# Supplementary Material for
Causal Inference with Treatment Measurement Error: A Nonparametric Instrumental Variable Approach

**Yuchen Zhu**[1] **Limor Gultchin**[2,3] **Arthur Gretton**[1] **Matt Kusner**[1] **Ricardo Silva**[1]

[1]Department of Computer Science, University College London, UK
[2]Department of Computer Science, University of Oxford, UK
[3]The Alan Turing Institute, London, UK

## 1 PSEUDOCODE

See Figure 5 for the pseudocode of our method.

## 2 RELAXING CLASSICAL MEASUREMENT ERROR ASSUMPTIONS

In Hu and Sasaki [2015], the authors assume the noise on the second measurement $N$, is an unknown monotonic polynomial function of $X$ with additive noise. The estimation procedure amounts of first identifying the polynomial function of $X$, before applying a technique similar to Schennach [2004]. In this work we take the first step to extend the estimator of Schennach [2004] to the confounded setting, and leave for future work the relaxation on the assumptions of the second measurement.

## 3 FURTHER ASSUMPTIONS ON KERNEL IDENTIFICATION

We also employ the following technical assumptions to enable causal effect estimation in the latent treatment setting.

**Assumption 8** $\mathcal{Z}, \mathcal{X}, \mathcal{M}, \mathcal{N}, \mathcal{Y}$ are measurable, separable Polish spaces.

Assumption 8 is a regularity condition that allows us to define the conditional mean embedding operator.

**Assumption 9** $Y$ is bounded.

**Assumption 10** $k(x, \cdot), k(m, \cdot), k(n, \cdot)$ are continuous, bounded by $\kappa > 0$, and their feature maps are measurable. (ii) $k(x, \cdot), k(m, \cdot), k(n, \cdot)$ are characteristic kernels.

Assumption 10 is a standard assumption employed in kernel causal learning (Singh et al. [2019] Mastouri et al. [2021]).

## 4 $s-$SAMPLE ESTIMATES

For clarity, we state the $s-$sample estimates for $\hat{\psi}_{\mathcal{P}_{X|z}}(\alpha)$, $\hat{\psi}_{\mathcal{P}_{N|z}}(\alpha)$, $\hat{\psi}_{\mathcal{P}_{M,N|z}}(\upsilon, \alpha)$, which are obtained from Kernel Ridge Regression, and the relevant derivatives below:

$$\hat{\psi}_{\mathcal{P}_{X|z}}(\alpha) = \sum_{j=1}^{s} \hat{\gamma}_X(z)_j e^{i\alpha x_j} \tag{29}$$

$$\hat{\psi}_{\mathcal{P}_{N|z}}(\alpha) = \sum_{j=1}^{s} \hat{\gamma}_X(z)_j e^{i\alpha n_j} \tag{30}$$

$$\hat{\psi}_{\mathcal{P}_{M,N|z}}(\alpha) = \sum_{j=1}^{s} \hat{\gamma}_{M,N}(z)_j e^{i(\upsilon m_j + \alpha n_j)} \tag{31}$$

*Accepted for the 38th Conference on Uncertainty in Artificial Intelligence* (UAI 2022).

**Algorithm 1** MEKIV training

```
Input: $M_1$, $M_2$, $N_1$, $N_2$, $Z_1$, $Z_2$, $Y_1$, $Y_2$ , kernelType:=RBF kernel

Step 1 Input: $M_1$, $N_1$, $Z_1$, kernelType
    $\hat{\gamma}_N \leftarrow$ KRR($N_1, Z_1$, kernelType) [Singh et al., 2019, Stage 1 estimate]
    Stack $M_1$ and $N_1$ to get $M_1N_1$.
    $\hat{\gamma}_{MN} \leftarrow$ KRR($M_1N_1, Z_1$, kernelType)
    return $\hat{\gamma}_{MN}, \hat{\gamma}_N$

Step 2 Input: kernelType, $\gamma_N^{(s_1)}$, $\gamma_{MN}^{(s_1)}$, $M_1$, $N_1$, $Z_2$, number of $\alpha$ samples := $C$,
    Take $q(\alpha)$ as the Inverse Fourier Transform of the kernel rescaled by $\frac{1}{2\pi}$.
    Take $\{\check{z}_j\}_{j=1}^{s_2}$ to be the set of data points in $Z_2$.
    $\{\hat{x}_j\}_{j=1}^{s_1}, \hat{\lambda}_X \leftarrow$ OptimiseX1($q(\alpha)$, $\gamma_N^{(s_1)}$, $\gamma_{MN}^{(s_1)}$, $M_1$, $N_1$, $\{\check{z}_j\}_{j=1}^{s_2}$, $C$)
    return $\{\hat{x}_j\}_{j=1}^{s_1}, \hat{\lambda}_X$

Step 3 Input: $\{\hat{x}_j\}_{j=1}^{s_1}, \hat{\lambda}_X$, $Y_2$, $Z_2$, $Z_1$
    $\xi \leftarrow$ KIVStage2Validation [Singh et al., 2019, A.5.2]
    $\hat{f} \leftarrow$ KIVStage2($\{\hat{x}_j\}_{j=1}^{s_1}, \hat{\lambda}_X, Y_2, Z_2, Z_1, \xi$)
    return $\hat{f}$
```

