Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Residual Importance Weighted Transfer Learning for High-dimensional Linear Regression

Junlong Zhao, Shengbin Zheng & Chenlei Leng

To cite this article: Junlong Zhao, Shengbin Zheng & Chenlei Leng (04 Jun 2026): Residual Importance Weighted Transfer Learning for High-dimensional Linear Regression, Journal of the American Statistical Association, DOI: 10.1080/01621459.2026.2623997

To link to this article: https://doi.org/10.1080/01621459.2026.2623997

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2026, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2026.2623997

Residual Importance Weighted Transfer Learning for High-dimensional Linear Regression

Junlong Zhao[a], Shengbin Zheng[a], and Chenlei Leng[b]

[a]School of Statistics, Beijing Normal University, Beijing, China; [b]Department of Applied Mathematics, Hong Kong Polytechnic University, Kowloon, Hong Kong

ABSTRACT

Transfer learning is an emerging paradigm for leveraging multiple sources to improve the statistical inference on a single target. In this article, we propose a novel approach named residual importance weighted transfer learning (RIW-TL) for high-dimensional linear models built on penalized likelihood. Compared to existing methods such as Trans-Lasso that selects sources in an (approximately) all-in-or-all-out manner, RIW-TL includes samples via importance weighting and thus may permit more effective sample use. To determine the weights, remarkably RIW-TL only requires the knowledge of one-dimensional densities dependent on residuals, thus overcoming the curse of dimensionality of having to estimate high-dimensional densities in naive importance weighting. We show that the oracle RIW-TL provides faster rate than its competitors and develop a cross-fitting procedure to estimate this oracle. We discuss variants of RIW-TL by adopting different choices for residual weighting. The theoretical properties of RIW-TL and its variants are established and compared with those of LASSO and Trans-Lasso. Extensive simulation and a real data analysis confirm its advantages. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received March 2025
Accepted January 2026

KEYWORDS
Density estimation; High-dimensional linear models; Importance weighting; Sample selection; Transfer learning

1. Introduction

Statistical techniques are most challenging in applications with small sample sizes, especially when accompanied by a large number of variables. While penalized regression has become a well-established approach for constructing sparse models and addressing high dimensionality, the inherent limitations of small sample size fundamentally constrain the statistical properties of any estimator.

Fortunately, even when the *target* sample size is small, multiple *source* datasets from related domains are often available, offering valuable auxiliary information. Leveraging these sources can enhance estimation accuracy if they are sufficiently similar to the target. However, incorporating non-informative sources may result in *negative transfer*, ultimately degrading performance. In this article, we focus on a particularly relevant transfer learning setup where source datasets are significantly larger than the target, and among them, some are informative—but we do not know which ones. Intuitively, this is the regime where improved rates of parameter estimation can be achieved, provided the informative sources are correctly identified. Otherwise, relying solely on the target data remains the safest approach.

We illustrate our method in the context of linear regression, where the number of variables exceeds the number of observations. The goal is to model the relationship between a response variable $y \in \mathbb{R}$ and a predictor vector $\boldsymbol{x} \in \mathbb{R}^p$. Suppose we are provided with the target data $\mathcal{S}^{(0)} = \{\boldsymbol{z}_i^{(0)} = ((\boldsymbol{x}_i^{(0)})^\top, y_i^{(0)})^\top,\ i = 1, \ldots, n_0\}$, where the observations follow the linear model

$$y_i^{(0)} = (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta}^{(0)} + \epsilon_i^{(0)},$$

with random noise $\epsilon_i^{(0)}$ that is independent of $\boldsymbol{x}_i^{(0)}$ and satisfies $\mathbb{E}(\epsilon_i^{(0)}) = 0$. Our objective is to estimate the unknown regression coefficients $\boldsymbol{\beta}^{(0)}$. Assume that $\boldsymbol{\beta}^{(0)}$ is sparse, meaning the cardinality of its support, $s_0 := |\mathrm{supp}(\boldsymbol{\beta}^{(0)})|$, satisfies $s_0 \ll n_0$ (or approximately so). The LASSO estimator (Tibshirani 1996)

$$\hat{\boldsymbol{\beta}}^{(0)}_{\mathrm{Lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\mathrm{argmin}} \frac{1}{2n_0} \sum_{i=1}^{n_0} \left\{ y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta} \right\}^2 + \lambda \|\boldsymbol{\beta}\|_1, \quad (1)$$

achieves the convergence rate $\|\hat{\boldsymbol{\beta}}^{(0)}_{\mathrm{Lasso}} - \boldsymbol{\beta}^{(0)}\|^2 = O_p(s_0 \log p / n_0)$ (Bickel, Ritov, and Tsybakov 2009), where $\|\cdot\|_1$ and $\|\cdot\|$ denote the ℓ_1 and ℓ_2 norm, respectively. For LASSO to achieve consistent estimation, a fundamental requirement is $n_0 \gg s_0 \log p$. In this article, we refer to the LASSO estimator as the one defined in (1), which relies solely on the target data $\mathcal{S}^{(0)}$.

CONTACT Chenlei Leng *chenlei.leng@polyu.edu.hk* Department of Applied Mathematics, Hong Kong Polytechnic University, Kowloon, Hong Kong.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

In our regression setup, we consider K independent source datasets, denoted as $\mathcal{S}^{(k)} = \{\boldsymbol{z}_i^{(k)} = ((\boldsymbol{x}_i^{(k)})^\top, y_i^{(k)})^\top,\ i = 1, \ldots, n_k\}$ for $k = 1, \cdots, K$, where each dataset consists of iid observations satisfying the model

$$y_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(k)} + \epsilon_i^{(k)},$$

with noise terms $\epsilon_i^{(k)}$ that are independent of $\boldsymbol{x}_i^{(k)}$ and satisfy $\mathbb{E}(\epsilon_i^{(k)}) = 0$.

Without loss of generality, we assume that for $k = 0, 1, \ldots, K$, the vectors $\boldsymbol{x}_i^{(k)}$ are centered, that is, $\mathbb{E}(\boldsymbol{x}_i^{(k)}) = 0$. Furthermore, we denote their covariance matrix by $\boldsymbol{\Sigma}^{(k)} = (\sigma_{ij}^{(k)})$ and assume that its eigenvalues are bounded away from both 0 and ∞. Let $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}$ represent the difference between the regression coefficients of the target and source k, and define $h_k = \|\boldsymbol{\delta}^{(k)}\|_1$ for $k \geq 1$. The informativeness of source k is inversely related to the magnitude of h_k, with smaller values of h_k indicating greater informativeness.

1.1. Transfer Learning

Transfer learning is a rapidly evolving paradigm that uses knowledge from source domains to improve performance in the target domain (Zhuang et al. 2021). In statistics, transfer learning has been extensively studied across various models (Bastani 2021; Tian and Feng 2022; Li, Cai, and Duan 2023; Li et al. 2024; Cai and Pu 2024; Zhang and Zhu 2025). A common theme in these works is the selection of source data at the source level, meaning that all observations from a given source k are either fully included or entirely excluded. We refer to this class of methods as the all-in-or-all-out approach, or the (approximately) all-in-or-all-out approach when further downstream model aggregation is employed, as in the case of Trans-Lasso.

For the high-dimensional linear regression model, Li, Cai, and Li (2021) introduced the oracle Trans-Lasso method, an all-in-or-all-out approach, under the assumption that the informative set $\mathcal{A}$, defined as

$$\mathcal{A} = \{1 \leq k \leq K : \|\boldsymbol{\delta}^{(k)}\|_1 \leq h\},$$

is known, where h is assumed to be small. The oracle Trans-Lasso estimator for the target coefficients $\boldsymbol{\beta}^{(0)}$ is given by

$$\hat{\boldsymbol{\beta}}^{(0)}_{\text{Trans}-\text{Lasso}} = \hat{\boldsymbol{w}}^{\mathcal{A}} - \hat{\boldsymbol{\delta}}^{\mathcal{A}},$$

where $\hat{\boldsymbol{w}}^{\mathcal{A}}$ is the LASSO estimator using data from the informative set $\cup_{k\in\mathcal{A}}\mathcal{S}^{(k)}$, and

$$\hat{\boldsymbol{\delta}}^{\mathcal{A}} = \underset{\boldsymbol{\delta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n_0}\sum_{i=1}^{n_0}\left\{y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top(\hat{\boldsymbol{w}}^{\mathcal{A}} - \boldsymbol{\delta})\right\}^2 + \lambda_\delta\|\boldsymbol{\delta}\|_1.$$

Supposing that $h \lesssim s_0\sqrt{\log p/n_0}$ and $n_0 \lesssim n_{\mathcal{A}}$, the oracle Trans-Lasso achieves a convergence rate of r_n, as shown in (16). However, in practice, the informative set $\mathcal{A}$ is unknown, and misidentification may lead to negative transfer. To mitigate this issue, the authors proposed Trans-Lasso, an averaging algorithm that aggregates estimates from multiple candidate informative sets. This results in a convergence rate of $r_n + \log K/n_0$, which is undesirable. In the most relevant case where n_0 is small relative to n_k, the second term dominates, yielding a rate comparable to that of the LASSO estimator in (1).

Furthermore, the effectiveness of Trans-Lasso hinges on a strong covariance homogeneity assumption—namely, that the covariance matrices of sources in $\mathcal{A}$ must be similar to that of the target. To improve robustness to covariate shift, Li et al. (2024) and He, Sun, and Li (2024) proposed estimators under the condition $\max_{1\leq k\leq K}\|\boldsymbol{\delta}^{(k)}\|_1 \ll s_0\sqrt{\log p/n_0}$. This condition, however, remains quite restrictive in practice. When it is violated, these methods may suffer from negative transfer. Thus, in scenarios where n_0 is small and the set of informative sources is unknown, a critical challenge arises: how to improve convergence rates while mitigating the risk of negative transfer. This motivates the following fundamental question:

How can we leverage source data more effectively in transfer learning beyond the (approximately) all-in-or-all-out approach?

1.2. Importance-Weighted Method

One approach to use all the observations in the source data is through importance weighting (Lu et al. 2022). Specifically, with the notation $\boldsymbol{z} = (\boldsymbol{x}^\top, y)^\top$, we have the following key observation:

$$\boldsymbol{\beta}^{(0)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, \mathbb{E}_{\boldsymbol{z}\sim f_k}\left\{\omega^{(k)}(\boldsymbol{z})\cdot(y - \boldsymbol{x}^\top\boldsymbol{\beta})^2\right\}, \tag{2}$$

with $\omega^{(k)}(\boldsymbol{z}) = f_0(\boldsymbol{z})/f_k(\boldsymbol{z})$, where f_0 and f_k represent the probability density functions (pdf) of $\boldsymbol{z}$ in the target and kth source, respectively. To estimate $\boldsymbol{\beta}^{(0)}$, a straightforward approach is to formulate a weighted least-squares loss function that aggregates all observations from both the target and source datasets. This can be enhanced by adding an ℓ_1 penalty on the estimand to encourage sparsity, similar to LASSO. Unlike the (approximately) all-in-or-all-out methods, this approach uses all source data, thereby avoiding the need for rigid inclusion or exclusion of entire datasets. Recently, the importance weighting approach has also been employed in different contexts to address covariate shift, including nonparametric regression and robust estimation (Ma, Pathak, and Wainwright 2023; Zhou et al. 2024; Cai, Li, and Liu 2024).

As seen in (2), a crucial step in applying importance weighting is estimating the density ratio, $f_0(\boldsymbol{z})/f_k(\boldsymbol{z})$. For low- to moderate-dimensional problems, numerous well-established methods exist for estimating these weights, often leveraging nonparametric techniques or relying on model assumptions (Lu et al. 2022, see). However, these methods become impractical in high-dimensional settings due to the curse of dimensionality. The foundation of our novel importance-weighted transfer learning approach lies in the following key property, which circumvents this issue by requiring only the estimation of one-dimensional densities.

Proposition 1. For $1 \leq k \leq K$, the (2) holds with

$$\omega^{(k)}(\boldsymbol{z}) = \frac{f_\epsilon(y - \boldsymbol{x}^\top\boldsymbol{\beta}^{(0)})}{f_{\epsilon^{(k)}}(y - \boldsymbol{x}^\top\boldsymbol{\beta}^{(k)})}, \tag{3}$$

where f_ϵ is the pdf of a univariate random variable ϵ, independent of $\boldsymbol{x}$ (i.e., the data in source k), and $f_{\epsilon^{(k)}}$ is the pdf of $\epsilon^{(k)}$ in the kth source data.

Note that in this proposition, we have omitted the dependence of ϵ on k for simplicity. Several remarks are in order. First, the unknown parameters $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(k)}$ in (3) can be estimated using penalized regression or related methods, using the target data and the source data respectively. Second, for the denominator of the weights, estimating $f_{\epsilon^{(k)}}$ can be based on the residuals $y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(k)}$ and is straightforward if $\epsilon^{(k)}$ follows a parametric distribution. Alternatively, without assuming a parametric form, $f_{\epsilon^{(k)}}$ can be estimated nonparametrically via univariate density estimation. The remainder of this article focuses on the nonparametric case, which is of particular methodological and theoretical interest, while we numerically explore the performance under the assumption of Gaussianity for $f_{\epsilon^{(k)}}$. Finally, in the numerator, ϵ can be any random variable with a mean of zero, offering flexibility in its choice. We discuss two examples: one with a symmetric distribution and another with a uniform distribution for ϵ. Since our approach fundamentally relies on residuals, we refer to it as Residual Importance Weighted Transfer Learning (RIW-TL).

