Causal Inference with Treatment Measurement Error: A Nonparametric Instrumental Variable Approach

Abstract

We propose a kernel-based nonparametric estimator for the causal effect when the cause is corrupted by error. We do so by generalizing estimation in the instrumental variable setting. Despite significant work on regression with measurement error, additionally handling unobserved confounding in the continuous setting is non-trivial: we have not seen any prior work. As a by-product of our investigation, we establish a connection between mean embeddings and characteristic functions, and how learning one simultaneously allows one to learn the other. This opens the way for kernel method research to leverage existing results in characteristic function estimation. Finally, we empirically show that our proposed method, MEKIV, improves over baselines and is robust under changes in the strength of measurement error and to the type of error distributions.

1 INTRODUCTION

Real world data poses many problems for causal effect estimation. Unmeasured confounding, the existence of hidden common causes of a treatment X and an outcome of interest Y, is a problem that lies at the heart of many applied sciences. Solving this problem led to a variety of approaches, the most common based on the idea of instrumental variables (IVs): an auxiliary variable Z independent of Y upon a perfect intervention on X [Pearl, 2009, Hernán and Robins, 2020], which is predictive of X but not caused by it.

A less commonly studied challenge is when *the treatment is not directly observed*. For instance, we may want to learn the effect of taking a drug ($X = 1$) against not taking it ($X = 0$), where we incentivize the patients to take it or not ($Z = 1$ vs $Z = 0$). It is not necessarily the case that $X = Z$, because the patients do it at home instead of a hospital with supervision, and so they may not comply with the incentive. This *non-compliance* problem is compounded with the *measurement error* problem: a self-reported measurement of taking ($M = 1$) or not taking ($M = 0$) the drug does not imply $X = M$, because the patient may be lying or be just forgetful. An instrumental variable approach to estimate an *average treatment effect* (ATE) such as $\mathbb{E}[Y \mid do(X = 1)] - \mathbb{E}[Y \mid do(X = 0)]$ [Pearl, 2009] may fail to give reliable results if our data consists of records of (Z, M, Y), but the assumption $X = M$ does not hold.

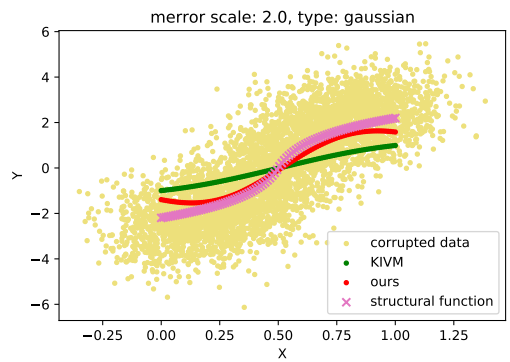


Figure 1: Comparison of curves fitted by our method and KIVM under a corrupted treatment measurement X against with true curve. KIVM is a method we will discuss in the sequel, which ignores that measurements of X are corrupted by additive noise.

A related issue happens when postulating latent *constructs* as causes. In a widespread example by [Bollen, 1989], a model for the effects of "industrialization level" of a country in its political freedom Y is considered. We may operationalize this construct by postulating a space of possible interventions Z on industrialization X that keep the relation between X and Y invariant. However, it remains the case that X is not directly observable but for indirect measurements M, such as the GDP or the proportion of labor force working in industry.

Of relevance, in both classes of indirect treatment measurement problem, is that the causal relation $\mathbb{E}[Y \mid do(x)]$ is

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

considered to be fundamental, with $\mathbb{E}[Y \mid do(m)]$ being either zero, or poorly defined, or of secondary interest (for instance, *redefining* GDP may as well have a genuine causal impact, but this intervention is not the motivation behind understanding the causal impact of industrialization levels). In particular, measurement mechanisms may change more easily than the relation between the putative cause and the outcome of interest (we may redefine GDP, or collect data where the phrasing and timing of our questioning of a patient's compliance varies in different communities, while assuming that the relation between X and Y is invariant). In a way, this measurement problem is a counterpart to why estimating intention-to-treat effects, i.e. $\mathbb{E}[Y \mid do(z)]$, is not in many cases the goal of an IV analysis, despite the policy-making implications.

The need to understand effects of the mismeasured quantities on other quantities of interest motivates the study of measurement error modeling [Carroll et al., 2006, Schennach, 2016, Hernán and Robins, 2020]. Famously, even in the linear (noncausal) regression case, naïvely regressing Y on a noisy measurement of X results in *attenuation error*, which essentially means that the regression coefficient will be underestimated due to the measurement error [Carroll et al., 2006]. An analogous phenomenon will take place when estimating causal effects. Figure 1 depicts a kernel method that attempts to estimate a X-Y dose-response curve, ignoring measurement error in X, compared against the curve found by the method we propose.

The nonlinear and confounded setting is an open domain to be explored. Schennach [2016] suggests that, in general, three measurements are needed to identify the full joint distribution of the measurements and the latent variable. However, in cases where we can make some assumptions on the error distribution, this can be reduced. Furthermore, we are not interested in the full joint distribution with the latent variable X, but only the parts which we need as components of the IV regression model. To that effect, we will assume that our problem follows the Markov properties of Figure 2: we are interested in the structural function $f(x) \equiv \mathbb{E}[Y \mid do(x)]$, where observationally $Y = f(x) + \epsilon$, the error term ϵ being correlated with treatment X. We assume that we have access to at least two treatment measurements, M and N, and an instrumental variable Z.

Our contribution is threefold:

- we propose an estimator for the structural function $f(x)$ without requiring latent variable modeling. The resulting method can be applied without restrictive assumptions in the likelihood, such as the requirement for Gaussian error terms;
- in particular, we provide a method to learn the conditional mean embedding [Muandet et al., 2017] of a latent variable distribution, which can be applied to many two-stage IV settings;

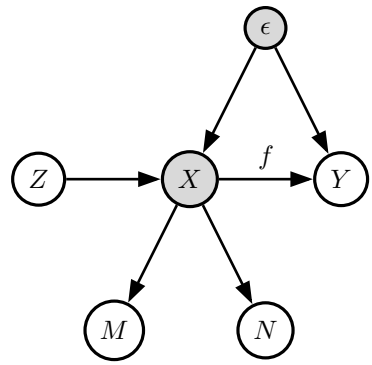


Figure 2: An instrumental variable model with confounded treatment X and Y, where the treatment is unobservable, but has indirect measurements M and N.

- we provide a connection between characteristic function methods and kernel methods for measurement error modelling.

