Partial Transfer Learning Under High-Dimensional Confounding

Xinhao Qu[1]

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

Abstract This article considers framework of linear regression with high-dimensional confounding, implementing Partial Transfer Learning (PTL) method for source detection and causal/structural parameters' estimation issue through either homogeneous or heterogeneous source transferring, asymptotics is derived and simulation reveals detection efficiency and the enhanced performance of PTL estimators. Empirical research also demonstrates PTL's well-application particularly in biological data.

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

1 Introduction

Transfer Learning facilitates scientific analyses in many aspects: from the perspective of machine learning, its intrinsic nature of knowledge exchange mimics the decision process of brains, promising its well-applications in machines' learning, neural network especially (Pan and Yang, 2009). From the perspective of data science, it refers to cheap and easily-obtained labels before modeling, enhancing the precision and efficiency of target estimation fundamentally, which also offers potential solutions to challenges of incomplete data, such as clinical data.

From the perspective of social science, however, transferability may be doubtful if not being carefully tackled, that largely serves as one major limitations within transfer learning technique. Usually in traditional machine learning literature, as in Weiss *et al.* (2016), scientists allow transferring in ad hoc tasks and would easily fail in other scenarios, thus these approaches are not applicable and lack of robustness in general. The procedure after confirming positiveness in transferring should also be seriously considered, but thanks to Li *et al.* (2022a), who introduces a statistical benchmark: a smoothing and de-biased framework for transfer learning in high-dimensional linear models, which minimizing the transferring gap through regularization. Scholars also dip into other different model settings, including high-dimensional generalized linear regression (Tian and Feng, 2022; Li *et al.*, 2023b), gaussian graphical model (Li *et al.*, 2022b) and so forth.

Among the aforementioned literature, however, the very first issue of transferability is still controversial, existing algorithms are varying among different settings. Indeed, such consideration reflects the trade-off between the 'exploitation' on target data and the 'exploration' on the source, similar to reinforcement learning. Preferring the side of 'exploitation', transferability would be so restrictive to promise, thus alleviating the shrinking sample size problem so mildly, but very likely avoiding bad/nagetive transfer, ensuring small bias. Preferring the side of 'exploration', the threshold of transferring is so low that may possibly lead to larger bias or bad/nagetive transfer, but promises estimation efficiency. Thus in order to find the equilibrium, I consider it partially, by allowing one subpart of the linear model to be completely dissimilar while remaining asymptotic closeness with respect to the other, while simultaneously conducting transfer within the similar subpart, which certainly offers flexible weighing between 'exploitation' and 'exploration' especially when it comes to empirical issues, giving large toleration for transferability.

Following this logic, I construct the model within a high-dimensional linear confounding framework, which is similar with semi-parametric modeling as in Robinson (1988). And the partially separated estimation for the dissimilar is fully based on Double/Debiased Machine Learning method, presented by Chernozhukov *et*

1. Supervisor: Wei Zhong, Wang Yanan Institute for Studies in Economics and Department of Statistics and Data Science, SOE, Xiamen University, Xiamen 361005, China.

al. (2018), which promises proper asymptotics in this article. Also due to possible covariate shift in nuisance, an independent estimation step is introduced, which combines Smoothly Clipped Absolute Deviation (SCAD) regularized estimation (Fan and Li, 2001) and cross-fitting procedure, similar jargon of which has shown in Fan *et al.* (2011), the oracle property of SCAD also guarantees asymptotic normality for Partial Transfer Learning (PTL) estimator in various scenarios.

2 Contribution

First and foremost, compared with the debiasing-based transfer learning strategy introduced by Li *et al.* (2021) initially, this paper considers a more general causal framework by combining potential confounding module, which is almost inevitable in reality, thus forming more stabilized modeling towards empirical applications.

Additionally, compared with Li *et al.* (2021), the restriction for similarity is also relaxed here by allowing total dissimilarity among partial parameters between the target and source data, offering larger insurance for transferability for real data sources, which also generalizes the applicable range in empirical analyses, since the data of which could hardly bear the entire similarity for every single parameter.

Moreover, PTL estimation is willing to be extended into federated case, which is considered in Li *et al.* (2023a) as well, where they packaged the sharable information into Hessian matrices in case of privacy leakage, but also adds extra cost when communicating. However, PTL approach purely requires the information of moments with respect to the nuisance, ensuring transition efficiency and privacy at the same time.

Finally, the introduced asymptotics largely mimics traditional theory in econometrics, which is free from concentration inequalities, offering another perspective of data fusion processes.

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 causal, 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. Notice that the finite causal part is without restriction of similarities, whose role thus is not confined to being 'causal', namely any distinctive part among target and sources could be allocated as representative 'causal' subpart.

4 Source Detection

This section intends for automatic detection of transferrable sources, note that the following algorithms are automatic in the sense that no subjective ranking is needed and the overall design is based on confidence interval cutting, similar techniques could be found in Tian and Feng (2022), which also combines a Cross Validation (CV) procedure. However, apart from CV, I also introduce a Bootstrapping alternative, which tends to show higher consistency although time-consuming in simulations.

