Statistica Sinica Preprint No: SS-2025-0202	
Title	Knowledge Transfer for Sparse Part Linear Models with Privacy Guarantee: Estimation, Inference and Multiple Testing
Manuscript ID	SS-2025-0202
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0202
Complete List of Authors	Zhengyu Zhu, Yibo Yan, Heng Lian and Riquan Zhang
Corresponding Authors	Riquan Zhang
E-mails	zhangriquan@163.com
Notice: Accepted author version.	

KNOWLEDGE TRANSFER FOR SPARSE PARTIAL LINEAR MODELS WITH PRIVACY GUARANTEE: ESTIMATION, INFERENCE AND MULTIPLE TESTING

Zhengyu Zhu[1,3,4,†], Yibo Yan[2,†], Heng Lian[3,4], Riquan Zhang[2,*]

[1] *School of Statistics, East China Normal University, Shanghai, China*

2 *School of Statistics and Data Science,*

Shanghai University of International Business and Economics, Shanghai, China

[3] *City University of Hong Kong Shenzhen Research Institute, Shenzhen, China*

[4] *Department of Mathematics, City University of Hong Kong, Hong Kong, China*

Abstract: **Transfer learning leverages knowledge from a source domain to enhance estimation or prediction accuracy in a target task. To strengthen data privacy protection when aggregating information across different sources and targets, differential privacy offers a promising solution. In this work, we propose a transfer learning framework for high-dimensional sparse partial linear models with a novel differential privacy guarantee. Our main algorithm consists of two steps. The first step constructs a surro-**

*Corresponding author. E-mail address: zhangriquan@163.com.
†Equally contributed.

gate linear model by removing the non-linear component in the target model. The second step applies noisy gradient aggregation to transfer information from source domains while preserving privacy guarantees. Theoretically, we establish a nearly optimal error bound for the proposed transfer method in partial linear model estimation, while incurring an acceptable privacy cost. Moreover, the debiased LASSO method is adopted to construct confidence intervals. Finally, we use an e-value based multiple testing approach to control the false discovery rate. The effectiveness of our method is demonstrated through simulation studies and further supported by its application to real-world data.

Key words and phrases: Transfer learning, partial linear model, RKHS, differential privacy, high-dimensional inference, e-Benjamini-Hochberg.

1. Introduction

Semiparametric regression is a well-established statistical tool that balances the flexibility of nonparametric models and the drawback of "curse of dimensionality" (Bickel et al., 1993; Kosorok, 2008). The topic of semiparametric regression estimation can be dated back to Engle et al. (1986). One of the most prominent examples among semiparametric models is the partial linear model (PLM), which can be formulated as follows:

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + g(\boldsymbol{W}) + \varepsilon. \tag{1.1}$$

Here $\boldsymbol{X} = (X_1, ..., X_p)^\top \in \mathbb{R}^p$, $Y \in \mathbb{R}$, $\boldsymbol{W} \in \mathbb{R}^d$, $g : \mathbb{R}^d \to \mathbb{R}$, and ε represents the Gaussian noise with zero mean and variance σ^2. Equation (1.1) provides the primary model, in which $\boldsymbol{\beta} \in \mathbb{R}^p$ is often the main parameter of interest. If the predictor $\boldsymbol{X}$ is exogenous conditional on the control variable $\boldsymbol{W}$, then $\boldsymbol{\beta}$ can be interpreted as the treatment effect in causal analysis or the "lift" parameter in business applications. A canonical problem in partial linear regression is to estimate and make statistical inference for the parameter in the presence of the high-dimensional setting. Among recent studies, Xie and Huang (2009), Zhu (2017) and Lian et al. (2019) focused on developing efficient methods to estimate both the parametric and nonparametric parts. From another perspective, Zhu et al. (2019), Tan et al. (2021), Liu et al. (2020) and Zhao et al. (2023) paid more attention to implement statistical inference on regression coefficients in the high-dimensional settings. Recently, some advanced work has been done to combine PLM with other regression models, including quantile regression model (Shi et al., 2026), single index model (Cui et al., 2025) and factor-augmented model (Shi et al., 2025). In addition, the PLM can also be generalized to different types of data structure, such as longitudinal data (Kim et al., 2017; Li et al., 2024) and functional data (Wong et al., 2019; Ling et al., 2025).

In recent years, transfer learning has emerged as a powerful technique for improving the performance of learning algorithms (Torrey and Shavlik, 2010). To facilitate knowledge transfer, researchers have been developing a range of theoretically grounded methods. For instance, Bastani (2021) proposed a two-step framework that first employs a surrogate contrast vector to extract transferable components, followed by a debiasing step. This approach has been proven to be effective in handling high-dimensional problems across multiple settings, see e.g., Li et al. (2022), Tian and Feng (2023), Zhang and Zhu (2025) and Wang and Yu (2025). A comprehensive overview of recent advances in statistical transfer learning is provided in Zhu et al. (2025). Despite the aforementioned progress in the transfer learning algorithm, its application to partial linear models remains relatively underexplored. From a different perspective, Cai et al. (2025), He et al. (2024) and Jiao et al. (2024) examined the scenarios where certain structures are shared between the source and target domains, allowing well-trained source models to be directly applied to the target task. Although these works provided a promising framework for tackling semiparametric problems, the verification of the underlying structural assumptions may pose challenges. As noted in Hu and Zhang (2023), replacing conventional estimation methods with an appropriate surrogate average can be beneficial, and the similar strategy has also been explored in Li et al. (2022) and Li et al. (2023). For all aforementioned approaches, it is crucial to carefully select transferable sources that exhibit sufficient similarity to the target domain. Overlooking potential incongruities among these sources can significantly impair learning performance in the target dataset—a detrimental outcome known as negative transfer, which has been well-documented in previous studies (Yao and Doretto, 2010; Hanneke and Kpotufe, 2019).

Concurrently, safeguarding data privacy constitutes a pivotal concern in real-world applications. Dwork et al. (2014) presented a rigorous method to analyze the effectiveness of privacy protection in randomized algorithms according to the differential privacy (DP) property. This property employs two parameters, (ϵ, δ), to assess the sensitivity of the outcome with respect to a single alternative observation in the input data. Such scheme has made significant developments and has found widespread application in various large-scale systems, including deep learning (Abadi et al., 2016) and reinforcement learning (Qiao and Wang, 2023). However, most previous research works focused on designing the privacy protected distributed algorithms instead of transfer learning algorithm, where we refer the readers to Duchi et al. (2018), Wang et al. (2023), Zhu et al. (2024) and Wei et al. (2020) for more details.

In this paper, we aim to integrate PLM inference with DP procedures within a transfer learning framework. Specifically, we focus on the following three key problems. First, we develop transfer learning methods for estimation and inference based on high-dimensional sparse PLMs. Most existing transfer learning approaches primarily focus on aggregated multi-source estimation, while downstream inference remains underexplored (Li et al., 2024). For classical PLM inference (i.e., target-only), reliance on a single data source can limit estimation efficiency, potentially leading to suboptimal performance. Second, the problem of achieving nearly optimal estimation and inference for PLMs under differential privacy constraints is of substantial relevance to this study. Inspired by Cai et al. (2026) and Zhang et al. (2024), we propose several algorithms with privacy protection for PLM, which extend beyond simple linear models with more flexibility. Additionally, we are devoted to establishing a false discovery rate (FDR) controlled multiple hypothesis testing procedure within this framework. Above all, our main contributions of this work are as follows.

- We introduce a novel transfer learning framework for the partial linear model that enhances the accuracy of the target parameter estimation. This framework facilitates knowledge transfer from multiple source domains to the target domain, thereby improving estimation precision. In addition, it incorporates an inference procedure for high-dimensional parameter estimation using the debiased LASSO method. Our proposed algorithms offer an effective approach for high-dimensional partially linear model estimation via transfer learning, as well as the construction of confidence intervals.

- In this study, we introduce a novel formulation of differential privacy, termed source-target-wise differential privacy (STDP), which offers enhanced adaptability and stronger privacy guarantees specifically tailored for transfer learning frameworks. Building upon this new framework, we establish rigorous theoretical foundations for the privacy guarantees of our proposed algorithms.

- We develop an e-value-based procedure for FDR control in multiple testing problems, which simultaneously ensures rigorous differential privacy guarantees throughout the inference procedure. Our approach enables FDR control at any user-specified level while achieving asymptotic power approaching 1 under appropriate conditions. To the best of our knowledge, this work represents the first systematic integration of e-values with differential privacy principles, offering substantial advantages in terms of both computational efficiency and theoretical tractability.

NOTATION: In this work, we denote $\mathbb{R}^p$ as p-dimensional real vector space, and use $\mathbb{I}(\cdot)$ to represent the indicator function. For any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, define $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^p \mathbb{I}(\beta_j \neq 0)$, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ and $\|\boldsymbol{\beta}\|_2 = (\sum_{j=1}^p \beta_j^2)^{1/2}$. For any given matrix $\boldsymbol{A}$, we use $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ to denote its minimum eigen value and maximum eigen value, respectively. Across this paper, we use $\mathbf{I}$ to represent the identity matrix and $\mathbf{e}_j$ to denote the unit vector with j-th element being 1. For two sequences of non-negative numbers $\{x_n\}_{n\geq 1}$ and $\{y_n\}_{n\geq 1}$, $x_n \lesssim y_n$ means that there exists some constant $C > 0$ independent of n such that $x_n \leq Cy_n$; $x_n \gtrsim y_n$ is equivalent to $y_n \lesssim x_n$; $x_n \asymp y_n$ is equivalent to $x_n \lesssim y_n$ and $y_n \lesssim x_n$. We use $C, c, c_0, c_1, \ldots$ to denote universal constants whose values may change from line to line. For any integer p, we denote $[p] := \{1, 2, \ldots, p\}$. Given real number R, denote Π_R the projection operator onto the ℓ_2-ball of radius R, i.e., $\Pi_R(x) = x \min\{1, R/\|x\|_2\}$, $x \in \mathbb{R}^p$, with the convention $\Pi_R(0) = 0$.

The remainder of this paper is constructed as follows: Section 2 depicts the problem and provides an estimation algorithm. Specifically, Section 2.1 formulates the model and data structure, Section 2.2 introduces the function space where the nonparametric part belongs to, and Section 2.3 provides the estimation procedure. Section 3 includes our main theoretical results as well as an inference procedure. Concretely, Section 3.1 shows the consistency of estimation, Section 3.2 introduces the debiased LASSO method and statistical inference, and Section 3.3 extends it to multiple hypothesis test with FDR control. Section 4 and Section 5 report simulation studies for numerical data and real data, respectively. We conclude in Section 6, and our proofs and additional statements and simulation results are presented in the supplementary material.

2. Background and Method

In this section, we present a novel transfer learning framework incorporating differential privacy for partial linear model, along with an innovative algorithm for knowledge transfer. The discussion begins with a formal problem setup, followed by a comprehensive exposition of the proposed algorithmic framework.

