbf{S}_m) X_0 X_0' \hat{\Phi}'}{T_0}) \|_2 &\lesssim h_W + \sqrt{\frac{p}{T}} + \sqrt{\frac{rp}{MT}} + \sqrt{\frac{p \log(p)}{T_0}} + \sqrt{\gamma} p h_\gamma \\ &\lesssim h_W + \sqrt{\frac{p}{T}} + \sqrt{\frac{rp}{MT}} + \sqrt{\frac{p \log(p)}{T_0}}. \end{aligned} \tag{26}$$

Then, under condition $C_W$, the upper bounds for all $m \in \mathcal{A}_W$ match the order in (11).

On the other hand, when $m \notin \mathcal{A}_W$, we note that the norm of this diagonal vector is lower bounded in the order of

$$\|\operatorname{diag}(\frac{Q_\gamma^\gamma(\mathbf{S}_m)X_0X_0'\widehat{\Phi}'}{T_0})\|_2 \gtrsim \beta\varphi^2\|\delta_m^W\|_F \gg \sqrt{\frac{p}{T}} + \sqrt{\frac{rp}{MT}} + \sqrt{\frac{p\log(p)}{T_0}}.$$

Hence, given Assumption 6, when $p \to \infty$ under the condition $C_W$, the probability that this lower bound succeeds the upper bound in (26) is approaching 1.

2. In terms of the sparse components, note

$$Y_0X_0^+ - \widehat{S}_m = W_0\widehat{\Phi} + W_0(\Phi - \widehat{\Phi}) + (S_m - \widehat{S}_m) - \delta_m^S + \epsilon_0X_0^+.$$

To minimize the dense effect from the low-rank component $\widehat{\Phi}$ (or $\Phi$), we consider the least absolute deviation (LAD) and regress $Y_0X_0^+ - \widehat{S}_m$ on $\widehat{\Phi}$. The residuals can be obtained as

$$\mathbf{W}_m = (W_0 - \widetilde{W}_m)\widehat{\Phi} + W_0(\Phi - \widehat{\Phi}) + (S_m - \widehat{S}_m) - \delta_m^S + \epsilon X_0^+.$$

In particular, with the LAD estimator definition

$$\widetilde{W}_m = \underset{W=\operatorname{diag}(w)}{\operatorname{argmin}} \|Y_0X_0^+ - \widehat{S}_m - W\widehat{\Phi}\|_1.$$

Denote by $\mathcal{I}_i = \{j : e_i'(S_m - \widehat{S}_m - \delta_m^S)e_j \neq 0\}$ and its corresponding sign vector $\mathbf{1}_{\mathcal{I}_i}$. Assumption 5 suggests $\frac{|\mathcal{I}_i|}{p} \lesssim \frac{s}{p^2} \ll 1$, and then according to Lemma C.2 the estimation error is bounded by

$$\|W_0 - \widetilde{W}\|_F \lesssim h_{LAD} \lesssim \max_i \left( \frac{\|e_iW_0(\Phi - \widehat{\Phi})\|_1}{\sqrt{p}} + \frac{|e_i'\epsilon_0X_0^+\mathbf{1}_{\mathcal{I}_i}|}{\sqrt{p}} \right).$$

Then using Cauchy-Schwartz and (24), with probability at least $1 - c_1\exp(-c_2\log(p))$, this is bounded by

$$h_{LAD} \lesssim \sqrt{\frac{r}{MT}} + \sqrt{\frac{\log(p)}{T_0}}.$$

Post-multiply $\mathbf{W}_m$ by $\frac{X_0X_0'}{T_0}$, and we consider the inner product between $|\widehat{S}_m|$ and $|\frac{\mathbf{W}_mX_0X_0'}{T_0}|$ (here the absolute value operator is element-wise). We have

$$\begin{aligned} \frac{\left|\langle \widehat{S}_m, \frac{\mathbf{W}_mX_0X_0'}{T_0}\rangle\right|}{\|\widehat{S}_m\|_F} &\lesssim h_S + \sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{T}} + \sqrt{\frac{s\log(p)}{T_0}} + \sqrt{s}h_{LAD} \\ &\lesssim h_S + \sqrt{\frac{rs}{MT}} + \sqrt{\frac{s\log(p)}{T}} + \sqrt{\frac{s\log(p)}{T_0}} \end{aligned} \tag{27}$$

Then, under condition $C_S$, the upper bounds for all $m \in \mathcal{A}_S$ match the order in (12).

On the other hand, if $m \notin \mathcal{A}_S$, we note that

$$\begin{aligned}
\frac{\langle |\widehat{S}_m|, |\frac{\mathbf{W}_m X_0 X_0'}{T_0}| \rangle}{\|\widehat{S}_m\|_F} &\gtrsim \frac{\langle |\widehat{S}_m|, |\frac{\delta_m^S X_0 X_0'}{T_0}| \rangle}{\|\widehat{S}_m\|_F} \\
&\gtrsim \frac{\langle |\widehat{S}_m|, |\delta_m^S| \operatorname{diag}(\operatorname{diag}(\frac{X_0 X_0'}{T_0})) \rangle}{\|\widehat{S}_m\|_F} \\
&\gtrsim \frac{\langle |S_m|, |\delta_m^S| \rangle - \langle |S_m - \widehat{S}_m|, |\delta_m^S| \rangle}{\|\widehat{S}_m\|_F} \\
&\gtrsim \frac{\langle |S_m|, |\delta_m^S| \rangle}{\|S_m\|_F} \cdot \frac{\|\widehat{S}_m\|_F - \|S_m - \widehat{S}_m\|_F}{\|\widehat{S}_m\|_F}.
\end{aligned}$$

Here we note that the pretrained model implies that $S_m - \widehat{S}_m$ is dominated by $S_m$ with respect to both support and magnitude. Hence given Assumption 6, when $p \to \infty$ under the condition $C_S$, the probability that this lower bound succeeds the upper bound in (27) is approaching 1.

□

# D Aggregation

In order to mitigate the potential misspecification issues from the detection of Proposition 2.1, we refer to Li et al. (2022) and propose to refine the estimators through an additional Q-aggregation step.

