Partial Transfer Learning Under High-Dimensional Confounding: Estimation, Prediction, and Efficiency

Xinhao Qu[1]

(School of Economics, Xiamen University, Xiamen 361005, China)

Abstract Knowledge transfer is of great importance as it implements the target referable information that would ultimately enhance model performance. To this end, I propose a learning framework based on partial transferring of parameters. The setting is under linear regression with high-dimensional confounding, where algorithms for source detection, parameter estimation and response prediction through either homogeneous or heterogeneous sources are designed. The Partial Transfer Learning (PTL) methodology utilizes Double/Debiased Machine Learning, oracle penalty such as SCAD and cross-fitting technique, whose estimation is demonstrated consistent and asymptotically normal. Further more, I present that PTL owns enhanced efficiency for tasks of estimation and prediction. Simulations showcase the detection efficiency and the superior performance of PTL estimators. Empirical research underscores the practical applicability of PTL, particularly in the realm of biological data analysis.

Keywords Partial Transfer Learning; Double/Debiased Machine Learning; Causal Inference

1 Introduction

Transfer Learning plays a pivotal role in advancing scientific analyses across various domains. When viewed through the lens of machine learning, it mirrors the cognitive process of knowledge transfer, resembling the decision-making mechanism of the human brain. This inherent characteristic holds great promise for its application in machine learning, particularly in neural networks (Pan and Yang, 2009).

From a data science perspective, Transfer Learning entails the utilization of cost-effective and readily accessible labels prior to modeling. This practice significantly enhances the precision and efficiency of target estimation, addressing issues of incomplete data, especially in domains like clinical and biological data with high-dimensional feasures. Related applications on data fusion process combines integrative analysis, including Wang *et al.* (2023) recently, and Wang *et al.* (2019) on omics data across multiple cancer types, an overview of such analyses is given by Ma *et al.* (2022), which also initiates my empirical analysis on GTEx database (`https://gtexportal.org/`).

However, from a social science standpoint, the transferability of these techniques may be questionable if not handled with care. This concern represents a major limitation within the domain of Transfer Learning. Traditional machine learning literature, as exemplified by Weiss *et al.* (2016), often permits ad hoc transfers, leading to challenges and a lack of robustness in broader scenarios.

Addressing the issue of transferability, scholars have proposed various approaches. Li *et al.* (2022a) introduced a statistical benchmark in the form of a smoothing and de-biased framework for Transfer Learning in High-Dimensional Linear Regression, which minimizes the transfer gap through regularization. Additionally, researchers have explored different model settings, including high-dimensional generalized linear regression (Tian and Feng, 2022; Li *et al.*, 2023b), and the Gaussian graphical model (Li *et al.*, 2022b).

Yet, even within this body of literature, the question of transferability remains contentious, with algorithms varying across different settings. This debate revolves around a trade-off between 'exploitation' of target data and 'exploration' of the source, or between bias and variance from statistical perspective, since restrictive transferability alleviates the issue of a shrinking sample size while avoids bad or negative

1. Thanks for advices from my supervisor Prof. Wei Zhong, Prof. Jingyuan Liu and Prof. Xingbai Xu at Wang Yanan Institute for Studies in Economics and Department of Statistics and Data Science, SOE, Xiamen University.

transfers and minimizing bias. Conversely, lowering the transfer threshold may potentially lead to larger bias or negative transfer but promising enhanced estimation efficiency.

To strike a balance, I propose a partial approach, allowing one subset of the linear model to be dissimilar while ensuring asymptotic closeness with respect to the other. This approach facilitates transfer within the similar subset, providing flexible weighting between the bias and variance trade-off especially in empirical contexts, and allowing for greater tolerance in transferability.

Following the logic, however, it's worth noting that the correlation between two separated subsets may lead to the failure of several classic estimators, especially in high-dimensional scenarios. For example, as pointed out by Zhao and Yu (2006), l_1 regularization is invalid due to failure in Irrepresentable Condition, the sufficient and necessary condition for selection consistency of Lasso. Additionally, the hat matrix for least square projection is largely singular in large p, small n scenarios.

Throughout the years, scientists presents modifications towards stableness of regularization methods, including Meinshausen and Bühlmann (2010)'s stability selection technique, Chernozhukov *et al.* (2018)'s Double-Debiased Machine Learning (DML) approach and so forth. And this paper will thus inherit the high-dimensional linear confounding framework as mentioned in Chernozhukov *et al.* (2018) for valid estimation issue, which is also a simplified version of semi-parametric modeling as in Robinson (1988), in order to handle unstableness.

Therefore, combining the aforementioned consideration of transferability and confounding structure, I formulate knowledge transferring strategy in a partial approach. With the prior information for the similar subset, I bound the transferable gap partially on single subset, identifying positive sources based on considerations of sample fluctuation and applying Cross-Validation and Bootstrap (Efron, 1994) to address randomness. Mimicking Tian and Feng (2022), a cutting interval for transferability is established for Partial Transfer Learning (PTL) method. And the sequential parameter estimation and outcome prediction issue are combing DML, Oracle Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001) and cross-fitting technique as in Fan *et al.* (2012) for proper asymptotics.

Under linearization and i.i.d. Data Generating Process (DGP), consistency and asymptotic normality are assured by law of large numbers and central limit theorem, as well as other fundamental theories of large samples. And compared with non-transferring, PTL reveals reduced asymptotic variance if the linear signal within the confounding is strong enough to be captured by DML. On the other hand, if such latent confounding is dominated by random noises, transferring may do harm to the estimator's efficiency asymptotically.

In summary, I contribute to previous researches on statistical transfer learning in the following: First and foremost, PTL considers a more general causal framework by combining potential confounding modules, forming reliable models for empirical applications compared to the smoothing and de-biasing framework introduced by Li *et al.* (2022a). Additionally, PTL relaxes the similarity restriction by allowing partially dissimilar parameters between the target and source data, offering greater flexibility. Moreover, PTL estimation is willing to be extended into federated case, offering a more efficient approach compared to Li *et al.* (2023a), which packages sharable information into Hessian matrices, potentially adding extra communication costs. Finally, the introduced properties under large samples mainly consider rate of convergence and efficiency enhancement, providing a different perspective on improved performance with knowledge transfer compared to Li *et al.* (2022a)'s minimax-lower-bound-type optimality.

2 Notation and Organization

In this paper, we will adopt specific notations. Matrices are denoted using capitalized bold letters (e.g., $\mathbf{X}$), while vectors are represented without bold formatting (e.g., X), and lowercase x is used for scalars. For a d-dimensional vector $X = (x_1, \ldots, x_d)^\top$, the function $dim(X)$ outputs its dimensionality, the l_q-norm is defined as $\|X\|_q = \left(\sum_{i=1}^{d} |x_i|^q\right)^{1/q}, q \in (0, 2]$ and the ∞-norm is represented as $\|X\|_\infty = \max_{i=1,\ldots,d} |x_i|$. For matrix $\mathbf{X}_{p\times q} = [x_{ij}]_{p\times q}$, $\mathbf{X} > 0$ signifies that the matrix is positive definite.

When dealing with two non-zero real sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, I use $a_n = o(b_n)$ to represent $|a_n/b_n| \to 0$ as $n \to \infty$, and $a_n = O(b_n)$ means that a_n/b_n converges to some positive constant. For sequences of random variable $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$, expression $x_n = O_p(y_n)$ means $|x_n/y_n| \xrightarrow{p} 0$ as $n \to \infty$, and notation $x_n = O_p(y_n)$ indicates that for any $\epsilon > 0$, there exists a positive constant C such that $\sup_n \mathbb{P}(|x_n/y_n| > C) \leq \epsilon$. Additionally, $x_n \xrightarrow{d} z$ asserts that $\sup_{B\in\mathcal{B}} |\mathbb{P}(x_n \in B) - \mathbb{P}(z \in B)| \to 0$ as $n \to \infty$ for $\mathcal{B}$ being the Borel set of real line. If x_n and z have the same distribution, I denote it as $x_n \overset{d}{\sim} z$, while $x_n \overset{i.i.d.}{\sim} z$ means that x_n is generated independently and is with identical distribution as z.

The expectation $\mathbb{E}$ and variance $\mathbb{V}$ are calculated based on all randomness while the sample-based counterparts are represented as $\widehat{\mathbb{E}}$, and $avar(\cdot)$ is used to denote the asymptotic variance. The symbol $\Phi^{-1}(p)$ refers to the p-th quantile point of the normal distribution.

The subsequent sections of this paper are organized as follows. Section 3 elucidates the underlying DGP for PTL within the context of a high-dimensional linear confounding framework, and it defines the informative source data in PTL. In Section 4, we introduce source detection algorithms that rely on cross-validated and bootstrapped cutting intervals. Sections 5 and 6 explore diverse scenarios encompassing both homogeneous and heterogeneous sources, where the transferable subset may vary among different sources. These sections encompass meticulously crafted algorithms and present the derived asymptotic results related to estimation consistency, asymptotic normality, and enhanced efficiency. Section 7 conducts comprehensive simulations that not only validate the enhanced performance of PTL but also corroborate the previously established asymptotic findings. In Section 8, we shift our focus to an empirical analysis of the GTEx database, demonstrating the tangible improvements in prediction made possible by PTL. Section 9 concludes the paper by reviewing the methodology and offering insights into promising avenues for future research. Additional simulation results, theoretical analyses, and all supporting proofs are relegated to the appendix for thorough examination.

3 Underlying Model

Target model is formatted as

$$Y^{(0)} = \mathbf{D}^{(0)}\rho_0 + \mathbf{X}^{(0)}\beta_0 + V^{(0)}, \quad \mathbb{E}[V^{(0)} \mid \mathbf{X}^{(0)}, \mathbf{D}^{(0)}] = 0 \tag{1}$$

$$\mathbf{D}^{(0)} = \mathbf{X}^{(0)}\boldsymbol{\gamma}_0 + \mathbf{U}^{(0)}, \quad \mathbb{E}[\mathbf{U}^{(0)} \mid \mathbf{X}^{(0)}] = 0 \tag{2}$$

where data $Y^{(0)} \in R^{n_0}$, $\mathbf{D}^{(0)} \in R^{n_0\times p}$, $\mathbf{X}^{(0)} \in R^{n_0\times q}$, and parameter $\rho_0 \in R^p$, $\beta_0 \in R^q$, $\boldsymbol{\gamma}_0 \in R^{q\times p}$ denote the impact of dissimilar causal, similar nuisance and confounding part on response, $V^{(0)} \in R^{n_0}$ and $\mathbf{U}^{(0)} \in R^{n_0\times p}$ are errors. Without loss of generality, p is fixed and finite.

k-th source model are formatted as

$$Y^{(k)} = \mathbf{D}^{(k)}\rho_0^{(k)} + \mathbf{X}^{(k)}\beta_0^{(k)} + V^{(k)}, \quad \mathbb{E}[V^{(k)} \mid \mathbf{X}^{(k)}, \mathbf{D}^{(k)}] = 0, \quad \forall k \in \{1, 2, ..., K\} \tag{3}$$

$$\mathbf{D}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\gamma}_0^{(k)} + \mathbf{U}^{(k)}, \quad \mathbb{E}[\mathbf{U}^{(k)} \mid \mathbf{X}^{(k)}] = 0, \quad \forall k \in \{1, 2, ..., K\} \tag{4}$$

where K is the total number of sources, and data matrix mimics the formation of target with sample size $n_k >> n_0,\ \forall k \in \{1, 2, ..., K\}$. Consider high-dimensional confounding: $q >> n_0$, but within the framework of sparsity.

The informative/transferrable source requires partially identical nuisance:

$$\mathcal{I}_{h_{n_k}} = \cup_{k=1}^{K} \left\{ \mathcal{S}_k : \|\beta_{0,\mathcal{J}_k} - \beta_{0,\mathcal{J}_k}^{(k)}\|_\infty = 0 \right\} \tag{5}$$

for $k \in \{1, 2, ..., K\}$, where $\{\mathcal{J}_k\}_{k=1}^K$ are index sets partitioning $\{1, 2, ..., q\}$, and $-\mathcal{J}_k$'s are the counterpart composing $\{1, 2, ..., q\}$. Also the above condition implies

$$\widehat{\mathcal{I}}_{h_{n_k}} = \cup_{k=1}^{K} \left\{ \mathcal{S}_k : \mathbb{P}\left(\beta_{0,\mathcal{J}_k} - \hat{\beta}_{\mathcal{J}_k}^{(k)} \leq \boldsymbol{h}_{n_k} \right) \to 1 \ as \ n_k \to \infty \right\} \tag{6}$$

if $\hat{\beta}_{\mathcal{J}_k}^{(k)}$ is $\sqrt{n_k}$-consistent estimator of $\beta_{0,\mathcal{J}_k}^{(k)}$ and $\boldsymbol{h}_{n_k} \to 0$ with slower rate than $O(1/\sqrt{n_k})$ as $n_k \to \infty$, where constant vector $\boldsymbol{h}_{n_k}$ measures the sample proximity of nuisance.

4 Source Detection

This section aims to facilitate the automatic detection of transferable sources. It is important to note that the following algorithms are designed to be fully automatic, requiring no subjective ranking. The overall design is rooted in the concept of confidence interval cutting. Similar techniques can be found in the work of Tian and Feng (2022), where they also incorporate a Cross Validation (CV) procedure. However, in addition to CV, I introduce an alternative approach using Bootstrapping, which tends to exhibit sharper consistency, albeit at the cost of being more time-consuming in simulations.

The first algorithm pertains to the implementation of CV. It is built upon the benchmark of DML estimator as outlined by Chernozhukov *et al.* (2018). In this algorithm, I calibrate the cutting point for determining the positivity of sources by accounting for estimation fluctuations. This is achieved by aggregating empirical variances from both the target and the source through the CV process. Additionally, I incorporate a multiplier of 0.025 upper quantile of standard normal distribution, creating a form akin to a confidence interval.

Meta-Algorithm-1.1 (Homogeneous Source Detection-Cross Validation)

1) *Partitioning* $\mathcal{S}_0$ *into* M *validation subsamples* $\mathcal{S}_0^{(1)}, ..., \mathcal{S}_0^{(M)}$ *and the corresponding* M *training subsamples are* $\mathcal{S}_0^{(-1)}, ..., \mathcal{S}_0^{(-M)}$*;*

2) *Apply DML to* $\mathcal{S}_0^{(-1)}, ..., \mathcal{S}_0^{(-M)}$ *and get* $\hat{\rho}^{(01)}, ..., \hat{\rho}^{(0M)}$*;*

3) *Compute* $\widehat{\mathbb{E}}\left[\widehat{f}_0(X^{(01)})\right], ..., \widehat{\mathbb{E}}\left[\widehat{f}_0(X^{(0M)})\right]$ *based on validation sample* $\mathcal{S}_0^{(1)}, ..., \mathcal{S}_0^{(M)}$*;*

4) *Partitioning* $\mathcal{S}_k$ *into* M *validation subsamples* $\mathcal{S}_k^{(1)}, ..., \mathcal{S}_k^{(M)}$ *and the corresponding* M *training subsamples are* $\mathcal{S}_k^{(-1)}, ..., \mathcal{S}_k^{(-M)}$*;*

5) *Apply DML to* $\mathcal{S}_k^{(-1)}, ..., \mathcal{S}_k^{(-M)}$ *and get* $\hat{\rho}^{(k1)}, ..., \hat{\rho}^{(kM)}$*;*

6) *Compute* $\widehat{\mathbb{E}}\left[\widehat{f}_k(X^{(k1)})\right], ..., \widehat{\mathbb{E}}\left[\widehat{f}_k(X^{(kM)})\right]$ *based on validation sample* $\mathcal{S}_k^{(1)}, ..., \mathcal{S}_k^{(M)}$*;*

7) *Compute*

$$\widehat{\mathbb{E}}\left[\widehat{f}_0(X^{(0)})\right] = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathbb{E}}\left[\widehat{f}_0(X^{(0m)})\right]$$

$$\widehat{\mathbb{E}}\left[\widehat{f}_k(X^{(k)})\right] = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathbb{E}}\left[\widehat{f}_k(X^{(km)})\right]$$

$$\hat{\sigma}_0^2 = \frac{1}{M-1}\sum_{m=1}^{M} \left(\widehat{\mathbb{E}}\left[\widehat{f}_0(X^{(0m)})\right] - \widehat{\mathbb{E}}\left[\widehat{f}_0(X^{(0)})\right] \right)^2$$

$$\hat{\sigma}_k^2 = \frac{1}{M-1}\sum_{m=1}^{M} \left(\widehat{\mathbb{E}}\left[\widehat{f}_k(X^{(km)})\right] - \widehat{\mathbb{E}}\left[\widehat{f}_k(X^{(k)})\right] \right)^2$$

$$\hat{\sigma}_{0k}^2 = \hat{\sigma}_0^2 + \hat{\sigma}_k^2,\ k \in \{1, ..., K\};$$

8) *Transferrable sources are*

$$\mathcal{I} = \cup_{k=1}^{K} \left\{ \mathcal{S}_k : \left| \widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(0)}) \right] - \widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(k)}) \right] \right| \leq C_0(\hat{\sigma}_{0k} \vee 0.01) \right\}.$$

where $\widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(km)}) \right] = \frac{n_k}{M} \sum_{i=1}^{\frac{n_k}{M}} \left[Y_i^{(km)} - \hat{\rho}^{(km)\top} D_i^{(km)} \right],\ k \in \{0, 1, ..., K\}$ for simplicity. Also note that $M = 3$ and constant $C_0 = \Phi^{-1}(0.975)$ typically in simulation.

