Statistica Sinica Preprint No: SS-2025-0335	
Title	Statistical Inference for High-dimensional Time Dependent Linear Models with Knowledge Transfer
Manuscript ID	SS-2025-0335
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0335
Complete List of Authors	Zongqi Liu, Shengji Jia and Xiao Guo
Corresponding Authors	Xiao Guo
E-mails	xiaoguo@ustc.edu.cn
Notice: Accepted author version.	

Statistical Inference for High-Dimensional Time Dependent Linear Models with Knowledge Transfer

Zongqi Liu[1], Shengji Jia[2], and Xiao Guo[1]

[1] *University of Science and Technology of China and*

[2] *Shanghai Lixin University of Accounting and Finance*

Abstract: This paper explores high-dimensional time dependent regression models within a transfer learning framework. Specifically, we develop an estimator for the regression parameters of the high-dimensional linear models in time series settings based on transfer learning, and establish the convergence rate of the proposed estimator. Our results reveal that leveraging auxiliary data can substantially enhance the convergence rate of the proposed estimator compared to the traditional single-task approaches. For statistical inference of the target regression coefficients, we propose a novel debiased method based on transfer learning that incorporates a banded estimator of the error autocovariance matrix and demonstrate its asymptotic normality. To mitigate the risk of negative transfer, we develop a transferable source detection algorithm that adapts to data dependence, guaranteeing correct selection of auxiliary samples that are sufficiently similar to the target samples. By leveraging information from multiple tasks, our method enhances both robustness and accuracy in the estimation process, ultimately improving statistical inference performance. Numerical simulations and a real-data experiment reveal significant improvements in estimation and inference accuracy compared to both the single-task Lasso regression and the transfer learning methods for independent data.

Key words and phrases: high-dimensional linear models; source detection; statistical inference; time series data; transfer learning

1. Introduction

In contemporary statistical applications, high-dimensional time series datasets from multiple related sources are commonly encountered. Such data structures are prevalent across various domains: regional economic indicators collected across different geographic areas, stock returns from companies within the same industry, environmental measurements from multiple monitoring stations, and clinical outcomes tracked across different medical centers, among others. However, the high dimensionality together with the temporal dependence poses unique statistical challenges that require specialized methodological developments beyond traditional approaches designed for independent observations.

Numerous studies have focused on high-dimensional regression models with time series data. For example, Wu and Wu (2016) examined high-dimensional linear models with dependent non-Gaussian errors and/or covariates, and demonstrated the asymptotic property for the Lasso estimator under a deterministic design, within the framework of functional dependence (Wu, 2005). Chernozhukov et al. (2021) addressed both temporal and cross-sectional dependence in high-dimensional regression systems, and proposed desparsified procedures for simultaneous inference on the parameters. Yuan and Guo (2022) derived the asymptotic normality of the debiased Lasso estimator via m-dependent approximations. Adamek et al. (2023) constructed a general inference framework for high-dimensional time series models under the near-epoch dependence assumption. More recently, Xia et al. (2024) studied inference for the low-dimensional parameters of the high-dimensional linear model under locally stationary error processes. However, these methods typically assume the availability of the target data and cannot leverage other relevant datasets to improve the performance of the target model. Moreover, as pointed out by Li (2020), debiased

estimators under single-task learning frameworks may still suffer from large bias. To address these limitations, integrating multiple datasets has become an increasingly promising strategy for enhancing estimation and inference accuracy in high-dimensional time series settings.

In the era of big data, it is often feasible to access related datasets alongside the limited data from the target task. Transfer learning aims to improve performance on the target task by leveraging knowledge from related source tasks (Torrey and Shavlik, 2010). This approach has been applied to various domains.

1. Transfer learning has been applied to many real-world applications. For instance, Zhao et al. (2013) presented a novel framework on active transfer learning for cross-system recommendations, Ma et al. (2015) applied transfer learning to atmospheric dust aerosol particle classification for enhancing global climate models, and Zhu et al. (2025) introduced a novel transfer learning framework for time series forecasting with Concept Echo State Network. Other empirical studies on transfer learning include medical diagnosis (Hajiramezanali et al., 2018), natural language processing (Pan and Yang, 2009; Devlin et al., 2018), and price prediction (Nguyen and Yoon, 2019; Xiao et al., 2017). For additional applications, refer to Weiss et al. (2016) and Zhu et al. (2025).

2. Theoretical studies on transfer learning have made progress in establishing consistency of estimators for high-dimensional models. For high-dimensional linear models, Tripuraneni et al. (2020) proposed an algorithm assuming all studies share a common low-dimensional representation, Bastani (2021) developed a two-step joint estimator for one source study, and Li et al. (2022) further investigated scenarios with multiple

auxiliary studies and established the minimax optimal rate. Other theoretical contributions include transfer learning for generalized linear models (Li et al., 2024; Tian and Feng, 2023), Gaussian graphical models (Li et al., 2023b), and representation transfer learning in semi-parametric regression (He et al., 2024). Nonetheless, most existing methods focus on i.i.d. data, whereas methods for time series are relatively limited.

3. Several recent studies investigated transfer learning for time series models. In particular, Lin et al. (2025) and Ma and Safikhani (2025) studied parameter estimation for vector autoregressive (VAR) models under the transfer learning framework. In addition, Duan et al. (2024) introduced "target-PCA", a transfer learning estimator that leverages auxiliary panel data to consistently and efficiently estimate latent factor models in large panels.

To summarize, research on transfer learning for high-dimensional time dependent linear models remains limited, despite its practical importance and theoretical complexity. Therefore, developing robust theoretical frameworks for transfer learning in high-dimensional time series regression is both essential and timely. Such developments would enable more effective utilization of multiple related datasets, even in the presence of complex temporal dependence, thereby achieving more accurate estimation and inference.

This work aims to study transfer learning for high-dimensional linear time series regression models. First, we develop a two-step transfer learning estimator for temporally dependent data by employing the functional dependence framework for the errors and covariates. Our theoretical analysis establishes that, when the target and auxiliary tasks are sufficiently similar and the number of auxiliary samples is large enough, the transfer learn-

ing estimator achieves a faster convergence rate than the single-task learning estimator. Second, we propose a new bias-corrected method, and prove the asymptotic normality of the debiased estimator for each individual coefficient. Furthermore, to mitigate the risk of negative transfer, i.e., the harm caused by transferring the sources that are far away from the target, we develop a transferable source detection algorithm, and demonstrate that the auxiliary samples that are sufficiently close to the target data can be correctly selected through this method.

The remainder of this paper is organized as follows. In Section 2, a debiasing algorithm is proposed for constructing statistical inference for the target parameters. Section 3 presents the theoretical analysis for the corresponding algorithm in the time series setting, including the convergence rate of the transfer learning estimator and the asymptotic normality of the debiased estimator. A feasible test statistic for hypothesis testing on the target parameter is developed by incorporating a banded estimator for the error autocovariance matrix. Asymptotic size and power of the proposed test statistics are demonstrated. In Section 4, we develop a transferable source detection algorithm to mitigate the risk of negative transfer and establish the consistency of the proposed algorithm. The simulations are presented in Section 5, and an analysis of a macroeconomic time series dataset is studied in Section 6. Finally, Section 7 concludes the paper. All technical proofs are provided in the Supplementary Material.

We finish this section with some notations. Denote by $\mathbf{I}_p$ the $p \times p$ identity matrix and $\boldsymbol{e}_j$ the j-th column of $\mathbf{I}_p$. Define $[M] = \{1, 2, \ldots, M\}$, $\{M\} = \{0, 1, \ldots, M\}$ and let $\mathbb{I}\{\cdot\}$ denote the indicator function. For a general positive semi-definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, let $\lambda_{\max}(\mathbf{\Sigma})$ and $\lambda_{\min}(\mathbf{\Sigma})$ denote the largest and smallest eigenvalues of $\mathbf{\Sigma}$, respectively.

We use the notation $a_n = O(b_n)$ and $a_n \lesssim b_n$ to indicate that for sufficiently large n, we have $|a_n/b_n| \leq c$, where c is some constant. Similarly, $a_n \asymp b_n$ means that as $n \to \infty$, $|a_n/b_n| \to c$ for some constant c. For a vector $\boldsymbol{v} = (v_1, \ldots, v_n)^\top$, its ℓ_1, ℓ_2, and ℓ_∞ norms are defined as $|\boldsymbol{v}|_1 = \sum_{i=1}^n |v_i|$, $|\boldsymbol{v}|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$, and $|\boldsymbol{v}|_\infty = \max_{1\leq i\leq n} |v_i|$, respectively. For an $n \times p$ matrix $A = (a_{ij})_{i\leq n, j\leq p}$, its spectral norm and Frobenius norm are defined as $|A|_2 = \max_{|\boldsymbol{v}|_2=1} |A\boldsymbol{v}|_2$ and $|A|_F = (\sum_{i=1}^n \sum_{j=1}^p a_{ij}^2)^{1/2}$, respectively. The element-wise maximum norm of A is denoted by $|A|_{\max} = \max_{i\leq n, j\leq p} |a_{ij}|$. For a random variable Z, we define $\|Z\|_q := \{\mathrm{E}(|Z|^q)\}^{1/q}$, where $q \geq 1$ is a constant. The sub-Gaussian norm of a random variable $u \in \mathbb{R}$ is defined as $\|u\|_{\psi_2} := \sup_{l\geq 1} l^{-1/2} (\mathrm{E}|u|^l)^{1/l}$, and the sub-Gaussian norm of a random vector $\boldsymbol{u} \in \mathbb{R}^p$ is defined as $\|\boldsymbol{u}\|_{\psi_2} := \sup_{|\boldsymbol{v}|_2=1,\, \boldsymbol{v}\in\mathbb{R}^p} \|\boldsymbol{v}^\top \boldsymbol{u}\|_{\psi_2}$. Throughout the paper, we use $c, c_0, c_1, C_1, \ldots$ to represent generic constants, which may vary across different statements.

2. Model and Method

2.1 High-dimensional linear model with auxiliary samples

In this paper, we study the problem of transfer learning in high-dimensional linear models where both predictors and errors exhibit serial dependence, and data are drawn from a target sample and multiple auxiliary samples. The target sample can be represented by the following p-dimensional model:

$$y_i^{(0)} = (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta}^* + \epsilon_i^{(0)}, \quad i = 1, \ldots, n_0, \tag{2.1}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top$ denotes the unknown regression parameters, and $\{\epsilon_i^{(0)}, \boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ represent time-dependent sequences with $\mathrm{E}(\epsilon_i^{(0)}) = 0$ and $\mathrm{E}(\boldsymbol{x}_i^{(0)}) = \boldsymbol{0}$. Here, p is the

number of variables, and n_0 is the number of observations in the target sample. In high-dimensional settings, the number of variables p can be much larger than the number of observations n_0. It is commonly assumed that $\boldsymbol{\beta}^*$ is sparse, meaning that the number of non-zero elements in $\boldsymbol{\beta}^*$, denoted by s_0, is significantly smaller than p. In the context of transfer learning, we also observe additional samples from M auxiliary studies. Specifically, $((\boldsymbol{x}_i^{(k)})^\top, y_i^{(k)})$ are derived from the auxiliary model given by

$$y_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w}^{(k)} + \epsilon_i^{(k)}, \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, M,$$

where $\boldsymbol{w}^{(k)} \in \mathbb{R}^p$ represents the regression coefficients of the k-th auxiliary study, and $\epsilon_i^{(k)}$ denotes the random error term. We assume that $\mathrm{E}(\epsilon_i^{(k)}) = 0$ and $\mathrm{E}(\boldsymbol{x}_i^{(k)}) = \mathbf{0}$. The regression coefficients $\boldsymbol{w}^{(k)}$ are typically unknown and may differ from the target coefficients $\boldsymbol{\beta}^*$. We define $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta}^* - \boldsymbol{w}^{(k)}$ as the similarity contrast vector between the k-th auxiliary study and the target study. For convenience, we set $\boldsymbol{\delta}^{(0)} = \mathbf{0}$. A smaller norm of $\boldsymbol{\delta}^{(k)}$ indicates a higher degree of similarity between $\boldsymbol{w}^{(k)}$ and $\boldsymbol{\beta}^*$. We define a set of auxiliary samples with sufficiently sparse contrast as follows:

$$\mathcal{A}_h = \{1 \le k \le M : |\boldsymbol{\delta}^{(k)}|_1 \le h\}.$$

The set $\mathcal{A}_h$ consists of the indices of the auxiliary samples whose ℓ_1-sparsity of the contrast vector does not exceed h. The subsequent analysis shows that, when h is sufficiently small, the information from $\mathcal{A}_h$ can improve the accuracy of both estimation and inference for $\boldsymbol{\beta}^*$. The number of auxiliary studies M is treated as a fixed integer in the following analysis.

