# Synthetic Potential Outcomes and Causal Mixture Identifiability

**Bijan Mazaheri**
Dartmouth Engineering
Hanover, NH
bijan.h.mazaher@dartmouth.edu

**Chandler Squires**
Carnegie Mellon University
Pittsburgh, PA
csquires@mit.edu

**Caroline Uhler**
LIDS, MIT, and Broad Institute
Cambridge, MA
cuhler@mit.ed

## Abstract

Heterogeneous data from multiple populations, sub-groups, or sources can be represented as a "mixture model" with a single latent class influencing all of the observed covariates. Heterogeneity can be resolved at different levels by grouping populations according to different notions of similarity. This paper proposes grouping with respect to the causal response of an intervention or perturbation on the system. This is distinct from previous notions, such as grouping by similar covariate values (e.g., clustering) or similar correlations between covariates (e.g., Gaussian mixture models). To solve the problem, we "synthetically sample" from a counterfactual distribution using higher-order multi-linear moments of the observable data. To understand how these "causal mixtures" fit in with more classical notions, we develop a hierarchy of mixture identifiability.

## 1 INTRODUCTION

### 1.1 Causal Inference

Causality encompasses both counterfactual (what could have been) and hypothetical (what could be) statements. If a patient is given treatment and cured, saying that treatment *caused* recovery is equivalent to saying that the patient would not have recovered without treatment (a counterfactual). Similarly, recommending a treatment carries with it the implication that the patient is better off with treatment than without.

**Latency of Counterfactuals** When hoping to recover causal effects, we are limited to data from a single world, revealing only one "potential outcome" (e.g., the patient recovers under treatment). As such, the counterfactual remains unobserved — for all we know, the treatment had no effect, and the patient would have recovered either way. If a treated patient's recovery is $Y^{(1)}$ and the untreated counterfactual is $Y^{(0)}$, then the individual treatment effect (ITE) is the difference between these "potential outcomes": $Y^{(1)} - Y^{(0)}$. The inherent latency of counterfactual and hypothetical outcomes is referred to as the "fundamental problem of causal inference" (Imbens and Rubin, 2015).

**Exchangeability** In order to estimate causal effects, we must pair examples deemed "exchangeable." For example, two twins given different treatments may be considered to be approximate counterfactuals of each other. When twins are not available, a common approach is to instead pair *populations* to identify an "average treatment effect" (ATE) $\mathbb{E}(Y^{(1)} - Y^{(0)})$. This can be achieved with a randomized control trial (RCT), which randomly assigns "treatment" and "control" groups to ensure exchangeability in expectation.

RCTs are not always feasible — it is unethical to withhold potentially lifesaving medicine from sick patients, and many economic policies must be implemented without a preliminary test. For such settings, methods for causal inference have developed across a number of fields (Pearl, 2009; Imbens and Rubin, 2015; Peters et al., 2017).

**Refining Exchangeability** Approaches to causal inference all rely on the insight that two individuals (or groups) need not be identical in every way to provide access to a counterfactual. Epidemiological methods often make use of assumptions like "unconfoundedness" (also called strong ignorability), which is formalized as $\{Y^{(1)}, Y^{(0)}\} \perp\!\!\!\perp T \mid \mathbf{X}$ for treatment $T$ and covariates $\mathbf{X}$ (Rosenbaum and Rubin, 1983). This assumption implies that no unobserved $U$ confounds both $T$ and $Y$, guaranteeing that two data entries with identical

$\mathbf{X} = \mathbf{x}$ and different $T = t$ are sufficiently exchangeable to be considered counterfactuals.

Relaxing unconfoundedness, graphical methods can be used to model the relationships between covariates (Lauritzen, 1996; Koller and Friedman, 2009). Such models give "adjustment sets" of covariates that are sufficient for exchangeability (Pearl, 2009). These approaches allow for exchangeability only with respect to a subset of $\mathbf{X}$ — which is sometimes necessary to exclude common effects of $T, Y$ (e.g., $T \to X \leftarrow Y$ must not include $X$ in the adjustment set). However, they still cannot handle unobserved confounding due to $U$ unless a valid adjustment set can fully resolve its effect on the ATE.

**Identifiability**    Tools for causal inference are often employed when there is no ground truth validation for success. Therefore, it is essential to guarantee the identifiability of the desired quantity, i.e., that infinite data will generate statistics that correspond to a single answer. Tian and Pearl (2002) gave a general identification criterion for ATEs with respect to graphically modeled unobserved confounding. This approach links causal identifiability to the existence of an observable metric of exchangeability.

## 1.2   Latent Heterogeneity

The principal culprit for entanglement between causation and correlation is latent heterogeneity. In the absence of an observable adjustment set, it is impossible to pair exchangeable points or distributions. For example, if groups with differing rates of recovery are also given treatment at different rates, then the apparent treatment effect will be biased. Treatment is often given to the most severe cases of a disease, resulting in treated patients being associated with poorer outcomes despite a positive treatment response. This is an example of Simpson's Paradox (Simpson, 1951).

**Formal Setting**    To study this setting rigorously, we consider a $k$-mixture to be given by:

- A binary treatment or action $T \in \{0, 1\}$.

- An outcome $Y$ (discrete or continuous).

- A list of covariates, which we will simplify to $\mathbf{X}, \mathbf{Z}$ (discrete or continuous), which can be scalar or vector-valued so long as certain identifiability constraints are met.

- $k$ classes encoded in a latent variable $U \in [k]$.

We will generally assume that $U$ has nontrivial causal arrows to all four $\{T, Y, \mathbf{X}, \mathbf{Z}\}$. These "nontrivial arrows" imply an "overlap" or "positivity" assumption

of $0 < \Pr(T \mid U) < 1$. We restrict ourselves to an "unconfounded" setting where $Y^{(t)} \perp\!\!\!\perp T \mid U$ for simplicity while noting that extending to adjustment sets of the form $\mathbf{C} \cup \{U\}$ for observable $\mathbf{C}$ is straightforward. In addition, we will assume $\mathbf{X}, \mathbf{Z}$ can be transformed into feature maps $\mathbf{\Phi}(\mathbf{X}), \mathbf{\Theta}(\mathbf{Z}) \in \mathbb{R}^k$ with distinct expectations within each $U$. That is $\mathbb{E}[\Phi \mid u] \neq \mathbb{E}[\Phi \mid u']$ whenever $u \neq u'$ for all $\Phi \in \mathbf{\Phi}(\mathbf{X})$ (and similarly for $\mathbf{\Theta}$). To simplify the presentation of our algorithm, we will assume $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^k$ instead of directly referring to their feature maps. The requirement of "distinctness" manifests as the invertibility of a matrix used in our algorithm, which is empirically verifiable.

## 1.3   Mixtures of Treatment Effects

We define the "mixture of treatment effects" (MTEs) problem to be the task of uniquely identifying conditional average treatment effects on a latent class $U$, e.g. $\mathbb{E}(Y^{(1)} - Y^{(0)} \mid u)$, in addition to the probability of each class $\Pr(u)$.

To illustrate the importance of identifying causal heterogeneity, consider two possible populations of bacterial infections. One population is heterogeneous, consisting of a sub-population that is resistant to an antibiotic, as well as a non-resistant sub-population. Another bacterial population is homogeneously partially resistant to the antibiotic, requiring a higher dosage to be killed. In the first setting, use of the antibiotic is ill-advised, as it will allow the resistant subpopulation to dominate. In the second setting, a larger dose is appropriate.

A similar problem emerges with virus vaccination. A vaccine may lose average effectiveness over time due to a loss in antibodies (a homogeneous loss in potency) or the emergence of a new resistant variant (a heterogeneous response). In this setting, the homogeneous response warrants a booster, while the heterogeneous response requires a new vaccine.

A final motivating task is that of evaluating the success of an intervention's implementation. For example, imperfect participation or a faulty shipment may create a heterogeneous response to a treatment, while general ineffectiveness may be homogeneous.

## 1.4   Identifiability Hierarchy

There are multiple notions of identifiability within causally heterogeneous settings. The first is that of heterogeneous treatment effects (HTEs) with respect to a subset of adjustment set $\mathbf{A}' \subseteq \mathbf{A}$, which are only identifiable if $\mathbf{A}'$ is observable. In principle, this means that *each entry* in a dataset can be grouped according to its conditional average treatment effect (CATE) indexed by a set of values for $\mathbf{A}'$. In our setup,

$\mathbf{A}' = \mathbf{A} = \{U\}$, and $U \in \mathbf{A}$ regardless of the causal structure on $(\mathbf{Z}, T, Y, \mathbf{X})$.

In contrast to the *endogenous* and *directly observed* heterogeneity of HTEs, the causal effects within MTEs are *exogenously* and *latently* heterogeneous. The primary computational challenge when recovering MTEs is the latency of $U$, which disallows the direct observation of CATEs with respect to $U$. Mixture models are considered identifiable when observed statistics map uniquely to the parameters of the model, i.e., $\Pr(U), \Pr(T, Y, \mathbf{X}, \mathbf{Z} \mid u)$, up to the $k!$ symmetric permutations of $U$. If a mixture model is identified, MTEs can be recovered with respect to adjustment sets, including unobserved $U$, even though such treatment effects cannot be assigned to individual entries of data.

At an even coarser resolution than MTEs, ATEs are considered identifiable when observed statistics map uniquely to the average causal effects, e.g., $\mathbb{E}(Y^{(1)} - Y^{(0)})$, which marginalizes out the effect of $U$.

These notions of identifiability are hierarchical. Clearly, data points that have been grouped by HTE can be separately analyzed to recover the CATEs needed for MTEs. Furthermore, MTEs can be marginalized into an ATE using the law of total probability. Meanwhile, negative controls (Miao et al., 2024) identify ATEs without giving access to MTEs or HTEs, and mixture models can identify MTEs without giving access to point-specific mixture membership. It is essential to develop methods at the level of the scientific goal — using a method that identifies a "more coarse-grained" level can lead to uncertainty, and using methods that are "more fine-grained" have more restrictive identifiability requirements.

Particular interest has been given to mixture model methods as tools for causal inference because they give access to the joint probability distribution between $U$ and $T, Y$, allowing the identification of both ATE and MTEs. However, two mixture models can easily generate identical ATEs. While less obvious, two mixture models can also generate identical MTEs, motivating an additional level of identifiability.