1.3. Contributions

Methodologically, this article introduces Residual Importance Weighted Transfer Learning (RIW-TL), a novel approach for transferring information from source data to the target. Unlike existing (approximately) all-in-or-all-out frameworks (Li, Cai, and Li 2021; Tian and Feng 2022, see), RIW-TL assigns a weight to each individual observation in the source data, rather than including or excluding entire sources. This shift from selection to weighting represents a paradigm change, offering a more nuanced and flexible method for incorporating source data, which enables us to leverage the information more efficiently. In particular, the weights, defined by a one-dimensional density ratio, effectively circumvent the curse of dimensionality.

Notably, RIW-TL is agnostic to covariate heterogeneity: Proposition 1 holds without requiring the marginal distribution of $\boldsymbol{x}^{(k)}$, the population version of the predictor in source k, to be the similar across sources—a key limitation of many existing methods, such as Trans-Lasso. This makes our framework adaptable to the *covariate shift* setting, where covariate distributions differ between source and target. This flexibility makes RIW-TL highly practical in a wide range of settings where such differences are common.

Theoretically, the oracle RIW-TL estimator with known weights achieves a convergence rate comparable to or better than the oracle Trans-Lasso and always outperforms LASSO. When informative sources are close to the target, its rate matches that of the oracle Trans-Lasso; when they moderately deviate, it performs better. See Table 1 for a detailed comparison with a single informative source.

In practice, when the oracle is unavailable and the weights in (3) must be estimated, we propose a cross-fitting procedure to implement RIW-TL. We explore both nonparametric estimation via kernel density estimation (Silverman 1978) and parametric estimation under the assumption of Gaussian errors. A comparison with Trans-Lasso at the end of Section 3 shows that RIW-TL consistently outperforms Trans-Lasso under mild conditions. Specifically, when no informative sources are available, RIW-TL prevents negative transfer, maintaining a convergence rate comparable to LASSO and Trans-Lasso. When informative sources are present, RIW-TL attains a significantly faster rate, making it particularly effective for transfer learning, especially when n_0 is small—where Trans-Lasso performs similarly to LASSO. However, theoretical results suggest that RIW-TL, when estimating the distribution of ϵ based on kernel density estimation of $\epsilon^{(k)}$, remains suboptimal compared to the oracle RIW-TL. To address this, we show in Section 4 that modeling ϵ as a uniform random variable with unknown endpoints allows RIW-TL to closely approximate the oracle rate while maintaining its advantages over Trans-Lasso and LASSO.

In developing RIW-TL, this article makes significant technical contributions to addressing a fundamental challenge in importance weighting methods: the issue of the ratio $f_\epsilon / f_{\epsilon^{(k)}}$ becoming unbounded. This problem is commonly encountered in literatures such as Cortes, Mansour, and Mohri (2010). To mitigate this, we introduce a sample selection strategy that ensures the weights, as defined in (3), remain bounded. However, this strategy introduces additional challenges in parameter estimation, particularly bias. When the sample selection subsets are known, we show that setting ϵ to follow a symmetric distribution eliminates bias completely. When the subsets must be estimated, bias can be reduced by modeling ϵ within a scale family and selecting parameters appropriately. Our rigorous analysis provides a robust solution to this complex issue, advancing the field of importance weighting methods.

1.4. Organization and Notations

The article is organized as follows. Section 2 analyzes the oracle RIW-TL. Section 3 develops a practical estimator via kernel density estimation and cross-fitting, with theoretical guarantees. Section 4 explores a variant with alternative weighting and sample selection, achieving oracle performance. Sections 5 and 6 present simulations and real data analyses, followed by a discussion in Section 7. Technical details and additional results are presented in the supplementary material.

Notations. Let $\{a_n\}$ and $\{b_n\}$ be two sequences of positive numbers. We use the notation $a_n \gtrsim b_n$ or $b_n \lesssim a_n$ to indicate that $\lim_{n\to\infty} a_n/b_n = C > 0$ for some constant C, and $a_n \asymp b_n$ to signify that they grow at the same rate. We write $a_n \ll b_n$

Table 1. The convergence rates of the two oracles with a single source having sample size n_1.

Regime	Oracle RIW-TL	Oracle Trans-Lasso
(*i*) $h \lesssim \sqrt{s_0 \log p/(n_0+n_1)}$	$s_0 \log p/(n_0+n_1)$	$s_0 \log p/(n_0+n_1)$
(*ii*) $\sqrt{s_0 \log p/(n_0+n_1)} \ll h \ll \sqrt{\log p/n_0}$	$s_0 \log p/(n_0+n_1)$	h^2
(*iii*) $\sqrt{\log p/n_0} \lesssim h \lesssim s_0\sqrt{\log p/n_0}$	$s_0 \log p/(n_0+n_1\{M_1/d_1 \wedge 1\})$	$h\sqrt{\log p/n_0}$

or $a_n = o(b_n)$ when $a_n/b_n \to 0$, and $a_n = O(b_n)$ when $\lim_{n\to\infty} a_n/b_n = C < \infty$ for some constant C. The stochastic versions of these notations are defined similarly, with $o_p(\cdot)$ and $O_p(\cdot)$. Let $a \wedge b$ denote the minimum of a and b. For a vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, we denote its ℓ_1, ℓ_2, and ℓ_∞ norms by $\|\boldsymbol{v}\|_1$, $\|\boldsymbol{v}\|$, and $\|\boldsymbol{v}\|_\infty$, respectively; the ℓ_0 norm of $\boldsymbol{v}$, denoted as $\|\boldsymbol{v}\|_0$, is defined as the number of nonzero element of $\boldsymbol{v}$. We define $[p] := \{1, \ldots, p\}$ for any integer p and use $\boldsymbol{v}_{\mathcal{C}}$ to denote the subvector of $\boldsymbol{v}$ whose entries are indexed by the set $\mathcal{C}$. In addition, the notation $|\mathcal{C}|$ serves to denote the cardinality of $\mathcal{C}$.

2. Residual Importance Weighting

This section introduces the oracle RIW-TL estimator, which serves as a benchmark for the best achievable rate under the assumption that the importance weights in (3) are known. Even with known weights, they may become degenerate, necessitating a careful selection of observations with bounded weights. Such selection impacts the loss function and its minimizer, potentially introducing bias. We further examine the choice of f_ϵ and discuss the notion of effective sample size underlying RIW-TL, which is key to understanding its theoretical properties.

For clarity, we define our parameter space as

$$\left\{\|\boldsymbol{\beta}^{(0)}\|_0 \le s_0,\ \|\boldsymbol{\beta}^{(k)}\|_0 \le s_k,\ \|\boldsymbol{\delta}^{(k)}\|_1 \le h_k,\ k = 1, \ldots, K\right\}, \tag{4}$$

for some constants s_0, s_k, and h_k. In other words, both $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(k)}$ are assumed to be sparse vectors. Rewriting the response $y_i^{(k)}$ as

$$\begin{aligned} y_i^{(k)} &= (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(k)} + \epsilon_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(0)} \\ &\quad + \{(\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}) + \epsilon_i^{(k)}\}, \end{aligned}$$

we define

$$\eta_i^{(k)} := (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}) = (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\delta}^{(k)}. \tag{5}$$

From Proposition 1, the importance weight for observation $\boldsymbol{z}_i^{(k)}$ is

$$\omega_i^{(k)} = \frac{f_\epsilon\left(y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(0)}\right)}{f_{\epsilon^{(k)}}\left(y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(k)}\right)} = \frac{f_\epsilon(\epsilon_i^{(k)} + \eta_i^{(k)})}{f_{\epsilon^{(k)}}(\epsilon_i^{(k)})}. \tag{6}$$

Here, $\eta_i^{(k)}$ captures the difference between $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\beta}^{(0)}$, adjusted by the predictor. Since the denominator in (6) may approach zero, $\omega_i^{(k)}$ can become unbounded. To mitigate this, we apply a sample selection procedure, retaining only observations with bounded weights:

$$\mathcal{I}_k = \{i \in [n_k] : |\epsilon_i^{(k)} + \eta_i^{(k)}| \le A,\ |\eta_i^{(k)}| \le M_k\}, \quad k = 1, \ldots, K, \tag{7}$$

where A and M_k's are positive tuning parameters. This selection ensures bounded weights under appropriate conditions and we will refer to $\mathcal{I}_k$ as *sample selection subsets*. We discuss the selection of A and M_k via cross-validation later.

While $\mathcal{I}_k$ is one possible selection criterion, alternatives exist, such as $\{i \in [n_k] : |\epsilon_i^{(k)}| \le A, |\eta_i^{(k)}| \le M_k\}$, which also constrains the weights. The choice of $\mathcal{I}_k$ impacts theoretical properties, as discussed in Section 3.1, particularly in Remark 5, with the alternative subset examined in Remark 3 in Section 2.1.

Remark 1. Our choice to clip on the residuals rather than the weights is primarily motivated by theoretical considerations to ensure the consistency of the weights estimator $\hat{\omega}^{(k)} = \hat{f}_\epsilon / \hat{f}_{\epsilon^{(k)}}$, where $\hat{f}_\epsilon$ and $\hat{f}_{\epsilon^{(k)}}$ represent the estimators of f_ϵ and $f_{\epsilon^{(k)}}$, respectively. Specifically, even when the weights remain bounded, the true density f_ϵ could be smaller than the estimation error $|\hat{f}_\epsilon - f_\epsilon|$, leading to inconsistency of the weights estimator. Further explanations are provided at Section S.4.1 in the supplementary materials.

To analyze the magnitude of $\omega_i^{(k)}$, we impose the following assumption on $f_{\epsilon^{(k)}}$.

Condition 1. For $k = 1, \ldots, K$, the density $f_{\epsilon^{(k)}}$ is strictly positive for all finite t and has a bounded first derivative.

This assumption is mild, as it permits $\inf_{t\in\mathbb{R}} f_{\epsilon^{(k)}}(t) = 0$ and holds for many common distributions, including the normal and t-distributions. Under this assumption, the following proposition ensures that both $f_{\epsilon^{(k)}}(\epsilon_i^{(k)})$ and $\omega_i^{(k)}$ remain bounded for all $i \in \mathcal{I}_k$.

Proposition 2. Under Condition 1, for $k = 1, \ldots, K$ and for all $i \in \mathcal{I}_k$, the following hold:

(i) $f_{\epsilon^{(k)}}(\epsilon_i^{(k)})$ is strictly bounded away from 0 (i.e., it is greater than a positive constant), and
(ii) $\omega_i^{(k)}$ is strictly bounded away from both 0 and infinity.

With bounded weights, we define the oracle RIW-TL estimator, assuming known weights and sample selection subsets. Let $\mathcal{I}_0 = [n_0]$ and set $\omega_i^{(0)} = 1$ for all target observations. The oracle RIW-TL estimator is given by

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{\text{ora}}^{(0)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} & \frac{1}{2(n_0 + \sum_{k=1}^K n_k)} \\ & \times \left\{\sum_{k=0}^K \sum_{i=1}^{n_k} \mathbb{I}(i \in \mathcal{I}_k)\omega_i^{(k)} \left[y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}\right]^2\right\} + \lambda\|\boldsymbol{\beta}\|_1, \end{aligned} \tag{8}$$

where λ is a tuning parameter. In other words, the oracle RIW-TL estimator is the solution to a penalized weighted least-squares problem, where the observations from the target data and all source data with bounded weights are incorporated.

Remark 2. In transfer learning, data privacy concerns may arise when the target and source datasets are stored at different sites and cannot be shared directly. In the oracle case, the optimization problem in (8) reduces to a classical LASSO problem. Distributed LASSO algorithms, such as those proposed in Mateos, Bazerque, and Giannakis (2010) and Zeng and Zhang (2022), can be employed to solve this problem without exchanging raw data, thereby offering stronger privacy protection.

When the weights $\omega_i^{(k)}$ and selected sets $\mathcal{I}_k$ in (8) are unknown, a privacy-preserving variant of RIW-TL can still be implemented. Specifically, the initial estimator $\tilde{\boldsymbol{\beta}}^{(0)}$, computed at the target site, can be transmitted to the K source sites. Each source site then locally estimates the weights and identifies the subset $\mathcal{I}_k$. Subsequently, these local statistics can be

aggregated—without revealing raw source data—to compute the final RIW-TL estimator using the same approach as in the oracle case.

2.1. The Choice of f_ϵ in (6)

We note that $\boldsymbol{\beta}^{(0)}$ is not the population minimizer of the expected loss in (8) when $\lambda = 0$, introducing some bias even in the ideal case. While explicitly characterizing this bias is interesting, it is largely irrelevant for establishing the oracle RIW-TL rate due to our proof strategy. Specifically, we use a basic inequality: since $\tilde{\boldsymbol{\beta}}^{(0)}_{\text{ora}}$ minimizes the loss in (8), its loss is no greater than that of $\boldsymbol{\beta}^{(0)}$. This technique is commonly used in analyzing LASSO-type estimators (see Bickel, Ritov, and Tsybakov 2009). For instance, when deriving the rate for the LASSO estimator in (1), a key step is upper-bounding the largest element in $\sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)} \epsilon_i^{(0)}$ with high probability.