This paper aims to perform statistical inference for the target parameter $\boldsymbol{\beta}^*$ based on the target data $\{(\boldsymbol{x}_i^{(0)})^\top, y_i^{(0)}\}_{i=1}^{n_0}$, in conjunction with auxiliary data $\{(\boldsymbol{x}_i^{(k)})^\top, y_i^{(k)}\}_{i=1}^{n_k}$, $k \in \mathcal{A}_h$, under a time series framework.

2.2 Dependence assumptions on $\boldsymbol{\epsilon}$ and X

To describe the dependence structure of the time series data, we adopt the functional dependence concept of Wu (2005); Wu and Wu (2016) and Zhang and Wu (2017). Assume that $\{\epsilon_i\}$ is stationary and satisfies

$$\epsilon_i = g(\ldots, \xi_{i-1}, \xi_i) = g(\mathcal{F}_i), \tag{2.2}$$

where $\{\xi_i\}$ are i.i.d. random variables, $\mathcal{F}_i = (\ldots, \xi_{i-1}, \xi_i)$ is a σ-field, and $g(\cdot)$ is a measurable function in $\mathbb{R}$ which makes that ϵ_i is well defined. The representation in (2.2) has a clear physical interpretation: the sequence $\{\xi_i\}$ serves as the input, and $\{\epsilon_i\}$ as the output of the system. This framework encompasses a wide class of stationary processes, including linear processes, their nonlinear transformations, and Volterra processes that involve interactions between the innovations.

Following Wu (2005) and Zhang and Wu (2017), we assume that $\|\epsilon_i\|_q < \infty$ for $q \geq 1$, and define the functional dependence measure $\delta_{i,q} = \|g(\mathcal{F}_i) - g(\mathcal{F}_i')\|_q$, where $\mathcal{F}_i' = (\ldots, \xi_{-1}, \xi_0', \xi_1, \ldots, \xi_i)$ is a coupled version of $\mathcal{F}_i$ with ξ_0' being an i.i.d. copy of ξ_0. Note that $\delta_{i,q}$ quantifies the dependence of ϵ_i on ξ_0. We impose the assumption of short-range dependence, such that

$$\Delta_{m,q} := \sum_{i=m}^{\infty} \delta_{i,q} < \infty. \tag{2.3}$$

For a fixed m, $\Delta_{m,q}$ represents the cumulative impact of ξ_0 on $\{\epsilon_i\}_{i \geq m}$, condition (2.3) ensures that the cumulative effect is finite. In order to account for dependence, for the process $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=-\infty}^{\infty}$, we define the dependence adjusted norm as:

$$\|\boldsymbol{\epsilon} \cdot\|_{q,\alpha} = \sup_{m \geq 0} (m+1)^{\alpha} \Delta_{m,q} = \sup_{m \geq 0} (m+1)^{\alpha} \sum_{i=m}^{\infty} \delta_{i,q}, \quad \alpha > 0.$$

Elementary calculations show that, if $\{\epsilon_i\}$ are i.i.d. random variables, then $\|\epsilon_0\|_q \le \|\boldsymbol{\epsilon}_\cdot\|_{q,\alpha} \le 2\|\epsilon_0\|_q$, suggesting that the dependence-adjusted norm is equivalent to the classical L_q norm in the independent case.

Similarly, we assume that the covariate process $\{\boldsymbol{x}_i\}$ is a high-dimensional stationary process of the form

$$\boldsymbol{x}_i = h(\mathcal{S}_i), \qquad \mathcal{S}_i = (\ldots, \boldsymbol{\eta}_{i-1}, \boldsymbol{\eta}_i), \tag{2.4}$$

where $\{\boldsymbol{\eta}_i\}$ are i.i.d. random vectors, and $h(\cdot) = (h_1(\cdot), \ldots, h_p(\cdot))^\top$ is a measurable function in $\mathbb{R}^p$. Similar to defining $\delta_{i,q}$, by assuming that $\boldsymbol{x}_i$ satisfies $\sup_{|\boldsymbol{v}|_2=1} \|\boldsymbol{v}^\top \boldsymbol{x}_i\|_\iota < \infty$, $\iota > 2$, we define the functional dependence measure

$$\phi_{i,\iota} = \sup_{|\boldsymbol{v}|_2=1} \|\boldsymbol{v}^\top h(\mathcal{S}_i) - \boldsymbol{v}^\top h(\mathcal{S}_i')\|_\iota,$$

where $\mathcal{S}_i'$ is a coupled version of $\mathcal{S}_i$ with $\boldsymbol{\eta}_0$ replaced by an i.i.d. copy $\boldsymbol{\eta}_0'$. Analogous to $\Delta_{m,q}$, we can define and assume

$$\Phi_{m,\iota} = \sum_{i=m}^{\infty} \phi_{i,\iota} < \infty, \qquad \|\boldsymbol{x} \cdot\|_{\iota,\alpha_x} = \sup_{m \ge 0} (m+1)^{\alpha_x} \Phi_{m,\iota}, \quad \text{for some} \quad \alpha_x > 0.$$

Next, we provide an example of high-dimensional time series to illustrate the cumulative functional dependence measure $\Phi_{m,\iota}$.

Example 1. *Suppose that* $\{\boldsymbol{\eta}_i\}$ *are* i.i.d. *sub-Gaussian random vectors with* $\|\boldsymbol{\eta}_i\|_{\psi_2} \le c_1$. *Let* A_i *be* $p \times p$ *coefficient matrices with real entries such that* $\sum_{i=0}^{\infty} |A_i|_F < \infty$. *Then by Kolmogorov's three-series theorem (Kolmogoroff, 1928), the linear process*

$$\boldsymbol{x}_t = \sum_{i=0}^{\infty} A_i \boldsymbol{\eta}_{t-i},$$

exists, and it is of the form (2.4) *with a linear functional* h. *As proved in Lemma 1 of the*

Supplementary Material,

$$\phi_{t,\iota} = \sup_{|\boldsymbol{v}|_2=1} \|\boldsymbol{v}^\top A_t(\boldsymbol{\eta}_0 - \boldsymbol{\eta}_0')\|_\iota \le c_0|A_t|_2, \qquad \Phi_{m,\iota} \le c_0 \sum_{i=m}^{\infty} |A_i|_2 < \infty,$$

where c_0 is a constant depending on ι and c_1.

The functional dependence measure provides a convenient framework that greatly simplifies the derivation of tail probability bounds under temporal dependence, which is particularly beneficial for our theoretical analysis. For a more detailed discussion of this dependence structure, see Wu and Wu (2016).

2.3 Statistical inference via debiased estimators

Let $X^{(k)} \in \mathbb{R}^{n_k \times p}$ and $\boldsymbol{y}^{(k)} \in \mathbb{R}^{n_k}$ be the design matrix and response vector of the k-th dataset, with $k = 0$ corresponding to the target dataset. We first estimate $\boldsymbol{\beta}^*$ by adopting the transfer learning method proposed by Li et al. (2022) under the assumption that $\mathcal{A}_h$ is known. The main strategy consists of two steps, corresponding to Steps 1 and 2 of Algorithm 1. The first step transfers information from all auxiliary datasets by pooling the data to obtain a rough estimator $\widehat{\boldsymbol{w}}^{\mathcal{A}_h}$. However, the probabilistic limit of $\widehat{\boldsymbol{w}}^{\mathcal{A}_h}$, denoted by $\boldsymbol{w}^{\mathcal{A}_h}$, differs from $\boldsymbol{\beta}^*$ as $\boldsymbol{w}^{(k)} \neq \boldsymbol{\beta}^*$ in general. The vector $\boldsymbol{w}^{\mathcal{A}_h}$ can be defined through the following moment condition:

$$\mathrm{E}\Big[\sum_{k\in\mathcal{A}_h} (X^{(k)})^\top (\boldsymbol{y}^{(k)} - X^{(k)}\boldsymbol{w}^{\mathcal{A}_h})\Big] = \mathbf{0}.$$

Denoting $\mathrm{E}[\boldsymbol{x}_i^{(k)}(\boldsymbol{x}_i^{(k)})^\top] = \boldsymbol{\Sigma}^{(k)}$ and $\alpha_k = n_k/n_{\mathcal{A}_h}$, $\boldsymbol{w}^{\mathcal{A}_h}$ has the following explicit form:

$$\boldsymbol{w}^{\mathcal{A}_h} = \boldsymbol{\beta}^* - \boldsymbol{\delta}^{\mathcal{A}_h},$$

where $\boldsymbol{\delta}^{\mathcal{A}_h} = \big(\sum_{k\in\mathcal{A}_h} \alpha_k \boldsymbol{\Sigma}^{(k)}\big)^{-1} \sum_{k\in\mathcal{A}_h} \alpha_k \boldsymbol{\Sigma}^{(k)} \boldsymbol{\delta}^{(k)}$ denotes the bias relative to $\boldsymbol{\beta}^*$. Then, the second step utilizes the target samples to correct the bias $\boldsymbol{\delta}^{\mathcal{A}_h}$.

Algorithm 1: Construction of Debiased Estimator

Input : Target data $(X^{(0)}, \boldsymbol{y}^{(0)})$, auxiliary datasets $\{(X^{(k)}, \boldsymbol{y}^{(k)})\}_{k\in\mathcal{A}_h}$, tuning parameters λ_w and λ_δ which will be specified in Theorem 1, constants $c_\Theta > 0$, $c_\gamma > 0$ and $1/6 < r < 1/4$

Output: Debiased estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{db}}$

Step 1: Compute

$$\widehat{\boldsymbol{w}}^{\mathcal{A}_h} = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min}\Big\{\frac{1}{2n_{\mathcal{A}_h}}\sum_{k\in\mathcal{A}_h}|\boldsymbol{y}^{(k)} - X^{(k)}\boldsymbol{w}|_2^2 + \lambda_w|\boldsymbol{w}|_1\Big\},$$

where $n_{\mathcal{A}_h} = \sum_{k\in\mathcal{A}_h} n_k$.