## 1.5 Contributions

Mixture model identifiability does not precisely correspond to MTE identifiability. That is, two different distributions for $\Pr(\mathbf{Z}, T, Y, X \mid u)$ could map to the same MTEs. In particular, correctly identifying the distribution on the proxies $X, Z$ is inconsequential for recovering the conditional average treatment effects with respect to $U$. For this reason, the conditions for MTE identifiability are *milder* than those for mixture identification. We present a straightforward differentiation between these conditions as a hierarchy on four
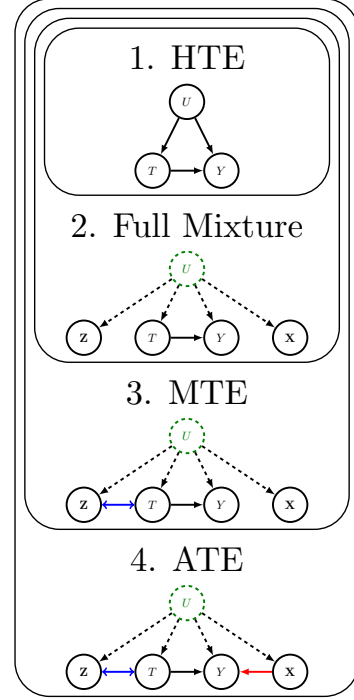


Figure 1: HTEs are only identifiable if $U$ is observed. Identification of the next three levels requires decreasingly restrictive graphical assumptions, demonstrated by the addition of an edge. $Z \leftrightarrow T$ indicates an arrow that could go in either direction (or a bidirected arrow from unobserved confounding).

levels:

1. Identification of HTEs, requiring observed $U$.

2. Identification of the mixture components, requiring $\mathbf{Z} \perp\!\!\!\perp (T, Y) \perp\!\!\!\perp \mathbf{X} \mid U$.

3. Identification of MTEs, requiring $\mathbf{Z} \perp\!\!\!\perp Y \mid T, U$ and $\mathbf{X} \perp\!\!\!\perp (Y, T) \mid U$.

4. Identification of ATEs, requiring $\mathbf{Z} \perp\!\!\!\perp Y \mid T, U$ and $\mathbf{X} \perp\!\!\!\perp T \mid U$.

The four notions of identifiability require decreasingly restrictive assumptions, which can be interpreted graphically as adding possible edges to the 5-vertex graph shown in Figure 1. The full requirements for identifiability at levels 3 and 4 are formally given in Theorems 1 and 2 after introducing some notation.[1]

Identification at levels 2 and 4 are addressed for categorical (discrete) $X, Z$ in previous works discussed

---

[1]The independence conditions needed for identification can be relaxed into a notion of exchangeability in expectation since we only require expectations to factorize. This will be discussed in more detail in Section 6.

in Section 1.6, which we expand to continuous $X, Z$ as a secondary result. Our primary contribution is to give the first identifiability result for level 3, including an algorithm called "synthetic potential outcomes" (SPOs) that we develop in Section 3. Section 4 then gives counterexamples that show distinctness for each level illustrated in Figure 1.

Whenever $U$ is not fully determined by $\mathbf{Z}, \mathbf{X}$, the observed supports of the latent classes overlap.[2] A consequence of this overlap is that identical entries can belong to two different classes. Hence, covariates are insufficient to assign individuals to MTE groups with perfect accuracy. This introduces class-misspecification biases (Loh and Kim, 2022) in addition to invalidating nearest-neighbor approaches (Suk et al., 2021). To handle this challenge, we develop SPOs as a method of moments.

## 1.6 Related Works

**Ladder of Causality** Pearl (2009) also gives a causal hierarchy: counterfactuals, interventions, associations. We further refine the space in between counterfactuals and interventions: HTEs correspond to counterfactuals that are indexed with respect to a known SCM, and ATEs correspond to interventions.

**HTEs** Recent work has studied HTEs for non-overlapping (i.e., qualitatively different) classes (Xie et al., 2012; Wendling et al., 2018; Künzel et al., 2019; Imai and Strauss, 2011). Popular approaches include causal forests (Wager and Athey, 2018) and two-step algorithms generalized by Nie and Wager (2021).

**MTEs** MTEs are latent classes defined by heterogeneous treatment effects that overlap in their observed covariates. Early attempts involve learning distinct parameters per latent class (Lyu et al., 2023; Kim et al., 2015; Kim and Steiner, 2015; Suk et al., 2021). All of these approaches make use of the unconfoundedness assumption, meaning they do not simultaneously deconfound the treatment effects that they recover (i.e., these approaches allow the latent class to affect outcome but not treatment). Our approach combines the task of recovering *and* deconfounding heterogeneous treatment effects.

**Proximal Causal Inference** Methods for proximal causal inference deconfound ATEs (but not MTEs) using "negative controls" that may depend on the unobserved confounder but do not influence (at least one of) treatment or outcome (Tchetgen et al., 2020). The

majority of work has focused on a continuous unobserved confounder (Miao et al., 2024; Tchetgen et al., 2020; Mastouri et al., 2021).

Proximal causal inference and negative controls have demonstrated effective application of the DAG setup from Figure 1 level 4, where $Z$ is a negative control effect and $X$ is a negative control outcome. See Shi et al. (2020) for a review as well as Kendrick Qijun Li and Tchetgen (2024) for a recent application to COVID-19 effectiveness. Our method of SPOs can be thought of as a generalization of Miao et al. (2018) to joint probabilities instead of conditionals, thereby allowing for continuous covariates. By providing insight into how negative controls synthetically copy potential outcomes, we also give a natural extension to MTEs.

**Mixtures** Latent variable methods have been employed to uncover the joint probability distribution between the latent class $U$, treatment $T$, and outcome $Y$. This approach was popularized by (Wang and Blei, 2019), though criticized by Ogburn et al. (2019, 2020) for not guaranteeing identifiability.

The identifiability of discrete mixture models has been studied extensively. In contrast to clustering approaches, this line of work uses the method of moments to handle overlapping probability densities. The primary focus has been on "mixtures of products," in which the observable variables are conditionally independent from each other when the latent $U$ is held constant. This problem can be thought of as a decomposition into rank 1 tensors, giving rise to a number of tensor methods spurred by the seminal work of Allman et al. (2009) and later followed up in Anandkumar et al. (2014). Outside of tensor methods, the current best[3] algorithm with provable guarantees is developed in Gordon et al. (2024). Gordon et al. (2023) recently generalized this setting to arbitrary DAGs.

**Tensor and Matrix Completion** SPOs are similar in concept to synthetic controls (Abadie et al., 2010; Abadie, 2021), synthetic interventions (Agarwal et al., 2020; Squires et al., 2022), and synthetic nearest-neighbors (Agarwal et al., 2023). However, these approaches use linear combinations of data *entries*, whereas we use linear combinations of features. We also use higher-order moments and do not attempt to recover any unit-specific information.

## 2 PRELIMINARIES

**Random Variables versus Assignments** We will use the capital Roman alphabet to denote random variables and boldface to denote sets of these random

---

[2]Formally this is described as positivity, i.e., $\Pr(u \mid \mathbf{z}, \mathbf{x}) > 0$.

[3]in sample and time complexity

variables. Corresponding lowercase letters denote assignments to those random variables. To shorten notation, $X = x$ is sometimes just written as $x$, e.g., $\Pr(y \mid x)$.

**Vectors of Probabilities and Expectations**
Whenever a quantity is left ambiguous in an expectation or probability, i.e., a boldface $\mathbf{X}, \mathbf{Z}$ or a capital $U$, we will first expand it into a column vector and then a matrix if a second ambiguity is given. That is,

$$\mathbf{E}[\mathbf{X}] \coloneqq \begin{pmatrix} \mathbb{E}(X_1) & \dots & \mathbb{E}(X_{|\mathbf{X}|}) \end{pmatrix}^\top, \tag{1}$$

$$\mathbf{P}[U] \coloneqq \begin{pmatrix} \Pr(U = 1) & \dots & \Pr(U = k) \end{pmatrix}^\top, \tag{2}$$

$$\mathbf{E}[X_i \mid U] \coloneqq \begin{pmatrix} \mathbb{E}(X_i \mid U = 1) & \dots & \mathbb{E}(X_i \mid U = k) \end{pmatrix}^\top. \tag{3}$$

$\mathbf{E}[\mathbf{X}Y]$ is a column vector of second-order multilinear moments, while $\mathbf{E}[\mathbf{X}, \mathbf{Z}] \in \mathbb{R}^{|\mathbf{X}| \times |\mathbf{Z}|}$ expands both sets of variables as a matrix:

$$\mathbf{E}[\mathbf{Z}, \mathbf{X}]_{ij} \coloneqq \mathbb{E}(Z_i X_j). \tag{4}$$

We will also give vectors of conditional expectations,

$$\mathbf{E}[\mathbf{X} \mid t] \coloneqq \begin{pmatrix} \mathbb{E}(X_1 \mid t) & \dots & \mathbb{E}(X_{|\mathbf{X}|} \mid t) \end{pmatrix}^\top. \tag{5}$$

$\mathbf{E}[\mathbf{X} \mid U]$ is a matrix with columns of $\mathbf{E}[\mathbf{X} \mid u]$, i.e.,

$$\mathbf{E}[\mathbf{X} \mid U]_{ij} \coloneqq \mathbb{E}(X_i \mid U = j). \tag{6}$$

**Observable Moments** Not all of our $\mathbf{E}[\cdot]$ vectors and matrices can be estimated directly by observable statistics. More specifically, we cannot estimate any vector or matrix of probabilities that involves $U$. To emphasize which matrices are observable, we will replace $\mathbf{E}$ with $\mathbf{M}$ for "observable moments." We will similarly replace scalar $\mathbb{E}(\cdot)$ with $M(\cdot)$.

$\mathbf{M}[\cdot]$ vectors can be decomposed as inner products between $\mathbf{E}[\cdot \mid U]$ vectors and $\mathbf{P}[U]$. That is,

$$\mathbf{M}[\mathbf{X}] = \mathbf{E}[\mathbf{X} \mid U]\mathbf{P}[U]. \tag{7}$$

When $\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \mid U$, we have a similar decomposition for second-order moment matrices:

$$\mathbf{M}[\mathbf{X}, \mathbf{Z}] = \mathbf{E}[\mathbf{X} \mid U]\,\mathbf{diag}(\mathbf{P}[U])\mathbf{E}[\mathbf{Z} \mid U]^\top. \tag{8}$$

Algorithmic computations will be restricted to using $\mathbf{M}[\cdot]$s. To prove correctness and analyze stability, we will decompose into $\mathbf{E}[\cdot]$ and $\mathbf{P}[\cdot]$, which exist hypothetically but cannot be accessed directly from observed statistics.

**Potential Outcomes** To understand the difference between *intervening* and *conditioning* on $T$, compare the following quantities:

$$\mathbb{E}(Y \mid t) = \mathbf{E}[Y \mid U, t]^\top \mathbf{P}[U \mid t], \tag{9}$$

$$\mathbb{E}(Y^{(t)}) = \mathbf{E}[Y \mid U, t]^\top \mathbf{P}[U]. \tag{10}$$

The crucial difference is that intervening on $T = t$ does change the right column vector. $\mathbb{E}(Y \mid t)$ is an observable moment that is estimated by counting the instances of $Y = 1$ when conditioning on $T = t$. $\mathbb{E}(Y^{(t)})$, in contrast, is not given by an observable moment because we cannot sample from the marginal $U$ while simultaneously drawing from the conditional $Y \mid U, t$. This is due to confounding, which shifts the distribution of $U$ between the treatment ($T = 1$) and control ($T = 0$) groups.

