Transfer Learning with General Estimating Equations

Han Yan

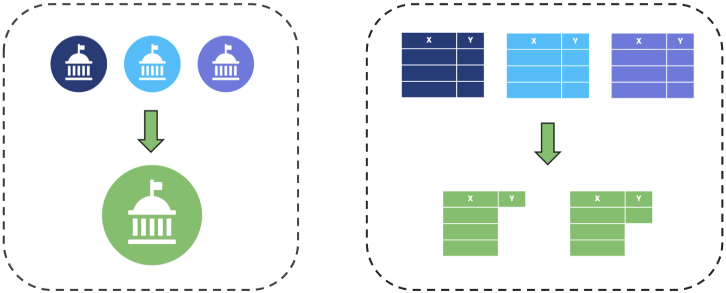
Joint work with Songxi Chen

Peking University

July 12, 2024

Motivation for TL

- High-quality labels often involve laborious human annotations or slow and expensive scientific measurements.
- **Transfer Learning:** utilize different but related source domain to facilitate the learning on the target domain?



Goal: Efficient and flexible method for the inference of general estimating equations (GEE) under the TL.

Problem Formulation

- **Source sample** $\mathcal{D}_S = \{\boldsymbol{Z}_i\}_{i=1}^{n} \sim P = P_{\boldsymbol{X}} \times P_{Y|\boldsymbol{X}}$ where $\boldsymbol{Z}_i = (\boldsymbol{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}$.

  Target sample $\mathcal{D}_T = \{\boldsymbol{X}_i\}_{i=n+1}^{N} \sim Q_{\boldsymbol{X}}$ with $N = n + m$, while the responses Y_i in $\mathcal{D}_T$ are *not* accessible.

- Our goal is the inference of a p-dimensional parameter $\boldsymbol{\theta}_0$ identified by the GEE

$$\mathbb{E}_Q\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}_0)\} = \mathbf{0}, \tag{1}$$

  where $Q = Q_{\boldsymbol{X}} \times Q_{Y|\boldsymbol{X}}$ and $\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}) = (g_1(\boldsymbol{Z}, \boldsymbol{\theta}), \ldots, g_r(\boldsymbol{Z}, \boldsymbol{\theta}))^{\mathrm{T}}$ with $r \geq p$.

- **Covariate shift TL:** we assume $P_{Y|\boldsymbol{X}} = Q_{Y|\boldsymbol{X}}$, while $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ can differ.

Density ratio weighting

- The most common method for the covariate shift is the density ratio weighting (DRW).
- Let $r_0(\boldsymbol{x}) = q_0(\boldsymbol{x})/p_0(\boldsymbol{x})$ be the density ratio of $Q_{\boldsymbol{X}}$ and $P_{\boldsymbol{X}}$. Then,

$$\mathbb{E}_P\{r_0(\boldsymbol{X})\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\} = \mathbb{E}_Q\left\{\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\right\},$$

- With a consistent $\widehat{r}(\boldsymbol{x})$, we can obtain an estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ from the DRW moment function $\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta},\widehat{r}) = \mathbf{0}$, where

$$\tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta},\widehat{r}) = \widehat{r}(\boldsymbol{X}_i)\mathbf{g}(\boldsymbol{Z}_i,\boldsymbol{\theta}) \quad \text{for } i = 1,\ldots,n, \tag{2}$$

with either the empirical likelihood or the GMM.

Drawbacks of the DRW:

- The accuracy of $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ crucially hinges on that of $\widehat{r}$.
- The EL ratio is *not* asymptotically χ^2 distributed, and requires a Bootstrap to approximate its asymptotic distribution.

Density ratio weighting

- The most common method for the covariate shift is the density ratio weighting (DRW).
- Let $r_0(\boldsymbol{x}) = q_0(\boldsymbol{x})/p_0(\boldsymbol{x})$ be the density ratio of $Q_{\boldsymbol{X}}$ and $P_{\boldsymbol{X}}$. Then,

$$\mathbb{E}_P\{r_0(\boldsymbol{X})\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\} = \mathbb{E}_Q\left\{\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\right\},$$

- With a consistent $\widehat{r}(\boldsymbol{x})$, we can obtain an estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ from the DRW moment function $\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta},\widehat{r}) = \mathbf{0}$, where

$$\tilde{\mathbf{g}}(\boldsymbol{Z}_i,\boldsymbol{\theta},\widehat{r}) = \widehat{r}(\boldsymbol{X}_i)\mathbf{g}(\boldsymbol{Z}_i,\boldsymbol{\theta}) \quad \text{for } i = 1,\ldots,n, \tag{2}$$

with either the empirical likelihood or the GMM.