2 RELATED WORK

Measurement bias takes several forms in causal inference. See [Hernán and Robins, 2020, Chapter 9] for a textbook overview, including examples motivating different error structures. For instance, there is a growing literature on causal inference by adjusting for confounders which are only measured via proxies [e.g Kuroki and Pearl, 2014, Battistin and Chesher, 2014, Kallus et al., 2018, Miao et al., 2018, Mastouri et al., 2021, Ghassami et al., 2022]. Because different components of a causal structure contribute differently to the target estimand, it is not surprising though that conditions on identifiability in our setup vary substantively from, e.g., those required by models of confounding proxies. In general, our work is related to the large field of latent variable distribution identification, particularly those involving Markovian assumptions [Allman et al., 2009]. Even closer is the literature on *error-in-variables* regression [Carroll et al., 2006], as it provides several results for identification and estimation in the related problem of regression analysis. Schennach [2016] provides a more up-to-date review, including some comments on its use in causal effect estimation. To the best of our knowledge, no direct connection between methods for continuous causal effect estimation with both unmeasured confounding and measurement error in the treatment is described in the literature. For partially identified models in *discrete* spaces, Finkelstein et al. [2021] provide an approach based on the linear programming formulation of Balke and Pearl [1994]. It should be pointed out that the term "instrumental variable" is sometimes used in the error-in-variables regression literature as the name for a second measurement of the missing regressor X [Carroll et al., 2006, Chapter 6]. Here it signifies a direct cause of X (the more standard definition in causal inference literature). A similar setting was studied in Gultchin et al. [2021], where Z operated as a crude intervention on a complex

cause X and causal effects under new Z interventions could be computed due to invariance. A family of methods for causal inference with corrupted data is presented by Agarwal and Singh [2021], including a linear error-in-variables formulation with riesz representers that can be interpreted as estimating causal effects when no confounding is present. An example of effect estimation with observed confounding adjustment is Song et al. [2015]. Finally, measurement error has also been considered in the causal graph discovery literature [e.g. Silva et al., 2006, Zhang et al., 2018]. In this case, all observed variables are measurements of some underlying latent causes, and the goal is to learn the causal structure among such latent variables. Parametric assumptions are usually necessary, and no causal effect estimates are provided when latent variables are themselves confounded by further hidden common causes.

3 BACKGROUND

Throughout, we use capital letters (e.g. A) to denote a random variable on a measurable space. We denote measurable spaces by calligraphic letters ($\mathcal{A}$), with one exception: $\mathcal{P}$, which we use to denote a probability measure. We use lowercase letters to denote realizations of a random variable ($A = a$). We will use the structural causal model (SCM) formulation of Pearl [2009], where causal relationships are represented as directed acyclic graphs (DAGs). The operator $do(\cdot)$ is defined in these models to describe the process of forcing a random variable to take a particular value, which isolates its effect on downstream variables (i.e. $\mathbb{E}[Y \mid do(A = a)]$ describes the isolated effect of A on Y).

Our goal is to estimate the average treatment effect (ATE) $\mathbb{E}[Y \mid do(X = x)]$ given the graph in Figure 2 (equivalent to the structural function $f : \mathcal{X} \longrightarrow \mathcal{Y}$). Here X, ϵ are unobserved. We only have access to an instrument Z, the effect Y, and corrupted measurements of X: M and N.

3.1 STRUCTURAL ASSUMPTIONS ON $p(x, y \mid z)$

When the treatment X is observed and an instrument Z is available, the structural function is identified by

Assumption 1 $Y = f(X) + \epsilon$ and $\mathbb{E}[\epsilon|Z] = 0$

Assumption 2 $p(x|z)$ is not constant in z.

Under Assumptions 1 and 2, the structural function satisfies the following equation $\mathcal{P}_{\mathcal{Z}}$-almost surely.:

$$\mathbb{E}[Y|Z] = \mathbb{E}[f(X)|Z] = \int f(x) d\mathcal{P}_{\mathcal{X}|\mathcal{Z}} \tag{1}$$

Typical methods fall into two categories: 1) two-stage methods (Singh et al. [2019a], Hartford et al. [2017], Xu et al. [2020]): first identify the conditional distribution directly or through estimating conditional expectations of basis functions; this is followed by identifying f under the identified conditional distribution of vector of conditional expected values of basis functions; 2) moment-based methods (Zhang et al. [2020], Bennett et al. [2019]): estimate f using moment conditions generated by the conditional moment restriction: $\mathbb{E}[(Y - f(X))g(Z)] = 0 \; \forall g$ measurable. A practical difference between two-stage methods and moment-based methods is that two-stage methods require separate data for each stage, and the first stage does not require the outcome observations Y, whereas moment-based methods require data for all variables simultaneously. In this work, we seek to identify the measurement process before identifying the structural function. We thus naturally adopt the two-stage framework since the measurement process requires only the instrument and the measurements, and not the outcome labels.

3.2 REPRODUCING KERNEL HILBERT SPACES

For any space $\mathcal{S} \in \{\mathcal{X}, \mathcal{Y}, \mathcal{M}, \mathcal{N}, \mathcal{Z}\}$, let $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be a positive definite kernel. We denote by ϕ its associated canonical feature map $\phi(x) = k(x, \cdot)$ for any $x \in \mathcal{S}$, and $\mathcal{H}_{\mathcal{S}}$ its corresponding Reproducing Kernel Hilbert Space (RKHS) of real-valued functions on $\mathcal{S}$. The space $\mathcal{H}_{\mathcal{S}}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{S}}}$. It satisfies two important properties: (i) $k(x, \cdot) \in \mathcal{H}_{\mathcal{S}}$ for all $x \in \mathcal{S}$, (ii) the reproducing property: for all $h \in \mathcal{H}_{\mathcal{S}}$ and $x \in \mathcal{S}, h(x) = \langle h, k(x, \cdot) \rangle_{\mathcal{H}_{\mathcal{S}}}$. For any distribution p on $\mathcal{S}$, $\mu_p := \int k(x, \cdot) p(x) dx$ is an element of $\mathcal{H}_{\mathcal{S}}$ and is referred to as the kernel mean embedding of p (Smola et al. [2007]). Similarly, for any conditional distribution $p(x|z)$, $\mu_{\mathcal{P}_{X|z}} := \int k(x, \cdot) p(x|z) dx$ is a *conditional mean embedding* (CME) of $p(x|z)$ (Song et al. [2009, 2013]); see Muandet et al. [2017] for a review.

3.3 STRUCTURE LEARNING USING KERNEL MEAN EMBEDDINGS

Provided that the structural function f lies in the RKHS $\mathcal{H}_{\mathcal{X}}$, then its conditional expectation under $\mathcal{P}_{X|Z}$ can be written as $\mathbb{E}[f(X)|Z] = \langle f, \mu_{\mathcal{P}_{X|Z}} \rangle_{\mathcal{H}_{\mathcal{X}}}$. In Singh et al. [2019b], the conditional mean embedding is estimated by the standard regression formula using the observed samples $\{z_j, x_j\}_{j=1}^{s_1}$ before the structural function f is estimated using a second-stage sample $\{\check{z}_j, \check{y}_j\}_{j=1}^{s_2}$. We present their solution here.