The first algorithm is with respect to CV implementation, based on the benchmark of Double/Debiased Machine Learning (DML) estimator (Chernozhukov *et al.*, 2018). I calibrate the cutting point for positiveness of sources with estimation fluctuation by adding up empirical variances from target and the source through CV, as well as multiplying 0.025 upper quantile of standard normal distribution as a form similar with 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, the empirical variance could also be obtained through Bootstrapping, thus the next algorithm applies this approach estimating the bound 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

I formulate a two-step estimation algorithm for linear model with high-dimensional confounding based on homogeneous sources in this section, which takes (1) similarity between mean functions of the nuisance, (2) similarity between confounding structures into consideration. Here homogeneity is with respect to the same causal nuisance separation between target and each source model. Without loss of generality, I only consider single source parametric models in subsection 5.1 and 5.2 for better illustration, algorithms for multiple sources are given in summary algorithms. Thus the informative auxiliary sample case is simplified as

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

I demonstrate consistency of algorithms when $p = 1$ in particular, which also take possible covariate shift into consideration, while the general case could be found in corollary and in appendix especially. Simultaneously, the following design could be easily extended for federated transfer learning scenario as in Li *et al.* (2023a), where the underlying Data Generating Process (DGP) of (1)-(4) is identical for each

site $m \in \{1, ..., M\}$, and information is computationally costless for sharing while privacy is also promised between sites.

5.1. Without Covariate Shift

For better illustration, the first subsection assumes no shift in covariates, although such consideration is hardly realistic. Based on Chernozhukov *et al.* (2018), I first implement Debised Machine Learning (DML) method to the source data, achieving debiased estimation of $\rho_0^{(k)}$, which is reserved for transferring. There are several reminders that may deserve consideration: First, I implicitly assume sparsity for both nuisance and confounding functions, thus I merely use the Lasso learner in DML. And if the generating formation is of other types, learner of Random Forest (RF), Gradient Boosting Decision Tree (GBDT) or Neural Network (NN) could be easily applied. Second, the causal part of source model is algorithmically of no interest, but it is a crucial bridge for unbiased nuisance estimation.

Consider (2) with additional intercept term derived from the source, I simply transfer the nuisance mean function to reconstruct the target response, then

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

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).

Additionally, although being overlooked in source detection algorithm **Meta-Algorithm-1.1** and **Meta-Algorithm-1.2**, the ranking of transferrable sources could serve as references for adaptive estimation, thus alleviating negative transfer by allocating various weights to sources. Such consideration is summarized into 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 possible shift in nuisance distribution, simply apply EER is far from consistent, thus parameter-wise transfer is formulated through smoothly clipped absolute deviation (SCAD) penalty from Fan and Li (2001), whose Oracle property ensures consistency here. The first step in **Meta-Algorithm-2.1** is inherited, however, the shifting covariates require more than de-confounding precaution, while the implementation of cross-fitting estimators $\hat{\rho}^{(k)}$ and $\hat{\beta}^{(k)}$ is added. Subsequently, I reconstruct response in source data as $Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)}$, and obtain oracle $\hat{\beta}^{(k)}$ through SCAD. Eventually I compute

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

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* $\hat{\beta}^{(k)} = \frac{1}{F}\sum_{f=1}^{F}\hat{\beta}^{(k,f)}$*;*

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** could 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 is revealed by the subsequent corollary and the proof in appendix.

(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 PTL estimators' consistency and asymptotic normality, where the detailed proofs are left in the appendix.

Theorem 1.1 (PTL on homogeneous sources without shift)

Under condition *(C1)-(C4)* in the appendix, we have:

$$\hat{\rho}_{\mathrm{PTL}} \xrightarrow{p} \rho_0 \tag{10}$$

as $n_0 \to \infty,\ n_k \to \infty$.

Theorem 1.2 (PTL on homogeneous sources with shift)

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

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

$$\sqrt{n_0}\left(\hat{\rho}*_{\mathrm{PTL}} - \rho_0\right) \xrightarrow{d} N(0, c_1\sigma_0^2) \tag{12}$$

as $n_0 \to \infty,\ n_k \to \infty$, 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, we have

$$\begin{aligned} \hat{\rho}*_{\mathrm{PTL},G} - \rho_0 &\xrightarrow{p} 0 \\ \sqrt{n_0}\left(\hat{\rho}*_{\mathrm{PTL},G} - \rho_0\right) &\xrightarrow{d} N\left(0, Q^{-1}\Sigma_0 Q^{-1}\right) \end{aligned} \tag{13}$$

as $n_0 \to \infty,\ n_k \to \infty$, where where $\Sigma_0 \equiv \mathbb{V}(D_i^{(0)} V_i^{(0)}) = \mathbb{E}(D_i^{(0)} D_i^{(0)^\top} V_i^{(0)^2})$, $Q \equiv \mathbb{E}(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)})$ invertible, subscript G indicates the general case. Also

$$\begin{aligned} Y_i^{(0)} - \widehat{Y}_i^{(0)} &= D_i^{(0)^\top}\left(\rho_0 - \hat{\rho}*_{\mathrm{PTL},G}\right) + X_i^{(0)^\top}\left(\beta_0 - \hat{\beta}^{(k)}\right) + V_i^{(0)} \\ &= O_p\left(n_0^{\frac{1}{2}}\right) + o_p\left(n_0^{\frac{1}{2}}\right) + O_p\left(n_k^{-\frac{1}{2}} + a_n\right) + V_i^{(0)} \\ &\overset{d}{\sim} V_i^{(0)} \end{aligned} \tag{14}$$

6 PTL on Heterogenous Sources

Based on condition (5) of informative auxiliary samples, I introduce the algorithm of hetero-source partial transfer learning for structural and treatment parameter ρ_0, here heterogeneity comes from the distinct causal and nuisance parts. In particular, for each S_k, treat each $\hat{\beta}_{0,j\notin\mathcal{J}_k}^{(k)}$ and $\rho_0^{(k)}$ as the 'causal' part in DML settings, thus implementing DML to get unbiased estimation of $\mathbb{E}\left[\hat{f}_{\mathcal{J}_k}^{(k)}(X^{(k)})\right]$'s would be similar as in subsection 5.1. Note that I implicitly assume exhaustiveness in heterogeneous sources, thus the ensembled nuisance will simply be $\sum_{k=1}^K \widehat{\mathbb{E}}\left[\hat{f}_{\mathcal{J}_k}^{(k)}(X^{(k)})\right]$. Similarly, the following algorithms are designed for $p = 1$ case, where the extension for federated PTL is trivial. The general case is summarized in corollary and in appendix particularly as complements.

