The Promises of Parallel Outcomes

Ying Zhou, Dingke Tang, Dehan Kong, Linbo Wang

Department of Statistical Sciences, University of Toronto, 700 University Avenue, Toronto, Ontario M5G 1X6, Canada

yingx.zhou@mail.utoronto.ca, dingke.tang@mail.utoronto.ca, dehan.kong@utoronto.ca, linbo.wang@utoronto.ca

October 17, 2022

Abstract

A key challenge in causal inference from observational studies is the identification and estimation of causal effects in the presence of unmeasured confounding. In this paper, we introduce a novel approach for causal inference that leverages information in multiple outcomes to deal with unmeasured confounding. The key assumption in our approach is conditional independence among multiple outcomes. In contrast to existing proposals in the literature, the roles of multiple outcomes in our key identification assumption are symmetric, hence the name parallel outcomes. We show nonparametric identifiability with at least three parallel outcomes and provide parametric estimation tools under a set of linear structural equation models. Our proposal is evaluated through a set of synthetic and real data analyses.

Keywords: Causal inference; Latent confounding; Multivariate outcome; Non-parametric identification.

1 Introduction

Unmeasured confounding poses a major threat to the validity of causal conclusions drawn from observational studies. Over the past few decades, there have been many frameworks developed to mitigate bias due to unmeasured confounding, such as the instrumental variable methods (e.g. Wright and Wright, 1928; Hernán and Robins, 2006; Wang and Tchetgen Tchetgen, 2018), causal structure learning (e.g. Spirtes et al., 2000), front-door adjustment (e.g. Pearl, 2009), invariant prediction (e.g. Peters et al., 2016), and negative controls (e.g. Shi et al., 2020b).

In this paper, we contribute to this effort by introducing a novel approach for causal inference with unmeasured confounding. Our approach leverages the information in multiple outcomes
that are influenced by the same exposure. The key identification assumption in our approach is the independence among the multiple outcomes conditional on the common exposure and both measured and unmeasured confounders. Compared to existing proposals in the literature that use multiple outcomes for causal inference, our approach is unique in that the roles of multiple outcomes in our key identification assumption are symmetric. This is appealing as in many modern applications, a priori, there is no reason to discriminate one outcome over the others. For example, in genetics applications studying associations between a risk factor and gene expression levels, gene expression data are often collected on multiple genes (Leek and Storey, 2007). In medical applications, with the abundance of electronic health records, investigators may now study the effect of a risk factor, such as off-label drug usage, on many health outcomes at the same time (Eguale et al., 2016). In financial applications, it is often of interest to study the implications of a particular policy on the returns of multiple stocks (Menchero et al., 2010).

As a specific example, in Section 6, we study the effect of smoking on serum vitamin C levels. To deal with unmeasured confounding between smoking and vitamin C, previous studies use genetic variants (Wehby et al., 2011) or cigarette price (Leigh and Schembri, 2004) as instrumental variables. As an alternative approach, we note that in the National Health and Nutrition Examination Survey 2005-2006, participants took a series of lab tests that measured a variety of health indexes, including serum vitamin C. Many of these health indexes are influenced by smoking, which motivates our developments below that leverage information in multiple parallel outcomes for identifying and estimating causal effects.

The idea of causal inference with multiple outcomes has been explored previously in the literature. For example, Mealli and Pacini (2013) and Mattei et al. (2013) used secondary outcomes to obtain tighter bounds for causal effects defined within certain subgroups known as principal strata. They assumed that a particular identification assumption, called the exclusion restriction, may be violated for the primary outcome of interest, but holds for a secondary outcome. Another line of research in this direction uses negative control outcomes, with the key assumption that the exposure has no causal effect on the negative control outcomes. Rosenbaum (1989) showed that a negative control outcome can be used to test for hidden confounding. With an additional variable known as negative control exposure, Miao et al. (2018a) further showed that the ACE is nonparametrically identifiable, and Shi et al. (2020a) developed a semiparametric inference procedure in the context of a categorical latent confounder and a binary exposure. Miao et al. (2018b, §6)
extended this framework by replacing the negative control outcome with a positive control outcome. They assumed that the exposure effect on the positive control outcome is non-zero, but known a priori. Our nonparametric identification strategies in Section 3 assume a similar causal structure to Miao et al. (2018b, §6) but do not require prior knowledge of the exposure effect.

2 A general framework for parallel outcomes

Let \( X \in \mathbb{R} \) be a scalar exposure, \( U \in \mathbb{R}^r \) be latent confounding variables and \( V \in \mathbb{R}^q \) be baseline covariates. Under the potential outcome framework, \( Y(x) = (Y^{(1)}(x), \ldots, Y^{(p)}(x))^T \in \mathbb{R}^p \) is the potential outcome had the subject received exposure \( x \). Following the stable unit treatment value assumption (Rubin, 1980), the observed outcome \( Y = Y(x) \) when \( X = x \). We are interested in estimating the mean potential outcome \( E\{Y(x)\} \) with \( n \) samples \( \{(X_i, Y_i^{(1)}, \ldots, Y_i^{(p)}); i = 1, \ldots, n\} \) independently drawn from the joint distribution of \( (X, Y^{(1)}, \ldots, Y^{(p)}) \).

If we had measured covariates \( U \) in addition to \( V \), then under the following assumptions, \( E\{Y(x)\} \) can be identified by \( E\{Y(x)\} = E_{U,V}E(Y \mid X = x, U, V) \).

**Assumption 1.** (Latent ignorability): For any \( x \), \( X \perp \perp Y(x) \mid (U, V) \).

**Assumption 2.** (Positivity): For any measurable set \( \tilde{X} \) such that \( \Pr(X \in \tilde{X}) > 0 \), we have \( \Pr(X \in \tilde{X} \mid U, V) > 0 \) almost surely.

When \( U \) is latent, however, \( E\{Y(x)\} \) is not identifiable without further assumptions. To address this problem, our approach leverages a conditional independence structure among the multiple outcomes \( Y^{(1)}, \ldots, Y^{(p)} \) to identify \( E\{Y(x)\} \). This structure is formally defined in Assumption 3.

**Assumption 3.** (Parallel outcomes): \( Y^{(1)}, Y^{(2)}, \ldots, Y^{(p)} \) are mutually independent conditional on \( (U, V, X) \): \( Y^{(1)} \perp \perp Y^{(2)} \perp \ldots \perp Y^{(p)} \mid (U, V, X) \).

Figure 1 gives the simplest causal diagram (Pearl, 2009) associated with the parallel-outcome model when \( p = 3 \). There is no directed or bi-directed edges among \( Y^{(j)} \)'s, encoding the conditional independence in Assumption 3. This assumption may be partially tested and relaxed under a set of linear structural equation models; see Remark 2 in §4.1.

If one considers \( Y^{(1)} \) as the primary outcome, then \( Y^{(2)} \) and \( Y^{(3)} \) in fact satisfy the conditions for negative control exposures as defined in Miao et al. (2018b). Despite this connection,
Figure 1: The simplest causal diagram associated with the parallel-outcome model when $p = 3$. The baseline covariates $V$ are omitted for brevity. Variables $X, Y^{(1)}, Y^{(2)}, Y^{(3)}$ are observed; $U$ is unobserved.

we coin the term parallel outcomes as in the literature, negative control exposures have commonly been used to refer to exposure variables; see Shi et al. (2020b) for examples. Moreover, unlike Assumption 3, usually no assumptions are imposed among negative control exposures. Nevertheless, it is worth pointing out that from this perspective, our results in Section 3 can also be interpreted as nonparametric identifiability of causal effects with a pair of negative control exposures that are conditional independent on $U$ and $X$.

3 Nonparametric identification

In this section, we study nonparametric identifiability of the mean potential outcome $E\{Y(x)\}$. We illustrate the main idea using the binary model in which $X, U, Y^{(j)}, j = 1, \ldots, p$ are all binary variables taking values 1 or 2, and then present identification results for the general categorical case. Our results extend to continuous models, which require more technicalities and will be deferred to the Supplementary Material §S1.

In the following, we suppress the dependence on baseline covariates $V$ for simplicity. For random variables $(W_1, \ldots, W_p)$, we denote their joint distribution as $\text{pr}(w_1, \ldots, w_p)$, the conditional distribution of $W_i$ given $W_j = w_j$ as $\text{pr}(w_i \mid w_j)$, and the marginal distribution of $W_i$ as $\text{pr}(w_i)$.

In the case where $p = 2$, it is easy to see that in the binary model, the observed data distribution $\text{pr}(x, y^{(1)}, y^{(2)})$ is determined by 7 parameters, while there are 11 unknown parameters involving $\text{pr}(u), \text{pr}(x \mid u)$ and $\text{pr}(y^{(j)} \mid u, x), j = 1, 2$. So nonparametric identification is generally not possible in this case; see the Supplementary Material §S2 for a counterexample.

We hence focus on the scenario where $p = 3$. A quick calculation shows that in this case, both the observed data distribution and unknown parameters reside in 15-dimensional spaces. We
shall show that under some additional conditions, there is indeed a one-to-one mapping between a subset of the parameter space and the observed data space. Consequently, these parameters are identifiable up to ordering, and the potential outcome distributions are identifiable.

Our identification approach is built on the matrix adjustment method used in Rothman et al. (2008), Hu (2008) and Kuroki and Pearl (2014). Note that from the parallel-outcome assumption, we have

\[ \Pr(y^{(2)}, y^{(3)} \mid x) = \sum_u \Pr(y^{(2)} \mid u, x) \Pr(y^{(3)} \mid u, x) \Pr(u \mid x), \] (1)
\[ \Pr(y^{(1)}, y^{(2)}, y^{(3)} \mid x) = \sum_u \Pr(y^{(1)} \mid u, x) \Pr(y^{(2)} \mid u, x) \Pr(y^{(3)} \mid u, x) \Pr(u \mid x). \] (2)

Following the matrix adjustment method, we rewrite (1) and (2) in the form of matrix multiplication. Let \( P(Y^{(2)}, Y^{(3)} \mid x) = \left( \Pr(Y^{(2)} = i, Y^{(3)} = j \mid X = x) \right)_{2 \times 2} \) be a \( 2 \times 2 \) matrix whose \((i, j)\)-th element is given by \( \Pr(Y^{(2)} = i, Y^{(3)} = j \mid X = x) \). Similarly, we let \( P(Y^{(2)} \mid U, x) = \left( \Pr(Y^{(2)} = i \mid U = j, X = x) \right)_{2 \times 2}, P(Y^{(3)} \mid U, x) = \left( \Pr(Y^{(3)} = i \mid U = j, X = x) \right)_{2 \times 2} \). We also let \( P_D(U \mid x) = \left( \Pr(U = i \mid X = x) \right)_{2 \times 2} \) be a diagonal matrix whose \((i, i)\)-th element is given by \( \Pr(U = i \mid X = x) \). We define \( P_D(y^{(1)} \mid U, x) = \left( \Pr(Y^{(1)} = y^{(1)} \mid U = i, X = x) \right)_{2 \times 2} \) similarly; here the subscript \( D \) refers to a diagonal matrix. Equations (1) and (2) can then be rewritten as

\[ P(Y^{(2)}, Y^{(3)} \mid x) = P(Y^{(2)} \mid U, x) P_D(U \mid x) P(Y^{(3)} \mid U, x)^T, \] (3)
\[ P(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x) = P(Y^{(2)} \mid U, x) P_D(y^{(1)} \mid U, x) P_D(U \mid x) P(Y^{(3)} \mid U, x)^T. \] (4)

To eliminate common terms in (3) and (4), we assume the following condition:

**Condition 1 (Full rank).** For all \( x \), \( P(Y^{(2)}, Y^{(3)} \mid x) \) is of full rank.

Under Condition 1, (3) and (4) imply that

\[ P(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x) P(Y^{(2)}, Y^{(3)} \mid x)^{-1} = P(Y^{(2)} \mid U, x) P_D(y^{(1)} \mid U, x) P(Y^{(2)} \mid U, x)^{-1}. \] (5)

The left-hand side of equation (5) can be identified from the observed data distribution, while the right-hand side is a canonical form of eigendecomposition. From the eigendecomposition of the
left-hand side of (5), we have that \( P_D(y^{(1)} | U, x) \) is identifiable from data up to permutation of the eigenvalues. To recover the ordering of eigenvalues, we assume the following condition.

**Condition 2 (No qualitative \( U\rightarrow X \) interaction).** For any fixed level \( y^{(1)} \) of \( Y^{(1)} \), \( 1 \leq u_1 < u_2 \leq k \), \( \Pr(y^{(1)} | U = u_2, x) - \Pr(y^{(1)} | U = u_1, x) \) are non-zero and have the same sign for all \( x \).