Step 2: Let the Trans-Lasso estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{w}}^{\mathcal{A}_h} + \widehat{\boldsymbol{\delta}}^{\mathcal{A}_h}$, where

$$\widehat{\boldsymbol{\delta}}^{\mathcal{A}_h} = \underset{\boldsymbol{\delta}\in\mathbb{R}^p}{\arg\min}\Big\{\frac{1}{2n_0}|\boldsymbol{y}^{(0)} - X^{(0)}(\widehat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta})|_2^2 + \lambda_\delta|\boldsymbol{\delta}|_1\Big\}.$$

Step 3: Construct the debiased estimator

$$\widehat{\beta}_j^{\mathrm{db}} = \widehat{\beta}_j + \widehat{\boldsymbol{\Theta}}_j^\top (X^{(0)})^\top(\boldsymbol{y}^{(0)} - X^{(0)}\widehat{\boldsymbol{\beta}})/n_0, \quad j = 1, \ldots, p, \tag{2.5}$$

where $\widehat{\boldsymbol{\Theta}}_j$ is obtained by solving the following constrained optimization problem

$$\widehat{\boldsymbol{\Theta}}_j \in \underset{\boldsymbol{\gamma}\in\mathbb{R}^p}{\arg\min}\ \boldsymbol{\gamma}^\top\widehat{\boldsymbol{\Sigma}}^{(0)}\boldsymbol{\gamma}$$
$$\text{subject to} \quad \begin{cases} |\widehat{\boldsymbol{\Sigma}}^{(0)}\boldsymbol{\gamma} - \boldsymbol{e}_j|_\infty \le c_\Theta\lambda_\delta, \\ \max_{1\le i\le n_0}|(\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\gamma}| \le c_\gamma n_0^r. \end{cases} \tag{2.6}$$

We next construct hypothesis tests for each component of $\boldsymbol{\beta}^*$. Under the general debiasing framework (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014), the debiased estimator is constructed according to (2.5) in Algorithm 1.

Define $\boldsymbol{\epsilon}^{(0)} = (\epsilon_1^{(0)}, \ldots, \epsilon_{n_0}^{(0)})$. Through simple algebraic manipulation, we obtain

$$\sqrt{n_0}(\widehat{\beta}_j^{\mathrm{db}} - \beta_j^*) = \widehat{\boldsymbol{\Theta}}_j^\top (X^{(0)})^\top \boldsymbol{\epsilon}^{(0)}/\sqrt{n_0} + \Lambda_j, \tag{2.7}$$

where

$$\Lambda_j = -\sqrt{n_0}(\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}}^{(0)} - \boldsymbol{e}_j^\top)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \tag{2.8}$$

is the remainder term. For the debiasing to be effective, it is crucial to choose an appropriate $\widehat{\boldsymbol{\Theta}}_j$ such that Λ_j is sufficiently small. We therefore propose a new estimator as in (2.6) specifically tailored to temporally dependent high-dimensional data.

The two constraints are linear in (2.6), leading to convex optimization and computationally efficiency. The first constraint ensures that $\widehat{\boldsymbol{\Theta}}_j$ approximates the j-th row of ${\boldsymbol{\Sigma}^{(0)}}^{-1}$, thereby controlling the magnitude of the remainder term Λ_j in Equation (2.8). The second constraint regulates the magnitude of $|(\boldsymbol{x}_i^{(0)})^\top \widehat{\boldsymbol{\Theta}}_j|$, which is critical for satisfying the Lindeberg condition. The objective function minimizes $\boldsymbol{\gamma}^\top \widehat{\boldsymbol{\Sigma}}^{(0)} \boldsymbol{\gamma}$ to control the magnitude of $|X^{(0)}\widehat{\boldsymbol{\Theta}}_j|_2^2/n_0$, which is also instrumental in satisfying the Lindeberg condition. The optimization problem in (2.6) is feasible, and Lemma 9 of the Supplementary Material establishes that $\boldsymbol{\gamma} = {\boldsymbol{\Sigma}^{(0)}}^{-1}_{\cdot,j}$ satisfies the two constraints in (2.6) with probability tending to one, provided that the tuning parameters c_Θ and c_γ are appropriately chosen.

Our estimator $\widehat{\boldsymbol{\Theta}}_j$ in (2.6) differs from that in Javanmard and Montanari (2014) primarily in the constraints imposed in the optimization problem. For the first constraint, the order of $c_\Theta \lambda_\delta$ is no smaller than $\sqrt{\log p/n_0}$ used in Javanmard and Montanari (2014), which reflects the slower tail decay of $|\widehat{\boldsymbol{\Sigma}}^{(0)}\boldsymbol{\gamma} - \boldsymbol{e}_j|_\infty$ under temporal dependence. For the second constraint, the parameter r is chosen to be strictly smaller than that in Javanmard and Montanari (2014), where $r \in (1/4, 1/2)$ is adopted. In contrast, we require $1/6 < r < 1/4$

to properly account for dependence. This more restrictive range is essential for establishing the asymptotic normality of the debiased estimator; see Section S4 of the Supplementary Material for further details. Zhu and Bradic (2018) and Li et al. (2024) also compute the correction vector $\widehat{\boldsymbol{\Theta}}_j$ using constraints akin to those in (2.6), but their approaches are based on an ℓ_1-minimization objective and likewise rely on the assumption of independent error terms. Moreover, the proposed debiasing framework is flexible and can also be applied to single-task high-dimensional linear time series models, where $\widehat{\boldsymbol{\beta}}$ may be replaced by the Lasso estimator for the single task. In comparison, the method developed by Yuan and Guo (2022) for time series settings is limited to scenarios where the design matrix $X^{(0)}$ is fixed.

3. Theoretical results

In this section, we present the theoretical guarantees for the proposed inference procedure. Section 3.1 establishes the convergence rate of the transfer learning estimator $\widehat{\boldsymbol{\beta}}$ under temporally dependent data, along with a detailed analysis of the asymptotic normality of the debiased estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{db}}$. However, the asymptotic distribution of $\widehat{\beta}_k^{\mathrm{db}}$ in Theorem 2 depends on the unknown autocovariance matrix of the error terms, $\boldsymbol{\Sigma}_{n_0}^{\epsilon}$. To address this challenge, we develop a banded estimator for $\boldsymbol{\Sigma}_{n_0}^{\epsilon}$ in Section 3.2, which facilitates valid statistical inference for transfer learning under temporal dependence.

3.1 Theoretical properties of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{db}}$

Formally, we consider the parameter space

$$\boldsymbol{\Theta}(s_0, h) = \{B = (\boldsymbol{\beta}^*, \boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(M)}) : |\boldsymbol{\beta}^*|_0 \leq s_0,\ \max_{k \in \mathcal{A}_h} |\boldsymbol{\delta}^{(k)}|_1 \leq h,\ \max_{k \in \mathcal{A}_h} |\boldsymbol{\delta}^{(k)}|_2^2 < c\},$$

where c denotes a positive constant. Define

$$C_{\boldsymbol{\Sigma}} = 1 + \max_{j\le p} \max_{k\in\mathcal{A}_h} \Big| \Big(\sum_{k\in\mathcal{A}_h} \alpha_k \boldsymbol{\Sigma}^{(k)} \Big)^{-1} \big(\boldsymbol{\Sigma}^{(k)} - \boldsymbol{\Sigma}^{(0)} \big) \boldsymbol{e}_j \Big|_1,$$

which measures the discrepancy between $\boldsymbol{\Sigma}^{(k)}$ and $\boldsymbol{\Sigma}^{(0)}$ for $k \in \mathcal{A}_h$. As discussed in Li et al. (2022), $C_{\boldsymbol{\Sigma}}$ is a constant if $\max_{1\le j\le p} |\boldsymbol{e}_j^\top (\boldsymbol{\Sigma}^{(k)} - \boldsymbol{\Sigma}^{(0)})|_0 \le C < \infty$, for all $k \in \mathcal{A}_h$. Examples include block diagonal $\boldsymbol{\Sigma}^{(k)}$ with constant block sizes or banded $\boldsymbol{\Sigma}^{(k)}$ with constant bandwidths for $k \in \{0\} \cup \mathcal{A}_h$. To establish our theoretical results, we introduce the following conditions.

Condition 1. *For all $k \in \{M\}$, there exist constants c_1 and c_2, such that $0 < c_1 < \lambda_{\min}(\boldsymbol{\Sigma}^{(k)}) \le \lambda_{\max}(\boldsymbol{\Sigma}^{(k)}) < c_2 < \infty$.*

Condition 2. *For each $k \in \{M\}$, the error sequence $\{\epsilon_i^{(k)}\}_{i=1}^{n_k}$ is independent of the covariate sequence $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k}$. Moreover, the sample pairs $(X^{(k)}, \boldsymbol{y}^{(k)})$, for all $k \in \{M\}$, are mutually independent.*

Condition 1 assumes that each $\boldsymbol{\Sigma}^{(k)}$ is positive definite, with eigenvalues uniformly bounded away from zero and infinity across all studies. Similar assumptions have been adopted in Li et al. (2022), Li et al. (2024), and Fan et al. (2011). Condition 2 states that the error term is independent of the covariate $\boldsymbol{x}_i^{(k)}$, which is a mild assumption in high-dimensional regression settings.

It is worth noting that although we impose certain conditions on auxiliary datasets not included in $\mathcal{A}_h$, these datasets are not utilized in estimating $\boldsymbol{\beta}^*$. The reason for imposing such conditions is that the Transferable Source Detection procedure in Section 4 involves all auxiliary datasets. To avoid redundant statements of assumptions, we incorporate all related conditions uniformly in Conditions 1 and 2.

In the following theorem, we establish the convergence rate of the Trans-Lasso estimator $\widehat{\boldsymbol{\beta}}$ under temporal dependence.

Theorem 1. *Assume that Conditions 1 and 2 hold, and that* $n_0 \ll n_{\mathcal{A}_h}$ *with* $C_{\boldsymbol{\Sigma}} h \leq c_1$. *Then the following asymptotic properties hold.*

(i) Assume that $\sup_{k\in\{M\}} \|\boldsymbol{x}^{(k)} \cdot\|_{\iota,\alpha_X} = N_x < \infty$ *and* $\sup_{k\in\{M\}} \|\boldsymbol{\epsilon}^{(k)} \cdot\|_{q,\alpha_e} = N_e < \infty$, *where* $q > 2$, $\iota > 4$ *and* $\alpha_X, \alpha_e > 0$. *Let* $\chi = 1$ *if* $\alpha_X > 1/2 - 2/\iota$ *and* $\chi = \iota/4 - \alpha_X \iota/2$ *if* $\alpha_X < 1/2 - 2/\iota$. *Further suppose that* $\tau = q\iota/(q+\iota) > 2$ *and let* $\alpha = \min(\alpha_X, \alpha_e)$. *Define* $\pi = 1$ *if* $\alpha > 1/2 - 1/\tau$ *and* $\pi = \tau/2 - \alpha\tau$ *if* $\alpha < 1/2 - 1/\tau$. *Let*

$$\lambda_{xx}^{\mathcal{A}} \asymp N_X^2 \max\{n_{\mathcal{A}_h}^{2\chi/\iota - 1} (p \log p)^{2/\iota}, (\log p / n_{\mathcal{A}_h})^{1/2}\},$$

$$\lambda_{x\epsilon}^{\mathcal{A}} \asymp N_X N_e \max\{n_{\mathcal{A}_h}^{\pi/\tau - 1} (p \log p)^{1/\tau}, (\log p / n_{\mathcal{A}_h})^{1/2}\},$$

$\lambda_w = \max\{\lambda_{xx}^{\mathcal{A}}, \lambda_{x\epsilon}^{\mathcal{A}}\}$, *and* λ_δ *is defined in the same way as* λ_w *but with* $n_{\mathcal{A}_h}$ *replaced by* n_0. *Let* $\lambda_X^0 \asymp N_X^2 \max\{n_0^{-(1-2\chi/\iota)} (p \log p)^{4/\iota}, \sqrt{\log p / n_0}\}$ *and further assume that* $s_0 \lambda_X^0 \leq c$ *for some sufficiently small constant* $c > 0$. *Then, with probability at least* $1 - c_1 (\log p)^{-1}$,

$$\frac{1}{n_0} |X^{(0)} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|_2^2 \lesssim s_0 \lambda_w^2 + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2, \tag{3.9}$$

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \lesssim s_0 \lambda_w + \frac{(C_{\boldsymbol{\Sigma}} h)^2}{\lambda_\delta} + C_{\boldsymbol{\Sigma}} h. \tag{3.10}$$

(ii) Assume that for some $\nu, \varrho \geq 0$,

$$K_\nu := \sup_{k\in\{M\}} \sup_{q\geq 2} q^{-\nu} \Delta_{0,q}^{(k)} < \infty, \qquad L_\varrho := \sup_{k\in\{M\}} \sup_{q\geq 2} q^{-\varrho} \Phi_{0,q}^{(k)} < \infty.$$

Define

$$\lambda_{xx}^{\mathcal{A}} \asymp L_\varrho^2 (\log p)^{(1+4\varrho)/2} / \sqrt{n_{\mathcal{A}_h}}, \quad \lambda_{x\epsilon}^{\mathcal{A}} \asymp K_\nu L_\varrho (\log p)^{(1+2\nu+2\varrho)/2} / \sqrt{n_{\mathcal{A}_h}},$$

$\lambda_w = \max\{\lambda_{xx}^{\mathcal{A}}, \lambda_{x\epsilon}^{\mathcal{A}}\}$, and λ_δ is the same as λ_w but with $n_{\mathcal{A}_h}$ replaced by n_0. Let $\lambda_X^0 \asymp L_\varrho^2 (\log p)^{(1+4\varrho)/2}/\sqrt{n_0}$ and further assume that $s_0 \lambda_X^0 \le c$ for some sufficiently small constant $c > 0$. Then, (3.9) and (3.10) hold with probability at least $1 - c_1 p^{-2}$.