On the other hand, an alternative method for obtaining the empirical variance is through Bootstrapping, and the subsequent algorithm employs this approach to estimate the bounds as well as the measure of nuisance similarity.

Meta-Algorithm-1.2 (Homogeneous Source Detection-Bootstrapping)

1) *Bootstrapping* $\mathcal{S}_0$ *into* $\mathcal{S}_0^{(1)}, ..., \mathcal{S}_0^{(B)}$*;*
2) *Apply DML to* $\mathcal{S}_0^{(1)}, ..., \mathcal{S}_0^{(B)}$ *and get* $\hat{\rho}^{(01)}, ..., \hat{\rho}^{(0B)}$*;*
3) *Compute* $\widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(01)}) \right], ..., \widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(0B)}) \right]$ *based on original sample* $\mathcal{S}_0$*;*

4) *Bootstrapping* $\mathcal{S}_k$ *into* $\mathcal{S}_k^{(1)}, ..., \mathcal{S}_k^{(B)}$*;*
5) *Apply DML to* $\mathcal{S}_k^{(1)}, ..., \mathcal{S}_k^{(B)}$ *and get* $\hat{\rho}^{(k1)}, ..., \hat{\rho}^{(kB)}$*;*

6) *Compute* $\widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(k1)}) \right], ..., \widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(kB)}) \right]$ *based on original sample* $\mathcal{S}_k$*;*
7) *Compute*

$$\widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(0)}) \right] = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(0b)}) \right]$$

$$\widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(k)}) \right] = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(kb)}) \right]$$

$$\hat{\sigma}_0^2 = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(0b)}) \right] - \widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(0)}) \right] \right)^2$$

$$\hat{\sigma}_k^2 = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(kb)}) \right] - \widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(k)}) \right] \right)^2$$

$$\hat{\sigma}_{0k}^2 = \hat{\sigma}_0^2 + \hat{\sigma}_k^2,\ k \in \{1, ..., K\};$$

8) *Transferrable sources are*

$$\mathcal{I} = \cup_{k=1}^{K} \left\{ \mathcal{S}_k : \left| \widehat{\mathbb{E}} \left[\widehat{f}_0(X^{(0)}) \right] - \widehat{\mathbb{E}} \left[\widehat{f}_k(X^{(k)}) \right] \right| \leq C_0(\hat{\sigma}_{0k} \vee 0.01) \right\}.$$

where $B = 10$ and constant $C_0 = \Phi^{-1}(0.975)$ in simulation.

5 PTL on Homogeneous Sources

In this section, I introduce a two-step estimation algorithm for a linear model characterized by high-dimensional confounding. This algorithm is designed to leverage homogeneous sources and specifically accounts for the similarity between the mean functions of the nuisance.

In this context, homogeneity is defined as the consistent separation of causal nuisance factors between

the target and each source model. To enhance clarity and provide a more illustrative explanation, I limit the discussion to single-source parametric models in subsections 5.1 and 5.2. Comprehensive algorithms for multiple sources are summarized separately. As a result, the informative auxiliary sample case is simplified as follows:

$$\mathcal{I}_{h_{n_k}} = \cup_{k=1}^{K} \left\{ \mathcal{S}_k : \left\| \beta_0 - \beta_0^{(k)} \right\|_{\infty} = 0 \right\} \tag{7}$$

I illustrate the consistency of the algorithms when the dimensionality parameter p is set to 1, with a specific focus on addressing potential covariate shift. Detailed discussions for the general case can be found in the following corollary and in the appendix section.

Moreover, it's worth noting that this design can be readily extended to a federated transfer learning scenario, as demonstrated in Li *et al.* (2023a). In this setting, the underlying DGP encompassing equations (1)-(4) remains identical for each site indexed as $m \in 1, ..., M$. Information sharing is computationally efficient, while privacy is meticulously safeguarded among the participating sites.

5.1. Without Covariate Shift

For the sake of clarity, the initial subsection operates under the assumption of no covariate shift, even though such an assumption is often far from realistic. Drawing from Chernozhukov *et al.* (2018), I begin by implementing the DML method on the source data, which leads to a debiased estimation of $\rho_0^{(k)}$. This estimate is reserved for the transfer process. A few key points merit attention: Firstly, I implicitly assume sparsity in both the nuisance and confounding functions, and therefore exclusively employ the Lasso learner within the DML framework. However, if the generative structure departs from this sparsity assumption, other learners such as Random Forest, Gradient Boosting Decision Tree, or Neural Network can be readily applied. Secondly, it's important to note that the causal component of the source model holds no algorithmic significance, but it plays a pivotal role in facilitating unbiased nuisance estimation.

Now, considering equation (2) with an additional intercept term derived from the source, I effectively transfer the nuisance mean function to reconstruct the target response. Subsequently,

$$\hat{\rho}_{\mathrm{PTL}} = \frac{\widehat{\mathbb{E}}\left(\widehat{Y}_i^{(0)}\right)}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} \tag{8}$$

where $\widehat{Y}_i^{(0)} \equiv Y_i^{(0)} - \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathbb{E}}\left(Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)}\right)$. Thus the consistency of $\hat{\rho}_{\mathrm{PTL}}$ is promised by well-calibration of nuisance. Summary algorithms are given in the following:

Meta-Algorithm-2.1 (Partial Transfer Learning with Homogeneous Source)

1) *Apply DML to the source data S_k and get estimation of $\hat{\rho}^{(k)}$;*

2) *Implement $Alg(\cdot)$ for $\widehat{Y}^{(0)}$ on $D^{(0)}$ and get estimation of ρ_0.*

where $Alg(\cdot)$ stands for the Empirical Expectation Ratio (EER) equation (8).

Furthermore, it is worth noting that while not explicitly addressed in the source detection algorithms **Meta-Algorithm-1.1** and **Meta-Algorithm-1.2**, the ranking of transferable sources can provide valuable insights for adaptive estimation. And this important aspect is encapsulated in the following algorithm:

Meta-Algorithm-2.2 (Adaptive Partial Transfer Learning with Homogeneous Source)

1) *Apply DML to the source data S_k and get estimation of $\hat{\rho}^{(k)}$;*

2) *Calculate adaptive weight $w_k = \frac{1}{rank(S_k)} / \sum_{k=1}^{K} \frac{1}{rank(S_k)}$ for each source;*

3) *Implement $Alg(\cdot)$ for $Y^{(0)} - \frac{1}{K}\sum_{k=1}^{K} w_k \widehat{\mathbb{E}}\left(Y^{(k)} - D^{(k)}\hat{\rho}^{(k)}\right)$ on $D^{(0)}$ and get $\hat{\rho}_{\mathrm{PTL}}$.*

5.2. Covariate Shift

Due to the potential shifts in nuisance distribution, a simplistic application of EER is unlikely to yield satisfactory results. To address this issue, a parameter-wise transfer strategy is formulated using the SCAD penalty, as introduced by Fan and Li (2001), which possesses the Oracle property, ensuring PTL's estimation consistency in this context.

The initial step in **Meta-Algorithm-2.1** remains unchanged; however, addressing the covariate shifts necessitates more than mere de-confounding precautions. Therefore, in this extended approach, the implementation of cross-fitting estimators $\hat{\rho}^{(k)}$ and $\hat{\beta}^{(k)}$ is introduced. Consequently, the response in the source data is reconstructed as $Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)}$ and oracle $\hat{\beta}^{(k)}$ is obtained through SCAD. The computation proceeds with

$$\hat{\rho^*}_{\mathrm{PTL}} = \frac{\widehat{\mathbb{E}}\left(\widehat{Y}_i^{*^{(0)}}\right)}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} \tag{9}$$

as the target causal estimation, where $\widehat{Y}_i^{*^{(0)}} \equiv Y_i^{(0)} - \frac{1}{K}\sum_{k=1}^{K}\left[\widehat{\mathbb{E}}\left(Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)}\right) - \widehat{\Delta}^{(k)^\top}\hat{\beta}^{(k)}\right]$, $\widehat{\Delta}^{(k)} \equiv \widehat{\mathbb{E}}\left(X_i^{(k)^\top}\right) - \widehat{\mathbb{E}}\left(X_i^{(0)^\top}\right)$. Thus forming the algorithm in the shifting case as

Meta-Algorithm-2.3 (Partial Transfer Learning with Homogeneous Source under Covariate Shift)

1) *Split each source* $S_k,\ k \in \{1, 2, ..., K\}$ *into* $F = 5$ *folds;*
2) *Apply DML to the source data* $S_{k,f}$ *and get estimation of* $\hat{\rho}^{(k,f)},\ f \in \{1, ..., F\}$*;*
3) *Apply SCAD to the source data* $S_{k,-f}$ *on response* $Y_i^{(k,-f)} - D_i^{(k,-f)}\hat{\rho}^{(k,-f)}$ *and get* $\hat{\beta}^{(k,f)}$*;*
4) *Cross-fitting estimator* $\hat{\rho}^{(k)} = \frac{1}{F}\sum_{f=1}^{F}\hat{\rho}^{(k,f)}$, $\hat{\beta}^{(k)} = \frac{1}{F}\sum_{f=1}^{F}\hat{\beta}^{(k,f)}$*;*
5) *Implement* $Alg(\cdot)$ *for* $\widehat{Y}_i^{*^{(0)}}$ *on* $D^{(0)}$ *and get* $\hat{\rho}$*.*

where $Alg(\cdot)$ here represents equation (9).

5.3. Remarks

(R1): SCAD method in all algorithms could be replaced by any sparse selection technique with Oracle property;

(R2): **Meta-Algorithm-2.1** and **Meta-Algorithm-2.3** can also be extended to multi-site cases, where for each $S_{(m,k)},\ m \in \{1, 2, ..., M\}$, same DGP as (1)-(4) is pre-assumed both for the target and source data sets. Thus the nuisance information of $\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}\left(Y_i^{(m,k)} - D_i^{(m,k)}\hat{\rho}^{(m,k)}\right)$ in **Meta-Algorithm-2.1** and $\frac{1}{K}\sum_{k=1}^{K}\left[\widehat{\mathbb{E}}\left(Y_i^{(m,k)} - D_i^{(m,k)}\hat{\rho}^{(m,k)}\right) - \widehat{\Delta}^{(m,k)^\top}\hat{\beta}^{(m,k)}\right]$ in **Meta-Algorithm-2.3** are willing to share among sites with privacy protected. Notice that the sharing of mean function could also be ensured unattackable if apply *ALGORITHM 3.1* in Cai *et al.* (2021) naturally through additional noise.

(R3): When $1 < p < n_0$ finite, the above PTL based on EER is no longer suitable for transfer, however, simply modifying $\widehat{Y}_i^{(0)}$ or $\widehat{Y}_i^{*^{(0)}}$ into $Y_i^{(0)} - X_i^{(0)}\hat{\beta}^{(k)}$ with $Alg(\cdot)$ being Ordinary Least Square (OLS) in **Meta-Algorithm-2.1** and **Meta-Algorithm-2.3** is enough for generalization, whose consistency and asymptotic normality is revealed by the subsequent corollary.

(R4): For consideration of possible covariate shift, the definition of $\widehat{\mathbb{E}}\left[\widehat{f}_k(X^{(km)})\right]$ in **Meta-Algorithm-1.1** and **Meta-Algorithm-1.2** in regard to source detection could easily be replaced by $\frac{n_k}{M}\sum_{i=1}^{\frac{n_k}{M}}\left[Y_i^{(km)} - \hat{\rho}^{(km)^\top}D_i^{(km)}\right] - \widehat{\Delta}^{(km)^\top}\hat{\beta}^{(km)},\ k \in \{0, 1, ..., K\}$.

5.4. Theorems

The following theorems and corollary demonstrate homogeneous PTL estimators' consistency and asymptotic normality, where the detailed proofs are left in the appendix A.

Theorem 1.1 (PTL on homogeneous sources without shift)

Under condition *(C1)-(C4)* in the appendix A.1,

$$\hat{\rho}_{\text{PTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \tag{10}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}_{\text{PTL}} \xrightarrow{p} \rho_0 \tag{11}$$

as $n_0 \to \infty$.

Theorem 1.2 (PTL on homogeneous sources with shift)

Under condition *(C1)-(C5)* in the appendix A.1,

$$\hat{\rho*}_{\text{PTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \tag{12}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho*}_{\text{PTL}} \xrightarrow{p} \rho_0 \tag{13}$$

$$\sqrt{n_0}\,(\hat{\rho*}_{\text{PTL}} - \rho_0) \xrightarrow{d} N(0, c_1\sigma_0^2) \tag{14}$$

as $n_0 \to \infty$, where σ_0^2 denotes the variance of noise V_i, $c_1 \equiv \frac{1}{\mathbb{E}^2\left(D_i^{(0)}\right)}$ constant, the superscript $*$ denotes the existence of nuisance shift.

Corollary 1 (Homogeneous PTL for general cases)

Under condition *(C1)-(C6)* in the appendix A.1,

$$\hat{\rho*}_{\text{PTL,G}} - \rho_0 \xrightarrow{p} \widehat{Q}^{-1}\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)} V_i^{(0)} \tag{15}$$

as $n_k \to \infty,\ n_0 = O(1)$, where $\widehat{Q}^{-1} \equiv \frac{1}{n_0}\sum_{i=1}^{n_0}\left(D_i^{(0)} D_i^{(0)^\top}\right)$. Further,

$$\hat{\rho*}_{\text{PTL},G} - \rho_0 \xrightarrow{p} 0$$

$$\sqrt{n_0}\left(\hat{\rho*}_{\text{PTL},G} - \rho_0\right) \xrightarrow{d} N\left(0, \sigma_0^2 Q^{-1}\right) \tag{16}$$

as $n_0 \to \infty,\ n_k \to \infty$ for independent random noises, where $Q \equiv \mathbb{E}\left(D_i^{(0)} D_i^{(0)^\top}\right)$ invertible, subscript G indicates the general case. Also,

$$\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\left(\widehat{Y}_i^{(0)} - Y_i^{(0)}\right) \xrightarrow{d} N\left(0, \sigma_0^2(1 + C_1^\top Q^{-1} C_1)\right) \tag{17}$$

where $C_1 \equiv \mathbb{E}\left(D_i^{(0)}\right)$.

6 PTL on Heterogenous Sources

Building upon the conditions outlined in equation (5) for informative auxiliary samples, I introduce the algorithm for hetero-source partial transfer learning, focusing on the structural and treatment parameter ρ_0. In this context, heterogeneity arises from the presence of distinct causal and nuisance components between sources. In particular, within each source S_k, I treat each $\hat{\beta}_{0,j\notin\mathcal{J}_k}^{(k)}$ and $\rho_0^{(k)}$ as the dissimilar subset, and the implementation of DML to get unbiased estimation for $\mathbb{E}\left[\hat{f}_{\mathcal{J}_k}^{(k)}(X^{(k)})\right]$'s is akin to the approach discussed in subsection 5.1. Note that without loss of generality, I implicitly assume exhaustiveness in the heterogeneous sources, resulting in the ensemble of the nuisance being simply $\sum_{k=1}^{K}\widehat{\mathbb{E}}\left[\hat{f}_{\mathcal{J}_k}^{(k)}(X^{(k)})\right]$. Furthermore, these algorithms are primarily designed for the $p = 1$ case, with the extension to federated Partial Transfer Learning being straightforward. The general case is comprehensively summarized in the corollary and is

elaborated upon in the appendix to provide complementary insights.

6.1. Without Covariate Shift

Even in cases where there is no shift in the nuisance distribution, the cross-fitting strategy is employed as a precaution to ensure valid large sample properties. The subsequent steps remain straightforward and mirror those discussed in subsection 5.2: transferring the nuisance mean function to the transformed target model, followed by obtaining

$$\hat{\rho}_{\text{HPTL}} = \frac{\widehat{\mathbb{E}}\left(\widehat{Y}_{i,H}^{(0)}\right)}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} \tag{18}$$

where $\widehat{Y}_{i,H}^{(0)} \equiv Y_i^{(0)} - \sum_{k=1}^{K} \widehat{\mathbb{E}}\left(Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right)$ and subscript H stands for Heterogeneity.