Now we denote $L_W = |\widehat{\mathcal{A}}_W|$ and $L_S = |\widehat{\mathcal{A}}_S|$ as the cardinalities of the detected informative index sets from Proposition 2.1, and construct sequences of the candidate sets sorted by incrementally adding the index with the least corresponding metrics $\mu_m$ and $\kappa_m$ respectively. For example,

$$\varnothing = \widehat{\mathcal{A}}_W^0 \subset \widehat{\mathcal{A}}_W^i \subset \cdots \subset \widehat{\mathcal{A}}_W^{L_W} = \widehat{\mathcal{A}}_W, \tag{28}$$

where $|\widehat{\mathcal{A}}_W^i| = i$, and $\widehat{\mathcal{A}}_W^i - \widehat{\mathcal{A}}_W^{i-1}$ is the index with the $i$-th smallest $\mu_m$. Similar subset sequence is constructed for $\widehat{\mathcal{A}}_S$.

We adopt a sequential split of the data $(X_0, Y_0)$, with the leading 80% ($T_0^{(1)} = \frac{4T_0}{5}$) training data $(X_0^{(1)}, Y_0^{(1)})$ and the trailing 20% ($T_0^{(2)} = \frac{T_0}{5}$) validation data $(X_0^{(2)}, Y_0^{(2)})$. For every pair of combinations $(\widehat{\mathcal{A}}_W^i, \widehat{\mathcal{A}}_S^j)$ we can then run the oracle version Algorithm 1 on $(X_0^{(1)}, Y_0^{(1)})$, and obtain the estimates $(\widehat{W}_{i,j}, \widehat{S}_{i,j})$. Define

$$F_i(W, S) = \frac{1}{T_0^{(i)}} \|(W\widehat{\Phi} + S)X_0^{(i)} - Y_0^{(i)}\|_F^2$$

for $i = 1, 2$. Then the estimation is finally completed with Q-aggregation on the remaining 20% validation sequence $(X_0^{(2)}, Y_0^{(2)})$,

$$\widehat{\beta}^q = \underset{\beta_{ij} \geqslant 0, \|\beta\|_1 = 1}{\operatorname{argmin}} \left( F_2(\sum_{i,j} \beta_{ij} \widehat{W}_{i,j}, \sum_{i,j} \beta_{ij} \widehat{S}_{i,j}) + \sum_{i,j} \beta_{ij} F_2(\widehat{W}_{i,j}, \widehat{S}_{i,j}) + \frac{2q}{T_0} \sum_{i,j} \beta_{ij} \log(\beta_{ij}) \right), \tag{29}$$

with the final estimates

$$\widehat{W}_0^q = \sum_{i,j} \widehat{\beta}_{ij}^q \widehat{W}_{i,j}, \quad \widehat{S}_0^q = \sum_{i,j} \widehat{\beta}_{ij}^q \widehat{S}_{i,j}. \tag{30}$$

**Algorithm 2:** Transfer Learning of $W_0$ and $S_0$

**Input:** Pretrained auxiliary model estimates $\{\widehat{\Phi}, (\widehat{W}_m, \widehat{S}_m)_{m=1}^M\}$, Target data $(X_0, Y_0)$, regularization parameters $\lambda$, Q-aggregation parameter $q$.

**Output:** Estimators $(\widehat{W}_0^q, \widehat{S}_0^q)$.

1 Plug in the pretrained estimators and detect the informative set according to Proposition 2.1.

2 Generate sequence of informative sets as described in (28).

3 For every pair of informative set $(\widehat{\mathcal{A}}_W^i, \widehat{\mathcal{A}}_S^j)$, solve $(\widehat{W}_{i,j}, \widehat{S}_{i,j})$ according to the oracle Algorithm 1.

4 Solve $\widehat{\beta}^q$ from (29).

5 Output target estimator $\widehat{W}_0^q = \sum_{i,j} \widehat{\beta}_{ij}^q \widehat{W}_{i,j}$ and $\widehat{S}_0^q = \sum_{i,j} \widehat{\beta}_{ij}^q \widehat{S}_{i,j}$.

**Theorem D.1.** *Assume the conditions of Proposition 2.1, and* $\|\Sigma_0\| \leqslant \sigma^2$ *for some* $\sigma^2 > 0$. *If* $T_0^{(2)} \gtrsim p$ *or* $T_0^{(2)} \gtrsim L_W L_S$, *then for any* $q \geqslant 2\sigma^2$, *there exist some constants* $c, C > 0$, *with probability at least* $1 - c\exp(-C\min(p, \log(L_W L_S)))$,

$$\|\widehat{W}_0^q - W_0\|_F^2 + \|\widehat{S}_0^q - S_0\|_F^2 \lesssim \min_{i,j} \left( \|\widehat{W}_{i,j} - W_0\|_F^2 + \|\widehat{S}_{i,j} - S_0\|_F^2 \right) + \frac{\log(L_W L_S)}{T_0}. \tag{31}$$

*Proof of Theorem D.1.* Note that the noise covariance with the presence of biased covariates $\widehat{\Phi}$ as opposed to $\Phi$ becomes

$$\widehat{\Sigma}_0 = \mathrm{Var}\left[\epsilon_t^0 + W_0(\widehat{\Phi} - \Phi)X_{t-1}^0\right] = \Sigma_0 + W_0(\widehat{\Phi} - \Phi)A_0\Sigma_0 A_0'(\widehat{\Phi} - \Phi)'W_0,$$

therefore the equivalent covariance scale keeps the same order, as

$$\|\widehat{\Sigma}_0\| \lesssim \sigma^2(1 + \frac{r}{MT}) \lesssim \sigma^2.$$

From Dai et al., 2012, Corollary 3.1, with probability at least $1 - \delta$, we have the prediction error bound

$$\begin{aligned}
&\frac{1}{T_0^{(2)}}\|(\widehat{W}_0^q\widehat{\Phi} + \widehat{S}_0^q)X_0^{(2)} - (W_0\widehat{\Phi} + S_0)X_0^{(2)}\|_F^2 \\
&\qquad \leqslant \min_{i,j} \frac{1}{T_0^{(2)}}\|(\widehat{W}_{i,j}\widehat{\Phi} + \widehat{S}_{i,j})X_0^{(2)} - (W_0\widehat{\Phi} + S_0)X_0^{(2)}\| + \frac{2q}{T_0}\log(L_W L_S/\delta).
\end{aligned}$$