Theorem 1 establishes the convergence rate of $\widehat{\boldsymbol{\beta}}$ under mild regularity conditions. Compared with the results of Theorem 4 in Li et al. (2022), which also adopts a transfer learning approach but assumes that the errors are i.i.d. sub-Gaussian variables, their estimation error bound satisfies

$$\frac{1}{n_0}|X^{(0)}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|_2^2 \lesssim s_0 \frac{\log p}{n_0 + n_{\mathcal{A}_h}} + \min\Big\{ s_0 \frac{\log p}{n_0},\, C_{\boldsymbol{\Sigma}} h \sqrt{\frac{\log p}{n_0}},\, (C_{\boldsymbol{\Sigma}} h)^2 \Big\}. \tag{3.11}$$

Due to the temporal dependence in the data, the magnitudes of λ_w and λ_δ in Theorem 1 are larger than $(\log p/(n_0 + n_{\mathcal{A}_h}))^{1/2}$ and $(\log p/n_0)^{1/2}$, respectively. As a result, the convergence rate established in (3.9) is generally slower than the bound given in (3.11). Nevertheless, our theoretical results are more broadly applicable, as they accommodate not only stationary error processes but also a wider class of covariate sequences $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k}$. In particular, Theorem 1 provides the theoretical guarantee for transfer learning in high-dimensional linear models with temporally dependent data. This generalization is nontrivial due to the combined challenges of high dimensionality and temporal dependence. We also remark that the ℓ_1-error bound in Theorem 1 plays a crucial role in enabling valid statistical inference for the target parameter $\boldsymbol{\beta}^*$, which will be further developed in Theorem 2. Moreover, under the setting of stationary Gaussian processes, we establish a sharper bound for $\widehat{\boldsymbol{\beta}}$.

Proposition 1. *Suppose that $\{\epsilon_i^{(k)}\}$ and $\{\boldsymbol{x}_i^{(k)}\}$ are Gaussian stationary processes with the form that*

$$\boldsymbol{x}_t^{(k)} = \sum_{i=0}^{\infty} A_i^{(k)} \boldsymbol{\eta}_{t-i}^{(k)}, \quad \boldsymbol{\eta}_i^{(k)} \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_k), \quad \sum_{i=0}^{\infty} |A_i^{(k)}|_F \le c_1 < \infty, \quad k \in \{M\},$$

$$\epsilon_t^{(k)} = \sum_{i=0}^{\infty} a_i^{(k)} \xi_{t-i}^{(k)}, \quad \xi_i^{(k)} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_k), \quad \sum_{i=0}^{\infty} |a_i^{(k)}| \leq c_2 < \infty, \quad k \in \{M\}.$$

Let $\lambda_w = c\sqrt{\log p / n_{\mathcal{A}_h}}$ *and* $\lambda_\delta = c\sqrt{\log p / n_0}$ *for some sufficiently large constant* c*. Further assume that* $C_{\boldsymbol{\Sigma}} h \lesssim (s_0\sqrt{\log p/n_0}) \wedge 1$ *and* $s_0 \log p / n_{\mathcal{A}_h} + C_{\boldsymbol{\Sigma}} h \sqrt{\log p / n_0} = o(1)$*. Then, under Conditions 1 and 2, it holds that*

$$\frac{1}{n_0}|X^{(0)}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|_2^2 \vee |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2^2 \lesssim s_0 \frac{\log p}{n_{\mathcal{A}_h}} + C_{\boldsymbol{\Sigma}} h \sqrt{\frac{\log p}{n_0}} \wedge (C_{\boldsymbol{\Sigma}} h)^2,$$

with probability at least $1 - c_1 p^{-2}$*.*

First, we compare our results with those of Basu and Michailidis (2015), who studied the convergence rates of the single-task linear model under Gaussian stationary processes. According to Proposition 3.3 of Basu and Michailidis (2015), under the conditions of Proposition 1, the performance of the single-task Lasso estimator $\widehat{\boldsymbol{\beta}}^l$ satisfies

$$\frac{1}{n_0}|X^{(0)}(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)|_2^2 \vee |\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_2^2 \lesssim s_0 \log p / n_0.$$

When $C_{\boldsymbol{\Sigma}} h \ll s_0\sqrt{\log p/n_0}$ and $n_0 \ll n_{\mathcal{A}_h}$, the bound in Proposition 1 is sharper than that of the single-task Lasso estimator in Basu and Michailidis (2015). In other words, when the source and target studies are sufficiently similar and the total auxiliary sample size is substantially larger than that of the target sample, incorporating auxiliary information can effectively improve the performance of the target model. Moreover, under the assumption $C_{\boldsymbol{\Sigma}} < \infty$, the rate in Proposition 1 attains the minimax optimal convergence rate over the parameter space $\boldsymbol{\Theta}(s_0, h)$, in view of Theorem 2 in Li et al. (2022) and Theorem 2 in Tian and Feng (2023).

Remark 1. *The condition* $C_{\boldsymbol{\Sigma}} h \leq c_1$ *required in Theorem 1 can be relaxed to* $h \leq c_1$ *by adopting the estimation approach in Li et al. (2024). However, this relaxation may, in*

certain cases, lead to slower convergence rates compared to the Trans-Lasso method, as well as increased computational cost. For more details, see Section S7 of the Supplementary Material.

We then establish the asymptotic distribution of the debiased transfer learning estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{db}}$.

Theorem 2. *Assume that the conditions of Theorem 1 hold.*

(i) For case (i) in Theorem 1, further suppose that $\iota > 6$, then the debiased estimator satisfies:

$$\frac{\sqrt{n_0}(\widehat{\beta}_k^{\mathrm{db}} - \beta_k^*)}{\sigma_k} = V_k + \frac{\Lambda_k}{\sigma_k},$$

where

$$V_k = \frac{1}{\sqrt{n_0}\sigma_k} \sum_{i=1}^{n_0} \widehat{\boldsymbol{\Theta}}_k^\top \boldsymbol{x}_i^{(0)} \epsilon_i^{(0)} \xrightarrow{D} N(0,1), \tag{3.12}$$

and

$$\mathbb{P}(|\Lambda_k| \gtrsim \sqrt{n_0}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)) \leq C_1 (\log p)^{-1}.$$

Here, $\sigma_k^2 = n_0^{-1}\widehat{\boldsymbol{\Theta}}_k^\top (X^{(0)})^\top \boldsymbol{\Sigma}_{n_0}^{\epsilon} X^{(0)} \widehat{\boldsymbol{\Theta}}_k$, with $\sigma_k^2 \geq c_0 - o_{\mathbb{P}}(1)$, and $\boldsymbol{\Sigma}_{n_0}^{\epsilon} = \mathrm{Cov}(\boldsymbol{\epsilon}^{(0)})$ is the autocovariance matrix of the error sequence.

(ii) For case (ii) in Theorem 1, the same result as in (3.12) *holds, but :*

$$\mathbb{P}(|\Lambda_k| \gtrsim \sqrt{n_0}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)) \leq C_1 p^{-2}.$$

In Theorem 2, σ_k^2 involves the autocovariance matrix $\boldsymbol{\Sigma}_{n_0}^{\epsilon}$ of the errors, which is unknown. This issue is mitigated by the consistent estimator for $\boldsymbol{\Sigma}_{n_0}^{\epsilon}$ that we develop in Section 3.2. We remark that minimizing $\boldsymbol{\gamma}^\top \widehat{\boldsymbol{\Sigma}}^{(0)} \boldsymbol{\gamma}$ in (2.6) also helps control the magnitude of σ_k^2; see Section S4 of the Supplementary Material for more details. Theorem 2

further decomposes the limiting distribution of $\widehat{\beta}_k^{\mathrm{db}}$ into two components: an asymptotically normal term V_k and a bias term Λ_k/σ_k. To establish the asymptotic normality of $n_0^{1/2}(\widehat{\beta}_k^{\mathrm{db}} - \beta_k^*)/\sigma_k$, it is required that $\Lambda_k/\sigma_k = o_{\mathbb{P}}(1)$, which holds under the condition $\sqrt{n_0}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\mathbf{\Sigma}}h + (C_{\mathbf{\Sigma}}h)^2) = o(1)$.

We also note that our debiasing approach differs fundamentally from the existing methods for sparse regression under dependence. Yuan and Guo (2022) employed a nodewise Lasso-based debiasing method under the assumption of a fixed design matrix. Xia et al. (2024) and Adamek et al. (2023) also adopted a nodewise Lasso-based debiasing approach. In contrast, our method is based on a constrained optimization framework and does not require the row-sparsity assumption on the inverse covariance matrix, making it more broadly applicable to time-dependent settings.

3.2 A feasible test statistic

As noted in Theorem 2, the limiting distribution of $\widehat{\beta}_k^{\mathrm{db}}$ depends on the unknown autocovariance matrix of the error terms and therefore cannot be directly used for hypothesis testing on the target regression coefficients. Thus, to conduct statistical inference on β_k^*, we develop a consistent estimator for $\mathbf{\Sigma}_{n_0}^{\epsilon}$.

Assume that $\{\epsilon_i^{(0)}\}_{i=1}^{n_0}$ is a stationary process with zero mean. Then the autocovariance function $\gamma_k^{\epsilon} = \mathrm{Cov}(\epsilon_0^{(0)}, \epsilon_k^{(0)})$ can be estimated by the sample autocovariance:

$$\widehat{\gamma}_k^{\epsilon} = \frac{1}{n_0}\sum_{i=1}^{n_0-|k|} e_i e_{i+|k|}, \quad k = 0, \pm 1, \ldots, \pm(n_0-1), \tag{3.13}$$

where $e_i = y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \widehat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n_0$, denotes the residuals obtained from the Trans-Lasso estimation. For any fixed k, under suitable conditions, we have $\widehat{\gamma}_k^{\epsilon} \xrightarrow{\mathbb{P}} \gamma_k^{\epsilon}$, as established in Lemma 11 of the Supplementary Material. However, entry-wise convergence

does not automatically imply that $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0} = (\widehat{\gamma}^{\epsilon}_{i-j})_{1\le i,j\le n_0}$ serves as a reliable estimator of $\mathbf{\Sigma}^{\epsilon}_{n_0} = (\gamma^{\epsilon}_{i-j})_{1\le i,j\le n_0}$, as noted by Hannan and Deistler (2012). To obtain a consistent estimator of $\mathbf{\Sigma}^{\epsilon}_{n_0}$, we adopt the banding strategy introduced by Bickel and Levina (2008) and develop a banded autocovariance matrix estimator $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l}$, defined as

$$\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l} = (\widehat{\gamma}^{\epsilon}_{i-j}\, \mathbb{I}_{|i-j|\le l})_{1\le i,j\le n_0},$$

where $l \ge 0$ is a banding parameter. The estimator $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l}$ retains the diagonal and the $2l$ main sub-diagonals of $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0}$, and for $l \ge n_0 - 1$, it simplifies to $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l} = \widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0}$. Under appropriate conditions on l, the consistency of $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l}$ is established in Lemma 1 as follows:

Lemma 1. *Assume that the conditions of Theorem 2 hold. If $l \to \infty$ and*

$$l(n_0^{2/q'-1} + s_0\lambda_{\boldsymbol{w}}\lambda_\delta + \lambda_\delta C_{\mathbf{\Sigma}} h + (C_{\mathbf{\Sigma}} h)^2) = o(1),$$

for some $2 < q' \le 4 \wedge q$, then we obtain:

$$|\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l} - \mathbf{\Sigma}^{\epsilon}_{n_0}|_2 = o_{\mathbb{P}}(1).$$

Lemma 1 establishes the consistency of the banded estimator for the error autocovariance matrix. Then, we can define a plug-in estimator for σ_k^2 as follows:

$$\widehat{\sigma}_k^2 = {n_0}^{-1}\widehat{\mathbf{\Theta}}_k^{\top}(X^{(0)})^{\top}\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l}X^{(0)}\widehat{\mathbf{\Theta}}_k.$$

The asymptotic distribution of the debiased Lasso estimator $\widehat{\beta}_k^{\mathrm{db}}$, incorporating the estimator $\widehat{\sigma}_k^2$, is presented below.