Drawbacks of the DRW:

- The accuracy of $\widehat{\boldsymbol{\theta}}^{\mathrm{drw}}$ crucially hinges on that of $\widehat{r}$.
- The EL ratio is *not* asymptotically χ^2 distributed, and requires a Bootstrap to approximate its asymptotic distribution.

Neyman orthogonal estimating function

- To alleviate the effect of the nuisance function estimation, we construct an estimating function which has the Neyman orthogonal property (Neyman, 1959; Chernozhukov et al., 2018).
- Let $\boldsymbol{W}_i = (\boldsymbol{X}_i, \delta_i Y_i, \delta_i)$, where $\delta_i = 0$ if the i-th observation belongs to the source sample and $\delta_i = 1$ otherwise.
- Let $\mathbf{m}_0(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}) | \boldsymbol{X} = \boldsymbol{x}\}$, the constructed estimating function is

$$\boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) = \frac{1-\delta_i}{1-\tau} \widehat{r}(\boldsymbol{X}_i)\{\mathbf{g}(\boldsymbol{Z}_i, \boldsymbol{\theta}) - \widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\} + \frac{\delta_i}{\tau} \widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta}), \tag{3}$$

  for $i = 1, \cdots, N$, where $\widehat{\boldsymbol{\eta}} = (\widehat{r}, \widehat{\mathbf{m}})$ is an estimate of $(r_0, \mathbf{m})$, δ_i is a binary indicator of whether the i-th observation belongs to the target sample or not.

Orthogonal estimating function

The following conditions are required for the sample and target populations.

Condition 1

(i) The covariate distributions $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ are absolutely continuous with densities $p_0(\boldsymbol{x})$ and $q_0(\boldsymbol{x})$ supported on $\mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact. (ii) The conditional distributions $P_{Y|\boldsymbol{X}=\boldsymbol{x}} = Q_{Y|\boldsymbol{X}=\boldsymbol{x}}$ for every $\boldsymbol{x} \in \mathcal{X}$.

Condition 2

(i) The parameter $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$ is the unique solution to the moment condition $\mathbb{E}_Q\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})\} = 0$. (ii) $\mathbb{E}_Q\{\sup_{\boldsymbol{\theta}\in\Theta} \|\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})\|_2^{\alpha}\} < \infty$ for some $\alpha > 2$. (iii) The eigenvalues of $\mathbb{E}_Q\{\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})^{\otimes 2}\}$ are bounded away from zero and infinity. (iv) $\mathbf{g}(\boldsymbol{z}, \boldsymbol{\theta})$ is continuously differentiable in a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$ with $\mathbb{E}_Q\{\sup_{\boldsymbol{\theta}\in\mathcal{N}} \|\partial\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\|_2\} < \infty$, and $\mathbb{E}_Q\{\partial\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\}$ is of full rank.

Orthogonal estimating function

Theorem 1

Under Conditions 1 and 2, the following results hold.

(i) $\mathbf{\Psi}(\boldsymbol{W}, \boldsymbol{\theta}_0, \boldsymbol{\eta})$ is Neyman orthogonal in the sense that

$$\frac{\partial}{\partial \tau} \mathbb{E}_F \{\mathbf{\Psi}(\boldsymbol{W}, \boldsymbol{\theta}_0, \boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0} = \mathbf{0}. \tag{4}$$

(ii) For any candidate $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta}) = (r(\boldsymbol{x}), \mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta}))$,

$$\|\mathbb{E}_F\{\mathbf{\Psi}(\boldsymbol{W}, \boldsymbol{\theta}_0, \boldsymbol{\eta})\}\|_1 \leq \|r(\boldsymbol{X}) - r_0(\boldsymbol{X})\|_{L_2(P_{\boldsymbol{X}})} (\sum_{j=1}^{r} \|m_j(\boldsymbol{X}, \boldsymbol{\theta}) - m_{0j}(\boldsymbol{X}, \boldsymbol{\theta})\|_{L_2(P_{\boldsymbol{X}})}).$$

$\mathbf{\Psi}(\boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{\eta})$ is robust against the estimation error of nuisance functions.

Challenges

The proposed estimating function

$$\mathbf{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) = \frac{1-\delta_i}{1-\tau}\widehat{r}(\boldsymbol{X}_i)\{\mathbf{g}(\boldsymbol{Z}_i, \boldsymbol{\theta}) - \widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\} + \frac{\delta_i}{\tau}\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})$$

s similar to that of the AIPW (Robins et al., 1994) and the double machine learning (DML, Chernozhukov et al., 2018), but is more challenging as we consider the GEE rather than only the average treatment effect.