The CME estimator of $\mathcal{P}_{X|z}$ is estimated using the samples $\{z_j, x_j\}_{j=1}^{s_1}$

$$\hat{\mu}^{(s_1)}_{\mathcal{P}_{X|z}} = \Phi_X (K_{ZZ} + s_1 \hat{\lambda} I)^{-1} \Phi'_Z \phi(z) \tag{2}$$

where $\hat{\lambda}$ is the ridge regression hyperparameter chosen using the validation procedure described in [Singh et al., 2019a, App.7.4.2]. K_{ZZ} denote the kernel matrix where $(K_{ZZ})_{jl} = k(z_j, z_l)$, $(\Phi_X)_{(:,j)} = \phi(x_j)$. This is precisely the adaptation of ridge regression to multi-dimensional feature spaces to the case where the number of features can be infinite.

Assumption 3 $f \in \mathcal{H}_{\mathcal{X}}$

Assuming $f \in \mathcal{H}_{\mathcal{X}}$, the estimated CME $\hat{\mu}^{(s_1)}_{\mathcal{P}_{X|z}}$ is used to learn the structural function f by solving the empirical ana-

logue of the following:

$$\mathbb{E}[Y|Z] = \langle f, \mu_{\mathcal{P}_{X|z}} \rangle_{\mathcal{H}_\mathcal{X}} \tag{3}$$

We solve for f via least squares in two stages: 1. Use $\{\check{z}_j, \check{y}_j\}_{j=1}^{s_1}$ to Monte-Carlo estimate $\mathbb{E}[Y|Z]$ and $\mu_{\mathcal{P}_{X|z}}$ (call the latter $\hat{\mu}^{(s_1)}_{\mathcal{P}_{X|z}}$); 2. Use $\{\check{z}_j, \check{y}_j\}_{j=1}^{s_2}$ to estimate f via

$$\hat{f}^{(s_2)}(x) = \hat{\beta}' K_{Xx} \tag{4}$$

$$\hat{\beta} = (VV' + s_2 \hat{\xi} K_{XX})^{-1} V \check{y} \tag{5}$$

$$V = K_{XX}(K_{ZZ} + s_1 \hat{\lambda} I)^{-1} K_{Z\check{Z}} \tag{6}$$

where $\hat{\xi}$ is a hyperparameter. Note that $\hat{\mu}^{(s_1)}_{\mathcal{P}_{X|z}}$ enters in eq. (6): $V_{jl} = \phi(x_j)^T \Phi_X (K_{ZZ} + s_1 \hat{\lambda} I)^{-1} \Phi_Z' \phi(\check{z}_l) = \hat{\mu}^{(s_1)}_{\mathcal{P}_{X|\check{z}_l}}(x_j)$. We refer the readers to Singh et al. [2019b] for the full derivation and for tuning $\hat{\xi}$.

This approach works when we observe treatment X. When X is unobserved, this is not possible. Thus, we propose a method to learn the CME directly from corrupted measurements of X; then, f is yielded as a mapping from the learnt CME to Y as in Eq. (4) to Eq. (6). Our method is detailed in Section 4. We note that the solution of Singh et al. [2019b] requires standard conditions for kernel causal learning, which we inherit. For clarity of presentation, we detail them in Appendix C.

3.4 CHARACTERISTIC FUNCTION IDENTIFICATION OF A LATENT VARIABLE USING MISMEASURED OBSERVATIONS

The main obstacle in the learning of the CME $\mu_{\mathcal{X}|\mathcal{Z}}$ is the lack of observed data of X. To this end, we first review a strongly related problem, which is to identify the characteristic function of $p(x|z)$ using corrupted observations M and N. The following assumptions are needed.

Assumption 4 Measurement errors enter additively:

$$M = X + \Delta M \tag{7}$$

$$N = X + \Delta N \tag{8}$$

Assumption 5 The measurement errors are uncorrelated with each other, ΔN is independent with X, and ϵ is uncorrelated with ΔN:

$$\mathbb{E}[\Delta M | X, \Delta N] = 0 \tag{9}$$

$$X \perp\!\!\!\perp \Delta N \tag{10}$$

$$\mathbb{E}[\epsilon | \Delta N] = 0 \tag{11}$$

As X is unobserved and can be redefined up to any invertible transformation, Eq. (7) is not imposing further constraints besides a monotonic relation between M and X in expectation. Eqs. (9) and (11) are weaker formulations of conditional independence statements $M \perp\!\!\!\perp \{X, \Delta N\}$ and $\epsilon \perp\!\!\!\perp \Delta N$.

Remark 1 Eq. (8) is a restrictive assumption. However, we point out that it can be relaxed and the relaxed setting can be reduced to our setting. Thus we focus on the simplified setting where future methods can extend form; we discuss one way to relax the assumption in Appendix B.

With two measurements, Schennach [2004] provides a constructive estimator for the moments of latent variables. Our work uses a special case of their theorem which we state below.

Assumption 6 $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|\Delta M|] < \infty$

Corollary 1 Given Assumptions 4-6, the characteristic function of X given $Z = z$, i.e. $\psi_{\mathcal{P}_{X|z}}(\alpha)$, is equal to

$$\overbrace{\mathbb{E}_{\mathcal{P}_{X|z}}[e^{i\alpha X}]}^{\psi_{\mathcal{P}_{X|z}}(\alpha):=} = \exp\left(\int_0^\alpha i \frac{\mathbb{E}[M e^{i\nu N}|z]}{\mathbb{E}[e^{i\nu N}|z]} d\nu\right). \tag{12}$$

Proof. Follows directly from [Schennach, 2004, Theorem 1], where the original phrased the equality for the marginal distribution $p(x, m, n)$ □

Since characteristic functions are exact representations of probability distributions, Corollary 1 says that we may model $\mathcal{P}_{X|z}$ through modelling $\mathcal{P}_{M,N|z}$ and $\mathcal{P}_{N|z}$, and the mathematical relation is specified by Eq. (12).

4 METHOD

In this section we will show how to recover $\mathbb{E}[Y \mid do(X = x)]$. To do so, recall that all we need is the kernel mean embedding $\mu_{\mathcal{P}_{X|Z}}$. We begin by demonstrating that estimating $\mu_{\mathcal{P}_{X|Z}}$ boils down to estimating the characteristic function $\psi_{\mathcal{P}_{X|Z}}$. We then introduce a trick for solving systems of equations that we call the *differentiation trick* which allows us to estimate $\psi_{\mathcal{P}_{X|Z}}$. Finally, we give a full procedure for estimating $\mathbb{E}[Y \mid do(X = x)]$ and describe advantages of our approach.