2.1 Problem Setup

In this part, we provide a canonical formulation of transfer learning. We consider the partial linear model as our target model, which shares structural similarities in its linear component with multiple source models. The effective parameters from the source models are determined by a linear relationship with observable covariates and responses. Formally, we define the source and target models as follows:

$$\textbf{Target:}\ \ Y^{(0)} = (\boldsymbol{X}^{(0)})^\top \boldsymbol{\beta}^{(0)} + g(\boldsymbol{W}^{(0)}) + \varepsilon^{(0)},$$

$$\textbf{Source:}\ \ Y^{(k)} = (\boldsymbol{X}^{(k)})^\top \boldsymbol{\beta}^{(k)} + \varepsilon^{(k)},\quad k = 1, \ldots, K,$$

where we observe the datasets $\mathcal{D}_0 = \{(\boldsymbol{X}_i^{(0)}, \boldsymbol{W}_i^{(0)}, Y_i^{(0)})\}_{i=1}^{n_0}$ and $\mathcal{D}_k = \{(\boldsymbol{X}_i^{(k)}, Y_i^{(k)})\}_{i=1}^{n_k}$ with $\boldsymbol{X}_i^{(k)} = (X_{i1}^{(k)}, ..., X_{ip}^{(k)})^\top \in \mathbb{R}^p$, $Y_i^{(k)} \in \mathbb{R}$ and $\boldsymbol{W}_i^{(0)} \in \mathbb{R}^d$. Moreover, let $\mathcal{D} = \cup_{k\in\{0\}\cup[K]}\mathcal{D}_k$. For the target domain, the model consists of a linear part with coefficient $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \beta_2^{(0)}, \ldots, \beta_p^{(0)})$ and the nonparametric part is $g : \mathbb{R}^d \to \mathbb{R}$. For the source domain, we define the collection of transferable domain as

$$\mathcal{A} = \{k \in [K] : \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}\|_2 \leq h\},$$

where $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^p$ for all $k \in [K]$. Here, $h > 0$ quantifies the level of transferability, ensuring that the similarity between the source and target domains is sufficiently high for effective knowledge transfer. Moreover, using this quantity together with the sparsity assumption, we can obtain equivalent characterizations under other norms. Under the high-dimensional setting, we typically assume $p \gg n_0$. It is also allowed that $n_0 < d$ as Zhu (2017) pointed out. The noise terms $\varepsilon^{(0)}$ and $\varepsilon^{(k)}$ are assumed to be independently and identically distributed from a Gaussian distribution, i.e., $\varepsilon^{(0)}$, $\varepsilon^{(k)} \sim \mathrm{N}(0, \sigma^2)$. The covariates $\boldsymbol{X}_i^{(k)}$ are sub-Gaussian with parameter $\tilde{\sigma}_k$, and denote $\tilde{\sigma} = \max_{k\in\{0\}\cup[K]}\{\tilde{\sigma}_k\}$. Furthermore, the target parameter $\boldsymbol{\beta}^{(0)}$ is assumed to be sparse, satisfying $\|\boldsymbol{\beta}^{(0)}\|_0 = s_0$.

The nonparametric function g and the conditional expectation $f_j := \mathbb{E}(X_{ij}^{(0)}|\boldsymbol{W}_i^{(0)})$ are assumed to belong to some reproducing kernel Hilbert space (RKHS), which will be specified later. We denote $N = n_{\mathcal{A}} + n_0$ and $n_{\mathcal{A}} = \sum_{k\in\mathcal{A}} n_k$. Typically, we consider the case that $n_{\mathcal{A}}$ significantly exceeds the sample size n_0 of the target domain. Our primary objective is to leverage data from the source domains to enhance the estimation accuracy within the target domain.

Despite leveraging transfer learning to improve algorithmic performance, we remain committed to developing statistical procedures that ensure strong privacy guarantees. As introduced by Dwork et al. (2014), differential privacy offers a mathematically rigorous framework for quantifying privacy. The standard definition is presented as follows.

Definition 1. (Dwork et al. (2014)) A randomized algorithm $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for all measurable sets $\mathcal{S}$ and all datasets $\mathcal{D}$ and all adjacent datasets $(\mathcal{D}, \mathcal{D}')$ differing in one observation, the following holds:

$$\mathbb{P}(\mathcal{M}(\mathcal{D}) \in \mathcal{S}) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(\mathcal{D}') \in \mathcal{S}) + \delta.$$

In particular, recent advances in differential privacy have extended this paradigm to the distributed computing environments (see, e.g., Liu et al. (2022); Wang and Nedić (2023); Cai et al. (2024, 2026)), addressing critical requirements in practical applications. However, research on differential privacy in transfer learning remains scarce. Motivated by Auddy et al. (2025), here we introduce our notion of **source-**

target-wise differential privacy (STDP).

Definition 2. Given dataset $\mathcal{D}_{all} = \cup_{k\in\{0\}\cup[K]}\mathcal{D}_k$, a randomized algorithm $\mathcal{M}$ satisfies (ϵ, δ)-source-target-wise differential privacy if for all measurable sets $\mathcal{S}$ and all adjacent datasets $(\mathcal{D}_k, \mathcal{D}'_k), \forall k \in \{0\} \cup [K]$ differing in one observation, the following holds:

$$\mathbb{P}(\mathcal{M}(\mathcal{D}_{all}) \in \mathcal{S}|\mathcal{D}_k, \mathcal{D}_{-k}) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(\mathcal{D}_{all}) \in \mathcal{S}|\mathcal{D}'_k, \mathcal{D}_{-k}) + \delta,$$

where $\mathcal{D}_{-k} := \mathcal{D}_{all} \backslash \mathcal{D}_k$.

Our proposed notion, STDP, provides privacy guarantees individually for each data source k, encompassing both the source and target domains in transfer learning. When $K = 0$ (indicating the target-only scenario), the STDP framework degenerates into the primary differential privacy setting established by Dwork et al. (2014).

2.2 RKHS-based Nonparametric Estimation

We begin the nonparametric estimation procedure by specifying the functional space that the nonparametric part belongs to. For a function space $\mathcal{F}$, let every function $f \in \mathcal{F}$ be a mapping $f : \mathcal{W} \to \mathbb{R}$, where $\mathcal{W}$ is the domain of control variables. Assume that $\mathcal{F}$ is equipped with a norm $\|\cdot\|_{\mathcal{F}}$ induced by an inner product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$. The RKHS is a Hilbert space of functions that is associated with a symmetric and positive definite kernel function $K(\cdot,\cdot)$ that satisfies the reproducing property:

$$\langle K(\cdot,\boldsymbol{W}), f(\cdot)\rangle_{\mathcal{F}} = f(\boldsymbol{W}), \quad \forall \boldsymbol{W} \in \mathcal{W},\ f \in \mathcal{F}.$$

To characterize the function space, we first introduce the conditional local complexity. Given a set of points $\{\boldsymbol{W}_i\}_{i=1}^n$, we define the empirical ball with radius $r_n > 0$ as $\mathbb{B}(r_n;\mathcal{F}) := \{f \in \partial\mathcal{F} : \sum_{i=1}^n [f(\boldsymbol{W}_i)]^2/n \le r_n^2, \|f\|_{\mathcal{F}} \le 1\}$, where $\partial\mathcal{F} = \{f = f' - f'' : f', f'' \in \mathcal{F}\}$. Then the conditional local complexity is defined as

$$\mathcal{G}_n(r_n;\mathcal{F}) := \mathbb{E}_{\xi}\left(\sup_{f\in\mathbb{B}(r_n;\mathcal{F})} |\frac{1}{n}\sum_{i=1}^n \xi_i f(\boldsymbol{W}_i)| \mid \{\boldsymbol{W}_i\}_{i=1}^n\right),$$

where $\xi_i, i = 1,\ldots,n$ are i.i.d sub-Gaussian variables with zero conditional mean $\mathbb{E}(\xi_i \mid \boldsymbol{W}_i) = 0$ and parameter $\sigma^\dagger$ (see Wainwright (2019) for more details). For any star-shaped class $\mathcal{F}$, Lemma A8 in Zhu et al. (2019) guarantees that the function $r_n \mapsto \mathcal{G}_n(r_n;\mathcal{F})/r_n$ is non-increasing on the interval $(0,\infty)$. Therefore, there exists some large enough $r_n > 0$ that satisfies the critical inequality as $\mathcal{G}_n(r_n;\mathcal{F}) \le r_n^2/(2\sigma^\dagger)$. According to Wainwright (2019), the critical radius is hence defined as

$$\tilde{r}_n = \inf_{r_n\in(0,\infty)}\left\{\mathcal{G}_n(r_n;\mathcal{F}) \le \frac{r_n^2}{2\sigma^\dagger}\right\}.$$

The nonparametric estimation error can therefore be upper bounded by the critical radius, according to Lemma S.1 in Zhu (2017).

Based on the above pieces, the initial nonparametric estimation on the target domain can be taken as the solution of the following optimization problem:

$$\hat{m}^{(init)} = \arg\min_{m \in \mathcal{F}} \left[\frac{1}{2n_0} \sum_{i=1}^{n_0} \{Y_i^{(0)} - m(\boldsymbol{W}_i^{(0)})\}^2 \right] + \frac{\lambda_0}{2} \|m\|_{\mathcal{F}}^2. \tag{2.2}$$

A nonparametric regression problem like (2.2) is a standard setup in many modern statistics books (Wainwright, 2019). The estimator $\hat{m}^{(init)}(\boldsymbol{w})$ is known to yield an empirical estimation error with respect to $\mathbb{E}(Y_i^{(0)} \mid \boldsymbol{W}_i^{(0)} = \boldsymbol{w})$ bounded above by a constant times the critical radius $\tilde{r}_n$.

2.3 Trans-DPPLM Method

We now introduce a privacy preserving transfer learning algorithm for the partial linear model, which estimates the high-dimensional target parameter $\boldsymbol{\beta}^{(0)}$ by leveraging information from auxiliary source models. To solve the partial linear regression model, a straightforward idea is to remove the effect caused by the $\boldsymbol{W}$-dependent component, in the target model, and then estimate the linear part. With the aim of improving the estimation efficiency, the transfer step extracts knowledge from auxiliary data to implement the linear estimation. The Gaussian mechanism (Dwork et al., 2014) is applied to the transfer step to ensure STDP.

The following provides a summarized illustration of the proposed method. First, a stabilized kernel ridge regression is solved on the target domain to estimate the main

Algorithm 1 Transfer High-dimensional Partial Linear Regression with Privacy Guarantees

INPUT: Data $\{(\boldsymbol{X}_i^{(0)}, \boldsymbol{W}_i^{(0)}, Y_i^{(0)})\}, \{(\boldsymbol{X}_i^{(k)}, Y_i^{(k)}), k \in [K]\}$, number of iterations T, step size ρ_1, privacy parameters (ϵ, δ), initialization $\boldsymbol{\beta}^0$, failure probability $\eta \in (0, 1/2)$, sparse level s', transferable set $\mathcal{A}$, constant L, truncation radius R_Y, R_0, R_k, R_d.