**Algorithm 2** Step 2: Learning the CME for $\mathcal{P}_{X|Z}$

```
Input: $q(\alpha)$, $\gamma_N^{(s_1)}$, $\gamma_{MN}^{(s_1)}$, $M_1$, $N_1$, number of $\alpha$ samples := $s_2$, $\{\check{z}_j\}_{j=1}^{s_2}$,
function OptimiseX1
    $\{\alpha_j, \check{z}_j, (w_{MN})_j\}_{j=1}^{(s_2)^2}$ = CreateTrainData($q(\alpha)$, $\hat{\gamma}_N$, $\hat{\gamma}_{MN}$, $M_1$, $N_1$ )
    initialize $\hat{X} = (M_1 + N_1)/2$
    initialize $\hat{\lambda}_X = \hat{\lambda}_N$
    while not converged do
        Use Eq. (19) to calculate $\{(w_X)_j\}_{j=1}^{(s_2)^2}$ from $\{\alpha_j, \check{z}_j\}_{j=1}^{(s_2)^2}$.
        Compute loss = MSE($\{(w_X)_j, (w_{MN})_j\}_{j=1}^{(s_2)^2}$)  ▷ Eq.(25)
        Compute $\nabla_{\hat{X}}(loss)$, $\nabla_{\hat{\lambda}_X}(loss)$
        $\hat{X} \leftarrow \hat{X} -$step$\times \nabla_{\hat{X}}(loss)$; $\hat{\lambda}_X \leftarrow \hat{\lambda}_X -$step$\times \nabla_{\hat{\lambda}_X}(loss)$
    end while
    return $\hat{X}$, $\hat{\lambda}_X$
end function

function CreateTrainData
    Sample $\{\alpha_j\}_{j=1}^{C}$ from $q(\alpha)$
    Take all pairs in $\{\alpha_j\}_{j=1}^{C} \times \{\check{z}_j\}_{j=1}^{s_2}$ to get $\{(\alpha_j, \check{z}_j)\}_{j=1}^{C\times s_2}$
    Substitute $\{(\alpha_j, \check{z}_j)\}_{j=1}^{C\times s_2}$, along with $M_1$, $N_1$, $\hat{\gamma}_N$, $\hat{\gamma}_{MN}$ into Eq. (20) to calculate the labels $\{w_j\}_{j=1}^{C\times s_2}$.
    return $\{\alpha_j, \check{z}_j, w_j\}_{j=1}^{C\times s_2}$
end function
```

Figure 5: Our proposed algorithm. Algorithm 1 outlines the end-to-end algorithm from training data to the structural estimator $\hat{f}$. Algorithm 2 outlines our main contribution, step 2 of the algorithm where we learn the CME for the latent variable $X$.

With:

$$\hat{\gamma}_X(z) = (K_{ZZ} + s\hat{\lambda}_X I)^{-1} K_{Zz} \tag{32}$$

$$\hat{\gamma}_N(z) = (K_{ZZ} + s\hat{\lambda}_N I)^{-1} K_{Zz} \tag{33}$$

$$\hat{\gamma}_{M,N}(z) = (K_{ZZ} + s\hat{\lambda}_{M,N} I)^{-1} K_{Zz} \tag{34}$$

And the derivatives:

$$\frac{\partial}{\partial \alpha} \hat{\psi}_{\mathcal{P}_{X|z}}(\alpha) = \sum_{j=1}^{s} i x_j \hat{\gamma}_X(z)_j e^{i\alpha x_j} \tag{35}$$