4.1 FROM KERNEL MEAN EMBEDDINGS TO CHARACTERISTIC FUNCTIONS

For simplicity, we limit our description to $\mathbb{R}$. However, all of the following arguments can be extended trivially to $\mathbb{R}^d, d > 1$. First recall the Fourier transform:

$$\tilde{h}(\alpha) = \frac{1}{2\pi} \int_{-\infty}^{\infty} h(x) e^{-i\alpha x} dx$$

and the inverse Fourier transform:

$$h(x) = \int_{-\infty}^{\infty} \tilde{h}(\alpha) e^{i\alpha x} d\alpha.$$

Further, we assume the following.

Assumption 7 (Symmetric, characteristic and translation-invariant kernels) $k(x, \cdot), k(m, \cdot), k(n, \cdot)$ are symmetric, characteristic and translation-invariant kernels.

Kernel symmetry is a standard assumption in ML as kernel functions are generally real. Characteristic kernels allow us to embed probability distributions uniquely in an RKHS. Translation-invariant kernels allow us to consider the probability measure associated with kernel functions.

Under Assumption 7, we can write $k(x,y) = k(x-y)$ and $k(t)$ is positive definite. By Bochner's theorem, we know that k can be written as the Fourier transform of a unique measure $\tilde{k}$:

$$k(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\alpha t}\tilde{k}(\alpha)d\alpha$$
$$\text{i.e. } k(x,y) = \int_{-\infty}^{\infty} e^{-i\alpha(x-y)}q(\alpha)d\alpha$$

where $q(\alpha) := \frac{1}{2\pi}\tilde{k}(\alpha)$. As illustrated in e.g. Fukumizu [2008], we may construct an RKHS on the entire real line using Fourier transforms as feature maps:

$$\mathcal{H}_X = \left\{ f \in \mathcal{L}^2(\mathbb{R}, dx) \middle| \int_{-\infty}^{\infty} \frac{\left|\tilde{f}(\alpha)\right|^2}{q(\alpha)} d\alpha < \infty \right\}$$
$$\langle f, g\rangle_{\mathcal{H}_X} = \int_{-\infty}^{\infty} \frac{\tilde{f}(\alpha)\overline{\tilde{g}(\alpha)}}{q(\alpha)} d\alpha$$

Now consider the Fourier transform of $k(x,\cdot)$, where x is fixed. Since we know that $k(x,y) = \int e^{-i\alpha x}e^{i\alpha y}q(\alpha)d\alpha$, by inspection we realise $\tilde{k}(x,\alpha) = e^{-i\alpha x}q(\alpha)$, recovering the identity that $k(x,y) = \int_{-\infty}^{\infty} \frac{e^{-i\alpha x}q(\alpha)e^{i\alpha y}q(\alpha)}{q(\alpha)} d\alpha = \langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}_X}$.

Recall the definition of the conditional mean embedding of $\mathcal{P}_{X|z}$ for a particular z: $\mu_{\mathcal{P}_{X|z}}(y) := \int k(x,y)p(x|z)dx$. When all variables are observed, the conditional mean embedding (CME) can be estimated by samples $\{x_j, z_j\}_{j=1}^s$:

$$\hat{\mu}^{(s)}_{X|z}(y) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z)k(x_j, y) \tag{13}$$

where

$$\hat{\gamma}^{(s)}_j(z) = (K_{ZZ} + s\hat{\lambda}^{(s)}I)^{-1}K_{Zz} \tag{14}$$

Taking the Fourier transform of $\hat{\mu}^{(s)}_{X|z}(y)$:

$$\begin{aligned}\tilde{\hat{\mu}}^{(s)}_{X|z}(\alpha) &= \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z)e^{-i\alpha x_j}q(\alpha) \\ &= q(\alpha)\underbrace{\sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z)e^{-j\alpha x_j}}_{=:\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(-\alpha)}\end{aligned}$$

Define the $s-$sample estimate of the characteristic function $\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(\alpha) := \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z)e^{i\alpha x_j}$ with $\{x_j\}_{j=1}^s \sim \mathcal{P}_{X|z}$.

Next, we show that $\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}} \longrightarrow \psi_{\mathcal{P}_{X|z}}$ 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 $\mathcal{H}_X$.

Theorem 1 (Convergence in CME is identical to convergence in characteristic function) . Let $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ be a symmetric, positive definite, and translationally invariant characteristic kernel, then for a (conditional) probability measure on $\mathcal{X}$, denoted $\mathcal{P}_{X|z}$, we have that $\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}} \longrightarrow \psi_{\mathcal{P}_{X|z}}$ 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, whenever either converges, the other converges at the same rate.

We provide the proof in Appendix F.

This means learning the characteristic function $\psi_{\mathcal{P}_{X|z}}$ in $\mathcal{L}^2(\mathbb{R}, q)$ simultaneously gives us a precise estimate of the kernel mean embedding $\mu_{\mathcal{P}_{X|z}}$ in $\mathcal{H}_X$.

4.2 LEARNING THE LATENT CHARACTERISTIC FUNCTION

We now show how to learn the latent characteristic function which will give us the latent kernel mean embedding.

Notation. To lighten notation, from now on we will use $\hat{f}$ to denote the empirical estimate of a quantity f, and only use $\hat{f}^{(s)}$ when we need to be specify the sample size s.

What if we are able to observe X? When X is observed, $\hat{\mu}^{(s)}_{\mathcal{P}_{X|z}}$ can be obtained directly and it can be shown that $\hat{\mu}^{(s)}_{\mathcal{P}_{X|z}} \longrightarrow \mu_{\mathcal{P}_{X|z}}$ as $s \longrightarrow \infty$. By Theorem 1, the same samples and λ which closely estimate the CME $\mu_{\mathcal{P}_{X|z}}$ would also closely estimate the characteristic function $\psi_{\mathcal{P}_{X|z}}$, and vice versa [1]. Thus, when s is suitably large, we can accurately approximate the right hand side of (12) as

$$\exp\left(\int_0^{\alpha} i\frac{\mathbb{E}[Me^{i\nu N}|z]}{\mathbb{E}[e^{i\nu N}|z]}\right) \approx \sum_{j=1}^{s} \hat{\gamma}_j(z)e^{i\alpha x_j} \tag{15}$$

where $\hat{\gamma}_j(z)$ is specified by Eq. (14). Recall that this term also depends on $\hat{\lambda}$. To make this explicit we write $\hat{\gamma}^{\hat{\lambda}}_j(z)$.

Solving for X. Given eq. 15 we make the following observation: given samples of $\{z_j\}_j$, the estimate $\hat{\psi}_{\mathcal{P}_{X|z}}$ only depends on $\{x_j\}_j$ and $\hat{\lambda}$.