## 3 SPOs

For ATEs, our goal is to recover $\mathbb{E}(Y^{(1)} - Y^{(0)})$. We will use $R \coloneqq Y^{(1)} - Y^{(0)}$ for the "response." For MTEs, we must recover $k$ CATEs, i.e. $\mathbf{E}[R \mid U]$, as well as $\mathbf{P}[U]$. We present SPOs as a method for identifying ATEs and MTEs.

**Theorem 1.** *The ATE given by $\mathbb{E}(R)$, is identifiable by SPOs if*

  *(i)* $\mathbf{Z} \perp\!\!\!\perp Y \mid T, U$,

  *(ii)* $\mathbf{X} \perp\!\!\!\perp T \mid U$,

  *(iii)* $\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid t]$ *is full rank for $t \in 0, 1$.*

**Theorem 2.** *The MTE given by $\mathbf{E}[R \mid U]$ and $\mathbf{P}[U]$ is identifiable by SPOs if*

  *(i)* $\mathbf{Z} \perp\!\!\!\perp Y \mid T, U$,

  *(ii)* $\mathbf{X} \perp\!\!\!\perp (Y, T) \mid U$,

  *(iii)* $\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid t]$ *is full rank for $t \in 0, 1$.*

For both theorems, conditions (i) and (ii) correspond to the graphs given in Figure 1. Notice that $\mathbf{Z} \perp\!\!\!\perp \mathbf{X} \mid U$ is implied in both cases if no other variables are present (i.e., no other active paths).

### 3.1 Main Insight

If we could sample from $Y^{(t)}$ *without* conditioning on $T$, we could access its expected value as an observable moment. Notice that if $\mathbf{X} \perp\!\!\!\perp T \mid U$, then $\mathbb{E}[\mathbf{X} \mid U, t] = \mathbb{E}[\mathbf{X} \mid U]$. Our approach will rely on the observation that $\mathbf{E}[Y^{(t)} \mid U]$ is a $k$-dimensional vector that can be expressed as a linear combination of the rows of

$\mathbf{E}[\mathbf{X} \mid U, t]$ provided that the matrix is full-rank, which will be guaranteed by $\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid t]$ being full rank due to its decomposition in Equation 8. Let this linear combination be given by coefficients $\boldsymbol{\alpha}$,

$$\mathbf{E}[Y^{(t)} \mid U] = \mathbf{E}[\mathbf{X} \mid U, t]^\top \boldsymbol{\alpha}. \qquad (11)$$

If $\mathbf{X} \perp\!\!\!\perp T \mid U$, the same $\boldsymbol{\alpha}$ can be applied to the unconditioned moments of $\mathbf{X}$ to access a "synthetic potential outcome":

$$\mathbf{M}[\mathbf{X}]^\top \boldsymbol{\alpha} = \mathbf{P}[U]^\top \mathbf{E}[\mathbf{X} \mid U]^\top \boldsymbol{\alpha} = \mathbb{E}(Y^{(t)}). \qquad (12)$$

In the next section, we show how to find $\boldsymbol{\alpha}$.

### 3.2 Finding SPOs

We want to find $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^\top$ such that, using unconfoundedness,

$$\mathbf{E}[Y^{(t)} \mid U]^\top = \mathbf{E}[Y \mid U, t]^\top = \mathbf{E}[\mathbf{X} \mid U, t]^\top \boldsymbol{\alpha}. \quad (13)$$

We begin by recalling that,

$$\begin{aligned} \mathbf{M}[Y \mid t] &= \mathbf{E}[Y \mid U, t]^\top \mathbf{P}[U \mid t] \\ &= \mathbf{E}[Y^{(t)} \mid U]^\top \mathbf{P}[U \mid t]. \end{aligned} \qquad (14)$$

Notice that the left row vector in this inner product is precisely what we want to "copy." To compute $\boldsymbol{\alpha}$, we can set up a system of linear equations that matches second-order moments with $\mathbf{Z}$:

$$\begin{aligned} \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid t]\boldsymbol{\alpha} &= \mathbf{M}[\mathbf{Z}Y \mid t] \\ \boldsymbol{\alpha} &= \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid t]^{-1}\mathbf{M}[\mathbf{Z}Y \mid t]. \end{aligned} \qquad (15)$$

To see how this moment-matching gives rise to a system of linear equations, apply the expansion given by Equation 8 to $\mathbf{M}[\mathbf{Z}, \mathbf{X}]$ on both sides (requires $\mathbf{Z} \perp\!\!\!\perp \mathbf{X} \mid U$), and expanding $\mathbf{M}[\mathbf{Z}Y \mid t]$ similarly using condition (i) from both theorems.

### 3.3 ATE Recovery

The average treatment effect is given by the difference between the moments of the potential outcomes,

$$\text{ATE} = \mathbb{E}(R) = E(Y^{(1)}) - E(Y^{(0)}). \qquad (16)$$

Both moments are accessible using SPO coefficients

$$\begin{aligned} \boldsymbol{\alpha}^{(1)} &:= \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 1]^{-1}\mathbf{M}[\mathbf{Z}, Y \mid T = 1], \\ \boldsymbol{\beta}^{(1)} &:= \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 0]^{-1}\mathbf{M}[\mathbf{Z}, Y \mid T = 0]. \end{aligned} \qquad (17)$$

The superscript 1 is used to signify that the coefficients are used to compute a first order-moment, as we will be generalizing this procedure later. We use $\boldsymbol{\gamma}^{(1)} = \boldsymbol{\alpha}^{(1)} - \boldsymbol{\beta}^{(1)}$ to find the ATE,

$$\mathbb{E}(R) = \mathbf{M}[\mathbf{X}]^\top \boldsymbol{\alpha}^{(1)} - \mathbf{M}[\mathbf{X}]^\top \boldsymbol{\beta}^{(1)} = \mathbf{M}[\mathbf{X}]^\top \boldsymbol{\gamma}^{(1)}. \quad (18)$$

### 3.4 MTE Recovery

We will denote the $\ell$th order moment as

$$\mathbb{E}(R^{\odot \ell}) = \mathbf{P}[U]^\top \mathbf{E}[R \mid U]^{\odot \ell} \qquad (19)$$

with $\odot \ell$ indicating an element-wise exponent. Notice that this quantity is different from $\mathbb{E}(R^\ell)$.

Information about mixed treatment effects is contained in these higher-order moments of $R$. For example, a homogeneous treatment effect of 0 will give $\mathbb{E}(R^{\odot 2}) = \mathbb{E}(R) = 0$, whereas a mixed treatment effect of half $+1/2$ and half $-1/2$ will give $\mathbb{E}(R^{\odot 2}) = 1/4$, despite an identical $\mathbb{E}(R) = 0$.

Using $\odot$ to denote element-wise multiplication, we slightly abuse notation to define

$$\mathbb{E}(Y^{(t)} R^{\odot \ell - 1}) := \Big(\mathbf{E}[Y^{(t)} \mid U] \odot \mathbf{E}[R^{\odot \ell - 1} \mid U]\Big)^\top \mathbf{P}[U]. \qquad (20)$$

We can then expand the $\ell$th order moment as

$$\mathbb{E}(R^{\odot \ell}) = \mathbb{E}(Y^{(1)} R^{\odot \ell - 1}) - \mathbb{E}(Y^{(0)} R^{\odot \ell - 1}). \qquad (21)$$

We will now estimate $\mathbb{E}(Y^{(t)} R^{\odot \ell - 1})$ the same way we estimated $\mathbb{E}(Y^{(t)})$: using coefficients $\boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}$ for a linear combination of moments with $\mathbf{X}$.

$$\begin{aligned} \mathbf{E}[Y^{(1)} \mid U] \odot \mathbf{E}[R^{\odot \ell - 1} \mid U] &= \mathbf{E}[\mathbf{X} \mid U, t]^\top \boldsymbol{\alpha}^{(\ell)}, \\ \mathbf{E}[Y^{(0)} \mid U] \odot \mathbf{E}[R^{\odot \ell - 1} \mid U] &= \mathbf{E}[\mathbf{X} \mid U, t]^\top \boldsymbol{\beta}^{(\ell)}. \end{aligned} \qquad (22)$$

An additional requirement of $Y \perp\!\!\!\perp \mathbf{X} \mid U$ means that

$$\mathbb{E}[Y X_i \mid U] = \mathbb{E}[Y \mid U] \odot \mathbb{E}[X_i \mid U] \ \forall X_i. \qquad (23)$$

We will compute $\boldsymbol{\gamma}^{(\ell)}$ with $\boldsymbol{\gamma}^{(\ell-1)}$ already computed (with the base case $\boldsymbol{\gamma}^{(1)}$ already covered for ATE recovery). We now rewrite Equation 22 using $\boldsymbol{\gamma}^{(\ell-1)}$ and second-order moments between $X_i$ and $Y$:

$$\begin{aligned} \mathbf{E}[\mathbf{X} \mid U, t]^\top \boldsymbol{\alpha}^{(\ell)} &= \mathbf{E}[Y^{(t)} \mid U] \odot \sum_i \boldsymbol{\gamma}_i^{(\ell-1)} \mathbf{E}[X_i \mid U, t] \\ &= \sum_i \boldsymbol{\gamma}_i^{(\ell-1)} \mathbf{E}[Y X_i \mid U, t]. \end{aligned} \qquad (24)$$

We again moment-match with $\mathbf{Z}$ to get

$$\begin{aligned} \boldsymbol{\alpha}^{(\ell)} &= \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 1]^{-1}\mathbf{M}[\mathbf{Z}, \mathbf{X}Y \mid T = 1]\boldsymbol{\gamma}^{(\ell-1)}, \\ \boldsymbol{\beta}^{(\ell)} &= \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 0]^{-1}\mathbf{M}[\mathbf{Z}, \mathbf{X}Y \mid T = 0]\boldsymbol{\gamma}^{(\ell-1)}. \end{aligned} \qquad (25)$$

Once we have computed $\boldsymbol{\gamma}^{(\ell)} = \boldsymbol{\alpha}^{(\ell)} - \boldsymbol{\beta}^{(\ell)}$, we can estimate the $\ell$th moment of the treatment effect,

$$\mathbb{E}(R^{\odot \ell}) = \mathbf{M}[\mathbf{X}]^\top \boldsymbol{\gamma}^{(\ell)}. \qquad (26)$$

Using these higher-order moments to recover a multiplicity of treatment effects reduces to the well-studied

"sparse Hausdorff moment problem," which seeks to recover $\mathbb{E}[R \mid U]$ and $\mathbf{P}[U]$. Notice that this new mixture model on $R$ only contains the parameters of interest and is, therefore, no longer beholden to the correct identification of $\Pr(\mathbf{X}, \mathbf{Z} \mid U)$.