- In general cases, such as the quantile regression, the conditional mean function $\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta})$ is parametric-dependent, so we have to estimate it at all possible $\boldsymbol{\theta}$, which is practically infeasible.
- Most existing estimation methods for the density ratio function $r(\boldsymbol{x})$ are not flexible enough to accommodate complex function structures.

Density ratio estimation

- Conventional methods, such as the kernel smoothing, suffer from instability and the curse of dimensionality.
- We employ a ϕ-divergence based density ratio estimation approach, which can be solved via an empirical risk minimization problem and can accommodates a variety of machine learning algorithms.
- The ϕ-divergence of Q from P is defined by:

$$D_\phi(Q\|P) = \int \phi\left(\frac{q_0(x)}{p_0(x)}\right) p_0(x)dx, \tag{5}$$

  where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a convex and lower semicontinuous function.
- Let ϕ_* be the Frenchel dual function of ϕ defined by $\phi_*(v) = \sup_{u\in\mathbb{R}}\{uv - \phi(u)\}$.

Density ratio estimation

For each ϕ-function, we define

$$\ell_{1,\phi}(r) = \phi_*\{\phi'(r)\} \text{ and } \ell_{2,\phi}(r) = \phi'(r). \tag{6}$$

The dual characteristic of $D_\phi(Q\|P)$ induces an identification condition for the density ratio r_0 as presented in the following lemma.

Lemma 1

For any convex and lower semicontinuous function $\phi : \mathbb{R}_+ \to \mathbb{R}$, the true density ratio satisfies

$$r_0 = \underset{r\in\mathcal{F}}{\arg\min}\, L_\phi(r) \text{ with } L_\phi(r) = \mathbb{E}_P\{\ell_{1,\phi}(r)\} - \mathbb{E}_Q\{\ell_{2,\phi}(r)\}, \tag{7}$$

where the candidate class $\mathcal{F}$ is any class of nonnegative functions that contains r_0.

Density ratio estimation

Table: Examples of ϕ-divergence, the associated Fenchel conjugate and the objective functions.

Divergence	$\phi(u)$	$\phi_*(v)$	$\ell_{1,\phi}(r)$	$\ell_{2,\phi}(r)$
Kullback-Leibler	$u\log(u)$	$\exp(v-1)$	r	$\log(r)+1$
Reverse KL	$-\log(u)$	$-1-\log(-v)$	$\log(r)+1$	$-r^{-1}$
Pearson χ^2	$(u-1)^2$	$v^2/4+v$	r^2-1	$2(r-1)$
Squared Hellinger	$(\sqrt{u}-1)^2$	$v/(v-1)$	$r^{\frac{1}{2}}-1$	$1-r^{-\frac{1}{2}}$

- With the two samples from P and Q, the density ratio r_0 can be estimated with the sample objective function

$$\widehat{r} = \operatorname*{arg\,min}_{r\in\mathcal{F}_N}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell_{1,\phi}\{r(\boldsymbol{X}_i)\} - \frac{1}{m}\sum_{i=n+1}^{N}\ell_{2,\phi}\{r(\boldsymbol{X}_i)\}\right\}. \tag{8}$$

- Different from the kernel smoothing method, (8) directly estimates the ratio function via an empirical risk minimization.
- It is noted that we not only obtain the estimator $\widehat{r}$, but also an estimate of the divergence $D_\phi(Q\|P)$ by the sample objective function with $\widehat{r}$.

Regularity conditions for density ratio estimation

Condition 3

There exist constants $B_1 > 0$ and $\beta_1 \geq 1$ such that the target function $r_0 \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$.

Condition 4

Let the pseudo-dimension (see Pollard, 1990) of $\mathcal{F}_N$ be $\mathrm{Pdim}(\mathcal{F}_N)$, then (i) $\mathrm{Pdim}(\mathcal{F}_N)\log(N) = o(N)$; and (ii) there exists a constant $c_2 > 0$ such that for large enough n, $\inf_{r\in\mathcal{F}_N} \|r - r_0\|_\infty \leq c_2 \mathrm{Pdim}(\mathcal{F}_N)^{-\frac{\beta_1}{d}}$. (iii) There exists a positive constant M_1 such that $\|r\|_\infty \leq M_1$ and $\|\ell''_{i,\phi}(r)\|_\infty \leq M_1$ for $i = 1, 2$ and for every $r \in \mathcal{F}_N$.

- For linear **sieve** function classes, the pseud-dimension $\mathrm{Pdim}(\mathcal{F}_N)$ equals to the number of basis functions (Chen, 2007).
- For **deep neural networks** (DNN) with the width W and depth L, the pseud-dimension $WL\log(W/L) \lesssim \mathrm{Pdim}(\mathcal{F}_n) \lesssim WL\log(W)$.
- The approximation error $\inf_{r\in\mathcal{F}_N} \|r - r_0\|_\infty \lesssim \mathrm{Pdim}(\mathcal{F}_N)^{-\frac{\beta}{d}}$ is attainable for both sieve functions and DNNs.