Under Condition 2, we can code \( U \) such that \( \Pr(Y^{(1)} = 1 | u, x) \) is strictly increasing in \( u \). We can hence identify \( \Pr(y^{(1)} | u, x) \) for each \( u \) from \( P_D(y^{(1)} | U, x) \). Furthermore, since by definition the columns of \( P(Y^{(2)} | U, x) \) sum up to 1, we can also identify \( P(Y^{(2)} | U, x) \) and hence \( \Pr(y^{(2)} | u, x) \) from equation (5). By symmetry, \( \Pr(y^{(3)} | u, x) \) is also identifiable. It then follows from (3) that \( P_D(U | x) \) and hence \( \Pr(u) \) is identifiable.

Finally, the potential outcome distributions can be identified from the g-formula:

\[
\Pr\{y^{(j)}(x)\} = \sum_u \Pr(y^{(j)} | x, u) \Pr(u), j = 1, 2, 3.
\]

These arguments also apply in a general categorical model, which gives us our main nonparametric identification result.

**Theorem 1.** Suppose that the latent confounder \( U \) and parallel outcomes \( Y^{(2)}, Y^{(3)} \) are all categorical variables with cardinality \( k \). Suppose further that Assumptions 1, 3 and Conditions 1 and 2 hold. Then for all \( x \), the potential outcome distributions \( \Pr\{y^{(j)}(x)\}, j = 1, 2, 3 \) are identifiable.

**Remark 1.** Kuroki and Pearl (2014)'s procedure can be used to establish identifiability of causal effects assuming the absence of arrow \( X \rightarrow Y^{(3)} \) in Figure 1; see Figure S1 in the Supplementary Material for an illustration of their causal diagram. Hu (2008) used a similar approach to deal with a measurement error problem, in which their interest lies in the causal effect of the latent exposure on the outcome; see Figure S2 in the Supplementary Material for an illustration of their causal diagram.

In Theorem 1, Condition 1 can be checked against observed data. Moreover, in a binary model where \( U, Y^{(2)}, Y^{(3)} \) are binary variables, Condition 1 holds if the causal diagram in Figure 1 is faithful (Pearl, 2009); see Lemma S1 in the Supplementary Material §S15 for more details. It also implies the positivity assumption 2 so the latter is not included in Theorem 1.

Condition 2 requires that for at least one of the parallel outcomes, labeled as \( Y^{(1)} \), the effect of latent confounder \( U \) on the outcome has the same direction across all exposure levels. In
other words, there is no qualitative interaction between $U$ and $X$ on the additive scale in the model for $Y^{(1)}$. This is much weaker than and implied by the treatment effect homogeneity assumption that there is no interaction between $U$ and $X$ on the additive scale in the outcome model, an assumption that is commonly invoked in contexts such as instrumental variable models (e.g. Wang and Tchetgen Tchetgen, 2018, A5.b). On the other hand, motivated by a reviewer’s comment, we clarify that Condition 2 is non-trivial either. Indeed, for each value of $x$, we can pick a permutation $\pi_{y,x}(\cdot)$ on $U$ such that $P(Y^{(1)} = y \mid U = \pi_{y,x}(i), x) < P(Y^{(1)} = y \mid U = \pi_{y,x}(j), x)$ if $i < j$. Condition 2 assumes that there exists such a permutation $\pi_{y}(\cdot)$ that applies to all values of $x$. In other words, although one can label $U$ in an arbitrary order, Condition 2 requires that such an order must be independent of the value of $x$.

Arguing as in the proof of Theorem 1, we may still arrive at local identifiability without Condition 2, that is, there is only a finite number of possible distributions for $\text{pr}\{y^{(j)}(x)\}, j = 1, 2, 3$ that are compatible with the observed data distribution. Condition 2 is then needed to achieve global identifiability. We provide an illustration in the Supplementary Material §S4.

In Theorem 1, we do not place restrictions on the cardinality of the exposure $X$ and outcome $Y^{(1)}$, so it applies as long as there are two outcomes with no fewer levels than the unmeasured confounder $U$. When the cardinality of $Y^{(2)}$ or $Y^{(3)}$ is larger than $k$, one can first combine different levels in $Y^{(2)}$ or $Y^{(3)}$ and then apply Theorem 1. In particular, with at least two continuous outcomes, Theorem 1 implies that the causal effects are nonparametrically identifiable assuming a categorical confounder with arbitrarily many levels. Furthermore, Theorem 1 can be generalized to more than three outcomes by choosing different sets of three outcomes and applying Theorem 1 repeatedly.

Our nonparametric identification results suggest that it is possible to consistently estimate causal effects under the parallel-outcome structure. In the Supplementary Material §S5-S6, we describe simple nonparametric estimating procedures under the discrete model, and evaluate their finite sample performance through simulations.

In the Supplementary Material §S1, we extend Theorem 1 to a continuous model. We however, point out that our nonparametric results do not apply to the linear structural equation models described later in (6) and (7), assuming that $p = 3$ and $U, \epsilon_X, \epsilon_j, j = 1, 2, 3$ follow Gaussian distributions. In particular, Condition S4 is violated in this model. See Supplementary Material §S7 for elaborations. This motivates our developments below.
4 Identification and estimation in linear structural equation models

4.1 Identification

In this section, we consider causal effect identification under a set of linear structural equation models with multiple latent confounders. We assume the following structural equation models:

\[ X = \alpha_X^T U + \epsilon_X, \]  
\[ Y^{(j)} = \alpha_j^T U + \beta_j X + \epsilon_j, \quad j = 1, \ldots, p, \]  

where \( U \in \mathbb{R}^r \) is an unobserved confounder and \( \epsilon_X, \epsilon_1, \ldots, \epsilon_p \) are random errors so that \( E(U\epsilon_X) = 0 \in \mathbb{R}^r, \text{Cov}(U) = I_r, (U, \epsilon_X) \perp \perp (\epsilon_1, \ldots, \epsilon_p), \epsilon_1, \ldots, \epsilon_p \) are mutually independent, and \( \text{Cov}(X, Y^{(j)}) \neq 0, j = 1, \ldots, p. \) All the random variables in these models are centered, so there are no intercepts. Let \( \sigma_X^2, \sigma_1^2, \ldots, \sigma_p^2 \) be the variance of \( \epsilon_X, \epsilon_1, \ldots, \epsilon_p, \) respectively. We are mainly interested in the parameters \( \beta_j, \) representing the causal effects of \( X \) on \( Y^{(j)}, j = 1, \ldots, p. \)

Remark 2. The mutual independence among \( \epsilon_1, \ldots, \epsilon_p \) implies Assumption 3. This condition may be partially checked by examining the error covariance matrix \( \text{Cov}(\epsilon) \) (e.g. Fan et al., 2013). See Section 6 for an illustration. If \( \text{Cov}(\epsilon) \) is not diagonal, then Assumption 3 is violated. In this case, one can find the maximal sub diagonal matrix of \( \text{Cov}(\epsilon) \) and apply our proposed method to the corresponding outcomes.

Remark 3. The condition that \( \text{Cov}(X, Y^{(j)}) \neq 0 \) can be checked from observed data. In the case that this condition is violated for some \( Y^{(j)}, \) one may apply the proposed method to the subset of outcomes with non-zero correlations with the exposure.

Substituting (6) into (7), we get

\[ Y^{(j)} = (\alpha_j^T + \beta_j \alpha_X^T)U + \beta_j \epsilon_X + \epsilon_j, \quad j = 1, \ldots, p. \]  

Let \( Y = (Y^{(1)}, \ldots, Y^{(p)})^T, \quad F = (\sigma_X^{-1} \epsilon_X, U_1, \ldots, U_r)^T, \quad \epsilon = (\epsilon_1, \ldots, \epsilon_p)^T. \) It follows that (i) \( F \perp \perp \epsilon; \) (ii) \( E(F) = 0, \text{Cov}(F) = I_{r+1}; \) (iii) \( E(\epsilon) = 0, \text{Cov}(\epsilon) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2). \) Hence (8) is
an orthogonal factor model with a matrix form

$$Y = \Gamma^* F + \epsilon,$$  \hspace{1cm} (9)

in which $\Gamma^*_{p \times (r+1)}$ is a loading matrix with the $j$-th row $\Gamma^*_j = (\sigma X \beta_j, \alpha_{j1} + \alpha X_1 \beta_j, \cdots, \alpha_{jr} + \alpha X_r \beta_j)$. Under the following mild condition, $\Gamma^*$ can be identified up to rotation (Anderson and Rubin, 1956, Thm 5.1). This condition also implies that $p \geq 2(r + 1) + 1$.

**Condition 3 (submatrix rank).** After removing any row, there remain two disjoint submatrices of $\Gamma^*$ of rank $r + 1$.

To further identify the rotation matrix, we assume the following sparsity conditions on the the causal parameter $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ and the loading matrix $\Gamma^*$.

**Condition 4 (sparsity).**

(i) $\|\beta\|_0 = s < p - r$;

(ii) The submatrix of $\Gamma^*$ corresponding to the $p - s$ outcomes with $\beta_j = 0$ has rank $r$; any other submatrix of $\Gamma^*$ consisting of $p - s$ rows has the full rank $r + 1$.

Condition 4 assumes that the causal parameter is sparse, while ruling out some too sparse $\Gamma^*$. In (ii), the submatrix corresponding to the $p - s$ outcomes with $\beta_j = 0$ is not of full rank because the entries in the first column are all zero. The other parts of Condition 4 are similar to the sparsity condition assumed in Wang et al. (2017), and in parallel to conditions (i) and (iii) in Miao et al. (2022, Theorem 3). We refer readers to Wang et al. (2017, Remark 2.2) and Miao et al. (2022) for discussions of this condition.

Under Conditions 3 and 4, one can identify the rows of $\Gamma^*$ whose first element is zero. In other words, one can identify the index set of outcomes with zero causal effects: $S_0 = \{ j : \beta_j = 0 \}$.

**Lemma 1.** Under models (6), (7) and Conditions 3, 4, we have

$$\|\Gamma^*_1\|_0 \leq \|(\Gamma^* R)_1\|_0$$

for any $(r + 1) \times (r + 1)$ rotation matrix $R$ with $R^T R = I$; here $A_j$ denotes the $j$-th column of the matrix $A$. Furthermore, for any $R \in \arg \min_{R^T R = I_{r+1}} \|(\Gamma^* R)_1\|_0$, we have $\{ j : (\Gamma^* R)_j = 0 \} = S_0.$
To identify the $s$ non-zero elements in $\beta$, we extend the two stage least squares method in Miao et al. (2018b) to accommodate multiple confounders. Following their terminology, we refer to the outcomes with indexes in $S_0$ as negative control outcomes, and the remaining outcomes as positive control outcomes. In the first stage, we regress a set of negative control outcomes on the exposure and positive control outcomes, and obtain the fitted values for these negative control outcomes. In the second stage, we regress the outcome of interest on the exposure and the fitted negative control outcomes. Condition 5 guarantees that the exposure and other regressors in the second stage are not co-linear.

**Condition 5 (noncollinearity).** Denote $[p] = \{1, \ldots, p\}$. Let $W = (Y^{(j_1)}, \ldots, Y^{(j_{p-s})}) \in \mathbb{R}^{p-s}$, where $\{j_1, \ldots, j_{p-s}\} = S_0$. For each $\ell \in [p] \setminus S_0$, $X$ is not a linear combination of $E(W \mid X, Z^\ell)$, where $Z^\ell = (Y^{(k_1)}, \ldots, Y^{(k_{s-1})}) \in \mathbb{R}^{s-1}$, with $\{k_1, \ldots, k_{s-1}\} = [p] \setminus \{j_1, \ldots, j_{p-s}, \ell\}$.

Our identification results under the linear structural equation models can be summarized as follows.

**Theorem 2.** Under models (6), (7) and Conditions 3, 4, and 5, the parameters $\beta_1, \ldots, \beta_p$ are identifiable.

### 4.2 Estimation

Throughout this section, we use upper-case letters to denote random vectors/matrices or constant matrices, and calligraphy letters denote sample vectors/matrices. If $S$ is a set of indexes, then $A_S$ denotes the columns of $A$ with column indexes in $S$. Suppose we observe independent and identically distributed data $\{(X_i, Y_i^{(1)}, \ldots, Y_i^{(p)}); i = 1, \ldots, n\}$ following models (6) and (7).

Based on (9), we have the factor model

$$Y_{n \times p} = F_{n \times (r+1)} \Gamma^* + E_{n \times p},$$

(10)

where $Y_{n \times p} = (Y_i^{(j)})_{n \times p}$, $\Gamma^*_{p \times (r+1)} = (\sigma_X \beta, \alpha_1 + \alpha_X \beta, \ldots, \alpha_r + \alpha_X \beta)$, $E_{n \times p} = (\epsilon_{i,j})_{n \times p}$, $F_{n \times (r+1)} = (\epsilon_X / \sigma_X, U_1, \ldots, U_r)$.