**Theorem 3** (Rabani et al. (2014)). $\mathbf{P}[U]$ *and distinct[4]* $\mathbf{E}[R \mid U]$ *can be uniquely identified using* $\mathbb{E}(R), \mathbb{E}(R^{\odot 2}), \ldots, \mathbb{E}(R^{\odot 2k-1})$.

Identifiability of a mixture of $k$ treatment effects requires $2k - 1$ moments and can be calculated using Prony's method (de Prony, 1795), as illustrated and analyzed in Gordon et al. (2020), or the matrix pencil method (Hua and Sarkar, 1990), which is outlined in Section 2.2 of Kim et al. (2019). The full algorithm pseudocode is given in Appendix A. Runtime and sample complexity analysis are detailed in Appendix B.

## 4 THE HIERARCHY OF IDENTIFIABILITY

We now give an in-depth discussion on the identifiability of each level of our hierarchy and why these levels are distinct. Having just discussed the identifiability of levels 3 and 4, we will briefly explain the identifiability of levels 1 and 2 with slight modifications of previous results. By nature of the hierarchy, a solution for a "harder" level can be used to obtain a solution for an "easier" level. We will also show the distinctness of each of these levels by giving two different models that resolve as two different "answers" for harder levels, but the same "answer" for easier levels.

To simplify things, all of our models will use Bernoulli $U, Z, X, T$, and $Y$, with $\mathbf{P}[U] = [1/2, 1/2]^\top$, and

$$\mathbf{P}[Z \mid U] = \mathbf{P}[X \mid U] = \begin{pmatrix} 1/4 \\ 3/4 \end{pmatrix}. \qquad (27)$$

The features $\mathbf{X}$ are therefore $X_1 = \mathbb{1}(X = 0)$ and $X_2 = \mathbb{1}(X = 1)$ and likewise for $\mathbf{Z}$. The resulting matrix $\mathbf{M}[\mathbf{Z}, \mathbf{X}]$ is full rank as required by Theorem 2.

**Level 1: HTEs** The identifiability of a CATE with respect to an observed adjustment set is perhaps the most well-studied of the levels. In our setup, we have

$$\mathbb{E}(Y^{(t)} \mid u) = \mathbb{E}(Y \mid t, u). \qquad (28)$$

This allows us to compute the CATE as $\mathbb{E}(Y^{(1)} \mid u, \mathbf{x}, \mathbf{z}) - \mathbb{E}(Y^{(0)} \mid u, \mathbf{x}, \mathbf{z})$, which can be assigned to all data points with $U = u, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}$ when $U$ is observed.

---

[4]i.e. $i \neq j \Rightarrow \mathbb{E}(R \mid u_i) \neq \mathbb{E}(R \mid u_j)$

**Level 2: Mixtures of Products** When $U$ is not observed, CATEs cannot be assigned to individual points. For example, if $(\mathbf{Z}, T, Y, \mathbf{X})$ are all discrete, then a strictly positive distribution will give nonzero probabilities for $\Pr(\mathbf{z}, t, y, \mathbf{x} \mid U = 1)$ and $\Pr(\mathbf{z}, t, y, \mathbf{x} \mid U = 0)$. If the $U = 0$ and $U = 1$ components exhibit different causal effects, then we can arbitrarily assign all $n$ points with covariates $(\mathbf{z}, \mathbf{x})$ to classes $U = 0$ and $U = 1$ so long as we obey the correct ratio of $\Pr(U = 0 \mid \mathbf{z}, \mathbf{x}) : \Pr(U = 1 \mid \mathbf{z}, \mathbf{x})$ in our labeling. Hence, any mixture component overlap makes HTE identifiability at the individual level impossible.

Despite this, under some circumstances, the conditional probability distributions $\Pr(\mathbf{Z}, T, Y, \mathbf{X} \mid U)$ and $\Pr(U)$ are still identifiable. One such circumstance is when $\mathbf{Z}$, $\mathbf{S}$, and $\mathbf{X}$ are conditionally independent given $U$, where $\mathbf{S} = (T, Y)$. Allman et al. (2009) showed this case to be identifiable, up to symmetries in permuting the labels of $U$, under certain algebraic assumptions on the conditional distribution matrices $\mathbf{P}[\mathbf{Z} \mid U]$, $\mathbf{P}[\mathbf{S} \mid U]$, and $\mathbf{P}[\mathbf{X} \mid U]$. As we show in Appendix D, this result can be extended to include the continuous setting by using moments instead of probabilities. After recovering the within-component conditional probabilities, $\Pr(\mathbf{Z}, T, Y, \mathbf{X} \mid U)$, we can also recover $\mathbb{E}(Y^{(t)} \mid u) = \mathbb{E}(Y \mid t, u)$.

**Level 3: MTEs** SPOs identify MTEs under the conditions given in Theorem 2. We give two models with invertible moments that identify the same MTE but represent different mixture models that are not distinguishable at level 2. In particular, we define two parameterized families:

First, for $\mu_{zt} \in [0, 1]$, we define

$$\mathbf{P}_{\mu_{zt}}[T = 1 \mid Z, U] = \frac{1}{4} \begin{pmatrix} 3 & 1 \\ 3 & 1 \end{pmatrix} + \frac{\mu_{zt}}{4} \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}. \quad (29)$$

Second, for $\mu_{xy} \in [0, 1]$, we define

$$\mathbf{P}_{\mu_{xy}}[Y^{(0)} = 1 \mid X, U] = \frac{1}{8} \begin{pmatrix} 7 & 1 \\ 7 & 1 \end{pmatrix} + \frac{\mu_{xy}}{8} \begin{pmatrix} 0 & 6 \\ -6 & 0 \end{pmatrix},$$
$$\mathbf{P}_{\mu_{xy}}[Y^{(1)} = 1 \mid X, U] = \frac{1}{8} \begin{pmatrix} 1 & 7 \\ 1 & 7 \end{pmatrix} + \frac{\mu_{xy}}{8} \begin{pmatrix} 0 & -6 \\ 6 & 0 \end{pmatrix}. \quad (30)$$
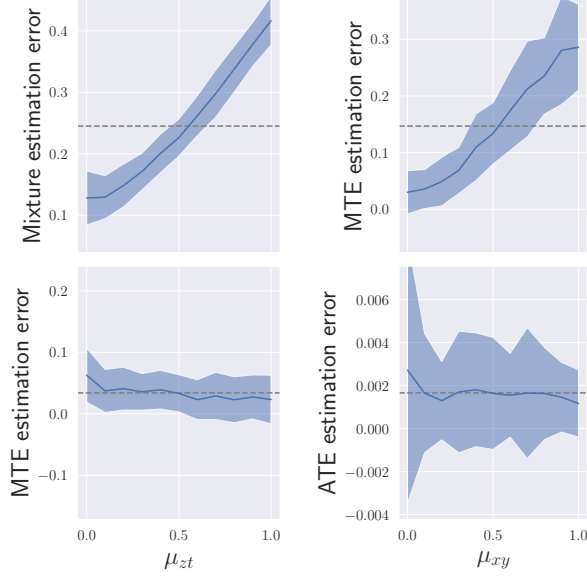
Thus, for $R = Y^{(1)} - Y^{(0)}$,

$$\mathbb{E}_{\mu_{xy}}[R \mid X, U] = \frac{1}{4} \begin{pmatrix} -3 & 3 \\ -3 & 3 \end{pmatrix} + \frac{\mu_{xy}}{4} \begin{pmatrix} 0 & -6 \\ 6 & 0 \end{pmatrix}. \quad (31)$$

The value of $\mu_{zt}$ controls how treatment assignment depends on $\mathbf{Z}$ and $U$, whereas the value of $\mu_{xy}$ controls how treatment effect depends on $\mathbf{X}$ and $U$.

We study the following two models:

$$\begin{aligned} \textbf{Model A:} \quad & \mu_{zt} = 0, \ \mu_{xy} = 0 \\ \textbf{Model B:} \quad & \mu_{zt} = 1, \ \mu_{xy} = 0 \end{aligned}$$

(a) **Level 2 vs. Level 3**    (b) **Level 3 vs. Level 4**

Figure 2: On the left, as we vary $\mu_{zt}$, the mixture estimation error increases, but the MTE estimation error is stable and close to zero. On the right, as we vary $\mu_{xy}$, MTE estimation error increases, but ATE estimation error is stable and close to zero. The blue line is the average error, and the shading covers one standard deviation. The dashed gray line is the mean over all parameter values.

In both models, since $\mu_{xy} = 0$, we have $Y \perp\!\!\!\perp X \mid U$, and $\mathbf{E}[R \mid U] = [-3/4, 3/4]$, i.e., the two models have the same mixture of treatment effects. However, for **Model A**, since $\mu_{zt} = 0$, we have $T \perp\!\!\!\perp Z \mid U$. Meanwhile, in **Model B**, since $\mu_{zt} = 1$, we have $T \perp\!\!\!\perp U \mid Z$. Thus, these two models trade off whether heterogeneity in treatment *allocation* is driven by $U$ or by $Z \leftrightarrow T$ and represent two different mixture models.

**Level 4: ATEs**   SPOs identify ATEs under the conditions given in Theorem 1. We define a third model:

$$\textbf{Model C:} \quad \mu_{zt} = 1,\ \mu_{xy} = 1$$

In this model, $Y \perp\!\!\!\perp U \mid X$, with $\mathbf{E}[R \mid X] = [-3/4, 3/4]$. A quick calculation shows that $\mathbf{E}[R \mid U] = [-3/8, 3/8]$. Hence, **Model C** has a different mixture of treatment effects than **Model B**, but both models have the same ATE, $\mathbb{E}(R) = 0$. Thus, **Model B** and **Model C** trade off whether heterogeneity in treatment *effect* is driven completely by $U$ or completely by $Y \leftarrow X$.

# 5   EMPIRICAL RESULTS

The previous section introduces a model in which the strength of $Z \leftrightarrow T$ and $X \rightarrow Y$ are controlled by $\mu_{zt}$ and $\mu_{xy}$. Recall that full mixture recovery (level 2) is identifiable when neither $Z \leftrightarrow T$ nor $X \rightarrow Y$ exist, MTEs (level 3) are identifiable with $Z \leftrightarrow T$ but no $X \rightarrow Y$, and ATEs are identifiable with both $Z \leftrightarrow T$ and $X \rightarrow Y$.