Estimation error bound of $\widehat{r}$

To quantity the estimation performance, we define empirical L_2 error of $\widehat{r}$ as

$$\mathcal{E}_N(\widehat{r}) = [N^{-1}\sum_{i=1}^{N}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2]^{1/2}. \tag{9}$$

Theorem 2

Under Conditions 1, 3, and 4, there exists a positive constant C_1 such that with probability at least $1-2e^{-t}$, for N large enough and any $t>0$,

$$\mathcal{E}_N(\widehat{r}) \leq C_1\left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N)\log(N)}{N}} + \inf_{r\in\mathcal{F}_N}\|r-r_0\|_\infty + \sqrt{\frac{t}{N}}\right). \tag{10}$$

Corollary 1

Under Conditions 1, 3, and 4, and taking $\mathrm{Pdim}(\mathcal{F}_N) = O(N^{-\frac{d}{2\beta_1+d}})$, *we have*

$$\mathcal{E}_N(\widehat{r}) = O_p\left(N^{-\frac{\beta_1}{2\beta_1+d}}\log^{\frac{1}{2}}(N)\right).$$

- In practice, we can specify the optimal $\mathrm{Pdim}(\mathcal{F}_N)$ with cross-validations.
- The above estimation error attains the minimax lower bound for density ratio estimation.

Multiple imputation

- The next goal is to estimate the conditional mean function $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}(\boldsymbol{X}, Y, \boldsymbol{\theta}) | \boldsymbol{X}\}$.
- **Directly estimating $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta})$ is not feasible, since it has to be estimated at infinitely many $\boldsymbol{\theta}$.**
- Wang and Chen (2009) proposed to make κ independent imputations $\{\tilde{Y}_\nu\}_{\nu=1}^{\kappa}$ from a kernel estimator

$$\widehat{F}(y|\boldsymbol{X}) = \sum_{i=1}^{n} \frac{K((\boldsymbol{X}_i - \boldsymbol{X})/h) I(Y_i \leq y)}{K((\boldsymbol{X}_i - \boldsymbol{X})/h)},$$

and then estimate $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta})$ by

$$\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}, \boldsymbol{\theta}) = \kappa^{-1} \sum_{\nu=1}^{\kappa} \mathbf{g}(\boldsymbol{X}, \tilde{Y}_\nu, \boldsymbol{\theta}).$$

- $\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}, \boldsymbol{\theta})$ is asymptotically equivalent to the Nadaraya–Watson estimator $\widehat{\mathbf{m}}(\boldsymbol{X}, \boldsymbol{\theta}) = \int \mathbf{g}(\boldsymbol{X}, Y, \boldsymbol{\theta}) d\widehat{F}(y|\boldsymbol{X})$.
- By sampling from the conditional distribution, multiple imputation bypasses estimating $\widehat{\mathbf{m}}(\boldsymbol{X}, \boldsymbol{\theta})$ explicitly at every $\boldsymbol{\theta}$.

Multiple imputation

- The next goal is to estimate the conditional mean function $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}(\boldsymbol{X}, Y, \boldsymbol{\theta})|\boldsymbol{X}\}$.
- **Directly estimating $\mathrm{m}(\boldsymbol{X}, \boldsymbol{\theta})$ is not feasible, since it has to be estimated at infinitely many $\boldsymbol{\theta}$.**
- Wang and Chen (2009) proposed to make κ independent imputations $\{\tilde{Y}_\nu\}_{\nu=1}^{\kappa}$ from a kernel estimator

$$\widehat{F}(y|\boldsymbol{X}) = \sum_{i=1}^{n} \frac{K((\boldsymbol{X}_i - \boldsymbol{X})/h) I(Y_i \leq y)}{K((\boldsymbol{X}_i - \boldsymbol{X})/h)},$$

and then estimate $\mathbf{m}(\boldsymbol{X}, \boldsymbol{\theta})$ by

$$\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}, \boldsymbol{\theta}) = \kappa^{-1} \sum_{\nu=1}^{\kappa} \mathbf{g}(\boldsymbol{X}, \tilde{Y}_\nu, \boldsymbol{\theta}).$$

- $\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}, \boldsymbol{\theta})$ is asymptotically equivalent to the Nadaraya–Watson estimator $\widehat{\mathbf{m}}(\boldsymbol{X}, \boldsymbol{\theta}) = \int \mathbf{g}(\boldsymbol{X}, Y, \boldsymbol{\theta}) d\widehat{F}(y|\boldsymbol{X})$.
- By sampling from the conditional distribution, multiple imputation bypasses estimating $\widehat{\mathbf{m}}(\boldsymbol{X}, \boldsymbol{\theta})$ explicitly at every $\boldsymbol{\theta}$.