6.1. Without Covariate Shift

If nuisance distribution is without shift, still, cross-fitting strategy is applied here for large sample properties' insurance. The following steps are trivial and similar with subsection 5.2: transfer the nuisance mean function to the transformed target model, and get

$$\hat{\rho}_{\mathrm{HPTL}} = \frac{\widehat{\mathbb{E}}\left(\widehat{Y}_{i,H}^{(0)}\right)}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} \tag{15}$$

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* $\hat{\beta}^{(k)} = \frac{1}{F}\sum_{f=1}^F \hat{\beta}^{(k,f)}$*;*
5) *Implement unbiased* $Alg(\cdot)$ *for* $\widehat{Y}_H^{(0)}$ *on* $D^{(0)}$ *and get estimation of* ρ_0*.*

where $Alg(\cdot)$ indicates equation (15).

6.2. Covariate Shift

Consider latent shift in nuisance covariates as in subsection 5.2, this part also largely mimics **Meta-Algorithm-3.1**, however, a multiple cross-fitting procedure is introduced due to 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 (15).

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** as long as the heterogeneity between target and each source is given, however, exhaustive searching may be needed if lacking such prior knowledge. Note that in empirical analysis, a data-driven pre-searching procedure is introduced for rough separation of causal and nuisance.

6.4. Theorems

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

Theorem 2.1 (PTL on heterogeneous sources without shift)

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

$$\hat{\rho}_{\text{HPTL}} - \rho_0 \xrightarrow{p} 0 \tag{16}$$

$$\sqrt{n_0}\left(\hat{\rho}_{\text{HPTL}} - \rho_0\right) \xrightarrow{d} N\left(0, c_1\sigma_0^2\right) \tag{17}$$

as $n_0 \to \infty$.

Theorem 2.2 (PTL on heterogeneous sources with shift)

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

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

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

as $n_0 \to \infty$.

Corollary 2 (Heterogeneous PTL for general cases)

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

$$\begin{aligned} \hat{\rho}*_{\mathrm{HPTL},G} - \rho_0 &\xrightarrow{p} 0 \\ \sqrt{n_0}\left(\hat{\rho}*_{\mathrm{HPTL},G} - \rho_0\right) &\xrightarrow{d} N\left(0, Q^{-1}\Sigma_0 Q^{-1}\right) \end{aligned} \tag{18}$$

as $n_0 \to \infty$. Also

$$\begin{aligned} Y_i^{(0)} - \widehat{Y}_i^{(0)} &= D_i^{(0)^\top}\left(\rho_0 - \hat{\rho}*_{\mathrm{HPTL},G}\right) + \sum_{k=1}^{K} X_{i,\mathcal{J}_k}^{(0)^\top}\left(\beta_{0,\mathcal{J}_k} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) + V_i^{(0)} \\ &= O_p\left(n_0^{\frac{1}{2}}\right) + K * o_p\left(n_0^{\frac{1}{2}}\right) + \sum_{k=1}^{K} O_p\left(n_k^{-\frac{1}{2}} + a_n\right) + V_i^{(0)} \\ &\overset{d}{\sim} V_i^{(0)} \end{aligned} \tag{19}$$

as $n_0 \to \infty$.

7 Simulation

In order to testify the validity of the above algorithms introduced, I conduct the following simulations.

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:

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

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

$\rho_0 = -0.8$, $\beta_0 = (\beta_{(1)}^\top, \beta_{(2)}^\top)^\top$ 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 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)*: $n_2 = 500$, $X_i^{(2)} \overset{i.i.d.}{\sim} N_q(-\mu_q, \mathbf{\Sigma}_{q\times q}^{(2)})$, where $\mathbf{\Sigma}_{q\times q}^{(2)}(i,j) = 0.2^{|i-j|}$, $V_i^{(2)} \overset{i.i.d.}{\sim} N(0,1)$, $U_i^{(2)} \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in \{1, 2, ..., n_2\}$, $\rho_0^{(2)} = 0.4$, $h_2 = 0.3, s_2 = 0.5$ under same sparsity structure above.

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

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

The oracle observation is that the positiveness 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.

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.* (2022) but with mild modification for

	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

equation (6) in their paper by

$$\hat{\delta}^{\mathcal{A}} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{2n_0} \left\| y^{(0)} - X^{(0)} \left(\hat{w}^{\mathcal{A}} + \delta \right) \right\|_2^2 + \lambda_\delta \| w_*^\top \delta \|_1 \right\} \tag{20}$$

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**

Result reveals: First, transferrable interval serves a automatic cut when selecting positive sources no matter for **Meta-Algorithm-1.1** (Cross Validation) or **Meta-Algorithm-1.2** (Bootstrap) methods. Second, Bootstrap tends to make the right choice more but also more time-consuming. Third, through comparison of different transferring strategies, PTL achieves the minimal RMSE, corresponding to 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 Root Mean Square Error (RMSE). For simplicity, two representative results are listed as tables, more simulations could be found in the appendix.

	$s = 0.05$	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**

	$s = 0.05$	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

From tables above, it is not hard to conclude that: First, debiasing benefits under the formation of irresistible confounding when DML is functioning well, which is just as expected. Second, transferring would also enhance the precision of causal parameter's estimation in case of positive sources. Third, RMSE has the tendency to decrease if either enlarging the positive source's sample size n_k, or promising better functioning of DML through small s. Fourth, in finite samples, the impact of narrowing 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.