Denote $(\widetilde{W}, \widetilde{S})$ as the pair of candidate estimators that satisfies

$$\|\widetilde{W} - W_0\|_F^2 + \|\widetilde{S} - S_0\|_F^2 = \min_{i,j} \left( \|\widehat{W}_{i,j} - W_0\|_F^2 + \|\widehat{S}_{i,j} - S_0\|_F^2 \right).$$

Consequently, with at least the same probability $1 - \delta$,

$$\begin{aligned}
&\frac{1}{T_0^{(2)}}\|(\widehat{W}_0^q\widehat{\Phi} + \widehat{S}_0^q)X_0^{(2)} - (\widetilde{W}\widehat{\Phi} + \widetilde{S})X_0^{(2)}\|_F^2 \\
&\qquad \lesssim \frac{1}{T_0^{(2)}}\|(\widetilde{W}\widehat{\Phi} + \widetilde{S})X_0^{(2)} - (W_0\widehat{\Phi} + S_0)X_0^{(2)}\|_F^2 + \frac{2q}{T_0}\log(L_W L_S/\delta).
\end{aligned}$$

Then, with the analogous arguments from Lemma A.1, and

$$\|\widehat{W}_0^q - W_0\|_F^2 + \|\widehat{S}_0^q - S_0\|_F^2 \lesssim \|\widehat{W}_0^q - \widetilde{W}\|_F^2 + \|\widehat{S}_0^q - \widetilde{S}\|_F^2 + \|\widetilde{W} - W_0\|_F^2 + \|\widetilde{S} - S_0\|_F^2,$$

and it suffices to bound $\|\widehat{W}_0^q - \widetilde{W}\|_F^2 + \|\widehat{S}_0^q - \widetilde{S}\|_F^2$.

Given the $\frac{\beta}{2}$-RSC of $\mathbf{G}_0$ in Lemma A.1, we have

$$\begin{aligned}\|\widehat{W}_0^q - \widetilde{W}\|_F^2 + \|\widehat{S}_0^q - \widetilde{S}\|_F^2 \leqslant{}& \frac{4}{\beta T_0^{(2)}}\|(\widehat{W}_0^q\hat{\Phi} + \widehat{S}_0^q)X_0^{(2)} - (\widetilde{W}\hat{\Phi} + \widetilde{S})X_0^{(2)}\|_F^2 \\ &+ \frac{4}{\beta}\left\langle \frac{X_0^{(2)}(X_0^{(2)})'}{T_0^{(2)}} - \Sigma_0, \left((\widehat{W}_0^q - \widetilde{W})\hat{\Phi} + \widehat{S}_0^q - \widetilde{S}\right)'\left((\widehat{W}_0^q - \widetilde{W})\hat{\Phi} + \widehat{S}_0^q - \widetilde{S}\right)\right\rangle.\end{aligned}$$

Then classical deviation bounds yield that if $T_0^{(2)} \gtrsim p$ or $T_0^{(2)} \gtrsim L_W L_S$, with probability at least $1 - c_2\exp(-c_3\min(p, L_W L_S))$, the inner product term above is bounded by

$$c_1\sqrt{\frac{\min(p, L_W L_S)}{T_0^{(2)}}}\left(\|\widehat{W}_0^q - \widetilde{W}\|_F^2 + \|\widehat{S}_0^q - \widetilde{S}\|_F^2\right).$$

When $c_1\sqrt{\frac{\min(p,L_W L_S)}{T_0^{(2)}}} \leqslant \frac{1}{2}$, we have with high probability

$$\begin{aligned}\|\widehat{W}_0^q - \widetilde{W}\|_F^2 + \|\widehat{S}_0^q - \widetilde{S}\|_F^2 &\lesssim \frac{1}{T_0^{(2)}}\|(\widehat{W}_0^q\hat{\Phi} + \widehat{S}_0^q)X_0^{(2)} - (\widetilde{W}\hat{\Phi} + \widetilde{S})X_0^{(2)}\|_F^2 \\ &\lesssim \frac{1}{T_0^{(2)}}\|(\widetilde{W}\hat{\Phi} + \widetilde{S})X_0^{(2)} - (W_0\hat{\Phi} + S_0)X_0^{(2)}\|_F^2 + \frac{2q}{T_0}\log(L_W L_S/\delta) \\ &\lesssim \left(\|\widetilde{W} - W_0\|_F^2 + \|\widetilde{S} - S_0\|_F^2\right) + \frac{2q}{T_0}\log(L_W L_S/\delta)\end{aligned}$$

Consequently, set $\delta$ to satisfy $-\log(\delta) \lesssim O_{\min}(p, \log(L_W L_S))$, with probability at least $1 - c\exp(-C\min(p, \log(L_W L_S)))$,

$$\begin{aligned}\|\widehat{W}_0^q - W_0\|_F^2 + \|\widehat{S}_0^q - S_0\|_F^2 &\lesssim \|\widehat{W}_0^q - \widetilde{W}\|_F^2 + \|\widehat{S}_0^q - \widetilde{S}\|_F^2 + \|\widetilde{W} - W_0\|_F^2 + \|\widetilde{S} - S_0\|_F^2 \\ &\lesssim \min_{i,j}\left(\|\widehat{W}_{i,j} - W_0\|_F^2 + \|\widehat{S}_{i,j} - S_0\|_F^2\right) + \frac{\log(L_W L_S)}{T_0}.\end{aligned}$$

□

## D.1 Greedy Aggregation

In cases where the informative indices sets are large, a greedy-flavor alternative, which avoids the exhaustive enumeration of all possible combinations of $(\widehat{\mathcal{A}}_W^i, \widehat{\mathcal{A}}_S^j)$, may be considered in practice to compromise estimation power for computational efficiency.