In (10), the number of factors $r + 1$ can be estimated using standard tools such as the Kaiser rule (Kaiser, 1960), while the loading matrix $\Gamma^*$ can be estimated up to rotation using the principal component method. Denote $\hat{\Gamma}$ an estimate of $\Gamma^*$ up to rotation.
Lemma 1 suggests that one can identify the set \( S_0 \) by solving an \( \ell_0 \)-minimization problem

\[
(x^*, w^*) = \arg \min_{x,w} \|x\|_0 \quad \text{subject to } w^T w = 1, x = \Gamma w,
\]

(11)

where \( \Gamma = \Gamma^* Q \) with \( Q \in \mathbb{R}^{(r+1) \times (r+1)} \) an unknown rotation matrix. It follows that \( S_0 = \{ j : x^*_j = 0 \} \). To account for the variability in \( \hat{\Gamma} \), we introduce a threshold \( \delta > 0 \) and solve the following problem instead:

\[
(\hat{y}^*, \hat{w}^*) = \arg \min_{y,w} \sum_{i=1}^p \mathbb{I}(|y_i| > \delta) \quad \text{subject to } w^T w = 1, y = \hat{\Gamma} w.
\]

(12)

We let

\[
\hat{S}_0 = \{ j : |\hat{y}_j^*| \leq \delta \}
\]

(13)

be an estimate of the set \( S_0 \), \( \hat{\Sigma} \in \mathbb{R}^{p \times p} \) be the sample covariance of \((Y_{i(1)}^*, \ldots, Y_{i(p)}^*)^T\), and \( \lambda_j(\hat{\Sigma}) \) be the \( j \)th eigenvalue of \( \hat{\Sigma} \). The threshold in (13) is set as \( \delta = \sqrt{2n^{-1} \log(p) \sigma^2} \) (Donoho and Johnstone, 1994), where \( \sigma^2 = p^{-1} \sum_{r+2}^p \lambda_j(\hat{\Sigma}) \). We include detailed derivation of \( \sigma^2 \) in the Supplementary Material §S8. If there exists a known scalar \( M > 0 \) such that \( M \geq \|\hat{y}^*\|_\infty \), then (12) can be reformulated as a mixed integer programming problem (Feng et al., 2018); see the Supplementary Material, §S9 for details. In the simulations and data application, we let \( M = 30 \).

It remains to estimate \( \beta_\ell \) for \( \ell \in [p] \setminus \hat{S}_0 \). A natural estimator of \( \beta_\ell \) is the two stage least squares estimator (Miao et al., 2018a). Specifically, let \( X \) be the sample vector of \( X \) and \( W = Y_{\hat{S}_0} \in \mathbb{R}^{n \times |\hat{S}_0|} \), so that each column of \( W \) corresponds to an outcome \( Y^{(j)} \) with \( \hat{\beta}_j = 0 \). For each \( \ell \in [p] \setminus \hat{S}_0 \), we let \( Z_\ell = Y_{\ell} \in \mathbb{R}^{n \times |T|} \), where \( T = [p] \setminus (\{\ell\} \cup \hat{S}_0) \). To estimate \( \beta_\ell \), one first regresses \( W \) on \((X, Z_\ell)\), and obtains the fitted value \( \hat{W} \). In the second step, one regresses \( Y_\ell \) on \((X, \hat{W})\), so that \( \hat{\beta}_\ell \) is the estimated coefficient of \( X \).

However, there may be collinearity among the regressors in the second step if \( |\hat{S}_0| > |T| \). This can happen, for example, if all the random variables in models (6) and (7) follow Gaussian distributions. To address this, we suggest using ridge regression in the second stage when regressing \( Y_\ell \) on \((X, \hat{W})\). Denote \( A = (X, Z_\ell) \in \mathbb{R}^{n \times (1+|T|)} \) the column-combined data matrix. We obtain \( \hat{W} = A(A^T A)^{-1} A^T W \). Let \( \hat{B} = (X, \hat{W}) \in \mathbb{R}^{n \times (1+|\hat{S}_0|)} \) be the column-combined data matrix. The proposed ridge estimator for \( \beta_\ell \) is \( \hat{\beta}_{\ell,ridge} = (\hat{B}^T \hat{B} + \lambda D)^{-1} \hat{B}^T Y_\ell \), where \( D = \text{diag}\{0, 1, \ldots, 1\} \) is a \((1 + |\hat{S}_0|) \times (1 + |\hat{S}_0|)\) diagonal matrix. The first diagonal element
of \( \mathcal{D} \) is 0 because we do not penalize the regressor \( \mathcal{X} \). In practice, the tuning parameter \( \lambda \) can be chosen by 10-fold cross-validation.

In summary, our estimator \( \hat{\beta} \) is defined as

\[
\hat{\beta}_j = \begin{cases} 
0 & j \in \hat{S}_0, \\
(\hat{B}^T \hat{B} + \lambda \mathcal{D})^{-1} \hat{B}^T \gamma & j \notin \hat{S}_0.
\end{cases}
\]

(14)

where \( \hat{S}_0 \) is defined in (13).

5 Simulation

In this section, we evaluate the performance of our proposed estimator (14). The data are generated following models (6) and (7). We consider a two-dimensional unobserved confounder \( U = (U_1, U_2)^T \). We set \( \sigma_X = 1 \) and \( \sigma_j = 1.5 + 0.25 \{(j + 2) \mod 3\} \) for \( j = 1, \ldots, p \), and generate \( (U_1, U_2, \sigma_X^{-1} \epsilon_X, \sigma_1^{-1} \epsilon_1, \ldots, \sigma_p^{-1} \epsilon_p)^T \) from \( \mathcal{N}(0, I_{p+3}) \). Define \( \gamma \in \mathbb{R}^{7p} \) so that \( \gamma_{(7k+1):(7k+7)} = (1.5, -1.8, 2.1, 2.4, -2.7, 3, -3.3), k = 0, \ldots, p - 1 \), where \( \gamma_{i:j} \) denotes the subvector of \( \gamma \) from the \( i \)th element to the \( j \)th element; so \( \gamma_{1:2p} \) is the first \( 2p \) entries of \( \gamma \). We set \( \alpha_X = (1, 1)^T, (\alpha_1^T, \alpha_2^T)^T = \gamma_{1:2p}, \beta_j = (-1)^j[1 + \{(j + 3) \mod 4\}] \) if \( 1 \leq j \leq 0.4p \), and \( \beta_j = 0 \) if \( 0.4p < j \leq p \). We consider \( p = 30, 60, 100 \) and \( n = 500, 1000, 2000 \) in our simulations.

In our simulations and data application, we use the function nScree in the package nFactors to estimate the number of factors \( r + 1 \) by the Kaiser rule, and the package POET (Fan et al., 2013) to estimate the error covariance matrix \( \text{Cov}(\epsilon) \) and check the conditional independence assumption 3. We solve the optimization problem (12) using the gurobi package.

In Table 1 we report the false positive rates and false negative rates over 100 Monte Carlo runs, defined as \( |\{j : \hat{\beta}_j \neq 0, \beta_j = 0\}|/|\{j : \hat{\beta}_j \neq 0\}| \) and \( |\{j : \hat{\beta}_j = 0, \beta_j \neq 0\}|/|\{j : \hat{\beta}_j = 0\}| \), respectively. One can see that our selection procedure for negative control outcomes is conservative in that our false negative rates are zero for all \( (n, p) \) combinations. In comparison, the false positive rate is non-zero, albeit very small, in the simulation settings considered here. This is desirable as consistent estimation of causal parameters relies on valid negative control outcomes. If \( \beta_j \neq 0 \) but \( \hat{\beta}_j = 0 \), then the estimates for all non-zero elements of \( \beta \) may be biased as \( Y^{(j)} \) is then mistaken as a negative control outcome in the ridge estimator (14). In comparison,
Table 1: False positive rate ×10000 and false negative rate ×10000 of the proposed selection method for the negative control outcomes

|              | False positive rate ×10000 | False negative rate ×10000 |
|--------------|----------------------------|----------------------------|
|              | p = 30                     | p = 60                     | p = 100                     | p = 30 | p = 60 | p = 100 |
| n = 500      | 58                         | 16                         | 10                         | 0      | 0      | 0       |
| n = 1000     | 54                         | 16                         | 54                         | 0      | 0      | 0       |
| n = 2000     | 37                         | 12                         | 5                          | 0      | 0      | 0       |

Table 2: Bias×100 (Standard error ×100) of the proposed ridge estimator for estimating the causal parameters $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4) = (-1, 2, -3, 4)$

|              | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ |
|--------------|------------------|------------------|------------------|------------------|
| p = 30       | n = 500          | 0.75(0.75)       | 1.26(0.98)       | -0.52(1.06)      | 0.98(0.79)      |
|              | n = 1000         | 0.03(0.60)       | 0.61(0.67)       | -0.25(0.88)      | 0.78(0.63)      |
|              | n = 2000         | -0.68(0.45)      | -0.12(0.49)      | 0.03(0.49)       | 0.10(0.36)      |
| p = 60       | n = 500          | -1.40(0.74)      | 0.52(0.90)       | -1.27(0.94)      | 1.43(0.83)      |
|              | n = 1000         | -0.58(0.55)      | 0.71(0.55)       | -1.81(0.84)      | 0.14(0.60)      |
|              | n = 2000         | 0.30(0.40)       | 0.04(0.47)       | -0.70(0.58)      | 0.15(0.40)      |
| p = 100      | n = 500          | -0.57(0.75)      | 0.91(0.86)       | 0.63(0.93)       | 0.60(0.83)      |
|              | n = 1000         | -1.06(0.59)      | 0.09(0.70)       | 1.10(0.71)       | 0.62(0.55)      |
|              | n = 2000         | 0.47(0.34)       | -0.30(0.38)      | -1.13(0.54)      | 0.45(0.41)      |

A false positive result does not introduce bias for the ridge estimator (14).

In Table 2, we report the average biases and the standard errors of the biases of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$, corresponding to the causal effects of the first four outcomes. We can see that the biases are small relative to their standard deviations in all the scenarios considered here.

6 Real data Application

Studies have found that smoking, including secondhand smoking, are associated with a high risk of vitamin C inadequacy (Institute of Medicine (US), 2000). In this section, we study how smoking affects serum vitamin C levels. As a byproduct, we also obtain the effects of smoking on a group of laboratory test results.

The data we use come from the National Health and Nutrition Examination Survey 2005-2006. The exposure of interest is the serum cotinine level. Cotinine is a major metabolite of nicotine, and the measurement of serum cotinine is a marker for active and passive smoking.
(Benowitz, 1996). We consider 20 laboratory results, including serum vitamin C level, as outcomes in our analysis. We provide detailed descriptions of these outcomes in the Supplementary Material, Table S3. These outcomes are selected from different categories in the laboratory data section using the criteria that they are representative of different health aspects, so that they are not likely to directly affect each other. A total of 1852 samples with ages between 12 and 85 are included in the analysis, among which 925 are females and 927 are males.

In our analysis, we adjust for observed confounders, including gender, age, and race, by regressing the log-transformed exposure and outcomes on these covariates. We take the residual corresponding to the exposure as $X$ and the residuals corresponding to the outcomes as $Y^{(j)}$, $j = 1, \ldots, 20$. We then regress each $Y^{(j)}$ on $X$ and keep the outcomes with estimated regression coefficient greater than $\hat{\sigma}_{(j)} \sqrt{2 \log 20}$, where $\hat{\sigma}_{(j)}$ is the standard error of the estimated regression coefficient. This leaves us with 11 outcomes, including acrylamide, vitamin C, cadmium, urinary thiocyanate, retinyl palmitate, osmolality, vitamin D, vitamin B12, blood mercury, enterolactone, and C-reactive protein.

After preprocessing, we apply our proposed methods in Section 4.2 to estimate the causal effects of serum cotinine level on the outcomes. Using the Kaiser rule, we select three factors in the model (10), suggesting a two-dimensional latent confounder $U$. We then estimate the error covariance matrix $\text{Cov}(\epsilon)$ using the R package POET. All the off-diagonal entries in the error covariance matrix are estimated to be zero, with diagonal elements \{0.31, 0.51, 0.33, 0.38, 0.70, 0.57, 0.57, 0.69, 0.81, 0.61, 0.58\}. This provides empirical evidence for Assumption 3. Our selection procedure described in (13) selects four positive control outcomes for which the causal effect of serum cotinine is potentially non-zero: acrylamide, vitamin C, cadmium, and urinary thiocyanate. Our causal effect estimation procedure described in (14) shows that increasing serum cotinine level by one fold causes acrylamide to increase by 8.6%(95% CI: [7.9%, 9.3%]), vitamin C to decrease by 3.9%(95% CI: [2.1%, 5.7%]), cadmium to increase by 11.7%(95% CI: [10.9%, 12.5%]), and urinary thiocyanate to increase by 9.2%(95% CI: [7.8%, 10.6%]). Note when reporting these confidence intervals, we do not take into account the randomness due to the selection of positive control outcomes. Our findings are consistent with the literature studying the effects of tobacco smoke on acrylamide (e.g. Mojska et al., 2016), vitamin C (e.g. Schectman et al., 1989; Preston et al., 2003), cadmium (e.g. Richter et al., 2017), and urinary thiocyanate (e.g. Ngogang et al., 1983).
7 Discussion

Our identification results in this paper show that the parallel-outcome framework provides a promising alternative to the multi-cause framework (Wang and Blei, 2019); the latter has attracted a lot of attention recently in the causal inference and machine learning community. It is also related to, but conceptually different from existing methods based on secondary outcomes such as negative controls or proximal causal inference. In particular, the roles of $Y^{(1)}, \ldots, Y^{(p)}$ in the definition of parallel outcomes are symmetric, which allows our method to be applied in a wide range of practical settings where multiple parallel outcomes are available. In fact, our framework has been implicitly assumed in genetic applications for multiple hypothesis testing with latent confounding. As pointed out by Leek and Storey (2008), this problem can be represented by a linear model relating a multivariate response to an exposure and latent factors. They also assume that the errors in the linear model are normally distributed with independent noises, which implies the parallel outcome structure defined in Assumption 3.