We now outline our strategy for ensuring a similar term remains stochastically small. For a random vector $(y_i^{(k)}, (\boldsymbol{x}_i^{(k)})^\top)^\top$ from source k, recall the decomposition:

$$y_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(0)} + (\epsilon_i^{(k)} + \eta_i^{(k)}),$$

where $\epsilon_i^{(k)} + \eta_i^{(k)}$ represents the random error. For the oracle RIW-TL estimator, we apply the basic inequality technique, requiring bounds for $\sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)} \epsilon_i^{(0)}$ (as in LASSO) and K additional terms, wherein the kth term takes the form:

$$\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \xi_i^{(k)}, \quad \xi_i^{(k)} := (\epsilon_i^{(k)} + \eta_i^{(k)}) \omega_i^{(k)} \mathbb{I}(i \in \mathcal{I}_k). \tag{9}$$

Since $\omega_i^{(k)}$ is bounded in Proposition 2, $\xi_i^{(k)}$ remains bounded for fixed $\boldsymbol{x}_i^{(k)}$. The expectation of $\xi_i^{(k)}$ is

$$\begin{aligned}\text{Bias}(f_\epsilon, \mathcal{I}_k) := \mathbb{E}_{\epsilon_i^{(k)}} \xi_i^{(k)} &= \int_{i \in \mathcal{I}_k} (\epsilon_i^{(k)} + \eta_i^{(k)}) f_\epsilon(\epsilon_i^{(k)} + \eta_i^{(k)}) d\epsilon_i^{(k)} \\ &= \int_{-A}^{A} t f_\epsilon(t) dt.\end{aligned}$$

If f_ϵ is symmetric, $\xi_i^{(k)}$ has mean zero for fixed $\boldsymbol{x}_i^{(k)}$, highlighting the advantage of a symmetric pdf for ϵ. Thus, (9) becomes a sum of bounded, mean-zero random variables, which can be stochastically controlled. Since this expectation guides the choice of f_ϵ and $\mathcal{I}_k$, we denote it as $\text{Bias}(f_\epsilon, \mathcal{I}_k)$. Building upon the above findings, we advance the following assumption on f_ϵ.

Condition 2. Assume that f_ϵ is the symmetric density function of a random variable $\epsilon^{(k)}$ that is independent of the source data and satisfies $\mathbb{E}(\epsilon) = 0$.

While f_ϵ can be any symmetric pdf, we focus on the following specific candidates in this article: (a) $f_\epsilon(t) = \left[f_{\epsilon^{(k)}}(t) + f_{\epsilon^{(k)}}(-t)\right]/2$, which is the symmetrized version of $f_{\epsilon^{(k)}}$; (b) More generally, $f_{\epsilon,\theta}(t) = \left[f_{\epsilon^{(k)}}(t/\theta) + f_{\epsilon^{(k)}}(-t/\theta)\right]/(2\theta)$, which arises from a scale family with parameter $1 \le \theta < \infty$; and (c) f_ϵ is the pdf of a uniform distribution. These choices are of particular interest compared to a Gaussian distribution for f_ϵ, though we only present numerical results for the latter.

Remark 3. An alternative sample selection subset is $\mathcal{I}_k' = \{i \in [n_k] : |\epsilon_i^{(k)}| \le A, |\eta_i^{(k)}| \le M_k\}$, for which the bias becomes

$$\begin{aligned}\text{Bias}(f_\epsilon, \mathcal{I}_k') &= \int_{-A}^{A} (\epsilon_i^{(k)} + \eta_i^{(k)}) f_\epsilon(\epsilon_i^{(k)} + \eta_i^{(k)}) \, d\epsilon_i^{(k)} \\ &= \int_{A-\eta_i^{(k)}}^{A+\eta_i^{(k)}} t f_\epsilon(t) \, dt,\end{aligned}$$

which is generally nonzero. When $\eta_i^{(k)}$ is small, the bias is approximately $2A f_\epsilon(A) \eta_i^{(k)}$. However, if the ℓ_1-norm of $\boldsymbol{\delta}^{(k)}$ is small on average, using f_ϵ as the pdf of a uniform distribution and selecting the sample subset $\mathcal{I}_k'$ provides certain advantages, as shown in Section 4.

2.2. Effective Sample Size

The oracle RIW-TL formulation implies that its effective sample size is $n_0 + \sum_{k=1}^K |\mathcal{I}_k|$, where the source data contribute the additional $\sum_{k=1}^K |\mathcal{I}_k|$ observations to the estimation of $\boldsymbol{\beta}^{(0)}$, relative to the LASSO estimator. Since $\mathcal{I}_k$ is a random set, we quantify the gain in sample size by evaluating the expectation of $|\mathcal{I}_k|$. Let $n_{\mathcal{I}_k} = |\mathcal{I}_k|$ represent the cardinality of $\mathcal{I}_k$, and define $\mathcal{I} = \cup_{k=1}^K \mathcal{I}_k$ and $n_{\mathcal{I}} = \sum_{k=1}^K n_{\mathcal{I}_k}$ being the total sample size used for transferring information from the source data. Recall that $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}$, and $\eta^{(k)} = (\boldsymbol{x}^{(k)})^\top(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}) = (\boldsymbol{x}^{(k)})^\top \boldsymbol{\delta}^{(k)}$, consistent with the notation $\eta_i^{(k)}$ in (5). We now proceed to quantify the effective sample size $n_{\mathcal{I}_k}$ from the kth source under the following sub-Gaussian conditions.

Condition 3. For $0 \le k \le K$, the $\epsilon_i^{(k)}$'s are iid sub-Gaussian random variables with mean zero. That is, there exist constants $\sigma^{(k)} > 0$ such that $\mathbb{E}[\exp(u\epsilon_i^{(k)})] \le \exp[u^2(\sigma^{(k)})^2/2]$ for all $u \in \mathbb{R}$.

Proposition 3. Suppose that for $k = 1, \ldots, K$, the covariates $\boldsymbol{x}^{(k)}$ are drawn from a multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma}^{(k)})$, where the eigenvalues of $\boldsymbol{\Sigma}^{(k)}$ satisfy $\kappa_l \le \lambda_{\min}(\boldsymbol{\Sigma}^{(k)}) \le \lambda_{\max}(\boldsymbol{\Sigma}^{(k)}) \le \kappa_u$ for some positive constants $0 < \kappa_l \le \kappa_u < \infty$. Under Condition 3, it holds that

$$\mathbb{E}(n_{\mathcal{I}_k}) \asymp n_k \min\{M_k/d_k, 1\},$$

where $d_k = \|\boldsymbol{\delta}^{(k)}\|$. In particular, when $M_k \gtrsim d_k$, we have $\mathbb{E}(n_{\mathcal{I}_k}) \asymp n_k$.

The bounded eigenvalue condition on the covariance matrix is standard in high-dimensional analysis. Proposition 3 describes how the effective sample size $\mathbb{E}(n_{\mathcal{I}_k})$ depends on M_k and d_k. When $M_k \gtrsim d_k$, the contribution from source k is roughly n_k, while for $M_k \le d_k$, it is $n_k M_k/d_k$, differing from traditional all-in-or-all-out methods.

To assess the contribution of transferrable data from sources to the target, we define the sample usage rate (SUR) as

$$\rho_{\mathcal{I}} := \frac{n_0 + \mathbb{E}(n_{\mathcal{I}})}{n_0 + \sum_{k=1}^K n_k},$$

which measures the proportion of the total data contributing to the knowledge transfer. A higher rate indicates greater use of

source data. In particular, under the conditions of Proposition 3, when $M_k/d_k \gtrsim 1$ for $k = 1, \ldots, K$, it follows that $\rho_{\mathcal{I}} \asymp 1$. This is evident from the fact that $\mathbb{E}(n_{\mathcal{I}}) = \sum_{k=1}^{K} \mathbb{E}(n_{\mathcal{I}_k})$, meaning that the effective sample size of RIW-TL is of the same order as the total sample size.

In practice, we often need the probability that an observation from source data contributes to estimating $\boldsymbol{\beta}^{(0)}$. While general characterization is challenging, we use a simulation for illustration. Recall that the oracle Trans-Lasso retains an observation from source k if $\|\boldsymbol{\delta}^{(k)}\|_1 \le h$, while RIW-TL retains it if $|\eta_i^{(k)}| = |(\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\delta}^{(k)}|$ is small. Since $|\eta_i^{(k)}| \le \|\boldsymbol{\delta}^{(k)}\|_1 \|\boldsymbol{x}_i^{(k)}\|_\infty$, we explore how the probability of using an observation depends on $\|\boldsymbol{\delta}^{(k)}\|_1$ and $\|\boldsymbol{x}_i^{(k)}\|_\infty$.

In the simulation, we set $K = 1$ and generate $\boldsymbol{z}_i^{(0)}$ and $\boldsymbol{z}_i^{(1)}$ from linear models, with $\epsilon_i^{(0)}$ and $\epsilon_i^{(1)}$ iid standard normal. The predictors are $\boldsymbol{x}_i^{(0)} \sim N(\mathbf{0}_p, \boldsymbol{\Sigma})$ and $\boldsymbol{x}_i^{(1)} \sim (1/3)N(-4\mathbf{1}_p, \boldsymbol{\Sigma}) + (1/3)N(\mathbf{0}_p, \boldsymbol{\Sigma}) + (1/3)N(2\mathbf{1}_p, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$. We set $\boldsymbol{\beta}^{(0)} = (\mathbf{1}_5^\top, \mathbf{0}_{95}^\top)^\top$ and $\boldsymbol{\beta}^{(1)} = (\mathbf{1}_5^\top, 0.2\mathbf{1}_l^\top, \mathbf{0}_{95-l}^\top)^\top$, where $l \in \{[5, 45], 5\}$, so that $\|\boldsymbol{\delta}^{(1)}\|_1 = 0.2l \in \{[1, 9], 1\}$, with $\{[a, b], c\}$ denoting the grid points from a to b with step size c.

We partition $\|\boldsymbol{x}^{(1)}\|_\infty$ and $\|\boldsymbol{\delta}^{(1)}\|_1$ into nine intervals as shown in Figure 1. For the constraint set in (7), we set $A = 3/2$ and $M_1 = 3$, defining $\mathcal{I}_1 = \{i \in [n_1] : |\epsilon_i^{(1)} + \eta_i^{(1)}| \le 3/2,\ |\eta_i^{(1)}| \le 3\}$. We generate 100 replicates to estimate the probabilities of an observation belonging to $\mathcal{I}_1$, discretized on a two-dimensional grid. As shown in Figure 1, the probability of retention decreases with increasing $\|\boldsymbol{\delta}^{(1)}\|_1$ or $\|\boldsymbol{x}_i^{(1)}\|_\infty$, and is never smaller than 0.2. This is contrasted with oracle Trans-Lasso, where probabilities are either 1 (if $\|\boldsymbol{\delta}^{(k)}\|$ is small) or 0.

2.3. Properties of the Oracle RIW-TL

For $k = 1, \ldots, K$, let $\mathbf{X}^{(0)} = (\boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_{n_0}^{(0)})^\top \in \mathbb{R}^{n_0 \times p}$ and $\mathbf{X}^{(k)} = (\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_{n_k}^{(k)})^\top \in \mathbb{R}^{n_k \times p}$ denote the design matrices of the target and source k, respectively. Define $\mathbf{I}_{\mathcal{I}_k}$ as the diagonal matrix with the ith diagonal element being the indicator function $\mathbb{I}(i \in \mathcal{I}_k)$. We examine the properties of the oracle estimator defined in (8) under the following condition.

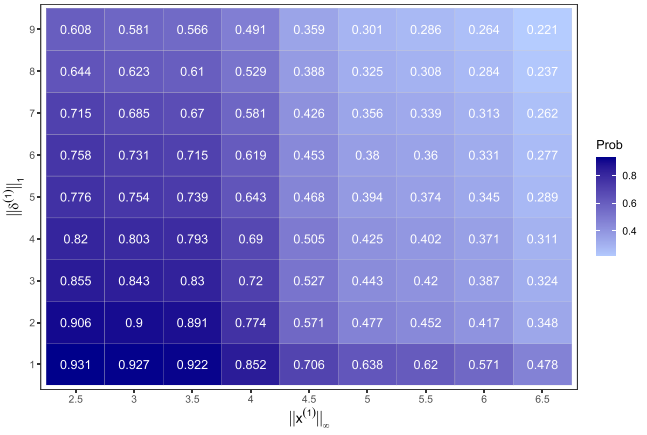

Figure 1. The probability that an observation from the source data is included in RIW-TL.

Condition 4. Let $\mathcal{H}_0 = \mathrm{supp}(\boldsymbol{\beta}^{(0)})$ be the support set of $\boldsymbol{\beta}^{(0)}$. There exists a positive constant ϕ such that

$$\inf_{\boldsymbol{v} \in \mathcal{E}(\mathcal{H}_0, 3)} \frac{\boldsymbol{v}^\top \left[(\mathbf{X}^{(0)})^\top \mathbf{X}^{(0)} + \sum_{k=1}^{K} (\mathbf{X}^{(k)})^\top \mathbf{I}_{\mathcal{I}_k} \mathbf{X}^{(k)} \right] \boldsymbol{v}}{(n_0 + n_{\mathcal{I}}) \|\boldsymbol{v}\|^2} \ge \phi^2,$$

where $\mathcal{E}(\mathcal{H}_0, 3) = \left\{ \boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}_{\mathcal{H}_0^c}\|_1 \le 3\|\boldsymbol{v}_{\mathcal{H}_0}\|_1 \right\}$.