Our empirical results explore the transition between identifiability with respect to $\mu_{zt}$ and $\mu_{xy}$, which gradually shift us between the regimes of identifiability for these levels. Each level is computed and evaluated using an algorithm and loss function tailored to the task:

- **Level 2**: Computed using PARAFAC in `tensorly` (Kossaifi et al., 2019) and evaluated using the total variation distance between $\Pr(Z, X, Y, T, U)$ and $\widehat{\Pr}(Z \mid U) \cdot \widehat{\Pr}(X \mid U) \cdot \widehat{\Pr}(Y, T \mid U) \cdot \widehat{\Pr}(U)$.
- **Level 3**: Computed using SPOs (all moments) and the matrix pencil method. Evaluated using the squared $\ell_2$ distance between the true MTE parameters $\mathbf{P}[U], \mathbf{E}[R \mid U]$ and their corresponding estimates.
- **Level 4**: Computed using SPOs for just the first moment and evaluated by the squared difference between the true ATE $\mathbb{E}(R)$ and its estimate.

In all experiments, we report results averaged over 100 runs for each parameter value. Each single run consists of taking 1,000 samples and running both methods on the empirical moments. A repository containing a Python implementation of our SPO method and code for replicating our experiments can be found here. For mixture recovery, we use the PARAFAC algorithm in `tensorly` (Kossaifi et al., 2019) to decompose the empirical probability distribution into factors, then normalize the factors into probability distributions $\widehat{\Pr}(U)$, $\widehat{\Pr}(Z \mid U)$, $\widehat{\Pr}(X \mid U)$ and $\widehat{\Pr}(Y, T \mid U)$.

**Level 2 vs. Level 3**   First, we demonstrate the separation between Level 2 and Level 3 by varying the parameter $\mu_{zt}$ from 0 to 1 while keeping $\mu_{xy} = 0$, i.e., we interpolate from **Model A** to **Model B**. As shown in Fig. 2a, our method of SPOs accurately recovers the mixture of treatment effects for all values of $\mu_{zt}$. In contrast, when we perform tensor decomposition, the performance of mixture recovery becomes significantly worse as $\mu_{zt}$ becomes larger.

**Level 3 vs. Level 4**   Next, we demonstrate the separation between Level 3 and Level 4 by varying the parameter $\mu_{xy}$ from 0 to 1 while keeping $\mu_{zt} = 1$,

i.e., we interpolate from **Model B** to **Model C**. As shown in Fig. 2b, our method accurately recovers the average treatment effect for all values of $\mu_{xy}$; however, as expected, the performance of MTE recovery becomes significantly worse as $\mu_{zt}$ becomes larger.

## 6   DISCUSSION

**Relaxing Assumptions for Identifiability**   The independence conditions given in Theorem 1 and Theorem 2 have a nice graphical correspondence in Figure 1, but can be relaxed to exchangeability with respect to expectations.[5] Notice that the decomposition given in Equation 8 only demands $\mathbb{E}[XZ] = \mathbb{E}[X \mid U]\mathbb{E}[Z \mid U]$ for all $Z \in \mathbf{Z}, X \in \mathbf{X}$. This is a milder condition than $\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \mid U$. As a result, Theorem 1 only requires

(i) $\mathbb{E}[ZY \mid U] = \mathbb{E}[Z \mid U]\mathbb{E}[Y \mid U]$ for all $Z \in \mathbf{Z}$,

(ii) $\mathbb{E}[XT \mid U] = \mathbb{E}[X \mid U]\mathbb{E}[T \mid U]$ for all $X \in \mathbf{X}$,

for ATE identifiability with SPOs. Similarly, Theorem 2 only requires

(i) $\mathbb{E}[ZY \mid U] = \mathbb{E}[Z \mid U]\mathbb{E}[Y \mid U]$ for all $Z \in \mathbf{Z}$,

(ii) $\mathbb{E}[XT \mid U] = \mathbb{E}[X \mid U]\mathbb{E}[T \mid U]$ and $\mathbb{E}[XY \mid U] = \mathbb{E}[X \mid U]\mathbb{E}[Y \mid U]$ for all $X \in \mathbf{X}$,

for MTE identifiability using SPOs.

**The Importance of Identifiability**   In the absence of validation sets, identifiability is essential. However, it is important to study identifiability at the correct granularity — too coarse-grained and we limit the information that we recover, too fine-grained and we may impose overly-stringent requirements. Empirical results show that the identifiability transitions are "soft," meaning that loosely violated requirements can still give accurate results. In Appendix B, we give sample complexity results for SPOs relative to the condition number of $\mathbf{M}[\mathbf{X}, \mathbf{Z}]$. Future studies of sample complexity may further resolve stability under mildly violated assumptions.

**Mechanism-based Classes**   In this paper, we argue that MTEs represent a fundamental granularity worth studying and show how their identifiability fits within the more well-studied frameworks and problems. MTEs are classes that are differentiated by their mechanisms, which are ensemble-level properties. Covariate values are often insufficient to differentiate relationships between observables, leading to distribution overlap that

many clustering approaches cannot appropriately address. It is, therefore, essential to consider parametric approaches and methods of moments when developing algorithms. These approaches are not only advantageous because of their increased power of identification (as demonstrated by the gap between level 1 and level 2) but also more private, as individuals are never assigned class memberships.

**Covariate Richness**   Finding a "sufficiently-rich" set of features $\mathbf{\Phi}(\mathbf{X}), \mathbf{\Theta}(\mathbf{Z})$ is left as an open problem. Many results in proximal causal inference assume that the negative controls are rich categorical variables, e.g., alphabets with cardinality $k$. This can be interpreted with respect to algebraic geometry. While additional parameters usually increase the difficulty of identification, more parameters within $\mathbf{X}, \mathbf{Z}$ give more degrees of freedom along which these random variables can differ between different classes. Each of these degrees of freedom can be made into a feature, which gives us access to more moments. Because we look at second-order moments between $\mathbf{Z}, \mathbf{X}$, a linear increase in parameters corresponds to a quadratic increase in the moments used to recover those parameters.

**Heterogeneity Detection**   Our approach is limited by the inherent instability of mixture models. As such, future work should develop statistical tests for the *presence* of (significantly different) MTEs, as in Pearl (2022), rather than guaranteeing full mixture recovery. This problem likely makes up an additional layer of identifiability. Such tests could be used for quality control (e.g., detecting an ineffective batch of medication), quantification of uncertainty in treatment response, and the detection of vaccine-resistant viral variants.

## References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case stud-

---

[5]Thank you to the anonymous reviewer who pointed this out to us.

ies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2023). Causal matrix completion. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3821–3826. PMLR.

Agarwal, A., Shah, D., and Shen, D. (2020). Synthetic interventions. *arXiv preprint arXiv:2006.07691*.

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables.

Anandkumar, A., Ge, R., Hsu, D. J., Kakade, S. M., Telgarsky, M., et al. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832.

de Prony, G. R. (1795). Essai experimental et analytique: sur les lois de la dilatabilite des fluides elastique et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, a differentes temperatures. *Journal Polytechnique ou Bulletin du Travail fait a l'Ecole Centrale des Travaux Publics*.

Gordon, S., Mazaheri, B., Rabani, Y., and Schulman, L. (2021). Source identification for mixtures of product distributions. In *Conference on Learning Theory*, pages 2193–2216. PMLR.

Gordon, S., Mazaheri, B., Rabani, Y., and Schulman, L. (2023). Causal inference despite limited global confounding via mixture models. In *2nd Conference on Causal Learning and Reasoning*.

Gordon, S., Mazaheri, B., Schulman, L. J., and Rabani, Y. (2020). The sparse hausdorff moment problem, with application to topic models. *arXiv preprint arXiv:2007.08101*.

Gordon, S. L., Jahn, E., Mazaheri, B., Rabani, Y., and Schulman, L. J. (2024). Identification of mixtures of discrete product distributions in near-optimal sample and time complexity. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2071–2091. PMLR.

Hua, Y. and Sarkar, T. K. (1990). Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(5):814–824.

Imai, K. and Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1):1–19.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Kendrick Qijun Li, Xu Shi, W. M. and Tchetgen, E. T. (2024). Double negative control inference in test-negative design studies of vaccine effectiveness. *Journal of the American Statistical Association*, 119(547):1859–1870.

Kim, J.-S. and Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy. In *Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014*, pages 293–306. Springer.

Kim, J.-S., Steiner, P. M., and Lim, W.-C. (2015). Mixture modeling methods for causal inference with multilevel data. *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*, pages 335–359.

Kim, Y., Koehler, F., Moitra, A., Mossel, E., and Ramnarayan, G. (2019). How many subpopulations is too many? exponential lower bounds for inferring population histories. In *Research in Computational Molecular Biology: 23rd Annual International Conference, RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings 23*, pages 136–157. Springer.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kossaifi, J., Panagakis, Y., Anandkumar, A., and Pantic, M. (2019). Tensorly: Tensor learning in python. *Journal of Machine Learning Research*, 20(26):1–6.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138.

Künzel, S. R., Walter, S. J., and Sekhon, J. S. (2019). Causaltoolbox—estimator stability for heterogeneous treatment effects. *Observational Studies*, 5(2):105–117.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Loh, W. W. and Kim, J.-S. (2022). Evaluating sensitivity to classification uncertainty in latent subgroup effect analyses. *BMC Medical Research Methodology*, 22(1):247.

Lyu, W., Kim, J.-S., and Suk, Y. (2023). Estimating heterogeneous treatment effects within latent class multilevel models: A bayesian approach. *Journal of Educational and Behavioral Statistics*, 48(1):3–36.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage

estimation and moment restriction. In *International conference on machine learning*, pages 7512–7523. PMLR.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Miao, W., Shi, X., Li, Y., and Tchetgen Tchetgen, E. J. (2024). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*, pages 1–12.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Ogburn, E. L., Shpitser, I., and Tchetgen, E. J. T. (2019). Comment on "blessings of multiple causes". *Journal of the American Statistical Association*, 114(528):1611–1615.

Ogburn, E. L., Shpitser, I., and Tchetgen, E. J. T. (2020). Counterexamples to" the blessings of multiple causes" by wang and blei. *arXiv preprint arXiv:2001.06555*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2022). Detecting latent heterogeneity. In *Probabilistic and causal inference: The works of judea pearl*, pages 483–506.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Rabani, Y., Schulman, L. J., and Swamy, C. (2014). Learning mixtures of arbitrary distributions over large discrete domains. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 207–224.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Shi, X., Miao, W., and Tchetgen, E. T. (2020). A selective review of negative control methods in epidemiology. *Current epidemiology reports*, 7:190–202.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Squires, C., Shen, D., Agarwal, A., Shah, D., and Uhler, C. (2022). Causal imputation via synthetic interventions. In *Conference on Causal Learning and Reasoning*, pages 688–711. PMLR.