Theorem 3. *Under the conditions of Lemma 1, for any $k \in \{1, \ldots, p\}$, we have*

$$|\widehat{\sigma}_k^2 - \sigma_k^2| = o_{\mathbb{P}}(1), \quad \sigma_k^2 \geq c_0 - o_{\mathbb{P}}(1),$$
$$\widehat{\zeta}_k = \frac{\sqrt{n_0}(\widehat{\beta}_k^{\mathrm{db}} - \beta_k^*)}{\widehat{\sigma}_k} = \widetilde{V}_k + \widetilde{\Lambda}_k,$$
$$\widetilde{V}_k = \frac{1}{\sqrt{n_0}\widehat{\sigma}_k} \sum_{i=1}^{n_0} \widehat{\boldsymbol{\Theta}}_k^\top \boldsymbol{x}_i^{(0)} \epsilon_i^{(0)} \xrightarrow{D} N(0,1),$$

and

$$\mathbb{P}\Big(|\widetilde{\Lambda}_k| \gtrsim \sqrt{n_0}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)\frac{1}{\widehat{\sigma}_k}\Big) \leq C_1(\log p)^{-1},$$

or

$$\mathbb{P}\Big(|\widetilde{\Lambda}_k| \gtrsim \sqrt{n_0}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)\frac{1}{\widehat{\sigma}_k}\Big) \leq C_1 p^{-2}.$$

Theorem 3 shows that, under the condition $\sqrt{n_0}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2) = o(1)$, $\widehat{\zeta}_k$ is asymptotically standard normal, which enables us to construct valid inference procedures for the target parameter. For any given $k \in \{1, \ldots, p\}$, the $100(1-a)\%$ confidence interval for β_k^* is given by

$$\Big[\widehat{\beta}_k^{\mathrm{db}} - \frac{\widehat{\sigma}_k}{\sqrt{n_0}}\Phi^{-1}\Big(1 - \frac{a}{2}\Big), \quad \widehat{\beta}_k^{\mathrm{db}} + \frac{\widehat{\sigma}_k}{\sqrt{n_0}}\Phi^{-1}\Big(1 - \frac{a}{2}\Big)\Big], \tag{3.14}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal CDF. Leveraging the duality between hypothesis testing and confidence intervals, we can construct the following asymptotically efficient hypothesis testing procedure to test

$$H_0 : \beta_k^* = u \qquad \textit{versus} \qquad H_1 : \beta_k^* \neq u.$$

We reject H_0 if the following condition holds:

$$\Big|\frac{\sqrt{n_0}}{\widehat{\sigma}_k}(\widehat{\beta}_k^{\mathrm{db}} - u)\Big| \geq z_{a/2},$$

where $z_{a/2}$ denotes the $1 - a/2$ quantile of the standard normal distribution.

4. Transferable source detection

Theoretical analysis suggests that the effectiveness of transfer learning is closely tied to the similarity measure $C_{\mathbf{\Sigma}}h$, which is typically unknown in practice. When $C_{\mathbf{\Sigma}}h$ is large, incorporating source data may deteriorate the estimation and inference accuracy for the target parameter. This phenomenon, known as negative transfer, occurs when the inclusion of auxiliary data leads to degraded performance on the target task (Weiss et al., 2016; Pan and Yang, 2009). To address this issue, we develop a detection algorithm tailored for scenarios where both the error sequences and design matrices exhibit temporal dependence. The proposed procedure is described in Algorithm 2. Specifically, we perform standard single-task Lasso regression on each task to obtain the corresponding regression coefficient estimators, denoted by $\widehat{\boldsymbol{w}}^{(k)}$, defined as:

$$\widehat{\boldsymbol{w}}^{(k)} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \Big\{ \frac{1}{n_k} |\boldsymbol{y}^{(k)} - X^{(k)}\boldsymbol{w}|_2^2 + 2\lambda_k |\boldsymbol{w}|_1 \Big\}. \tag{4.15}$$

Several methods have been developed to mitigate the effect of negative transfer. For instance, Li et al. (2022) and Li et al. (2024) introduced different aggregated estimators that combine the transfer learning estimator with the single-task Lasso estimator. Tian and Feng (2023) proposed a procedure that detects and filters out potentially harmful auxiliary samples prior to model fitting. Li et al. (2023a) studied negative transfer in a multi-task learning setting by learning a surrogate model to predict the performance of source-task subsets and selecting relevant sources accordingly. However, these approaches are developed under the assumption of independent error terms and thus are not directly applicable to the temporally dependent data setting considered in this work. Related negative-transfer mitigation strategies in multi-task learning include feature decomposition (Zhou et al.,

Algorithm 2: Transferable Source Detection

Input : Target data $(X^{(0)}, \boldsymbol{y}^{(0)})$, auxiliary datasets $\{(X^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^{M}$, constants

$C > 0, C_1 > 0$

Output: Estimated transferable source set $\widehat{\mathcal{A}}_h$

Step 1: Split the target dataset into two equal parts in order: $(X^{(0)}_{(1)}, \boldsymbol{y}^{(0)}_{(1)})$, $(X^{(0)}_{(2)}, \boldsymbol{y}^{(0)}_{(2)})$.

Step 2: Estimate regression coefficients:

(i) For each k, obtain $\widehat{\boldsymbol{w}}^{(k)}$ by solving the Lasso problem (4.15) on $(X^{(k)}, \boldsymbol{y}^{(k)})$.

(ii) Obtain $\widehat{\boldsymbol{\beta}}^{l}_{(1)}$ and $\widehat{\boldsymbol{\beta}}^{l}_{(2)}$ by solving (4.15) on $\{X^{(0)}_{(1)}, \boldsymbol{y}^{(0)}_{(1)}\}$ and $\{X^{(0)}_{(2)}, \boldsymbol{y}^{(0)}_{(2)}\}$ respectively.

Step 3: Compute prediction losses. For any vector $\boldsymbol{w}$, define

$$\widehat{L}_i(\boldsymbol{w}) = \frac{1}{n_0/2}|\boldsymbol{y}^{(0)}_{(i)} - X^{(0)}_{(i)}\boldsymbol{w}|_2^2, \quad i = 1, 2.$$

Then compute:

(i) $\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) = \frac{1}{2}(\widehat{L}_1(\widehat{\boldsymbol{w}}^{(k)}) + \widehat{L}_2(\widehat{\boldsymbol{w}}^{(k)}))$;

(ii) $\widehat{L}(\widehat{\boldsymbol{\beta}}^{l}) = \frac{1}{2}(\widehat{L}_1(\widehat{\boldsymbol{\beta}}^{l}_{(2)}) + \widehat{L}_2(\widehat{\boldsymbol{\beta}}^{l}_{(1)}))$.

Step 4: Construct the transferable source set. Let $\Delta = |\widehat{L}_1(\widehat{\boldsymbol{\beta}}^{l}_{(2)}) - \widehat{L}_2(\widehat{\boldsymbol{\beta}}^{l}_{(1)})|$, and define

$$\widehat{\mathcal{A}}_h \leftarrow \{k \neq 0 : \widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^{l}) \leq C \cdot \max\{\Delta, C_1\}\}.$$

2023) and alignment-based methods (Wu et al., 2020), which are largely empirical and lack theoretical guarantees for statistical inference. More recently, Ma and Safikhani (2025) developed a screening algorithm to mitigate negative transfer in the time series setting. However, their method is tailored to VAR models and focuses on excluding detrimental auxiliary samples, without providing guarantees on the retention of all auxiliary sources

that are sufficiently similar to the target.

It is worth emphasizing that Algorithm 2 does not require the input of h, and the target data can be sequentially partitioned into any number of equal parts in Step 1; the two-fold split adopted here is merely for simplicity. Furthermore, we show that, under suitable regularity conditions, the proposed transferable source detection algorithm can accurately recover the level-h transferable set $\mathcal{A}_h$ with high probability. Define the sparsity of the auxiliary regression coefficient $\boldsymbol{w}^{(k)}$ as $|\boldsymbol{w}^{(k)}|_0 = s_k$.

Condition 3. *For $k \notin \mathcal{A}_h$, there exists a sufficiently large constant $c > 0$ such that*

$$|\boldsymbol{\delta}^{(k)}|_2^2 \geq c[(s_0\lambda_\delta^2 + s_k\lambda_\delta\lambda_k + s_k\lambda_k^2 + \lambda_\delta + \sqrt{s_k}\lambda_k|\boldsymbol{\delta}^{(k)}|_2) \vee 1].$$

Moreover, there exists a sufficiently small constant $\varepsilon_0 > 0$ such that, for sufficiently large n_0 and n_k, the following inequality holds for any $k \in \mathcal{A}_h$:

$$s_0\lambda_\delta^2 + s_k\lambda_\delta\lambda_k + s_k\lambda_k^2 + \lambda_\delta + h^2 < \varepsilon_0.$$

Condition 3 consists of two parts. The first part ensures that for sources not in $\mathcal{A}_h$, there exists a sufficiently large gap between the auxiliary sample regression coefficient $\boldsymbol{w}^{(k)}$ and the true coefficient $\boldsymbol{\beta}^*$ of the target data. This discrepancy primarily determines the magnitude of $\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l)$ for $k \notin \mathcal{A}_h$. The second part of Condition 3 guarantees that for $k \in \mathcal{A}_h$, the difference $\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l)$ remains sufficiently small as the sample sizes tend to infinity. Further details on the behavior of $\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l)$ are provided in Section S6 of the Supplementary Material. A similar restriction appears in Assumption 5 of Tian and Feng (2023).