In related work, Wang et al. (2017) further assume that the latent confounding factors may be represented as a linear transformation of the exposure variable plus a Gaussian error. They develop an approach to identify the direct effect from an exposure to multiple outcomes in the presence of unmeasured mediators. Their model and ours can be seen as reparametrizations of each other. For example, when $r = 1$, equation (6) can be rewritten as $U = a_X X + \epsilon_U$, where $a_X = \alpha_X (\alpha_X^2 + \sigma_X^2)^{-1}$, $\epsilon_U = (\alpha_X^2 + \sigma_X^2)^{-1} \sigma_X^2 U - (\alpha_X^2 + \sigma_X^2)^{-1} \alpha_X \epsilon_X$, and $\text{Cov}(X, \epsilon_U) = 0$. Similarly, their model can be rewritten in the form of equation (6). The key difference between Wang et al. (2017)’s and our parametric identification results in Section 4.1 is that we require a much weaker sparsity condition than Wang et al. (2017). In particular, Wang et al. (2017) assume $\|\beta\|_0 \leq (p-r)/2$, while our Condition 4 requires that $\|\beta\|_0 < p-r$. This improvement in sparsity is possible due to a novel identification approach we develop.

The parallel-outcome assumption 3 might be violated if there are common latent mediators $M$ for the exposure effect on the outcomes. In this case, the causal effects are still identifiable if one assumes linear structural equation models on the relationships among $U, X, M, Y$; see the Supplementary Material, §S14 for detailed discussions.
Supplementary material

The Supplementary Material contains nonparametric identification results in the continuous model, and methods and simulations for nonparametric estimation in the discrete model. It also contains additional examples and proofs of all the theorems and lemmas.

References

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, 3.5:111–151.

Benowitz, N. L. (1996). Cotinine as a biomarker of environmental tobacco smoke exposure. *Epidemiologic Reviews*, 18(2):188–204.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Eguale, T., Buckeridge, D. L., Verma, A., Winslade, N. E., Benedetti, A., Hanley, J. A., and Tamblyn, R. (2016). Association of off-label drug use and adverse drug events in an adult population. *JAMA Internal Medicine*, 176(1):55–63.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4):603–680.

Feng, M., Mitchell, J. J., Pang, J.-S., Shen, X., and Wachter, A. (2018). Complementarity formulations of $\ell_0$-norm optimization. *Pacific Journal of Optimization*, 14(2):273–305.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist’s dream? *Epidemiology*, 17(4):360–372.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61.

Institute of Medicine (US) (2000). *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. National Academies Press (US), Washington (DC).
Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):1724–1735.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

Leigh, J. P. and Schembri, M. (2004). Instrumental variables technique: cigarette price provided better estimate of effects of smoking on SF-12. *Journal of Clinical Epidemiology*, 57(3):284–293.

Mattei, A., Li, F., Mealli, F., et al. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, 7(4):2336–2360.

Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108(503):1120–1131.

Menchero, J., Morozov, A., and Shepard, P. (2010). Global equity risk modeling. In *Handbook of Portfolio Construction*, pages 439–480. Springer.

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

Miao, W., Hu, W., Ogburn, E. L., and Zhou, X.-H. (2022). Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–15.

Miao, W., Shi, X., and Tchetgen Tchetgen, E. J. (2018b). A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*. 

17
Mojska, H., Gielecińska, I., and Cendrowski, A. (2016). Acrylamide content in cigarette mainstream smoke and estimation of exposure to acrylamide from tobacco smoke in Poland. *Annals of Agricultural and Environmental Medicine*, 23(3):456–461.

Ngogang, J., Eben-Moussi, E., and Raisonnier, A. (1983). Salivary, urinary and plasma thiocyanate in smokers and non-smokers. *Pathologie-Biologie*, 31(3):155–160.

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

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012.

Preston, A. M., Rodriguez, C., Rivera, C. E., and Sahai, H. (2003). Influence of environmental tobacco smoke on vitamin C status in children. *The American Journal of Clinical Nutrition*, 77(1):167–172.

Richter, P., Faroon, O., and Pappas, R. S. (2017). Cadmium and cadmium/zinc ratios and tobacco-related morbidities. *International Journal of Environmental Research and Public Health*, 14(10):1154.

Rosenbaum, P. R. (1989). The role of known effects in observational studies. *Biometrics*, 45(2):557–569.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Schectman, G., Byrd, J. C., and Gruchow, H. W. (1989). The influence of smoking on vitamin C status in adults. *American Journal of Public Health*, 79(2):158–162.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. (2020a). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B*, 82(2):521–540.
Shi, X., Miao, W., and Tchetgen Tchetgen, E. J. (2020b). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports, 7*(4):190–202.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000). *Causation, Prediction, and Search*. MIT press.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics, 45*(5):1863–1894.

Wang, L. and Tchetgen Tchetgen, E. J. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B, 80*(3):531–550.

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

Wehby, G., Fletcher, J. M., Lehrer, S. F., Moreno, L. M., Murray, J. C., Wilcox, A., and Lie, R. T. (2011). A genetic instrumental variables analysis of the effects of prenatal smoking on birth weight: Evidence from two samples. *Biodemography and Social Biology, 57*(1):3–32.

Wright, P. G. and Wright, S. (1928). *The tariff on animal and vegetable oils*. Macmillan Company, New York.
Abstract

The Supplementary Material is organized as follows. Section S1 contains non-parametric identification results in the continuous model. Section S2 contains an example that shows that the causal effects are not identifiable with a categorical latent confounder and two parallel outcomes. Section S3 includes two causal diagrams for related works. In Section S4, we use an example to show local identifiability without Condition 2 in the categorical model. Sections S5 and S6 contain methods and simulations for non-parametric estimation in the discrete model. In Section S7, we explain why Condition S4 is violated in a linear Gaussian model. We derive the threshold in (12) in Section S8, give the reformulation of (12) to a mixed integer programming problem in Section S9, and provide a description of the outcomes in the real data example in Section S10. Sections S11-S13 contain proofs of theorems and lemmas in the main text. Section S14 extends models (6) and (7) to accommodate common latent mediators among outcomes. Sections S15 and S16 contain additional lemmas and related proofs, and Sections S17 and S18 contain proofs of Theorem S1 and Theorem S2.

S1 Non-parametric identification with a continuous latent confounder

S1.1 Main identification results

In most empirical studies, the latent confounder $U$ is continuous. In parallel to the eigendecomposition of matrix $P(y^{(1)}, Y^{(2)}, Y^{(3)} | x)P(Y^{(2)}, Y^{(3)} | x)^{-1}$, we use spectral decomposition of the corresponding integral operator to identify the mean potential outcomes $E\{Y^{(j)}(X = x)\}, j = 1, 2, 3$. 

1
1, 2, 3. Specifically, let \( \mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3, \mathcal{U}, \mathcal{X} \) denote the support of random variables \( Y^{(1)}, Y^{(2)}, Y^{(3)}, U, \) and \( X, \) respectively. Similar to Theorem 1, we do not place restrictions on the cardinality of the exposure \( X, \) but assume that outcomes \( Y^{(1)}, Y^{(2)}, Y^{(3)} \) are continuous random variables. Throughout this section, we assume all the density functions are bounded, which is formally stated as below (Hu and Schennach, 2008, A1).

**Condition S1.** The joint distribution of \( X \) and \( U, Y^{(1)}, Y^{(2)}, Y^{(3)} \) admits a bounded density with respect to the product measure of some dominating measure \( \mu \) defined on \( \mathcal{X}, \) and the Lebesgue measure on \( \mathcal{Y}^{(1)} \times \mathcal{Y}^{(2)} \times \mathcal{Y}^{(3)} \times \mathcal{U}. \) For any \( A, B \subset \{U, Y^{(1)}, Y^{(2)}, Y^{(3)}\} \) such that \( A \cap B = \emptyset \) and \( x \in \mathcal{X}, \) the conditional density \( f(A \mid B, X = x) \) is bounded.

To facilitate the statement of our results, we first introduce the concept of an integral operator. Let \( \mathcal{L}^1_{\text{bnd}}(\mathcal{A}) \) be the set of all bounded absolutely integrable function over the support \( \mathcal{A}. \) An integral operator denoted by \( T_{k(a,b)} \) is an operator mapping \( g \in \mathcal{L}^1_{\text{bnd}}(\mathcal{B}) \) to \( T_{k(a,b)} g \in \mathcal{L}^1_{\text{bnd}}(\mathcal{A}), \) defined as \( (T_{k(a,b)}g)(a) = \int_{\mathcal{B}} k(a,b) g(b) \, db; \) here \( k(a,b) \) is called the kernel of the operator. This generalizes matrix multiplication in the discrete case. Similar to matrix transpose, we let \( (T_{k(a,b)})^T g) = \int_{\mathcal{A}} k(a,b) g(a) \, da. \) In parallel to (3) and (4), we have

\[
T_{f(Y^{(2)}, Y^{(3)} \mid x)} = T_{f(Y^{(2)} \mid U, x)} \Delta_{f(U \mid x)} T_{f(Y^{(3)} \mid U, x)}^T, \tag{S1}
\]

\[
T_{f(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)} = T_{f(Y^{(2)} \mid U, x)} \Delta_{f(y^{(1)} \mid U, x)} \Delta_{f(U \mid x)} T_{f(Y^{(3)} \mid U, x)}^T. \tag{S2}
\]

The multiplication operators in (S1) and (S2), defined as \( \Delta_{m(U)} g(U) = m(U) g(U), \) generalize the diagonal matrices in (3) and (4). Under Condition S2 below, we have

\[
T_{f(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)}^{-1} T_{f(Y^{(2)} \mid U, x)}^{-1} = T_{f(Y^{(2)} \mid U, x)} \Delta_{f(y^{(1)} \mid U, x)} T_{f(Y^{(2)} \mid U, x)}^{-1}. \tag{S3}
\]

By the spectral decomposition of the left-hand side of (S2), under Conditions S3 and S4 below, we can identify \( E\{Y^{(j)}(x)\}, j = 1, 2, 3 \) in a similar way as in Section 3.

**Condition S2.** \( f(Y^{(2)} \mid U, x) \) is bounded complete in \( Y^{(2)}, f(Y^{(2)}, Y^{(3)} \mid x) \) is bounded complete in \( Y^{(3)}, \) and \( f(Y^{(3)} \mid Y^{(2)}, x) \) is bounded complete in \( Y^{(3)}; \) here a function \( k(a, b) \) is said to be bounded complete in \( a \) if for all \( f(b) \in \mathcal{L}^1_{\text{bnd}}, \int k(a, b) f(b) \, db = 0 \) implies \( f(b) = 0. \)

**Condition S3.** If there exists \( x \) such that \( Y^{(1)} \mid u_1, x \equiv Y^{(1)} \mid u_2, x, \) then \( u_1 = u_2. \)

**Condition S4.** There exists a known continuous functional \( M \) such that \( M\{f(Y^{(2)} \mid U, X)(u, x)\} = \)
\[ g^{-1}\{h_1(u) + h_2(x)\}, \] where \( h_1 \) is one-to-one, \( h_1(u) \) is bounded from above or below for \( u \in \mathcal{U} \), and \( g \) is a known continuous link function.

**Theorem S1.** Suppose that the latent confounder \( U \) and parallel outcomes \( Y^{(1)}, Y^{(2)}, Y^{(3)} \) are all continuous variables. Suppose further that Assumptions 1, 3, and Conditions S1-S4 hold. Then for all \( x \), the mean potential outcomes \( E\{Y^{(j)}(X = x)\}, j = 1, 2, 3 \) are identifiable.

The bounded completeness assumption in Condition S2 generalizes the full rank condition 1; this assumption was often used for nonparametric identification in the causal inference literature (e.g. Miao et al., 2018; Yang et al., 2019). If the kernel \( k(a, b) \) of an operator \( T_{k(a,b)} \) is bounded complete, then \( T_{k(a,b)} \) is injective. Hence it plays a similar role to the full rank condition in the discrete case; see §S1.2 for more discussions.