Suk, Y., Kim, J.-S., and Kang, H. (2021). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes. *Journal of Educational and Behavioral Statistics*, 46(3):323–347.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, Y. and Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, 37(23):3309–3324.

Xie, Y., Brand, J. E., and Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347.

## Checklist

1. For all models and algorithms presented, check if you include:

    (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

    (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

    (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

    (a) Statements of the full set of assumptions of all theoretical results. [Yes]

    (b) Complete proofs of all theoretical results. [Yes]

    (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

    (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]

    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

    (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

    (a) Citations of the creator If your work uses existing assets. [Not Applicable]

    (b) The license information of the assets, if applicable. [Not Applicable]

    (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

    (d) Information about consent from data providers/curators. [Not Applicable]

    (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

    (a) The full text of instructions given to participants and screenshots. [Not Applicable]

    (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

    (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# A Algorithm Pseudocode

---
**Algorithm 1:** Recovers MTEs using SPOs.

---
**Input:** $\mathbf{M}[\cdot]$ matrices computed by estimating conditional expected values.
**Result:** $\gamma^{(1)}, \ldots, \gamma^{(2k-1)}$
/* Compute $\gamma^{(1)}$ coefficients for $\mathbf{X}$ using reference $\mathbf{Z}$. */
1 **Function** FirstMomentCoeffs($\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T]$, $\mathbf{M}[\mathbf{Z}, Y \mid T]$):
2    $\quad \alpha^{(1)} \leftarrow \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 1]^{-1}\mathbf{M}[\mathbf{Z}, Y \mid T = 1]$
3    $\quad \beta^{(1)} \leftarrow \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 0]^{-1}\mathbf{M}[\mathbf{Z}, Y \mid T = 0]$
4    $\quad$ **return** $\alpha^{(1)} - \beta^{(1)}$

/* Compute $\gamma^{(\ell)}$ for $\mathbf{X}$ using reference $\mathbf{Z}$ and previous $\gamma^{(\ell-1)}$. */
5 **Function** NextMomentCoeffs($\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T]$, $\mathbf{M}[\mathbf{Z}, \mathbf{X}Y \mid T]$, $\gamma^{(\ell-1)}$):
6    $\quad \alpha^{(\ell)} \leftarrow \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 1]^{-1}\mathbf{M}[\mathbf{Z}, \mathbf{X}Y \mid T = 1]\gamma^{(\ell-1)}$
7    $\quad \beta^{(\ell)} \leftarrow \mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 0]^{-1}\mathbf{M}[\mathbf{Z}, \mathbf{X}Y \mid T = 0]\gamma^{(\ell-1)}$
8    $\quad$ **return** $\alpha^{(\ell)} - \beta^{(\ell)}$
9 $\gamma^{(1)} \leftarrow$ FirstMomentCoeffs($\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T]$, $\mathbf{M}[\mathbf{Z}, Y \mid T]$)
10 $\nu_1 \leftarrow \mathbf{M}[\mathbf{X}]^{\top}\gamma^{(1)}$
11 **for** $\ell \leftarrow 2$ **to** $2k - 1$ **do**
12    $\quad \gamma^{(\ell)} \leftarrow$ NextMomentCoeffs($\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T]$, $\mathbf{M}[\mathbf{Z}, \mathbf{X}Y \mid T]$, $\gamma^{(\ell-1)}$)
13    $\quad \nu_\ell \leftarrow \mathbf{M}[\mathbf{X}]^{\top}\gamma^{(\ell)}$
14 $\mathbf{P}[U], \mathbf{E}[R \mid U] \leftarrow$ PronyOrPencil($\nu_1, \ldots, \nu_{2k-1}$)

---

# B Analysis

In this section, we will analyze the sample and time complexity of SPOs. The time complexity of computing an SPO is $\mathcal{O}(k^4)$, which is mild compared to the time required to process the approach's data demands. For this reason, the *sample* complexity is the dominating factor when it comes to run-time.

## B.1 Time Complexity

The computation of an SPO involves inverting a $k \times k$ matrix and multiplying it. This can be done in $\mathcal{O}(k^3)$ time. When computing higher order moments of the treatment effect, we will pick up another factor of $k$. The final step of applying the matrix pencil or Prony's method involves solving for an eigensystem, which is again $\mathcal{O}(k^3)$. Altogether, the time complexity becomes $\mathcal{O}(k^4)$, which is relatively mild when considering the time involved in processing the data.

## B.2 Sample Complexity

The sample complexity for a single SPO computation is given in Theorem 4.

**Theorem 4** (SPO Sample Complexity). *Let $\pi := \min_t \Pr(T = t)$ and let $\lambda_k$ be the smallest eigenvalue of $\mathbf{E}[\mathbf{X} \mid U]$. The sample complexity of calculating $\mathbb{E}(Y^{(t)})$ using SPOs is $\mathcal{O}(k^6\pi^{-1}\lambda_k^{-2})$.*

The formal proof is delayed to Appendix C. We will now discuss a few notable aspects of the sample complexity.

First, the dependence on $k^6$ shows the sample-complexity-dependent nature of this approach — the time needed to compute a result will primarily scale with the data needed to accurately recover a solution. The sample complexity's dependence on $\pi^{-1}$ shows how computing SPOs for rare treatments scales in difficulty, as it limits the data that we have to calculate $\alpha$'s and compute the SPO.

The primary veil of sample complexity lies in $\lambda_k^{-2}$, which essentially says that $\mathbf{E}[\mathbf{X} \mid U]$ must be well-conditioned. Fortunately, the condition number of $\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid t]$ upper bounds the condition number of $\mathbf{E}[\mathbf{X} \mid U]$, which allows us to ensure the stability of the approach in practice.

**Higher Order Moments** To recover MTEs, we must use solutions for SPOs at lower orders to bootstrap higher-order moments. This process accumulates errors, particularly with respect to the dependence on the condition number of $\mathbf{E}[\mathbf{X} \mid U]$. Gordon et al. (2021) outlined an approach that involves only $\mathcal{O}(\log(k))$ steps within this chain, which involves incrementing $\ell$ by more than 1 using *two* synthetic copies (requiring another $\mathbf{X}'$ with $\mathbf{X}' \perp\!\!\!\perp (Y, T) \mid U$).

Gordon et al. (2020) and Kim et al. (2019) outline the sample complexities of Prony's and the matrix pencil method, respectively. A notable factor is a minimum difference between treatment effects across components — clearly, sub-populations with identical or close-to-identical treatment effects will be hard to disentangle. The stability is similarly dependent on the lowest probability sub-population, as rare populations require more data to resolve. We note that these dependencies are not of serious concern to practitioners who are primarily concerned with the *existence* of nontrivial differences in MTEs as a motivator for further investigation.

## C  Sample Complexity Proof

We will now provide an analysis of the empirical error propagation for our method. $\mathbf{M}$ vectors and matrices can be estimated empirically on a dataset $D$, denoted $\hat{\mathbf{M}}^D[\cdot]$. We will also use $\hat{P}[\cdot], \hat{\mathbf{E}}[\cdot]$ vectors and matrices that cannot be directly observed, but still exist hypothetically for anyone who would have observed $U$.

Equation 15 tells us how to compute the coefficients for a synthetic bit in the presence of perfect statistics. In reality, the computation we will be performing is on imperfect empirical estimates,

$$\hat{\boldsymbol{\alpha}} = \hat{\mathbf{M}}[\mathbf{Z}, \mathbf{X} \mid t]^{-1}\hat{\mathbf{M}}[\mathbf{Z}Y \mid t]. \tag{32}$$

Expanding these moments into unobservable $\hat{\mathbf{P}}[\cdot]$ matrices using Equation 8 gives

$$\begin{aligned}
\hat{\mathbf{M}}[\mathbf{Z}, \mathbf{X} \mid t] &= \hat{\mathbf{E}}[\mathbf{Z} \mid U, t]\,\mathbf{diag}(\hat{\mathbf{P}}[U \mid t])\hat{\mathbf{E}}[\mathbf{X} \mid U, t]^{\top}, \\
\hat{\mathbf{M}}[\mathbf{Z}, Y \mid t] &= \hat{\mathbf{E}}[\mathbf{Z} \mid U, t]\,\mathbf{diag}(\hat{\mathbf{P}}[U \mid t])\hat{\mathbf{E}}[Y \mid U, t].
\end{aligned} \tag{33}$$

Substituting these expansions and recalling that $\mathbf{X} \perp\!\!\!\perp T \mid U$ simplifies Equation 32,

$$\hat{\boldsymbol{\alpha}}^{\top} = \hat{\mathbf{E}}[Y \mid U, t]^{\top}\hat{\mathbf{E}}[\mathbf{X} \mid U, t]^{-1} = \hat{\mathbf{E}}[Y \mid U, t]^{\top}\hat{\mathbf{E}}[\mathbf{X} \mid U]^{-1}. \tag{34}$$

An important observation from Equation 34 is that $\hat{\boldsymbol{\alpha}}$ no longer depends on $\mathbf{Z}$. That is, we do not have to worry about the stability of sampled statistics from $\mathbf{Z}$ because any noise is canceled out by the inversion (so long as we do not deviate into non-invertable $\hat{\mathbf{P}}[\mathbf{Z} \mid U, t]$). This significantly simplifies our analysis.

We will analyze the stability of potential outcome expectation estimations computed using the empirical version of Equation 12. We will distinguish between the original dataset $D$ and the post-selected $D_t$, which is formed from the subset of points for which $T = t$. We want to analyze

$$\hat{\mathbb{E}}(Y^{(t)}) = \hat{\mathbf{M}}^D[\mathbf{X}]^{\top}\hat{\boldsymbol{\alpha}}^{D_t}. \tag{35}$$

Statistics computed on the $D_t$ dataset will be less accurate than those on the full $D$ dataset because of the smaller sample-size, so we have expressed our sample complexity in terms of $\min_i \mathbf{P}[T]_i$.

### C.1  Proof of Theorem 4

We use $\delta$ to denote the difference between the true and empirical values, i.e., $\boldsymbol{\delta}\hat{\mathbf{E}}^D[\cdot] := \hat{\mathbf{E}}^D[\cdot] - \mathbf{E}[\cdot]$. To analyze the sample complexity of synthetic potential outcomes, we will give a probably approximately correct bound by separately (1) analyzing the propagation of errors from observed statistics to the recovered $\mathbb{E}[Y^{(t)}]$ and (2) determining the data needed to keep those errors in check. To tackle (1), we will work under Assumption 5.

**Assumption 5.** *For all $V \in \mathbf{X} \cup \mathbf{Z} \cup \{T, Y\}$ we assume $\delta\hat{\mathbb{E}}^D(V), \delta\hat{\mathbb{E}}^{D_t}(V) \leq \varepsilon$. Similarly we assume $\delta\hat{P}^D(U), \delta\hat{P}^{D_t}(U) \leq \varepsilon$.*

Lemma 6 gives us an upper bound on the error in a synthetic potential outcome.