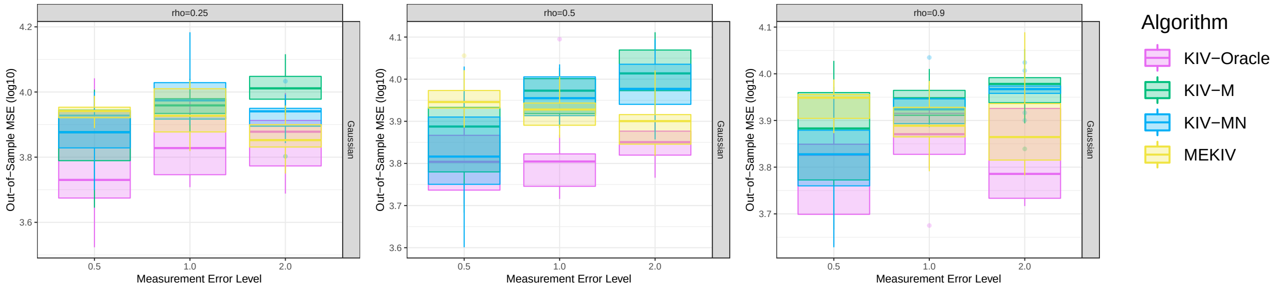
$$\frac{\partial}{\partial v}\bigg|_{v=0} \hat{\psi}_{\mathcal{P}_{M,N|z}}(\alpha, v) = \sum_{j=1}^{s} i m_j \hat{\gamma}_{M,N}(z)_j e^{i\alpha n_j} \tag{36}$$

## 5 DEMAND DESIGN - FURTHER RESULTS

See Figure 6 for further results on Demand design with Gaussian measurement error.

## 6 PROOFS

**Proof of Theorem 1.**



(a) $\rho = 0.25$

(b) $\rho = 0.5$

(c) $\rho = 0.9$

Figure 6: Demand design - Gaussian Measurement Error.

*Proof.* First we note that by Fubini's theorem the Fourier Transform of the (ground truth) conditional mean embedding $\mu_{\mathcal{P}_{X|z}}$ can be computed as:

$$\tilde{\mu}_{\mathcal{P}_{X|z}}(\alpha) = q(\alpha)\psi_{\mathcal{P}_{X|z}}(-\alpha) \tag{37}$$

$$\begin{aligned}
&\|\hat{\mu}^{(s)}_{\mathcal{P}_{X|z}} - \mu_{\mathcal{P}_{X|z}}\|_{\mathcal{H}_{\mathcal{X}}} && (38)\\
&= \int_{-\infty}^{\infty} \frac{\left|\tilde{\hat{\mu}}^{(s)}_{\mathcal{P}_{X|z}}(\alpha) - \tilde{\mu}_{\mathcal{P}_{X|z}}(\alpha)\right|^2}{q(\alpha)} d\alpha && (39)\\
&= \int_{-\infty}^{\infty} q(\alpha)\left|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(-\alpha) - \psi_{\mathcal{P}_{X|z}}(-\alpha)\right|^2 d\alpha && (40)
\end{aligned}$$

Since $k$ is a symmetric kernel i.e. even function, $q(\alpha) = \frac{1}{2\pi}\tilde{k}(\alpha)$ is a real and even measure, so

$$\begin{aligned}
&\int_{-\infty}^{\infty} q(\alpha)\left|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(-\alpha) - \psi_{\mathcal{P}_{X|z}}(-\alpha)\right|^2 d\alpha && (41)\\
&= \int_{-\infty}^{\infty} q(\alpha)\left|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(\alpha) - \psi_{\mathcal{P}_{X|z}}(\alpha)\right|^2 d\alpha && (42)\\
&= \|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(\alpha) - \psi_{\mathcal{P}_{X|z}}(\alpha)\|_{\mathcal{L}^2(\mathbb{R},q)} && (43)
\end{aligned}$$

Consequentially, whenever $\|\hat{\mu}^{(s)}_{\mathcal{P}_{X|z}} - \mu_{\mathcal{P}_{X|z}}\|_{\mathcal{H}_{\mathcal{X}}} < \epsilon$, $\|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}} - \psi_{\mathcal{P}_{X|z}}\|_{\mathcal{L}^2(\mathbb{R},q)} < \epsilon$. and vice versa. Therefore, $\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}} \longrightarrow \psi_{\mathcal{P}_{X|z}}$ in $\mathcal{L}^2(\mathbb{R}, q)$ if and only if $\hat{\mu}^{(s)}_{\mathcal{P}_{X|z}} \longrightarrow \mu_{\mathcal{P}_{X|z}}$ in $\mathcal{H}_{\mathcal{X}}$. Moreover, if they converge, the convergence happen at the same rate. □

**Proof of Theorem 2.**

*Proof.* Since there is a bijection between characteristic functions and probability distributions, we only have to show that the characteristic function satisfying Eq. (17) is unique.