Summary algorithm is given in the following:

Meta-Algorithm-3.1 (Partial Transfer Learning with Heterogeneous Sources)

1) *Split each source* $S_k,\ k \in \{1, 2, ..., K\}$ *into* $F = 5$ *folds;*
2) *Apply DML to the source data* $S_{k,f}$ *and get estimation of* $\hat{\rho}^{(k,f)},\ f \in \{1, ..., F\}$*;*
3) *Apply SCAD to the source data* $S_{k,-f}$ *on response* $Y_i^{(k,-f)} - D_i^{(k,-f)}\hat{\rho}^{(k,-f)}$ *and get* $\hat{\beta}^{(k,f)}$*;*
4) *Cross-fitting estimator* $\hat{\rho}^{(k)} = \frac{1}{F}\sum_{f=1}^{F}\hat{\rho}^{(k,f)},\ \hat{\beta}^{(k)} = \frac{1}{F}\sum_{f=1}^{F}\hat{\beta}^{(k,f)}$*;*
5) *Implement unbiased* $Alg(\cdot)$ *for* $\widehat{Y}_H^{(0)}$ *on* $D^{(0)}$ *and get estimation of* ρ_0*.*

where $Alg(\cdot)$ indicates equation (18).

6.2. Covariate Shift

In the scenario where latent shifts in nuisance covariate are taken into account, as discussed in subsection 5.2, this subsection closely resembles **Meta-Algorithm-3.1**. However, here, I introduce a multi-level cross-fitting procedure to address the inherent heterogeneity.

Meta-Algorithm-3.2 (Partial Transfer Learning with Heterogeneous Source under Covariate Shift)

1) *Split each source* $S_k,\ k \in \{1, 2, ..., K\}$ *into* $F = 3$ *folds;*
2) *Apply DML to the source data* $S_{k,f=1}$ *and get estimation of* $\hat{\rho}^{(k,1),a}$*;*
3) *Apply SCAD to* $S_{k,2}$ *on response* $Y_i^{(k,2)} - D_i^{(k,2)}\hat{\rho}^{(k,1),a}$ *and get* $\hat{\beta}_{-\mathcal{J}_k}^{(k,2),a}$*;*
4) *Apply SCAD to* $S_{k,3}$ *on response* $Y_i^{(k,3)} - D_i^{(k,3)}\hat{\rho}^{(k,1),a} - X_i^{(k,3)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k,2),a}$ *and get* $\hat{\beta}_{\mathcal{J}_k}^{(k,3),a}$*;*
5) *Exhaust* $A \equiv F! = 6$ *paths to get the multiple cross-fitting estimator*

$$\hat{\rho}^{(k)} = \frac{1}{A}\sum_{a=1}^{A}\hat{\rho}^{(k),a},\ \hat{\beta}_{-\mathcal{J}_k}^{(k)} = \frac{1}{A}\sum_{a=1}^{A}\hat{\beta}_{-\mathcal{J}_k}^{(k),a},\ \hat{\beta}_{\mathcal{J}_k}^{(k)} = \frac{1}{A}\sum_{a=1}^{A}\hat{\beta}_{\mathcal{J}_k}^{(k),a};$$

6) *Implement* $Alg(\cdot)$ *for* $\widehat{Y}_H^{*(0)}$ *on* $D^{(0)}$ *and get* $\hat{\rho}$*.*

where $\widehat{Y}_{i,H}^{*(0)} \equiv Y_i^{(0)} - \sum_{k=1}^{K}\left[\widehat{\mathbb{E}}\left(Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) - \widehat{\Delta}_{\mathcal{J}_k}^{(k)\top}\hat{\beta}_{\mathcal{J}_k}^{(k)}\right]$, $Alg(\cdot)$ here represents equation (18).

6.3. Remarks

(R5): Again, if assuming identical DGP among sites as Li *et al.* (2023a), then mean functions of

$$\sum_{k=1}^{K}\widehat{\mathbb{E}}\left(Y_i^{(m,k)} - D_i^{(m,k)}\hat{\rho}^{(m,k)} - X_{i,-\mathcal{J}_{(m,k)}}^{(m,k)\top}\hat{\beta}_{-\mathcal{J}_{(m,k)}}^{(m,k)}\right)$$

in **Meta-Algorithm-3.1** and

$$\sum_{k=1}^{K}\left[\widehat{\mathbb{E}}\left(Y_i^{(m,k)} - D_i^{(m,k)}\hat{\rho}^{(m,k)} - X_{i,-\mathcal{J}_{(m,k)}}^{(m,k)\top}\hat{\beta}_{-\mathcal{J}_{(m,k)}}^{(m,k)}\right) - \widehat{\Delta}_{\mathcal{J}_{(m,k)}}^{(m,k)\top}\hat{\beta}_{\mathcal{J}_{(m,k)}}^{(m,k)}\right]$$

in **Meta-Algorithm-3.2** are harmless to share between different sites, thus automatically extending to federated PTL.

(R6): For general cases when $1 < p < n_0$, $\widehat{Y}_{i,H}^{(0)}$ and $\widehat{Y}_{i,H}^{*(0)}$ could be modified to $Y_i^{(0)} - \sum_{k=1}^{K} X_{i,\mathcal{J}_k}^{(0)^\top} \hat{\beta}_{\mathcal{J}_k}^{(k)}$ and $Alg(\cdot)$ in **Meta-Algorithm-3.1** and **Meta-Algorithm-3.2** are simply OLS.

(R7): The source detection algorithms for heterogeneous case should be similar with **Meta-Algorithm-1.1** or **Meta-Algorithm-1.2** when the heterogeneity between the target and each source is well-defined. However, in cases where such prior knowledge is lacking, an exhaustive search may be necessary. It's important to highlight that in my empirical analysis on GTEx data, a data-driven pre-searching procedure is introduced to facilitate a rough separation of the causal and nuisance components.

6.4. Theorems

The following theorems and corollaries serve to illustrate the consistency and asymptotic normality of the Partial Transfer Learning (PTL) estimators. The comprehensive proofs for these results are provided in the appendix B and C. Furthermore, in comparison to performing DML solely on the target dataset, PTL is demonstrated to be more efficient when the confounding factors are not predominantly influenced by random noise.

Theorem 2.1 (PTL on heterogeneous sources without shift)

Under condition *(C1)-(C5)* in the appendix A.1,

$$\hat{\rho}_{\text{HPTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \tag{19}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}_{\text{HPTL}} - \rho_0 \xrightarrow{p} 0 \tag{20}$$

$$\sqrt{n_0}\,(\hat{\rho}_{\text{HPTL}} - \rho_0) \xrightarrow{d} N\left(0, c_1\sigma_0^2\right) \tag{21}$$

as $n_0 \to \infty$.

Theorem 2.2 (PTL on heterogeneous sources with shift)

Under condition *(C1)-(C5)* in the appendix A.1,

$$\hat{\rho}^*_{\text{HPTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \tag{22}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}^*_{\text{HPTL}} - \rho_0 \xrightarrow{p} 0 \tag{23}$$

$$\sqrt{n_0}\,(\hat{\rho}^*_{\text{HPTL}} - \rho_0) \xrightarrow{d} N\left(0, c_1\sigma_0^2\right) \tag{24}$$

as $n_0 \to \infty$.

Corollary 2 (Heterogeneous PTL for general cases)

Under condition *(C1)-(C6)* in the appendix A.1,

$$\hat{\rho}^*_{\text{HPTL,G}} - \rho_0 \xrightarrow{p} \widehat{Q}^{-1}\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)} V_i^{(0)} \tag{25}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}^*_{\text{HPTL},G} - \rho_0 \xrightarrow{p} 0 \tag{26}$$

$$\sqrt{n_0}\left(\hat{\rho}*_{\mathrm{HPTL},G}-\rho_0\right)\xrightarrow{d}N\left(0,\sigma_0^2Q^{-1}\right) \tag{27}$$

as $n_0\to\infty$. Also,

$$\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\left(\widehat{Y}_{i,\mathrm{H}}^{(0)}-Y_i^{(0)}\right)\xrightarrow{d}N\left(0,\sigma_0^2(1+C_1^\top Q^{-1}C_1)\right) \tag{28}$$

as $n_0\to\infty$.

Proposition (DML for High-Dimensional Linear Regression)

Under condition *(C1)* in the appendix

$$\sqrt{n_0}\left(\hat{\rho}_{\mathrm{DML}}-\rho_0\right)\xrightarrow{d}N\left(0,c_2\sigma_0^2\right) \tag{29}$$

as $n_0\to\infty$, where $c_2\equiv\frac{1}{\mathbb{V}\left(U_i^{(0)}\right)}$ quantifies the variability in the confounding caused by random noise. This proposition can be readily derived using *THEOREM 3.1* from Chernozhukov *et al.* (2018). Consequently, the PTL estimator is shown to have reduced variance asymptotically:

Corollary 3 (Efficiency for PTL estimation)

Under condition *(C1)-(C5)* in the appendix,

$$avar\left(\hat{\rho}*_{\mathrm{PTL}}\right)=avar\left(\hat{\rho}*_{\mathrm{HPTL}}\right)<avar\left(\hat{\rho}_{\mathrm{DML}}\right) \tag{30}$$

if

$$\mathbb{V}\left(\gamma_0^\top X_i^{(0)}\right)>\mathbb{V}\left(U_i^{(0)}\right) \tag{31}$$

Notice that $\mathbb{V}(\gamma_0^\top X_i^{(0)})$ reflects the linear component that contributes to the confounding signal. Consequently, when strong signals in the existing confounding are harnessed through PTL, the estimator effectively decreases the estimator's asymptotic variance. As a result, PTL proves to be more efficient when dealing with large samples. However, when the linear confounding is akin to random noise, knowledge transfer might introduce additional fluctuations to the target, ultimately inflating the asymptotic variance and leading to a loss in efficiency.

Likewise for $p>1$

$$avar\left(\hat{\rho}*_{\mathrm{PTL,G}}\right)=avar\left(\hat{\rho}*_{\mathrm{HPTL,G}}\right)<avar\left(\hat{\rho}_{\mathrm{DML,G}}\right) \tag{32}$$

if

$$\boldsymbol{\gamma}_0\neq\mathbf{0}_{q\times p} \tag{33}$$

where $\mathbf{0}_{q\times p}$ denotes the zero matrix with dimension q by p.

Combine **Corollary 1** and **Corollary 2**, for consideration of prediction, PTL also owns enhanced efficiency as long as the existence of informative linear confounding:

Corollary 4 (Efficiency for PTL prediction)

Under condition *(C1)-(C6)* in the appendix,

$$avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_i^{(0)}\right)=avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_{i,\mathrm{H}}^{(0)}\right)<avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_{i,\mathrm{DML}}^{(0)}\right) \tag{34}$$

if (24) holds. Detailed proofs are in appendix C.

7 Simulation

7.1. PTL with Detected Sources

This section tends to verify **Meta-Algorithm-1.1** and **Meta-Algorithm-1.2**, Data Generating Process (DGP) for target data follows:

$n_0 = 60$, $q = 200$, $\mu_q = (10, ..., 10)^\top$, $\mathbf{\Sigma}^{(ij)}_{q\times q} = 0.5^{|i-j|}$, $\forall i, j \in \{1, 2, ..., q\}$,

$X_i^{(0)} \overset{i.i.d.}{\sim} N_q(\mu_q, \mathbf{\Sigma}_{q\times q})$, $V_i^{(0)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(0)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_0\}$,

$\rho_0 = -0.8$, $\beta_0 = (\beta_{(1)}^\top, \beta_{(2)}^\top)^\top$ where $\beta_{(1)}$ is a non-zero vector with dimension 20 and $\beta_{(2)}$ is a zero vector with dimension 180,

$\gamma_0 = (\gamma_{(1)}^\top, \gamma_{(2)}^\top)^\top$ where $\gamma_{(1)}$ is a non-zero vector with value being $seq(0.1, 1, 0.1)$, and $\gamma_{(2)}$ is a zero vector with dimension 190.

And for distinct sources, I enrich the diversifications by the following settings with:

Source 1 *(nearly positive)*: $n_1 = 500$, $X_i^{(1)} \overset{i.i.d.}{\sim} N_q(\mu_q, \mathbf{\Sigma}_{q\times q})$, $V_i^{(1)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(1)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_1\}$, $\rho_0^{(1)} = 0.8$, $\|\beta_0 - \beta_0^{(1)}\|_\infty \equiv h_1 = 0.1, \|\gamma_0 - \gamma_0^{(1)}\|_\infty \equiv s_1 = 0.5$ under the same sparsity structure above.

Source 2 *(neutral)*: $n_2 = 500$, $X_i^{(2)} \overset{i.i.d.}{\sim} N_q(-\mu_q, \mathbf{\Sigma}^{(2)}_{q\times q})$, where $\mathbf{\Sigma}^{(2)}_{q\times q}(i,j) = 0.2^{|i-j|}$, $V_i^{(2)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(2)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_2\}$, $\rho_0^{(2)} = 0.4$, $h_2 = 0.3, s_2 = 0.5$ under same sparsity structure above.

Source 3 *(confounding-negative)*: $n_3 = 500$, $X_i^{(3)} \overset{i.i.d.}{\sim} N_q(\mu_q, \mathbf{\Sigma}^{(3)}_{q\times q})$, where $\mathbf{\Sigma}^{(3)}_{q\times q}(i,j) = 0.8^{|i-j|}$, $V_i^{(3)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(3)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_3\}$, $\rho_0^{(3)} = 1.8$, $h_3 = 0.5, s_3 = 5$ under the same sparsity structure above.

Source 4 *(nuisance-negative)*: $n_4 = 500$, $X_i^{(4)} \overset{i.i.d.}{\sim} N_q(-\mu_q, \mathbf{\Sigma}^{(4)}_{q\times q})$, where $\mathbf{\Sigma}^{(4)}_{q\times q}(i,j) = 0.8^{|i-j|}$, $V_i^{(4)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(4)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_4\}$, $\rho_0^{(4)} = 1.8$, $h_4 = 1, s_4 = 0.5$ under same sparsity structure above.

The oracle observation is that the transferability of sources should obey Source 1 > Source 2 > Source 3 and 4. Let $M = 3$, $B = 10$ and $C_0 = \Phi^{-1}(0.975)$ in **Meta-Algorithm-1.1** and **Meta-Algorithm-1.2**, simulate for 100 times over 4 sources and the following tables compare these two algorithms as well as the sequentially partial transfer learning result.

	Detected Sources	Total Time
Cross Validation	17/100 None, 83/100 **Source 1**	87.3906 mins
Bootstrap	7/100 None, 93/100 **Source 1**	238.5247 mins

Following table also compares PTL with other published methods: Non-Trans Lasso and Non-Trans DML, which only use information from the target data; All-Trans PTL, which arbitrarily transfers all sources without detection; and Partial Trans-Lasso, which inherits Li *et al.* (2022a) but with mild modification for equation (6) in their paper by

$$\hat{\delta}^{\mathcal{A}} = \arg\min_{\delta\in\mathbb{R}^p} \left\{ \frac{1}{2n_0} \left\| y^{(0)} - X^{(0)} \left(\hat{w}^{\mathcal{A}} + \delta \right) \right\|_2^2 + \lambda_\delta \|w_*^\top \delta\|_1 \right\} \tag{35}$$

where $w_* \equiv (0, 1, 1, ..., 1)^\top$ is the partial penalty factor in order to match the prior knowledge on causal with PTL. The Root Mean Square Error (RMSE) comparisons are shown as below:

		Debiased	Transferred	Detected	RMSE
Cross Validation	Non-Trans Lasso	✗	✗	✗	1.3246
	Non-Trans DML	✓	✗	✗	0.4006
	All-Trans PTL	✓	✓	✗	1.8694
	Partial Trans-Lasso	✗	✓	✓	1.0300
	PTL	✓	✓	✓	**0.3823**
Bootstrap	Non-Trans Lasso	✗	✗	✗	1.1507
	Non-Trans DML	✓	✗	✗	0.4040
	All-Trans PTL	✓	✓	✗	1.8817
	Partial Trans-Lasso	✗	✓	✓	1.0007
	PTL	✓	✓	✓	**0.3985**

First, the transferrable interval effectively automates the selection of positive sources, whether employing the Cross Validation method in **Meta-Algorithm-1.1** or the Bootstrap method in **Meta-Algorithm-1.2**. Second, the Bootstrap method appears to enhance the accuracy of source selection but also requires a more time-intensive process. Third, when comparing various transfer strategies, PTL seems to stand out by achieving the lowest RMSE, particularly in the homogeneous case.

7.2. Oracle PTL Under Homogeneous Sources

The subsequent simulations are set on already-detected sources due to demonstration in subsection 7.1, thus for **Meta-Algorithm-2.1**, I design the following DGP for target data:

$n_0 = 20$, $q = 200$, $\mu_q = (10, ..., 10)^\top$, $\mathbf{\Sigma}_{q\times q}^{(ij)} = 0.5^{|i-j|}$, $\forall i, j \in \{1, 2, ..., q\}$,
$X_i^{(0)} \overset{i.i.d.}{\sim} N_q(\mu_q, \mathbf{\Sigma}_{q\times q})$, $V_i^{(0)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(0)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_0\}$,
$\rho_0 = -0.8$, $\beta_0 = (\beta_{(1)}^\top, \beta_{(2)}^\top)^\top$ where $\beta_{(1)}$ is a non-zero vector with dimension 20 and $\beta_{(2)}$ is a zero vector with dimension 180,
$\gamma_0 = (\gamma_{(1)}^\top, \gamma_{(2)}^\top)^\top$ where $\gamma_{(1)}$ is a non-zero vector of value $seq(0.1, 1, 0.1)$, and $\gamma_{(2)}$ is a zero vector with dimension 190.