Conditional Density Estimation

- The key to the multiple imputations is the conditional density estimation.
- Since the conditional density function is essentially a density ratio between $p_0(y, \boldsymbol{x})$ over $p_0(\boldsymbol{x})$, it is natural to employ the ϕ-divergence based density estimation.
- To ensure the support of the denominator density covers that of the numerator density, we introduce an auxiliary variable $\tilde{Y} \sim \tilde{P}_Y$ and express the conditional density as

$$p_0(y|\boldsymbol{x}) = \frac{p_0(y, \boldsymbol{x})}{p_0(\boldsymbol{x})} = \frac{p_0(y, \boldsymbol{x})}{p_0(\boldsymbol{x})\tilde{p}_0(y)}\tilde{p}_0(y) =: \tilde{r}_0(y, \boldsymbol{x})\tilde{p}_0(y). \tag{11}$$

Estimating $p_0(y|\boldsymbol{x})$ amounts to estimating the $\tilde{r}_0(y, \boldsymbol{x})$, the density ratio between $P_{\boldsymbol{X},Y}$ and $P_{\boldsymbol{X}} \times \tilde{P}_Y$.

Conditional Density Estimation

- Let $\mathcal{G}_N$ be a $(d+1)$-dimensional candidate function class that satisfies Condition 6 below, then the density ratio $\tilde{r}_0(y, \boldsymbol{x})$ can be estimated via the following sample criterion

$$\widehat{\tilde{r}}(y, \boldsymbol{x}) = \underset{p \in \mathcal{G}_N}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{1,\phi}\{p(\tilde{Y}_i, \boldsymbol{X}_i)\} - \frac{1}{n} \sum_{i=1}^{n} \ell_{2,\phi}\{p(Y_i, \boldsymbol{X}_i)\} \right\}, \tag{12}$$

where $\{\tilde{Y}_i\}_{i=1}^n$ are independently sampled from $\tilde{P}_Y$.

- With $\widehat{\tilde{r}}(y, \boldsymbol{x})$, the conditional density is estimated by

$$\widehat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \widehat{\tilde{r}}(y, \boldsymbol{x}) \tilde{p}_0(y). \tag{13}$$

- Using the conditional density estimator $\widehat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$, for any $\boldsymbol{X}_i \in \{\boldsymbol{X}_l\}_{l=1}^N$, we generate a sample $\{\tilde{Y}_i^{\nu}\}_{\nu=1}^{\kappa}$ independently from $\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)$. Then, the imputed moment function is

$$\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}_i, \boldsymbol{\theta}) = \frac{1}{\kappa} \sum_{\nu=1}^{\kappa} \mathbf{g}(\boldsymbol{X}_i, \tilde{Y}_i^{\nu}, \boldsymbol{\theta}).$$

Estimation error bound of $\widehat{\mathbf{m}}_\kappa$

Condition 5

(i) The support of $\tilde{P}_Y$ covers that of P_Y, and (ii) the density function of $\tilde{P}_Y$ is uniformly bounded. (iii) There exist constants $B_2 > 0$ and $\beta_2 \geq 1$ such that the true conditional density function $p_{Y|\boldsymbol{X}} \in \mathcal{H}^{\beta_2}(\mathcal{Y} \times \mathcal{X}, B_2)$. (iv) $\inf_{y \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}} p_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) > 0$.

Condition 6

The pseudo-dimension of $\mathcal{G}_N$ satisfies (i) $\mathrm{Pdim}(\mathcal{G}_N) \log(N) = o(N)$, and (ii) there exists a constant $c_3 > 0$ such that for large enough n, $\inf_{p \in \mathcal{G}_N} \|p - p_{Y|\boldsymbol{X}}\|_\infty \leq c_3 \mathrm{Pdim}(\mathcal{G}_N)^{-\frac{\beta_2}{d+1}}$. (iii) There exists a positive constant M_2 such that $\|p\|_\infty \leq M_2$ and $\|\ell''_{i,\phi}(p)\|_\infty \leq M_2$ for $i = 1, 2$ for every $p \in \mathcal{G}_N$.

Condition 7

There exists a positive constant $\sigma_g > 0$ such that $\mathbb{E}\{\exp(\lambda \|\mathbf{g}(\boldsymbol{Z}, \boldsymbol{\theta}\|^2) | \boldsymbol{X} = \boldsymbol{x}\} < \exp(\lambda \sigma_g^2)$ for all $0 \leq \lambda \leq \sigma_g^{-2}$ for each $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{x} \in \mathcal{X}$.