Therefore, we can solve for $\{x_j\}_j, \hat{\lambda}$ by minimising the discrepancy between both sides of Eq. (15) over $\{x_j\}_j, \hat{\lambda}$:

[1] This should be possible for z from an unseen distribution $\mathcal{P}_{\tilde{Z}}$ provided the unseen distribution has the same support as the training distribution $\mathcal{P}_Z$.

$$\{\hat{x}_j\}_j, \hat{\lambda}_X = \underset{\{x_j\}_j, \hat{\lambda}}{\operatorname{argmin}} \mathbb{E}_{q(\alpha), \mathcal{P}_{\check{Z}}} \left[\left(\sum_{j=1}^{s} \hat{\gamma}_j^{\hat{\lambda}}(\check{Z}) e^{i\alpha \hat{x}_j} - \eta \right)^2 \right] \tag{16}$$

$$\text{with } \eta = \exp \int_0^{\alpha} \left(i \frac{\mathbb{E}[M e^{i\nu N} | \check{Z}]}{\mathbb{E}[e^{i\nu N} | \check{Z}]} d\nu \right)$$

Where the estimation is taken over $q(\alpha)$ and $\mathcal{P}_{\check{Z}}$ because the convergence of characteristic function, had the $X-$samples been observed, is in $\mathcal{L}^2(\mathbb{R}, q)$ and the $\check{Z}$ distribution does not have to equal to the one used to learn the CME, as long as the two have the same support. To estimate η requires two components of approximation: a) finite-sample approximation of $\mathbb{E}[e^{i\nu N} | \check{Z}]$ and $\mathbb{E}[M e^{i\nu N} | \check{Z}]$, b) computation of the integral $\int_0^{\alpha} \left(i \frac{\mathbb{E}[M e^{i\nu N} | \check{Z}]}{\mathbb{E}[e^{i\nu N} | \check{Z}]} d\nu \right)$, given a). While it is possible to use numerical methods such as quadrature to approximate the integral, we propose to save the second component by differentiation.

The Differentiation Trick. We now describe a trick for handling intractable integrals when solving a system of equations. First, let us reproduce eq. (12) below

$$\overbrace{\mathbb{E}_{\mathcal{P}_{X|z}}[e^{i\alpha X}]}^{\psi_{\mathcal{P}_{X|z}}(\alpha) :=} = \exp \left(\int_0^{\alpha} i \frac{\mathbb{E}[M e^{i\nu N} | z]}{\mathbb{E}[e^{i\nu N} | z]} d\nu \right).$$

We can take the natural logarithm and differentiate both sides of eq. (12), and substitute the samples of $\check{\mathcal{Z}}$:

$$\frac{\mathbb{E}[X e^{i\alpha X} | \check{z}]}{\mathbb{E}[e^{i\alpha X} | \check{z}]} = \frac{\mathbb{E}[M e^{i\alpha N} | \check{z}]}{\mathbb{E}[e^{i\alpha N} | \check{z}]} \tag{17}$$

Since differentiation is a many-to-1 operation, we need to verify that the solution to Eq. (17) is also the solution to Eq. (12).

Lemma 1 Considering differentiable functions $\mathbb{C}^n \to \mathbb{C}$. Denote $f'(x) := \frac{d}{dx} f(x)$. Then if $f' = g'$ and $f(a) = g(a) = b, a, b \in \mathbb{C}$, then $f = g$.

Proof. If $f' = g'$, then $f = g + C$ for some $C \in \mathbb{C}$. But $f(a) - g(a) = b - b = 0$, so $C = 0$. □

Theorem 2 The (conditional) distribution of X, denoted by $\mathcal{P}_{X|z}$, which satisfies Eq. (17) is unique, and therefore is the same as the solution to (12).

The proof relies on the fact that characteristic functions are always 1 at $\alpha = 0$ (Appendix F).

When should one use the differentiation trick? When estimation for the target function/parameter requires evaluating an intractible integral, one can think of using the differentation trick. Lemma 1 specifies one condition where this can be done. Note that there are more situations where the differentation trick can be applied, such as when all functions in the target class have the same normalization constant. We summarize three situations where the differentation trick can be applied:

- When the target function class is itself normalized, or fixed at certain input values. Examples of this which may be of interest to machine learning practitioners are: a) probability densities, which always integrates to 1, b) cumulative distributions, which is always 1 at ∞.
- When an invertible transformation of the function class is normalized or fixed at certain inputs. In those cases, one can in principle solve the problem in the normalized function class, and then apply the invertible transformation to go back to original class.
- When the constant can be fixed: having a normalized class of functions essentially stops us from gaining a constant through differentiation. If we have *a priori* knowledge on what the constant is, we can also go ahead and use the differentation trick. This point can be absorbed into the first bullet, but we state it here for clarity.

Towards a sample-based estimator. As discussed, we may replace $\mathbb{E}[e^{i\alpha X} | \check{z}]$ and $\mathbb{E}[e^{i\alpha N} | \check{z}]$ with their finite-sample estimates $\hat{\psi}_{\mathcal{P}_{X|\check{z}}}$ and $\hat{\psi}_{\mathcal{P}_{N|\check{z}}}$. For $\mathbb{E}[X e^{i\alpha X} | \check{z}]$ and $\mathbb{E}[M e^{i\alpha N} | \check{z}]$, we realise that $\mathbb{E}[X e^{i\alpha X} | \check{z}] = \frac{\partial}{\partial \alpha} \mathbb{E}[e^{i\alpha X} | \check{z}]$, and $\mathbb{E}[M e^{i\alpha N} | \check{z}] = \frac{\partial}{\partial \upsilon} \Big|_{\upsilon=0} \mathbb{E}[e^{i(\alpha N + \upsilon M)} | \check{z}]$. Thus, we replace them with $\frac{\partial}{\partial \alpha} \hat{\psi}_{\mathcal{P}_{X|z}}(\alpha)$ and $\frac{\partial}{\partial \upsilon} \Big|_{\upsilon=0} \hat{\psi}_{\mathcal{P}_{M,N|z}}(\alpha, \upsilon)$ respectively. The full expressions of $s-$sample estimates for $\hat{\psi}_{\mathcal{P}_{X|z}}(\alpha)$, $\hat{\psi}_{\mathcal{P}_{N|z}}(\alpha)$, $\hat{\psi}_{\mathcal{P}_{M,N|z}}(\upsilon, \alpha)$ and the relevant derivatives are stated in Appendix D.

Therefore, we arrive at the new objective function:

$$\{\hat{x}_j\}_{j=1}^s, \hat{\lambda}_X = \underset{\{x_j\}_{j=1}^s, \hat{\lambda}_X}{\operatorname{argmin}} \mathbb{E}_{q(\alpha), \mathcal{P}_{\check{Z}}} \left[\left(w_X(\alpha, \check{Z}) - w_{MN}(\alpha, \check{Z}) \right)^2 \right] \tag{18}$$

$$\text{with} \quad w_X(\alpha, \check{Z}) = \frac{\sum_{j=1}^s x_j \hat{\gamma}_X(\check{Z})_j e^{i\alpha x_j}}{\sum_{j=1}^s \hat{\gamma}_X(\check{Z})_j e^{i\alpha x_j}} \tag{19}$$

$$w_{MN}(\alpha, \check{Z}) = \frac{\sum_{j=1}^s m_j \hat{\gamma}_{M,N}(\check{Z})_j e^{i\alpha n_j}}{\sum_{j=1}^s \hat{\gamma}_N(\check{Z})_j e^{i\alpha n_j}} \tag{20}$$

w_{MN} is the sample estimate for the integrand in η from Eq. (16). We can interpret the output values of w_{MN} as the labels for the supervised learning task defined by Eq. (18), the $(\alpha, \check{z})$ as inputs, and the $\{x_j\}$ and $\hat{\lambda}_X$ are the parameters. As soon as we have obtained the optimal $\{\hat{x}_j\}_{j=1}^s$ and $\hat{\lambda}_X$, we can substitute into Eq. (13) and Eq. (14) to obtain the CME estimate $\hat{\mu}_{\mathcal{P}_{X|z}}$.