Nonparametric estimation:

for $k = 0$ **do**

  Compute $\hat{m}^{(init)} = \underset{m \in \mathcal{F}}{\arg\min} \left[\frac{1}{2n_0} \sum_{i=1}^{n_0} \{Y_i^{(0)} - m(\boldsymbol{W}_i^{(0)})\}^2\right] + \frac{\lambda_0}{2} \|m\|_{\mathcal{F}}^2$.

  Residualization: denote $\hat{\boldsymbol{m}}^{(init)}(\boldsymbol{W}^{(0)}) = (\hat{m}^{(init)}(\boldsymbol{W}_1^{(0)}), \ldots, \hat{m}^{(init)}(\boldsymbol{W}_{n_0}^{(0)}))^\top$, $\breve{\boldsymbol{Y}}^{(0)} = \Pi_{R_Y}(\boldsymbol{Y}^{(0)} - \hat{\boldsymbol{m}}^{(init)}(\boldsymbol{W}^{(0)})) + \boldsymbol{\Delta}_Y, \boldsymbol{\Delta}_Y \sim \mathrm{N}(\mathbf{0}, \frac{8\log(5/\delta) R_Y^2}{(\epsilon/2)^2} \mathbf{I}_{n_0})$.

end for

for $k \in \mathcal{A}$ **do**

  $\breve{Y}_i^{(k)} = Y_i^{(k)}$.

end for

Transferred parametric estimation with DP:

Randomly split each dataset into T parts of equal sizes. Denote the corresponding index sets as $\mathcal{I}^{(k)[t]}, t = 0, \ldots, T-1$.

for $t = 0, \ldots, T-1$ **do**

  for $k \in \{0\} \cup \mathcal{A}$ **do**

    Update the gradient with noise $\boldsymbol{\Delta}_t^{(k)} \sim \mathrm{N}\left(\mathbf{0}, \frac{8\log(5/\delta) T^2 R_d^2 R_k^2}{n_k^2 (\epsilon/2)^2} \mathbf{I}_p\right)$:

$$\tilde{\boldsymbol{G}}_k^t = \frac{n_k}{N} \left(\frac{1}{\lfloor n_k/T \rfloor} \sum_{i \in \mathcal{I}^{(k)[t]}} \Pi_{R_k}\left((\boldsymbol{X}_i^{(k)})^\top \boldsymbol{\beta}^t - \breve{Y}_i^{(k)} \right) \Pi_{R_d}(\boldsymbol{X}_i^{(k)}) + \boldsymbol{\Delta}_t^{(k)} \right).$$

  end for

  Update $\boldsymbol{\beta}^{t+0.5} = \boldsymbol{\beta}^t - \rho_1 \sum_{k \in \{0\} \cup \mathcal{A}} \tilde{\boldsymbol{G}}_k^t$.

  Update $\boldsymbol{\beta}^{t+1} = \mathrm{HT}\left(\boldsymbol{\beta}^{t+0.5}, s'\right)$.

end for

OUTPUT: $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^T$.

$\boldsymbol{W}$-dependent component of the response. Based on this initial projection, we construct a private surrogate response for the target by subtracting the fitted nonparametric part, applying truncation, and then adding Gaussian noise. As a result, the subsequent transfer step is carried out using the privatized surrogate target response, rather than the raw target outcome, which ensures privacy protection already at the residualization stage. Although the fitted vector $\hat{\boldsymbol{m}}^{(init)}(\boldsymbol{W}^{(0)})$ depends on all target observations, the released residualization query is the projected vector $\Pi_{R_Y}(\boldsymbol{Y}^{(0)} - \hat{\boldsymbol{m}}^{(init)}(\boldsymbol{W}^{(0)}))$. Hence any two neighboring target datasets produce two vectors in the ℓ_2-ball of radius R_Y, so the global ℓ_2-sensitivity of this query is at most $2R_Y$. To implement the transfer learning procedure, at each iteration, each dataset from different domains is split into T parts. Let the sample index set of the t-th partition from the k-th domain be denoted by $\mathcal{I}^{(k)[t]}$, where $\mathcal{I}^{(k)[t]} \subset [n_k]$. These index sets form a partition of $[n_k]$, i.e., $\bigcup_{t=0}^{T-1} \mathcal{I}^{(k)[t]} = [n_k]$ and $\mathcal{I}^{(k)[t]} \cap \mathcal{I}^{(k)[t']} = \emptyset$ for any $t \neq t'$. Without loss of generality, assume that these subsets are of equal size. Then the gradients are aggregated over the target and informative source domains, with Gaussian perturbations added to each gradient update to preserve privacy. The aggregated gradient steers the estimator toward the target parameter, while the hard-thresholding algorithm (HT, see Algorithm S1 in the supplementary material) is applied to enforce sparsity. After T iterations, the estimator of the linear coefficient $\boldsymbol{\beta}^{(0)}$ is given by $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^T$ in Algorithm 1. By the Gaussian mechanism together with the composition principle of differential privacy (Cai et al., 2021), the resulting estimator satisfies (ϵ, δ)-STDP.

3. Theoretical Results

In this section, we will establish the statistical properties of the proposed method.

3.1 Consistency of Trans-DPPLM

At the outset of this part, we begin by imposing some regularity assumptions.

Assumption 1. For all $k \in \{0\} \cup [K]$, $j = 1, \ldots, p$, $\{X_{ij}^{(k)}\}_{i=1}^{n_k}$ are independent sub-Gaussian variables with parameter at most $\tilde{\sigma}$. The noise terms $\varepsilon_i^{(k)}$ are i.i.d. from a Gaussian distribution, i.e., $\varepsilon_i^{(k)} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$. Assume $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^p$, $\mathbb{E}(\boldsymbol{X}_i^{(k)}) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{X}_i^{(k)}(\boldsymbol{X}_i^{(k)})^\top) = \boldsymbol{\Sigma}^{(k)}$, $0 < L_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}^{(k)}) \leq \lambda_{\max}(\boldsymbol{\Sigma}^{(k)}) \leq L_0 < \infty$ with some absolute constant $L_0 \geq 1$.

Assumption 2. For all $k \in \mathcal{A}$, $\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}\|_1 \lesssim \sqrt{s_0}\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}\|_2$, where s_0 denotes the sparsity level of $\|\boldsymbol{\beta}^{(0)}\|_0$, i.e., $\|\boldsymbol{\beta}^{(0)}\|_0 \leq s_0$.

Assumption 3. (i) The conditional expectations satisfy: $\mathbb{E}(Y_i^{(0)} \mid \boldsymbol{W}_i^{(0)}) \in \mathcal{F}$, $\mathbb{E}(X_{ij}^{(0)} \mid \boldsymbol{W}_i^{(0)}) \in \mathcal{F}$, where $\mathcal{F}$ is the RKHS with critical radius $\tilde{r}_n$;

(ii) $\mathcal{F}$ is star-shaped, that is, for any $f \in \partial\mathcal{F}$ and $\kappa \in [0,1], \kappa f \in \partial\mathcal{F}$;

(iii) The critical radius satisfies $\tilde{r}_n^2 = O((\log p/n_0)^{1/2})$.

Assumptions 1 and 2 are regular conditions that can be seen in the context of high-dimensional statistics (Li et al., 2022, 2024). Assumption 3(i) requires the relevant conditional mean functions to belong to the RKHS class $\mathcal{F}$, and Assumption 3(ii) is a mild geometric condition ensuring the well-posedness of the optimization problems in (2.2) and (3.4); similar conditions also appear in Zhu et al. (2019). Assumption 3(iii) is a compatibility condition on the complexity of the nonparametric component. Since the critical radius $\tilde{r}_n$ is defined through the localized complexity of the RKHS class, we keep it in this abstract form in the main statement. The requirement $\tilde{r}_n^2 = O((\log p/n_0)^{1/2})$ ensures that the error from the nonparametric residualization step is sufficiently small relative to the high-dimensional scale in the subsequent analysis. For concrete function classes, this condition can be verified by standard localized-complexity arguments. In particular, for low-dimensional smooth function classes (for example, when $\boldsymbol{W}_i$ is low-dimensional and g belongs to a Sobolev-type RKHS), the required order of $\tilde{r}_n$ follows from the localized-complexity arguments; see Zhu et al. (2019) for representative technical arguments.

Theorem 1. Under Assumptions 1-3, for any $\eta \in (0,1)$, if the tuning parameter $\lambda_0 \asymp (\log p/n_0)^{1/2}$, iteration times $T \asymp \log N$, step size $\rho_1 = 0.9(1 - 0.296/L_0^4)/L_0$, sparse level $4.18 L_0^4 s_0 \le s' \lesssim s_0$, initial estimate $\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^{(0)}\|_2 \lesssim C$, truncation radius $R_Y = 50(\log(p/\eta))^{1/2}$, $R_0 = s' R_d + R_Y(\log(n_0/\eta)\log(1/\delta))^{1/2}/\epsilon$, $R_k = 2(\log(18 n_k/\eta)(\sigma^2 + h))^{1/2}$, and $R_d = (p\log(18N/\eta))^{1/2}\tilde{\sigma}$, then we have:

(i) Algorithm 1 is (ϵ,δ)-differential privacy guaranteed for both source and target, i.e., (ϵ,δ)-STDP. (ii) The output $\hat{\boldsymbol{\beta}}$ from Algorithm 1 satisfies

$$\mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\|_2 \lesssim (\mathrm{B})) \ge 1 - \eta,$$

where (B) is $\sqrt{\frac{s_0 d\log(p/\eta)\log(N)}{N}}+\frac{\sqrt{|\mathcal{A}|ps_0\log(1/\delta)\log^5(Np/\eta)}}{N\epsilon}+h$.

Corollary 1. *Under the same conditions as in Theorem 1, if* $h \lesssim (s_0 d\log p/n_0)^{1/2}$, $\log(N)/N \lesssim 1/n_0$, *and* $\epsilon \gtrsim (n_0 p\log(1/\delta)/(Nd))^{1/2}$, *the initial estimator* $\hat{\boldsymbol{\beta}}$ *satisfies that* $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(0)}\|_2 = O_p((s_0 d\log p/n_0)^{1/2})$ *and* $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(0)}\|_1 = O_p(s_0(d\log p/n_0)^{1/2})$.

The upper bound shown in (B) consists of three parts: The first term represents the statistical error, the second term is the privacy cost and the third term is the transfer cost. For the privacy cost, it is shown to be a nearly optimal rate (Cai et al., 2021), differing only by a logarithmic factor. For the transfer cost, the result is comparable to that of Theorem 1 in Li et al. (2022), in which the minimax optimality for linear regression is established. As the corollary indicates, if the transferable source and target are similar enough (i.e., $h \lesssim (s_0 d\log p/n_0)^{1/2}$), the proposed estimator will enjoy a similar convergence rate compared to LASSO.