Based on the observation on large tolerance of s, I further do 100 simulations based on debiasing approach under condition (5). Measured by RMSE, the result shows: Firstly, transferring performance would be worse than non-transferring when $s \geq 7$. Secondly, the above deteriorative situation could not be alleviated by enlarging sample size, namely, sample size of the source would be a curse. And the reason is that DML's actually not debiasing in this setting, a violation of *ASSUMPTION 3.2 (SCORE REGULARITY)* in Chernozhukov *et al.* (2018) occurred.

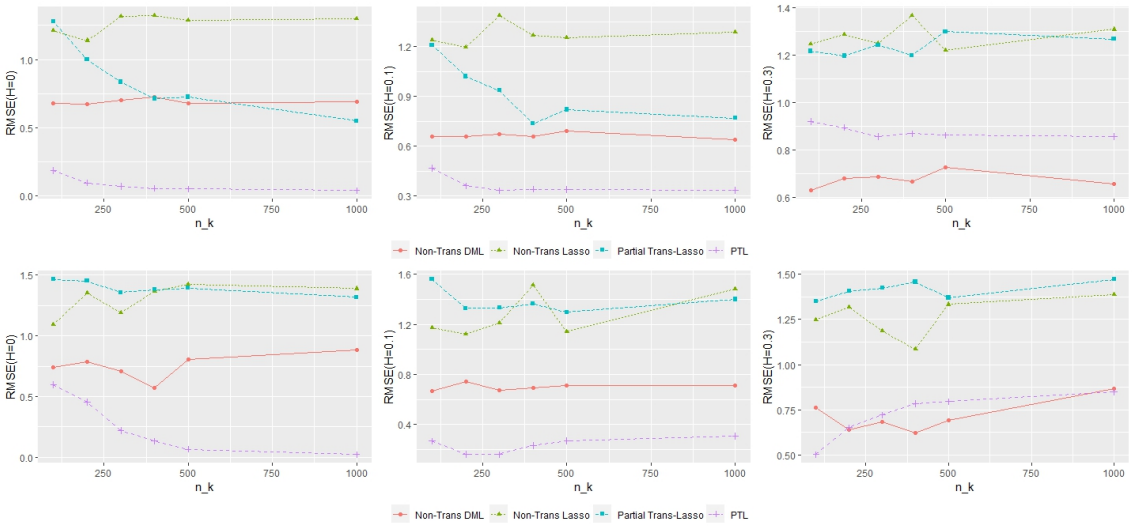
		Transferred	Debiased	$n_k = 500$	$n_k = 1000$	$n_k = 2000$
$s = 5$	Non-Trans Lasso	✗	✗	1.2674	1.2641	1.1865
	Non-Trans DML	✗	✓	0.6753	0.6647	0.6894
	Partial Trans-Lasso	✓	✗	1.2106	1.2572	1.2935
	PTL	✓	✓	**0.6036**	**0.3304**	**0.2309**
$s = 6$	Non-Trans Lasso	✗	✗	1.3218	1.2773	1.2588
	Non-Trans DML	✗	✓	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

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 :



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 $\beta^{(2)}_{(1)_b}$ satisfying $\|\beta_{(1)_b} - \beta^{(2)}_{(1)_b}\|_\infty = h$ and no requirement for $\beta^{(2)}_{(1)_a}$, 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 'max.diff's are given, representing the 'causal' part's learning performance, i.e.

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

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

Applying **Meta-Algorithm-3.1**, the results are as follow, which also includes the comparison with Partial Trans-Lasso method, however, here w_*, as in (23), adapts to $(0, 0, 0, 0, 1, 1, ..., 1)^\top$. More simulations could be found in the appendix.

$s = 0.05$	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

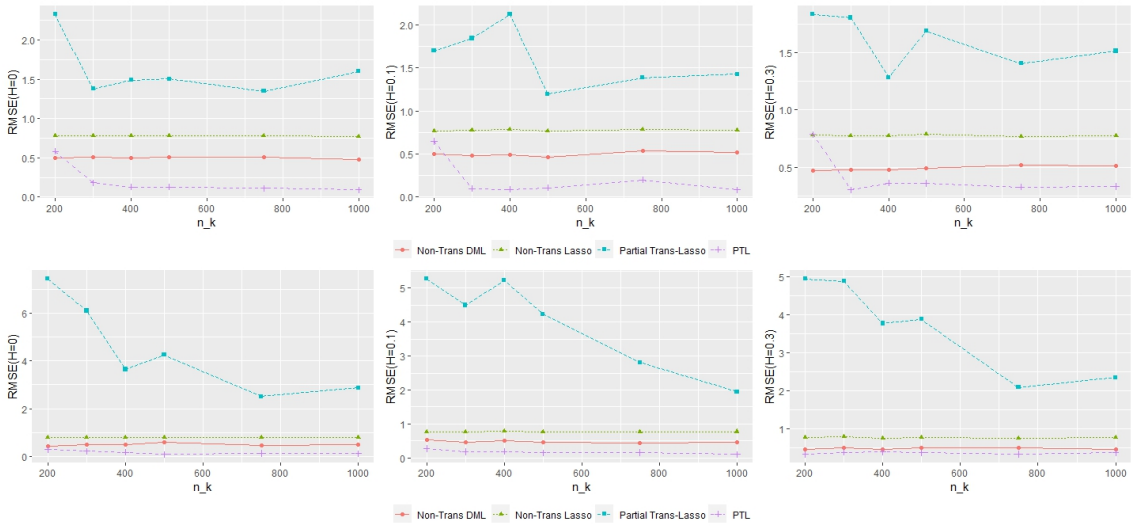
The above tables show: First, same pattern of n_k, h, s has shown in the decreasing tendency of RMSE

compared with homogeneous case. Second, malfunction of DML happens earlier (with smaller s) than homogeneous counterpart since the backdoor path dominates quicker if causal signals are further blurred by segregated sources.