Theorem 4. *Suppose that the conditions of Theorem 1 and Condition 3 hold. Define λ_k to be the value obtained by replacing n_0 in λ_δ with n_k. Furthermore, assume that there*

exist constants $C_1 > 0$ *and* $C_2 > 0$ *such that* $|\boldsymbol{\beta}^*|_2 \le C_1$ *and* $|\boldsymbol{w}^{(k)}|_2 \le C_2$ *for all* $k \in [M]$. *Then, for any* $\delta > 0$, *there exist positive constants* $C(\delta)$, $C_1(\delta)$, *and* $N(\delta)$ *such that when the thresholds* $C = C(\delta)$ *and* $C_1 = C_1(\delta)$ *are employed in Step 4 of Algorithm 2, and* $n_{\min} > N(\delta)$, *we have*

$$\mathbb{P}(\widehat{\mathcal{A}}_h = \mathcal{A}_h) \ge 1 - \delta.$$

Theorem 4 establishes that, under appropriate conditions, our transferable set detection algorithm can accurately recover the level-h transferring set $\mathcal{A}_h$, in the sense that $\widehat{\mathcal{A}}_h = \mathcal{A}_h$, with high probability. As shown in Lemma 12 of the Supplementary Material, under the functional dependence framework, the convergence rate of the single-task Lasso estimator satisfies

$$|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_2^2 = O_{\mathbb{P}}(s_0\lambda_\delta^2), \qquad |\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_1 = O_{\mathbb{P}}(s_0\lambda_\delta).$$

Hence, for the proposed transfer learning procedure to achieve a faster convergence rate than the Lasso estimator applied solely to the target data, a sufficient condition is that $C_{\boldsymbol{\Sigma}} h \ll \sqrt{s_0}\,\lambda_\delta$ and $n_{\mathcal{A}_h} \gg n_0$. Accordingly, if we further assume that the set $\mathcal{A}_h$ satisfies $C_{\boldsymbol{\Sigma}} h \ll \sqrt{s_0}\,\lambda_\delta$, as similar conditions are also required in Tian and Feng (2023), then incorporating the auxiliary datasets indexed by $\widehat{\mathcal{A}}_h$ into the transfer learning procedure described in Section 2.3 leads to a strictly faster convergence rate compared with applying the Lasso solely to the target dataset. We refer to this procedure as SD Trans-Lasso.

Remark 2. *Suppose that there exists a transferring set* $\mathcal{A}_h$ *such that* $C_{\boldsymbol{\Sigma}} h \lesssim \sqrt{s_0}\lambda_\delta$. *For the estimator* $\widehat{\boldsymbol{\beta}}^{\mathrm{SD}}$ *obtained via SD Trans-Lasso, under the conditions of Theorem 4, we have*

$$\frac{1}{n_0}|X^{(0)}(\widehat{\boldsymbol{\beta}}^{\mathrm{SD}} - \boldsymbol{\beta}^*)|_2^2 \lesssim \left(s_0\lambda_w^2 + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2\right) \wedge s_0\lambda_\delta^2,$$

and

$$|\widehat{\boldsymbol{\beta}}^{\mathrm{SD}} - \boldsymbol{\beta}^*|_1 \lesssim \left(s_0\lambda_w + \tfrac{(C_{\boldsymbol{\Sigma}}h)^2}{\lambda_\delta} + C_{\boldsymbol{\Sigma}}h\right) \wedge s_0\lambda_\delta,$$

with probability tending to 1 *as* $n_{\min} \to \infty$.

Moreover, for the estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{SD}}$, the statistical inference framework previously developed for $\widehat{\boldsymbol{\beta}}$ can be directly applied to construct the corresponding bias-corrected estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{SDdb}}$ and to establish its asymptotic normality. Specifically, let $\widehat{\beta}_j^{\mathrm{SDdb}}$ denote the debiased estimator defined in Equation (2.5), with $\widehat{\boldsymbol{\beta}}$ replaced by $\widehat{\boldsymbol{\beta}}^{\mathrm{SD}}$, and let $\widehat{\gamma}_k^{\epsilon}$ in Equation (3.13) be computed based on the residuals

$$e_i^{\mathrm{SD}} = y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \widehat{\boldsymbol{\beta}}^{\mathrm{SD}}, \quad i = 1, \ldots, n_0.$$

Then, the asymptotic normality of $\widehat{\beta}_j^{\mathrm{SDdb}}$ can be established in the same manner. The details of this procedure are omitted here for brevity.

5. Simulation

In this section, we evaluate the practical performance of the proposed methods—the Trans-Lasso method introduced in Section 2.3 and the SD Trans-Lasso method presented in Section 4—in comparison with the single-task learning approach of Yuan and Guo (2022), through comprehensive simulation studies. The results highlight the substantial benefits of leveraging transfer learning. Furthermore, to assess the effectiveness of our statistical inference procedure, we include a comparison with the method of Tian and Feng (2023), which also adopts a transfer learning approach but assumes independent errors in high-dimensional generalized linear models. This comparison further highlights the importance of accounting for error autocovariance in the presence of temporal dependence. We set

the target sample size to $n_0 = 300$, and consider a high-dimensional setting with $p = 500$ covariates. The auxiliary datasets consist of $M = 20$ studies, each with sample size $n_j = 200$ for $j \in [M]$. For each $k \in \{M\}$, we first independently generate $\boldsymbol{z}_i^{(k)} \sim N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}^{(k)})$, and then construct the covariates as $\boldsymbol{x}_i^{(k)} = 0.3\,\boldsymbol{x}_{i-1}^{(k)} + \sqrt{1-0.3^2}\boldsymbol{z}_i^{(k)}$. The error process follows an ARMA(1,3) model across all datasets:

$$\epsilon_i^{(k)} = 0.15\epsilon_{i-1}^{(k)} + \xi_i^{(k)} + 0.4\xi_{i-1}^{(k)} + 0.25\xi_{i-2}^{(k)} + 0.7\xi_{i-3}^{(k)}, \quad \xi_i^{(k)} \overset{\text{i.i.d.}}{\sim} N(0,1), \quad i = 1, \ldots, n_k.$$

Two configurations of the covariance matrices are considered.

(a) For each $k \in [M]$, we construct a Toeplitz matrix whose first row is given by

$$\widetilde{\boldsymbol{\Sigma}}_{1,\cdot}^{(k)} = \Big(1, \underbrace{1/(k+1), \ldots, 1/(k+1)}_{2k-1 \text{ times}}, \underbrace{0, \ldots, 0}_{p-2k \text{ times}}\Big),$$

moreover, $\widetilde{\boldsymbol{\Sigma}}^{(0)} = \mathbf{I}_p$.

(b) We consider an equi-correlated structure for $\widetilde{\boldsymbol{\Sigma}}^{(0)}$, where $\widetilde{\boldsymbol{\Sigma}}_{j,k}^{(0)} = 0.3$ for $j \neq k$ and $\widetilde{\boldsymbol{\Sigma}}_{j,j}^{(0)} = 1$. For each $k \in [M]$, we generate a random matrix $A^{(k)} \in \mathbb{R}^{p \times p}$, where each entry independently takes the value 0.1 with probability 0.1 and 0 with probability 0.9. The covariance matrix is then defined as

$$\widetilde{\boldsymbol{\Sigma}}^{(k)} = (A^{(k)})^\top A^{(k)} + \mathbf{I}_p, \quad \text{for } k = 1, \ldots, M.$$

In both settings (a) and (b), the rows of $X^{(k)}$ are correlated, and the covariance matrices $\boldsymbol{\Sigma}^{(k)}$ vary across sources for $k = 0, \ldots, M$. In setting (a), $\widetilde{\boldsymbol{\Sigma}}^{(0)}$ is sparse, whereas in setting (b), it is non-sparse. Details on the selection of tuning parameters are provided in Section S2 of the Supplementary Material, including the bandwidth for estimating the error autocovariance matrix, the Lasso regularization parameters, and other implementation

details. We also consider additional temporal dependence settings, with the corresponding results reported in Section S1 of the Supplementary Material.

5.1 Estimation performance for $\boldsymbol{\beta}^*$

For the target study, the coefficient vector is set to be $\boldsymbol{\beta}^* = (0.3, \ldots, 0.3, 0, \ldots, 0)^\top$, with $|\boldsymbol{\beta}^*|_0 = 10$. For the auxiliary regression coefficients, we differentiate between informative and non-informative sources. If $k \in \mathcal{A}_h$, we define the heterogeneous components as $H_k = \{1, \ldots, 100\}$, and for $j = 1, \ldots p$, set

$$w_j^{(k)} = \beta_j^* + \xi_j^{(k)} \mathbb{I}(j \in H_k), \quad \text{where } \xi_j^{(k)} \overset{\text{i.i.d.}}{\sim} N(0, h/40).$$

Here $h \in \{1, 2, 3\}$ controls the similarity level between the auxiliary and target coefficients. For non-informative sources $k \notin \mathcal{A}_h$, we use the same heterogeneous component $H_k = \{1, \ldots, 100\}$, but for $j = 1, \ldots, p$, let

$$w_j^{(k)} = \beta_j^* + \xi_j^{(k)} \mathbb{I}(j \in H_k), \quad \text{where } \xi_j^{(k)} \overset{\text{i.i.d.}}{\sim} N(0, 1/3).$$

We vary the number of informative sources $|\mathcal{A}_h| \in \{0, 4, 8, 12, 16, 20\}$ to examine the impact of transferability.

In Figures 1 and 2, we report the mean squared prediction error (MSPE) and the sum of absolute errors (SAE) for each estimator $\boldsymbol{b}$, defined as $|X^{(0)}(\boldsymbol{b} - \boldsymbol{\beta}^*)|_2^2/n_0$ and $|\boldsymbol{b} - \boldsymbol{\beta}^*|_1$, respectively, under different covariance structures. Each point in the figures represents an average over 500 independent replications.

As expected, the performance of the method proposed by Yuan and Guo (2022) remains stable as $|\mathcal{A}_h|$ increases, no matter whether the covariance matrix $\widetilde{\boldsymbol{\Sigma}}^{(0)}$ is sparse or not. In contrast, the estimation errors of Trans-Lasso and SD Trans-Lasso decrease notably as

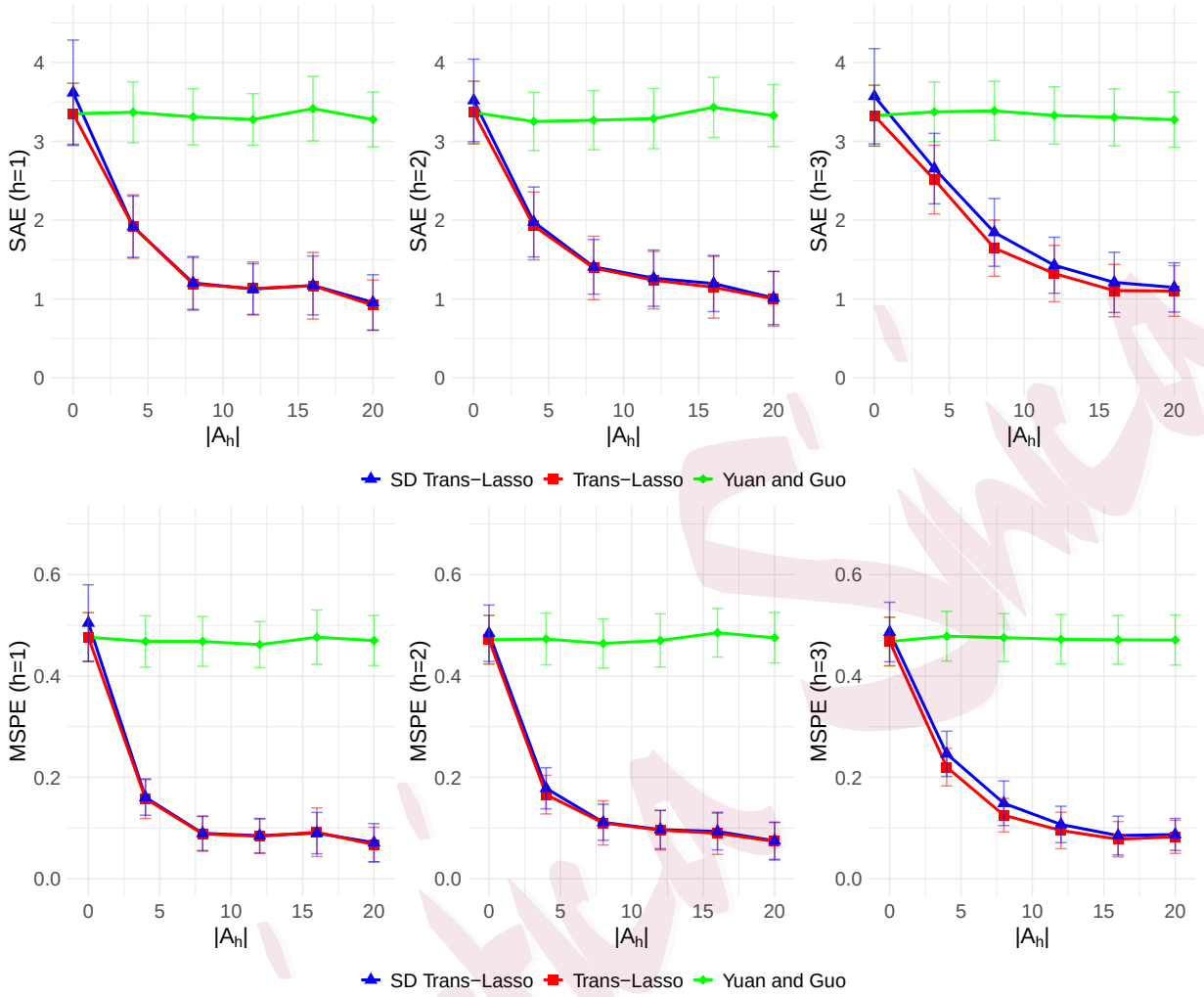

Figure 1: Estimation errors of Yuan and Guo (2022), Trans-Lasso, and SD Trans-Lasso in setting (a). The two rows correspond to SAE(i) and MSPE(ii) respectively. Error bars denote standard deviations divided by 2.5.