Condition 4 is a mild variant of the restricted eigenvalue (RE) condition. Assuming the covariates $\boldsymbol{x}_i^{(k)}$ are sub-Gaussian, the aggregated Gram matrix $\sum_{k=0}^{K} (\mathbf{X}^{(k)})^\top \mathbf{I}_{\mathcal{I}_k} \mathbf{X}^{(k)}$ can be viewed as being formed by $n_0 + n_{\mathcal{I}}$ iid samples drawn from a mixed sub-Gaussian distribution. Since the RE condition holds in probability under such designs (Negahban et al. 2012), Condition 4 is satisfied with high probability as $n_0 + n_{\mathcal{I}} \to \infty$. More details on the mixed distribution can be seen in Section S.4.2 of the supplementary materials.

Under this assumption, we derive the convergence rate of the oracle RIW-TL estimator.

Theorem 1 (Convergence rate of oracle RIW-TL). Assume Conditions 1–4 are satisfied, and that $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k}$ for $k = 0, \ldots, K$ are fixed. Let $\mathcal{I}_k$ and $\omega_i^{(k)}$ be known. Define $\lambda \asymp \rho_{\mathcal{I}} \sqrt{\log p/(n_0 + \mathbb{E}(n_{\mathcal{I}}))}$, where $\rho_{\mathcal{I}}$ is the sample usage rate. If $\min_{0 \le k \le K} n_k \to \infty$, then the oracle estimator $\tilde{\boldsymbol{\beta}}_{ora}^{(0)}$ satisfies

$$\|\tilde{\boldsymbol{\beta}}_{ora}^{(0)} - \boldsymbol{\beta}^{(0)}\|^2 = O_p\left(\frac{s_0 \log p}{n_0 + \mathbb{E}(n_{\mathcal{I}})} \right).$$

Theorem 1 shows that the convergence rate of RIW-TL in squared ℓ_2-norm scales with $n_0 + \mathbb{E}(n_{\mathcal{I}})$ rather than just n_0, as in LASSO. When $\mathbb{E}(n_{\mathcal{I}}) \gg n_0$, the rate improves significantly. As shown in Proposition 3, $\mathbb{E}(n_{\mathcal{I}}) \asymp \sum_{k=1}^{K} n_k \min\{M_k/d_k, 1\}$, where $d_k = \|\boldsymbol{\delta}^{(k)}\|$. If $M_k \gtrsim d_k$ uniformly over k, the convergence rate becomes $s_0 \log p/(n_0 + \sum_{k=1}^{K} n_k)$, maximizing the use of data across all sources. If $M_k/d_k \ll 1$ uniformly over k, then $\mathbb{E}(n_{\mathcal{I}})$ is of order $\sum_{k=1}^{K} n_k M_k/d_k$, potentially exceeding n_0 if $n_k \gg n_0 d_k/(K M_k)$. Even in this case, the oracle RIW-TL still outperforms LASSO, with the rate depending on d_k.

Comparison of LASSO, Oracle Trans-Lasso, and Oracle RIW-TL. We compare the oracle Trans-Lasso from Li, Cai, and Li (2021) and oracle RIW-TL in a simplified setup with a single source ($K = 1$), assuming M_1 is constant. For simplicity, we set $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{\Sigma}^{(1)}$, ensuring the oracle Trans-Lasso achieves its best rate.

The oracle Trans-Lasso uses optimally chosen tuning parameters with a known support set $\mathcal{A}$, while the oracle RIW-TL assumes known weights $\omega_i^{(k)}$'s and sample subsets $\mathcal{I}_k$'s. We present the convergence rates by calculating the squared ℓ_2-norm of the difference between the oracle estimator and the true parameter $\boldsymbol{\beta}^{(0)}$.

To analyze the behavior of these oracles under different regimes, we categorize the range of h (i.e., $\|\boldsymbol{\delta}^{(1)}\|_1$) into three scenarios, with corresponding rates in Table 1. In the first regime, where $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(1)}$ are close, both methods exhibit the same rate, $s_0 \log p/(n_0 + n_1)$, corresponding to an effective sample size of $n_0 + n_1$. In regime (*ii*), where h increases but remains smaller than $\sqrt{\log p/n_0}$, the oracle Trans-Lasso rate is slower than that

of oracle RIW-TL, which retains the rate $s_0 \log p/(n_0 + n_1)$. In regime (*iii*), where $\sqrt{\log p/n_0} \lesssim h \lesssim s_0\sqrt{\log p/n_0}$, the oracle Trans-Lasso rate is comparable to that of LASSO under the condition $s_0 = O(1)$, while the oracle RIW-TL can still outperform LASSO if $n_1 M_1/d_1 \gg n_0$. Furthermore, when $h \gg s_0\sqrt{\log p}$, the oracle Trans-Lasso becomes ineffective due to the empty informative set, while our method still holds the same rate as that in Regime (*iii*).

Thus, unlike all-in-or-all-out methods like oracle Trans-Lasso, RIW-TL has the advantage of incorporating observations individually according to their weights, making it more flexible, particularly when there is significant heterogeneity between sources and the target. Both methods exploit sparsity in $\boldsymbol{\beta}^{(k)}$ ($k \geq 0$), but in different ways to leverage this information effectively.

Oblivious to Covariate Heterogeneity. The theorem and the subsequent theory hold even when the covariates $\boldsymbol{x}^{(k)}$ follow different distributions across sources. In contrast, methods like Li, Cai, and Li (2021) assume that heterogeneity in the design matrices is small or moderate, and similar assumptions are made in Tian and Feng (2022) for transfer learning in generalized linear models. The RIW-TL method is design-oblivious because it uses importance weights that depend solely on the conditional distribution of $y^{(k)}$ given $\boldsymbol{x}^{(k)}$.

3. RIW-TL in Practice

We have discussed the oracle RIW-TL, which relies on known $\mathcal{I}_k$ for sample selection and $\omega_i^{(k)}$ for the selected observations. In practice, these quantities are rarely available. This section presents a data-driven approach to estimate both sets of unknowns.

Recall the definition of $\mathcal{I}_k$ in (7), involving constraints on $\eta_i^{(k)}$ and $\epsilon_i^{(k)} + \eta_i^{(k)}$. Since $y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(0)} = \epsilon_i^{(k)} + \eta_i^{(k)}$, we estimate them using a preliminary estimate of $\boldsymbol{\beta}^{(0)}$. Similarly, for $\eta_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)})$, we use preliminary estimates of $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(k)}$. For the weight in (6), we replace $\epsilon_i^{(k)}$ with the residual $y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(k)}$, using the preliminary estimate of $\boldsymbol{\beta}^{(k)}$. If $f_{\epsilon^{(k)}}$ is Gaussian, we can estimate its variance, though this may suffer from misspecification. We address the more challenging case of a nonparametric $f_{\epsilon^{(k)}}$, using standard density estimators from the extensive literatures. For the numerator, we define $f_\epsilon(t) = [f_{\epsilon^{(k)}}(t) + f_{\epsilon^{(k)}}(-t)]/2$, making f_ϵ source-dependent. Alternatively, the general scale family $f_{\epsilon,\vartheta}$, as in Section 2.1, can be applied. We use kernel density estimation to estimate $f_{\epsilon^{(k)}}$ and SCAD (Fan and Li 2001) for preliminary estimates of $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(k)}$. The comparison of RIW-TL with different initial estimates can be found in Section S.5.5 of the supplementary materials.

However, directly plugging these estimates into (8) to estimate $\boldsymbol{\beta}^{(0)}$ leads to non-ignorable bias. This issue was also highlighted by Fan, Guo, and Hao (2012) in the context of high-dimensional regression, where using plug-in estimators for variance results in underestimation. To address this, they proposed a refitting procedure, which inspired our cross-fitting strategy outlined below. Specifically, we begin by randomly splitting the data in source k into three (roughly) equal-sized subsets, denoted as $\mathcal{D}_{kj}$ for $k = 0, 1, \ldots, K$ and $j = 1, 2, 3$. The algorithm is formally presented in Algorithm 1.

Algorithm 1 Cross-fitting Algorithm for RIW-TL with Kernel Density Estimation

Input: Target data $\mathcal{S}^{(0)}$ and K source datasets $\{\mathcal{S}^{(k)}\}_{k=1}^K$.

Output: $\hat{\boldsymbol{\beta}}^{(0)}$.

Step 1 (Construct initial estimators): Estimate $\boldsymbol{\beta}^{(k)}$ as $\tilde{\boldsymbol{\beta}}^{(k)}$ using observations indexed by $\mathcal{D}_{k1}$ via SCAD for $k = 0, \ldots, K$.

Step 2 (Estimate the density function of $\epsilon^{(k)}$): Estimate $f_{\epsilon^{(k)}}$ with kernel density estimation on residuals from $\mathcal{D}_{k2}$ as

$$\hat{f}_{\epsilon^{(k)}}(t) = \frac{1}{|\mathcal{D}_{k2}| b_k} \sum_{i \in \mathcal{D}_{k2}} K\left(\frac{|t - \hat{\epsilon}_i^{(k)}|}{b_k}\right), \tag{10}$$

where $K(\cdot)$ is the kernel function with bandwidth b_k for source k, and $\hat{\epsilon}_i^{(k)} = y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \tilde{\boldsymbol{\beta}}^{(k)}$.

Step 3 (Estimate the weights and conduct sample selection): Estimate the weights $\omega_i^{(k)}$ for $i \in \mathcal{D}_{k3}$ as

$$\hat{\omega}_i^{(k)} = \frac{\hat{f}_\epsilon(y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \tilde{\boldsymbol{\beta}}^{(0)})}{\hat{f}_{\epsilon^{(k)}}(y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \tilde{\boldsymbol{\beta}}^{(k)})}, \tag{11}$$

where $\hat{f}_\epsilon(t) = [\hat{f}_{\epsilon^{(k)}}(t) + \hat{f}_{\epsilon^{(k)}}(-t)]/2$ is the estimator for $f_\epsilon(t)$. The estimated subset for $\mathcal{D}_{k3}$ is

$$\hat{\mathcal{I}}_{k3} = \Big\{ i \in \mathcal{D}_{k3} : |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \tilde{\boldsymbol{\beta}}^{(0)}| \leq A,\ |(\boldsymbol{x}_i^{(k)})^\top (\tilde{\boldsymbol{\beta}}^{(k)} - \tilde{\boldsymbol{\beta}}^{(0)})| \leq M_k \Big\}.$$

Step 4 (Construct estimators): Set $\hat{\mathcal{I}}_{03} = \mathcal{D}_{03}$ and $\omega_i^{(0)} = 1$ for $i \in \hat{\mathcal{I}}_{03}$. The final estimator $\hat{\boldsymbol{\beta}}_1^{(0)}$ is given by

$$\hat{\boldsymbol{\beta}}_1^{(0)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_{03}| + \sum_{k=1}^K |\mathcal{D}_{k3}|} \times \left\{ \sum_{k=0}^K \sum_{i \in \mathcal{D}_{k3}} \mathbb{I}\{i \in \hat{\mathcal{I}}_{k3}\} \hat{\omega}_i^{(k)} (y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta})^2 \right\} + \lambda \|\boldsymbol{\beta}\|_1. \tag{12}$$

Step 5 (Alternate datasets $\mathcal{D}_{kj}$'s): Permute the order of $\mathcal{D}_{k1}$, $\mathcal{D}_{k2}$, and $\mathcal{D}_{k3}$ to get estimators $\hat{\boldsymbol{\beta}}_2^{(0)}$ and $\hat{\boldsymbol{\beta}}_3^{(0)}$. The final estimator is

$$\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\boldsymbol{\beta}}_1^{(0)} + \hat{\boldsymbol{\beta}}_2^{(0)} + \hat{\boldsymbol{\beta}}_3^{(0)})/3. \tag{13}$$

In Step 1, preliminary estimates of $\boldsymbol{\beta}^{(k)}$ are obtained using the first data subset. Step 2 estimates the residual densities via kernel density estimation on the second subset. Step 3 constructs the sample selection sets and computes weights for the third subset. In Step 4, a single RIW-TL estimate is obtained, with the regularization parameter λ selected via cross-validation (CV) method. This procedure is repeated across three permutations of the data splits and averaged, as detailed in Step 5. Regarding the selection of bandwidth parameter b_k, we determine it using classical methods such as the plug-in procedure based on data

from source k. This plug-in procedure can be implemented via the function `kde` from the R package `ks`. Additionally, for the truncation parameters $\{A, M_k\}$ used in $\mathcal{I}_k$, we select them by employing the J-fold CV approach. To ease computation, we set $M_k \equiv 2A$ throughout our studies.

Remark 4. To further reduce the computation in selecting hyperparameters, we provide an adaptive rule based on the initial estimators $\tilde{\boldsymbol{\beta}}^{(0)}$ and $\tilde{\boldsymbol{\beta}}^{(k)}$. For the parameter A, since it is theoretically required to be a large constant, we recommend to set it as the 85% quantile of the sequence $\{|y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \tilde{\boldsymbol{\beta}}^{(0)}|\}_{i=1}^{n_k}$ when selecting data from source k. For M_k, we set it as $M_k = c/(1 + \|\tilde{\boldsymbol{\beta}}^{(k)} - \tilde{\boldsymbol{\beta}}^{(0)}\|_1)$ for some constant $c > 0$. Based on our numerical results, we recommend $c = 10$, though one may also select c via cross-validation. Finally, we employ the optimal bandwidth criterion developed by Silverman (1986) to determine b_k in the kernel estimators of the density functions. To validate the effectiveness of the proposed adaptive rules, we conduct simulations, with results reported in Section S.5.4 of the supplementary materials.