Such approach starts from $(i, j) = (0, 0)$, and monotonically moves towards $(i, j) = (L_W, L_S)$, with every single step determined by minimizing the evaluation among $\{F_1(\widehat{W}_{i+1,j}, \widehat{S}_{i+1,j}), F_1(\widehat{W}_{i,j+1}, \widehat{S}_{i,j+1})\}$. The detailed pseudo code algorithm is summarized in Algorithm 3.

**Algorithm 3:** Greedy Q-Aggregation

**Input:** Pretrained auxiliary model estimates $\{\widehat{\Phi}, (\widehat{W}_m, \widehat{S}_m)_{m=1}^M\}$, Target data $(X_0, Y_0)$, Candidate informative sets $\widehat{\mathcal{A}}_W$ and $\widehat{\mathcal{A}}_S$, regularization parameter $\lambda$, Q-aggregation parameter $q$.

**Output:** Estimators $(\widehat{W}_0, \widehat{S}_0)$.

1 Initialize $i = j = k = 0$, solve $(\widehat{W}^k, \widehat{S}^k)$ with Algorithm 1 given $\mathcal{A}_W = \widehat{\mathcal{A}}_W^i$ and $\mathcal{A}_S = \widehat{\mathcal{A}}_S^j$.

2 If $i < L_W$, solve $(\widehat{W}_{i+1,j}, \widehat{S}_{i+1,j})$ and evaluate $f_{i+1,j} = F_1(\widehat{W}_{i+1,j}, \widehat{S}_{i+1,j})$. Otherwise set $f_{i+1,j} = \infty$

3 If $j < L_S$, solve $(\widehat{W}_{i,j+1}, \widehat{S}_{i,j+1})$, and evaluate $f_{i,j+1} = F_1(\widehat{W}_{i,j+1}, \widehat{S}_{i,j+1})$. Otherwise set $f_{i,j+1} = \infty$.

4 If $f_{i+1,j} < f_{i,j+1}$, update $i \mapsto i+1$, otherwise update $j \mapsto j+1$. Update $k \mapsto k+1$.

5 Assign $(\widehat{W}^k, \widehat{S}^k) = (\widehat{W}_{i,j}, \widehat{S}_{i,j})$.

6 Repeat steps 2 to 5 until $(i, j) = (L_W, L_S)$, and $k = L_W + L_S$.

7 Solve $\widehat{\beta}^q$ from (29), with the candidate estimates $(\widehat{W}^k, \widehat{S}^k)_{k=0}^{L_W+L_S}$.

8 Output target estimator $\widehat{W}_0^q = \sum_k \widehat{\beta}_k^q \widehat{W}^k$ and $\widehat{S}_0^q = \sum_k \widehat{\beta}_k^q \widehat{S}^k$.

# E Simulation Supplements

As pointed in Remark 1.3, the most significant gains of TL arise when at least some auxiliary models lie within an estimation confidence neighborhood of the target parameters. Recall that the low-rank plus sparse regression using the target data alone typically returns estimators concentrated within the classical confidence region of radii $\sqrt{\frac{p}{T_0}}$ and $\sqrt{\frac{s\log(p)}{T_0}}$. This motivates us to construct the similarity bounds $(h_W, h_S)$ as functions of some effective lengths $T_e = (T_e^W, T_e^S)$ as

$$h_W = O(\sqrt{\frac{p}{T_e^W}}), \qquad h_S = O(\sqrt{\frac{s\log(p)}{T_e^S}}). \tag{32}$$

Specifically,

1. If $T_e^W \ll T_0$ ($T_e^S \ll T_0$), the parameters $W_m$ ($S_m$) of the informative auxiliary models are likely to fall outside the confidence neighborhood of the estimation on the target data alone. As a result, it becomes uncertain whether there can be any gain from transfer.

2. If $T_e^W \gg T_0$ ($T_e^S \gg T_0$), the similarity bounds generally give a tighter cluster pattern than the confidence neighborhood of the estimation on the target data. Then with high probability, the performance of the proposed TL pipeline can be superior than that without transfer.

Consequently, we use $T_e$ as the proxies for $(h_W, h_S)$ to analyze the estimation errors and other metrics including prediction accuracies.

## E.1 Data Generating Mechanism

The data generating process for the simulations are primarily specified with the set of parameters $(M, p, T, T_0, \omega_W, \omega_S, T_e^W, T_e^S, s)$, where $\omega_W = \frac{|\mathcal{A}_W|}{M}$ and $\frac{|\mathcal{A}_S|}{M}$[1].

Here we briefly cover the realization of the cluster structure as described in (3) and (4).

[1]The rank constraint $r$ and the sparsity level $s$, being of minor interest in later discussion, are fixed as $(r, s) = (\frac{p}{5}, p)$ and omitted in the specification.

1. We first randomly sample the indices sets $(\mathcal{A}_W, \mathcal{A}_S)$ of the informative auxiliary models from 1 to $M$ respectively.

2. Then, accordingly for $m \in \{\mathcal{A}_W, 0\}$, with $w_m = \text{diag}(W_m)$ being the vectors of diagonal entries, fix a deterministic vector $\bar{w} \in \mathbb{R}^p$ and randomly sample

$$w_m \sim N(\bar{w}, \frac{1}{T_e^W} I_p).$$

3. For $m \in \{\mathcal{A}_S, 0\}$, sample a common $s$-support $\mathcal{P}$, fix a deterministic vector $\bar{s} \in \mathbb{R}^s$, and assign the entries of $S_m$ on $\mathcal{P}$ using length-$s$ vectors $s_m$ drawn from

$$s_m \sim N(\bar{s}, \frac{\log(p)}{T_e^S} I_s).$$

4. For generality, we may additionally allow some $S_m$ for $m \in \{\mathcal{A}_S, 0\}$ to exhibit a relatively small portion of nonzero perturbations outside the common support $\mathcal{P}$.

5. The remaining routine for generating model parameters and sampling time series follows analogously from Xu and Michailidis (2025).

If we temporarily ignore the nonzero perturbation outside the support $\mathcal{P}$ (step 4), then the above construction guarantees that the resulting cluster structure satisfies

$$\mathbb{E}[\|\delta_m^W\|_F^2] = \frac{4p}{T_e^W}, \qquad \mathbb{E}[\|\delta_m^S\|_F^2] = \frac{4s\log(p)}{T_e^S},$$

for $m \in \mathcal{A}_W$ and $m \in \mathcal{A}_S$ respectively.