Also simulating case of covariate shift, 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 reveals the following results, which is similar with non-shifting case as well.

$s = 0.05$	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, which originates from 838 donors of 49 human tissues, revealing expression levels among various genes over different types of tissues. Similarly, I consider one subset 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 adds up to 545 covariate genes in total. Also 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 refers to Tissue Sampling Sites in GTEx (`https://gtexportal.org/home/samplingSitePage`), the overall sample size of the aforementioned subsample is 2,642.

8.1. Analytical Strategy

In order to discover the underlying functional status of CNS, bioscientist particularly pays attention to possible relationships in genes' expression levels, especially among those aforementioned brain tissues. Statistically speaking, it is one of the priorities to predict one response gene's expression level using other CNS genes as covariates.

Inheriting the framework as in Li *et al.* (2022), here I allocate `JAM2`, one type of protein coding gene located on chromosome 21, as the response, the lack of which may cause malfunction in lymphocyte homing to secondary lymphoid organs (Johnson-Léger *et al.*, 2002), whose mutation also lays a potential threatening factor for primary familial brain calcification (Cen *et al.*, 2020; Schottlaender *et al.*, 2020). Thus I consider the hidden relationship between `JAM2` and other MODULE_137 genes, hoping for a better prediction performance in-between.

Due to multiple brain tissues mentioned as above, I treat each of them as the target respectively, while the rest is assigned as the sources to be detected. Initially, I presumed that all sources are informative since the overall data set is within brains, which intuitively fits the in-distribution transfer proposition, the prediction result also demonstrates this claim. Thus I apply the following strategy for 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 logic lies in the consideration of exterior attack, where the target sample could be thought as such impact originates outside of the underlying DGP of nuisance, only those with sustainable estimation could ultimately be considered in-distribution nuisance.

Based on 13 target samples with average size of 203, through cross-validation technique similar in Li *et al.* (2022), one fold of each target is reserved as validation, **Meta-Algorithm-1.1** and **Meta-Algorithm-2.3** in general cases are automatically applied here for the prediction purpose.

8.2. Prediction Performance of PTL for `JAM2` Expression

This subsection compares the prediction performance of All-Trans PTL and Detected PTL with Non-Trans DML by computing their relative prediction error, on the other hand, the comparison ratio between Partial Trans-Lasso and Partial Lasso is also shown for better illustration of degree of enhancement based on the previous literature.

The first modeling path is with respect to the baseline model of Non-Trans DML, which simply apply DML algorithm merely on the target sample, lacking the informative knowledge partially from the detected source. Two modeling strategies are compared with this baseline: All-Trans PTL and Detected PTL. The first of which arbitrarily considers all sources left transferrable and then apply **Meta-Algorithm-2.3** in general cases for prediction, which lacks the procedure of detection, however, such consideration makes some sense since 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 step of source detection by applying **Meta-Algorithm-1.1**, the following table shows their relative rate of prediction performance as

well as the detection result.

The second modeling path aims at better illustration through the comparison with previous methods' comparative performance. 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, denote such modified version as Partial Trans-Lasso and compare it with Adaptive Lasso with penalty factor to be 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, namely Amygdala, Anterior Cingulate Cortex, Caudate, Cerebellar Hemisphere, Cerebellum, Cortex, Frontal Cor- tex, Hippocampus, Hippocampus, Nucleus Accumbens, Putamen, Spinal cord and Substantia nigra in order. For 'Null' element in the second column, it refers to $\forall j \in \{1, ..., p+q\}$, $|\hat{\theta}_j^{(k)} - \hat{\theta}_j^{(0k)}| \leq h_{0k}$, thus fitting the scenario of Li *et al.* (2022) and there is no need to apply PTL in such case. 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 is clearly shown in the table, due to dearth of information under pure target data (whose largest sample size is 255), Non-Trans methods lacks their prediction power in most cases, also on the other hand, casual transferring of All-Trans PTL, remains several positive transfer though, largely incurs negativeness through transfer. And Partial Trans-Lasso loses prediction accuracy possibly due to underlying confounding structure among genes.

However, Detected PTL outperforms in all detectable cases, with an average of 22.65% improve of prediction accuracy compared with Non-Trans DML, an average of 16.30% increase of relative error rate compared with All-Trans PTL and an average of 34.76% enhancement of error rate performance compared with Partial Trans-Lasso, indicating the overall enhanced performance of de-confounded partial knowledge transfer.

9 Discussion

This article analyzes the statistical modeling strategy under the scenario where there exist additional data sets as references for the label-sparse target set, with the consideration of partially mismatching parameters and ubiquitous confounding factors under high-dimension. Refer the transferrable and the counterpart parameter set as the 'nuisance' and 'causal', I propose the transferring criterion merely on the nuisance

similarity, developing sequential detection and estimation algorithms with respect to the causal parameters. Simulations and application based on GTEx data all demonstrate the efficiency and enhanced performance of parameter estimation and response prediction for detected PTL.

Among ongoing transfer learning literatures, restriction for parameter similarities is sometimes too strong to be applicable, however, PTL offers a solution under prior information of 'nuisance' 'causal' separation for both sources and the target data. However, such field knowledge is sometimes lacking, and the computation cost for source detection under such circumstance could be challenging. Hence, the matching problem without the help of prior knowledge remains open for future research.

On the other hand, a totally data-driven approach based on Trans-Lasso framework may be helpful in handling the aforementioned matching issue. Through plugging an extra weight as an adaptive regularization within the naive Lasso penalty, the transferring would be automatically 'partial', similar with modification as (20) but with w_* non-constant. Such topic should be interesting for further research, although it may lack proper interpretations in empirical cases.