3.1. Properties of RIW-TL

To analyze the properties of RIW-TL obtained via cross-fitting, we first need to examine the sample selection subset $\hat{\mathcal{I}}_{kj}$ for each $j = 1, 2, 3$, which estimates

$$\mathcal{I}_{kj} = \{i \in \mathcal{D}_{kj} : |\epsilon_i^{(k)} + \eta_i^{(k)}| \le A,\ |\eta_i^{(k)}| \le M_k\},$$

as well as the properties of the weight estimators in (11). To investigate the properties of $\hat{\mathcal{I}}_{kj}$, we define its "lower bound" as

$$\mathcal{I}_{kj}^- = \{i \in \mathcal{D}_{kj} : |\epsilon_i^{(k)} + \eta_i^{(k)}| \le A - \alpha_n,\ |\eta_i^{(k)}| \le M_k - \alpha_n\},$$

where $\alpha_n = o_p(1)$ (explicit expression is given in (B.1) of the supplementary material). For each $k = 1, \ldots, K$ and $j = 1, 2, 3$, we show that $\mathcal{I}_{kj}^- \subseteq \hat{\mathcal{I}}_{kj}$ with high probability and that $\mathbb{E}(n_{\mathcal{I}_{kj}^-}) \asymp \mathbb{E}(n_{\mathcal{I}_k}) \asymp \mathbb{E}(n_{\hat{\mathcal{I}}_{kj}})$ (see Proposition S1 in the supplementary material). To establish the properties of the RIW-TL estimator, we define $\mu_{\max} = \max_{k\in[K]} \max_{i\in\mathcal{D}_k} \|\boldsymbol{x}_i^{(k)}\|_\infty$ and assume the following conditions.

Condition 5. For each $j = 1, 2, 3$, the RE Condition 4 holds with $\mathcal{I}_k$ replaced by $\mathcal{I}_{kj}^-$ for all k, and $n_{\mathcal{I}}$ replaced by $\sum_{k=1}^K n_{\mathcal{I}_{kj}^-}$.

Condition 6. For $k = 0, 1, \ldots, K$, the initial estimator $\tilde{\boldsymbol{\beta}}^{(k)}$ of $\boldsymbol{\beta}^{(k)}$ is consistent, meaning the ℓ_1 rate of convergence satisfies $\|\tilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}\|_1 = O_p(\gamma_k)$, where $\gamma_k = o(1)$.

Condition 7. The kernel $K(\cdot)$ is symmetric about zero and has a bounded continuous first derivative. The bandwidths b_k satisfies $b_k = o(1)$ and $b_k^2 \gg \mu_{\max}\gamma_k$ for $k = 1, \ldots, K$.

Condition 5 ensures that the restricted eigenvalue condition holds for $\mathcal{I}_{kj}^-$. In Condition 6, $\boldsymbol{\beta}^{(k)}$ can be exactly or approximately sparse. If $\boldsymbol{\beta}^{(k)}$ is exactly sparse with $s_k = |\mathrm{supp}(\boldsymbol{\beta}^{(k)})|$ and $\tilde{\boldsymbol{\beta}}^{(k)}$ is estimated via LASSO or SCAD, then $\gamma_k = s_k\sqrt{\log p/n_k}$. For the debiased LASSO estimator (Zhang and Zhang 2014), we have $\gamma_k = s_k\sqrt{1/n_k}$. In Condition 7, $\mu_{\max}\gamma_k$ accounts for $|\hat{\epsilon}_i^{(k)} - \epsilon_i^{(k)}|$, influenced by the estimation error of $\tilde{\boldsymbol{\beta}}^{(k)}$. Assuming $\mu_{\max}\gamma_k \sim n_k^{-c_1}$, the optimal bandwidth $b_k \asymp n_k^{-1/5}$ applies when $2/5 < c_1 \le 1/2$, whereas under-smoothing is needed for $0 < c_1 \le 2/5$. Thus, if $\tilde{\boldsymbol{\beta}}^{(k)}$ is sufficiently accurate, the standard bandwidth is valid; otherwise, under-smoothing is required.

Define $\tilde{\omega}_i^{(k)} = f_\epsilon(\epsilon_i^{(k)} + \eta_i^{(k)} - r_i^{(k)0})/f_{\epsilon^{(k)}}(\epsilon_i^{(k)})$, where $r_i^{(k)0} = (\boldsymbol{x}_i^{(k)})^\top(\tilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)})$ for all $i \in \mathcal{D}_k$ and $k = 1, \ldots, K$. Under Conditions 1, 6, and 7, it can be shown that $\max_{i\in\hat{\mathcal{I}}_{kj}} |\hat{\omega}_i^{(k)}/\tilde{\omega}_i^{(k)} - 1| = O_p(q_k)$ and $\max_{i\in\hat{\mathcal{I}}_{kj}} |\hat{\omega}_i^{(k)}/\omega_i^{(k)} - 1| = o_p(1)$ for each $j = 1, 2, 3$, where q_k is independent of n_0 and satisfies $q_k \to 0$ as $n_k \to \infty$. Details are referred to in Lemma S2 of the supplementary materials.

Remark 5. A subtle issue arises regarding bias in the estimated subsets $\hat{\mathcal{I}}_{kj}$. For the true subset $\mathcal{I}_k$ in (7), the bias $\mathrm{Bias}(f_\epsilon, \mathcal{I}_k)$ vanishes when f_ϵ is symmetric. However, for $\hat{\mathcal{I}}_{kj}$, estimation errors introduce a small bias. Taking $\hat{\mathcal{I}}_{k3}$ as an example, when f_ϵ is symmetric, the bias simplifies to

$$\begin{aligned}\mathrm{Bias}(f_\epsilon, \hat{\mathcal{I}}_{k3}) &= \mathbb{E}\left[(\epsilon_i^{(k)} + \eta_i^{(k)})\tilde{\omega}_i^{(k)}\mathbb{I}(i \in \hat{\mathcal{I}}_{k3})\right] \\ &= r_i^{(k)0}\int_{-A}^{A} f_\epsilon(t)\,dt := C_{f_\epsilon} r_i^{(k)0},\end{aligned}$$

where $r_i^{(k)0} = (\boldsymbol{x}_i^{(k)})^\top(\tilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)})$ and $C_{f_\epsilon} = \int_{-A}^{A} f_\epsilon(t)\,dt$. This bias depends on $r_i^{(k)0}$, linking the estimation error of $\tilde{\boldsymbol{\beta}}^{(0)}$ to the final convergence rate. If $f_\epsilon = f_{\epsilon,\theta}$, the bias is at most $O(|r_i^{(k)0}|/\theta)$, which is small for large θ, highlighting the advantage of a scale family distribution. Moreover, for any $i \in \hat{\mathcal{I}}_{k3}$, we have

$$\begin{aligned}|r_i^{(k)0}| &\le M_k + |(\boldsymbol{x}_i^{(k)})^\top(\tilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)})| + |(\boldsymbol{x}_i^{(k)})^\top(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)})| \\ &= O_p\left(\mu_{\max}(M_k + \gamma_k + h_k)\right),\end{aligned}$$

where $h_k = \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(0)}\|_1$ and γ_k is the convergence rate of $\tilde{\boldsymbol{\beta}}^{(k)}$.

Combining the errors on weights with the bounds for $\hat{\mathcal{I}}_{kj}$, we establish the convergence of the RIW-TL estimator. Recall $\mu_{\max} = \max_{k\in[K]} \max_{i\in\mathcal{D}_k} \|\boldsymbol{x}_i^{(k)}\|_\infty$. If $\boldsymbol{x}_i^{(k)}$ follows a Gaussian distribution, then $\mu_{\max} = O_p(\sqrt{\log p})$. For bounded predictors in the ℓ_∞ norm, $\mu_{\max} = O(1)$, which is common in applications such as image (Shorten and Khoshgoftaar 2019) and gene expression data (Viñals, Lié, and Bryson 2022). Theorem 2 provides the convergence rate of the RIW-TL estimator computed using the cross-fitting algorithm.

Theorem 2 (Convergence rate of RIW-TL). Suppose that Conditions 1–7 are satisfied and that the covariates $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k}$ for $k = 0, \ldots, K$ are fixed. Furthermore, assume (i) For $k = 1, \ldots, K$, $n_k \gg n_0$, such that $\gamma_k \ll \gamma_0$ and $q_k \ll \gamma_0$, and (ii) $\mu_{\max} = O_p(1)$. Let $\lambda = 2(\lambda_n^{(1)} + \lambda_n^{(2)})$ with

$$\lambda_n^{(1)} \asymp \rho_{\mathcal{I}} \sqrt{\frac{\log p}{n_0 + \mathbb{E}(n_{\mathcal{I}})}},$$

$$\lambda_n^{(2)} \asymp C_{f_\epsilon} \rho_{\mathcal{I}} \left\{ (\pi_0 \gamma_0) \wedge \sum_{k=1}^{K} \pi_{M_k} (M_k + h_k) \right\},$$

where $\pi_{M_k} = \mathbb{E}(n_{\mathcal{I}_k})/(n_0 + \mathbb{E}(n_{\mathcal{I}}))$ depends on M_k, $\pi_0 = \sum_{k=1}^{K} \pi_{M_k}$, $C_{f_\epsilon} = \int_{-A}^{A} f_\epsilon(t)\, dt$, and $\rho_{\mathcal{I}}$ is defined in Section 2.2. As $\min_{0 \le k \le K} n_k \to \infty$, it follows that

$$\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)}\|^2 = O_p \left\{ \frac{s_0 \log p}{n_0 + \mathbb{E}(n_{\mathcal{I}})} + C_{f_\epsilon}^2 s_0 \left((\pi_0 \gamma_0) \wedge \sum_{k=1}^{K} \pi_{M_k} (M_k + h_k) \right)^2 \right\}. \tag{14}$$

The rate above consists of two terms: the first corresponds to the oracle RIW-TL estimator, while the second accounts for the bias arising from sample selection. The term $(\pi_0 \gamma_0) \wedge \sum_{k=1}^{K} \pi_{M_k} (M_k + h_k)$ indicates that our estimator avoids negative transfer. Since $\pi_0 < 1$, it can outperform the initial estimator when the sources are distant, which is particularly useful when the selected sample size is not large. As shown in Proposition 3, when $n_k \equiv n_1$ and $M_k \le d_k$, the term π_{M_k} can be expressed as $\pi_{M_k} := (M_k/d_k)/(n_0 + \sum_{k=1}^{K} (M_k/d_k))$, representing the sampling proportion for source k. Specifically, the larger d_k (or h_k) is, the smaller M_k should be selected to reduce the sampling proportion of source k, thereby minimizing the bias term $\pi_{M_k} h_k$. Consequently, the term $\sum_{k=1}^{K} \pi_{M_k} (M_k + h_k)$ in (14) can be made small through appropriate choices of M_k, particularly when informative sources are available.

In addition, as noted in Remark 5, the constant C_{f_ϵ} satisfies $0 < C_{f_\epsilon} \le 1$ and can be made small by using $f_{\epsilon,\theta}$ with a large θ. Regarding the tuning parameter λ, it depends on the quantities $\mathbb{E}(n_{\mathcal{I}_k})$'s and h_k's that can be effectively estimated by $n_{\hat{\mathcal{I}}_{k2}}$ and $\hat{h}_k = \|\tilde{\boldsymbol{\beta}}^{(k)} - \tilde{\boldsymbol{\beta}}^{(0)}\|$ using the data-splitting procedure. More details can be found at Section S.4.3 in the supplementary materials. In practice, cross-validation is generally applied to select λ.

Remark 6. Assumptions (*i*) and (*ii*) simplify the presentation of Theorem 2. A more general version without these assumptions is given in the supplementary materials. Additionally, if $\epsilon^{(k)}$ follows a Gaussian distribution, kernel density estimation errors vanish, as stated in Proposition S2 of the supplementary material.

Now let us consider the common case where n_0 is small and informative sources exist. For simplicity, we examine $K = 2$, with one informative and another non-informative source; results for general K appear in Corollary S1 of the supplementary material. Recall $h_k = \|\boldsymbol{\delta}^{(k)}\|_1$ and $d_k = \|\boldsymbol{\delta}^{(k)}\|$ for $k \ge 1$. Corollary 1 shows that the RIW-TL estimator attains a faster rate than using only the target data.

Corollary 1. Assume the conditions in Theorem 2 and Proposition 3 hold, along with: (*i*) $n_1 = n_2 = N$; (*ii*) $d_1 \le h_1 \le h \ll \gamma_0 \lesssim d_2 \le h_2$; and (*iii*) $h_2/d_2 \lesssim h_1/d_1$. Let $f_\epsilon = f_{\epsilon,\theta}$ and $M_k \equiv M$ for $k = 1, 2$ with $M = d_1 \wedge d_2$. Then,

$$\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)}\|^2 = O_p \left(\frac{s_0 \log p}{n_0 + N} + \frac{s_0 h^2}{\theta^2} \right), \tag{15}$$

where $1 \le \theta < \infty$ is the scale parameter in $f_{\epsilon,\theta}$.

The conditions in Corollary 1 are mild. Condition (*i*) assumes equal sample sizes for simplicity, but this can easily be extended to the case where n_k's differ. Condition (*ii*) imposes $d_k \le h_k$ for $k = 1, 2$, which is trivial, and $h_1 \le h \ll \gamma_0 \lesssim h_2$, implying that source 1 is informative. Specifically, $h \ll \gamma_0$ is necessary for Trans-Lasso to improve, assuming LASSO is the initial estimator. Condition (*iii*) is mild and holds, for instance, when $h_2 = O(d_2)$. In summary, Corollary 1 shows the rate of RIW-TL always outperforms that of the LASSO.