**Lemma 6.** *Under Assumption* 5,

$$\delta\hat{\mathbb{E}}(Y^{(t)}) \leq 2k^4\varepsilon \left\|\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\right\| + 2k^2\varepsilon + k^2\varepsilon^2. \tag{36}$$

*Proof.* First expand Equation 35,

$$\hat{\mathbb{E}}(Y^{(t)}) = \hat{\mathbf{E}}^{D_t}[Y \mid U, t]^\top \hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\hat{\mathbf{E}}^{D}[\mathbf{X} \mid U]\hat{\mathbf{P}}^{D}[U]. \tag{37}$$

Observe that $\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\hat{\mathbf{E}}^{D}[\mathbf{X} \mid U]$ involves *two different empirical estimates* of $\mathbf{E}[\mathbf{X} \mid U]$, one of which is inverted. Define the "error matrix,"

$$\mathbf{\Delta} := \hat{\mathbf{E}}^{D}[\mathbf{X} \mid U] - \hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U], \tag{38}$$

and expand to get

$$\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\hat{\mathbf{E}}^{D}[\mathbf{X} \mid U] = \hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}(\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U] + \mathbf{\Delta}) = \mathbf{I} + \hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\mathbf{\Delta}. \tag{39}$$

Recall that $\mathbb{E}(Y^{(t)}) = \delta\mathbf{E}[Y \mid U, t]^\top \delta\mathbf{P}[U]$ and write the empirical computation in terms of the "correct" (in blue) values plus their errors (in red),

$$\hat{\mathbb{E}}(Y^{(t)}) = (\mathbf{E}[Y \mid U, t]^\top + \delta\hat{\mathbf{E}}^{D_t}[Y \mid U, t]^\top)(\mathbf{I} + \hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\mathbf{\Delta})(\mathbf{P}[U] + \delta\hat{\mathbf{P}}^{D}[U]), \tag{40}$$

This gives us an expression for the difference between the true and computed do-intervention,

$$\begin{aligned}\delta\hat{\mathbb{E}}(Y^{(t)}) =& \hat{\mathbf{E}}^{D_t}[Y \mid U, t]^\top \hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\mathbf{\Delta}\hat{\mathbf{P}}^{D}[U] + \mathbf{E}[Y \mid U, t]^\top \delta\hat{\mathbf{P}}^{D}[U] \\ &+ \delta\hat{\mathbf{E}}^{D_t}[Y \mid U, t]^\top \mathbf{P}[U] + \delta\hat{\mathbf{E}}^{D_t}[Y \mid U, t]^\top \delta\hat{\mathbf{P}}^{D}[U].\end{aligned} \tag{41}$$

In order to upper bound the error in Equation 41 using Cauchy-Schwartz, we use Assumption 5 to find the following bounds on norms.

$$\left\|\hat{\mathbf{E}}^{D_t}[Y \mid U, t]\right\| \leq k$$
$$\|\mathbf{\Delta}\| \leq 2k^2\varepsilon$$
$$\left\|\hat{\mathbf{P}}[U]\right\| \leq k$$
$$\left\|\mathbf{E}^{D_t}[Y \mid U, t]\right\| \leq k$$
$$\left\|\delta\hat{\mathbf{P}}^{D}[U]\right\| \leq k\varepsilon$$
$$\left\|\delta\hat{\mathbf{E}}^{D_t}[Y \mid U, t]\right\| \leq k\varepsilon$$
$$\|\mathbf{P}[U]\| \leq k$$

Applying these inequalities gives

$$\delta\hat{\mathbb{E}}(Y^{(t)}) \leq 2k^3\varepsilon \left\|\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\right\| + 2k^2\varepsilon + k^2\varepsilon^2. \tag{42}$$

$\square$

All that remains is to upper bound $\left\|\hat{\mathbf{P}}^{D_t}[\mathbf{X} \mid U]^{-1}\right\|$. Lemma 7 shows how this is related to the condition number of $\mathbf{E}[\mathbf{X} \mid U]$.

**Lemma 7.** *Let $\lambda_k$ be the smallest eigenvalue of $\mathbf{E}^{D_t}[\mathbf{X} \mid U]$. Then,*

$$\left\|\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\right\| \leq \frac{1}{\lambda_k - k^2\varepsilon}. \tag{43}$$

*Proof.* The operator norm of an the inverse of $\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]$ can be upper bounded by $1/\hat{\lambda}_k$ where $\hat{\lambda}_k$ is the smallest eigenvalue of $\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]$. We know that $\lambda_k - \hat{\lambda}_k$ cannot be any greater than the operator norm of the difference, $\left\|\delta\hat{\mathbf{E}}^{D_t}(\mathbf{X} \mid U)\right\| \leq k\varepsilon$. Therefore we have

$$\left\|\hat{\mathbf{E}}^{D_t}[\mathbf{X} \mid U]^{-1}\right\| \leq \frac{1}{\lambda_k - k\varepsilon}. \tag{44}$$

$\square$

The proof of Theorem 4 now involves analyzing the sample complexity needed to achieve Assumption 5 with an $\varepsilon$ such that $\delta\hat{\mathbb{E}}(Y^{(t)})$ does not depend on $k$ or $\lambda_k$. Hoeffding's inequality gives an upper bound on the probability of failing Assumption 5 for an arbitrary variable $V$ on $n$ samples:

$$\Pr(|\hat{V} - \mathbb{E}[V]| \geq \varepsilon) \leq 2\exp(-2\varepsilon^2 n). \tag{45}$$

In order to ensure that this upper bound remains constant with respect to $k$ and $\lambda_k$ we must have $n$ proportional to $1/\varepsilon^2$. For our error $\delta\hat{\mathbb{E}}(Y^{(t)})$ to remain constant, Lemmas 6 and 7 require $\varepsilon \leq \frac{\lambda_k - k^2}{k^4}$, which consequently gives the sample complexity in Theorem 4 (other than the $\pi^{-1}$, which is due to the data-constraints of conditioning on treatment, as explained in the main paper).

## D  Mixture and ATE identifiability with moments

### D.1  Mixture identifiability with moments (Level 2)

Here, we generalize the proof of mixture identifiability from Allman et al. (2009) to also include continuous-valued feature maps. We recall that the *Kruskal rank* of a matrix $M$, denoted $\mathrm{rank}_K(M)$, is the largest number $r$ such that all sets of $r$ columns in $M$ are linearly independent.

We use the following notation for tensors: given $\mathbf{w} \in \mathbb{R}^k$, $\mathbf{A} \in \mathbb{R}^{a \times k}$, $\mathbf{B} \in \mathbb{R}^{b \times k}$, and $\mathbf{C} \in \mathbb{R}^{c \times k}$, we define the order-3 tensor $\mathbf{T} = [\mathbf{w}; \mathbf{A}, \mathbf{B}, \mathbf{C}] \in \mathbb{R}^{a \times b \times c}$ as

$$\mathbf{T}_{ij\ell} = \sum_{k'=1}^{k} w_{k'} \cdot A_{ik'} \cdot B_{jk'} \cdot C_{\ell k'}$$

Recall from Section 2 the definitions for the matrices $\mathbf{E}[\mathbf{Z} \mid U] \in \mathbb{R}^{d_1 \times k}$ and $\mathbf{E}[\mathbf{X} \mid U] \in \mathbb{R}^{d_2 \times k}$:

$$\mathbf{E}[\mathbf{Z} \mid U]_{iu} = \mathbb{E}(Z_i \mid U = u)$$

$$\mathbf{E}[\mathbf{X} \mid U]_{ju} = \mathbb{E}(X_j \mid U = u)$$

Define a new random vector $\mathbf{S} = (\mathbb{1}_{T=0}, \mathbb{1}_{T=1}, Y\mathbb{1}_{T=0}, Y\mathbb{1}_{T=1})$, and a new matrix $\mathbf{E}[\mathbf{S} \mid U] \in \mathbb{R}^{4 \times k}$:

$$\mathbf{E}[\mathbf{S} \mid U]_{\ell u} = \mathbb{E}(S_\ell \mid U = u)$$

We define the third-order tensor of moments $\mathbf{M}[\mathbf{Z}, \mathbf{X}, \mathbf{S}] \in \mathbb{R}^{d_1 \times d_2 \times 4}$

$$\mathbf{M}[\mathbf{Z}, \mathbf{X}, \mathbf{S}]_{ijt} = \mathbb{E}[Z_i \cdot X_j \cdot S_\ell]$$

**Lemma 8.** *Let $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{S}$ be independent given $U$. Let*

$$I_1 = \mathrm{rank}_K(\mathbf{E}[\mathbf{Z} \mid U]), \quad I_2 = \mathrm{rank}_K(\mathbf{E}[\mathbf{X} \mid U]), \quad and \quad I_3 = \mathrm{rank}_K(\mathbf{E}[\mathbf{S} \mid U]).$$

*Assume that $I_1 + I_2 + I_3 \geq 2k + 2$. Then, given $\mathbf{M}[\mathbf{Z}, \mathbf{X}, \mathbf{S}]$, we can generically identify $\mathbf{P}(U)$, $\mathbf{E}[\mathbf{Z} \mid U]$, $\mathbf{E}[\mathbf{X} \mid U]$, and $\mathbf{E}[\mathbf{S} \mid U]$ up to permutation of the labels of $U$.*

*Proof.* For convenience, we define the following extensions:

$$\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_{d_1}, 1],$$
$$\mathbf{X}' = [X_1, X_2, \ldots, X_{d_2}, 1], \text{ and}$$
$$\mathbf{S}' = [S_1, S_2, S_3, S_4, 1],$$

along with

$$\mathbf{E}[\mathbf{Z}' \mid U]_{iu} = \mathbb{E}(Z_i' \mid U = u),$$
$$\mathbf{E}[\mathbf{X}' \mid U]_{ju} = \mathbb{E}(X_j' \mid U = u), \text{ and}$$
$$\mathbf{E}[\mathbf{S}' \mid U]_{\ell u} = \mathbb{E}(S_\ell' \mid U = u),$$

and finally, $\mathbf{M}[\mathbf{Z}', \mathbf{X}', \mathbf{S}'] \in \mathbb{R}^{(d_1+1) \times (d_2+1) \times 5}$ as

$$\mathbf{M}[\mathbf{Z}', \mathbf{X}', \mathbf{S}']_{ij\ell} = \mathbb{E}[Z'_i \cdot X'_j \cdot S'_\ell].$$

Note that $\mathbf{E}[\mathbf{Z}' \mid U]$ is equal to $\mathbf{E}[\mathbf{Z} \mid U]$ concatenated with an additional row of 1's. Generically, adding this row will not decrease the Kruskal rank, i.e., $\mathrm{rank}_K(\mathbf{E}[\mathbf{Z}' \mid U]) = I_1$, $\mathrm{rank}_K(\mathbf{E}[\mathbf{S}' \mid U]) = I_2$, and $\mathrm{rank}_K(\mathbf{E}[\mathbf{S}' \mid U]) = I_3$.