$|\mathcal{A}_h|$ increases, highlighting their ability to effectively leverage auxiliary information. The SD Trans-Lasso method closely mirrors the behavior of Trans-Lasso, indicating that the proposed transferable source detection algorithm can accurately identify the informative source set $\mathcal{A}_h$. As h increases, the estimation task becomes more difficult, leading to larger

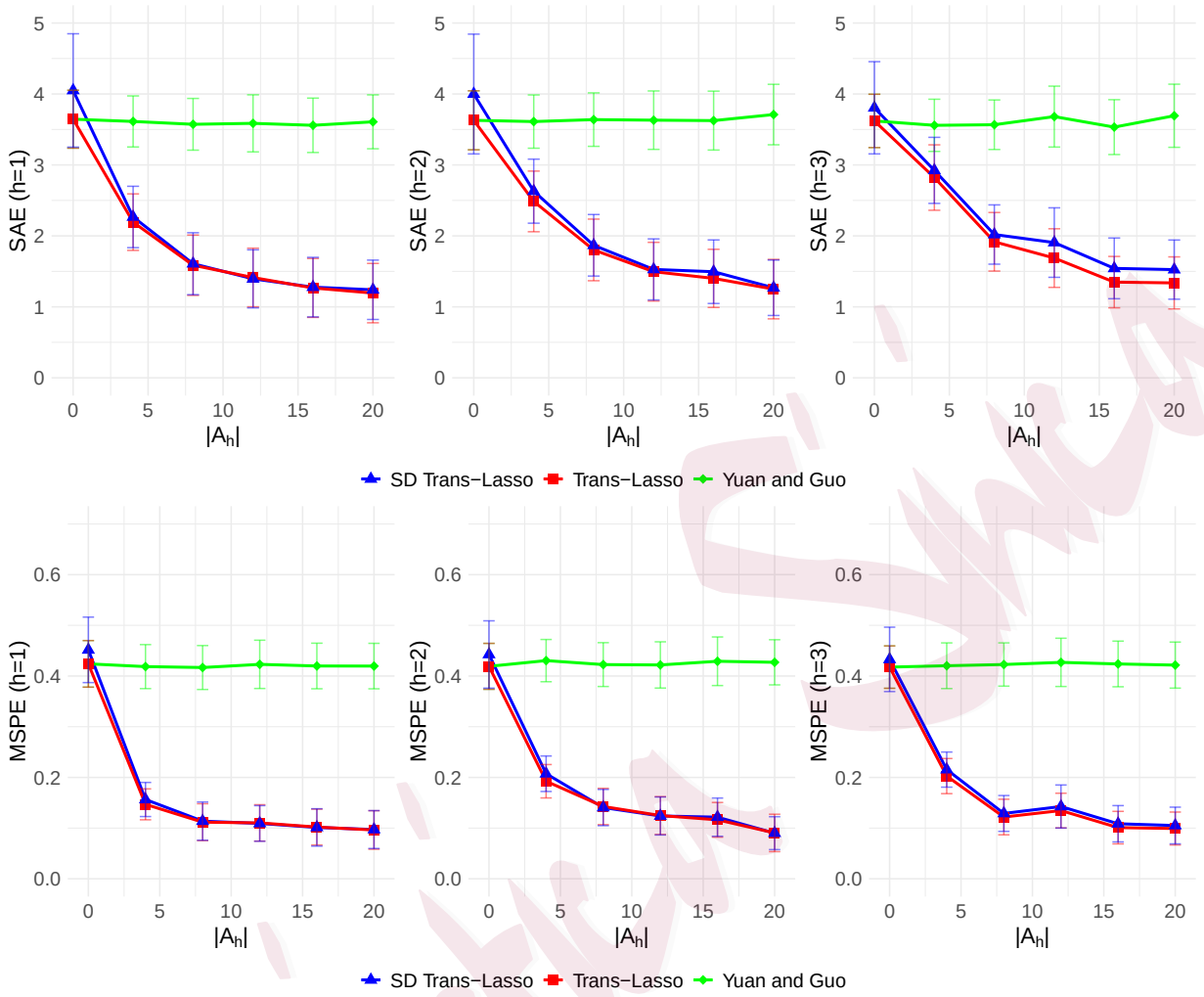

Figure 2: Estimation errors of Yuan and Guo (2022), Trans-Lasso, and SD Trans-Lasso in setting (b). The two rows correspond to SAE(i) and MSPE(ii) respectively. Error bars denote standard deviations divided by 2.5.

errors for both transfer learning methods. Nonetheless, for both SAE and MSPE, Trans-Lasso consistently achieves lower estimation error than the baseline single-task method of Yuan and Guo (2022), further demonstrating the benefits of incorporating auxiliary data.

5.2 Confidence intervals for $\boldsymbol{\beta}_j^*$

To assess the performance of our method in statistical inference, we construct 95% two-sided confidence intervals for the target coefficients β_j^*, with $j = 1, \ldots, p$.

The true target coefficient vector is specified as $\boldsymbol{\beta}^* = (0.5, \ldots, 0.5, 0, \ldots, 0)^\top$, with $|\boldsymbol{\beta}^*|_0 = 10$. For auxiliary regression coefficients, if $k \in \mathcal{A}_h$, we define $H_k = \{1, \ldots, 100\}$ and generate

$$w_j^{(k)} = \beta_j^* + \xi_j^{(k)} \mathbb{I}(j \in H_k), \quad \xi_j^{(k)} \overset{\text{i.i.d.}}{\sim} N(0, h/30), \quad j = 1, \ldots, p,$$

with $h \in \{1, 2, 3\}$. For $k \notin \mathcal{A}_h$, we similarly define $H_k = \{1, \ldots, 100\}$ and draw

$$w_j^{(k)} = \beta_j^* + \xi_j^{(k)} \mathbb{I}(j \in H_k), \quad \xi_j^{(k)} \overset{\text{i.i.d.}}{\sim} N(0, 1/3), \quad j = 1, \ldots, p.$$

We consider $|\mathcal{A}_h| \in \{4, 8, 12, 16, 20\}$.

A total of 600 Monte Carlo repetitions are conducted. Table 1 reports the results under setting (a), where the covariance matrix $\widetilde{\boldsymbol{\Sigma}}^{(0)}$ is row-sparse. Table 2 presents the results under setting (b), where $\widetilde{\boldsymbol{\Sigma}}^{(0)}$ is not row-sparse. For comparison, we implement the method of Tian and Feng (2023) using the `glmtrans` R package provided by the authors.

We observe that for true signals with $\beta_j^* = 0.5$, our debiased SD Trans-Lasso estimator achieves reliable confidence interval coverage under both settings and significantly outperforms the methods of Yuan and Guo (2022) and Tian and Feng (2023). The improvement over Yuan and Guo (2022) is largely due to the smaller remaining bias of our debiased estimator, as demonstrated in Section S1 of the Supplementary Material. The advantage over Tian and Feng (2023) primarily stems from the more accurate estimation of the error autocovariance matrix. As indicated by (3.14), the length of the confidence

interval is directly influenced by $\widehat{\sigma}_k^2$, which in turn depends on the accuracy of the estimator for $\mathbf{\Sigma}_{n_0}^{\epsilon}$. In both Table 1 and Table 2, the confidence intervals constructed by Tian and Feng (2023) are substantially shorter than those produced by Yuan and Guo (2022) and our debiased SD Trans-Lasso method, highlighting the critical role of accurate error autocovariance estimation in achieving valid and efficient statistical inference. For the case where $\beta_j^* = 0$, under setting (b), all methods achieve reliable coverage. However, under setting (a), the method proposed by Tian and Feng (2023) performs worse than the debiased SD Trans-Lasso method and the method of Yuan and Guo (2022). More discussion about this simulation study can be found in the Supplementary Material, Section S1.

6. A real data study

In this section, we apply the proposed method to macroeconomic data collected from New York, California, and Texas, with the goal of predicting unemployment rates based on the fitted models. The dataset comprises monthly time series observations from January 2009 to January 2019, covering 212 variables and 121 time points. Given that the data pertain to economic development across different U.S. states, it is reasonable to expect a certain degree of correlation among the datasets. However, the underlying data-generating mechanisms may differ across regions. Therefore, transfer learning offers a natural and promising approach for this application. The original data are publicly available from the Federal Reserve Economic Data (FRED®) repository (`https://fred.stlouisfed.org/`).

Prior to analysis, we performed standard preprocessing procedures. Specifically, variables with incomplete recordings due to delayed initiation of data collection were removed, resulting in a total of 185 retained variables. In all analyses, the unemployment rate was

Table 1: Average coverage probabilities for $\beta_8^* = 0.5$ and $\beta_{16}^* = 0$ in setting (a), with average confidence interval lengths in brackets.

h	$\lvert\mathcal{A}_h\rvert$	Yuan and Guo (2022)		Debiased SD Trans-Lasso		Tian and Feng (2023)	
		β_8^*	β_{16}^*	β_8^*	β_{16}^*	β_8^*	β_{16}^*
1	4	0.872 (0.333)	0.943 (0.331)	0.933 (0.335)	0.948 (0.334)	0.832 (0.232)	0.900 (0.232)
1	8	0.890 (0.336)	0.952 (0.337)	0.947 (0.339)	0.960 (0.337)	0.818 (0.232)	0.898 (0.233)
1	12	0.902 (0.337)	0.937 (0.338)	0.948 (0.337)	0.937 (0.337)	0.895 (0.232)	0.887 (0.232)
1	16	0.890 (0.336)	0.938 (0.337)	0.943 (0.339)	0.950 (0.338)	0.852 (0.232)	0.893 (0.232)
1	20	0.878 (0.332)	0.952 (0.331)	0.947 (0.333)	0.945 (0.333)	0.872 (0.232)	0.890 (0.232)
2	4	0.875 (0.335)	0.942 (0.333)	0.905 (0.339)	0.953 (0.337)	0.800 (0.232)	0.895 (0.232)
2	8	0.895 (0.337)	0.960 (0.336)	0.950 (0.340)	0.945 (0.340)	0.825 (0.233)	0.920 (0.232)
2	12	0.875 (0.333)	0.955 (0.332)	0.945 (0.336)	0.940 (0.334)	0.853 (0.232)	0.900 (0.232)
2	16	0.897 (0.332)	0.947 (0.332)	0.943 (0.332)	0.950 (0.333)	0.867 (0.232)	0.912 (0.232)
2	20	0.892 (0.334)	0.948 (0.334)	0.945 (0.336)	0.955 (0.335)	0.867 (0.232)	0.910 (0.232)
3	4	0.900 (0.332)	0.945 (0.333)	0.950 (0.340)	0.938 (0.340)	0.887 (0.232)	0.890 (0.232)
3	8	0.858 (0.331)	0.947 (0.333)	0.923 (0.335)	0.953 (0.337)	0.850 (0.232)	0.923 (0.232)
3	12	0.912 (0.339)	0.943 (0.339)	0.950 (0.338)	0.945 (0.337)	0.867 (0.232)	0.897 (0.232)
3	16	0.873 (0.333)	0.945 (0.333)	0.943 (0.336)	0.943 (0.335)	0.847 (0.232)	0.892 (0.232)
3	20	0.887 (0.337)	0.950 (0.335)	0.925 (0.339)	0.952 (0.339)	0.853 (0.232)	0.913 (0.232)

used as the response variable, while the remaining variables served as predictors. All predictor variables were standardized before model fitting.