Remark 1. The step size $\rho_1 = 0.9(1-0.296/L_0^4)/L_0$ also appears in Proposition 19 of Li et al. (2024), which is particularly chosen to ensure $((10\xi/(9(1-\xi))+17/5) > 242L_0^4/9$ as shown in the proof. The truncation levels, which are relative to the sub-Gaussian parameters for $\boldsymbol{X}$, Y and ε, are also tuneable as Cai et al. (2021), Cai et al. (2026) and Zhang et al. (2024) suggested. The assumptions on sparsity level can also be seen in Li et al. (2024) and Cai et al. (2021) with the same scale of sL_0^4. Our sample size requirement $N \gtrsim s_0$ results from the specific combination of the Gaussian privatized gradient scheme, the truncation levels, and the iterative hard-thresholding structure used here. A similar $N \gtrsim s_0$ scaling can also be observed in Li et al. (2024) under a comparable privatized scheme.

Remark 2. The quantity h is a user-specified similarity threshold that characterizes the required level of transferability, rather than an unknown population parameter that needs to be estimated. In practice, for a given h, one may first apply a transferable source detection procedure, such as the method in Tian and Feng (2023), to construct an estimated transferable set $\hat{\mathcal{A}}$, and then run Algorithm 1 using the selected sources in $\hat{\mathcal{A}}$. In this sense, h serves as a thresholding parameter for source selection, while the actual transfer step is carried out on the detected transferable set. The tuning parameters involving h are calibrated according to the chosen value, and a larger h generally leads to a less favorable theoretical bound.

3.2 Debiased Estimator and Statistical Inference

Building upon Algorithm 1, this part presents an algorithm to conduct statistical inference on $\boldsymbol{\beta}$ via a debiasing procedure, accompanied by a theoretical result that establishes the asymptotic normality. In addition, we also show that the construction of confidence intervals satisfies the STDP property.

First, we introduce several key variables essential for carrying out the inference procedure. Given the estimated coefficient vector $\hat{\boldsymbol{\beta}}$ with privacy parameters $(\epsilon/4, \delta/4)$ obtained from Algorithm 1, we proceed to refine the nonparametric component by solving the following optimization problem:

$$\hat{g} = \arg\min_{g \in \mathcal{F}} \left[\frac{1}{2n_0} \sum_{i=1}^{n_0} \{ Y_i^{(0)} - g(\boldsymbol{W}_i^{(0)}) - (\boldsymbol{X}_i^{(0)})^\top \hat{\boldsymbol{\beta}} \}^2 + \frac{\lambda_g}{2} \|g\|_{\mathcal{F}}^2 \right]. \tag{3.3}$$

After obtaining $\hat{g}$, we remove the nonparametric component of $Y^{(0)}$ and define $\check{Y}_i^{(0)} = Y_i^{(0)} - \hat{g}(\boldsymbol{W}_i^{(0)})$, which estimates the residualized response $Y_i^{(0)} - g(\boldsymbol{W}_i^{(0)})$. Similarly, the projection of $\boldsymbol{X}^{(0)}$ is denoted as $\tilde{\boldsymbol{X}}_i^{(0)} = \boldsymbol{X}_i^{(0)} - \mathbb{E}(\boldsymbol{X}_i^{(0)} \mid \boldsymbol{W}_i^{(0)})$ with precision matrix $\boldsymbol{\Theta} = \mathbb{E}(\tilde{\boldsymbol{X}}_i^{(0)} (\tilde{\boldsymbol{X}}_i^{(0)})^\top)^{-1}$. We estimate the conditional expectation $\mathbb{E}(X_{ij}^{(0)} \mid \boldsymbol{W}_i^{(0)})$ by

$$\hat{f}_j = \arg\min_{f \in \mathcal{F}} [\frac{1}{2n_0} \sum_{i=1}^{n_0} \{ X_{ij}^{(0)} - f(\boldsymbol{W}_i^{(0)}) \}^2], \ j = 1, \ldots, p. \tag{3.4}$$

Then, based on the estimation, we define the adjusted covariates $\check{\boldsymbol{X}}_i^{(0)} = \boldsymbol{X}_i^{(0)} - \hat{\boldsymbol{f}}(\boldsymbol{W}_i^{(0)})$. Here $\hat{\boldsymbol{f}}(\boldsymbol{W}_i^{(0)}) = (\hat{f}_1(\boldsymbol{W}_i^{(0)}), \hat{f}_2(\boldsymbol{W}_i^{(0)}), ..., \hat{f}_p(\boldsymbol{W}_i^{(0)}))^\top$. For convenience, we concatenate each vector by column and denote $\hat{\boldsymbol{f}}(\boldsymbol{W}^{(0)}) = (\hat{\boldsymbol{f}}(\boldsymbol{W}_1^{(0)})^\top, \ldots, \hat{\boldsymbol{f}}(\boldsymbol{W}_{n_0}^{(0)})^\top)^\top \in \mathbb{R}^{n_0 \times p}$. To apply the debiased method for β_j, the j-th column of precision matrix $\boldsymbol{\Theta}_j$ should be estimated first. The computation procedure for $\hat{\boldsymbol{\Theta}}_j$ will be specified in detail later. Once $\hat{\boldsymbol{\Theta}}_j$ is obtained, given $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^\top$, then the debiased estimator is computed as per

$$\hat{\beta}_j^{(de)} = \hat{\beta}_j + \frac{1}{n_0} \sum_{i=1}^{n_0} \Pi_{\tilde{R}}((\check{\boldsymbol{X}}_i^{(0)})^\top \hat{\boldsymbol{\Theta}}_j)(\Pi_{\tilde{R}}(\check{\boldsymbol{Y}}_i^{(0)}) - \Pi_{\tilde{R}}((\boldsymbol{X}_i^{(0)})^\top \hat{\boldsymbol{\beta}})). \tag{3.5}$$

Furthermore, we add a noise to the debiased estimator to ensure privacy, i.e.,

$$\hat{\beta}_j^{(dp)} = \hat{\beta}_j^{(de)} + \tilde{\Delta}_j, \tag{3.6}$$

where $\tilde{\Delta}_j \sim \mathrm{N}(0, 16\{\frac{4\tilde{R}^2}{n_0}\}^2 \frac{2\log(10/\delta)}{(\epsilon/4)^2})$. Theoretically, $\hat{\beta}_j^{(dp)}$ is asymptotically normal, as will be established in Theorem 2 in the following on the basis of Assumption 4.

Assumption 4. (i) The precision matrix $\mathbf{\Theta} := (\Theta_{i,j})_{p\times p}$ has bounded eigenvalues: $0 < \tilde{L}_0^{-1} \leq \lambda_{\min}(\mathbf{\Theta}) \leq \lambda_{\max}(\mathbf{\Theta}) \leq \tilde{L}_0 < \infty$.

(ii) The j-th column of $\mathbf{\Theta}$ is sparse: $\|\mathbf{\Theta}_j\|_0 \leq \tilde{s}_j$ for all $j = 1, \ldots, p$.

(iii) The following technical conditions hold:

$$\sqrt{n_0}\, \tilde{s}_j (s_0 \vee 1) \left(\tilde{r}_n^2 \vee \frac{\log p}{n_0} \right) = o(1), \tag{3.7}$$

$$\sqrt{n_0} \left\{ \tilde{s}_j^2 (s_0 \vee 1) \frac{\log p}{n_0} \vee \tilde{s}_j^3 (s_0 \vee 1) \left(\tilde{r}_n^2 \vee \frac{\log p}{n_0} \right) \sqrt{\frac{\log p}{n_0}} \right\} = o(1), \tag{3.8}$$

$$\left\{ (\tilde{s}_j^2 \vee 1) \sqrt{\frac{\log p}{n_0}} \vee (\tilde{s}_j^3 \vee 1) \left(\tilde{r}_n^2 \vee \frac{\log p}{n_0} \right) \right\} = o(1). \tag{3.9}$$

The first two items of this assumption are necessary for most debiased methods that require solving the precision matrix, such as Tian and Feng (2023) and Li et al. (2023). And the third part of this assumption is a technical condition as a complement to Assumption 3 (iii), which can also be found in Zhu (2017) and Zhu et al. (2019).

Theorem 2. Under Assumptions 1-4 and conditions in Corollary 1, let $\bar{s} = \max\{s_0, \tilde{s}_1, \ldots, \tilde{s}_p\}$,

$\rho_2 = L_0/6$, $T \asymp \rho_2 \log n_0$, $\tilde{R} \asymp 2L_0(d\log(18n_0/\eta)\sigma^2+\log(p/\eta))^{1/2}$, $\tilde{s}' \asymp \bar{s}$, $\bar{s}\log p \log n_0 = o(n_0)$, $\bar{s}\log p(\log(1/\delta))^{1/2}(\log n_0)^{7/2}/\epsilon = o(n_0)$, $(\bar{s}\log p\log(1/\delta)/n_0)^{1/2}\log^{3/2} n_0 \lesssim \epsilon$ and $\|\boldsymbol{\beta}^{(0)}\|_2 \leq C$. Suppose $d \ll p$, then we have

$$\frac{\sqrt{n_0}(\hat{\beta}_j^{(dp)} - \beta_j^{(0)})}{V_j} \xrightarrow{D} \mathrm{N}(0,1), \tag{3.10}$$

where $V_j^2 = \Theta_{j,j}\sigma^2$ is the asymptotic variance of $\hat{\beta}_j^{(dp)}$ with estimation computed by Algorithm 2. Moreover, $\hat{\beta}_j^{(dp)}$ satisfies (ϵ,δ)-STDP.

Algorithm 2 Differentially Private Asymptotic Variance Estimation of β_j

1: **INPUT:** Adjusted dataset $\{(\boldsymbol{X}_i^{(0)}, \check{\boldsymbol{X}}_i^{(0)}, \check{Y}_i^{(0)}, \boldsymbol{W}_i^{(0)})\}_{i\in[n_0]}$, initial estimation $\hat{\boldsymbol{\beta}}$ with privacy parameters $(\epsilon/4, \delta/4)$, step size ρ_2, noise scale B, number of iterations T, sparsity level $\tilde{s}'$, truncation radius $\tilde{R}$, initial value $\boldsymbol{\Theta}_j^0$, feasibility parameter C .