### E.2 Correctly Specified Setup Discussion Continued

We have the following settings to be examined under the correctly specified model, while only the results for first setting are summarized in Section 3.1 of the main paper for space concerns. We here provide a complete table of summary statistics for comprehensive comparison.

1. Separate informative sources with $(\omega_W, \omega_S, T_e^W, T_e^S) = (0.4, 0.3, 400, 400)$, especially set $\mathcal{A}_W = \{1, 2, 3, 4\}$ and $\mathcal{A}_S = \{5, 6, 7\}$.

2. All the sparse components are informative: $(\omega_W, \omega_S, T_e^W, T_e^S) = (0, 1, 400, 400)$.

3. All the low-rank components are informative: $(\omega_W, \omega_S, T_e^W, T_e^S) = (1, 0, 400, 400)$.

Table 3 suggests that most of the gains seem to be related to the transfer on the sparse component $S_0$, as the some prior about the support can make a big difference. This is also discovered and discussed later in the tradeoff comparison Section E.3. Our conjecture is that the lower-dimensional dense rescaling $W_0$–being $p$-dimension compared to the $p^2$-dimensional $S_0$–is relatively easier for estimation given the target data alone. Thus, transfer learning tends to play a less critical role when those rescaling factors are the only component that has informative sources from the auxiliary panel.

Table 3: Some summary metrics about the simulation studies of Section 3.1 and Section E.2.

| | | Example 1 | | | | Example 2 | | | | Example 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Algorithm \ $T_0$ | 100 | 400 | 1600 | 6400 | 100 | 400 | 1600 | 6400 | 100 | 400 | 1600 | 6400 |
| Computation Cost | Local LS | 44.484 (1.164) | 71.608 (0.983) | 176.666 (3.687) | 617.858 (20.225) | 44.179 (1.381) | 71.645 (1.305) | 176.958 (3.896) | 624.353 (16.611) | 44.183 (1.360) | 71.244 (1.414) | 177.218 (4.021) | 625.432 (15.896) |
| | PVAR Refit | 197.516 (102.957) | 187.013 (125.633) | 236.564 (113.712) | 287.610 (94.310) | 214.974 (102.781) | 139.272 (56.768) | 143.539 (40.578) | 136.053 (12.291) | 68.611 (45.099) | 58.872 (34.055) | 50.834 (6.239) | 50.291 (2.780) |
| | **Nominal TL** | 0.113 (0.004) | 0.114 (0.004) | 0.122 (0.006) | 0.143 (0.022) | 0.112 (0.004) | 0.112 (0.004) | 0.120 (0.006) | 0.139 (0.011) | 0.106 (0.005) | 0.115 (0.026) | 0.126 (0.007) | 0.147 (0.008) |
| | **Full TL** | 11.155 (0.989) | 12.607 (1.039) | 15.301 (0.760) | 19.576 (0.878) | 6.120 (0.313) | 6.923 (0.248) | 8.012 (0.304) | 10.107 (0.583) | 6.276 (0.167) | 7.752 (0.220) | 9.984 (0.390) | 13.409 (0.521) |
| Estimation Error | Local LS | 2.255 (0.241) | 0.880 (0.070) | 0.300 (0.022) | 0.090 (0.005) | 2.113 (0.230) | 0.809 (0.066) | 0.280 (0.022) | 0.085 (0.005) | 2.843 (0.260) | 0.936 (0.068) | 0.281 (0.018) | 0.078 (0.005) |
| | PVAR Refit | 4.101 (0.505) | 0.425 (0.055) | 0.197 (0.017) | 0.170 (0.006) | 5.003 (0.485) | 0.537 (0.063) | 0.187 (0.018) | 0.151 (0.009) | 2.346 (0.262) | 0.722 (0.068) | 0.571 (0.026) | 0.554 (0.019) |
| | **Nominal TL** | 1.468 (0.227) | 0.428 (0.064) | 0.176 (0.023) | 0.102 (0.006) | 1.378 (0.309) | 0.377 (0.059) | 0.152 (0.012) | 0.100 (0.004) | 2.194 (0.265) | 0.971 (0.133) | 0.462 (0.058) | 0.155 (0.021) |
| | **Full TL** | 0.831 (0.145) | 0.383 (0.063) | 0.172 (0.017) | 0.098 (0.009) | 1.287 (0.273) | 0.371 (0.061) | 0.159 (0.014) | 0.100 (0.004) | 1.800 (0.143) | 0.983 (0.104) | 0.449 (0.061) | 0.154 (0.020) |
| PMSE | Local LS | 0.872 (0.047) | 0.852 (0.033) | 0.850 (0.033) | 0.846 (0.036) | 0.858 (0.043) | 0.836 (0.032) | 0.838 (0.033) | 0.834 (0.036) | 0.600 (0.039) | 0.581 (0.026) | 0.575 (0.026) | 0.573 (0.026) |
| | PVAR Refit | 0.902 (0.050) | 0.845 (0.032) | 0.848 (0.033) | 0.847 (0.036) | 0.903 (0.047) | 0.833 (0.031) | 0.837 (0.033) | 0.835 (0.036) | 0.597 (0.037) | 0.578 (0.026) | 0.578 (0.026) | 0.580 (0.027) |
| | **Nominal TL** | 0.858 (0.045) | 0.844 (0.032) | 0.848 (0.033) | 0.846 (0.036) | 0.846 (0.040) | 0.830 (0.032) | 0.836 (0.033) | 0.834 (0.035) | 0.590 (0.037) | 0.581 (0.026) | 0.576 (0.025) | 0.574 (0.026) |
| | **Full TL** | 0.846 (0.045) | 0.843 (0.032) | 0.848 (0.033) | 0.846 (0.036) | 0.844 (0.043) | 0.830 (0.032) | 0.836 (0.033) | 0.834 (0.035) | 0.586 (0.035) | 0.581 (0.026) | 0.576 (0.026) | 0.574 (0.026) |
| | Oracle | 0.832 (0.043) | 0.836 (0.032) | 0.845 (0.032) | 0.845 (0.036) | 0.822 (0.038) | 0.823 (0.031) | 0.834 (0.032) | 0.832 (0.035) | 0.556 (0.031) | 0.567 (0.025) | 0.570 (0.025) | 0.572 (0.026) |
| CRPS | Local LS | 3.374 (0.432) | 3.308 (0.401) | 3.282 (0.376) | 3.298 (0.403) | 3.362 (0.432) | 3.251 (0.359) | 3.202 (0.395) | 3.252 (0.372) | 2.506 (0.257) | 2.401 (0.207) | 2.455 (0.215) | 2.414 (0.200) |
| | PVAR Refit | 3.500 (0.456) | 3.283 (0.395) | 3.277 (0.382) | 3.298 (0.403) | 3.505 (0.460) | 3.252 (0.333) | 3.203 (0.397) | 3.254 (0.367) | 2.497 (0.235) | 2.402 (0.209) | 2.458 (0.220) | 2.440 (0.209) |
| | **Nominal TL** | 3.337 (0.433) | 3.274 (0.391) | 3.279 (0.387) | 3.296 (0.402) | 3.278 (0.407) | 3.226 (0.333) | 3.204 (0.397) | 3.258 (0.368) | 2.476 (0.238) | 2.414 (0.209) | 2.455 (0.220) | 2.418 (0.202) |
| | **Full TL** | 3.300 (0.437) | 3.277 (0.385) | 3.277 (0.382) | 3.298 (0.400) | 3.284 (0.409) | 3.226 (0.337) | 3.200 (0.400) | 3.258 (0.366) | 2.464 (0.232) | 2.413 (0.216) | 2.458 (0.222) | 2.417 (0.201) |
| | Oracle | 3.226 (0.414) | 3.248 (0.389) | 3.266 (0.374) | 3.291 (0.402) | 3.221 (0.398) | 3.203 (0.341) | 3.191 (0.393) | 3.249 (0.370) | 2.384 (0.223) | 2.372 (0.194) | 2.430 (0.215) | 2.412 (0.200) |