Recall that in the discrete case, we can code \( U \) such that \( \text{pr}(Y^{(1)} = 1 \mid u, x) \) is strictly increasing in \( u \) for a fixed \( x \). Condition 2 is used to ensure that \( \text{pr}(Y^{(1)} = 1 \mid u, x) \) is strictly increasing in \( u \) for all values of \( x \), so that we can code \( U \) in a consistent way across different values of \( x \). Otherwise we cannot apply the formula \( \text{pr}(u_i) = E_X\{\text{pr}(u_i \mid x)\} \) to identify \( \text{pr}(u_i) \), which is needed in identifying the potential outcome distributions by the g-formula \( \text{pr}\{y^{(j)}(x)\} = \sum_u \text{pr}\{y^{(j)} \mid x, u\} \text{pr}(u), j = 1, 2, 3 \). When \( U \) takes value in the real line, the eigenvalues \( f(y^{(1)} \mid U, x) \) are continuous functions of \( U \). Even if we can code \( U \) such that for all \( x, f(y^{(1)} \mid u, x) \) is strictly increasing in \( u \), it is in general not possible to identify \( f(y^{(1)} \mid u, x) \) for a specific \( u \) from the set \( \{f(y^{(1)} \mid u, x) : u \in \mathcal{U}\} \). This is because there may be infinitely many increasing one-to-one mappings from \( \mathcal{U} \) to \( \mathcal{U} \). To identify \( f(y^{(1)} \mid u, x) \) for a specific \( u \), we instead employ Conditions S3 and S4.

Similar to Condition 2, which is used to identify the distribution of \( U \) up to relabelling in the discrete case, Conditions S3 and S4 are used to identify the distribution of a continuous variable \( \tilde{U} \), which is a one-to-one mapping of \( U \). Condition S3 is similar to the non-zero part in Condition 2. Unlike Condition 2 which is about the eigenvalues \( \{ f(y^{(1)} \mid u, x) : u \in \mathcal{U} \} \), Condition S4 is about the eigenfunctions \( \{ e_u(y^{(2)}) : u \in \mathcal{U} \} \), where \( e_u : y^{(2)} \in \mathcal{Y}^{(2)} \mapsto f(y^{(2)} \mid u, x) \in \mathbb{R}_+ \). We aim to index each eigenfunction to a specific \( u \) in a way that the index is not affected by \( x \), and Condition S4 serves this purpose. It assumes the effect of \( U \) on \( Y^{(2)} \) is bounded (from above or below) on a known scale, and \( U \) has no effect modification on the relationship between \( X \) and \( Y^{(2)} \) on the same scale. This is stronger than Condition 2 for the discrete case, which only
requires the direction of the confounding effect of $U$ on $Y^{(1)}$ is the same for all values of $x$. For example, Condition S4 holds if $E(Y^{(2)} \mid U, X) = h_1(U) + h_2(X)$, or $\log \{\text{var}(Y^{(2)} \mid U, X)\} = h_1(U) + h_2(X)$, where $h_1$ is one-to-one, $h_1(U)$ is bounded, and $h_2$ is an arbitrary function of $X$.

### S1.2 Discussion on injectivity and bounded completeness

Let $\mathcal{L}^1$ be the set of all absolutely integrable function over the support, and $\mathcal{L}^1_{\text{bnd}}$ be the set of bounded functions in $\mathcal{L}^1$. We say that a function $k(a, b)$ is bounded complete in $a$ if for all $f(b) \in \mathcal{L}^1_{\text{bnd}}$, $\int_B k(a, b)f(b) \, db = 0$ implies $f(b) = 0$. We say that a linear operator $T_{k(a,b)}$ is injective as a mapping from $\mathcal{G}(B)$ to $\mathcal{G}(A)$ if $(T_{k(a,b)}f)(a) = (T_{k(a,b)}g)(a)$ implies $f = g$.

**Proposition S1.** When $\mathcal{G} = \mathcal{L}^1_{\text{bnd}}$, $T_{k(a,b)}$ is injective if and only its kernel $k(a, b)$ is bounded complete in $a$.

**Proof of Proposition S1.**

\[
T_{k(a,b)} \text{ is injective}
\iff (T_{k(a,b)}f)(a) = (T_{k(a,b)}g)(a) \Rightarrow f = g
\iff [T_{k(a,b)}(f - g)](a) = 0 \Rightarrow f - g = 0
\iff k(a, b) \text{ is bounded complete in } a.
\]

Another closely related concept is completeness. We say that a function $k(a, b)$ is complete in $a$ if for all $f(b) \in \mathcal{L}^1$, $\int_B k(a, b)f(b) \, db = 0$ implies $f(b) = 0$. By definition, completeness implies bounded completeness. Completeness has been studied and used in literature. For example, exponential families are known to be complete, see Lemma S6. We would like to mention that, usually exponentially family is defined as the distribution with the probability density function (or probability mass function) of the form $f_X(x \mid \theta) = \psi(\theta)h(x)\exp\{\eta(\theta)^T\lambda(x)\}$, where $\theta$ is the parameter. In our context, the conditional distribution $f_{a\mid b}(a \mid b; \theta)$ is called exponential family in the sense that $b$ is considered as the parameter, while $\theta$ is taken as fixed.

**Remark S1.** Suppose that

\[
f_{b\mid a}(b \mid a; \theta_1) = \psi_1(a; \theta_1)h_1(b; \theta_1)\exp\{\eta_1(a; \theta_1)^T\lambda_1(b; \theta_1)\},
\]
then the joint density of \((a, b)\) has the form

\[
\begin{align*}
    f_{a,b}(a, b) &= f_{b|a}(b \mid a; \theta_1) f_a(a; \theta_2) \\
    &= f_a(a; \theta_2) \psi_1(a; \theta_1) h_1(b; \theta_1) \exp\{\eta_1(a; \theta_1)^T \lambda_1(b; \theta_1)\}.
\end{align*}
\]

By Lemma S6, if it holds that

1) \(\psi_1(a; \theta_1) > 0, h_1(b; \theta_1) > 0\);
2) \(\eta_1(a; \theta_1)\) is one-to-one in \(a\), \(\lambda_1(b; \theta_1)\) is one-to-one in \(b\);
3) \(\eta_1(a; \theta_1)\) contains an open set, \(\lambda_1(b; \theta_1)\) contains an open set.

Then \(f_{a,b}(a, b)\) is complete in \(a\), and \(f_{b|a}(b, a)\) is complete in \(b\).

**Remark S2.** There are not many families of distributions beyond the exponential family that are known to be complete. However, the weaker concept of bounded completeness has been studied and known to encompass a larger family of distributions. For example, the location family generated by an absolutely continuous distribution (with respect to Lebesgue measure) is bounded complete if and only if its characteristic function is zero free (Mattner, 1993).

**S2 Counterexample for non-parametric identification with a categorical latent confounder and two outcomes considered in Section 3**

In this section, we show that with a categorical latent confounder considered in Section 3, the potential outcome distribution cannot be identified with two outcomes. We will start with the binary model in which \(U, Y^{(1)}, Y^{(2)}\) are binary variables, and then generalize to the case where \(U, Y^{(1)}, Y^{(2)}\) have \(k\) categories, where \(k \geq 2, k \in \mathbb{Z}\).

When there are two outcomes \(Y^{(1)}\) and \(Y^{(2)}\) in the binary model, the joint distribution of
\((X, Y^{(1)}, Y^{(2)}, U)\) is determined by 11 parameters, which we set as
\[
\begin{align*}
    \text{pr}(U = 1) &= 0.6, \quad \text{pr}(X = 1 \mid U = 1) = 0.7, \quad \text{pr}(X = 1 \mid U = 2) = 0.4, \\
    \text{pr}(Y^{(1)} = 1 \mid U = 1, X = 1) &= 0.1, \quad \text{pr}(Y^{(1)} = 1 \mid U = 2, X = 1) = 0.2, \\
    \text{pr}(Y^{(1)} = 1 \mid U = 1, X = 2) &= 0.3, \quad \text{pr}(Y^{(1)} = 1 \mid U = 2, X = 2) = 0.4, \\
    \text{pr}(Y^{(2)} = 1 \mid U = 1, X = 1) &= 0.1, \quad \text{pr}(Y^{(2)} = 1 \mid U = 2, X = 1) = 0.2, \\
    \text{pr}(Y^{(2)} = 1 \mid U = 1, X = 2) &= 0.3, \quad \text{pr}(Y^{(2)} = 1 \mid U = 2, X = 2) = 0.4.
\end{align*}
\]

Then the joint distribution of \((X, Y^{(1)}, Y^{(2)})\) is
\[
\begin{align*}
    \text{pr}(X = 1, Y^{(1)} = 1, Y^{(2)} = 1) &= 0.0106, \quad \text{pr}(X = 1, Y^{(1)} = 1, Y^{(2)} = 2) = 0.0634, \\
    \text{pr}(X = 1, Y^{(1)} = 2, Y^{(2)} = 1) &= 0.0634, \quad \text{pr}(X = 1, Y^{(1)} = 2, Y^{(2)} = 2) = 0.4426, \\
    \text{pr}(X = 2, Y^{(1)} = 1, Y^{(2)} = 1) &= 0.0546, \quad \text{pr}(X = 2, Y^{(1)} = 1, Y^{(2)} = 2) = 0.0954, \\
    \text{pr}(X = 2, Y^{(1)} = 2, Y^{(2)} = 1) &= 0.0954, \quad \text{pr}(X = 2, Y^{(1)} = 2, Y^{(2)} = 2) = 0.1746.
\end{align*}
\]

Given the observed joint distribution of \((X, Y^{(1)}, Y^{(2)})\), the values of \(\text{pr}(x), \text{pr}(y^{(1)} \mid x), \text{pr}(y^{(2)} \mid x), \text{pr}(y^{(1)}, y^{(2)} \mid x)\) are also determined. These values, also denoted by \(\text{pr}(\cdot)\), will be used to construct another set of parameters which share the same observed distribution as the case above.

If we fix \(X = 1\), the conditional independence \(Y^{(1)} \perp Y^{(2)} \mid (U, X = 1)\) enables us to construct 3 equations by equating the observed distribution with expression of unknown parameters, i.e.,
\[
\begin{align*}
    \text{pr}(Y^{(1)} = 1, Y^{(2)} = 1 \mid X = 1) &= \sum_u \text{pr}(Y^{(1)} = 1 \mid u, X = 1) \text{pr}(Y^{(2)} = 1 \mid u, X = 1) \text{pr}(u \mid X = 1), \\
    \text{pr}(Y^{(1)} = 1, Y^{(2)} = 2 \mid X = 1) &= \sum_u \text{pr}(Y^{(1)} = 1 \mid u, X = 1) \text{pr}(Y^{(2)} = 2 \mid u, X = 1) \text{pr}(u \mid X = 1), \\
    \text{pr}(Y^{(1)} = 2, Y^{(2)} = 1 \mid X = 1) &= \sum_u \text{pr}(Y^{(1)} = 2 \mid u, X = 1) \text{pr}(Y^{(2)} = 1 \mid u, X = 1) \text{pr}(u \mid X = 1).
\end{align*}
\]

There are 5 unknown parameters: \(\text{pr}(U = 1 \mid X = 1), \text{pr}(Y^{(1)} = 1 \mid U = 1, X = 1), \text{pr}(Y^{(1)} = 1 \mid U = 2, X = 1), \text{pr}(Y^{(2)} = 1 \mid U = 1, X = 1)\), and \(\text{pr}(Y^{(2)} = 1 \mid U = 2, X = 1)\).

We will show that we can set \(\text{pr}(U = 1 \mid X = 1)\) and \(\text{pr}(Y^{(1)} = 1 \mid U = 1, X = 1)\) arbitrarily, and solve the remaining 3 parameters from the 3 equations. This set of parameters is denoted by \(\text{pr}^*(\cdot)\) to distinguish them from the parameter values we set at the beginning.
Let \( \text{pr}^*(U = 1 \mid X = 1) = 0.3 \) and \( \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 1) = 0.4 \), then

\[
\text{pr}^*(Y^{(1)} = 1 \mid U = 2, X = 1) = \frac{\text{pr}(Y^{(1)} = 1 \mid X = 1) - \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 1)\text{pr}^*(U = 1 \mid X = 1)}{1 - \text{pr}^*(U = 1 \mid X = 1)}
\]

=0.0108,

\[
\text{pr}^*(Y^{(2)} = 1 \mid U = 2, X = 1)
\]

\[
= \frac{\text{pr}(Y^{(1)} = 1, Y^{(2)} = 1 \mid X = 1) - \text{pr}(Y^{(2)} = 1 \mid X = 1)\text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 1)}{\{\text{pr}^*(Y^{(1)} = 1 \mid U = 2, X = 1) - \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 1)\} \{1 - \text{pr}^*(U = 1 \mid X = 1)\}}
\]

=0.1203,

\[
\text{pr}^*(Y^{(2)} = 1 \mid U = 1, X = 1)
\]

\[
= \frac{\text{pr}(Y^{(2)} = 1 \mid X = 1) - \text{pr}^*(Y^{(2)} = 1 \mid U = 2, X = 1)\{1 - \text{pr}^*(U = 1 \mid X = 1)\}}{\text{pr}^*(U = 1 \mid X = 1)}
\]

=0.1447.