2: Randomly split each dataset into T parts of equal sizes. Denote the corresponding index sets as $\mathcal{I}^{(0)[t]}, t = 0, \ldots, T-1$.
3: **for** t in 0 to $T-1$ **do**
4: Compute $\boldsymbol{\Theta}_j^{t+0.5} = \boldsymbol{\Theta}_j^t - \rho_2(\mathbf{e}_j - \sum_{i\in\mathcal{I}^{(0)[t]}} \check{\boldsymbol{X}}_i^{(0)} \Pi_{\tilde{R}}((\check{\boldsymbol{X}}_i^{(0)})^\top \boldsymbol{\Theta}_j^t)/\lfloor n_0/T \rfloor)$.
5: Update $\boldsymbol{\Theta}_j^{t+1} = \Pi_C(\mathrm{NoisyHT}(\boldsymbol{\Theta}_j^{t+0.5}, \tilde{s}', \frac{\rho_2 TB}{n_0}, \frac{\epsilon}{4T}, \frac{\delta}{4T}))$.
6: **end for**
7: Let $\hat{\Theta}_{j,j} = \Theta_{j,j}^T$, calculate the variance as per $\hat{V}_j^{(dp)} = (\hat{\Theta}_{j,j}\hat{\sigma}^2)^{1/2}$, where $\hat{\sigma}^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(\Pi_{\tilde{R}}(\check{Y}_i^{(0)}) - \Pi_{\tilde{R}}((\boldsymbol{X}_i^{(0)})^\top\hat{\boldsymbol{\beta}}))^2 + \Delta_\sigma,\ \Delta_\sigma \sim \mathrm{N}(0, 16\{\frac{2(2\tilde{R})^2}{n_0}\}^2 \frac{2\log(10/\delta)}{(\epsilon/4)^2})$.
8: **OUTPUT:** $\hat{V}_j^{(dp)}$

Based on Theorem 2, we can now establish a differentially private confidence interval for the regression coefficient β_j. The whole procedure is summarized as follows. Given dataset $\{(\boldsymbol{X}_i^{(0)}, \check{\boldsymbol{X}}_i^{(0)}, \check{Y}_i^{(0)}, \boldsymbol{W}_i^{(0)})\}_{i=1}^{n_0}$, where $\check{Y}_i^{(0)} = Y_i^{(0)} - \hat{g}(\boldsymbol{W}_i^{(0)})$ and $\check{X}_{ij}^{(0)} = X_{ij}^{(0)} - \hat{f}_j(\boldsymbol{W}_i^{(0)})$, the effects of control variables are adjusted through (3.3) and (3.4). Similarly, the target dataset is divided into T disjoint subsets of equal size. At t-th round, the variable $\boldsymbol{\Theta}_j^t$ is iteratively updated based on each subset. An intermediate update $\boldsymbol{\Theta}_j^{t+0.5}$ is calculated as the result of one-step gradient descent. Using a projection function Π_C, followed by a noisy hard-thresholding algorithm (NoisyHT, see Algorithm S2 in the supplementary), the privacy and sparsity of $\boldsymbol{\Theta}_j^{t+1}$ are ensured. The asymptotic variance of $\hat{\beta}_j^{(dp)}$, denoted as V_j, is estimated by $\hat{V}_j^{(dp)} = (\hat{\Theta}_{j,j}\hat{\sigma}^2)^{1/2}$, where the noise-adjusted variance $\hat{\sigma}^2$ incorporates a Gaussian noise term $\Delta_\sigma \sim \mathrm{N}(0, 16\{\frac{2(2\tilde{R})^2}{n_0}\}^2 \frac{2\log(10/\delta)}{(\epsilon/4)^2})$. Finally, the differentially private confidence interval is constructed as

$$\widehat{\mathrm{CI}}(j) = [\hat{\beta}_j^{(dp)} - \frac{\hat{V}_j^{(dp)} \mathrm{z}_{1-\alpha/2}}{\sqrt{n_0}}, \hat{\beta}_j^{(dp)} + \frac{\hat{V}_j^{(dp)} \mathrm{z}_{1-\alpha/2}}{\sqrt{n_0}}],$$

where $\mathrm{z}_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Gaussian distribution. Algorithm 2 returns $\hat{V}_j^{(dp)}$ as the final output, which is (ϵ,δ)-STDP. Then, by Theorem 2 and the post-processing property of differential privacy, the proposed confidence interval guarantees both statistical validity and differential privacy.

Corollary 2. *Under the same conditions as in Theorem 2, DP-debiased estimation $\boldsymbol{\beta}^{(dp)}$ with privacy parameters (ϵ,δ), we have:*

(i) $\lim_{n_0\to\infty} \mathbb{P}(\beta_j^{(0)} \in \widehat{\mathrm{CI}}(j)) = 1-\alpha$. *(ii) The process of deriving $\widehat{\mathrm{CI}}(j)$ is (ϵ,δ)-STDP.*

Compared to the non-private counterpart of the debiased term in Zhang and Zhang (2014) and Ning and Liu (2017), it could be found that our confidence interval has a similar form with additional noise injected to ensure privacy. When the noise level is low, the confidence interval closely approximates the non-private counterpart, allowing us to nearly achieve privacy without incurring additional costs. In contrast, when the privacy level is high, the confidence interval has a larger length to reach the same confidence level. Although the proposed confidence interval does not enjoy a first-order reduction in width from transfer learning under the current inference construction, its finite-sample centering and overall empirical performance may still be influenced by the quantity and quality of the transferable source domains through their effect on the preliminary estimator and the debiasing remainder.

Remark 3. Most of the parameters in Theorem 2 are selected to meet the requirement in Lemma 4.1 Cai et al. (2026). The truncation level is similarly chosen by the same argument in Remark 1. The bounding condition, $\|\boldsymbol{\beta}^{(0)}\|_2 \leq C$, also appears in Condition 3.2 in Cai et al. (2026).

3.3 Multiple Testing with e-BH Procedure

In this part, we propose a novel differentially private framework for large-scale multiple hypothesis testing in high-dimensional settings, which provides rigorous false discovery rate control guarantees. We are interested in the testing problem

$$H_0 : \beta_j^{(0)} = 0 \quad v.s. \quad H_j : \beta_j^{(0)} \neq 0, \quad j = 1, ..., p. \tag{3.11}$$

Consider $\mathcal{J}_1 = \{j \in [p] : \beta_j^{(0)} \neq 0\}$, and denote $\mathcal{J}_0 = \mathcal{J}_1^c$ as its complement. We would like to discover a rejection index list $\hat{\mathcal{J}}_1$, where we define $\mathrm{FDP} = \frac{|\mathcal{J}_1^c \cap \hat{\mathcal{J}}_1|}{|\hat{\mathcal{J}}_1| \vee 1}$, and $\mathrm{FDR} = \mathbb{E}(\mathrm{FDP})$.

In recent years, there has been a growing body of research on differentially private FDR control. Dwork et al. (2021) and Xia and Cai (2023) achieved privacy-preserving FDR control by adding noise to logarithmic transformed p-values, while Pournaderi and Xiang (2021) and Cai et al. (2026) employed data splitting and knockoff techniques to ensure differential privacy. However, directly incorporating p-values with differentially private procedures remains challenging, and the data-splitting approach requires partitioning of the data, which reduces estimation efficiency. Meanwhile, the knockoff-based approach relies on a specific dimensionality condition ($p < n_0$), which limits its applicability to certain problems.

Unlike previous work, we propose an e-value-based procedure to realize differentially private FDR control. The e-variable, introduced in Wang and Ramdas (2022), is a random variable that maps data to $[0, \infty]$. In contrast to a p-variable—which traditionally corresponds to the test ϕ_α and maps data to $[0, 1]$ with small *Probability* under the null hypothesis (i.e., the p-value)—an e-variable instead has a small *Expectation* under the null (specifically, an expectation less than 1). For any given dataset, the realization of e-variable is called the e-value. However, constructing exact e-values is not always straightforward; in such cases, several types of generalized e-values are also meaningful. For example, Ren and Barber (2024) pointed out that relaxed e-values are sufficient to guarantee FDR control under the e-Benjamini-Hochberg (e-BH) procedure. Here we introduce another generalized e-values developed by Yu et al. (2024), which is known as the asymptotic e-value.

Definition 3. (Yu et al. (2024)) For multiple hypothesis $H_1, \ldots, H_p$, we have non-negative test statistics $e_1, \ldots, e_p$. The sample size is denoted as n_0 and the index set of null hypotheses is $\mathcal{J}_0$. For $j \in \mathcal{J}_0$, if $\limsup_{n_0, p \to \infty} \mathbb{E}(e_j) \leq 1$, we then refer to e_j as an asymptotic e-value.

Building on the debiased estimator derived in the previous subsection, we now construct the following statistic to test (3.11):

$$\hat{e}_j = \frac{1}{2}(\exp\{\sqrt{n_0}\hat{\beta}_j^{(dp)} - (\hat{V}_j^{(dp)})^2/2\} + \exp\{-\sqrt{n_0}\hat{\beta}_j^{(dp)} - (\hat{V}_j^{(dp)})^2/2\}). \tag{3.12}$$

Note that whenever $\beta_j^{(0)} = 0$, Theorem 2 gives $T_{n,j} := \sqrt{n_0}\hat{\beta}_j^{(dp)} \xrightarrow{D} \mathrm{N}(0, V_j^2)$, where $V_j^2 = (\boldsymbol{\Theta})_{j,j}\sigma^2$. With the supplementary verification of exponential integrability and the consistency of the variance estimator, $\mathbb{E}(\hat{e}_j) \to 1$. Thus, for $j \in \mathcal{J}_0$, $\hat{e}_j$ is an asymptotic e-value. After applying $\hat{e}_j$ with the e-BH procedure (Wang and Ramdas, 2022), we obtain the privatized FDR control procedure summarized in Algorithm 3.

The e-BH algorithm computes the e-values only for variables in the candidate set $\widehat{\mathcal{S}}$, whose size is at most s', and sets $\hat{e}_j = 0$ for $j \notin \widehat{\mathcal{S}}$. Following the privacy-accounting idea of Dwork et al. (2021), the private statistics are released only for the selected candidates rather than for all p variables, so the privacy composition is taken over

Algorithm 3 FDR Control Multiple Hypothesis Testing (e-BH)

1: **INPUT:** Candidate set $\widehat{\mathcal{S}} = \mathrm{supp}(\hat{\boldsymbol{\beta}}) \subset [p]$ with $|\widehat{\mathcal{S}}| \leq s'$, private pairs $(\hat{\beta}_j^{(dp)}, \hat{V}_j^{(dp)})$, $j \in \widehat{\mathcal{S}}$, each with privacy parameters $(\epsilon/(s'+1), \delta/(s'+1))$, FDR level $q \in (0,1)$

2: Set $\hat{e}_j = 0$ for $j \notin \widehat{\mathcal{S}}$.

3: **for** $j \in \widehat{\mathcal{S}}$ **do**

4: Release the private pair $(\hat{\beta}_j^{(dp)}, \hat{V}_j^{(dp)})$ jointly as per Algorithm 2.