## E.3 Tradeoff Between Cluster Radii and Target Length

In this section we provide numerical evidences to supplement our discussions on the tradeoff between the cluster concentration rate of the generating mechanism $(T_e^W, T_e^S)$ and the problem setups, including panel size $M$, transferrable sources $(\mathcal{A}_W, \mathcal{A}_S)$, and target series length $T_0$. In particular, we fix the dimension $p = 50$, the portions of transferrable auxiliary sources $\omega_W = \omega_S = 0.2$, and consider $M \in \{20, 50\}$, $T_0 \in \{150, 300, 600\}$, and $(T_e^W, T_e^S) \in \{100, 400\} \times \{100, 400\}$.

Particularly from Table 4, we highlight the following observations.

1. The gain from our proposed TL is most notable under the ideal settings where $T_0 \ll T_e^W$ or $T_0 \ll T_e^S$. In contras, when $T_0$ increases and eventually exceeds the effective lengths–in the setting with $T_0 = 600$ and $M = 20$–the reported statistics suggest that the advantage of *Full TL* over *Nominal TL* diminishes or even disappears.

2. The algorithm is more sensitive to the cluster radius associated with the rescaling component $W_0$ $(T_e^W)$, than to that of the sparse component $S_0$ $(T_e^S)$. This may be because even less concentrated auxiliary $S_m$ matrices still implicitly carry valuable information about the support, which can help accelerate the otherwise intensive task of sparse structure recovery.

3. In most of these correctly specified settings, the TL framework, even in its nominal form, outperforms the *Local LS* benchmark, as long as some the relevant pretrained auxiliary panel is available.

## E.4 Robustness Discussion Continued

For the misspecified dense model (13), here are the simulation results after increasing $\omega_W$ and $\omega_S$ to 0.5. Compared to the settings in Section 3.2, the proposed approach is shown by Figure 4 to exhibit more persistent improvement for problems with longer target time series ($T_0 = 400$). This is partially due to the fact that the pretrained low-rank component $\widehat{\Phi}$ fitted from the PVAR model tends to align more closely with the dominant cluster patterns in the auxiliary panel, consequently better captures the baseline structure of the target model. However, it needs to be addressed that when the target model deviates substantially from the dominant cluster in the auxiliary panel, the TL approach may lead to degraded performances.

Table 4: Some summary metrics about the tradeoff concerning target length $T_0$, panel size $M$, and the distance to informative sources, which is given by the effective lengths proxy $(T_e^W, T_e^S)$ as in (32).