Estimation error bound of $\widehat{\mathbf{m}}_\kappa$

Define the empirical L_2 error of $\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}, \boldsymbol{\theta})$ as

$$\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = \sum_{j=1}^{r}[N^{-1}\sum_{i=1}^{N}\{\widehat{m}_{\kappa j}(\boldsymbol{X}_i, \boldsymbol{\theta}) - m_{0j}(\boldsymbol{X}_i, \boldsymbol{\theta})\}^2]^{1/2}, \tag{14}$$

where $\widehat{m}_{\kappa j}$ and m_{0j} are the j-th component of $\widehat{\mathbf{m}}_\kappa$ and $\mathbf{m}_0$, respectively.

Theorem 3

Under Conditions 1, 5–7 and taking $\mathrm{Pdim}(\mathcal{G}_N) = O(N^{-\frac{d+1}{2\beta_2+d+1}})$ and $\kappa \gtrsim N$, for any $\boldsymbol{\theta} \in \Theta$,

$$\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = O_p\left(N^{-\frac{\beta_2}{2\beta_2+d+1}} \log^{\frac{3}{2}}(N)\right).$$

Empirical likelihood inference

- Using the orthogonal moment function $\boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$ with $\widehat{\boldsymbol{\eta}}(\boldsymbol{X}_i, \boldsymbol{\theta}) = (\widehat{r}(\boldsymbol{X}_i), \widehat{\mathbf{m}}_\kappa(\boldsymbol{X}_i, \boldsymbol{\theta}))$, the EL estimator of $\boldsymbol{\theta}_0$ is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} L_N(\boldsymbol{\theta}) \tag{15}$$

where $L_N(\boldsymbol{\theta})$ is the profile EL

$$L_N(\boldsymbol{\theta}) = \sup\left\{\prod_{i=1}^N p_i \;\middle|\; p_i \geq 0,\ \sum_{i=1}^N p_i = 1,\ \sum_{i=1}^N p_i \boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\boldsymbol{X}_i, \boldsymbol{\theta})) = \mathbf{0}\right\}. \tag{16}$$

- Let $\boldsymbol{\Gamma} = \mathbb{E}\{\partial\boldsymbol{\Psi}(\boldsymbol{W}, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0)/\partial\boldsymbol{\theta}\}$, $\boldsymbol{\Omega} = \mathbb{E}\{\boldsymbol{\Psi}(\boldsymbol{W}, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0)^{\otimes 2}\}$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}$.

Theorem 4

Under Conditions 1 and 2, if the estimation errors satisfy

$$\mathcal{E}_N(\widehat{r}) + \mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(1) \text{ and } \mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(N^{-\frac{1}{2}}), \tag{17}$$

for every $\boldsymbol{\theta} \in \Theta$, then we have

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{18}$$

Empirical likelihood inference

- Under Conditions 3–7 where r_0 and $p_{Y|\boldsymbol{X}}$ have the smoothness of β_1 and β_2, respectively, then (17) is attainable provided that

$$\frac{\beta_1}{2\beta_1+d}+\frac{\beta_2}{2\beta_2+d+1}>\frac{1}{2}, \tag{19}$$

- The asymptotic variance of $\widehat{\boldsymbol{\theta}}$ reaches the semiparametric efficiency bound.

Let the log EL ratio be $\ell_N(\boldsymbol{\theta})=-\log\{L_N(\boldsymbol{\theta})/N^{-N}\}$ and $R_N(\boldsymbol{\theta}_0)=2\ell_N(\boldsymbol{\theta}_0)-2\ell_N(\widehat{\boldsymbol{\theta}})$.

Theorem 5

Under the same conditions as in Theorem 4, as $N\to\infty$,

$$R_N(\boldsymbol{\theta}_0)\xrightarrow{d}\chi_r^2.$$

- In the presence of nuisance functions, the log EL ratio no longer necessarily converges weakly to a central χ^2 distribution but may be a weighted sum of χ^2 distributions, whose critical values require Bootstrap to approximate (e.g., Chen et al., 2024).
- We overcome such a situation and restore Wilks' theorem due to the orthogonality of the estimating function.

Extension to growing dimensions

- We consider the inference for $\boldsymbol{\theta}$ with the presence of a high dimensional covariate.
- Without structural assumptions, the convergence rates of $\widehat{r}$ and $\widehat{\mathbf{m}}_\kappa$ attains the corresponding minimax lower bounds.
- There have been increasing studies indicating that high-dimensional data tend to be supported on low-dimensional manifolds in many applications, such as image analysis and natural language processing (Goodfellow et al., 2016).