Now we repeat the process for \( X = 2 \). Let \( \text{pr}^*(U = 1 \mid X = 2) = 0.2 \) and \( \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 2) = 0.6 \), then

\[
\text{pr}^*(Y^{(1)} = 1 \mid U = 2, X = 2)
\]

\[
= \frac{\text{pr}(Y^{(1)} = 1 \mid X = 2) - \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 2)\text{pr}^*(U = 1 \mid X = 2)}{1 - \text{pr}^*(U = 1 \mid X = 2)}
\]

=0.2964,

\[
\text{pr}^*(Y^{(2)} = 1 \mid U = 2, X = 2)
\]

\[
= \frac{\text{pr}(Y^{(1)} = 1, Y^{(2)} = 1 \mid X = 2) - \text{pr}(Y^{(2)} = 1 \mid X = 2)\text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 2)}{\{\text{pr}^*(Y^{(1)} = 1 \mid U = 2, X = 2) - \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 2)\} \{1 - \text{pr}^*(U = 1 \mid X = 2)\}}
\]

=0.3471,

\[
\text{pr}^*(Y^{(2)} = 1 \mid U = 1, X = 2)
\]

\[
= \frac{\text{pr}(Y^{(2)} = 1 \mid X = 2) - \text{pr}^*(Y^{(2)} = 1 \mid U = 2, X = 2)\{1 - \text{pr}^*(U = 1 \mid X = 2)\}}{\text{pr}^*(U = 1 \mid X = 2)}
\]

=0.3975.
In addition, we have

\[ \text{pr}^*(U = 1) = \text{pr}^*(U = 1 \mid X = 1)\text{pr}(X = 1) + \text{pr}^*(U = 1 \mid X = 2)\text{pr}(X = 2) = 0.258, \]
\[ \text{pr}^*(X = 1 \mid U = 1) = \frac{\text{pr}^*(U = 1 \mid X = 1)\text{pr}(X = 1)}{\text{pr}^*(U = 1)} = 0.6744, \]
\[ \text{pr}^*(X = 1 \mid U = 2) = \frac{\text{pr}^*(U = 2 \mid X = 1)\text{pr}(X = 1)}{1 - \text{pr}^*(U = 1)} = 0.5472. \]

To sum up, the parameter values

\[ \text{pr}^*(U = 1) = 0.258, \quad \text{pr}^*(X = 1 \mid U = 1) = 0.6744, \quad \text{pr}^*(X = 1 \mid U = 2) = 0.5472, \]
\[ \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 1) = 0.4, \quad \text{pr}^*(Y^{(1)} = 1 \mid U = 2, X = 1) = 0.0108, \]
\[ \text{pr}^*(Y^{(1)} = 1 \mid U = 1, X = 2) = 0.6, \quad \text{pr}^*(Y^{(1)} = 1 \mid U = 2, X = 2) = 0.2964, \]
\[ \text{pr}^*(Y^{(2)} = 1 \mid U = 1, X = 1) = 0.1447, \quad \text{pr}^*(Y^{(2)} = 1 \mid U = 2, X = 1) = 0.1203, \]
\[ \text{pr}^*(Y^{(2)} = 1 \mid U = 1, X = 2) = 0.3975, \quad \text{pr}^*(Y^{(2)} = 1 \mid U = 2, X = 2) = 0.3471, \]

give the same joint distribution of \((X, Y^{(1)}, Y^{(2)})\) as the first set. However, these two sets of values indicate different causal effects. The first set gives \(\text{pr}\{Y^{(1)}(X = 1) = 1\} = 0.14\) and \(\text{pr}\{Y^{(1)}(X = 2) = 1\} = 0.34\), while the second set has \(\text{pr}^*\{Y^{(1)}(X = 1) = 1\} = 0.111\) and \(\text{pr}^*\{Y^{(1)}(X = 2) = 1\} = 0.375\).

The above construction of a counterexample can be generalized to the case where \(Y^{(1)}, Y^{(2)}, U\) all have \(k\) categories. In that case, we fix \(X\) at a certain \(x\), then we have \(k^2 - 1\) equations by equating the observed \(\text{pr}(y^{(1)}, y^{(2)} \mid x)\) with expression of parameters. The unknown parameters are \((k - 1)\) of \(\text{pr}(u \mid x)\), \(k(k - 1)\) of \(\text{pr}(y^{(1)} \mid u, x)\), and \(k(k - 1)\) of \(\text{pr}(y^{(2)} \mid u, x)\). So there is a total of \((k - 1)(2k + 1)\) parameters. This leaves \((k - 1)(2k + 1) - (k^2 - 1) = k(k - 1)\) degrees of freedom. Therefore, one can set \(k(k - 1)\) parameters freely. For example, we can first specify the values of \(\text{pr}^*(U = u \mid x)\), \(u = 1, \ldots, k - 1\), which takes \(k - 1\) degrees of freedom. Then, we assign the values of \(\text{pr}^*(Y^{(1)} = i \mid U = u, x)\), \(i = 1, \ldots, k - 1; u = 1, \ldots, k - 1\), which takes the remaining \((k - 1)^2\) degrees of freedom.

After we specify the values of the \(k(k-1)\) parameters, we can solve the remaining parameters. First, we solve \(\text{pr}^*(Y^{(1)} = i \mid U = k, x)\), \(i = 1, \ldots, k - 1\) from the following \(k - 1\) equations

\[ \text{pr}(Y^{(1)} = i \mid x) = \sum_{u=1}^{k} \text{pr}^*(Y^{(1)} = i \mid U = u, x)\text{pr}^*(U = u \mid x), \]

where \(i = 1, \ldots, k - 1\). We then solve \(\text{pr}^*(Y^{(2)} = j \mid U = u, x)\), \(j = 1, \ldots, k - 1; u = 1, \ldots, k\)
from the following $k(k-1)$ equations

$$
\text{pr}(Y^{(1)} = i, Y^{(2)} = j \mid x) = \sum_{u=1}^{k} \text{pr}^*(Y^{(1)} = i \mid U = u, x) \text{pr}^*(Y^{(2)} = j \mid U = u, x) \text{pr}^*(U = u \mid x),
$$

where $i = 1, \ldots, k$; $j = 1, \ldots, k - 1$. This is a linear system with $k(k-1)$ equations and $k(k-1)$ parameters. It has a unique solution if the coefficient matrix is invertible. We can repeat the process for each $x$, and get $\text{pr}^*(u) = \sum_x \text{pr}^*(u \mid x) \text{pr}(x)$ as well as $\text{pr}^*(x \mid u) = \text{pr}^*(u \mid x) \text{pr}(x) / \text{pr}(u)^*$. In this way, we can find a set of parameters denoted by $\text{pr}^*(\cdot)$ which is compatible with the observed distribution of $(X, Y^{(1)}, Y^{(2)})$, but gives different potential outcome distributions $\text{pr}\{y^{(1)}(x)\}$ and $\text{pr}\{y^{(2)}(x)\}$ from the true values.

### S3 Causal diagrams for related works

Figure S1 is a causal diagram for the model in Kuroki and Pearl (2014). Figure S2 is a causal diagram for the model in Hu (2008).

---

**Figure S1:** A causal diagram associated with the measurement error model of Kuroki and Pearl (2014). They identified the effect of exposure $X$ on outcome $Y$, with two proxies of the latent confounder $U$, denoted as $Z$ and $W$.

---

**Figure S2:** A causal diagram associated with the measurement error model of Hu (2008). They identified the effect of a latent exposure $X^*$ on outcome $Y$, with an exposure proxy $X$, an instrument $Z$ and measured independent variables $W$. 
In this section, we use an example to show that in the categorical model in Section 3, local identifiability may be achieved without Condition 2, and Condition 2 is then needed to achieve global identifiability.

We consider the binary model as in Section 3. Without Condition 2, it is possible that

Case I: \[ \text{pr}(y^{(1)} | U = 1, X = 1) < \text{pr}(y^{(1)} | U = 2, X = 1) \]
\[ \text{pr}(y^{(1)} | U = 1, X = 2) < \text{pr}(y^{(1)} | U = 2, X = 2) \]

or,

Case II: \[ \text{pr}(y^{(1)} | U = 1, X = 1) > \text{pr}(y^{(1)} | U = 2, X = 1) \]
\[ \text{pr}(y^{(1)} | U = 1, X = 2) > \text{pr}(y^{(1)} | U = 2, X = 2) \]

or,

Case III: \[ \text{pr}(y^{(1)} | U = 1, X = 1) < \text{pr}(y^{(1)} | U = 2, X = 1) \]
\[ \text{pr}(y^{(1)} | U = 1, X = 2) > \text{pr}(y^{(1)} | U = 2, X = 2) \]

or,

Case IV: \[ \text{pr}(y^{(1)} | U = 1, X = 1) > \text{pr}(y^{(1)} | U = 2, X = 1) \]
\[ \text{pr}(y^{(1)} | U = 1, X = 2) < \text{pr}(y^{(1)} | U = 2, X = 2) \].

Note that Case I and Case II are equivalent by permuting the labels of U, but they are not equivalent to Case III/IV.

More specifically, note that identification of causal effects is based on the following equations:

\[ \text{pr}(y^{(1)} | X = 1) = \text{pr}(y^{(1)} | U = 1, X = 1) \text{pr}(U = 1 | X = 1) \]
\[ + \text{pr}(y^{(1)} | U = 2, X = 1)\{1 - \text{pr}(U = 1 | X = 1)\} \]
\[ \text{pr}(y^{(1)} | X = 2) = \text{pr}(y^{(1)} | U = 1, X = 2) \text{pr}(U = 1 | X = 2) \]
\[ + \text{pr}(y^{(1)} | U = 2, X = 2)\{1 - \text{pr}(U = 1 | X = 2)\}. \] (S4)

Suppose that \{\text{pr}(y^{(1)} | U = 1, X = 1), \text{pr}(y^{(1)} | U = 2, X = 1)\} = \{0.1, 0.6\}, and
\{\text{pr}(y^{(1)} | U = 1, X = 2), \text{pr}(y^{(1)} | U = 2, X = 2)\} = \{0.4, 0.8\}, \text{pr}(y^{(1)} | X = 1) = \text{pr}(y^{(1)} | X = 2) = 0.5 \text{ and pr}(X = 1) = 0.5.
Then we have under Case I, (S4) becomes

\[ 0.5 = 0.1 \text{pr}(U = 1 \mid X = 1) + 0.6(1 - \text{pr}(U = 1 \mid X = 1)) \]

\[ 0.5 = 0.4 \text{pr}(U = 1 \mid X = 2) + 0.8(1 - \text{pr}(U = 1 \mid X = 2)), \]

so that \( \text{pr}(U = 1 \mid X = 1) = 0.2, \text{pr}(U = 1 \mid X = 2) = 0.75, \text{pr}(U = 1) = 0.2 \times 0.5 + 0.75 \times 0.5 = 0.475, \) and \( \text{pr}\{y(1)\} = \text{pr}(y(1) \mid X = 1, U = 1)\text{pr}(U = 1) + \text{pr}(y(1) \mid X = 1, U = 2)\text{pr}(U = 2) = 0.1 \times 0.475 + 0.6 \times 0.525 = 0.3625. \)

Under Case II, we have \( \text{pr}(U = 1 \mid X = 1) = 0.8, \text{pr}(U = 1 \mid X = 2) = 0.25, \text{pr}(U = 1) = 0.8 \times 0.5 + 0.25 \times 0.5 = 0.525, \) and \( \text{pr}\{y(1)\} = 0.6 \times 0.525 + 0.1 \times 0.475 = 0.3625. \)

Under Case III, we can obtain similarly that \( \text{pr}(U = 1 \mid X = 1) = 0.2, \text{pr}(U = 1 \mid X = 2) = 0.25, \text{pr}(U = 1) = 0.2 \times 0.5 + 0.25 \times 0.5 = 0.225, \) and \( \text{pr}\{y(1)\} = 0.1 \times 0.225 + 0.6 \times 0.775 = 0.4875. \)

Under case IV, we have \( \text{pr}(U = 1 \mid X = 1) = 0.8, \text{pr}(U = 1 \mid X = 2) = 0.75, \text{pr}(U = 1) = 0.8 \times 0.5 + 0.75 \times 0.5 = 0.775, \) and \( \text{pr}\{y(1)\} = 0.6 \times 0.775 + 0.1 \times 0.225 = 0.4875. \)

One can now see clearly that there are two possible values for \( \text{pr}\{y(1)\} \), so \( \text{pr}\{y(1)\} \) is only locally identifiable without Condition 2. Since only Case I and Case II are allowed by Condition 2, and these two cases have the same value for \( \text{pr}\{y(1)\} \), the global identifiability is achieved under Condition 2.