5: Compute

$$\hat{e}_j = \frac{1}{2}(\exp\{\sqrt{n_0}\hat{\beta}_j^{(dp)} - (\hat{V}_j^{(dp)})^2/2\} + \exp\{-\sqrt{n_0}\hat{\beta}_j^{(dp)} - (\hat{V}_j^{(dp)})^2/2\}).$$

6: **end for**

7: Decide the threshold

$$k^\star = \max\left(\left\{j \in [p] : \frac{j\hat{e}_{[j]}}{p} \geq \frac{1}{q}\right\} \cup \{0\}\right), \qquad e^\star = \begin{cases} \hat{e}_{[k^\star]}, & k^\star \geq 1, \\ +\infty, & k^\star = 0. \end{cases}$$

8: **OUTPUT:** $\hat{\mathcal{J}}_1 = \{j \in [p] : \hat{e}_j \geq e^\star\}$

$s'+1$ private releases. For each selected variable j, the released private statistic is the joint pair $(\hat{\beta}_j^{(dp)}, \hat{V}_j^{(dp)})$, and the corresponding e-value is obtained by post-processing this pair. Therefore, the composition counts one release for constructing the candidate set and at most s' joint pair releases, rather than two separate releases per candidate. The subsequent step closely follows the classical BH procedure proposed by Benjamini and Hochberg (1995), with the only difference being the use of e-values in place of p-values. The rejection rule first finds the largest rank $k^\star = \max(\{j \in [p] : j\hat{e}_{[j]}/p \geq 1/q\} \cup \{0\})$, where $\hat{e}_{[j]}$ is the j-th largest value in $\{\hat{e}_1, \ldots, \hat{e}_p\}$. It then sets the threshold $e^\star = \hat{e}_{[k^\star]}$ if $k^\star \geq 1$, and $e^\star = +\infty$ otherwise. Finally, the rejection index set is given by $\hat{\mathcal{J}}_1 = \{j \in [p] : \hat{e}_j \geq e^\star\}$. According to the post-processing principle of differential privacy, the Algorithm 3 is (ϵ, δ)-STDP as well.

Theorem 3. Under the same conditions as in Theorem 2, applied to arbitrary non-negative random e-variables $\hat{e}_1, \ldots, \hat{e}_p$ computed by (3.12) and $q \in (0, 1)$, the e-BH procedure satisfies $\limsup_{n_0, p \to \infty} \mathrm{FDR} \leq q$. Moreover, the e-BH procedure is (ϵ, δ)-STDP.

This theorem establishes the validity of our method for FDR control. Compared to Benjamini and Hochberg (1995), our approach does not rely on additional assumptions about the dependence structure of $H_1, \ldots, H_p$, thus extending its applicability to a wider range of settings. From another perspective, one may also conduct the differential private BH procedure based on the p-value as proposed by Dwork et al. (2021) to control the FDR. We compare our method with Dwork et al. (2021) in the simulation part.

4. Simulation

In this section, we conduct a series of simulations to evaluate the performance of the proposed algorithms, as discussed in the previous sections. First, we validate the accuracy of point estimation and demonstrate the efficiency of the transfer step in Algorithm 1. Second, to assess the effectiveness of Algorithm 2 and verify Theorem 2, we employ the same model used for point estimation and evaluate the empirical coverage of the confidence intervals. Third, we compare the e-BH procedure in Algorithm 3 with the conventional p-value-based method and compare their relative performance.

4.1 Examples on Estimation Efficiency

In this part, we consider the estimation accuracy of the three statistics provided in the literature, which are initial estimator $\hat{\boldsymbol{\beta}}$ computed by Algorithm 1, debiased estimator $\hat{\boldsymbol{\beta}}^{(de)}$ and its privacy form $\hat{\boldsymbol{\beta}}^{(dp)}$. Our simulations are generally based on the following models:

$$\textbf{Target:}\ \ Y^{(0)} = (\boldsymbol{X}^{(0)})^\top \boldsymbol{\beta}^{(0)} + 4\sin(2\pi W_1^{(0)}) + 4\cos(2\pi W_2^{(0)}) + \varepsilon, \quad j = 1, ..., p,$$

$$\textbf{Source:}\ \ Y^{(k)} = (\boldsymbol{X}^{(k)})^\top \boldsymbol{\beta}^{(k)} + \varepsilon, \quad k = 1, \ldots, K.$$

We generate $\boldsymbol{X}^{(0)} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^{(0)})$, $\boldsymbol{X}^{(k)} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^{(k)})$, $\boldsymbol{X}^{(0)} \perp \boldsymbol{W}^{(0)}, W_1^{(0)}, W_2^{(0)} \sim \mathrm{Unif}([0,1])$, where $\boldsymbol{\Sigma}_{i,j}^{(0)} = (\omega^{(0)})^{|i-j|}$ and $\boldsymbol{\Sigma}_{i,j}^{(k)} = (\omega^{(k)})^{|i-j|}$ are Toepliz matrices shared by the target and source domains, respectively. ε is generated from the standard Gaussian distribution. We also assume that the sample size is the same across all source domains, i.e., $n_1 = n_2 = \cdots = n_K = n$. The default DP parameters are $\epsilon = 0.5$ and $\delta = 1/n^{1.1}$, which will be modified in Configuration 2. The default correlation is set to be $\omega^{(0)} = \omega^{(k)} = 0.6$, which will be modified in Configuration 3. The linear coefficients are $\boldsymbol{\beta}^{(0)} = (\mathbf{1}_{s_0}, \mathbf{0}_{p-s_0})^\top$ with $s_0 = 10$ and the source coefficients are generated as per

$$\beta_j^{(k)} = \begin{cases} \beta_j^{(0)} - 0.3\mathbb{I}(j \in \tilde{\mathcal{J}}^{(k)}), & \text{if } k \in \mathcal{A}, \\ \beta_j^{(0)} - 0.5\mathbb{I}(j \in \tilde{\mathcal{J}}^{(k)}), & \text{if } k \notin \mathcal{A}, \end{cases}$$

where $\tilde{\mathcal{J}}^{(k)} \subseteq [p]$ is a randomly selected subset from different sources, satisfying $|\tilde{\mathcal{J}}^{(k)}| =$

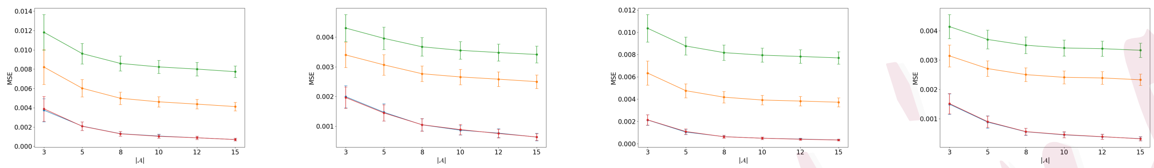
4. Similar coefficient settings can be found in Li et al. (2022). For different setups, we verify the robustness of Algorithm 1 across different parameters, with input sparse level $s' = 15$.

- **Configuration 1** : Fix the (ϵ, δ) privacy parameter, vary (n, n_0, p) by selecting $n \in \{500, 1000\}$, $n_0 \in \{500, 1000\}$ and $p \in \{500, 1000\}$.

- **Configuration 2**: Fix $(n, n_0, p) = (1000, 1000, 1000)$, vary the privacy parameter $\epsilon \in \{0.3, 0.4, 0.6, 0.7\}$.

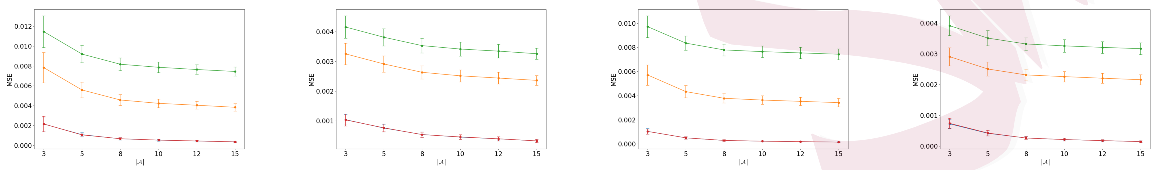
- **Configuration 3**: Fix $(n, n_0, p) = (1000, 1000, 1000)$, generate the target covariance matrix by $\omega^{(0)} = 0.5$, and vary the source correlation parameter $\omega^{(k)} \in \{0.2, 0.3, 0.4, 0.6\}$.

The B-spline method is applied to solve the optimization problem (2.2), (3.3) and (3.4) as Lian (2020) suggested. We conduct 500 independent experiments for each setting, and evaluate the performance of the algorithm by the averaged mean square error (MSE). In the first setting, we explore the effectiveness of transfer learning by increasing the number of transferable sources, following a similar approach to Tian and Feng (2023). We compare three estimators $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^{(de)}$ and $\hat{\boldsymbol{\beta}}^{(dp)}$, which are named as Initial, Debiased, and DP-Debiased, respectively. The benchmark, initial estimation without any noise (Zero Noise Estimation, ZNE) is also presented. Figure 1 shows that when $p = 500$, the accuracy of the estimations increases gradually as the number of transferable sources increases. Similar pattern is also observed in Figure 2 with higher



(500, 500, 500) (500, 1000, 500) (1000, 500, 500) (1000, 1000, 500)

Figure 1: MSE under Configuration 1 with changing (n, n_0, p).



(500, 500, 1000) (500, 1000, 1000) (1000, 500, 1000) (1000, 1000, 1000)

Figure 2: MSE under Configuration 1 with changing (n, n_0, p).

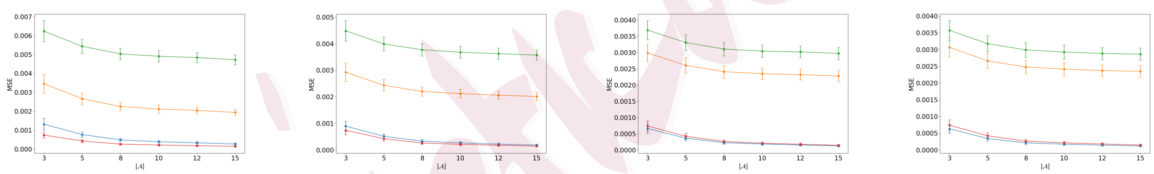


Figure 3: MSE under Configuration 2 with $\epsilon = (0.3, 0.4, 0.6, 0.7)$ from left to right.

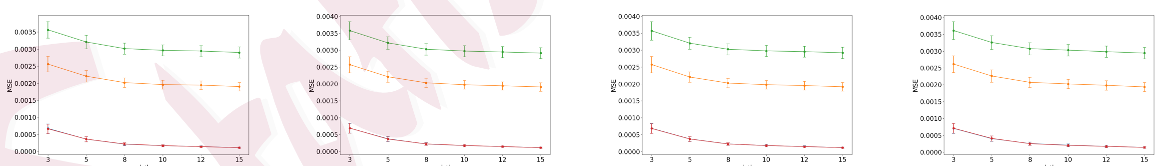
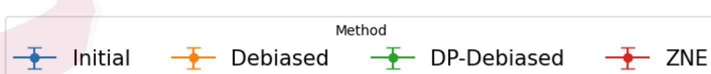


Figure 4: MSE under Configuration 3 with $\omega^{(k)} = (0.2, 0.3, 0.4, 0.6)$ from left to right.



dimension $p = 1000$, which implies the utility of transfer learning.