REFERENCE

[1] Cai, T. T., and H. M. Pu (2022). Transfer learning for nonparamet ricregression: non-asymptotic minimax analysis and adaptive procedure. *Submitted to the Annals of Statistics.*

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

[3] Cai, T. T., and H. J. Wei. (2021) Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), 100-128.

[4] 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.

[5] Chernozhukov, V., W. K. Newey, and R. Singh (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90, 967-1027.

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

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

[8] Li, S., T. 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.

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

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

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

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

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

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

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

[16] Strickler J. H., T. Yoshino, R. P. 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.

[17] 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.

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

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

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

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

APPENDIX

5.A Large Sample Properties for PTL on Homogeneous Sources

5.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):

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

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 $n_0 \to \infty$ with order $n_0 = o(n_k)$, $q = O(n_k)$.

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

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

(C5): Based on condition (A)-(C) and Theorem 1 in Fan and Li (2001), for k_{th} source after de-confounding, we have

$$\left\|\hat{\beta}_{\text{SCAD}} - \beta_0\right\| = O_p(n_k^{-\frac{1}{2}} + a_n) \tag{21}$$

where $a_n = \max\left\{p'_{\lambda_n}\left(|\beta_{j0}|\right) : \beta_{j0} \neq 0\right\} \to 0$ 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.

5.A.2. Theorem

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

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

as $n_0 \to \infty,\ n_k \to \infty$.

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

$$\begin{aligned}
\hat{\rho}_{\text{PTL}} - \rho_0 &= \frac{\widehat{\mathbb{E}}\left[Y_i^{(0)} - \widehat{\mathbb{E}}\left(Y_i^{(1)} - D_i^{(1)}\hat{\rho}^{(1)}\right)\right]}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} - \rho_0 \\
&= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\rho_0 D_i^{(0)} + \beta_0^\top X_i^{(0)} + V_i^{(0)}\right)}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} - \rho_0 - \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\left[\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) D_i^{(1)} + \beta_0^{(1)^\top} X_i^{(1)} + V_i^{(1)}\right]}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \\
&= \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\beta_0^\top X_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1}\beta_0^{(1)^\top} X_i^{(1)}\right] - \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) D_i^{(1)}\right] \\
&+ \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\left[\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1} V_i^{(1)}\right] \qquad (22) \\
&= \frac{1}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}}\beta_0^\top\left[\mu^{(0)} - \mu^{(1)} + o_p(1)\right] - \frac{\mathbb{E}\left(D_i^{(1)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{1}{n_1}\sum_{i=1}^{n_1} V_i^{(1)}\right] \\
&= o_p(1) + o_p(1) + o_p(1) + o_p(1) \\
&= o_p(1)
\end{aligned}$$

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

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

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

as $n_0 \to \infty,\ n_k \to \infty$, 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{23}$$

under condition *(C1)-(C5)*. Also

$$\begin{aligned}
\sqrt{n_0}\left(\hat{\rho}*_{\mathrm{PTL}} - \rho_0\right) &= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}\right)^\top\frac{\sqrt{n_0}}{\sqrt{n_1}}\sqrt{n_1}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) - \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right)^\top\frac{\sqrt{n_0}}{\sqrt{n_1}}\sqrt{n_1}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right)\right] \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left[\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)} - \frac{\sqrt{n_0}}{\sqrt{n_1}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} V_i^{(1)}\right] - \frac{\mathbb{E}\left(D_i^{(1)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\frac{\sqrt{n_0}}{\sqrt{n_1}}\sqrt{n_1}\left(\rho_0^{(1)} - \hat{\rho}^{(1)}\right) \\
&= o_p(1) + o_p(1) + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)}\right) + o_p(1) + o_p(1) \\
&\xrightarrow{d} N(0, c_1\sigma_0^2)
\end{aligned} \tag{24}$$

as $n_0 \to \infty$.

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

$$\hat{\rho}*_{\mathrm{PTL},G} - \rho_0 \xrightarrow{p} 0$$
$$\sqrt{n_0}\left(\hat{\rho}*_{\mathrm{PTL},G} - \rho_0\right) \xrightarrow{d} N\left(0, Q^{-1}\Sigma_0 Q^{-1}\right)$$

as $n_0 \to \infty,\ n_k \to \infty$, where subscript G indicates the general case. Also

$$\begin{aligned} Y_i^{(0)} - \widehat{Y}_i^{(0)} &= D_i^{(0)^\top}\left(\rho_0 - \hat{\rho}*_{\mathrm{PTL},G}\right) + X_i^{(0)^\top}\left(\beta_0 - \hat{\beta}^{(k)}\right) + V_i^{(0)} \\ &= O_p\left(n_0^{\frac{1}{2}}\right) + o_p\left(n_0^{\frac{1}{2}}\right) + O_p\left(n_k^{-\frac{1}{2}} + a_n\right) + V_i^{(0)} \\ &\overset{d}{\sim} V_i^{(0)} \end{aligned} \tag{25}$$

Proof.