Condition 8 (Approximate manifold support)

The covariate distributions $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ are concentrated on $\mathcal{M}_\rho$, a ρ-neighborhood of $\mathcal{M} \subset \mathcal{X}$, where $\mathcal{M}$ is a compact $d_{\mathcal{M}}$-dimensional Riemannian manifold (Lee, 2006) and $\mathcal{M}_\rho = \{\boldsymbol{x} \in \mathcal{X} : \inf\{\|\boldsymbol{x} - \boldsymbol{y}\|_2 : \boldsymbol{y} \in \mathcal{M}\} \le \rho\},\ \rho \in (0, 1)$.

The dimension $d_{\mathcal{M}}$ of the manifold $\mathcal{M}$ can be regarded as an intrinsic dimension. We allow the nominal dimension d to diverge with N, while taking $d_{\mathcal{M}}$ as a fixed constant.

Circumventing the curse of dimensionality

- Jiao et al. (2023) established that the fully connected DNNs can adaptively estimate a smooth function with the manifold assumption, hence alleviating the curse of dimensionality.
- We choose the function classes $\mathcal{F}_N$ and $\mathcal{G}_N$ as the DNNs with the ReLU activation function. The widths for $\mathcal{F}_N$ and $\mathcal{G}_N$ are specified as W_1 and W_2, and the depths are specified as D_1 and D_2, respectively.
- Let $\tilde{d}_{\mathcal{M}} = O(d_{\mathcal{M}} \log(d/\delta)/\delta^2)$ be an integer such that $d_{\mathcal{M}} \leq \tilde{d}_{\mathcal{M}} < d$, where $\delta \in (0,1)$.

Theorem 6

Under Conditions 3–8, let the widths and depths of $\mathcal{F}_N$ and $\mathcal{G}_N$ be

$$W_i = 114(\lfloor \beta_i \rfloor + 1)^2 \tilde{d}_{\mathcal{M}}^{\lfloor \beta_i \rfloor + 1} \quad \text{and} \quad D_i = 21(\lfloor \beta_i \rfloor + 1)^2 N^{\tilde{d}_{\mathcal{M}}/2(\tilde{d}_{\mathcal{M}}+2\beta_i)} \lceil \log_2(8N^{\tilde{d}_{\mathcal{M}}/2(\tilde{d}_{\mathcal{M}}+2\beta_i)}) \rceil,$$

for $i = 1$ and 2. Then, the estimation errors of $\widehat{r}$ and $\widehat{\mathbf{m}}_{\boldsymbol{\theta}}$ satisfy

$$\mathcal{E}_N(\widehat{r}) = O_p\left(d^{\frac{1}{2}} N^{-\frac{\beta_1}{\tilde{d}_{\mathcal{M}}+2\beta_1}} \log^{\frac{1}{2}}(N)\right) \quad \text{and}$$
$$\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = O_p\left((d+1)^{\frac{1}{2}} N^{-\frac{\beta_2}{(\tilde{d}_{\mathcal{M}}+1+2\beta_2)}} \log^{\frac{3}{2}}(N)\right), \quad \text{respectively.} \tag{20}$$

Circumventing the curse of dimensionality

Theorem 7

Under Conditions 1–8 and suppose that $d = O(N^k)$ for some $k \geq 0$ that satisfies

$$\frac{\beta_1}{2\beta_1 + \tilde{d}_{\mathcal{M}}} + \frac{\beta_2}{2\beta_2 + \tilde{d}_{\mathcal{M}} + 1} > \frac{2+k}{4}. \tag{21}$$

If $r^3 p^2 N^{-1} = o(1)$ and $r^3 N^{2/\alpha - 1} = o(1)$, where $\alpha > 2$ is the order of moment defined in Condition 2, then as $r, p, N \to \infty$, (i) for any $\boldsymbol{u}_n \in \mathbb{R}^p$ with unit L_2-norm,

$$\sqrt{N}\boldsymbol{u}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0,1); \tag{22}$$

(ii) the EL ratio statistic $R_N(\boldsymbol{\theta}_0)$ satisfies

$$(2r)^{-\frac{1}{2}}\{R_N(\boldsymbol{\theta}_0) - r\} \xrightarrow{d} \mathcal{N}(0,1). \tag{23}$$

The above asymptotic distributions of $\widehat{\boldsymbol{\theta}}$ and $R_N(\boldsymbol{\theta}_0)$ recover those in Chang et al. (2015) in the absence of the nuisance functions.

Case study: TL for O_3 pollutions

We demonstrate that the proposed method is well-suited for the transfer learning of the inference for the O_3 levels.