Overall, in Configuration 1, the MSE for initial estimation and ZNE are comparable and lowest among the estimators at any level of transferrable sources number, which implies a transfer efficient utility in parametric estimation of PLM. Specifically, in the high-dimensional setting ($n_0 < p$), $\hat{\boldsymbol{\beta}}$ also performs similarly to the other settings with the same trend as $|\mathcal{A}|$ increases. Note that the DP-Debiased estimator is relatively close to the Debiased estimator without noise, and has an acceptable difference with ZNE, indicating that the added noise provides reasonable privacy protection with minimal loss in statistical efficiency.

To further illustrate the relationship between privacy loss and estimation accuracy, we consider the second setting with varying privacy parameters. Figures 3 show that as ϵ decreases, more accuracy is sacrificed to achieve stronger privacy guarantees, which aligns with our theoretical findings.

In Figure 4, we show the robustness of transfer learning via a heterogeneous setting. We switch our target model a little bit with a new correlation $\omega^{(0)} = 0.5$ while the others remain the same as in model (4.1). In this setting, we shift the correlation of sources from 0.2 to 0.6. Again, we illustrate the estimation accuracy under different settings. The change in parameters has minimal impact on the performance of all three estimators, indicating the robustness of the proposed algorithm.

We present additional simulations in the supplementary material to validate the efficacy of transfer learning.

4.2 Asymptotic Normality and Statistical Inference

To illustrate the asymptotic normality of our estimation and the validity of confidence interval, we conduct a simulation setting the same as Configuration 1. We pick the settings of $(n, n_0, p) = (500, 1000, 1000)$ and $(n, n_0, p) = (1000, 1000, 1000)$ to show the histograms of $\hat{\beta}_1^{(dp)} - \beta_1$, $\hat{\beta}_{15}^{(dp)} - \beta_{15}$ and $\hat{\beta}_{30}^{(dp)} - \beta_{30}$. In addition to the histograms, we also compute the empirical coverage probabilities for active index, inactive index and overall rate, which are defined as:

$$CP_{\mathcal{S}_0} = \frac{1}{s_0} \sum_{j \in \mathcal{S}_0} \widehat{CP}(\beta_j^{(0)}), \quad CP_{\mathcal{S}_0^C} = \frac{1}{p - s_0} \sum_{j \in \mathcal{S}_0^c} \widehat{CP}(\beta_j^{(0)}), \quad \text{and} \quad CP = \frac{1}{p} \sum_{j \in [p]} \widehat{CP}(\beta_j^{(0)}),$$

where $\mathcal{S}_0$ represents the support set, comprising the non-zero indices of $\boldsymbol{\beta}$. $\widehat{CP}(\beta_j^{(0)}) := \sum_{i=1}^{500} \mathbb{I}_i(\beta_j^{(0)} \in \widehat{\mathrm{CI}}(j))/500$ denotes the estimated inclusion rate, calculated empirically over 500 repetitions. The results are summarized in Table 1. Similar to Zhang et al. (2024), we also report the average length of the confidence intervals.

From the histograms in Figure 5, we observe that the distribution of $\hat{\beta}_j^{(dp)}$ closely approximates a standard Gaussian distribution (red line in the figure). According to Table 1, the length of the confidence intervals remains relatively stable but shows a slight decrease, indicating that the efficiency of the intervals improves as more transferable sources are incorporated. This suggests that as the number of transferable sources increases, the resulting confidence intervals become more efficient and stable. Another important trend lies in the empirical coverage probabilities. Both $CP_{\mathcal{S}_0}$ and $CP_{\mathcal{S}_0^c}$ con-

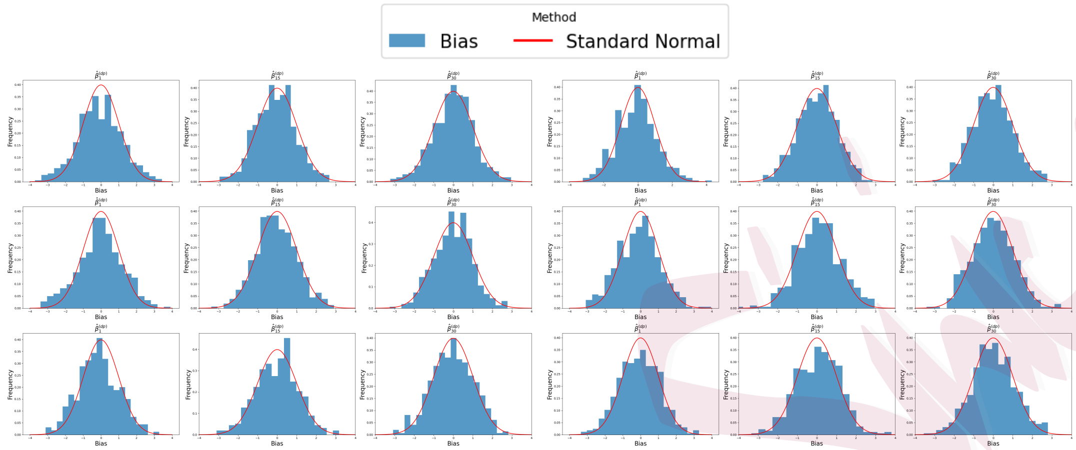


Figure 5: Left: $(n, n_0, p) = (500, 1000, 1000)$; Right: $(n, n_0, p) = (1000, 1000, 1000)$

n	n_0	$\lvert\mathcal{A}\rvert$	$p = 500$				$p = 1000$			
			length	$CP_{\mathcal{S}_0}$	$CP_{\mathcal{S}_0^c}$	CP	length	$CP_{\mathcal{S}_0}$	$CP_{\mathcal{S}_0^c}$	CP
500	500	3	0.3458	0.7218	0.9497	0.9452	0.3306	0.6054	0.9463	0.9429
		5	0.3039	0.7692	0.9525	0.9488	0.2860	0.6792	0.9472	0.9446
		10	0.2731	0.8170	0.9552	0.9525	0.2562	0.7508	0.9506	0.9486
		15	0.2631	0.8656	0.9580	0.9561	0.2446	0.7824	0.9495	0.9479
1000	500	3	0.3083	0.7876	0.9496	0.9464	0.2865	0.6972	0.9453	0.9428
		5	0.2761	0.8446	0.9526	0.9505	0.2570	0.7444	0.9471	0.9451
		10	0.2556	0.8686	0.9545	0.9528	0.2384	0.7810	0.9482	0.9465
		15	0.2499	0.8824	0.9552	0.9538	0.2329	0.8044	0.9485	0.9470
500	1000	3	0.2486	0.9176	0.9674	0.9664	0.2473	0.9036	0.9693	0.9687
		5	0.2357	0.9248	0.9666	0.9657	0.2335	0.9044	0.9693	0.9687
		10	0.2213	0.9400	0.9674	0.9668	0.2170	0.9352	0.9683	0.9679
		15	0.2153	0.9538	0.9681	0.9679	0.2118	0.9418	0.9697	0.9694
1000	1000	3	0.2400	0.9376	0.9670	0.9665	0.2341	0.9284	0.9694	0.9690
		5	0.2252	0.9460	0.9681	0.9677	0.2173	0.9394	0.9686	0.9683
		10	0.2115	0.9544	0.9666	0.9664	0.2068	0.9530	0.9690	0.9688
		15	0.2067	0.9580	0.9656	0.9655	0.2019	0.9554	0.9683	0.9682

Table 1: Table for the coverage probability

sistently improve as $|\mathcal{A}|$ increases, suggesting that the transferring algorithm performs better in the inference task for both active and inactive features. Overall, CP exhibits a steady increase, further reinforcing the notion that incorporating additional information enables the model to generate more accurate and reliable confidence intervals. These trends highlight a conclusion: incorporating knowledge from transferable sources leads to better accuracy, stability, and coverage on target model.

4.3 Multiple Hypothesis Testing with FDR Control

In this part, we continue to investigate the performance of multiple hypothesis testing procedure with STDP guarantee. Through the e-value introduced by (3.12), our simulation follows the Algorithm 3 step by step. In particular, the candidate set is taken as $\widehat{\mathcal{S}} = \mathrm{supp}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the preliminary private sparse estimator obtained from Algorithm 1. We compute the private pairs and the corresponding e-values only for $j \in \widehat{\mathcal{S}}$, and set $\hat{e}_j = 0$ for $j \notin \widehat{\mathcal{S}}$, as in Algorithm 3. Based on the previous two parts, here we can easily obtain the e-values in Configuration 1 with known $\hat{\beta}_j^{(dp)}$ and $\hat{V}_j^{(dp)}$.

In Figures 6 and 7, we illustrate the utility of FDR control for both proposed method (blue lines) and p-value-based method (Dwork et al. (2021), orange lines). The FDR here takes the average of 500 repetitions. According to Figures 6 and 7, both methods achieve an FDR below the nominal level when the sample size is sufficiently large. However, under Configuration 1, the p-value-based method exhibits slightly higher inflation compared to the proposed e-value-based method. Moreover, the power

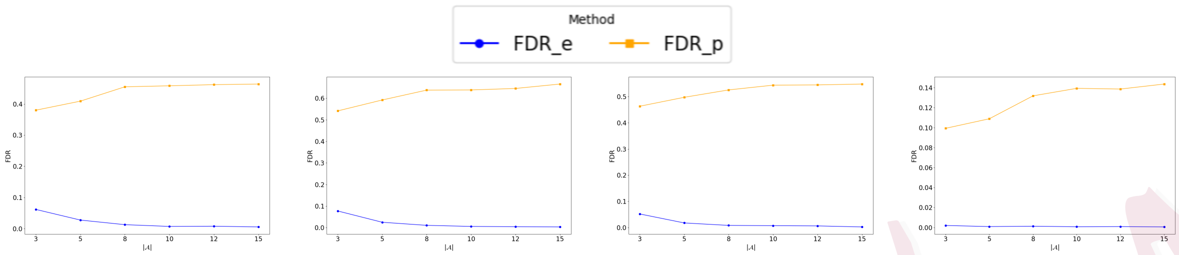


Figure 6: FDR under Configuration 1 with (n, n_0, p) changing as Figure 1.

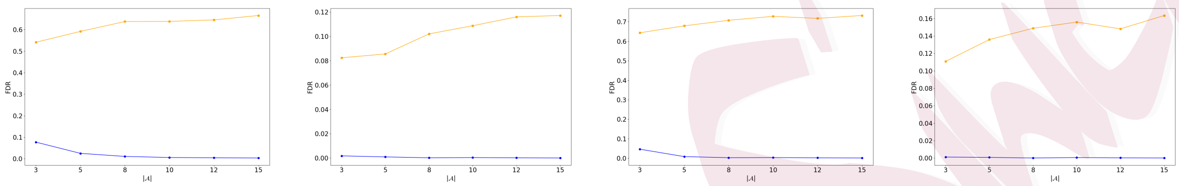


Figure 7: FDR under Configuration 1 with (n, n_0, p) changing as Figure 2.

of all these methods is close to 1.