| Metrics | M | $T_0$ | 150 | | | | 300 | | | | 600 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $(T_e^W, T_e^S)$ | (100, 100) | (100, 400) | (400, 100) | (400, 400) | (100, 100) | (100, 400) | (400, 100) | (400, 400) | (100, 100) | (100, 400) | (400, 100) | (400, 400) |
| Computation Cost (seconds) | 20 | Local LS | 8.485 (0.286) | 8.445 (0.170) | 8.482 (0.222) | 8.369 (0.292) | 10.501 (0.257) | 10.668 (0.290) | 10.614 (0.300) | 10.580 (0.294) | 14.160 (0.289) | 14.262 (0.272) | 14.218 (0.294) | 14.398 (0.411) |
| | | **Nominal TL** | 0.207 (0.006) | 0.214 (0.006) | 0.207 (0.016) | 0.202 (0.018) | 0.211 (0.016) | 0.222 (0.016) | 0.210 (0.011) | 0.202 (0.006) | 0.210 (0.016) | 0.222 (0.015) | 0.213 (0.020) | 0.207 (0.010) |
| | | **Full TL** | 2.505 (0.155) | 2.545 (0.157) | 2.440 (0.145) | 2.466 (0.140) | 2.664 (0.135) | 2.715 (0.145) | 2.624 (0.147) | 2.583 (0.133) | 2.709 (0.157) | 2.827 (0.127) | 2.702 (0.154) | 2.718 (0.139) |
| | 50 | Local LS | 8.492 (0.232) | 8.347 (0.187) | 8.546 (0.229) | 8.413 (0.202) | 10.422 (0.246) | 10.556 (0.293) | 10.599 (0.194) | 10.683 (0.273) | 14.186 (0.358) | 14.241 (0.375) | 14.310 (0.364) | 14.359 (0.345) |
| | | **Nominal TL** | 0.200 (0.008) | 0.186 (0.007) | 0.215 (0.015) | 0.214 (0.008) | 0.206 (0.016) | 0.195 (0.021) | 0.216 (0.005) | 0.221 (0.018) | 0.204 (0.019) | 0.194 (0.014) | 0.220 (0.017) | 0.223 (0.019) |
| | | **Full TL** | 5.649 (0.287) | 5.294 (0.304) | 5.810 (0.334) | 5.867 (0.388) | 6.041 (0.342) | 5.755 (0.358) | 6.227 (0.259) | 6.105 (0.338) | 6.252 (0.273) | 5.944 (0.202) | 6.588 (0.387) | 6.228 (0.286) |
| Relative Estimation Errors $\frac{\|\hat{A}_0 - A_0\|_F}{\|A_0\|_F}$ | 20 | Local LS | 0.551 (0.014) | 0.666 (0.018) | 0.563 (0.015) | 0.539 (0.016) | 0.414 (0.010) | 0.517 (0.014) | 0.426 (0.011) | 0.404 (0.010) | 0.308 (0.007) | 0.391 (0.009) | 0.317 (0.007) | 0.299 (0.008) |
| | | **Nominal TL** | 0.243 (0.020) | 0.248 (0.026) | 0.271 (0.020) | 0.228 (0.019) | 0.157 (0.017) | 0.170 (0.015) | 0.171 (0.017) | 0.146 (0.011) | 0.096 (0.009) | 0.120 (0.010) | 0.112 (0.009) | 0.105 (0.007) |
| | | **Full TL** | 0.233 (0.019) | 0.227 (0.019) | 0.250 (0.020) | 0.168 (0.014) | 0.154 (0.016) | 0.170 (0.017) | 0.172 (0.019) | 0.132 (0.011) | 0.101 (0.009) | 0.121 (0.011) | 0.113 (0.011) | 0.101 (0.009) |
| | 50 | Local LS | 0.532 (0.016) | 0.458 (0.012) | 0.595 (0.014) | 0.611 (0.017) | 0.400 (0.010) | 0.344 (0.008) | 0.460 (0.012) | 0.473 (0.012) | 0.296 (0.006) | 0.257 (0.006) | 0.351 (0.008) | 0.355 (0.008) |
| | | **Nominal TL** | 0.261 (0.024) | 0.196 (0.031) | 0.281 (0.021) | 0.285 (0.023) | 0.181 (0.014) | 0.133 (0.016) | 0.197 (0.013) | 0.187 (0.018) | 0.119 (0.010) | 0.089 (0.008) | 0.135 (0.010) | 0.112 (0.011) |
| | | **Full TL** | 0.219 (0.019) | 0.145 (0.012) | 0.252 (0.025) | 0.201 (0.015) | 0.158 (0.013) | 0.122 (0.010) | 0.181 (0.017) | 0.169 (0.013) | 0.108 (0.009) | 0.088 (0.010) | 0.126 (0.012) | 0.124 (0.014) |
| PMSE | 20 | Local LS | 0.804 (0.030) | 0.911 (0.026) | 0.829 (0.028) | 0.793 (0.028) | 0.781 (0.024) | 0.891 (0.025) | 0.807 (0.024) | 0.767 (0.026) | 0.769 (0.021) | 0.877 (0.020) | 0.790 (0.020) | 0.758 (0.019) |
| | | **Nominal TL** | 0.750 (0.026) | 0.861 (0.022) | 0.776 (0.021) | 0.736 (0.024) | 0.749 (0.022) | 0.858 (0.022) | 0.775 (0.022) | 0.734 (0.021) | 0.750 (0.019) | 0.859 (0.019) | 0.771 (0.019) | 0.738 (0.018) |
| | | **Full TL** | 0.749 (0.025) | 0.859 (0.022) | 0.774 (0.022) | 0.730 (0.023) | 0.749 (0.022) | 0.858 (0.022) | 0.775 (0.021) | 0.733 (0.022) | 0.750 (0.019) | 0.859 (0.019) | 0.772 (0.019) | 0.738 (0.018) |
| | | Oracle | 0.736 (0.024) | 0.851 (0.022) | 0.761 (0.021) | 0.722 (0.022) | 0.743 (0.021) | 0.854 (0.022) | 0.769 (0.021) | 0.728 (0.021) | 0.747 (0.019) | 0.856 (0.019) | 0.769 (0.019) | 0.735 (0.018) |
| | 50 | Local LS | 0.778 (0.031) | 0.657 (0.031) | 0.861 (0.026) | 0.880 (0.028) | 0.757 (0.023) | 0.644 (0.023) | 0.846 (0.021) | 0.859 (0.023) | 0.745 (0.021) | 0.636 (0.018) | 0.835 (0.020) | 0.843 (0.020) |
| | | **Nominal TL** | 0.728 (0.025) | 0.614 (0.025) | 0.815 (0.022) | 0.831 (0.024) | 0.727 (0.019) | 0.619 (0.020) | 0.815 (0.019) | 0.828 (0.021) | 0.728 (0.020) | 0.621 (0.016) | 0.817 (0.018) | 0.825 (0.018) |
| | | **Full TL** | 0.724 (0.025) | 0.610 (0.024) | 0.811 (0.021) | 0.825 (0.025) | 0.726 (0.019) | 0.619 (0.020) | 0.814 (0.019) | 0.827 (0.021) | 0.727 (0.020) | 0.621 (0.016) | 0.817 (0.018) | 0.825 (0.018) |
| | | Oracle | 0.712 (0.023) | 0.602 (0.024) | 0.798 (0.020) | 0.817 (0.024) | 0.719 (0.019) | 0.613 (0.020) | 0.808 (0.018) | 0.822 (0.020) | 0.724 (0.019) | 0.619 (0.016) | 0.814 (0.018) | 0.822 (0.018) |
| CRPS | 20 | Local LS | 4.151 (0.379) | 4.587 (0.458) | 4.121 (0.417) | 3.973 (0.410) | 3.922 (0.342) | 4.438 (0.430) | 4.026 (0.391) | 3.930 (0.312) | 3.961 (0.364) | 4.353 (0.394) | 4.025 (0.356) | 3.855 (0.321) |
| | | **Nominal TL** | 3.900 (0.352) | 4.356 (0.405) | 3.889 (0.369) | 3.746 (0.359) | 3.783 (0.300) | 4.310 (0.372) | 3.885 (0.365) | 3.780 (0.290) | 3.882 (0.349) | 4.292 (0.393) | 3.940 (0.327) | 3.762 (0.291) |
| | | **Full TL** | 3.891 (0.342) | 4.352 (0.414) | 3.881 (0.360) | 3.715 (0.355) | 3.784 (0.297) | 4.304 (0.372) | 3.891 (0.357) | 3.779 (0.291) | 3.886 (0.350) | 4.292 (0.390) | 3.945 (0.328) | 3.764 (0.290) |
| | | Oracle | 3.841 (0.350) | 4.302 (0.404) | 3.832 (0.352) | 3.681 (0.344) | 3.750 (0.294) | 4.285 (0.377) | 3.857 (0.357) | 3.757 (0.284) | 3.871 (0.346) | 4.289 (0.388) | 3.929 (0.318) | 3.749 (0.284) |
| | 50 | Local LS | 3.963 (0.359) | 3.449 (0.279) | 4.321 (0.425) | 4.398 (0.429) | 3.896 (0.353) | 3.456 (0.307) | 4.204 (0.428) | 4.324 (0.381) | 3.803 (0.274) | 3.355 (0.243) | 4.133 (0.401) | 4.215 (0.375) |
| | | **Nominal TL** | 3.760 (0.327) | 3.268 (0.223) | 4.118 (0.395) | 4.200 (0.412) | 3.777 (0.324) | 3.345 (0.282) | 4.059 (0.405) | 4.176 (0.387) | 3.746 (0.276) | 3.301 (0.221) | 4.056 (0.373) | 4.128 (0.353) |
| | | **Full TL** | 3.747 (0.336) | 3.254 (0.233) | 4.112 (0.383) | 4.165 (0.393) | 3.767 (0.324) | 3.344 (0.286) | 4.054 (0.402) | 4.176 (0.383) | 3.747 (0.277) | 3.304 (0.222) | 4.054 (0.370) | 4.138 (0.362) |
| | | Oracle | 3.690 (0.329) | 3.228 (0.221) | 4.054 (0.370) | 4.129 (0.383) | 3.729 (0.314) | 3.318 (0.279) | 4.025 (0.405) | 4.149 (0.375) | 3.734 (0.281) | 3.292 (0.217) | 4.033 (0.371) | 4.119 (0.360) |