It is insightful to compare the rate of RIW-TL estimator with the (oracle) Trans-Lasso under the conditions of Corollary 1, where the informative set $\mathcal{A} = \{1\}$ in Li, Cai, and Li (2021). The Trans-Lasso achieves the following rate:

Trans-Lasso Rate:

$$\underbrace{\frac{s_0 \log p}{n_0 + n_{\mathcal{A}}} + \frac{s_0 \log p}{n_0} \wedge (C_\Sigma h) \sqrt{\frac{\log p}{n_0}} \wedge (C_\Sigma h)^2}_{\text{oracle Trans-Lasso}} + \frac{\log K}{n_0}, \tag{16}$$

where $C_\Sigma \ge 1$ measures the similarity between $\boldsymbol{\Sigma}^{(0)}$ and $\boldsymbol{\Sigma}^{(1)}$ and a larger value of C_Σ indicates a greater disparity.

Using the LASSO as the initial estimator and assuming a finite number of sources and that $s_0 = O(1)$, we conclude the following:

- Comparison with oracle Trans-Lasso: When C_Σ is of order $O(1)$, the rate of RIW-TL is similar to that of oracle Trans-Lasso. If $\boldsymbol{\Sigma}^{(0)}$ is dissimilar to $\boldsymbol{\Sigma}^{(k)}$'s, C_Σ can grow large, slowing the oracle Trans-Lasso rate.
- Comparison with Trans-Lasso: If $\gamma_0 = O(\sqrt{\log p/n_0})$, then $h \ll \sqrt{\log p/n_0}$ by condition (*ii*) of Corollary 1. For small n_0, Trans-Lasso has the rate $C_\Sigma^2 h^2 + n_0^{-1}$, while RIW-TL has the rate $s_0 \log p/(n_0 + N) + h^2$, which is similar or better. In particular, if $h \ll n_0^{-1/2}$ or covariate heterogeneity is large, RIW-TL achieves faster convergence than Trans-Lasso.

Moreover, we compare our convergence rate with that of He, Sun, and Li (2024), which achieves the minimax lower bound over the space $\boldsymbol{\Theta} = \{\|\boldsymbol{\beta}^{(0)}\|_0 \le s_0, \|\boldsymbol{\delta}^{(k)}\|_1 \le h_k,\ 1 \le k \le K\}$ under the assumption that $h_k \asymp h_1$ for $k \ge 2$. Our parameter space in (4) is a subspace of $\boldsymbol{\Theta}$ and coincides with $\boldsymbol{\Theta}$ when $\boldsymbol{\delta}^{(k)}$ (for $k \ge 1$) are exactly ℓ_0-sparse—that is, when each $\boldsymbol{\beta}^{(k)}$ is exactly sparse. Notably, the RIW-TL estimator matches the same rate as that of He, Sun, and Li (2024) when $h_1 \lesssim \sqrt{\log p/n_0}$ and $s_0 = O(1)$, even allowing some h_k to be large. In this sense, our method attains the minimax lower bound within our parameter space under mild conditions. More discussions can be found in Section S.4.4 of the supplementary materials.

While RIW-TL has desirable properties, its convergence rate is slower than that of the oracle version in Theorem 1. As shown in Theorem 2, optimal theoretical properties require careful selection of M_k, which can be challenging and time-consuming.

In the next section, we introduce a variant of RIW-TL, where f_ϵ is the density of a uniform distribution and a suitable sample selection set is used. This variant achieves the same convergence rate as the oracle version when the source sample sizes are sufficiently large.

4. An Alternative RIW-TL

This section introduces a variant of RIW-TL, referred to as RIW-TL-U, where the error term ϵ follows a uniform distribution, $\epsilon \sim U[-T, T]$, with density $f_\epsilon(t) = (2T)^{-1}\mathbb{I}(|t| \le T)$ and T as a tuning parameter. We first discuss the convergence rate of the oracle RIW-TL-U with known weights and sample selection subsets, and then consider the rate when these quantities are estimated.

Our findings show that the convergence rate of RIW-TL-U is independent of the rate of $\tilde{\boldsymbol{\beta}}^{(0)}$, as long as it is consistent. This is because only a consistent estimator of the endpoints is required for the numerator in (6), reducing the dependence on n_0. However, an additional bias term appears in the rate, which remains small when h_k's are small on average. Under this setting, RIW-TL-U can achieve the same convergence rate as the oracle version.

The development in this section closely follows the previous ones. However, due to the use of a different distribution for f_ϵ, we introduce new notations for consistency. First, we express the weight associated with data $\boldsymbol{z}_i^{(k)}$ as

$$\omega_{i,T}^{(k)} = \frac{f_\epsilon(\epsilon_i^{(k)} + \eta_i^{(k)})}{f_{\epsilon^{(k)}}(\epsilon_i^{(k)})} = \frac{1}{2Tf_{\epsilon^{(k)}}(\epsilon_i^{(k)})}\mathbb{I}(|\epsilon_i^{(k)} + \eta_i^{(k)}| \le T). \tag{17}$$

For sample selection, we introduce another type of subset:

$$\mathcal{I}_k' = \{i \in \mathcal{D}_k : |\epsilon_i^{(k)}| \le A,\ |\eta_i^{(k)}| \le M_k\}, \qquad k = 1, \dots, K, \tag{18}$$

where A and M_k are parameters such that $A + M_k < T$. Unlike Theorem 2, which involves $M_k + h_k$ in the error rate, the term M_k is removed when f_ϵ follows a uniform distribution, as shown in Theorems 3 and 4. Hence, we set $M_k \equiv M$ for all $k = 1, \dots, K$ in this section. Define $n_{\mathcal{I}_k'} = |\mathcal{I}_k'|$, $\mathcal{I}' = \cup_{k=1}^K \mathcal{I}_k'$, and $n_{\mathcal{I}'} = \sum_{k=1}^K n_{\mathcal{I}_k'}$. The sample usage rate about $\mathcal{I}'$ can be denoted as $\rho_{\mathcal{I}'} = (n_0 + \mathbb{E}(n_{\mathcal{I}'}))/(n_0 + \sum_{k=1}^K n_k)$.

Similar to the oracle estimator $\tilde{\boldsymbol{\beta}}_{\text{ora}}^{(0)}$, we define the oracle RIW-TL-U estimator $\tilde{\boldsymbol{\beta}}_{T,\text{ora}}^{(0)}$ by replacing $(\mathcal{I}_k, \omega_i^{(k)})$ with $(\mathcal{I}_k', \omega_{i,T}^{(k)})$ in (8) and present the convergence rate as follows.

Theorem 3. Assume Conditions 1–4 hold, with fixed covariates $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k}$ for $k = 0, \dots, K$. Suppose the sets $\mathcal{I}_k'$ and weights $\omega_{i,T}^{(k)}$ are known for all $i \in \mathcal{I}_k'$, and that $\mu_{\max} = O_p(1)$. Let $\lambda = 2(\lambda_n^{(1)} + \lambda_n^{(2)})$ where

$$\lambda_n^{(1)} \asymp \rho_{\mathcal{I}'}\sqrt{\frac{\log p}{n_0 + \mathbb{E}(n_{\mathcal{I}'})}}, \quad \lambda_n^{(2)} \asymp C_{A,T}\rho_{\mathcal{I}'}\sum_{k=1}^K \pi_M^{(k)} h_k,$$

with $\pi_M^{(k)} = \mathbb{E}(n_{\mathcal{I}_k'})/(n_0 + \mathbb{E}(n_{\mathcal{I}'}))$ and $C_{A,T} = A/T$. As $\min_{0\le k\le K} n_k \to \infty$, we have

$$\|\tilde{\boldsymbol{\beta}}_{T,\text{ora}}^{(0)} - \boldsymbol{\beta}^{(0)}\|^2 = O_p\left\{\frac{s_0 \log p}{n_0 + \mathbb{E}(n_{\mathcal{I}'})} + C_{A,T}^2 s_0\left(\sum_{k=1}^K \pi_M^{(k)} h_k\right)^2\right\}.$$

If $M \lesssim \gamma_0$, we can replace $\sum_{k=1}^K \pi_M^{(k)} h_k$ by $(\pi_0'\gamma_0) \wedge \sum_{k=1}^K \pi_M^{(k)} h_k$, where $\pi_0' = \sum_{k=1}^K \pi_M^{(k)} < 1$, avoiding negative transfer.

Compared to the oracle RIW-TL in Theorem 1, the oracle RIW-TL-U introduces an additional term depending on $h_{\text{ave}} = \sum_{k=1}^K \pi_M^{(k)} h_k$. When $h_{\text{ave}} \lesssim \sqrt{\log p/(n_0 + \mathbb{E}(n_{\mathcal{I}'}))}$, the rate achieves the same order as the oracle RIW-TL in Theorem 1. Otherwise, the oracle RIW-TL-U performs worse than the oracle RIW-TL.

When the weights are unknown and the sets $\mathcal{I}_k'$ are unobserved, we estimate the weights using a cross-fitting procedure as follows:

$$\hat{\omega}_{i,T}^{(k)} = \frac{1}{2T\hat{f}_{\epsilon^{(k)}}(\hat{\epsilon}_i^{(k)})}\mathbb{I}(|\hat{\epsilon}_i^{(k)} + \hat{\eta}_i^{(k)}| \le T), \quad i \in \mathcal{D}_{k3},$$

where the denominator is estimated in the same way as (10). The estimated sample selection sets are given by

$$\hat{\mathcal{I}}_{k3}' = \{i \in \mathcal{D}_{k3} : |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \tilde{\boldsymbol{\beta}}^{(k)}| \le A,\ |\hat{\eta}_i^{(k)}| \le M\}.$$

Here, A and M are parameters such that $A + M - 2\theta_0 \le T$, where $\theta_0 > 0$ is a small constant. This condition ensures that both the weights $\omega_{i,T}^{(k)}$ and their estimates $\hat{\omega}_{i,T}^{(k)}$ are nonzero with high probability for all $i \in \hat{\mathcal{I}}_{k3}'$. It can be shown that $\max_{i\in\hat{\mathcal{I}}_{kj}'} |\hat{\omega}_{i,T}^{(k)}/\omega_{i,T}^{(k)} - 1| = O_p(q_k)$, where q_k is independent of n_0 and satisfies $q_k \to 0$ as $n_k \to \infty$. Details are provided in Lemma S5 of the supplementary materials.

Similar to $\hat{\boldsymbol{\beta}}_1^{(0)}$ in (12), we define the estimator $\hat{\boldsymbol{\beta}}_{T,1}^{(0)}$ by replacing $(\hat{\mathcal{I}}_{k3}, \hat{\omega}_i^{(k)})$ with $(\hat{\mathcal{I}}_{k3}', \omega_{i,T}^{(k)})$. Similarly, we define $\hat{\boldsymbol{\beta}}_{T,2}^{(0)}$ and $\hat{\boldsymbol{\beta}}_{T,3}^{(0)}$. The overall RIW-TL-U estimator is $\hat{\boldsymbol{\beta}}_T^{(0)} = (\hat{\boldsymbol{\beta}}_{T,1}^{(0)} + \hat{\boldsymbol{\beta}}_{T,2}^{(0)} + \hat{\boldsymbol{\beta}}_{T,3}^{(0)})/3$, with its convergence rate established in Theorem 4.

Theorem 4 (Convergence rate of RIW-TL-U). Suppose the conditions of Theorem 3 and Conditions 5–7 hold. Additionally, assume: for $k = 1, \dots, K$, $n_k \gg n_0$ such that $\gamma_k \ll \gamma_0$ and $q_k \ll \gamma_0$. Let $\lambda = 2(\lambda_n^{(1)} + \lambda_n^{(2)})$, where

$$\lambda_n^{(1)} \asymp \rho_{\mathcal{I}'}\sqrt{\frac{\log p}{n_0 + \mathbb{E}(n_{\mathcal{I}'})}}, \quad \lambda_n^{(2)} \asymp C_{A,T}\rho_{\mathcal{I}'}\sum_{k=1}^K \pi_M^{(k)} h_k,$$

with $\pi_M^{(k)} = \mathbb{E}(n_{\mathcal{I}_k'})/(n_0 + \mathbb{E}(n_{\mathcal{I}'}))$ and $C_{A,T} = A/T$. As $\min_{k\in[K]} n_k \to \infty$, for any fixed $M > 0$, it follows that

$$\|\hat{\boldsymbol{\beta}}_T^{(0)} - \boldsymbol{\beta}^{(0)}\|^2 = O_p\left\{\frac{s_0 \log p}{n_0 + \mathbb{E}(n_{\mathcal{I}'})} + C_{A,T}^2 s_0\left(\sum_{k=1}^K \pi_M^{(k)} h_k\right)^2\right\}.$$

In particular, if $M \lesssim \gamma_0$, the term $\sum_{k=1}^K \pi_M^{(k)} h_k$ can be replaced by $(\pi_0'\gamma_0) \wedge \sum_{k=1}^K \pi_M^{(k)} h_k$.

Theorem 4 provides the convergence rate for the RIW-TL-U estimator. When the sample sizes n_k's ($k \geq 1$) are sufficiently large such that both q_k and γ_k tend to zero, the convergence rate of RIW-TL-U matches that of the oracle estimator in Theorem 3. Additionally, under the conditions of Corollary 1, the RIW-TL-U estimator can achieve a convergence rate similar to that in (15), with full details given in Corollary S2 of the supplementary materials.