5. Real Data

We apply our transfer learning framework to real-world air quality data collected from 20 monitoring stations in Shenzhen, China, spanning January $1, 2021$ through October $17, 2024$. Each station records daily concentrations of five air-pollutant components (PM_{10}, $PM_{2.5}$, SO_2, CO, and O_3) along with meteorological measurements: air temperature, relative humidity, wind speed (mph) with its bearing in degrees clockwise from true north, and precipitation intensity (inches/hour). Focusing on moderate to heavily polluted days ($PM_{2.5} > 35\,\mu g/m^3$ and $PM_{10} > 50\,\mu g/m^3$), we treat these weather variables as non-linear inputs and use the remaining air-pollutant concentrations as linear predictors to model PM_{10} at a target site (station 1360A). Several auxiliary stations

(9 of 19) are treated as transferable source domains in this study as shown in Table 2.

Station ID	**Sample size**
1360A (target)	31
1356A	74
1357A	113
1358A	88
1361A	97
1363A	18
1364A	32
1393A	97
1394A	109
1396A	55

Table 2: Stations selected.

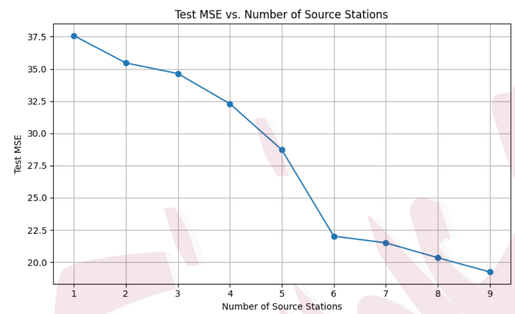


Figure 8: Test MSE at target 1360A using Algorithm 1.

We split the dataset on target site 1360A into a training set (80%) and a test set (20%). We combine the training data set from the target site and the full data from each source site as input to train the partial linear model with Algorithm 1. We evaluate the performance of Algorithm 1 by mean square error of predictions on the test data set from the target site. The result presented in Figure 8 demonstrates the utility of transfer learning in this real case, as the test error generally decreases with increasing number of transfer sources.

6. Discussion

In this work, we propose a transfer learning algorithm for estimation and inference in high-dimensional sparse PLM under differential privacy protection. Our method not only enhances estimation efficiency in the target domain but also addresses inference problem based on the asymptotic distribution of estimation. We tailor the standard DP definition to fit our transfer learning framework and incorporate a multiple hypothesis testing procedure that rigorously controls the FDR.

Looking ahead, several promising research directions remain. First, based on the flexibility of PLM, transfer learning algorithms in PLM could be extended to accommodate other data structure scenarios involving longitudinal data and functional data. Second, beyond the current linear-transfer setting, it would be interesting to study more general semiparametric transfer models in which the source domains also contain domain-specific nonparametric nuisance components, and to investigate how to characterize their transferability while balancing the resulting additional statistical and privacy costs. Finally, while our e-value-based FDR control procedure exhibits strong empirical power in simulations, establishing its theoretical power guarantees remains an important open challenge.

Supplementary Materials

The online Supplementary Material contains some theoretical statements, auxiliary results, all technical proofs and additional simulation results. The code supporting this paper is available at `https://github.com/moondanced/Trans-DPPLM`.

Acknowledgments

We would like to thank the Editor, the Associate Editor, and the two anonymous reviewers for their valuable comments and constructive suggestions, which led to significant improvements in the paper. Yibo Yan's research is supported by the National Natural Science Foundation of China (12401390). Riquan Zhang's research is supported by the National Natural Science Foundation of China (12371272, 12531013).

References

Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318.

Auddy, A., T. T. Cai, and A. Chakraborty (2025). Minimax and adaptive transfer learning for nonparametric classification under distributed differential privacy constraints. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf070.

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Management Science 67*(5), 2964–2984.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology 57*(1), 289–300.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Volume 4. Springer.

Cai, T., M. Li, and M. Liu (2025). Semi-supervised triply robust inductive transfer learning. *Journal of the American Statistical Association 120*(550), 1037–1047.

Cai, T. T., A. Chakraborty, and L. Vuursteen (2024). Federated nonparametric hypothesis testing with differential privacy constraints: Optimal rates and adaptive tests. *arXiv preprint arXiv:2406.06749*.

Cai, T. T., A. Chakraborty, and L. Vuursteen (2026). Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. *Journal of the American Statistical Association*, in press.

Cai, T. T., Y. Wang, and L. Zhang (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics 49*(5), 2825–2850.

Cai, Z., S. Li, X. Xia, and L. Zhang (2026). Differentially private estimation and inference in high-dimensional regression with FDR control. *Journal of Machine Learning Research 27*, 1–54.

Cui, S., X. Guo, and Z. Zhang (2025). Estimation and inference in ultrahigh-dimensional partially linear single-index models. *Science China Mathematics 68*(8), 1807–1840.

Duchi, J. C., M. I. Jordan, and M. J. Wainwright (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association 113*(521), 182–201.

Dwork, C., A. Roth, et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*(3–4), 211–407.

Dwork, C., W. Su, and L. Zhang (2021). Differentially private false discovery rate control. *Journal of Privacy and Confidentiality 11*(2).

Engle, R. F., C. W. Granger, J. Rice, and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association 81*(394), 310–320.

Hanneke, S. and S. Kpotufe (2019). On the value of target data in transfer learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

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

Hu, X. and X. Zhang (2023). Optimal parameter-transfer learning by semiparametric model averaging. *Journal of Machine Learning Research 24*(358), 1–53.

Jiao, Y., H. Lin, Y. Luo, and J. Z. Yang (2024). Deep transfer learning: Model framework and error analysis. *arXiv preprint arXiv:2410.09383*.

Kim, S., D. Zeng, and J. M. Taylor (2017). Joint partially linear model for longitudinal data with informative drop-outs. *Biometrics 73*(1), 72–82.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*, Volume 61. Springer.

Li, M., Y. Tian, Y. Feng, and Y. Yu (2024). Federated transfer learning with differential privacy. *arXiv preprint arXiv:2403.11343*.

Li, N., Y. Fei, and X. Zhang (2024). Partial linear model averaging prediction for longitudinal data. *Journal of Systems Science and Complexity 37*(2), 863–885.

Li, S., T. T. Cai, and H. Li (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(1), 149–173.

Li, S., T. T. Cai, and H. Li (2023). Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association 118*(543), 2171–2183.

Lian, H. (2020). Asymptotics of the non-parametric function for b-splines-based estimation in partially linear models. *International Statistical Review 88*(1), 142–154.

Lian, H., K. Zhao, and S. Lv (2019). Projected spline estimation of the nonparametric function in high-dimensional partially linear models for massive data. *The Annals of Statistics 47*(5), 2922–2949.

Ling, N., Y. Yang, and Q. Peng (2025). Partial linear quantile regression model with incompletely observed functional covariates. *Journal of Nonparametric Statistics 37*(3), 713–739.

Liu, W., X. Mao, and X. Zhang (2022). Fast and robust sparsity learning over networks: A decentralized surrogate median regression approach. *IEEE Transactions on Signal Processing 70*, 797–809.

Liu, Y., S. Zhang, S. Ma, and Q. Zhang (2020). Tests for regression coefficients in high dimensional partially linear models. *Statistics & Probability Letters 163*, 108772.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics 45*(1), 158–195.

Pournaderi, M. and Y. Xiang (2021). Differentially private variable selection via the knockoff filter. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE.

Qiao, D. and Y.-X. Wang (2023). Offline reinforcement learning with differential privacy. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Advances in Neural Information Processing Systems*, Volume 36, pp. 61395–61436. Curran Associates, Inc.

Ren, Z. and R. F. Barber (2024). Derandomised knockoffs: Leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology 86*(1), 122–154.

Shi, H., W. Yang, N. Zhou, and X. Guo (2026). Inference for partially linear quantile regression models in ultrahigh dimension. *Communications in Mathematics and Statistics 14*(3), 495–540.

Shi, Y., M. Hao, Y. Tang, and X. Guo (2025). Estimation and inference of high-dimensional partially linear regression models with latent factors. *arXiv preprint arXiv:2501.06529*.

Tan, F., X. Jiang, X. Guo, and L. Zhu (2021). Testing heteroscedasticity for regression models based on projections. *Statistica Sinica 31*(2), 625–646.

Tian, Y. and Y. Feng (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association 118*(544), 2684–2697.

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

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Volume 48. Cambridge University Press.

Wang, D., L. Hu, H. Zhang, M. Gaboardi, and J. Xu (2023). Generalized linear models in non-interactive local differential privacy with public data. *Journal of Machine Learning Research 24*(132), 1–57.

Wang, F. and Y. Yu (2025). Transfer learning for piecewise-constant mean estimation: Optimality, l_1-and l_0-penalization. *Biometrika*, asaf018.

Wang, R. and A. Ramdas (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(3), 822–852.

Wang, Y. and A. Nedić (2023). Tailoring gradient methods for differentially private distributed optimization. *IEEE Transactions on Automatic Control 69*(2), 872–887.

Wei, K., J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security 15*, 3454–3469.

Wong, R. K., Y. Li, and Z. Zhu (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Associa-*

tion 114(525), 406–418.

Xia, X. and Z. Cai (2023). Adaptive false discovery rate control with privacy guarantee. *Journal of Machine Learning Research 24*(252), 1–35.

Xie, H. and J. Huang (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics 37*(2), 673–696.

Yao, Y. and G. Doretto (2010). Boosting for transfer learning with multiple sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1855–1862. IEEE.

Yu, C., R. Ming, M. Xiao, and Z. Wang (2024). A flexible approach: Variable selection procedures with multilayer FDR control via e-values. *arXiv preprint arXiv:2409.17039*.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(1), 217–242.

Zhang, Y. and Z. Zhu (2025). Transfer learning for high-dimensional quantile regression via convolution smoothing. *Statistica Sinica 35*, 939–958.

Zhang, Z., R. Nakada, and L. Zhang (2024). Differentially private federated learning: Servers trustworthiness, estimation, and statistical inference. *arXiv preprint arXiv:2404.16287*.

Zhao, F., N. Lin, and B. Zhang (2023). A new test for high-dimensional regression coefficients in partially linear models. *Canadian Journal of Statistics 51*(1), 5–18.

Zhu, L., M. Ding, V. Aggarwal, J. Xu, and D. Wang (2024). Improved analysis of sparse linear regression in local differential privacy model. In *12th International Conference on Learning Representations*.

Zhu, Y. (2017). Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients. *The Annals of Statistics 45*(5), 2274–2298.

Zhu, Y., Z. Yu, and G. Cheng (2019). High dimensional inference in partially linear models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2760–2769. PMLR.

Zhu, Z., Y. Yan, L. Gefei, and R. Zhang (2025). Recent developments on statistical transfer learning. *International Statistical Review*, in press.