Now Eq. (17) can be rewritten as

$$\begin{aligned}
\frac{d\psi_{\mathcal{P}_{X|\check{z}}}(\alpha)/d\alpha}{\psi_{\mathcal{P}_{X|\check{z}}}(\alpha)} &= ig(\alpha) && (44)\\
\frac{d}{d\alpha}\log\big(\psi_{\mathcal{P}_{X|\check{z}}}(\alpha)\big) &= ig(\alpha) && (45)\\
\text{with } g(\alpha) &= \frac{\mathbb{E}[Me^{i\alpha N}|\check{z}]}{\mathbb{E}[e^{i\alpha N}|\check{z}]} && (46)
\end{aligned}$$

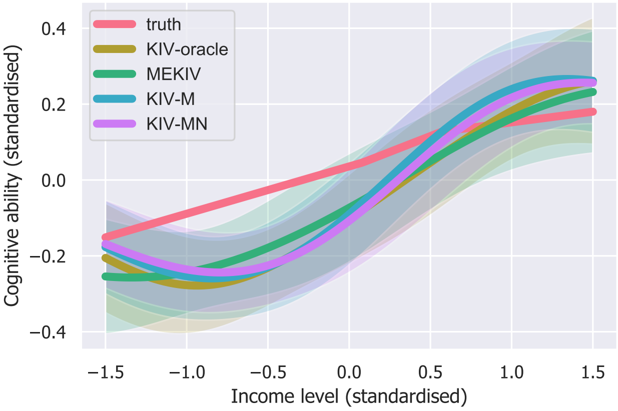


Figure 7: Dahl-Lochner Income on Cognitive outcome

Now suppose there is another characteristic function $\psi(\alpha)$ which also satisfies Eq. (17) for all $\alpha \in \mathbb{R}$, i.e.

$$\frac{d}{d\alpha}\log(\psi(\alpha)) = ig(\alpha) \tag{47}$$

Let $f(\alpha) = \log\big(\psi_{\mathcal{P}_{X|z}}(\alpha)\big)$, $g(\alpha) = \log(\psi(\alpha))$. $f' = g'$. But since characteristic functions are always $1$ at $\alpha = 0$, $f(0) = g(0) = \log(1) = 0$. So by Lemma 1 $f = g$. Since $\log$ is an invertible function whose inverse is $\exp$, we must have $\psi_{\mathcal{P}_{X|z}} = \exp(f) = \exp(g) = \psi$. i.e. the solution to Eq. (17) is unique. □

# 7 REAL-WORLD EXPERIMENT: INCOME ON CHILDREN'S OUTCOME

| Method | MSE |
|---|---|
| KIV-Oracle | 0.0345 ± 0.0190 |
| MEKIV | 0.0295 ± 0.0144 |
| KIV-M | 0.0318 ± 0.0199 |
| KIV-MN | 0.0310 ± 0.0142 |

Table 1: MSE, income-on-children's-outcome data

As described in Section 6 of the main paper, we apply our algorithm to the dataset described in Dahl and Lochner [2012]. In order to obtain causal ground truth, we fit a simulation model to the observed data, obtaining the structural equation $f$. We then generate data from the fitted simulation model, for which we now have access to causal ground truth. We then run MEKIV along with the baselines on the generated dataset. Table 1 present the results. We observe that the performance across all methods do not differ much, and in particular the perturbations around the average MSE overlap. This prompts us to look into the performance of the learnt estimators and we plot the estimated $\mathbb{E}[Y|do(A)]$ in Figure 7. In Figure 7, we observe that in fact none of the methods work well, including KIV-oracle. This suggests that the instrumental variable is only weakly associated with the input. A simple analysis on the dataset suggests exactly this: the average increase in average yearly income from 1985 to 2000 is around \$2000, whereas the largest increase between the EITC credit rate of two consecutive years is about 10%, which corresponds to only a 10% portion of the increase in income.

## References

Gordon B. Dahl and Lance Lochner. The impact of family income on child achievement: Evidence from the earned income tax credit. *American Economic Review*, 102(5):1927–56, May 2012. doi: 10.1257/aer.102.5.1927. URL https://www.aeaweb.org/articles?id=10.1257/aer.102.5.1927.

Yingyao Hu and Yuya Sasaki. Closed-form estimation of nonparametric models with non-classical measurement errors. *Journal of Econometrics*, 185(2):392–408, 2015. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2014.11.004. URL https://www.sciencedirect.com/science/article/pii/S0304407614002796.

A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. Kusner, A. Gretton, and K. Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, pages 7512–7523, 2021.

Susanne M. Schennach. Estimation of nonlinear models with measurement error. *Econometrica*, 72(1):33–75, 2004. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3598849.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.