Single source data are generated from:
$n_k = \{100, 200, 500\}$, $X_i^{(k)} \overset{i.i.d.}{\sim} N_q(\mu_q, \mathbf{\Sigma}_{q\times q})$, $V_i^{(k)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(k)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_k\}$,
$\rho_0^{(k)} = 0.8$, $\beta_0^{(k)} = (\beta_{(1)}^{(k)^\top}, \beta_{(2)}^{(k)^\top})^\top$ where $\beta_{(1)}^{(k)}$ are non-zero vector with dimension 20 and $\beta_{(2)}^{(k)}$ are zero vector with dimension 180. In order to prove PTL's robustness with respect to h_k among finite samples, here I simulate for $h_k \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$,
$\gamma_0^{(k)} = (\gamma_{(1)}^{(k)^\top}, \gamma_{(2)}^{(k)^\top})^\top$ where $\gamma_{(1)}^{(k)}$ are non-zero vector with dimension 10 and $\gamma_{(2)}^{(k)}$ are zero vector with dimension 190, $s_k \in \{0.05, 0.1, 0.5, 1, 5\}$. For simplicity, the subscript k is ignored in the following.

Following tables are displaying results under different s's with 100 simulation times through RMSE. For simplicity, two representative results are listed as tables, more simulations can be found in the appendix D.1.

$n_0 = 20,\ s = 0.05$		Transferred	Debiased	$h = 0$	$h = 0.05$	$h = 0.1$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2100	1.1985	1.2386
	Non-Trans DML	✗	✓	0.6806	0.6507	0.6602
	Partial Trans-Lasso	✓	✗	1.2806	1.2517	1.2106
	PTL	✓	✓	**0.1846**	**0.3303**	**0.4691**
$n_k = 200$	Non-Trans Lasso	✗	✗	1.1392	1.2039	1.1955
	Non-Trans DML	✗	✓	0.6716	0.6880	0.6576
	Partial Trans-Lasso	✓	✗	0.9995	0.9941	1.0211
	PTL	✓	✓	**0.0941**	**0.1965**	**0.3623**
$n_k = 500$	Non-Trans Lasso	✗	✗	1.2887	1.2085	1.2523
	Non-Trans DML	✗	✓	0.6767	0.7156	0.6907
	Partial Trans-Lasso	✓	✗	0.7246	0.7068	0.8209
	PTL	✓	✓	**0.0509**	**0.1771**	**0.3401**
$n_k = 1000$	Non-Trans Lasso	✗	✗	1.2981	1.3923	1.2894
	Non-Trans DML	✗	✓	0.6903	0.6996	0.6405
	Partial Trans-Lasso	✓	✗	0.5497	0.6903	0.7704
	PTL	✓	✓	**0.0384**	**0.1658**	**0.3349**

$n_0 = 20,\ s = 0.05$		Transferred	Debiased	$h = 0.2$	$h = 0.3$	$h = 0.5$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2173	1.2466	1.2473
	Non-Trans DML	✗	✓	0.7224	**0.6285**	**0.7017**
	Partial Trans-Lasso	✓	✗	1.2545	1.2172	1.0577
	PTL	✓	✓	**0.6492**	0.9205	1.4496
$n_k = 200$	Non-Trans Lasso	✗	✗	1.2912	1.2876	1.2783
	Non-Trans DML	✗	✓	0.6729	**0.6803**	**0.7154**
	Partial Trans-Lasso	✓	✗	1.1442	1.1977	1.3256
	PTL	✓	✓	**0.6129**	0.8945	1.4389
$n_k = 500$	Non-Trans Lasso	✗	✗	1.3044	1.2201	1.2990
	Non-Trans DML	✗	✓	0.6729	**0.7267**	**0.6849**
	Partial Trans-Lasso	✓	✗	1.0547	1.3001	1.4785
	PTL	✓	✓	**0.5785**	0.8623	1.4120
$n_k = 1000$	Non-Trans Lasso	✗	✗	1.2111	1.3094	1.2655
	Non-Trans DML	✗	✓	0.6498	0.6557	**0.6653**
	Partial Trans-Lasso	✓	✗	1.0615	1.2689	1.5015
	PTL	✓	✓	**0.5759**	**0.8574**	1.4091

Analyzing the information presented in the tables, several key observations can be made. Firstly, as anticipated, debiasing proves beneficial in mitigating the effects of the underlying confounding. Secondly, the knowledge transfer contributes to increased precision in estimating causal parameters, notably when sources are positively involved. Thirdly, PTL owns the robustness for limited n_0 and mild dissimilarity from h, where RMSE remains superior over Non-Trans DML. Lastly, in finite sample scenarios, it becomes evident that the influence of reducing the parameter s, the backdoor effects towards $D^{(0)}$, is incommensurate with narrowing h, the latter is certainly more influential to the transferring efficiency, since for $h = 0.5$, all transferring are negative. Related details could be found in the appendix D.1.

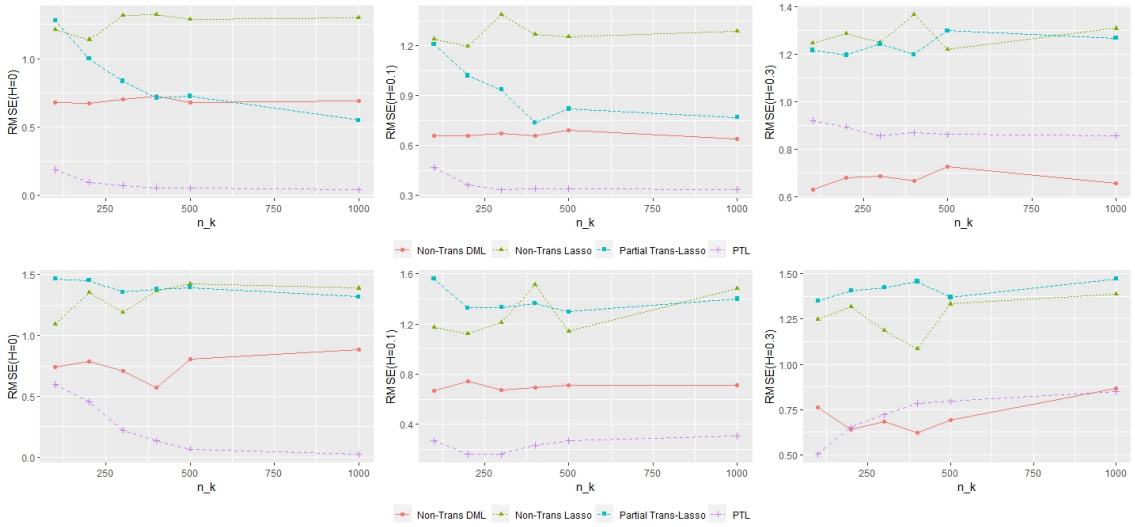
For **Meta-Algorithm-2.3**, the simulation design is identical with the above, except that $X_i^{(0)} \overset{i.i.d.}{\sim} N_q(\mu_q^{(0)}, \mathbf{\Sigma}_{q\times q})$ and $X_i^{(k)} \overset{i.i.d.}{\sim} N_q(\mu_q^{(k)}, \mathbf{\Sigma}_{q\times q})$ with distribution shift $\|\mu_q^{(0)} - \mu_q^{(k)}\|_\infty = 5$. Results for 20

simulations are as follows, which is similar with non-shifting case.

	$s = 0.05$	Transferred	Debiased	$h = 0$	$h = 0.05$	$h = 0.1$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.0937	1.3775	1.1728
	Non-Trans DML	✗	✓	0.7402	0.6071	0.6702
	Partial Trans-Lasso	✓	✗	1.4640	1.4733	1.5589
	PTL	✓	✓	**0.5983**	**0.4799**	**0.2677**
$n_k = 200$	Non-Trans Lasso	✗	✗	1.3512	1.2868	1.1252
	Non-Trans DML	✗	✓	0.7887	0.7289	0.7448
	Partial Trans-Lasso	✓	✗	1.4510	1.4022	1.3279
	PTL	✓	✓	**0.4550**	**0.2883**	**0.1620**
$n_k = 500$	Non-Trans Lasso	✗	✗	1.4248	1.1696	1.1418
	Non-Trans DML	✗	✓	0.8083	0.7437	0.7146
	Partial Trans-Lasso	✓	✗	1.3944	1.3261	1.2977
	PTL	✓	✓	**0.0655**	**0.1140**	**0.2710**
$n_k = 1000$	Non-Trans Lasso	✗	✗	1.3885	1.3363	1.4818
	Non-Trans DML	✗	✓	0.8821	0.7601	0.7092
	Partial Trans-Lasso	✓	✗	1.3178	1.3157	1.4010
	PTL	✓	✓	**0.0219**	**0.1518**	**0.3094**

	$s = 0.05$	Transferred	Debiased	$h = 0.2$	$h = 0.3$	$h = 0.5$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.3141	1.2467	1.1557
	Non-Trans DML	✗	✓	0.6784	0.7642	**0.7481**
	Partial Trans-Lasso	✓	✗	1.4517	1.3490	1.3694
	PTL	✓	✓	**0.2359**	**0.5031**	1.2225
$n_k = 200$	Non-Trans Lasso	✗	✗	1.2031	1.3181	1.3300
	Non-Trans DML	✗	✓	0.7105	**0.6423**	**0.7719**
	Partial Trans-Lasso	✓	✗	1.3253	1.4066	1.3588
	PTL	✓	✓	**0.2881**	0.6514	1.3483
$n_k = 500$	Non-Trans Lasso	✗	✗	1.2966	1.3324	1.2010
	Non-Trans DML	✗	✓	0.6580	**0.6940**	**0.7510**
	Partial Trans-Lasso	✓	✗	1.3675	1.3692	1.4247
	PTL	✓	✓	**0.5155**	0.7964	1.4042
$n_k = 1000$	Non-Trans Lasso	✗	✗	1.1655	1.3886	1.3117
	Non-Trans DML	✗	✓	0.7362	0.8667	0.8083
	Partial Trans-Lasso	✓	✗	1.4804	1.4705	1.6085
	PTL	✓	✓	**0.5565**	**0.8476**	1.4147

The following graph depicts the case of homogeneous PTL without shift (first row) and with shift (second row) compared other methods, which clearly suggests that PTL has sharper rate of convergence compared with Partial Trans-Lasso as long as $\frac{n_0}{n_k} = o(1)$.



7.3. Oracle PTL Under Heterogeneous Sources

Then I implement **Meta-Algorithm-3.1** for a more sophisticated simulation design, and DGP for target data follows subsection 7.2, but with: $dim(\beta_{(1)}) = 4$ and $dim(\gamma_{(1)}) = 5$;

Typifying heterogeneity, source 1 generates with:
$n_1 = 600$, $X_i^{(1)} \overset{i.i.d.}{\sim} N_q(\mu_q, \mathbf{\Sigma}_{q\times q})$, $V_i^{(1)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(1)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i, j \in \{1, 2, ..., n_0\}$,
$\rho_0^{(1)} = 3$, $dim(\beta_{(1)}^{(1)}) = dim(\beta_{(1)})$, $dim(\gamma_{(1)}^{(1)}) = dim(\gamma_{(1)})$,
its heterogeneity lies in its confounding separation:
$\beta_{(1)}^{(1)} \equiv (\beta_{(1)_a}^{(1)^\top}, \beta_{(1)_b}^{(1)^\top})^\top$ with $dim(\beta_{(1)_a}^{(1)^\top}) = 1$, $dim(\beta_{(1)_b}^{(1)^\top}) = 3$ satisfying $\|\beta_{(1)_a} - \beta_{(1)_a}^{(1)}\|_\infty \equiv h \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$, $\gamma_0^{(1)}$ follows subsection 7.2 with $s = \{0.05, 0.1, 0.5, 1, 5\}$.

Source 2 generates with:

$n_2 = 300$, same dimensionality attribution as source 1, but with $\beta_{(1)_b}^{(2)}$ satisfying $\|\beta_{(1)_b} - \beta_{(1)_b}^{(2)}\|_\infty = h$ and no requirement for $\beta_{(1)_a}^{(2)}$, which illustrates its heterogeneity.

Under 20 simulations for each case, by comparing RMSE of estimators under DML debiasing and with or without source data transfer, precision of estimators as well as the measure of 'max.diff's for estimation performance of dissimilar subset are given, i.e.

$$\text{max.diff.s\_1} \equiv \max_{i=\{1,2,3\}} \left|\beta_{(1)_{b,i}}^{(1)} - \widehat{\beta}_{(1)_{b,i}}^{(1)}\right| = \left\|\beta_{(1)_b}^{(1)} - \widehat{\beta}_{(1)_b}^{(1)}\right\|_\infty$$

$$\text{max.diff.s\_2} \equiv \left|\beta_{(2)_a}^{(2)} - \widehat{\beta}_{(2)_a}^{(2)}\right| = \left\|\beta_{(2)_a}^{(2)} - \widehat{\beta}_{(2)_a}^{(2)}\right\|_\infty$$

Utilizing **Meta-Algorithm-3.1**, the results are as follow, the results are summarized below, along with a comparison to the Partial Trans-Lasso method. Notably, in this context, the weight vector w_*, as in (23), adapts to $(0, 0, 0, 0, 1, 1, ..., 1)^\top$. More simulations could be found in the appendix D.2.

$s = 0.05$	Transferred	Debiased	$h = 0$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7902		
Non-Trans DML	✗	✓	0.4714		
Partial Trans-Lasso	✓	✗	1.3901		
PTL	✓	✓	**0.1253**	0.1323	0.0772
			$h = 0.05$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7853		
Non-Trans DML	✗	✓	0.4732		
Partial Trans-Lasso	✓	✗	1.5821		
PTL	✓	✓	**0.1542**	0.1224	0.1029
			$h = 0.1$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7711		
Non-Trans DML	✗	✓	0.4585		
Partial Trans-Lasso	✓	✗	2.0903		
PTL	✓	✓	**0.0938**	0.1083	0.0827

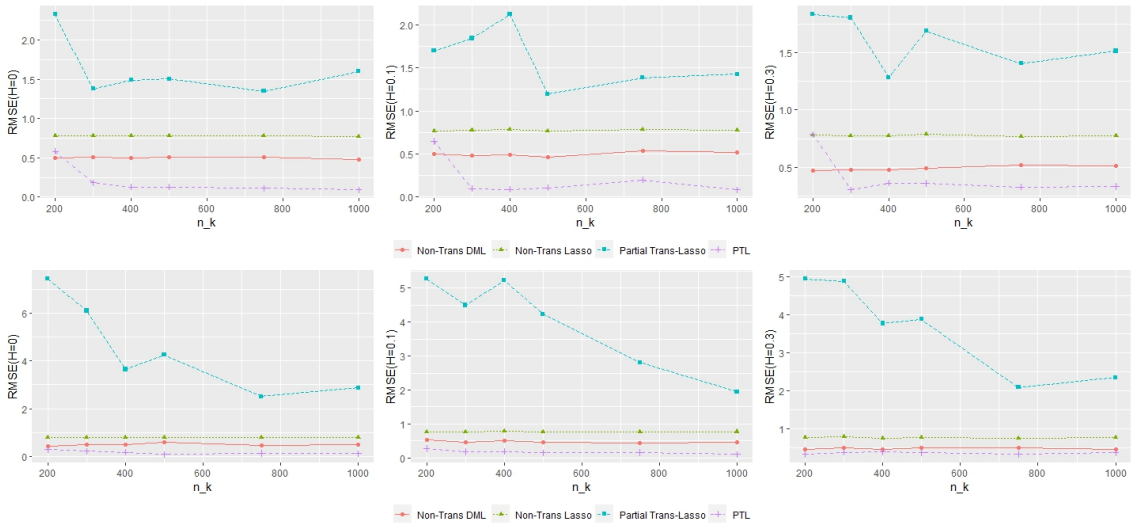
$s = 0.05$	Transferred	Debiased	$h = 0.2$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7836		
Non-Trans DML	✗	✓	0.5065		
Partial Trans-Lasso	✓	✗	1.9340		
PTL	✓	✓	**0.2032**	0.1039	0.0899
			$h = 0.3$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7919		
Non-Trans DML	✗	✓	0.5168		
Partial Trans-Lasso	✓	✗	2.1399		
PTL	✓	✓	**0.3387**	0.0850	0.0889
			$h = 0.5$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7837		
Non-Trans DML	✗	✓	**0.4591**		
Partial Trans-Lasso	✓	✗	2.2585		
PTL	✓	✓	0.6821	0.1186	0.0940

The above tables show: Firstly, robustness in non-sharp n_0 or minor h is revealed for PTL. Secondly, there is a consistent pattern observed in the decreasing trend of RMSE across varying values of n_k, h, s when compared to the homogeneous case. This suggests that improving the size of source sample, diminishing h and reducing s all contribute to more accurate estimates, aligning with the trend observed in the homogeneous scenario. Thirdly, a distinct trend emerges where the malfunction of DML occurs earlier, signaled by smaller values of s, in contrast to its homogeneous counterpart. This effect is attributed to the swifter dominance of the back-door path when causal signals are further obscured by disparate sources, underscoring the importance of maintaining clarity in transfer learning scenarios. Related details could be found in the appendix D.2.