- **Source domain:** Beijing, Xian, and Jinan;

  Target domain: Taiyuan
- **Study period**: spring (March 1 to May 31) of 2018.
- **Response:** O_3 levels;

  Covariates: meteorological variables and PM.

To investigate the performances of the TL, we assumed only the covariate variables of the target domain Taiyuan were observable during their implementations, while the true O_3 levels of the target sample were used to evaluate the quality of the transfer learning.

Imputations for O_3 of the target domain

Performance of the multiple imputations for O_3 of the target sample is demonstrated in Figure 1, which verifies that the conditional density of the target sample was similar to that of the source, but also shows that our multiple imputation method produced high-quality surrogates for the O_3 on the target domain.

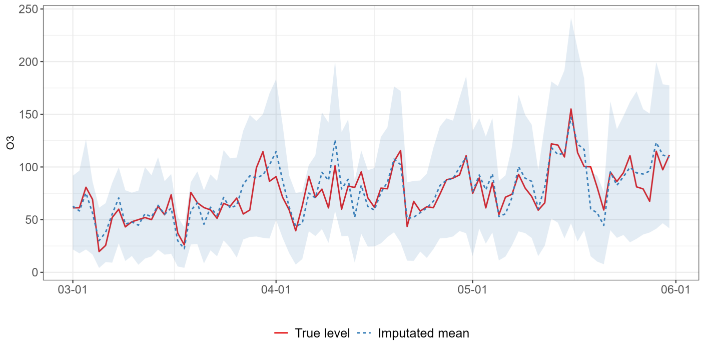


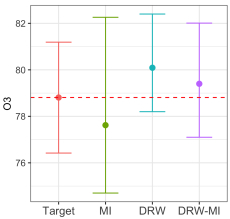
Figure: Illustration for the results of the multiple imputations for O_3 the target sample. The upper and lower boundaries of the blue region are the 2.5% and 97.5% empirical quantiles of the 200 imputations. The blue dotted line is the empirical mean of the imputed values. The red line indicates the true O_3 levels of the target sample.

Inference for O_3 of the target domain

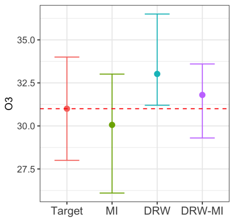
We considered the estimation and inference for the mean and the α-quantiles ($\alpha = 25\%, 50\%$, and 75%) of the O_3 of the target domain in Taiyuan. The methods include the multiple imputation (MI), the density ratio weighting (DRW), and the proposed method (DRW-MI).

Figure: Estimation and 95% confidence intervals for the mean and three quantiles of the O_3 of the target population obtained from the target sample, the multiple imputations (MI), the density ratio weighting (DRW), and the density ratio weighting with multiple imputations (DRW-MI), respectively. As a comparison baseline, the red dotted line indicates the estimated value of the O_3 with the target sample.

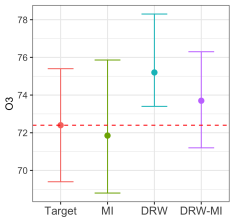
(a) Mean



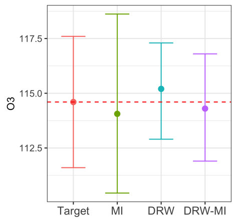
(b) 25%-quantile



(c) 50%-quantile



(d) 75%-quantile



Summary

1. We construct a Neyman orthogonal estimating function for the covariate shift, which is more robust against nuisance function estimation errors compared with existing methods.
2. We propose novel methods for nuisance function estimation that enable the use of flexible nonparametric tools, including generic ML algorithms.
3. With a multiple imputation strategy, we overcome the challenge that one of the nuisances $\mathbf{m}(\boldsymbol{x}, \boldsymbol{\theta})$ is parametric-dependent, namely it has to be estimated at infinitely many $\boldsymbol{\theta}$.
4. By employing the EL method, the proposed estimation is shown to be semi-parametric efficient. The log EL ratio statistics admits Wilks' theorem which greatly facilitates the inference, while existing methods commonly require Bootstraps.
5. We also discuss a growing dimension scenario and adopt deep neural networks to mitigate the curse of dimensionality.

Thank You!

References I

Chang, J., Chen, S. X., and Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185(1):283–304.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Chen, X., Liu, Y., Ma, S., and Zhang, Z. (2024). Causal inference of general treatment effects using neural networks with a diverging number of confounders. *Journal of Econometrics*, 238(1):105555.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716.

Lee, J. M. (2006). *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and Statsitics*, pages 213–234.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, volume 2. Institute of Mathematical Statistics.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 37(1):490 – 517.