4.3 ALGORITHM

We propose *MEKIV*: **M**easurement-**E**rror-corrected **K**ernel **I**nstrumental **V**ariable regression. Two independent samples are needed: $\{z_j, m_j, n_j\}_{j=1}^{s_1}$ and $\{\check{z}_j, \check{y}_j\}_{j=1}^{s_2}$.

Thanks to Theorem 1, In step 1 of the MEKIV, we use $\{z_j, m_j, n_j\}_{j=1}^{s_1}$ to compute the sample estimates of the conditional kernel mean embeddings of $\mathcal{P}_{N|z}$ and $\mathcal{P}_{M,N|z}$, which in large sample size is guaranteed to converge to the ground truth Singh et al. [2019b]. By Theorem 1, this also gives us a sample estimate of the characteristic functions which converges in $\mathcal{L}^2$ of their measures induced by their respective kernels.

Step 2 of the MEKIV learns the characteristic function of $\mathcal{P}_{X|Z}$ by optimising for the X samples using the training objective in Eq. (18). Again by Theorem 1, a good estimate of the characteristic function gives us a good estimate of the conditional kernel mean embedding.

In Step 3, MEKIV uses the learnt kernel conditional mean embedding and the second samples $\{\check{z}_j, \check{y}_j\}_{j=1}^{s_2}$ to estimate the structural function f - equivalent to the stage 2 of the KIV (Singh et al. [2019b]).

The pseudocode of our complete algorithm can be found in Algorithm 1 and 2

Step 1. From the first sample $\{z_j, m_j, n_j\}_{j=1}^{s_1}$, learn the conditional mean embedding of $p(m|z)$ and $p(m, n|z)$ using the result stated in Eq. (2), Section 3.3:

$$\hat{\mu}^{(s_1)}_{\mathcal{P}_{N|z}}(\cdot) = \sum_{j=1}^{s_1} (\hat{\gamma}^{(s_1)}_N(z))_j k(n_j, \cdot), \quad \text{with} \tag{21}$$

$$\hat{\gamma}^{(s_1)}_N(z) = (K_{ZZ} + s_1 \hat{\lambda}_N I)^{-1} K_{Zz} \tag{22}$$

Similarly, it can be shown that:

$$\hat{\mu}^{(s_1)}_{\mathcal{P}_{M,N|z}}(\cdot) = \sum_{j=1}^{s_1} (\hat{\gamma}_{M,N}(z))_j k((m_j, n_j), \cdot), \quad \text{with} \tag{23}$$

$$\hat{\gamma}^{(s_1)}_{M,N}(z) = (K_{ZZ} + s_1 \hat{\lambda}_{M,N} I)^{-1} K_{Zz} \tag{24}$$

Remark 2 (23) allows the use of product kernels.

Step 2. After obtaining from Step 1 the quantities: $\hat{\gamma}_N$ and $\hat{\gamma}_{MN}$, Step 2 creates samples $\{\alpha_j\}$, $\{\check{z}_j\}$ and $\{(w_{MN})_j\}$. To this end, Step 2 samples $\{\alpha_j\}_{j=1}^{s_2}$ from $q(\alpha)$, and uses $\{\check{z}_j\}_{j=1}^{s_2}$ unseen in Step 1. In general, $\{\check{z}_j\}_{j=1}^{s_2}$ can be drawn from any distribution $\mathcal{P}_{\check{Z}}$ with the same support as $\mathcal{P}_Z$. To maximize sample usage, we take all pairs in the cross product $\{\alpha_j\}_{j=1}^{s_2} \times \{z_j\}_{j=1}^{s_2}$, giving $(s_2)^2$ pairs: $\{\alpha_j, \check{z}_j\}_{j=1}^{(s_2)^2}$ - here we overload notation $\{\check{z}_j\}$ to be both before and after taking the cross product. We input each pair of $\{\alpha_j, \check{z}_j\}$ into Eq. (20) to generate the labels $\{(w_{MN})_j\}_{j=1}^{(s_2)^2}$. The process of sampling $\{\alpha_j\}$ from $q(\alpha)$ has a close connection with the Random Fourier Features literature (Bach [2017], Sriperumbudur and Szabó [2015], Rahimi and Recht [2007]).

We now seek $\{x_j\}_{j=1}^{s_1}$ and $\hat{\lambda}_X$ in order to minimize the following objective, which is the empirical analogue of Eq. (18):

$$\{\hat{x}_j\}_{j=1}^{s_1}, \hat{\lambda}_X = \underset{\{x_j\}_{j=1}^{n}, \hat{\lambda}_X}{\operatorname{argmin}} \sum_{j=1}^{(s_2)^2} \left[(w_X(\alpha_j, \check{z}_j) - (w_{MN})_j)^2 \right] \tag{25}$$

For clarity, Step 2 is illustrated in Algorithm 2.

Step 3. Given estimates of $\{x_j\}_{j=1}^{s_1}$ and $\hat{\lambda}_X$, we obtain the empirical estimate $\hat{mu}_{\mathcal{P}_{X|z}}$. Along with the samples $\{\check{z}_j, \check{y}_j\}_{j=1}^{s_2}$, we obtain the solution for $\hat{f}^{s_1}$. The prcedure is identical to the Stage 2 estimation of KIV Singh et al. [2019b], for which we stated the derived estimator in Section 3.3. Our solution for f is:

$$\hat{f}^{(s_2)}(x) = (\hat{\beta})' K_{\hat{X}x} \tag{26}$$

$$\text{with } \hat{\beta} = (VV' + s_2 \hat{\xi} K_{\hat{X}\hat{X}})^{-1} V \check{y} \tag{27}$$

$$V = K_{\hat{X}\hat{X}} (K_{ZZ} + s_1 \hat{\lambda} I)^{-1} K_{Z\check{Z}} \tag{28}$$

4.4 ADVANTAGES OF MEKIV

We highlight the benefits of MEKIV:

- MEKIV is **free of distributional assumptions**: as long as the measurement error satisfy the mean-independence conditions in Eq. (9)-(11), the distributions can have any shape.
- **Computational efficiency**: MEKIV models only the CME of $\mathcal{P}_{X|Z}$, and in particular, no modelling of the full joint distribution $\mathcal{P}_{X|Z,M,N}$ as is commonly done in standard latent variable modelling.
- **Ease of implementation**: Unlike standard latent variable modelling, which is typically hard to train due to the large number of hyperparameters, MEKIV is easy to implement and works stably without large efforts in tuning.