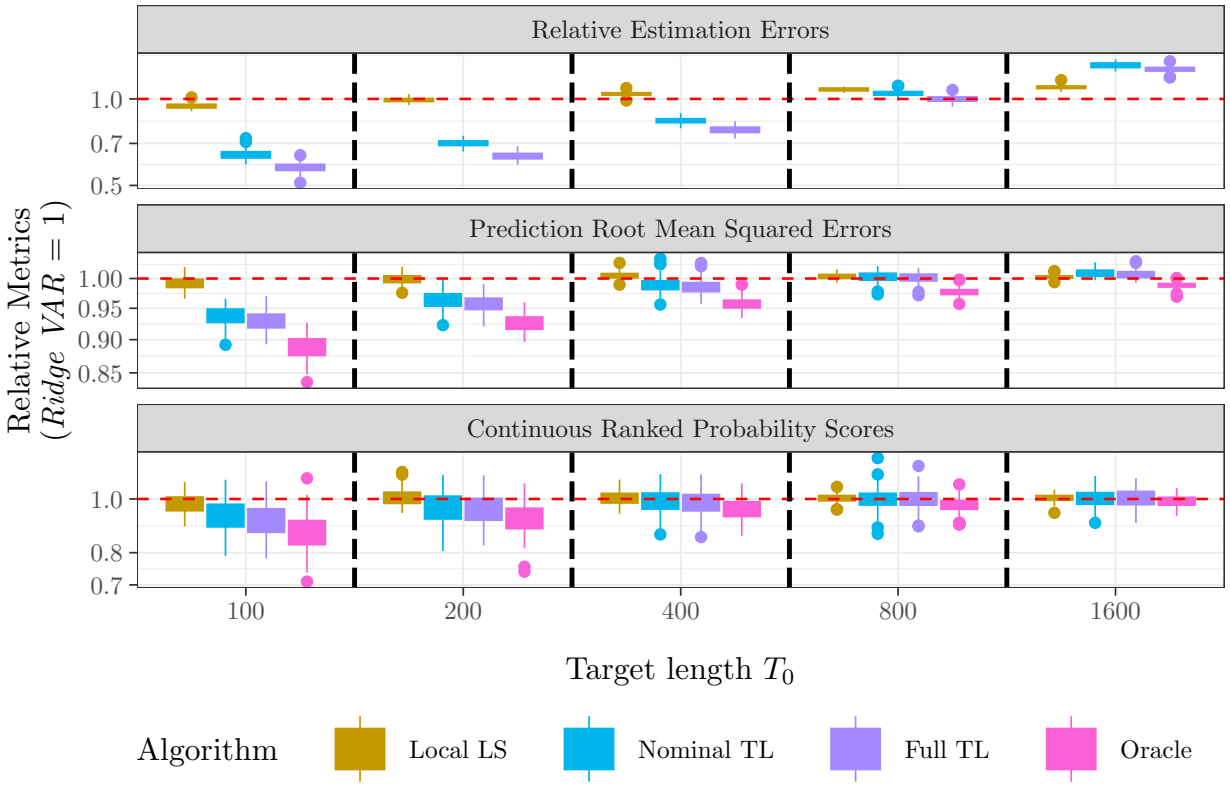


Figure 4: The summary metrics on a misspecified panel of dense VAR models (13), but with increased $\omega_W = \omega_S = 0.5$ for more informative sources.