On the other hand, I also simulate under the context of covariate shifting, where $X_i^{(0)} \overset{i.i.d.}{\sim} N_q(\mu_q^{(0)}, \mathbf{\Sigma}_{q\times q})$ and $X_i^{(k)} \overset{i.i.d.}{\sim} N_q(\mu_q^{(k)}, \mathbf{\Sigma}_{q\times q})$ for $k \in \{1, 2\}$ with distribution shift $\|\mu_q^{(0)} - \mu_q^{(1)}\|_\infty = 5, \|\mu_q^{(0)} - \mu_q^{(2)}\|_\infty = 10$. 20 times of simulation reveal the following results, which is similar with non-shifting case.

$s = 0.05$	Transferred	Debiased	$h = 0$	$h = 0.05$	$h = 0.1$
Non-Trans Lasso	✗	✗	0.7804	0.7894	0.7606
Non-Trans DML	✗	✓	0.5053	0.4729	0.4832
Partial Trans-Lasso	✓	✗	2.4033	2.9191	2.2174
PTL	✓	✓	**0.1496**	**0.1499**	**0.1324**
			$h = 0.2$	$h = 0.3$	$h = 0.5$
Non-Trans Lasso	✗	✗	0.7883	0.7797	0.7664
Non-Trans DML	✗	✓	0.4786	0.4940	**0.4830**
Partial Trans-Lasso	✓	✗	2.1831	2.4523	1.6958
PTL	✓	✓	**0.1749**	**0.3469**	0.6928

The following graph depicts the case of heterogeneous PTL without shift (first row) and with shift (second row) compared other methods :



8 Application

This section intends to exemplify PTL's enhanced prediction performance in real-data scenario. Here I typically inherit Li *et al.* (2022a), which contains subsamples of Genotype-Tissue Expression (GTEx) data (`https://gtexportal.org/`). The overall GTEx database combines 1,207,976 observations of 38,187 genes, obtained from 838 donors of 49 human tissues, revealing expression levels among various genes over different types of tissues. In this study, I focus on a specific subset of genes integrated as MODULE_137 (`https://www.gsea-msigdb.org/gsea/msigdb/cards/MODULE_137.html`), which serves the purpose of interacting with Central Nervous System (CNS) in human brains and is complemented by 545 covariate genes in total. On the other hand, tissues related to human brains are picked out automatically, including Amygdala, Anterior Cingulate Cortex, Caudate, Cerebellar Hemisphere, Cerebellum, Cortex, Frontal Cortex, Hippocampus, Hippocampus, Nucleus Accumbens, Putamen, Spinal Cord and Substantia Nigra, whose categorization aligns with the Tissue Sampling Sites in GTEx (`https://gtexportal.org/home/samplingSitePage`), the overall sample size of the aforementioned subsample amounts to 2,642.

8.1. Analytical Strategy

In the pursuit of unraveling the intricate functioning within the CNS, researchers place a particular emphasis on scrutinizing potential relationships among the expression levels of genes, especially within the previously mentioned brain tissues. From a statistical perspective, a fundamental priority involves the prediction of a target gene's expression level using other CNS genes as covariate.

In line with the framework introduced by Li *et al.* (2022a), the gene `JAM2`, situated on chromosome 21 and encoding a protein crucial for lymphocyte homing to secondary lymphoid organs (Johnson-Léger *et al.*, 2002), serves as the chosen response, whose mutation also lays a potential threatening factor for primary familial brain calcification (Cen *et al.*, 2020; Schottlaender *et al.*, 2020). Consequently, the aim is to explore the concealed relationships between `JAM2` and other genes from MODULE_137, hoping for a better prediction performance in-between.

Given the diverse brain tissues mentioned earlier, each is considered individually as a target, while the remaining brain tissues are designated as potential sources to be detected. Initially, I presumed that all sources are informative since the dataset primarily encompasses brain tissues, aligning with the notion of in-distribution transfer. The resulting predictions validate this assumption to some degree, leading to the application of the following strategy for the selection of partial transferring:

Strategy-8.1 (Separation)

1) *Apply Lasso to combined source data* $S^{(k)}$ *and get all parameters'estimation* $\hat{\theta}^{(k)} \equiv (\hat{\rho}^{(k)}, \hat{\beta}^{(k)^\top})^\top$*;*

2) *Apply Lasso to mixed target and source data* $S^{(0k)}$ *and get all parameters'estimation* $\hat{\theta}^{(0k)}$*;*

3) *For* $|\hat{\theta}_j^{(k)} - \hat{\theta}_j^{(0k)}| \leq h_{0k},\ j \in \{1, ..., p+q\}$*, assign them as the transferring nuisance.*

where h_{0k} is set to be 0.2 based on the simulation result. The rationale behind lies in the consideration of exterior attack, where the target sample is viewed as an impact that potentially emanates from DGPs without covariates of similar subsets. Consequently, only those with sustainable estimation could ultimately be considered in-distribution nuisance.

Based on 13 target samples with average size of 203, a cross-validated prediction error could be calculated, **Meta-Algorithm-1.1** and **Meta-Algorithm-2.3** are then applied here for prediction purpose.

8.2. Prediction Performance of PTL for `JAM2` Expression

This subsection evaluates the prediction performance of All-Trans PTL and Detected PTL in comparison to Non-Trans DML by calculating their relative prediction errors. Additionally, to provide a clearer illustration of the degree of improvement, I present the comparison ratios between Partial Trans-Lasso and Partial Lasso, drawing upon insights from prior literature.

The first modeling path is centered around the baseline model of Non-Trans DML, which exclusively employs the DML algorithm on the target sample, lacking the informative knowledge partially from the detected source. This baseline is then compared to two modeling strategies: All-Trans PTL and Detected PTL. The former considers all sources as transferrable and applies **Meta-Algorithm-2.3** for prediction, without the step of detection, however, such consideration makes some sense given that all sources come from other brain tissues, which may avoid negativeness from transferring to some degree. And on the other hand, the second approach inserts an additional source detection layer by applying **Meta-Algorithm-1.1**. The following table presents the relative prediction performance ratios and the outcomes of the detection process.

The second modeling path aims at better illustration through comparing prediction performance with previous established methods of Trans-Lasso in Li *et al.* (2022a). However, due to prior knowledge of partial

transferring, here Partial Trans-Lasso method modifies Trans-Lasso method as in the simulation:

$$\hat{\delta}^{\mathcal{A}} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{2n_0} \left\| y^{(0)} - X^{(0)} \left(\hat{w}^{\mathcal{A}} + \delta \right) \right\|_2^2 + \lambda_\delta \| w_*^\top \delta \|_1 \right\}$$

with $w_* \equiv (0, 1, 1, ..., 1)^\top$ for fair comparison. Based on detected source, compare Partial Trans-Lasso with Adaptive Lasso with penalty factor being w_* based merely on target sample, the relative prediction error rate is also present in the table below.

target	pre-specified causal	detected source	All-Trans PTL	Partial Trans-Lasso	Detected PTL
1	CYP2F1	2,4,8,9,10	1.1640	1.1138	**0.6833**
2	CYP2F1	1,5,8	1.0284	1.3874	**0.8554**
3	CYP2F1	7,12,13	0.8515	1.1529	**0.8136**
4	CYP2F1	Null	—	—	—
5	Null	—	—	—	—
6	APOC4	1,3,7,8,9,11,13	0.9094	1.0544	**0.5915**
7	CYP2F1	3,6,11,12,13	0.7975	0.8176	**0.7310**
8	Null	—	—	—	—
9	CYP2F1	1,5,6,8,10,11	1.0045	1.6802	**0.8537**
10	CYP2F1	1,8,9	0.9433	0.8858	**0.8232**
11	CYP2F1	6,7,9	0.9908	1.2003	**0.8981**
12	CYP2F1	3,7,13	0.7030	0.8429	**0.6672**
13	CYP2F1	3,7,12	0.9726	1.0760	**0.8179**

where number 1-13 represents the overall tissue types in brain, for 'Null' element in the second column, it refers to $|\hat{\theta}_j^{(k)} - \hat{\theta}_j^{(0k)}| \leq h_{0k},\ \forall j \in \{1, ..., p+q\}$, which fits into the scenario of Li *et al.* (2022a) and thus is no need to apply PTL. For 'Null' element in the third column, it denotes no sources among the rest 12 data sets are detected by **Meta-Algorithm-1.1**, where transfer learning is certainly out of use.

As the table clearly illustrates, the Non-Trans methods struggle to make accurate predictions in most instances, primarily due to the lack of sufficient information within the pure target dataset, where the largest sample size is 255. Conversely, the indiscriminate transfer approach of All-Trans PTL, while still yielding some positive transfer effects, tends to introduce overdue impacts through the transfer process, leading to eventual negativeness. Additionally, the decrease in prediction accuracy observed in the Partial Trans-Lasso method might be attributed to the presence of underlying confounding structures among the genes.

In stark contrast, Detected PTL consistently outperforms the other methods in all detectable cases. It achieves an average improvement of 22.65% in prediction accuracy compared to Non-Trans DML, a 16.30% reduction in the relative error rate when compared to All-Trans PTL, and a remarkable 34.76% enhancement in error rate performance in comparison to Partial Trans-Lasso. These results underscore the overall superior performance of the de-confounded partial knowledge transfer approach.

9 Discussion

This article delves into the statistical modeling strategy within a scenario where additional datasets serve as references for the label-sparse target set. This analysis is carried out while considering the presence of partially mismatched parameters and pervasive confounding factors in high-dimensional linear settings. The subset of parameters that is transferable are referred to as 'nuisance', while those representing the causal aspects are considered dissimilar between the target and source data. The proposed approach centers on a informative criterion with identical nuisance and it encompasses the development of sequential detection and estimation algorithms, focusing on the causal parameters. Simulations and the application

of this methodology to GTEx dataset consistently demonstrate the efficiency and improved performance in parameter estimation and response prediction, particularly for the detected PTL approach.

Among ongoing transfer learning literature, certain restrictions based on parameter similarities can be overly stringent and thus challenging to apply, while PTL presents a viable solution through leveraging prior knowledge of the partial dissimilarity. Nevertheless, there are scenarios where such field-specific knowledge is lacking, and the computational cost associated with causal identification under these circumstances can be formidable. Hence, the open question of addressing the matching problem without the aid of prior knowledge remains a topic for future research.

On the other hand, a data-driven approach rooted in the Trans-Lasso framework may offer a solution to the identification issue. By introducing an adaptive regularization term within the penalty, the transfer process can be made automatically 'partial', akin to the modification outlined in equation (35), albeit with a non-constant weight vector, denoted as w_*. This avenue presents an intriguing direction for further research, although it may require more detailed empirical interpretations to fully grasp its practical implications.

Furthermore, it is worth noting that modeling partial knowledge transfer without the necessity of linearization also warrants exploration. A notable example in this direction is the work of Liao *et al.* (2023) on nonparametric transferring on partial information, which leveraged prognostic adjustment to transfer nuisance statistics between historical source and targeted trial data. While this approach serves for improved prediction and coverage of Average Treatment Effect, it may lack proper interpretation, particularly since their formulation involves ignored parameter estimation issue within the super learner framework. As a result, there is a need for an eclectic semi-parametric modeling strategy that can strike a balance between prediction and interpretation. This, therefore, presents a compelling field for future research and further investigations in this direction are eagerly anticipated.

REFERENCE

[1] Cai, 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.

[2] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 1, 1.

[3] Eraslan, G., D. Eugene, A. Shankara, F. Evgenij, S. Ayshwarya, E. Fiskin, A. Subramanian, M. Slyper, J. Wang, N. Wittenberghe, J. Rouhana, J. Waldman, O. Ashenberg, M. Lek, D. Dionne, T. Win, M. Cuoco, O. Kuksenko, A. Tsankov, P. Branton, J. Marshall, A. Greka, G. Getz, A. Segrè, F. Aguet, O. Rosen, K. Ardlie, and A. Regev (2016). Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science*, 376(6594), 4290.

[4] Efron, B., and R. Tibshirani (1994). An Introduction to the Bootstrap. *Chapman and Hall/CRC*.

[5] Fan, J., S. Guo, and N. Hao (2012). Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression. *Journal of the Royal Statistical Society: Series B*, 74(1), 37-65.

[6] Fan, J., and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348-1360.

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

[8] Li, S., T. Cai, and H. Li (2022b). Transfer learning in large-scale graphical models with false discovery rate control. *Journal of the American Statistical Association (accepted)*.

[9] Li, S., T. Cai, and R. Duan (2023a). Targeting underrepresented populations in precision medicine: a federated transfer learning approach. *Annals of Applied Statistics (accepted)*.

[10] Li, S., L. Zhang, T. Cai, and H. Li (2023b). Estimation and inference in high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association (accepted)*.

[11] Liao, L., A. E. Hubbard, and A. Schuler (2023). Transfer Learning With Efficient Estimators to Optimally Leverage Historical Data in Analysis of Randomized Trials. *arXiv:2305.19180*.

[12] Liu, Y., S. He, Y. Chen, Y. Liu, F. Feng, W. Liu, Q. Guo, L. Zhao, and H. Sun (2020). Overview of AKR1C3: inhibitor achievements and disease insights. *Journal of Medicinal Chemistry*, 63(20), 11305-11329.

[13] Ma, C., M. Wu, and S. Ma (2022). Analysis of cancer omics data: a selective review of statistical techniques. *Briefings in Bioinformatics*, 23(2), 585.

[14] Pan, S., and Q. Yang (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

[15] Robinson, M. (1988). Root-N-consistent semi-parametric regression. *Econometrica*, 56: 931-54.

[16] Shalit, U., D. Fredrik, and S. David (2017). Estimating individual treatment effect: generalization bounds and algorithms. *International Conference on Machine Learning*.

[17] Strickler, J., T. Yoshino, R. Graham, S. Siena, and T. Bekaii-Saab (2022). Diagnosis and treatment of ERBB2-positive metastatic colorectal cancer: A review. *JAMA Oncology*, 8(5), 760 – 769.

[18] Tian, Y., and Y. Feng (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2022.2071278.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267-288.

[20] Wang, F., D. Liang, Y. Li, and S. Ma (2023). Prior information-assisted integrative analysis of multiple datasets. *Bioinformatics*, 39(8), 452.

[21] Wang, X., J. Oliva, J. Schneider, and B. Póczos (2016). Nonparametric risk and stability analysis for multi-task learning problems. *IJCAI*, 2146-2152.

[22] Wang, S., M. Wu, and S. Ma (2019). Integrative Analysis of Cancer Omics Data for Prognosis Modeling. *Genes*, 10(8), 604.

[23] Weiss, K., T. Khoshgoftaar, and D. Wang (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 9.

[24] Zhao, P., and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541-2563.

APPENDIX

A Large Sample Properties for PTL on Homogeneous Sources

A.1. Condition

(C1): Denote sample size as $n_k,\ k \in \{0, 1, ..., K\}$, $\hat{\rho}_{\text{DML}}$ is the causal estimator through DML with *i.i.d.* sample $S_i^{(k)} \equiv \{Y_i, X_i\}_{i=1}^{n_k}$, then under DGP of (1)-(4):

$$\hat{\rho}_{\text{DML}}^{(k)} - \rho_0^{(k)} = O_p(n_k^{-\frac{1}{2}})$$

which is given by Chernozhukov *et al.* (2018). Note that this is a higher order condition, which also promises well-functioning of DML estimators under proper s_{n_k} and other regular assumptions.

(C2): $\forall k \in \{1, ..., K\},\ n_k \to \infty$ and $n_0 = o(n_k)$.

(C3): $\forall k \in \{0, ..., K\}$, $\mathbb{E}\left(X_i^{(k)}\right) = \mu^{(k)} < \infty$, $\mathbb{E}\left(D_i^{(k)}\right) < \infty$, $\mathbb{E}\left(X_i^{(0)} X_{i,\mathcal{J}_k}^{(0)\top}\right) < \infty$, also $\mathbb{E}\left(X_i^{(0)} X_i^{(0)\top}\right) < \infty$, $\mathbb{E}\left(D_i^{(0)} D_i^{(0)\top}\right) < \infty$ nonsingular, $\mathbb{E}\left(U_i^{(0)} U_i^{(0)\top}\right) < \infty$ nonsingular.