We consider each of the three states—New York, California, and Texas—as the target study in turn, treating the other two as source studies. For each state, we designate data from January 2009 to June 2016 (a total of 90 observations) as the training set, and use the

Table 2: Average coverage probabilities for $\beta_8^* = 0.5$ and $\beta_{16}^* = 0$ in setting (b), with average confidence interval lengths in brackets.

h	$\lvert\mathcal{A}_h\rvert$	Yuan and Guo (2022)		Debiased SD Trans-Lasso		Tian and Feng (2023)	
		β_8^*	β_{16}^*	β_8^*	β_{16}^*	β_8^*	β_{16}^*
1	4	0.893 (0.403)	0.958 (0.403)	0.937 (0.392)	0.955 (0.392)	0.775 (0.197)	0.950 (0.196)
1	8	0.892 (0.403)	0.943 (0.404)	0.932 (0.390)	0.932 (0.391)	0.755 (0.197)	0.932 (0.196)
1	12	0.897 (0.404)	0.928 (0.404)	0.942 (0.392)	0.925 (0.394)	0.865 (0.197)	0.930 (0.196)
1	16	0.908 (0.408)	0.940 (0.407)	0.953 (0.395)	0.955 (0.395)	0.878 (0.197)	0.957 (0.196)
1	20	0.898 (0.406)	0.955 (0.405)	0.948 (0.395)	0.955 (0.395)	0.903 (0.197)	0.940 (0.196)
2	4	0.878 (0.405)	0.937 (0.406)	0.925 (0.396)	0.945 (0.395)	0.623 (0.197)	0.932 (0.196)
2	8	0.915 (0.407)	0.935 (0.406)	0.943 (0.397)	0.950 (0.396)	0.683 (0.197)	0.947 (0.196)
2	12	0.873 (0.401)	0.948 (0.403)	0.953 (0.390)	0.940 (0.393)	0.765 (0.197)	0.938 (0.196)
2	16	0.915 (0.403)	0.943 (0.406)	0.958 (0.393)	0.938 (0.394)	0.913 (0.197)	0.942 (0.196)
2	20	0.903 (0.404)	0.917 (0.404)	0.932 (0.393)	0.927 (0.392)	0.872 (0.197)	0.932 (0.196)
3	4	0.907 (0.402)	0.937 (0.405)	0.942 (0.395)	0.948 (0.396)	0.810 (0.197)	0.937 (0.196)
3	8	0.910 (0.403)	0.933 (0.403)	0.945 (0.396)	0.948 (0.396)	0.585 (0.197)	0.948 (0.196)
3	12	0.890 (0.402)	0.930 (0.404)	0.935 (0.393)	0.950 (0.393)	0.800 (0.197)	0.945 (0.197)
3	16	0.915 (0.403)	0.937 (0.403)	0.960 (0.395)	0.940 (0.396)	0.790 (0.197)	0.927 (0.196)
3	20	0.887 (0.405)	0.937 (0.406)	0.923 (0.394)	0.937 (0.394)	0.762 (0.198)	0.932 (0.196)

remaining observations as the test set. We first fit the proposed model on the training set and denote the set of variables with nonzero estimated coefficients as $\widehat{\mathcal{S}}_1$. Next, we conduct hypothesis testing for each component in $\widehat{\mathcal{S}}_1$, testing the null hypothesis $H_{0j}\colon \beta_j^* = 0$ against the alternative $H_{1j}\colon \beta_j^* \neq 0$, using a significance level of 0.01. Variables in $\widehat{\mathcal{S}}_1$ with p-values less than 0.01 are retained, and we denote their index set as $\widehat{\mathcal{S}}_2$. Finally, we construct

a linear regression model on the test set using only the variables in $\widehat{\mathcal{S}}_2$ to predict the unemployment rate, yielding the final fitted model.

We compare the prediction performance of our proposed debiased SD Trans-Lasso with that of two benchmark methods: the single-task learning method in Yuan and Guo (2022), and the transfer learning method under the independent error assumption proposed by Tian and Feng (2023). Prediction accuracy is evaluated using the Predicted Mean Squared Error (PMSE), calculated as $\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2/n$, where y_i denotes the actual unemployment rate and $\widehat{y}_i$ is the predicted value from the corresponding model.

As shown in Table 3, our SD Trans-Lasso consistently achieves lower prediction errors compared to both the single-task and the independent-error transfer learning methods. Additionally, the predicted versus actual unemployment rates for Texas, California, and New York over July 2016–January 2019 are presented in Figures 3–5 of Section S1 in the Supplementary Material.

Table 3: Comparison of PMSE Values Across Different Methods

Method	Yuan and Guo (2022)	SD Trans-Lasso	Tian and Feng (2023)
New York	0.036	0.016	0.214
California	0.023	0.010	0.040
Texas	0.022	0.005	0.090

7. Conclusion

Our main contribution lies in advancing transfer learning for high-dimensional time series regression. Specifically, we first establish the convergence rate of the transfer learning

estimator under temporally dependent data. To facilitate valid statistical inference for the target regression coefficients $\boldsymbol{\beta}^*$, we propose a novel debiasing procedure and establish the asymptotic normality of the resulting debiased estimator, relying on the consistency of a banded estimator for the error autocovariance matrix. To alleviate the impact of negative transfer, we further develop a transferable source detection algorithm that selectively retains source samples exhibiting sufficient similarity to the target domain. Finally, we validate the proposed methodology through extensive simulation studies and a real data application.

Several promising directions remain for future research. One important direction is to investigate whether negative transfer can be mitigated through adaptive weighting of auxiliary samples—assigning larger weights to samples whose regression coefficients are more similar to those of the target task, and smaller weights to those that are less similar. Another limitation of the current approach is that it does not shorten the length of the confidence intervals, which asymptotically scale as $1/\sqrt{n_0}$, as with many existing transfer learning methods. Addressing this limitation may require the development of new debiasing techniques, which presents a valuable avenue for further investigation.

Supplementary Material

The Supplementary Material includes detailed proofs of the main theorems and necessary lemmas, additional numerical results, and comprehensive guidelines for tuning parameter selection.

Acknowledgements

We thank the editor, associate editor and two referees for their helpful comments and suggestions. Jia's research was partially supported by National Natural Science Foundation of China, Grant 12501374, and Shanghai Natural Science Foundation, Grant 25ZR1402404. Guo's research was supported by the National Natural Science Foundation of China, Grant 12471267.

References

Adamek, R., Smeekes, S., and Wilms, I. (2023). Lasso inference for high-dimensional time series. *Journal of Econometrics*, 235:1114–1143.

Bastani, H. (2021). Predicting with proxies: transfer learning in high dimension. *Management Science*, 67:2964–2984.

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

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227.

Chernozhukov, V., Karl Härdle, W., Huang, C., and Wang, W. (2021). Lasso-driven inference in time and space. *The Annals of Statistics*, 49:1702–1735.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duan, J., Pelger, M., and Xiong, R. (2024). Target PCA: Transfer learning large dimensional panel data. *Journal of Econometrics*, 244:105–121.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39:3320–3356.

Hajiramezanali, E., Zamani Dadaneh, S., Karbalayghareh, A., Zhou, M., and Qian, X. (2018). Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. *Neural Information Processing Systems*, 31:9133–9142.

Hannan, E. J. and Deistler, M. (2012). *The Statistical Theory of Linear Systems*. SIAM, New York.

He, B., Liu, H., Zhang, X., and Huang, J. (2024). Representation transfer learning for semiparametric regression. *arXiv preprint arXiv:2406.13197*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Kolmogoroff, A. (1928). Über die Summen durch den Zufall bestimmter unabhängiger Größen. *Mathematische Annalen*, 99:309–319.

Li, D., Nguyen, H. L., and Zhang, H. R. (2023a). Identification of negative transfers in multitask learning using surrogate models. *arXiv preprint arXiv:2303.14582*.

Li, S. (2020). Debiasing the debiased lasso with bootstrap. *Electronic Journal of Statistics*, 14:2298–2337.

Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84:149–173.

Li, S., Cai, T. T., and Li, H. (2023b). Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, 118:2171–2183.

Li, S., Zhang, L., Cai, T. T., and Li, H. (2024). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 119:1274–1285.

Lin, Y., Zhu, Q., and Li, G. (2025). Improving time series estimation and prediction via transfer learning. *arXiv*

preprint arXiv:2510.25236.

Ma, M. and Safikhani, A. (2025). Transfer learning for high-dimensional reduced rank time series models. *Proceedings of Machine Learning Research*, 258:2926–2934.

Ma, Y., Gong, W., and Mao, F. (2015). Transfer learning used to analyze the dynamic evolution of the dust aerosol. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 153:119–130.

Nguyen, T.-T. and Yoon, S. (2019). A novel approach to short-term stock price movement prediction using transfer learning. *Applied Sciences*, 9:4745.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118:2684–2697.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI global, Hershey, Pennsylvania.

Tripuraneni, N., Jordan, M., and Jin, C. (2020). On the theory of transfer learning: the importance of task diversity. *Neural Information Processing Systems*, 33:7852–7862.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42:1166–1202.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3:1–40.

Wu, S., Zhang, H. R., and Ré, C. (2020). Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102:14150–14154.

Wu, W. B. and Wu, Y. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10:352–379.

Xia, J., Chen, Y., and Guo, X. (2024). Inference for high-dimensional linear models with locally stationary error processes. *Journal of Time Series Analysis*, 45:78–102.

Xiao, J., Hu, Y., Xiao, Y., Xu, L., and Wang, S. (2017). A hybrid transfer learning model for crude oil price forecasting. *Statistics and Its Interface*, 10:119–130.

Yuan, P. and Guo, X. (2022). High-dimensional inference for linear model with correlated errors. *Metrika*, 85:21–52.

Zhang, C. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76:217–242.

Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45:1895–1919.

Zhao, L., Pan, S., Xiang, E., Zhong, E., Lu, Z., and Yang, Q. (2013). Active transfer learning for cross-system recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27:1205–1211.

Zhou, J., Yu, Q., Luo, C., and Zhang, J. (2023). Feature decomposition for reducing negative transfer: a novel multi-task learning method for recommender system (student abstract). *Proceedings of the AAAI conference on Artificial Intelligence*, 37:16390–16391.

Zhu, Y. and Bradic, J. (2018). Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12:3312–3364.

Zhu, Y., Yu, W., and Li, X. (2025). A multi-objective transfer learning framework for time series forecasting with Concept Echo State Networks. *Neural Networks*, 186:107272.

Zongqi Liu

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei,

Anhui 230026, China

E-mail: lzq20030713@mail.ustc.edu.cn

Shengji Jia

School of Statistics and Mathematics; Interdisciplinary Research Institute of Data Science, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China

E-mail: 20200026@lixin.edu.cn

Xiao Guo

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, Anhui 230026, China

E-mail: xiaoguo@ustc.edu.cn