$$\begin{aligned} \hat{\rho}*_{\mathrm{PTL},G} - \rho_0 &= \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\widehat{Y}^{*^{(0)}} - \rho_0 \\ &= \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\left(Y^{(0)} - \mathbf{X}^{(0)}\hat{\beta}^{(1)}\right) - \rho_0 \\ &= \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\left(\mathbf{X}^{(0)}\beta_0 - \mathbf{X}^{(0)}\hat{\beta}^{(1)} + V^{(0)}\right) \\ &= \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\mathbf{X}^{(0)}\left(\beta^{(1)} - \hat{\beta}^{(1)}\right) + \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}V^{(0)} \\ &\xrightarrow{p} 0 \end{aligned} \tag{26}$$

as $n_0 \to \infty,\ n_k \to \infty$. Also

$$\begin{aligned} \sqrt{n_0}(\hat{\rho}*_{\mathrm{PTL},G} - \rho_0) &= \widehat{Q}^{-1}\mathbf{D}^{(0)^\top}\mathbf{X}^{(0)}\frac{\sqrt{n_0}}{\sqrt{n_1}}\sqrt{n_1}\left(\beta^{(1)} - \hat{\beta}^{(1)}\right) + \widehat{Q}^{-1}\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} D_i^{(0)}V_i^{(0)} \\ &\xrightarrow{d} N\left(0, Q^{-1}\Sigma_0 Q^{-1}\right) \end{aligned} \tag{27}$$

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

6.A Large Sample Properties for PTL on Heterogeneous Sources

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

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

as $n_0 \to \infty$.

Proof.

$$\begin{aligned}
\hat{\rho}_{\text{HPTL}} - \rho_0 &= \frac{\widehat{\mathbb{E}}\left[Y_i^{(0)} - \sum_{k=1}^K \widehat{\mathbb{E}}\left(Y_i^{(k)} - D_i^{(k)}\hat{\rho}^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right)\right]}{\widehat{\mathbb{E}}\left(D_i^{(0)}\right)} - \rho_0 \\
&= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(D_i^{(0)}\rho_0 + X_i^{(0)\top}\beta_0 + V_i^{(0)}\right)}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} - \rho_0 \\
&\quad - \frac{\sum_{k=1}^K \frac{1}{n_k}\sum_{i=1}^{n_k}\left[D_i^{(k)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) + \left(X_i^{(k)\top}\beta_0^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) + V_i^{(k)}\right]}{\frac{1}{n_0}\sum_{i=1}^{n_0} D_i^{(0)}} \\
&= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \times \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i^{(0)}\right)^\top \beta_0 - \sum_{k=1}^K \frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) \\
&\quad - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^K \frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^{(k)\top}\beta_0^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) \\
&\quad + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^K\left(\frac{1}{n_k}\sum_{i=1}^{n_k} V_i^{(k)}\right)
\end{aligned} \tag{28}$$

since

$$\begin{aligned}
\sum_{k=1}^K \frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^{(k)\top}\beta_0^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\hat{\beta}_{-\mathcal{J}_k}^{(k)}\right) &= \sum_{k=1}^K \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\sum_{j=1}^q x_{ij}^{(k)}\beta_{0j}^{(k)} - \sum_{j\notin\mathcal{J}_k} x_{ij}^{(k)}\hat{\beta}_j^{(k)}\right) \\
&= \sum_{k=1}^K \frac{1}{n_k}\sum_{i=1}^{n_k}\left[\sum_{j=1}^q x_{ij}^{(k)}\beta_{0j}^{(k)} - \sum_{j\notin\mathcal{J}_k} x_{ij}^{(k)}\left(\beta_{0j}^{(k)} + \hat{\beta}_j^{(k)} - \beta_{0j}^{(k)}\right)\right] \\
&= \sum_{k=1}^K \frac{1}{n_k}\sum_{i=1}^{n_k}\left[\sum_{j\in\mathcal{J}_k} x_{ij}^{(k)}\beta_{0j}^{(k)} - \sum_{j\notin\mathcal{J}_k} x_{ij}^{(k)}\left(\hat{\beta}_j^{(k)} - \beta_{0j}^{(k)}\right)\right] \\
&= \sum_{k=1}^K \frac{1}{n_k}\sum_{i=1}^{n_k}\left[X_{i,\mathcal{J}_k}^{(k)\top}\beta_{0,\mathcal{J}_k}^{(k)} - X_{i,-\mathcal{J}_k}^{(k)\top}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right)\right]
\end{aligned} \tag{29}$$

thus

$$\begin{aligned}
\hat{\rho}_{\text{HPTL}} - \rho_0 &= \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^K\left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_{i,\mathcal{J}_k}^{(0)} - \frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,\mathcal{J}_k}^{(k)}\right)^\top\left(\beta_{0,\mathcal{J}_k} - \beta_{0,\mathcal{J}_k}^{(k)}\right) \\
&\quad - \sum_{k=1}^K \frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) \\
&\quad + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^K\left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,-\mathcal{J}_k}^{(k)}\right)^\top\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right) \\
&\quad + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)}\sum_{k=1}^K\left(\frac{1}{n_k}\sum_{i=1}^{n_k} V_i^{(k)}\right) \\
&= o_p(1) + o_p(1) + o_p(1) + o_p(1) + o_p(1) = o_p(1)
\end{aligned} \tag{30}$$

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

$$
\begin{aligned}
\sqrt{n_0}(\hat{\rho}_{\text{HPTL}} - \rho_0) &= -\sum_{k=1}^{K} \frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k}\left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_{i,-\mathcal{J}_k}^{(k)}\right)^{\top} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k}\left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{\sqrt{n_0}}{\sqrt{n_k}} \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} V_i^{(k)}\right) \\
&= o_p(1) + o_p(1) + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} V_i^{(0)}\right) + o_p(1) \\
&\xrightarrow{d} N\left(0, c_1\sigma_0^2\right)
\end{aligned}
\tag{31}
$$

as $n_0 \to \infty$ under *(C1)-(C5)*.

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

$$\hat{\rho}^*_{\mathrm{HPTL}} - \rho_0 \xrightarrow{p} 0$$

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

as $n_0 \to \infty$.