We now compare the convergence rates of RIW-TL-U and RIW-TL when the sample sizes n_k ($k \geq 1$) are large, causing both γ_k and q_k to converge to zero. In this case, Theorem 2 shows that the error for RIW-TL is of order $[(\pi_0\gamma_0) \wedge \sum_{k=1}^K \pi_{M_k}(M_k+h_k)]^2$, while Theorem 4 gives the error for RIW-TL-U as $(\sum_{k=1}^K \pi_M^{(k)} h_k)^2$.

First, RIW-TL is more robust against negative transfer than RIW-TL-U. It guarantees an error no worse than γ_0^2, independent of M_k and h_k. On the other hand, RIW-TL-U can only achieve this error bound when either $M_k \lesssim \gamma_0$ (from Theorem 4) or when $\sum_{k=1}^K \pi_M^{(k)} h_k \lesssim \gamma_0$. Second, RIW-TL-U is more efficient than RIW-TL when h_k's are small but M_k's are large. In this case, while RIW-TL may incur an error rate of γ_0^2, the RIW-TL-U estimator maintains a more favorable error rate.

5. Simulation

In this section, we numerically compare RIW-TL with Trans-Lasso and LASSO using synthetic data. Specifically, we implement the following methods: RIW-TL as described in Section 3, RIW-TL-U from Section 4, and a variant of RIW-TL, denoted RIW-TL-P. In RIW-TL-P, we assume that $\epsilon^{(k)}$ follows a Gaussian distribution, with its variance estimated from the sample, and f_ϵ is the symmetrization of the estimated $f_{\epsilon^{(k)}}$.

In our setup, we set $p = 200$, $n_0 = 150$, $K = 10$, and $n_k = 600$ for $k = 1, \ldots, K$. Simulation results for a broader range of n_0, n_k, and p are provided in the supplementary materials. The errors $\epsilon_i^{(0)}$ and $\epsilon_i^{(k)}$ are independently generated from the standard normal distribution, and the response-predictor pairs are generated according to the linear model. To simulate realistic scenarios with varying informativeness of the source data, we consider different configurations for generating $\boldsymbol{x}_i^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ for $k = 0, \ldots, K$.

Covariates. The source data covariates $\boldsymbol{x}_i^{(k)}$ ($k = 1, \ldots, K$) are generated from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = (0.5^{|l-l'|})_{l,l'=1}^p$. For the target data covariates $\boldsymbol{x}_i^{(0)}$, we generate them either from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ or from a multivariate central t-distribution with covariance matrix $\boldsymbol{\Sigma}$ and five degrees of freedom. In the former case, the marginal distribution of the covariates is identical for both the source and target datasets. However, the conditional distribution of y given $\boldsymbol{x}$ differs between the source and target datasets if $\boldsymbol{\beta}^{(k)} \neq \boldsymbol{\beta}^{(0)}$, commonly referred to as *posterior shift*. In the latter case, both the marginal distribution of $\boldsymbol{x}$ and the conditional distribution of y given $\boldsymbol{x}$ differ between the source and target datasets, referred to as *full distribution shift*. We will analyze these two types of shifts in separate plots.

Coefficients. For the target data, we set $\boldsymbol{\beta}^{(0)} = (\mathbf{1}_{s_0}^\top, \mathbf{0}_{p-s_0}^\top)^\top$ with $s_0 = 10$. For the source data, we vary $\boldsymbol{\beta}^{(k)}$ based on the number and magnitude of differing entries from $\boldsymbol{\beta}^{(0)}$. Following Li, Cai, and Li (2021), we define an index set $\mathcal{B} = [m_{\mathcal{B}}]$, where $m_{\mathcal{B}} \in \{0, 2, 4, 6, 8, 10\}$, and specify $\boldsymbol{\beta}^{(k)}$ as follows:

- If $k \in \mathcal{B}$, set $\beta_j^{(k)} = \beta_j^{(0)} - 0.5$ for $j \in T_k$ and otherwise $\beta_j^{(k)} = \beta_j^{(0)}$, where T_k is a random subset of $\{s_0 + 1, \ldots, p\}$ with $|T_k| = d$ to be set later;
- If $k \notin \mathcal{B}$, set $\beta_j^{(k)} = \beta_j^{(0)} - 1$ for $j \in [s_0]$, $\beta_j^{(k)} = \beta_j^{(0)} - 0.5$ for $j \in U_k$, and otherwise $\beta_j^{(k)} = \beta_j^{(0)}$, where U_k is a random subset of $\{s_0 + 1, \ldots, p\}$ with $|U_k| = 2s_0$.

In the above configuration, the magnitude of the difference between the entries of $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(k)}$ is fixed at either 0.5 or 1, while the number of differing entries depends on the values of $m_{\mathcal{B}}$ and d. Generally, a larger value of d corresponds to a greater difference between $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(k)}$. On the other hand, the values of $\boldsymbol{\beta}^{(k)}$ for $k \in \mathcal{B}$ are relatively close to $\boldsymbol{\beta}^{(0)}$, and as $m_{\mathcal{B}}$ increases, a larger number of sources are relatively close to the target. To ensure the generality of our method, we also consider cases where the magnitude of the differences is random, and where $\epsilon_i^{(0)}$ and $\epsilon_i^{(k)}$ follow different distributions. The simulation results for these cases, which are similar to those presented here, can be found in Section S.5.1 of the supplementary materials.

For the tuning parameters appearing in the constraints of $\mathcal{I}_k$ in RIW-TL, we set $M = 2A$ and tune M on a grid with a step size of 0.5 in the interval $[1, 3]$. The bandwidth parameters h_k's are determined by the strategy introduced prior to Remark 4. For RIW-TL-U, we set $M = 2A = 2T/3$ and select M via cross-validation from the interval $[1, 3]$ with a step size of 0.5.

To evaluate performance, we compute the sum of squared errors (SSE), defined as $\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\|^2$ for any estimator $\check{\boldsymbol{\beta}} \in \mathbb{R}^p$, and the sample usage rate (SUR). Since the Trans-Lasso approach lacks a formal definition of SUR, we report only the SURs of our methods and the LASSO for clarity. We consider two simulation scenarios based on the parameters d and $m_{\mathcal{B}}$. In the first scenario, we vary $m_{\mathcal{B}} \in \{0, 2, 4, 6, 8, 10\}$ while fixing d at 4 or 8. In the second scenario, we fix $m_{\mathcal{B}}$ at 4 or 8 and vary d on a grid from 0 to 32. For each configuration, we repeat the experiment 200 times and report the average SSE and SUR. Results for posterior shift are presented in Figure 2, and for full distribution shift in Figure 3. Based on these results, we draw the following conclusions.

LASSO is inferior to transfer learning methods. Figures 2 and 3 show that LASSO consistently exhibits the worst SSE and smallest SUR, as it does not use source data. In contrast, transfer learning methods outperform LASSO by leveraging source data. Performance improves as $m_{\mathcal{B}}$ increases or d decreases, as more source data becomes similar to the target.

RIW-TL estimators outperform Trans-Lasso, especially when d is large or $m_{\mathcal{B}}$ is small. Smaller $m_{\mathcal{B}}$ or larger d suggests fewer source datasets are similar to the target, often resulting in larger $\|\boldsymbol{\delta}^{(k)}\|_1$. Figures 2 and 3 highlight that RIW-TL and its variants (RIW-TL-P, RIW-TL-U) significantly outperform Trans-Lasso in these scenarios, confirming our theory that the method uses sample data effectively even when $\|\boldsymbol{\delta}^{(k)}\|_1$'s are large. RIW-TL-U, in particular, outperforms all others, likely due to its weaker conditions on initial estimators.

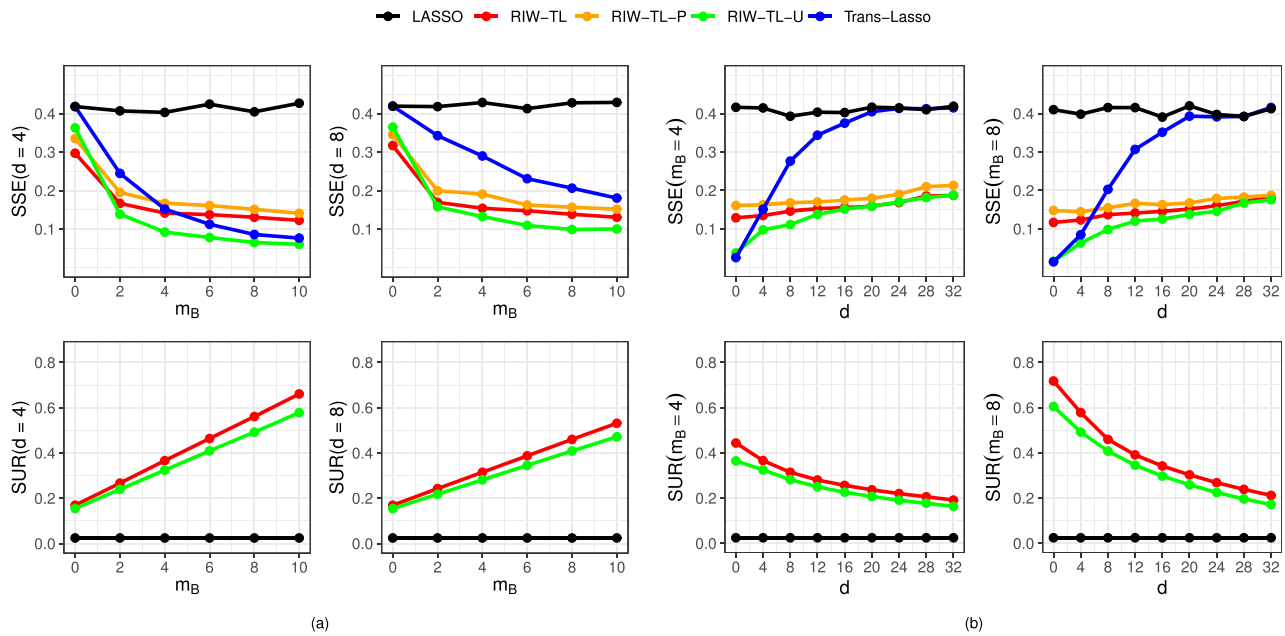

Figure 2. Posterior Shift: (a) Estimation errors (top row) and sample usage rates (bottom row) versus $m_{\mathcal{B}}$ for different d values. (b) Estimation errors (top row) and sample usage rates (bottom row) versus d for different $m_{\mathcal{B}}$ values. In the SUR plots, RIW-TL-P lines overlap with those of RIW-TL and are not visible.

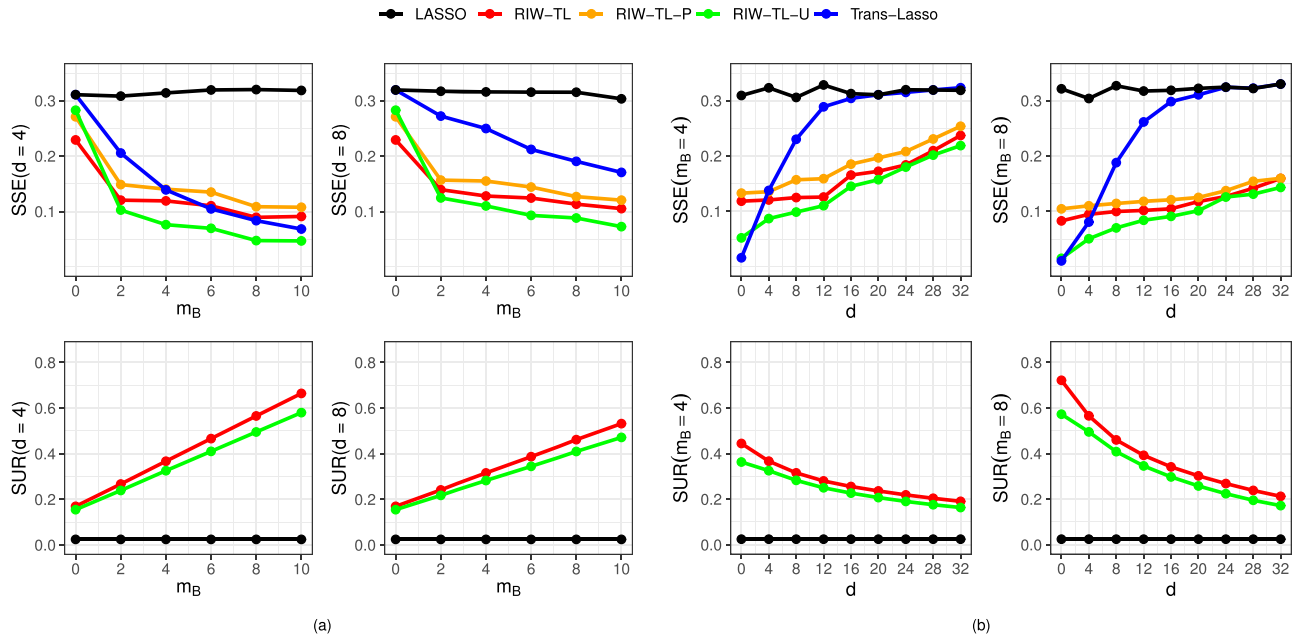

Figure 3. Full Distribution Shift: (a) Estimation errors (top row) and sample usage rates (bottom row) versus $m_{\mathcal{B}}$ for different d values. (b) Estimation errors (top row) and sample usage rates (bottom row) versus d for different $m_{\mathcal{B}}$ values. In the SUR plots, RIW-TL-P lines overlap with RIW-TL and are not visible.