(C4): Condition (5) of partially identical nuisance is satisfied.

(C5): Based on condition (A)-(C) and Theorem 1 in Fan and Li (2001), we have

$$\left\| \hat{\beta}_{\text{SCAD}}^{(k)} - \beta_0^{(k)} \right\| = O_p(n_k^{-\frac{1}{2}} + a_n) \tag{36}$$

for k_{th} source after consistent de-confounding, where $a_n = \max\left\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\right\} \to 0$ as $n_k \to \infty$ under the SCAD penalty.

(C6): $\Sigma_0 \equiv \mathbb{V}\left(D_i^{(0)} V_i^{(0)}\right) = \mathbb{E}\left(D_i^{(0)} D_i^{(0)\top} V_i^{(0)2}\right) < \infty$ positive definite.

A.2. Theorem

Theorem 1.1 (PTL on homogeneous sources without shift)
Under condition *(C1)-(C4)*,

$$\hat{\rho}_{\text{PTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}_{\text{PTL}} \xrightarrow{p} \rho_0$$

as $n_0 \to \infty$.

Proof. Without loss of generality, consider single positive source $K = 1$, based on *i.i.d.* DGP

$$\begin{aligned}
\hat{\rho}_{\text{PTL}} - \rho_0 &= \frac{\widehat{\mathbb{E}}\left[Y_i^{(0)} - \widehat{\mathbb{E}}\left(Y_i^{(1)} - D_i^{(1)}\hat{\rho}^{(1)}\right)\right]}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} - \rho_0 \\
&= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\rho_0 D_i^{(0)} + \beta_0^\top X_i^{(0)} + V_i^{(0)}\right)}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} - \rho_0 - \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\left[\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) D_i^{(1)} + \beta_0^{(1)\top} X_i^{(1)} + V_i^{(1)}\right]}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \\
&= \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\beta_0^\top X_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1}\beta_0^{(1)\top} X_i^{(1)}\right] - \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) D_i^{(1)}\right] \\
&+ \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1} V_i^{(1)}\right] \\
&= \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\beta_0^\top\left[\mu^{(0)} - \mu^{(1)} + o_p(1)\right] - \frac{\mathbb{E}\left(D_i^{(1)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1} V_i^{(1)}\right] \\
&= o_p(1) + o_p(1) + o_p(1) + o_p(1) = o_p(1)
\end{aligned} \tag{37}$$

□

Theorem 1.2 (PTL on homogeneous sources with shift)
Under condition *(C1)-(C5)*,

$$\hat{\rho}*_{\mathrm{PTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}*_{\mathrm{PTL}} \xrightarrow{p} \rho_0$$

$$\sqrt{n_0}\left(\hat{\rho}*_{\mathrm{PTL}} - \rho_0\right) \xrightarrow{d} N(0, c_1\sigma_0^2)$$

as $n_0 \to \infty$, where σ_0^2 denotes the variance of noise V_i, $c_1 \equiv \frac{1}{\mathbb{E}^2\left(D_i^{(0)}\right)}$ constant, the superscript $*$ denotes the existence of nuisance shift.

Proof.

$$\begin{aligned}
\hat{\rho}*_{\mathrm{PTL}} - \rho_0 &= \frac{\widehat{\mathbb{E}}\left\{Y_i^{(0)} - \left[\widehat{\mathbb{E}}\left(Y_i^{(1)} - D_i^{(1)}\hat{\rho}^{(1)}\right) - \widehat{\Delta}^\top\hat{\beta}^{(1)}\right]\right\}}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} - \rho_0 \\
&= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\rho_0 D_i^{(0)} + \beta_0^\top X_i^{(0)} + V_i^{(0)}\right)}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} - \rho_0 \\
&- \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\left[\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right)D_i^{(1)} + {\beta_0^{(1)}}^\top X_i^{(1)} + V_i^{(1)}\right] - {\hat{\beta}^{(1)}}^\top\widehat{\Delta}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \\
&= \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\beta_0^\top X_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1}\left({\beta_0^{(1)}}^\top X_i^{(1)} - {\hat{\beta}^{(1)}}^\top\widehat{\Delta}\right)\right] \\
&+ \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1} V_i^{(1)}\right] - \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right)D_i^{(1)}\right] \\
&= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\beta_0^\top\left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right) - {\hat{\beta}^{(1)}}^\top\left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right) + {\hat{\beta}^{(1)}}^\top\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}\right) - {\beta_0^{(1)}}^\top\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}\right)\right] \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1} V_i^{(1)}\right] - \frac{\mathbb{E}\left(D_i^{(1)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) \\
&= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}\right)^\top\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) - \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right)^\top\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)\right] \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1} V_i^{(1)}\right] - \frac{\mathbb{E}\left(D_i^{(1)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) \\
&= o_p(1) + o_p(1) + o_p(1) + o_p(1) + o_p(1) = o_p(1)
\end{aligned} \tag{38}$$

under condition *(C1)-(C5)*, also promised by cross-fitting:

$$\begin{aligned}
\sqrt{n_0}\left(\hat{\rho}*_{\mathrm{PTL}} - \rho_0\right) &= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}\right)^\top\frac{\sqrt{n_0}}{\sqrt{n_1}}\sqrt{n_1}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) - \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right)^\top\frac{\sqrt{n_0}}{\sqrt{n_1}}\sqrt{n_1}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)\right] \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{\sqrt{n_0}}{\sqrt{n_1}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} V_i^{(1)}\right] - \frac{\mathbb{E}\left(D_i^{(1)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\frac{\sqrt{n_0}}{\sqrt{n_1}}\sqrt{n_1}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) \\
&= o_p(1) + o_p(1) + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)}\right) + o_p(1) + o_p(1) \\
&\xrightarrow{d} N(0, c_1\sigma_0^2)
\end{aligned} \tag{39}$$

as $n_0 \to \infty$. □

Corollary 1 (Homogeneous PTL for general cases)
Under condition *(C1)-(C6)*,

$$\hat{\rho}^*_{\mathrm{PTL,G}} - \rho_0 \xrightarrow{p} \widehat{Q}^{-1} \frac{1}{n_0} \sum_{i=1}^{n_0} D_i^{(0)} V_i^{(0)}$$

as $n_k \to \infty,\ n_0 = O(1)$, where $\widehat{Q}^{-1} \equiv \frac{1}{n_0} \sum_{i=1}^{n_0} \left(D_i^{(0)} D_i^{(0)^\top} \right)$. Further,

$$\hat{\rho}^*_{\mathrm{PTL},G} - \rho_0 \xrightarrow{p} 0$$

$$\sqrt{n_0} \left(\hat{\rho}^*_{\mathrm{PTL},G} - \rho_0 \right) \xrightarrow{d} N\left(0, \sigma_0^2 Q^{-1}\right)$$

as $n_0 \to \infty$ for independent random noises, where $Q \equiv \mathbb{E}\left(D_i^{(0)} D_i^{(0)^\top} \right)$ invertible, subscript G indicates the general case. Also,

$$\begin{aligned}
\frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \left(\widehat{Y}_i^{(0)} - Y_i^{(0)} \right) &= \frac{1}{n_0} \sum_{i=1}^{n_0} D_i^{(0)^\top} \sqrt{n_0} \left(\hat{\rho}^*_{\mathrm{PTL},G} - \rho_0 \right) + \frac{1}{n_0} \sum_{i=1}^{n_0} X_i^{(0)^\top} \sqrt{n_0} \left(\hat{\beta}^{(k)} - \beta_0 \right) - \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} V_i^{(0)} \\
&= \frac{1}{n_0} \sum_{i=1}^{n_0} D_i^{(0)^\top} \sqrt{n_0} \left(\hat{\rho}^*_{\mathrm{PTL},G} - \rho_0 \right) + \frac{1}{n_0} \sum_{i=1}^{n_0} X_i^{(0)^\top} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k} \left(\hat{\beta}^{(k)} - \beta_0 \right) - \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} V_i^{(0)} \\
&\xrightarrow{d} N\left(0, \sigma_0^2 (1 + C_1^\top Q^{-1} C_1)\right)
\end{aligned} \tag{40}$$

where $C_1 \equiv \mathbb{E}\left(D_i^{(0)} \right)$.

Proof.

$$\begin{aligned}
\hat{\rho}^*_{\mathrm{PTL},G} - \rho_0 &= \left(\mathbf{D}^{(0)^\top} \mathbf{D}^{(0)} \right)^{-1} \mathbf{D}^{(0)^\top} \widehat{Y}^{*(0)} - \rho_0 \\
&= \left(\mathbf{D}^{(0)^\top} \mathbf{D}^{(0)} \right)^{-1} \mathbf{D}^{(0)^\top} \left(Y^{(0)} - \mathbf{X}^{(0)} \hat{\beta}^{(1)} \right) - \rho_0 \\
&= \left(\mathbf{D}^{(0)^\top} \mathbf{D}^{(0)} \right)^{-1} \mathbf{D}^{(0)^\top} \left(\mathbf{X}^{(0)} \beta_0 - \mathbf{X}^{(0)} \hat{\beta}^{(1)} + V^{(0)} \right) \\
&= \left(\mathbf{D}^{(0)^\top} \mathbf{D}^{(0)} \right)^{-1} \mathbf{D}^{(0)^\top} \mathbf{X}^{(0)} \left(\beta^{(1)} - \hat{\beta}^{(1)} \right) + \left(\mathbf{D}^{(0)^\top} \mathbf{D}^{(0)} \right)^{-1} \mathbf{D}^{(0)^\top} V^{(0)} \\
&\xrightarrow{p} 0
\end{aligned} \tag{41}$$

as $n_0 \to \infty$. Also for $\widehat{Q} \equiv \frac{1}{n_0} \sum_{i=1}^{n_0} (D_i^{(0)} D_i^{(0)^\top})$

$$\begin{aligned}
\sqrt{n_0}(\hat{\rho}^*_{\mathrm{PTL},G} - \rho_0) &= \widehat{Q}^{-1} \frac{\mathbf{D}^{(0)^\top} \mathbf{X}^{(0)}}{n_0} \frac{\sqrt{n_0}}{\sqrt{n_1}} \sqrt{n_1} \left(\beta^{(1)} - \hat{\beta}^{(1)} \right) + \widehat{Q}^{-1} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} D_i^{(0)} V_i^{(0)} \\
&\xrightarrow{d} N\left(0, Q^{-1} \Sigma_0 Q^{-1}\right) = N\left(0, \sigma_0^2 Q^{-1}\right)
\end{aligned} \tag{42}$$

as $n_0 \to \infty$ under *(C1)-(C6)*, where $\Sigma_0 \equiv \mathbb{V}(D_i^{(0)} V_i^{(0)}) = \mathbb{E}(D_i^{(0)} D_i^{(0)^\top} V_i^{(0)^2})$, $Q \equiv \mathbb{E}(D_i^{(0)} D_i^{(0)^\top})$ invertible. □

B Large Sample Properties for PTL on Heterogeneous Sources

Theorem 2.1 (PTL on heterogeneous sources without shift)
Under condition *(C1)-(C5)*,

$$\hat{\rho}_{\text{HPTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}_{\text{HPTL}} - \rho_0 \xrightarrow{p} 0$$
$$\sqrt{n_0}\left(\hat{\rho}_{\text{HPTL}} - \rho_0\right) \xrightarrow{d} N\left(0, c_1\sigma_0^2\right)$$

as $n_0 \to \infty$.

Proof. Since

$$\begin{aligned}
\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^{(k)\top}\beta_0^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) &= \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\sum_{j=1}^{q} x_{ij}^{(k)}\beta_{0j}^{(k)} - \sum_{j\notin\mathcal{J}_k} x_{ij}^{(k)}\hat{\beta}_j^{(k)}\right)\\
&= \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left[\sum_{j=1}^{q} x_{ij}^{(k)}\beta_{0j}^{(k)} - \sum_{j\notin\mathcal{J}_k} x_{ij}^{(k)}\left(\beta_{0j}^{(k)} + \hat{\beta}_j^{(k)} - \beta_{0j}^{(k)}\right)\right]\\
&= \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left[\sum_{j\in\mathcal{J}_k} x_{ij}^{(k)}\beta_{0j}^{(k)} - \sum_{j\notin\mathcal{J}_k} x_{ij}^{(k)}\left(\hat{\beta}_j^{(k)} - \beta_{0j}^{(k)}\right)\right]\\
&= \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left[X_{i,\mathcal{J}_k}^{(k)\top}\beta_{0,\mathcal{J}_k}^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right)\right]
\end{aligned} \tag{43}$$

Thus

$$\begin{aligned}
\hat{\rho}_{\text{HPTL}} - \rho_0 &= \frac{\widehat{\mathbb{E}}\left[Y_i^{(0)} - \sum_{k=1}^{K}\widehat{\mathbb{E}}\left(Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right)\right]}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} - \rho_0\\
&= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(D_i^{(0)}\rho_0 + X_i^{(0)\top}\beta_0 + V_i^{(0)}\right)}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} - \rho_0\\
&- \frac{\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left[D_i^{(k)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) + \left(X_i^{(k)\top}\beta_0^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) + V_i^{(k)}\right]}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\\
&= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \times \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right)^{\top}\beta_0 - \sum_{k=1}^{K}\frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right)\\
&- \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^{(k)\top}\beta_0^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right)\\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} V_i^{(k)}\right)\\
&= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_{i,\mathcal{J}_k}^{(0)} - \frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,\mathcal{J}_k}^{(k)}\right)^{\top}\left(\beta_{0,\mathcal{J}_k} - \beta_{0,\mathcal{J}_k}^{(k)}\right)\\
&- \sum_{k=1}^{K}\frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,-\mathcal{J}_k}^{(k)}\right)^{\top}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right)\\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} V_i^{(k)}\right)\\
&= o_p(1) + o_p(1) + o_p(1) + o_p(1) + o_p(1) = o_p(1)
\end{aligned} \tag{44}$$

since $\cup_{k=1}^{K}\mathcal{J}_k = \{1, 2, ..., q\}$, then similarly

$$\begin{aligned}
\sqrt{n_0}(\hat{\rho}_{\text{HPTL}} - \rho_0) = &-\sum_{k=1}^{K} \frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,-\mathcal{J}_k}^{(k)}\right)^{\top} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{\sqrt{n_0}}{\sqrt{n_k}} \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} V_i^{(k)}\right) \\
&= o_p(1) + o_p(1) + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)}\right) + o_p(1) \\
&\xrightarrow{d} N\left(0, c_1\sigma_0^2\right)
\end{aligned} \tag{45}$$

as $n_0 \to \infty$ under *(C1)-(C5)* and independence through cross-fitting. □

Theorem 2.2 (PTL on heterogeneous sources with shift)
Under condition *(C1)-(C5)*,

$$\hat{\rho}^*_{\text{HPTL}} - \rho_0 \xrightarrow{p} \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}^*_{\text{HPTL}} - \rho_0 \xrightarrow{p} 0$$
$$\sqrt{n_0}\left(\hat{\rho}^*_{\text{HPTL}} - \rho_0\right) \xrightarrow{d} N\left(0, c_1\sigma_0^2\right)$$

as $n_0 \to \infty$.