**S5 Non-parametric estimation under a categorical model**

In this part, we consider the estimation problem assuming that both the exposure \( X \) and parallel outcomes \( Y^{(1)}, Y^{(2)}, Y^{(3)} \) are categorical variables. Given identifiability, one may use the generalized least squares to estimate the probabilities \( \text{pr}(y^{(j)} \mid u, x), j = 1, 2, 3, \text{pr}(x \mid u) \) and \( \text{pr}(u) \) from equations

\[
\text{pr}(y^{(1)}, y^{(2)}, y^{(3)}, x) = \sum_u \text{pr}(y^{(1)} \mid u, x)\text{pr}(y^{(2)} \mid u, x)\text{pr}(y^{(3)} \mid u, x)\text{pr}(x \mid u)\text{pr}(u), \tag{S5}
\]

subject to the constraints that

\[
\sum_u \text{pr}(u) = 1, \sum_x \text{pr}(x \mid u) = 1, \sum_{y^{(j)}} \text{pr}(y^{(j)} \mid u, x) = 1, \text{pr}(y^{(1)} \mid u, x) \text{ is strictly increasing in } u. \tag{S6}
\]

The distributions \( \text{pr}\{y^{(j)}(x)\}, j = 1, 2, 3 \) can then be estimated using plug-in estimators.
Solving the generalized least squares problem in practice, however, can be challenging as the equations in (S5) are non-linear and involve a large number of parameters. As a result, the optimization problem may be very time-consuming and even worse, the results can be highly sensitive to the choice of starting values. To solve this problem, we propose to first follow the identification procedure in Section 3 and obtain initial plug-in estimates of the probabilities $\Pr(y^{(j)} | u, x), j = 1, 2, 3$, $\Pr(x | u)$ and $\Pr(u)$. This step is computationally very efficient. We then use these initial estimates as warm starts for the generalized least squares procedure. Our proposed estimating procedure is summarized in Algorithm 1.

Theorem S2 summarizes the main theoretical properties of our proposed estimating procedure. In particular, it shows that the starting values $\tilde{\Pr}(u), \tilde{\Pr}(x | u)$ and $\tilde{\Pr}(y^{(j)} | u, x)$ are root-n consistent for their population counterparts, and that the resulting estimator $\hat{\Pr}\{y^{(j)}(x)\}$ is consistent and asymptotically normal. A key step in the proof of the first statement is Lemma S2, which shows that there exists an analytic continuation of eigendecomposition to its neighborhood. The second statement follows from standard M-estimation theory.

**Theorem S2.** Under the conditions of Theorem 1 and standard regularity conditions, we have that if $X$ and $Y^{(1)}$ are also categorical variables, then

1. For each $u, x, y^{(j)}, j = 1, 2, 3$, $\tilde{\Pr}(u) - \Pr(u) = O_p(n^{-1/2})$, $\tilde{\Pr}(x | u) - \Pr(x | u) = O_p(n^{-1/2})$, $\tilde{\Pr}(y^{(j)} | u, x) - \Pr(y^{(j)} | u, x) = O_p(n^{-1/2})$;
2. For each $x, y^{(j)}, j = 1, 2, 3$, $\sqrt{n}(\hat{\Pr}\{y^{(j)}(x)\} - \Pr(y^{(j)}(x))) \rightarrow_d N(0, \sigma_{jx}^2)$, where $\sigma_{jx}^2$ is the asymptotic variance of $\hat{\Pr}\{y^{(j)}(x)\}$.

### S6 Simulations for binary models

In this part, the unmeasured confounder $U$ is simulated from a Bernoulli distribution taking value 1 with probability 0.65. Conditional on $U$, the exposure and outcomes are generated from the
1 For each $x$ and $y^{(1)}$:

1-1 Estimate $P(Y^{(2)}, Y^{(3)} \mid x)$ and $P(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)$ by their empirical estimates, denoted by $\overline{P}(Y^{(2)}, Y^{(3)} \mid x)$ and $\overline{P}(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)$;

1-2 Apply eigendecomposition on $\overline{P}(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)\overline{P}(Y^{(2)}, Y^{(3)} \mid x)^{-1}$, and use its eigenvalues and associated normalized eigenvectors (with $\ell_1$-norm 1) to construct $\tilde{P}_D(y^{(1)} \mid U, x)$, $\tilde{P}(Y^{(2)} \mid U, x)$ and $\tilde{P}(Y^{(3)} \mid U, x)$ subject to the constraints in (S6);

1-3 Estimate $P_D(U \mid x)$ by

$$\tilde{P}_D(U \mid x) = \tilde{P}(Y^{(2)} \mid U, x)^{-1}\overline{P}(Y^{(2)}, Y^{(3)} \mid x)\{\tilde{P}(Y^{(3)} \mid U, x)^r\}^{-1};$$  \hfill (S7)

2 Estimate $pr(u)$ and $pr(x \mid u)$ by

$$\tilde{pr}(u) = \sum_x \tilde{pr}(u \mid x)\overline{pr}(x), \quad \tilde{pr}(x \mid u) = \tilde{pr}(u \mid x)\overline{pr}(x) / \tilde{pr}(u),$$  \hfill (S8)

where $\overline{pr}(x)$ is the empirical estimate of $pr(x)$;

3 Find $\hat{pr}(u)$, $\hat{pr}(x \mid u)$ and $\hat{pr}(y^{(j)} \mid u, x)$ that minimize

$$\sum_{x, y^{(1)}, y^{(2)}, y^{(3)}} \left\{ pr(y^{(1)}, y^{(2)}, y^{(3)} \mid x)\overline{pr}(x) - \sum_u pr(y^{(1)} \mid u, x)pr(y^{(2)} \mid u, x)pr(y^{(3)} \mid u, x)pr(x \mid u)pr(u) \right\}^2$$

subject to the constraints in (S6), using $\tilde{pr}(u)$, $\tilde{pr}(x \mid u)$ and $\tilde{pr}(y^{(j)} \mid u, x)$ as starting values;

4 For each $x$, estimate the potential outcome distribution $pr\{y^{(j)}(x)\}$, $j = 1, 2, 3$ by

$$\hat{pr}\{y^{(j)}(x)\} = \sum_u \hat{pr}(y^{(j)} \mid u, x)\hat{pr}(u).$$

**Algorithm 1:** A generalized least squares procedure with warm starts for estimating the potential outcome distributions.
following models:

$$(\text{pr}(X = 1 \mid U = u) : u = 1, 2) = (0.4, 0.7),$$

$$(\text{pr}(Y^{(1)} = 1 \mid U = u, X = x) : u = 1, 2, x = 1, 2) = \begin{pmatrix} 0.4 & 0.3 \\ 0.7 & 0.6 \end{pmatrix},$$

$$(\text{pr}(Y^{(2)} = 1 \mid U = u, X = x) : u = 1, 2, x = 1, 2) = \begin{pmatrix} 0.25 & 0.45 \\ 0.45 & 0.75 \end{pmatrix}, \quad (S9)$$

$$(\text{pr}(Y^{(3)} = 1 \mid U = u, X = x) : u = 1, 2, x = 1, 2) = \begin{pmatrix} 0.15 & 0.25 \\ 0.45 & 0.65 \end{pmatrix},$$

where for $j = 1, 2, 3$, $\text{pr}(Y^{(j)} = 1 \mid U = u, X = x)$ is given by the $(u, x)$-th element of the corresponding matrix. In addition to Algorithm 1, we also implement two comparison methods for estimating $\text{pr}\{y^{(j)}(x)\}$, $j = 1, 2, 3$, $x = 1, 2$: (1) Crude: empirical conditional probabilities $\tilde{\text{pr}}(y^{(j)} \mid x)$; (2) Random start: instead of the warm starting values $\tilde{\text{pr}}(u)$, $\tilde{\text{pr}}(x \mid u)$ and $\tilde{\text{pr}}(y^{(j)} \mid u, x)$, we use random starting values from a uniform distribution on $[0, 1]$ in Step 3 of Algorithm 1. All simulation results are based on 1000 Monte-Carlo runs, each with a sample size $n = 1000$. One can see from Table S1 that the crude estimator is severely biased, while using the warm starts as proposed substantially improves the performance of the generalized least squares procedure.

Table S1: Bias×100 (Standard deviation ×100) of various estimators for estimating the potential outcome distributions. The sample size is 1000

| potential outcome | crude   | random start | warm start |
|-------------------|---------|--------------|------------|
| $\text{pr}\{Y^{(1)}(X = 1) = 1\}$ | -4.02 (2.21) | -2.71 (11.8) | -1.11 (10.4) |
| $\text{pr}\{Y^{(1)}(X = 2) = 1\}$ | 4.06 (2.22) | 7.43 (12.0) | 2.32 (9.11) |
| $\text{pr}\{Y^{(2)}(X = 1) = 1\}$ | -2.75 (1.97) | 19.2 (22.0) | 0.87 (9.56) |
| $\text{pr}\{Y^{(2)}(X = 2) = 1\}$ | 4.08 (2.24) | 5.78 (9.63) | 1.84 (8.85) |
| $\text{pr}\{Y^{(3)}(X = 1) = 1\}$ | -4.16 (1.89) | 24.5 (22.0) | 0.02 (10.5) |
| $\text{pr}\{Y^{(3)}(X = 2) = 1\}$ | 5.47 (2.23) | 11.3 (9.36) | 3.19 (10.9) |
Table S2: Bias×100 (Standard deviation ×100) of the proposed estimator for estimating the average causal effect \( \text{pr}\{Y^{(j)}(X = 2) = 1\} - \text{pr}\{Y^{(j)}(X = 1) = 1\}, j = 1, 2, 3\), when \( X \) and \( Y^{(3)} \) may be confounded by variable \( W \). The sample size is 1000

| \((\alpha_X^X, \alpha_3^W)\) | \(Y^{(1)}\) | \(Y^{(2)}\) | \(Y^{(3)}\) |
|----------------------|----------------|----------------|----------------|
| (0, 0)               | 4.8 (18)       | 1.4 (15)       | 3.6 (17)       |
| (1, 1)               | 6.3 (20)       | 1.6 (15)       | 7.9 (16)       |
| (2, 2)               | 6.6 (23)       | -0.04 (17)     | 20.7 (17)      |
| (3, 3)               | 5.1 (24)       | 0.7 (19)       | 36.1 (19)      |
| (4, 4)               | 2.6 (25)       | -1.4 (21)      | 46.5 (20)      |

S7  **Explanation on linear Gaussian structural equation models violating Condition S4**

Consider the following linear Gaussian structural equation models with three parallel outcomes,

\[
X = \alpha_X U + \epsilon_X, \quad Y^{(j)} = \alpha_j U + \beta_j X + \epsilon_j, \quad j = 1, 2, 3,
\]

where \((U, \epsilon_X, \epsilon_1, \epsilon_2, \epsilon_3)^T \sim \mathcal{N}(0, \text{diag}(1, \sigma_X^2, \sigma_1^2, \sigma_2^2, \sigma_3^2))\). We will show that in this case, Condition S4 is violated.

The conditional distribution of \(Y^{(2)}\) given \((U, X) = (u, x)\) is a Gaussian distribution with mean \(\alpha_2 u + \beta_2 x\) and variance \(\sigma_2^2\). Since a Gaussian distribution is totally characterized by its mean and variance, for any continuous functional \(M\), we have \(M\{f(Y^{(2)} \mid U, X)(u, x)\} = \tilde{q}(\alpha_2 u + \beta_2 x, \sigma_2^2) = q(\alpha_2 u + \beta_2 x)\), where \(\tilde{q}, q\) are continuous functions, and the second equality is by the fact \(\sigma_2^2\) is a constant. We want to show if \(g \circ q(\alpha_2 u + \beta_2 x) = h_1(u) + h_2(x)\), then \(h_1(u)\) is a linear function.

Let \(f = g \circ q\), then \(f\) is also continuous by the continuity of \(g\) and \(q\). We have \(f(\alpha_2 u + \beta_2 x) = h_1(u) + h_2(x)\). Let \(x = 0\), we have \(f(\alpha_2 u) = h_1(u) + h_2(0)\). Let \(u = 0\), we have \(f(\beta_2 x) = h_2(x) + h_1(0)\). Let \(u = x = 0\), we have \(f(0) = h_1(0) + h_2(0)\).