RIW-TL type estimators can still leverage some information in difficult cases. As shown in the second rows of Figures 2 and 3, the SURs of RIW-TL type methods increase with $m_{\mathcal{B}}$ and decrease with d. Specifically, in difficult cases where $m_{\mathcal{B}} = 0$ or d is large (indicating a significant difference between sources and target), RIW-TL type methods can still make use of some data from sources (i.e., larger SURs than the LASSO) and thereby achieve a smaller SSE than the LASSO method.

Assuming a parametric distribution for $f_{\epsilon^{(k)}}$ does not offer significant advantages. In both the simulations here and in the supplementary materials, RIW-TL-P tends to underperform compared to RIW-TL and RIW-TL-U, even when the error distributions of $\epsilon^{(k)}$ are Gaussian. This suggests that the univariate density estimation in RIW-TL and RIW-TL-U is already accurate, or that symmetrizing $f_{\epsilon^{(k)}}$ to obtain f_{ϵ} is not effective for Gaussian errors. Therefore, we recommend using RIW-TL or RIW-TL-U in practice.

6. Real Data Analysis

In this section, we apply our method to the Genotype-Tissue Expression (GTEx) dataset, available at https://gtexportal.org/. This choice is motivated by Li, Cai, and Li (2021), where Trans-Lasso was shown to enhance LASSO prediction performance.

The GTEx dataset includes gene expression levels for 38,187 genes across 49 tissues from 838 human donors. We focus on predicting the expression of the JAM2 gene (Junctional adhesion molecule B) in brain tissue, using other genes from the central nervous system (CNS) as predictors. Mutations in JAM2 are linked to primary familial brain diseases (Cen et al. 2020). JAM2 expression is measured across all 49 tissues, and we select the 40 tissues with more than 150 measurements of JAM2 expression.

We treat the relationship between JAM2 and other CNS genes in each of the nine brain tissues as a separate target model. A complete list of target tissues and their sample sizes are given Table S1 in the supplementary materials. Specifically, we analyze nine target models, each using the remaining 31 tissues as source models, with a total source sample size of 12,386. The covariates are derived from genes in the enriched MODULE_137 pathway. There are no missing values in any of the 40 tissues, resulting in 1089 covariates.

After obtaining the initial estimator $\tilde{\boldsymbol{\beta}}^{(k)}$ for each source tissue using SCAD, we test the normality of the residuals. Based on the p-values in Figure S6 of the supplementary materials, most source tissues show p-values less than 0.05, suggesting that the normality assumption for $\epsilon^{(k)}$ is not appropriate. Therefore, we focus on RIW-TL and RIW-TL-U. The tuning scheme from Section 5 is applied. For a reliable comparison, we randomly split the target tissue data into training and testing sets with a ratio 7:3 and compute the average prediction errors over 100 replications. The relative prediction error (RPE) is evaluated against LASSO, with smaller values indicating better performance. As demonstrated in Figure 4, across different target tissues, RIW-TL, RIW-TL-U, and Trans-Lasso exhibit superior performance compared to LASSO. Their relative prediction errors remain consistently below 1, with average relative improvements reaching 29.8% for Trans-Lasso, 38.7% for RIW-TL, and 43.4% for RIW-TL-U.

Additionally, we compare the performance of these methods in terms of selected variables. Define the following metrics relative to the LASSO estimator $\hat{\boldsymbol{\beta}}_{\text{Lasso}} = (\hat{\beta}_{L,1}, \ldots, \hat{\beta}_{L,p})^\top$:

$$\text{S} = \frac{1}{p}\#\{j : \check{\beta}_j \neq 0\}, \quad \text{PR} = \frac{\#\{j : \check{\beta}_j \neq 0, \hat{\beta}_{L,j} \neq 0\}}{\#\{j : \hat{\beta}_{L,j} \neq 0\}},$$

$$\text{NR} = \frac{\#\{j : \check{\beta}_j \neq 0, \hat{\beta}_{L,j} = 0\}}{\#\{j : \hat{\beta}_{L,j} = 0\}}.$$

Here, S represents the sparsity rate, PR is the positive rate (percentage of nonzero coefficients in LASSO correctly estimated), and NR is the negative rate (percentage of zero coefficients in LASSO incorrectly estimated as nonzero).

The results are summarized in Table 2, including the relative prediction error defined earlier. From the table, we observe that both RIW-TL and RIW-TL-U, as well as Trans-Lasso, select more variables than LASSO. This is expected, as transfer learning methods increase the sample size by incorporating data from source tissues, facilitating the identification of more variables through penalized likelihood. When comparing the two RIW-TL methods (RIW-TL and RIW-TL-U) to Trans-Lasso, we find that our methods typically produce less sparse models with higher positive and negative rates. Overall, RIW-TL and RIW-TL-U achieve smaller relative prediction errors across all target tissues compared to LASSO and Trans-Lasso, with RIW-TL-U yielding the smallest errors, except for C.B.ganglia.

7. Discussion

We introduced RIW-TL, a novel transfer learning method for high-dimensional linear regression, which uniquely weights residuals based on their importance. RIW-TL enhances transfer precision by applying individualized weighting, simplifies implementation through univariate density estimation, is robust to covariance heterogeneity, and outperforms competing methods in both oracle and practical settings.

Our framework opens several exciting avenues for future research. We aim to extend RIW-TL to quantile regression to handle heterogeneity and heavy-tailed distributions in source and target domains. Additionally, we envision extending RIW-TL to generalized linear models and applying it in semi-supervised and unsupervised learning. These directions and further generalizations of RIW-TL will be explored in future work.

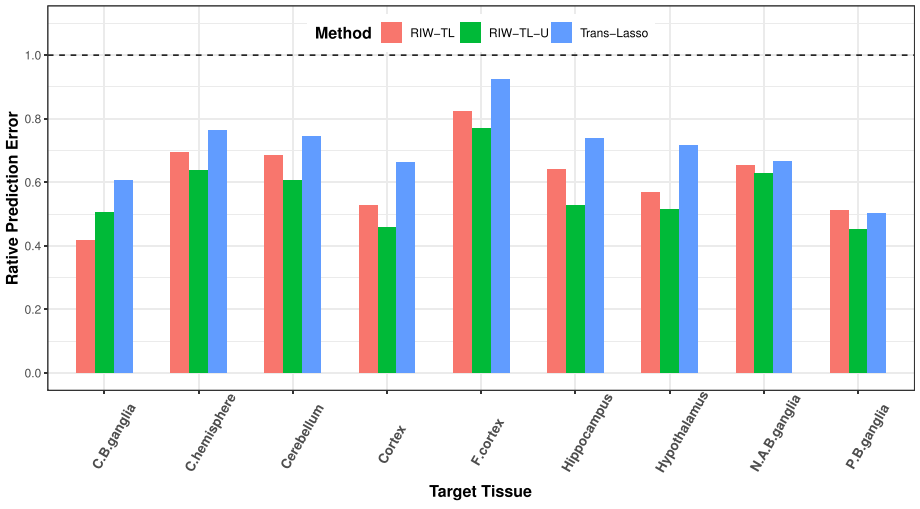

Figure 4. The prediction errors of RIW-TL-type methods and the Trans-Lasso relative to that of LASSO across nine different target tissues.

Table 2. Further results on variable selection and the relative prediction error (RPE) in data analysis.

Method	S	PR	NR	RPE	S	PR	NR	RPE	S	PR	NR	RPE
	C.B.gangli				C.hemisphere				Cerebellu			
RIW-TL	0.312	0.794	0.468	0.416	0.252	0.379	0.248	0.695	0.812	0.880	0.810	0.283
RIW-TL-U	0.378	0.676	0.313	0.505	0.429	0.690	0.422	0.637	0.813	0.840	0.812	0.206
Trans-Lasso	0.144	0.365	0.140	0.605	0.121	0.138	0.121	0.764	0.124	0.321	0.119	0.344
Lasso	0.031	1	0	1	0.027	1	0	1	0.023	1	0	1
	Cortex				F.cortex				Hippocampus			
RIW-TL	0.253	0.619	0.245	0.527	0.565	0.663	0.557	0.824	0.511	0.643	0.507	0.640
RIW-TL-U	0.529	0.619	0.527	0.458	0.590	0.687	0.582	0.768	0.594	0.714	0.591	0.527
Trans-Lasso	0.129	0.571	0.120	0.661	0.142	0.181	0.139	0.922	0.096	0.214	0.092	0.739
Lasso	0.019	1	0	1	0.076	1	0	1	0.026	1	0	1
	Hypothalamus				N.A.B.ganglia				P.B.gang			
RIW-TL	0.619	0.516	0.622	0.567	0.307	0.357	0.304	0.654	0.496	0.567	0.494	0.511
RIW-TL-U	0.775	0.613	0.780	0.514	0.694	0.750	0.797	0.629	0.587	0.767	0.788	0.450
Trans-Lasso	0.123	0.226	0.120	0.715	0.111	0.196	0.106	0.664	0.120	0.200	0.118	0.502
Lasso	0.028	1	0	1	0.051	1	0	1	0.028	1	0	1

Supplementary Materials

The supplementary material provides all the technical proofs, methodological details, additional numerical results, and the program code for all the results.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was supported by the National Natural Science Foundation of China (No.12371288, 12131006), the Fundamental Research Funds for the Central Universities.

ORCID

Chenlei Leng https://orcid.org/0000-0001-5703-9617

References

Bastani, H. (2021), "Predicting with Proxies: Transfer Learning in High Dimension," *Management Science*, 67, 2964–2984. [2]

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [1,5]

Cai, T., Li, M., and Liu, M. (2024), "Semi-Supervised Triply Robust Inductive Transfer Learning," *Journal of the American Statistical Association*, 120, 1037–1047. [2]

Cai, T. T., and Pu, H. (2024), "Transfer Learning for Nonparametric Regression: Non-Asymptotic Minimax Analysis and Adaptive Procedure," arXiv preprint arXiv:2401.12272. [2]

Cen, Z., Chen, Y., Chen, S., Wang, H., Yang, D., Zhang, H., et al. (2020), "Biallelic Loss-of-Function Mutations in JAM2 Cause Primary Familial Brain Calcification," *Brain*, 143, 491–502. [13]

Cortes, C., Mansour, Y., and Mohri, M. (2010), "Learning Bounds for Importance Weighting," in *Advances in Neural Information Processing Systems 24 (NIPS 2010)* (Vol. 1(9)), pp. 442–450. [3]

Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 74, 37–65. [7]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American statistical Association*, 96, 1348–1360. [7]

He, Z., Sun, Y., and Li, R. (2024), "Transfusion: Covariate-Shift Robust Transfer Learning for High-Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 238, 703–711. [2,9]

Li, S., Cai, T., and Duan, R. (2023), "Targeting Underrepresented Populations in Precision Medicine: A Federated Transfer Learning Approach," *The Annals of Applied Statistics*, 17, 2970–2992. [2]

Li, S., Cai, T. T., and Li, H. (2021), "Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality," *Journal of the Royal Statistical Society*, Series B, 84, 149–173. [2,3,6,7,9,11,12]

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. [2]

Lu, N., Zhang, T., Fang, T., Teshima, T., and Sugiyama, M. (2022), "Rethinking Importance Weighting for Transfer Learning," in *Federated and Transfer Learning* (Vol. 27), pp. 185–231, Cham: Springer. [2]

Ma, C., Pathak, R., and Wainwright, M. J. (2023), "Optimally Tackling Covariate Shift in RKHS-based Nonparametric Regression," *The Annals of Statistics*, 51, 738–761. [2]

Mateos, G., Bazerque, J. A., and Giannakis, G. B. (2010), "Distributed Sparse Linear Regression," *IEEE Transactions on Signal Processing*, 58, 5262–5276. [4]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of M-estimators with Decomposable Regularizers," *Statistical Science*, 27, 538–557. [6]

Shorten, C., and Khoshgoftaar, T. (2019), "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, 6, 1–48. [8]

Silverman, B. (1978), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives," *The Annals of Statistics*, 6, 177–184. [3]

——— (1986), *Density Estimation for Statistics and Data Analysis* (Vol. 26), Boca Raton, FL: CRC Press. [8]

Tian, Y., and Feng, Y. (2022), "Transfer Learning under High-Dimensional Generalized Linear Models," *Journal of the American Statistical Association*, 118, 2684–2697. [2,3,7]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Viñals, R., Lié, P., and Bryson, K. (2022), "Adversarial Generation of Gene Expression Data," *Bioinformatics*, 38, 730–737. [8]

Zeng, W., and Zhang, R. (2022), "A Distributed Algorithm for Lasso Variable Selection," *Chinese Journal of Applied Probability and Statistics*, 38, 99–110. [4]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [8]

Zhang, Y., and Zhu, Z. (2025), "Transfer Learning for High-Dimensional Quantile Regression via Convolution Smoothing," *Statistical Sinica*, 35, 939–958. [2]

Zhou, D., Liu, M., Li, M., and Cai, T. (2024), "Doubly Robust Augmented Model Accuracy Transfer Inference with High Dimensional Features," *Journal of the American Statistical Association*, 120, 524–534. [2]

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., and Zhu, H. (2021), "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, 109, 43–76. [2]