Proof.

$$\begin{aligned}
\hat{\rho}^*_{\text{HPTL}} - \rho_0 &= \frac{\widehat{\mathbb{E}}\left\{Y_i^{(0)} - \sum_{k=1}^{K}\left[\widehat{\mathbb{E}}\left(Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) - \widehat{\Delta}_{\mathcal{J}_k}^{(k)\top}\hat{\beta}_{\mathcal{J}_k}^{(k)}\right]\right\}}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} - \rho_0 \\
&= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(D_i^{(0)}\rho_0 + X_i^{(0)\top}\beta_0 + V_i^{(0)}\right)}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} - \rho_0 \\
&- \frac{\sum_{k=1}^{K}\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\left[D_i^{(k)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) + \left(X_i^{(k)\top}\beta_0^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) + V_i^{(k)}\right] - \widehat{\Delta}_{\mathcal{J}_k}^{(k)\top}\hat{\beta}_{\mathcal{J}_k}^{(k)}\right\}}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \\
&= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right)^{\top}\beta_0 - \sum_{k=1}^{K}\frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) \\
&- \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^{K}\left[\left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_i^{(k)}\right)^{\top}\beta_0^{(k)} - \left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,-\mathcal{J}_k}^{(k)}\right)^{\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,\mathcal{J}_k}^{(k)} - \frac{1}{n_0}\sum_{i=1}^{n_0} X_{i,\mathcal{J}_k}^{(0)}\right)^{\top}\hat{\beta}_{\mathcal{J}_k}^{(k)}\right] \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} V_i^{(k)}\right)
\end{aligned} \tag{46}$$

Then

$$
\begin{aligned}
\hat{\rho}^{*}_{\mathrm{HPTL}}-\rho_0 &= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}X_i^{(0)}\right)^{\top}\beta_0-\sum_{k=1}^{K}\frac{\mathbb{E}\left(D_i^{(k)}\right)+o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\rho_0^{(k)}-\hat{\rho}^{(k)}\right)\\
&-\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left[X_{i,\mathcal{J}_k}^{(k)\top}\beta_{0,\mathcal{J}_k}^{(k)}-X_{i,-\mathcal{J}_k}^{(k)\top}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)}-\beta_{0,-\mathcal{J}_k}^{(k)}\right)\right]\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left[\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,\mathcal{J}_k}^{(k)}-\frac{1}{n_0}\sum_{i=1}^{n_0}X_{i,\mathcal{J}_k}^{(0)}\right)^{\top}\hat{\beta}_{\mathcal{J}_k}^{(k)}\right]\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}V_i^{(0)}\right)-\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}V_i^{(k)}\right)\\
&=-\sum_{k=1}^{K}\frac{\mathbb{E}\left(D_i^{(k)}\right)+o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\rho_0^{(k)}-\hat{\rho}^{(k)}\right)+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,-\mathcal{J}_k}^{(k)}\right)^{\top}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)}-\beta_{0,-\mathcal{J}_k}^{(k)}\right)\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left[\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,\mathcal{J}_k}^{(k)}-\frac{1}{n_0}\sum_{i=1}^{n_0}X_{i,\mathcal{J}_k}^{(0)}\right)^{\top}\hat{\beta}_{\mathcal{J}_k}^{(k)}\right]\\
&-\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left[\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,\mathcal{J}_k}^{(k)}-\frac{1}{n_0}\sum_{i=1}^{n_0}X_{i,\mathcal{J}_k}^{(0)}\right)^{\top}\beta_{0,\mathcal{J}_k}^{(k)}\right]\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}V_i^{(0)}\right)-\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}V_i^{(k)}\right)\\
&=-\sum_{k=1}^{K}\frac{\mathbb{E}\left(D_i^{(k)}\right)+o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\rho_0^{(k)}-\hat{\rho}^{(k)}\right)+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,-\mathcal{J}_k}^{(k)}\right)^{\top}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)}-\beta_{0,-\mathcal{J}_k}^{(k)}\right)\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,\mathcal{J}_k}^{(k)}-\frac{1}{n_0}\sum_{i=1}^{n_0}X_{i,\mathcal{J}_k}^{(0)}\right)^{\top}\left(\hat{\beta}_{\mathcal{J}_k}^{(k)}-\beta_{0,\mathcal{J}_k}^{(k)}\right)\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}V_i^{(0)}\right)-\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}V_i^{(k)}\right)\\
&=o_p(1)+o_p(1)+o_p(1)+o_p(1)+o_p(1)=o_p(1)
\end{aligned}
\tag{47}
$$

Also,

$$
\begin{aligned}
\sqrt{n_0}(\hat{\rho}^{*}_{\mathrm{HPTL}}-\rho_0) &= -\sum_{k=1}^{K}\frac{\mathbb{E}\left(D_i^{(k)}\right)+o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\frac{\sqrt{n_0}}{\sqrt{n_k}}\sqrt{n_k}\left(\rho_0^{(k)}-\hat{\rho}^{(k)}\right)\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,-\mathcal{J}_k}^{(k)}\right)^{\top}\frac{\sqrt{n_0}}{\sqrt{n_k}}\sqrt{n_k}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)}-\beta_{0,-\mathcal{J}_k}^{(k)}\right)\\
&-\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}X_{i,\mathcal{J}_k}^{(k)}-\frac{1}{n_0}\sum_{i=1}^{n_0}X_{i,\mathcal{J}_k}^{(0)}\right)^{\top}\frac{\sqrt{n_0}}{\sqrt{n_k}}\sqrt{n_k}\left(\hat{\beta}_{\mathcal{J}_k}^{(k)}-\beta_{0,\mathcal{J}_k}^{(k)}\right)\\
&+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}V_i^{(0)}\right)-\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\sum_{k=1}^{K}\left(\frac{\sqrt{n_0}}{\sqrt{n_k}}\frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k}V_i^{(k)}\right)\\
&=o_p(1)+o_p(1)+o_p(1)+\frac{1}{\mathbb{E}\left(D_i^{(0)}\right)+o_p(1)}\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}V_i^{(0)}\right)+o_p(1)\xrightarrow{d}N\left(0,c_1\sigma_0^2\right)
\end{aligned}
\tag{48}
$$

as $n_0\to\infty$ under condition *(C1)-(C5)*, due to independent estimation between $\hat{\rho}^{(k)}$, $\hat{\beta}_{\mathcal{J}_k}^{(k)}$ and $\hat{\beta}_{-\mathcal{J}_k}^{(k)}$, for $k\in\{1,...,K\}$. □

Corollary 2 (Heterogeneous PTL for general cases)
Under condition *(C1)-(C6)*,

$$\hat{\rho}^*_{\mathrm{HPTL,G}} - \rho_0 \xrightarrow{p} \widehat{Q}^{-1} \frac{1}{n_0} \sum_{i=1}^{n_0} D_i^{(0)} V_i^{(0)}$$

as $n_k \to \infty,\ n_0 = O(1)$. Further,

$$\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0 \xrightarrow{p} 0$$
$$\sqrt{n_0}\left(\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0\right) \xrightarrow{d} N\left(0, \sigma_0^2 Q^{-1}\right)$$

as $n_0 \to \infty$. Also,

$$\begin{aligned}
\frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \left(\widehat{Y}_{i,H}^{(0)} - Y_{i,H}^{(0)}\right) &= \frac{1}{n_0} \sum_{i=1}^{n_0} D_i^{(0)\top} \sqrt{n_0}\left(\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0\right) + \sum_{k=1}^{K} \frac{1}{n_0} \sum_{i=1}^{n_0} X_{i,\mathcal{J}_k}^{(0)\top} \sqrt{n_0}\left(\hat{\beta}_{\mathcal{J}_k}^{(k)} - \beta_{0,\mathcal{J}_k}\right) - \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} V_i^{(0)} \\
&= \frac{1}{n_0} \sum_{i=1}^{n_0} D_i^{(0)\top} \sqrt{n_0}\left(\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0\right) + \sum_{k=1}^{K} \frac{1}{n_0} \sum_{i=1}^{n_0} X_{i,\mathcal{J}_k}^{(0)\top} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k}\left(\hat{\beta}_{\mathcal{J}_k}^{(k)} - \beta_{0,\mathcal{J}_k}\right) - \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} V_i^{(0)} \\
&\xrightarrow{d} N\left(0, \sigma_0^2 (1 + C_1^\top Q^{-1} C_1)\right)
\end{aligned} \tag{49}$$

as $n_0 \to \infty$.

Proof.

$$\begin{aligned}
\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0 &= \left(\mathbf{D}^{(0)\top}\mathbf{D}^{(0)}\right)^{-1} \mathbf{D}^{(0)\top} \widehat{Y}_H^{*(0)} - \rho_0 \\
&= \left(\mathbf{D}^{(0)\top}\mathbf{D}^{(0)}\right)^{-1} \mathbf{D}^{(0)\top} \left(Y^{(0)} - \sum_{k=1}^{K} \mathbf{X}_{\mathcal{J}_k}^{(0)} \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) - \rho_0 \\
&= \left(\mathbf{D}^{(0)\top}\mathbf{D}^{(0)}\right)^{-1} \mathbf{D}^{(0)\top} \left(\mathbf{X}^{(0)}\beta_0 - \sum_{k=1}^{K} \mathbf{X}_{\mathcal{J}_k}^{(0)} \hat{\beta}_{\mathcal{J}_k}^{(k)} + V^{(0)}\right) \\
&= \sum_{k=1}^{K} \left(\mathbf{D}^{(0)\top}\mathbf{D}^{(0)}\right)^{-1} \mathbf{D}^{(0)\top} \mathbf{X}_{\mathcal{J}_k}^{(0)} \left(\beta_{0,\mathcal{J}_k} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) \\
&+ \left(\mathbf{D}^{(0)\top}\mathbf{D}^{(0)}\right)^{-1} \mathbf{D}^{(0)\top} V^{(0)} \\
&= \sum_{k=1}^{K} \left(\mathbf{D}^{(0)\top}\mathbf{D}^{(0)}\right)^{-1} \mathbf{D}^{(0)\top} \mathbf{X}_{\mathcal{J}_k}^{(0)} \left(\beta_{0,\mathcal{J}_k}^{(k)} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) \\
&+ \left(\mathbf{D}^{(0)\top}\mathbf{D}^{(0)}\right)^{-1} \mathbf{D}^{(0)\top} V^{(0)} \\
&\xrightarrow{p} 0
\end{aligned} \tag{50}$$

as $n_0 \to \infty$. Also

$$\begin{aligned}
\sqrt{n_0}(\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0) &= \sum_{k=1}^{K} \widehat{Q}^{-1} \frac{\mathbf{D}^{(0)\top}\mathbf{X}_{\mathcal{J}_k}^{(0)}}{n_0} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k}\left(\beta_{0,\mathcal{J}_k}^{(k)} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) \\
&+ \widehat{Q}^{-1} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} D_i^{(0)} V_i^{(0)} \\
&\xrightarrow{d} N\left(0, Q^{-1}\Sigma_0 Q^{-1}\right) = N\left(0, \sigma_0^2 Q^{-1}\right)
\end{aligned} \tag{51}$$

as $n_0 \to \infty$ under *(C1)-(C6)*, where $\Sigma_0 \equiv \mathbb{V}(D_i^{(0)} V_i^{(0)}) = \mathbb{E}(D_i^{(0)} D_i^{(0)\top} V_i^{(0)^2})$, $Q \equiv \mathbb{E}(D^{(0)} D^{(0)\top})$ invertible. □

C Efficiency for PTL

Corollary 3 (Efficiency for PTL estimation)
Under condition *(C1)-(C5)*,

$$avar\left(\hat{\rho}*_{\mathrm{PTL}}\right) = avar\left(\hat{\rho}*_{\mathrm{HPTL}}\right) < avar\left(\hat{\rho}_{\mathrm{DML}}\right)$$

if

$$\mathbb{V}\left(\gamma_0^\top X_i^{(0)}\right) > \mathbb{V}\left(U_i^{(0)}\right)$$

Likewise for $p > 1$

$$avar\left(\hat{\rho}*_{\mathrm{PTL,G}}\right) = avar\left(\hat{\rho}*_{\mathrm{HPTL,G}}\right) < avar\left(\hat{\rho}_{\mathrm{DML,G}}\right)$$

if

$$\boldsymbol{\gamma}_0 \neq \mathbf{0}_{q\times p}$$

where $\mathbf{0}_{q\times p}$ denotes the zero matrix with dimension q by p.

Proof. For high-dimensional linear regression, its score function under least square loss is

$$\partial\ell_\theta^{(0)}(Y, D, X \mid \rho, \beta) = \left(Y^{(0)} - D^{(0)}\rho - {X^{(0)}}^\top \beta_0\right) D^{(0)} \tag{52}$$

$$\partial\ell_\beta^{(0)}(Y, D, X \mid \rho, \beta) = \left(Y^{(0)} - D^{(0)}\rho - {X^{(0)}}^\top \beta\right) X^{(0)} \tag{53}$$

Through approach of concentrating-out as in Chernozhukov *et al.* (2018), the Neyman orthogonal score is given by

$$\begin{aligned}\psi^{(0)}(Y, D, X \mid \rho, \beta, \gamma) &= \left(Y^{(0)} - D^{(0)}\rho - {X^{(0)}}^\top \beta\right)\left(D^{(0)} - {X^{(0)}}^\top \gamma\right) \\ &= -D^{(0)}\left(D^{(0)} - {X^{(0)}}^\top \gamma\right) * \rho + \left(Y^{(0)} - {X^{(0)}}^\top \beta\right)\left(D^{(0)} - {X^{(0)}}^\top \gamma\right) \\ &\equiv \psi_a^{(0)}(Y, D, X \mid \gamma) * \rho + \psi_b^{(0)}(Y, D, X \mid \beta, \gamma)\end{aligned} \tag{54}$$

Jacobian matrix

$$J^{(0)} \equiv \mathbb{E}\left[\psi_a^{(0)}\left(Y, D, X \mid \gamma = \gamma_0\right)\right] = -\mathbb{E}\left(D^{(0)}U^{(0)}\right) \tag{55}$$

or

$$J^{(0)} \equiv \partial_{\rho'}\left\{\mathbb{E}\left[\psi^{(0)}\left(Y, D, X \mid \rho, \beta = \beta_0, \gamma = \gamma_0\right)\right]\right\}\Big|_{\rho=\rho_0} = -\mathbb{E}\left(D^{(0)}U^{(0)}\right)$$

under exchangeability. Then without knowledge transfer, by *THEOREM 3.1* in Chernozhukov *et al.* (2018),

$$\sqrt{n_0}\left(\hat{\rho}_{\mathrm{DML}} - \rho_0\right) \xrightarrow{d} N\left(0, c_2\sigma_0^2\right) \tag{56}$$

as $n_0 \to \infty$, where $c_2 \equiv \frac{1}{\mathbb{V}\left(U_i^{(0)}\right)}$, since

$$J^{(0)} = -\mathbb{E}\left(D^{(0)}U^{(0)}\right) = -\mathbb{E}\left[\left(\gamma_0^\top X^{(0)} + U^{(0)}\right)U^{(0)}\right] = -\mathbb{E}\left({U^{(0)}}^2\right) \tag{57}$$

$$\mathbb{E}\left[\psi^{(0)}\left(Y, D, X \mid \rho_0, \beta_0, \gamma_0\right)\psi^{(0)}\left(Y, D, X \mid \rho_0, \beta_0, \gamma_0\right)^\top\right] = \mathbb{E}\left({U^{(0)}}^2{V^{(0)}}^2\right) = \mathbb{E}\left({U^{(0)}}^2\right)\mathbb{E}\left({V^{(0)}}^2\right) \tag{58}$$

by independence of random noises. Remind that for PTL, the asymptotic variance is with

$$c_1 \equiv \frac{1}{\mathbb{E}^2\left(D_i^{(0)}\right)} = \frac{1}{\mathbb{E}^2\left(\gamma_0^\top X_i^{(0)} + U_i^{(0)}\right)} = \frac{1}{\mathbb{E}^2\left(\gamma_0^\top X_i^{(0)}\right)} = \frac{1}{\mathbb{V}\left(\gamma_0^\top X_i^{(0)}\right)}$$

under scaling. Thus if

$$\mathbb{V}\left(\gamma_0^\top X_i^{(0)}\right) > \mathbb{V}\left(U_i^{(0)}\right)$$

enhanced efficiency then reveals.