Proof.

$$\begin{aligned}
\hat{\rho}^*_{\mathrm{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) \\
&= \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)
\end{aligned} \tag{32}$$

Then

$$\begin{aligned}
\hat{\rho}^*_{\text{HPTL}} - \rho_0 &= -\sum_{k=1}^{K} \frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} X_{i,-\mathcal{J}_k}^{(k)}\right)^\top \left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} X_{i,\mathcal{J}_k}^{(k)} - \frac{1}{n_0} \sum_{i=1}^{n_0} X_{i,\mathcal{J}_k}^{(0)}\right)^\top \left(\hat{\beta}_{\mathcal{J}_k}^{(k)} - \beta_{0,\mathcal{J}_k}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} V_i^{(k)}\right) \\
&= o_p(1) + o_p(1) + o_p(1) + o_p(1) + o_p(1) \\
&= o_p(1)
\end{aligned} \tag{33}$$

also

$$\begin{aligned}
\sqrt{n_0}(\hat{\rho}^*_{\text{HPTL}} - \rho_0) &= -\sum_{k=1}^{K} \frac{\mathbb{E}\left(D_i^{(k)}\right) + o_p(1)}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k} \left(\rho_0^{(k)} - \hat{\rho}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} X_{i,-\mathcal{J}_k}^{(k)}\right)^\top \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k} \left(\hat{\beta}_{-\mathcal{J}_k}^{(k)} - \beta_{0,-\mathcal{J}_k}^{(k)}\right) \\
&- \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} X_{i,\mathcal{J}_k}^{(k)} - \frac{1}{n_0} \sum_{i=1}^{n_0} X_{i,\mathcal{J}_k}^{(0)}\right)^\top \frac{\sqrt{n_0}}{\sqrt{n_k}} \sqrt{n_k} \left(\hat{\beta}_{\mathcal{J}_k}^{(k)} - \beta_{0,\mathcal{J}_k}^{(k)}\right) \\
&+ \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} V_i^{(0)}\right) - \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \sum_{k=1}^{K} \left(\frac{\sqrt{n_0}}{\sqrt{n_k}} \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} V_i^{(k)}\right) \\
&= o_p(1) + o_p(1) + o_p(1) + \frac{1}{\mathbb{E}\left(D_i^{(0)}\right) + o_p(1)} \left(\frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} V_i^{(0)}\right) + o_p(1) \\
&\xrightarrow{d} N\left(0, c_1 \sigma_0^2\right)
\end{aligned} \tag{34}$$

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)*, we have

$$\begin{aligned} \hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0 &\xrightarrow{p} 0 \\ \sqrt{n_0}\left(\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0\right) &\xrightarrow{d} N\left(0, Q^{-1}\Sigma_0 Q^{-1}\right) \end{aligned} \tag{35}$$

as $n_0 \to \infty$. Thus

$$\begin{aligned} Y_i^{(0)} - \widehat{Y}_i^{(0)} &= D_i^{(0)^\top}\left(\rho_0 - \hat{\rho}^*_{\mathrm{HPTL},G}\right) + \sum_{k=1}^{K} X_{i,\mathcal{J}_k}^{(0)^\top}\left(\beta_{0,\mathcal{J}_k} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) + V_i^{(0)} \\ &= O_p\left(n_0^{\frac{1}{2}}\right) + K * o_p\left(n_0^{\frac{1}{2}}\right) + \sum_{k=1}^{K} O_p\left(n_k^{-\frac{1}{2}} + a_n\right) + V_i^{(0)} \\ &\overset{d}{\sim} V_i^{(0)} \end{aligned} \tag{36}$$

as $n_0 \to \infty$.

Proof.

$$\begin{aligned} \hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0 &= \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\widehat{Y}_H^{*(0)} - \rho_0 \\ &= \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\left(Y^{(0)} - \sum_{k=1}^{K}\mathbf{X}_{\mathcal{J}_k}^{(0)}\hat{\beta}_{\mathcal{J}_k}^{(k)}\right) - \rho_0 \\ &= \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\left(\mathbf{X}^{(0)}\beta_0 - \sum_{k=1}^{K}\mathbf{X}_{\mathcal{J}_k}^{(0)}\hat{\beta}_{\mathcal{J}_k}^{(k)} + V^{(0)}\right) \\ &= \sum_{k=1}^{K}\left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\mathbf{X}_{\mathcal{J}_k}^{(0)}\left(\beta_{0,\mathcal{J}_k} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) \\ &+ \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}V^{(0)} \\ &= \sum_{k=1}^{K}\left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}\mathbf{X}_{\mathcal{J}_k}^{(0)}\left(\beta_{0,\mathcal{J}_k}^{(k)} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) \\ &+ \left(\mathbf{D}^{(0)^\top}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)^\top}V^{(0)} \\ &\xrightarrow{p} 0 \end{aligned} \tag{37}$$

as $n_0 \to \infty$. Also

$$\begin{aligned} \sqrt{n_0}(\hat{\rho}^*_{\mathrm{HPTL},G} - \rho_0) &= \sum_{k=1}^{K}\widehat{Q}^{-1}\mathbf{D}^{(0)^\top}\mathbf{X}_{\mathcal{J}_k}^{(0)}\frac{\sqrt{n_0}}{\sqrt{n_k}}\sqrt{n_k}\left(\beta_{0,\mathcal{J}_k}^{(k)} - \hat{\beta}_{\mathcal{J}_k}^{(k)}\right) \\ &+ \widehat{Q}^{-1}\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}D_i^{(0)}V_i^{(0)} \\ &\xrightarrow{d} N\left(0, Q^{-1}\Sigma_0 Q^{-1}\right) \end{aligned} \tag{38}$$

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

7.A More Simulation Results

7.2.A 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	$H = 0.01$	$H = 0.05$	$H = 0.1$
$n_k = 100$	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

	$s = 5$	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

7.3.A 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