5 EXPERIMENTS

In this section we evaluate empirical performance of MerrorKIV across multiple designs and against baselines. In particular, we compare to three baselines: A) KernelIV Singh et al. [2019b] with ground truth X provided from an oracle (KIV-Oracle); B) KernelIV taking M as the true treatment observations (KIV-M); C) since taking the average of independent errors reduces the error variance, we also compare with KernelIV taking $(M + N)/2$ as the true treatment observations (KIV-MN).

We run each estimator on three designs. The *linear* design Chen and Christensen [2017] involves learning the structural function $f(x) = 4x - 2$, where X is unseen and we are only given corrupted measurements of treatment (M, N), continuous instrument Z, and observations of outcome variables

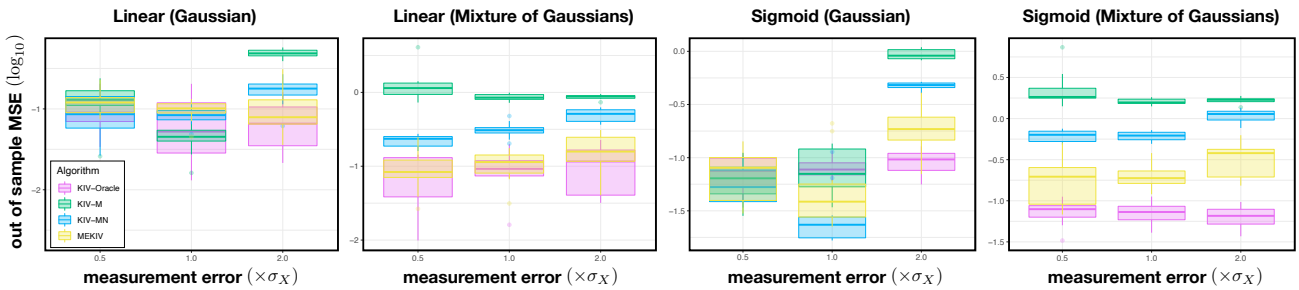


Figure 3: Out of sample MSEs ($\log_{10}$) for linear and sigmoid designs.

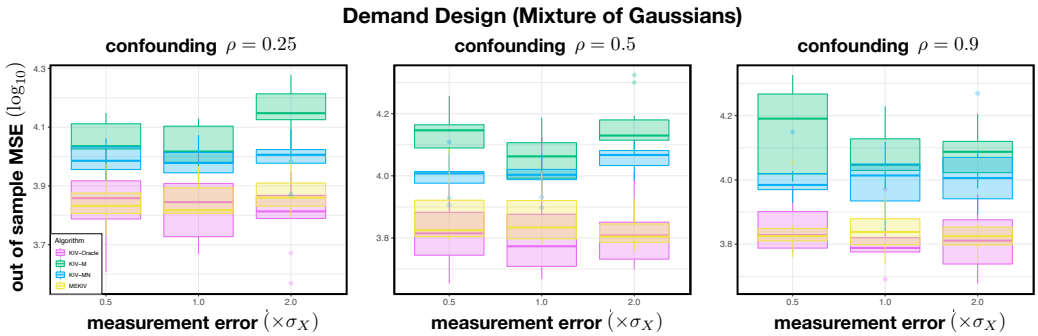


Figure 4: Out of sample MSEs ($\log_{10}$) for demand design.

Y which is confounded with the true treatments X. The *sigmoid* design Chen and Christensen [2017] involves learning the structural $f(x) = \ln(|16x - 8| + 1) \cdot sgn(x - 0.5)$ under the same data generating process otherwise. The *demand* design Hartford et al. [2017] involves learning demand function $h(p, t, s) = 100 + (10 + p) \cdot s \cdot \psi(t) - 2p$ where $\psi(t)$ is a complicated nonlinear function. An data tuple including the ground truth treatments consists of (Y, P, T, S, C) where Y is sales, P is price, T is time of year, S is customer sentiment (a discrete variable), and C parameterizes the supply cost shift. A parameter $\rho \in \{0.25, 0.5, 0.9\}$ calibrates the confounding level of P by supply-side market forces. We set $X := (P, T, S)$ and instruments are $Z := (C, T, S)$.

Since the originally proposed design is one where X is observed, we construct M and N from X and we mask X from all algorithms except KIVoracle. For the demand design where X is 3-dimensional, we mask only the dimension corresponding to P. For each design we construct M and N from adding noise on X. We analyze the robustness of MerrorKIV in two dimensions. First, we vary the measurement error distribution: we implement a *Gaussian* additive noise design and a multi-modal *Mixture of Gaussian* additive noise design where we mix two Gaussian distributions, centred at twice the standard deviation of X away from 0 on either side. Second, for each measurement error distribution, we vary their standard deviation. For both designs, we set the standard deviation of the Gaussian(s) to be $\{0.5, 1., 2.\}$ times the standard deviation of the ground truth X.

For each algorithm, design, measurement error distribution (merror type) and standard deviation (merror level), we implement 10 simulations and calculate MSE with respect to the true structural function f. Figure 3 4 6 plots the results in each design, measurement error distribution type, and measurement error level. We expect KIV-Oracle to be the best across all methods and its performance is viewed as an upper bound for the other algorithms. MerrorKIV beats all other baselines in the highest measurement error level setting and is robust to non-classical measurement error as demonstrated by its performance under Mixture of Gaussian error.

6 CONCLUSION

We propose MEKIV, an instrumental variable approach for confounded structural learning when the treatment variable is measured with error. We clarify a connection between mean embedding learning and characteristic function learning, showing that the two can be done simultaneously. In constructing our algorithm, we introduce the 'derivative trick' which allows target function recovery while avoiding the computation of an intractable integral. Our method performs well on both classical and non-classical measurement error, and is robust over increasing measurement error levels.

References

A. Agarwal and R. Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv:2107.02780*, 2021.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.*, 18(1):714–751, jan 2017. ISSN 1532-4435.

A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds and applications. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI1994)*, pages 46–54, 1994.

Erich Battistin and Andrew Chesher. Treatment effect estimation with covariate measurement error. *Journal of Econometrics*, 178(2):707–715, 2014.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems 32*, pages 3564–3574. Curran Associates, Inc., 2019.

K. Bollen. *Structural Equations with Latent Variables*. Wiley & Sons, 1989.

R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2006. ISBN 9781420010138. URL https://books.google.co.uk/books?id=9kBx5CPZCqkC.

Xiaohong Chen and Timothy M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression, 2017.