For general cases when $p > 1$,

$$J^{(0)} = -\mathbb{E}\left(D^{(0)}U^{(0)^\top}\right) = -\mathbb{E}\left[\left(\gamma_0^\top X^{(0)} + U^{(0)}\right)U^{(0)^\top}\right] = -\mathbb{E}\left(U^{(0)}U^{(0)^\top}\right) \tag{59}$$

$$\mathbb{E}\left[\psi^{(0)}\left(Y, D, X \mid \rho_0, \beta_0, \gamma_0\right)\psi^{(0)}\left(Y, D, X \mid \rho_0, \beta_0, \gamma_0\right)^\top\right] = \mathbb{E}\left(U^{(0)}V^{(0)}V^{(0)^\top}U^{(0)^\top}\right) = \mathbb{E}\left(U^{(0)}U^{(0)^\top}\right)\sigma_0^2 \tag{60}$$

denote $\widetilde{Q} \equiv \mathbb{E}\left(U^{(0)}U^{(0)^\top}\right)$, the enhanced efficiency is achieved

$$avar\left(\hat{\rho}_{\mathrm{DML,G}}\right) = \mathbb{E}^{-1}\left(U^{(0)}U^{(0)^\top}\right)\sigma_0^2/n_0 \equiv \widetilde{Q}^{-1}\sigma_0^2/n_0 \geq Q^{-1}\sigma_0^2/n_0 = avar\left(\hat{\rho}*_{\mathrm{PTL,G}}\right) = avar\left(\hat{\rho}*_{\mathrm{HPTL,G}}\right) \tag{61}$$

since

$$\mathbb{E}\left(D^{(0)}D^{(0)^\top}\right) - \mathbb{E}\left(U^{(0)}U^{(0)^\top}\right) = \boldsymbol{\gamma}_0^\top\mathbb{E}\left(X^{(0)}X^{(0)^\top}\right)\boldsymbol{\gamma}_0 \geq 0 \tag{62}$$

Notice that efficiency is strictly improved if $\boldsymbol{\gamma}_0$ is not zero matrix, and the degree of which relies on the magnitude of linear coefficients. □

Corollary 4 (Efficiency for PTL prediction)
Under condition *(C1)-(C6)*,

$$avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_i^{(0)}\right) = avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_{i,\mathrm{H}}^{(0)}\right) < avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_{i,\mathrm{DML}}^{(0)}\right)$$

if

$$\boldsymbol{\gamma}_0 \neq \mathbf{0}_{q\times p}$$

Proof. From **Corollary 3**, it is easy to show that

$$\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\left(\widehat{Y}_{i,\mathrm{DML}}^{(0)} - Y_i^{(0)}\right) \xrightarrow{d} N\left(0, \sigma_0^2(1 + C_1^\top\widetilde{Q}^{-1}C_1)\right) \tag{63}$$

For

$$\boldsymbol{\gamma}_0 \neq \mathbf{0}_{q\times p}$$

similar to **Corollary 3**, PTL has

$$avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_i^{(0)}\right) = avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_{i,\mathrm{H}}^{(0)}\right) = \sigma_0^2(1+C_1^\top Q^{-1}C_1) < \sigma_0^2(1+C_1^\top\widetilde{Q}^{-1}C_1) = avar\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\widehat{Y}_{i,\mathrm{DML}}^{(0)}\right) \tag{64}$$

□

D More Simulation Results

D.1 Oracle PTL Under Homogeneous Sources

	$s = 0.1$	Transferred	Debiased	$H = 0.01$	$H = 0.05$	$H = 0.1$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.3318	1.1739	1.2778
	Non-Trans DML	✗	✓	0.7227	0.6888	0.6348
	Partial Trans-Lasso	✓	✗	1.3919	1.4382	1.4672
	PTL	✓	✓	**0.1721**	**0.2809**	**0.4205**
$n_k = 200$	Non-Trans Lasso	✗	✗	1.3324	1.2184	1.3351
	Non-Trans DML	✗	✓	0.6619	0.6560	0.6565
	Partial Trans-Lasso	✓	✗	1.0080	1.1253	1.0271
	PTL	✓	✓	**0.0958**	**0.2127**	**0.3606**
$n_k = 500$	Non-Trans Lasso	✗	✗	1.3368	1.2287	1.2615
	Non-Trans DML	✗	✓	0.6837	0.6972	0.6594
	Partial Trans-Lasso	✓	✗	0.6129	0.6180	0.7538
	PTL	✓	✓	**0.0571**	**0.1752**	**0.3303**

	$s = 0.1$	Transferred	Debiased	$H = 0.2$	$H = 0.3$	$H = 0.5$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2263	1.1316	1.3172
	Non-Trans DML	✗	✓	0.6463	**0.7033**	**0.6960**
	Partial Trans-Lasso	✓	✗	1.4383	1.3939	1.3712
	PTL	✓	✓	**0.6436**	0.9016	1.4334
$n_k = 200$	Non-Trans Lasso	✗	✗	1.4082	1.2507	1.3560
	Non-Trans DML	✗	✓	0.7180	**0.7270**	**0.6991**
	Partial Trans-Lasso	✓	✗	1.1984	1.2941	1.5383
	PTL	✓	✓	**0.5931**	0.8661	1.4019
$n_k = 500$	Non-Trans Lasso	✗	✗	1.2840	1.1741	1.2274
	Non-Trans DML	✗	✓	0.6291	**0.6324**	**0.6483**
	Partial Trans-Lasso	✓	✗	1.0115	1.2542	1.7153
	PTL	✓	✓	**0.5666**	0.8486	1.4038

	$s = 0.5$	Transferred	Debiased	$H = 0.01$	$H = 0.05$	$H = 0.1$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2433	1.2424	1.2475
	Non-Trans DML	✗	✓	0.6315	0.7385	0.7136
	Partial Trans-Lasso	✓	✗	1.6385	1.6866	1.6753
	PTL	✓	✓	**0.4528**	**0.5077**	**0.6864**
$n_k = 200$	Non-Trans Lasso	✗	✗	1.3509	1.2335	1.2745
	Non-Trans DML	✗	✓	0.7673	0.7167	0.7109
	Partial Trans-Lasso	✓	✗	1.5128	1.5499	1.6117
	PTL	✓	✓	**0.2108**	**0.3114**	**0.4515**
$n_k = 500$	Non-Trans Lasso	✗	✗	1.1781	1.2570	1.1802
	Non-Trans DML	✗	✓	0.7320	0.6925	0.6629
	Partial Trans-Lasso	✓	✗	1.4612	1.4908	1.5015
	PTL	✓	✓	**0.1312**	**0.2293**	**0.3910**

	$s = 0.5$	Transferred	Debiased	$H = 0.2$	$H = 0.3$	$H = 0.5$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2028	1.2471	1.2835
	Non-Trans DML	✗	✓	**0.6256**	**0.6923**	**0.7237**
	Partial Trans-Lasso	✓	✗	1.6379	1.7047	1.6912
	PTL	✓	✓	0.8687	1.1553	1.6863
$n_k = 200$	Non-Trans Lasso	✗	✗	1.1982	1.2160	1.3201
	Non-Trans DML	✗	✓	**0.6409**	**0.6937**	**0.6738**
	Partial Trans-Lasso	✓	✗	1.6370	1.6774	1.6682
	PTL	✓	✓	0.6803	0.9640	1.5130
$n_k = 500$	Non-Trans Lasso	✗	✗	1.2939	1.1675	1.2764
	Non-Trans DML	✗	✓	0.6702	**0.6711**	**0.7033**
	Partial Trans-Lasso	✓	✗	1.6085	1.6412	1.7811
	PTL	✓	✓	**0.6020**	0.8858	1.4485

	$s = 1$	Transferred	Debiased	$H = 0.01$	$H = 0.05$	$H = 0.1$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2639	1.1912	1.2227
	Non-Trans DML	✗	✓	0.6376	0.7253	**0.6499**
	Partial Trans-Lasso	✓	✗	1.6264	1.6356	1.6098
	PTL	✓	✓	**0.6149**	**0.7202**	0.7813
$n_k = 200$	Non-Trans Lasso	✗	✗	1.2579	1.1723	1.3631
	Non-Trans DML	✗	✓	0.7218	0.6546	0.7269
	Partial Trans-Lasso	✓	✗	1.5814	1.5320	1.6432
	PTL	✓	✓	**0.3034**	**0.3645**	**0.4999**
$n_k = 500$	Non-Trans Lasso	✗	✗	1.1996	1.2521	1.2535
	Non-Trans DML	✗	✓	0.6960	0.7276	0.6601
	Partial Trans-Lasso	✓	✗	1.5856	1.6017	1.5321
	PTL	✓	✓	**0.1629**	**0.2523**	**0.4195**

	$s = 1$	Transferred	Debiased	$H = 0.2$	$H = 0.3$	$H = 0.5$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.1927	1.1661	1.1908
	Non-Trans DML	✗	✓	**0.7077**	**0.6776**	**0.6689**
	Partial Trans-Lasso	✓	✗	1.6413	1.6099	1.6379
	PTL	✓	✓	1.0923	1.2656	1.8148
$n_k = 200$	Non-Trans Lasso	✗	✗	1.2866	1.2359	1.1910
	Non-Trans DML	✗	✓	**0.6657**	**0.6983**	**0.6979**
	Partial Trans-Lasso	✓	✗	1.5906	1.6277	1.6770
	PTL	✓	✓	0.7498	0.9909	1.5979
$n_k = 500$	Non-Trans Lasso	✗	✗	1.2522	1.3203	1.3028
	Non-Trans DML	✗	✓	0.7492	**0.6698**	**0.7039**
	Partial Trans-Lasso	✓	✗	1.6039	1.6167	1.7066
	PTL	✓	✓	**0.6591**	0.9366	1.4953

	$s = 5$	Transferred	Debiased	$H = 0.2$	$H = 0.3$	$H = 0.5$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2611	1.2280	1.2303
	Non-Trans DML	✗	✓	**0.6426**	**0.6819**	**0.6986**
	Partial Trans-Lasso	✓	✗	1.5574	1.5515	1.5487
	PTL	✓	✓	2.6584	2.6751	3.0720
$n_k = 200$	Non-Trans Lasso	✗	✗	1.2056	1.3233	1.2614
	Non-Trans DML	✗	✓	**0.6518**	**0.6717**	**0.7103**
	Partial Trans-Lasso	✓	✗	1.5973	1.5763	1.4974
	PTL	✓	✓	1.3549	1.5710	2.2290
$n_k = 500$	Non-Trans Lasso	✗	✗	1.2015	1.2525	1.2790
	Non-Trans DML	✗	✓	**0.6859**	**0.6790**	**0.7205**
	Partial Trans-Lasso	✓	✗	1.5763	1.5427	1.5472
	PTL	✓	✓	0.9960	1.1669	1.8138

	$s = 5$	Transferred	Debiased	$H = 0.01$	$H = 0.05$	$H = 0.1$
$n_k = 100$	Non-Trans Lasso	✗	✗	1.2555	1.2328	1.2831
	Non-Trans DML	✗	✓	**0.6623**	**0.6916**	**0.7087**
	Partial Trans-Lasso	✓	✗	1.5864	1.5658	1.5743
	PTL	✓	✓	2.1543	2.3213	2.2494
$n_k = 200$	Non-Trans Lasso	✗	✗	1.3003	1.2525	1.3276
	Non-Trans DML	✗	✓	**0.6965**	**0.6347**	**0.6868**
	Partial Trans-Lasso	✓	✗	1.6174	1.5200	1.5677
	PTL	✓	✓	0.9183	1.0783	1.4178
$n_k = 500$	Non-Trans Lasso	✗	✗	1.1734	1.1789	1.3549
	Non-Trans DML	✗	✓	0.7521	0.6816	**0.7017**
	Partial Trans-Lasso	✓	✗	1.6276	1.6021	1.5934
	PTL	✓	✓	**0.5534**	**0.5792**	0.7818

Based on the observation on large tolerance of s, I further do 100 simulations based on debiasing approach under condition (5). The results reveal that the performance of transferring sources deteriorates to the point where it is worse than non-transferring when $s \geq 7$, which could not be alleviated by enlarging sample size, indicating a violation of *ASSUMPTION 3.2 (SCORE REGULARITY)* in Chernozhukov *et al.* (2018) occurred.

	$n_0 = 20,\ h = 0.05$	Transferred	Debiased	$n_k = 500$	$n_k = 1000$	$n_k = 2000$
$s = 5$	Non-Trans Lasso	✗	✗	1.2674	1.2641	1.1865
	Non-Trans DML	✗	✓	0.6753	0.6647	0.6894
	Partial Trans-Lasso	✓	✗	1.2106	1.2572	1.2935
	PTL	✓	✓	**0.6036**	**0.3304**	**0.2309**
$s = 6$	Non-Trans Lasso	✗	✗	1.3218	1.2773	1.2588
	Non-Trans DML	✗	✓	0.6948	0.7034	0.7118
	Partial Trans-Lasso	✓	✗	1.3079	1.3565	1.3127
	PTL	✓	✓	**0.6875**	**0.5783**	**0.4639**
$s = 7$	Non-Trans Lasso	✗	✗	1.2788	1.2353	1.2338
	Non-Trans DML	✗	✓	**0.6422**	**0.7635**	**0.6846**
	Partial Trans-Lasso	✓	✗	1.2889	1.3512	1.3487
	PTL	✓	✓	1.8534	2.1546	2.1730
$s = 8$	Non-Trans Lasso	✗	✗	1.1608	1.2221	1.2614
	Non-Trans DML	✗	✓	**0.7153**	**0.7432**	**0.6496**
	Partial Trans-Lasso	✓	✗	1.3902	1.3900	1.2895
	PTL	✓	✓	3.5444	4.0315	4.0399

D.2 Oracle PTL Under Heterogeneous Sources

$s = 0.1$	Transferred	Debiased	$H_1 = 0.01$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7742		
Non-Trans DML	✗	✓	0.4509		
Partial Trans-Lasso	✓	✗	1.1938		
PTL	✓	✓	**0.1731**	0.1941	0.1091
			$H_1 = 0.05$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7869		
Non-Trans DML	✗	✓	0.5108		
Partial Trans-Lasso	✓	✗	1.1022		
PTL	✓	✓	**0.1392**	0.2022	0.0971
			$H_1 = 0.1$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7835		
Non-Trans DML	✗	✓	0.5267		
Partial Trans-Lasso	✓	✗	1.1489		
PTL	✓	✓	**0.3259**	0.2018	0.1263

$s = 0.1$	Transferred	Debiased	$H_1 = 0.2$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7895		
Non-Trans DML	✗	✓	0.5134		
Partial Trans-Lasso	✓	✗	1.1679		
PTL	✓	✓	**0.4023**	0.1929	0.1074
			$H_1 = 0.3$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7811		
Non-Trans DML	✗	✓	**0.4564**		
Partial Trans-Lasso	✓	✗	1.0445		
PTL	✓	✓	0.5904	0.2050	0.1063
			$H_1 = 0.5$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7738		
Non-Trans DML	✗	✓	**0.5014**		
Partial Trans-Lasso	✓	✗	1.9148		
PTL	✓	✓	0.9075	0.1998	0.1165

$s = 0.5$	Transferred	Debiased	$H_1 = 0.01$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7750		
Non-Trans DML	✗	✓	0.4940		
Partial Trans-Lasso	✓	✗	0.4831		
PTL	✓	✓	**0.2432**	0.2421	0.1863
			$H_1 = 0.05$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7727		
Non-Trans DML	✗	✓	0.4873		
Partial Trans-Lasso	✓	✗	0.5143		
PTL	✓	✓	**0.1576**	0.2710	0.1566
			$H_1 = 0.1$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7756		
Non-Trans DML	✗	✓	0.4798		
Partial Trans-Lasso	✓	✗	0.5984		
PTL	✓	✓	**0.4118**	0.2590	0.2181

$s = 0.5$	Transferred	Debiased	$H_1 = 0.2$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7531		
Non-Trans DML	✗	✓	**0.4926**		
Partial Trans-Lasso	✓	✗	0.4955		
PTL	✓	✓	0.6651	0.2766	0.2436
			$H_1 = 0.3$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7460		
Non-Trans DML	✗	✓	0.4660		
Partial Trans-Lasso	✓	✗	**0.4247**		
PTL	✓	✓	0.7777	0.3046	0.2560
			$H_1 = 0.5$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7633		
Non-Trans DML	✗	✓	**0.4265**		
Partial Trans-Lasso	✓	✗	0.5590		
PTL	✓	✓	1.1733	0.2930	0.3291

$s = 1$	Transferred	Debiased	$H_1 = 0.01$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7689		
Non-Trans DML	✗	✓	0.5050		
Partial Trans-Lasso	✓	✗	0.5567		
PTL	✓	✓	**0.3041**	0.3472	0.1595
			$H_1 = 0.05$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7739		
Non-Trans DML	✗	✓	0.4857		
Partial Trans-Lasso	✓	✗	0.5659		
PTL	✓	✓	**0.3477**	0.3652	0.1733
			$H_1 = 0.1$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7863		
Non-Trans DML	✗	✓	**0.4803**		
Partial Trans-Lasso	✓	✗	0.5791		
PTL	✓	✓	0.5938	0.3753	0.2077

$s = 1$	Transferred	Debiased	$H_1 = 0.2$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7806		
Non-Trans DML	✗	✓	**0.5317**		
Partial Trans-Lasso	✓	✗	0.5373		
PTL	✓	✓	0.7803	0.3577	0.2187
			$H_1 = 0.3$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7865		
Non-Trans DML	✗	✓	**0.5052**		
Partial Trans-Lasso	✓	✗	0.5913		
PTL	✓	✓	0.9342	0.3933	0.2436
			$H_1 = 0.5$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7788		
Non-Trans DML	✗	✓	**0.4668**		
Partial Trans-Lasso	✓	✗	0.6063		
PTL	✓	✓	1.2355	0.4173	0.3654

$s = 5$	Transferred	Debiased	$H_1 = 0.01$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7590		
Non-Trans DML	✗	✓	**0.4571**		
Partial Trans-Lasso	✓	✗	0.7298		
PTL	✓	✓	3.6556	2.0786	2.3589
			$H_1 = 0.05$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7664		
Non-Trans DML	✗	✓	**0.4976**		
Partial Trans-Lasso	✓	✗	0.7526		
PTL	✓	✓	3.5571	2.0313	2.5309
			$H_1 = 0.1$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7828		
Non-Trans DML	✗	✓	**0.5055**		
Partial Trans-Lasso	✓	✗	0.7366		
PTL	✓	✓	3.1903	2.0698	2.3837

$s = 5$	Transferred	Debiased	$H_1 = 0.2$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7726		
Non-Trans DML	✗	✓	**0.4284**		
Partial Trans-Lasso	✓	✗	0.7145		
PTL	✓	✓	2.5915	2.0831	2.2644
			$H_1 = 0.3$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7744		
Non-Trans DML	✗	✓	**0.5015**		
Partial Trans-Lasso	✓	✗	0.7248		
PTL	✓	✓	2.5490	2.0461	2.0683
			$H_1 = 0.5$	max.diff.s_1	max.diff.s_2
Non-Trans Lasso	✗	✗	0.7524		
Non-Trans DML	✗	✓	**0.4669**		
Partial Trans-Lasso	✓	✗	0.7371		
PTL	✓	✓	2.2989	1.9236	1.8187