Then, since $\mathbf{Z}, \mathbf{X}$, and $\mathbf{S}$ are conditionally independent given $U$, we have

$$\mathbf{M}[\mathbf{Z}', \mathbf{X}', \mathbf{S}'] = [\mathbf{P}(U); \mathbf{E}[\mathbf{Z}' \mid U], \mathbf{E}[\mathbf{X}' \mid U], \mathbf{E}[\mathbf{S}' \mid U]]$$

Under the assumption that $I_1 + I_2 + I_3 \geq 2k + 2$, the Kruskal rank theorem (Kruskal, 1977) guarantees that the above rank-$k$ decomposition of $\mathbf{M}[\mathbf{Z}', \mathbf{X}', \mathbf{S}']$ is unique up to permutation of the labels of $U$, and simultaneous scaling of the columns of each matrix. Under the additional constraint that the last rows of $\mathbf{E}[\mathbf{Z}' \mid U]$, $\mathbf{E}[\mathbf{Z}' \mid U]$, and $\mathbf{E}[\mathbf{Z}' \mid U]$ are equal to one, the scaling indeterminacy is removed; i.e., the parameters are recoverable up to permutation. □

Given $\mathbf{P}(U)$ and $\mathbf{E}[\mathbf{S} \mid U]$, we can compute MTEs. For example, we can compute the expected potential outcome under $T = 0$ for each subgroup $u$ as follows:

$$\begin{aligned}
\mathbb{E}[Y^{(0)} \mid U = u] &= \frac{\mathbb{E}(S_3 \mid U = u)}{\mathbb{E}(S_1 \mid U = u)} \\
&= \frac{\mathbb{E}(Y \mathbb{1}_{T=0} \mid U = u)}{\mathbb{E}(\mathbb{1}_{T=0} \mid U = u)} \\
&= \mathbb{E}(Y \mid U = u, T = 0),
\end{aligned}$$

where the last line is a standard change of measure.

## D.2 ATE identifiability with moments (Level 4)

We now explore how to use additional features in $\mathbf{X}$ using a pseudoinverse. Given a matrix $\mathbf{A}$, we let $\mathbf{A}^+$ denote the Moore-Penrose pseudoinverse of $\mathbf{A}$.

**Lemma 9.** *Let* $\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \mid T, U$, *let* $Y \perp\!\!\!\perp \mathbf{Z} \mid T, U$, *and let* $\mathbf{X} \perp\!\!\!\perp T \mid U$. *Assume that* $\mathbf{M}[\mathbf{X}, \mathbf{Z} \mid t]$ *has a left inverse. Then*

$$\mathbb{E}[Y^{(t)}] = \mathbf{M}[Y, \mathbf{Z} \mid t] \cdot \mathbf{M}[\mathbf{X}, \mathbf{Z} \mid t]^+ \cdot \mathbf{E}[\mathbf{X}]$$

*Proof.* First, for any $t$, since $\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \mid T, U$, we have by the law of total probability that

$$\mathbf{M}[\mathbf{X}, \mathbf{Z} \mid t] = \mathbf{E}[\mathbf{X} \mid U, t] \cdot \mathbf{diag}(\mathbf{P}(U \mid t)) \cdot \mathbf{E}[\mathbf{Z} \mid U, t]^\top$$

Similarly, for any $t$, since $Y \perp\!\!\!\perp \mathbf{Z} \mid T, U$, we have

$$\mathbf{M}[Y, \mathbf{Z} \mid t] = \mathbf{E}[Y \mid U, t] \cdot \mathbf{diag}(\mathbf{P}(U \mid t)) \cdot \mathbf{E}[\mathbf{Z} \mid U, t]^\top$$

Thus, if $\mathbf{M}[\mathbf{X}, \mathbf{Z} \mid t]$ has a left inverse,

$$\mathbf{M}[Y, \mathbf{Z} \mid t] \cdot \mathbf{M}[\mathbf{X}, \mathbf{Z} \mid t]^+ = \mathbf{E}[Y \mid U, t] \cdot \mathbf{E}[\mathbf{X} \mid U, t] \tag{46}$$

Now, we prove the result:

$$\begin{aligned}
\mathbb{E}[Y^{(t)}] &= \mathbf{E}[Y \mid U, t] \cdot \mathbf{P}[U] \\
&= \mathbf{E}[Y \mid U, t] \cdot \mathbf{E}[\mathbf{X} \mid U, t]^+ \cdot \mathbf{E}[\mathbf{X} \mid U, t] \cdot \mathbf{P}[U] && \text{(Multiply by the identity)} \\
&= \mathbf{M}[Y, \mathbf{Z}] \cdot \mathbf{M}[\mathbf{X}, \mathbf{Z}]^+ \cdot \mathbf{E}[\mathbf{X} \mid U, t] \cdot \mathbf{P}[U] && \text{(Using (46))} \\
&= \mathbf{M}[Y, \mathbf{Z}] \cdot \mathbf{M}[\mathbf{X}, \mathbf{Z}]^+ \cdot \mathbf{E}[\mathbf{X} \mid U] \cdot \mathbf{P}[U] && \text{(Since } \mathbf{X} \perp\!\!\!\perp T \mid U) \\
&= \mathbf{M}[Y, \mathbf{Z}] \cdot \mathbf{M}[\mathbf{X}, \mathbf{Z}]^+ \cdot \mathbf{E}[\mathbf{X}]
\end{aligned}$$

□

# E  Additional experiments

We have implemented our approach for computing SPOs and provide the following experiments on synthetic data.

**Experiment 1: ATEs**  We sample Bernoulli $U$ ($k = 2$) uniformly as well as 4 covariates $|\mathbf{X}| = |\mathbf{Z}| = 2$ from non-identical Bernoulli distributions that depend only on $U$. For both experiments, we have the following causal mechanisms:

$$\Pr(Z_1 = 1 \mid U) = 0.2 + 0.3 \cdot U$$
$$\Pr(Z_2 = 1 \mid U) = 0.28 + 0.3 \cdot (1 - U)$$
$$\Pr(X_1 = 1 \mid U) = 0.36 + 0.3 \cdot U$$
$$\Pr(X_2 = 1 \mid U) = 0.44 + 0.3 \cdot (1 - U)$$

These choices ensure that the matrices $\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 1]$ and $\mathbf{M}[\mathbf{Z}, \mathbf{X} \mid T = 0]$ in Algorithm 1 are relatively well-conditioned. We sample $T, Y$ according to

$$\Pr(T = 1) = \frac{3}{4} - \frac{U}{2}$$
$$\Pr(Y = 1) = \frac{1}{4} + \frac{T}{4} + \frac{\mathbb{1}_{U=T}}{4}. \tag{47}$$

We do this for 100 runs of 100,000 samples each and compute the difference between the true and calculated ATE in each run, reporting the results in Figure 3(a). As expected, the method accurately estimates the average treatment effect $\mathbb{E}[Y^{(1)} - Y^{(0)}]$, despite unobserved confounding.

**Experiment 2: MTEs**  We sample from the same generative model as above, this time taking 100 runs of 500,000 samples each (for each parameter value) to account for the additional complexity of the task. We report



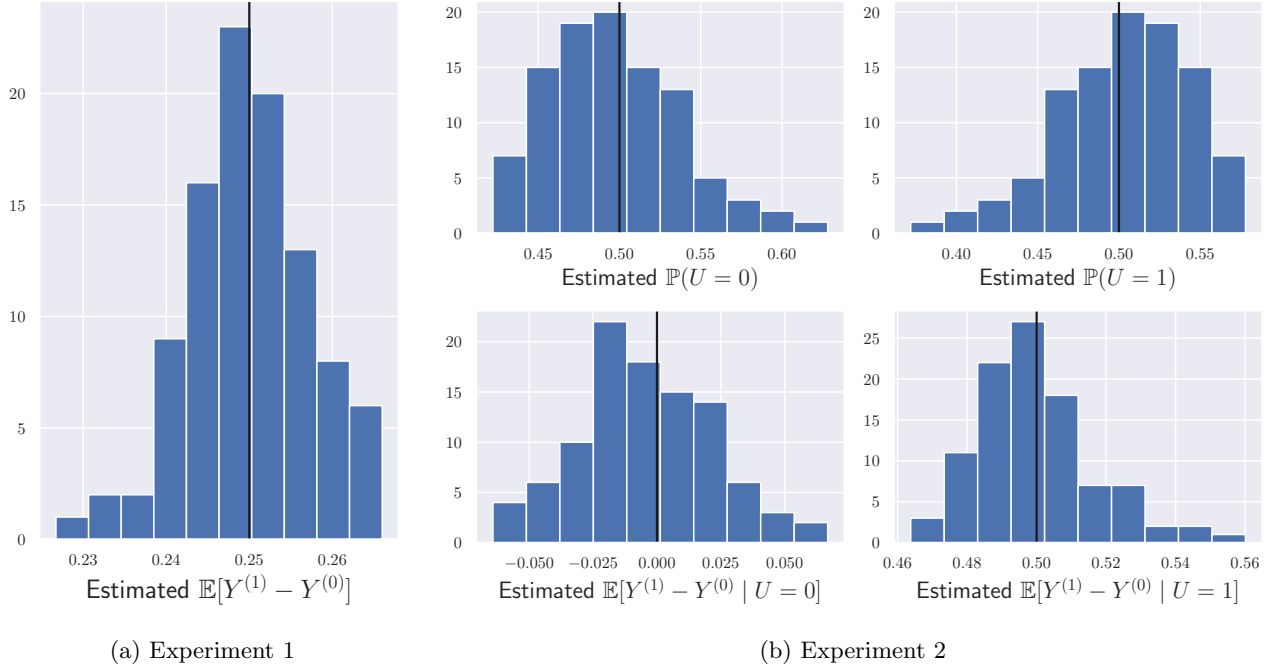(a) Experiment 1                    (b) Experiment 2

Figure 3: **Synthetic Potential Outcomes accurately recover the ATE (average treatment effects), as well as the decomposition of the ATE into MTEs (mixed treatment effects).** In each plot, the true value is shown as a black vertical line, and the estimated values from 100 runs are shown as a histogram. See text for details.

the results in Figure 3(b), showing that our method accurately decomposes the ATE into a mixture of two distinct underlying treatment effects. Most importantly, the method recovers a clear separation between the two treatment effects, exposing the heterogeneity underlying the ATE that was recovered in Figure 3(a).