N. Finkelstein, R. Adams, S. Saria, and I. Shpitser. Partial identifiability in discrete data with measurement error. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, PMLR 161, page 1798–1808, 2021.

Kenji Fukumizu. Elements of positive definite kernels and reproducing kernel hilbert spaces, October 2008.

A. Ghassami, A. Ying, I. Shpitser, and E Tchetgen Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Limor Gultchin, David Watson, Matt Kusner, and Ricardo Silva. Operationalizing complex causes: A pragmatic view of mediation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3875–3885. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/gultchin21a.html.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1414–1423. PMLR, 2017.

M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Yingyao Hu and Yuya Sasaki. Closed-form estimation of nonparametric models with non-classical measurement errors. *Journal of Econometrics*, 185(2):392–408, 2015. URL https://EconPapers.repec.org/RePEc:eee:econom:v:185:y:2015:i:2:p:392-408.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization. *Advances in neural information processing systems*, 31, 2018.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101 (2):423–437, 03 2014. ISSN 0006-3444. doi: 10.1093/biomet/ast066. URL https://doi.org/10.1093/biomet/ast066.

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.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 08 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy038. URL https://doi.org/10.1093/biomet/asy038.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

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.

Susanne M. Schennach. Recent advances in the measurement error literature. *Annual Review of*

Economics, 8(1):341–377, 2016. doi: 10.1146/annurev-economics-080315-015058. URL https://doi.org/10.1146/annurev-economics-080315-015058.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

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

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

Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer-Verlag, 2007.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

S. Song, S. M. Schennach, and H. White. Estimating nonseparable models with mismeasured endogenous variables. *Quantitative Economics*, 6:749–794, 2015.

Bharath K. Sriperumbudur and Zoltán Szabó. Optimal rates for random fourier features. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1144–1152, Cambridge, MA, USA, 2015. MIT Press.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. 10 2020.

K. Zhang, M. Gong, J. Ramsey, K. Batmanghelich, P. Spirtes, and C. Glymour. Causal discovery with linear non-gaussian models under measurement error: Structural identifiability results. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 2020.

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., 2019a, Stage 1 estimate]
Stack M_1 and N_1 to get $M_1 N_1$.
$\hat{\gamma}_{MN} \leftarrow$ KRR($M_1 N_1$, 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 α 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., 2019a, A.5.2]
$\hat{f} \leftarrow$ KIVStage2($\{\hat{x}_j\}_{j=1}^{s_1}, \hat{\lambda}_X$, Y_2, Z_2, Z_1, ξ)
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 α 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 $\nabla_{\hat{X}}(loss)$, $\nabla_{\hat{\lambda}_X}(loss)$
    $\hat{X} \leftarrow \hat{X} - \text{step} \times \nabla_{\hat{X}}(loss)$; $\hat{\lambda}_X \leftarrow \hat{\lambda}_X - \text{step} \times \nabla_{\hat{\lambda}_X}(loss)$
  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.

A PSEUDOCODE

See Figure 5 for the pseudocode of our method.

B 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.

C FURTHER ASSUMPTIONS ON KERNEL IDENTIFICATION

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

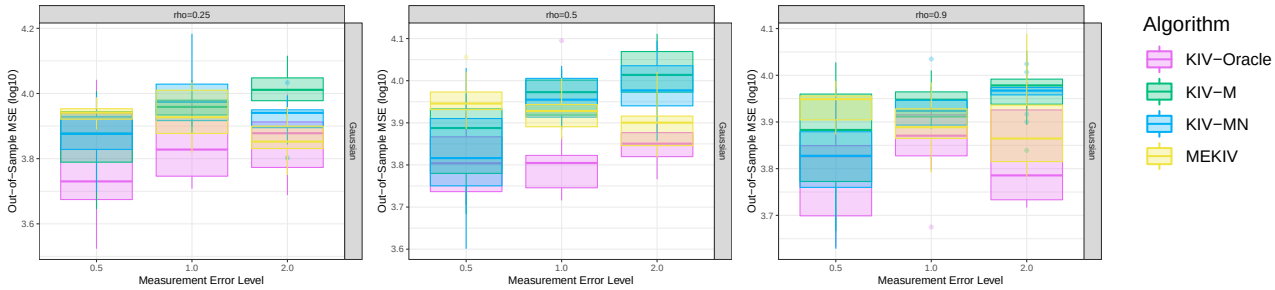
Assumption 8 $\mathcal{Z}, \mathcal{X}, \mathcal{M}, \mathcal{N}, \mathcal{Y}$ are measurable, separable Polish spaces.

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. [2019a] Mastouri et al. [2021]).



(a) $\rho = 0.25$ (b) $\rho = 0.5$ (c) $\rho = 0.9$

Figure 6: Demand design - Gaussian Measurement Error.

D $s-$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:

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

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

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

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:

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

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

E DEMAND DESIGN - FURTHER RESULTS

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

F PROOFS

Proof of Theorem 1.

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:

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

$$\|\hat{\mu}^{(s)}_{\mathcal{P}_{X|z}} - \mu_{\mathcal{P}_{X|z}}\|_{\mathcal{H}_\mathcal{X}} \tag{38}$$

$$= \int_{-\infty}^{\infty} \frac{\left|\tilde{\hat{\mu}}^{(s)}_{\mathcal{P}_{X|z}}(\alpha) - \tilde{\mu}_{\mathcal{P}_{X|z}}(\alpha)\right|^2}{q(\alpha)} d\alpha \tag{39}$$

$$= \int_{-\infty}^{\infty} q(\alpha) \left|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(-\alpha) - \psi_{\mathcal{P}_{X|z}}(-\alpha)\right|^2 d\alpha \tag{40}$$

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

$$\int_{-\infty}^{\infty} q(\alpha) \left|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(-\alpha) - \psi_{\mathcal{P}_{X|z}}(-\alpha)\right|^2 d\alpha \tag{41}$$

$$= \int_{-\infty}^{\infty} q(\alpha) \left|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(\alpha) - \psi_{\mathcal{P}_{X|z}}(\alpha)\right|^2 d\alpha \tag{42}$$

$$= \|\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(\alpha) - \psi_{\mathcal{P}_{X|z}}(\alpha)\|_{\mathcal{L}^2(\mathbb{R},q)} \tag{43}$$

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

$$\frac{d\psi_{\mathcal{P}_{X|\check{z}}}(\alpha)/d\alpha}{\psi_{\mathcal{P}_{X|\check{z}}}(\alpha)} = ig(\alpha) \tag{44}$$

$$\frac{d}{d\alpha}\log\big(\psi_{\mathcal{P}_{X|\check{z}}}(\alpha)\big) = ig(\alpha) \tag{45}$$

$$\text{with } g(\alpha) = \frac{\mathbb{E}[Me^{i\alpha N}|\check{z}]}{\mathbb{E}[e^{i\alpha N}|\check{z}]} \tag{46}$$

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. □