Therefore, \(f(\alpha_2 u) + f(\beta_2 x) - f(0) = h_1(u) + h_2(x)\). Let \(\tilde{f}(t) = f(t) - f(0)\), then \(\tilde{f}(0) = 0\), \(\tilde{f}(\alpha_2 u + \beta_2 x) = f(\alpha_2 u + \beta_2 x) - f(0) = h_1(u) + h_2(x) - f(0) = f(\alpha_2 u) + f(\beta_2 x) - 2f(0) = \tilde{f}(\alpha_2 u) + \tilde{f}(\beta_2 x)\). By Lemma S3, \(\tilde{f}(t) = ct\), where \(c \in \mathbb{R}\) is a constant. So \(f(t) = ct + f(0)\) and \(f(\alpha_2 u + \beta_2 x) = c(\alpha_2 u + \beta_2 x) + f(0) = c\alpha_2 u + c\beta_2 x + f(0)\). By comparing with \(f(\alpha_2 u + \beta_2 x) = h_1(u) + h_2(x)\), we conclude \(h_1\) is a linear function of the form \(h_1(u) = au + b\). If \(a = 0\), then \(h_1\) is...
not one-to-one; if \( a \neq 0 \), then \( h_1(u) \) is neither bounded from above nor below when \( U \sim \mathcal{N}(0, 1) \). Both cases violate Condition S4.

### S8 Derivation of the threshold in (12)

In (12), we need to set the threshold \( \delta \). Under the high dimensional framework, when distinguishing non-zero signals from noises, the threshold \( \delta \) is often set as \( \delta = \sqrt{2n^{-1} \log(p) \sigma^2} \), where \( \sigma^2 \) is the error variance (Donoho and Johnstone, 1994).

In our case, we consider the factor model

\[
Y = \Gamma F + \varepsilon,
\]

i.e., model (9) in the main paper. Recall in this model, we have (i) \( F \perp \perp \varepsilon \); (ii) \( E(F) = 0, \text{Cov}(F) = I_{r+1} \); (iii) \( E(\varepsilon) = 0, \text{Cov}(\varepsilon) = \Lambda = \text{diag}(\sigma^2_1, \ldots, \sigma^2_p) \). Let \( \Sigma = \text{Cov}(Y) \) and \( \Lambda = \text{Cov}(\varepsilon) \), one can write \( \Sigma = \Gamma \Gamma^\top + \Lambda \).

The error term in model (S10) is a \( p \)-dimensional vector \( \varepsilon \), and each component has a variance \( \sigma^2_j \) for \( j = 1, \ldots, p \). Thus, we choose \( \sigma^2 = p^{-1} \sum_{j=1}^p \sigma^2_j \).

In practice, \( \sigma^2 \) is unknown, and needs to be estimated. This boils down to estimating \( \Sigma \), \( \Gamma \), and \( \Lambda \). Denote \( \widehat{\Sigma} \) the sample covariance of the \( (Y_i^{(1)}, \ldots, Y_i^{(p)}) \), \( i = 1, \ldots, n \). The eigen decomposition of \( \widehat{\Sigma} \) is

\[
\widehat{\Sigma} = \sum_{j=1}^p \widehat{\lambda}_j \widehat{v}_j \widehat{v}_j^\top,
\]

where \( \widehat{\lambda}_j \)'s are the eigenvalues and \( \widehat{v}_j \)'s are orthonormal eigenvectors. For the loading matrix \( \Gamma \), we use the principal component method to estimate it. Specifically, we have \( \widehat{\Gamma} \widehat{\Gamma}^\top = \sum_{j=1}^{\hat{r}+1} \widehat{\lambda}_j \widehat{v}_j \widehat{v}_j^\top \), where \( \hat{r} + 1 \) is the estimated number of factors. And the estimate of \( \Lambda \) is taken as \( \widehat{\Lambda} = \text{diag}(\widehat{\Sigma} - \widehat{\Gamma} \widehat{\Gamma}^\top) \).

Therefore, we can obtain \( \widehat{\sigma}^2 \) by

\[
\widehat{\sigma}^2 = p^{-1} \sum_{j=1}^p \widehat{\sigma}^2_j = p^{-1} \text{tr}(\widehat{\Lambda}) = p^{-1} \text{tr}(\widehat{\Sigma} - \widehat{\Gamma} \widehat{\Gamma}^\top) = p^{-1} \text{tr}(\sum_{j=\hat{r}+2}^p \widehat{\lambda}_j \widehat{v}_j \widehat{v}_j^\top)
\]

\[
=p^{-1} \sum_{j=\hat{r}+2}^p \text{tr}(\widehat{\lambda}_j \widehat{v}_j \widehat{v}_j^\top) = p^{-1} \sum_{j=\hat{r}+2}^p \widehat{\lambda}_j \text{tr}(\widehat{v}_j \widehat{v}_j^\top) = p^{-1} \sum_{j=\hat{r}+2}^p \widehat{\lambda}_j.
\]
S9 Reformulation of (12) to a mixed integer programming problem

We assume there exists a known scalar $M > 0$ such that $M \geq \|y^*\|_{\infty}$ for an optimal solution $y^*$ to (12). By introducing binary variables $\zeta = (\zeta_1, \ldots, \zeta_p)^T \in \{0, 1\}^p$ with $\zeta_j = \mathbb{I}(y_j > \delta)$, the problem (12) can be reformulated as the following mixed integer programming problem (Feng et al., 2018)

$$\min_{w, y, \zeta} \quad 1^T_p \zeta = \sum_{j=1}^{p} \zeta_j$$
$$\text{subject to} \quad w^T w = 1, \quad y = \hat{w}, \quad -M \zeta - \delta \leq y \leq M \zeta + \delta, \quad \zeta \in \{0, 1\}^p.$$  

S10 Description of outcomes in the real data example

In Table S3, we include the detailed descriptions of each outcome used in Section 6.

S11 Proof of Theorem 1

Proof. Condition 2 allows us to rearrange $\{u_1, \ldots, u_k\}$ in ascending order in terms of $\text{pr}(y^{(1)} | u_i, x)$ for fixed $y^{(1)}$ and $x$, and relabel it as $\{u_{(1)}, \ldots, u_{(k)}\}$. By Assumption 3, we have

$$\text{pr}(y^{(2)}, y^{(3)} | x) = \sum_{i=1}^{k} \text{pr}(y^{(2)} | u_{(i)}, x) \text{pr}(y^{(3)} | u_{(i)}, x) \text{pr}(u_{(i)} | x),$$

(S11)

$$\text{pr}(y^{(1)}, y^{(2)}, y^{(3)} | x) = \sum_{i=1}^{k} \text{pr}(y^{(1)} | u_{(i)}, x) \text{pr}(y^{(2)} | u_{(i)}, x) \text{pr}(y^{(3)} | u_{(i)}, x) \text{pr}(u_{(i)} | x).$$

(S12)

Let $P(Y^{(2)}, Y^{(3)} | x) = (\text{pr}(Y^{(2)} = i, Y^{(3)} = j | X = x))_{k \times k}$ be a $k$ by $k$ matrix whose $(i, j)$-th element is given by $\text{pr}(Y^{(2)} = i, Y^{(3)} = j | X = x)$. Similarly, we let $P(Y^{(2)} | U, x) = (\text{pr}(Y^{(2)} = i | U = u_{(j)}, X = x))_{k \times k}$, $P(Y^{(3)} | U, x) = (\text{pr}(Y^{(3)} = i | U = u_{(j)}, X = x))_{k \times k}$. Additionally, $P_D(y^{(1)} | U, x) = \text{diag}(\text{pr}(y^{(1)} | U = u_{(1)}, x), \ldots, \text{pr}(y^{(1)} | U = u_{(k)}, x))$, $P_D(U | x) = \text{diag}(\text{pr}(U = u_{(1)} | x), \ldots, \text{pr}(U = u_{(k)} | x))$ are $k \times k$ diagonal matrices. Then in the form of matrix multiplication, (S11) and (S12) can be written as

$$P(Y^{(2)}, Y^{(3)} | x) = P(Y^{(2)} | U, x) P_D(U | x) P(Y^{(3)} | U, x)^T,$$

(S13)

$$P(y^{(1)}, Y^{(2)}, Y^{(3)} | x) = P(Y^{(2)} | U, x) P_D(U | x) P_D(y^{(1)} | U, x) P(Y^{(3)} | U, x)^T.$$  

(S14)
Table S3: Description of outcomes in the real data application

| Acrylamide          | a toxic and potentially cancer-causing chemical which is formed in high amounts in many types of food prepared/cooked at high temperatures. |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Vitamin C           | a water-soluble vitamin. It is an essential nutrient involved in the repair of tissue, the formation of collagen, and the enzymatic production of certain neurotransmitters. |
| Cadmium             | a chemical element with the symbol Cd and atomic number 48.                                                                       |
| Urinary thiocyanate | a biomarker of cyanide exposure from tobacco smoke or diet. It can disrupt thyroid function by competitively inhibiting iodide uptake. |
| Retinyl palmitate   | the most abundant form of vitamin A storage in animals.                                                                             |
| Osmolality          | plasma osmolality measures the body’s electrolyte–water balance.                                                                  |
| Vitamin D           | a group of fat-soluble secosteroids responsible for increasing intestinal absorption of calcium, magnesium, and phosphate, and many other biological effects. |
| Vitamin B12         | a water-soluble vitamin involved in metabolism. It is one of eight B vitamins. It is required by animals, which use it as a cofactor in DNA synthesis, in both fatty acid and amino acid metabolism. |
| Mercury, total      | a chemical element with the symbol Hg and atomic number 80                                                                        |
| Enterolactone       | a organic compound classified as an enterolignan. It is formed by the action of intestinal bacteria on plant lignan precursors present in the diet. |
| C-reactive protein  | an annular pentameric protein found in blood plasma, whose circulating concentrations rise in response to inflammation.          |
| Total bilirubin     | a red-orange compound that occurs in the normal catabolic pathway that breaks down heme in vertebrates. Elevated levels may indicate liver damage or disease. |
| Perchlorate, urine  | a polyatomic anion that can disrupt thyroid function by competitively inhibiting iodide uptake.                                  |
| Creatinine, urine   | a breakdown product of creatine phosphate from muscle and protein metabolism. High levels of creatinine can indicate that the kidney is not working well. |
| Glucose, serum      | the amount of glucose in the fluid portion of the blood. It is the simplest and most direct single test available to test for diabetes. |
| Iron, refrigerated  | used in the diagnosis and treatment of diseases such as iron deficiency anemia, chronic renal disease, and hemochromatosis         |
| Serum total IgE anti-body | the amount of IgE antibodies in the blood and is the sum of all the forms of IgE. Total IgE testing is used to help diagnose some health conditions including certain types of infections and immune disorders. |
| 2,4,5-trichlorophenol | metabolite of several organochlorine pesticides.                                                                                  |
| Mono-2-ethyl-5-carboxypentyl phthalate | used extensively as plasticizers in a wide range of applications such as children’s toys, food packaging, personal care products, and medical supplies. |
| Urinary bisphenol A | used in the manufacture of polycarbonate plastics and epoxy resins, which have been used in baby bottles, as protective coatings on food containers, and as composites and sealants in dentistry. |
By Condition 1, \( P(Y^{(2)}, Y^{(3)} | x) \) is invertible. Then from (S13) and (S14) we have
\[
P(y^{(1)}, Y^{(2)}, Y^{(3)} | x)x^{-1} = P(Y^{(2)} | U, x)P_D(y^{(1)} | U, x)P(Y^{(2)} | U, x)^{-1}.
\]
Using the technique of eigendecomposition, \( P_D(y^{(1)} | U, x) \) and \( P(Y^{(2)} | U, x) \) can be recovered in the sense that the eigenvalues of \( P(y^{(1)}, Y^{(2)}, Y^{(3)} | x)xP(Y^{(2)}, Y^{(3)} | x)^{-1} \) in descending order are the diagonal elements of \( P_D(y^{(1)} | U, x) \), i.e. \( \text{pr}(y^{(1)} | U = u_{(i)}, x) \), and the corresponding eigenvectors scaled to \( \ell_1 \)-norm one are columns of \( P(Y^{(2)} | U, x) \). By symmetry, \( P(Y^{(3)} | U, x) \) can be identified likewise. Therefore, following (S13),
\[
P_D(U | x) = P(Y^{(2)} | U, x)^{-1}P(Y^{(2)}, Y^{(3)} | x)\{P(Y^{(3)} | U, x)^T\}^{-1} \text{ is identified.}
\]
We can then identify \( \text{pr}(u_{(i)}) = E_X\{\text{pr}(u_{(i)} | x)\} \). Finally, the potential outcome distributions can be identified from
\[
\text{pr}\{y^{(j)}(x)\} = \sum_{i=1}^{k} \text{pr}(y^{(j)} | x, u_{(i)})\text{pr}(u_{(i)}), j = 1, 2, 3.
\]

\[\square\]

**S12 Proof of Lemma 1**

*Proof.* Recall that
\[
\Gamma_{pX(r+1)}^* = \begin{pmatrix}
\sigma_{cX}\beta_1 & \alpha_{11} + \alpha_{X1}\beta_1 & \cdots & \alpha_{1r} + \alpha_{Xr}\beta_1 \\
\sigma_{cX}\beta_2 & \alpha_{21} + \alpha_{X1}\beta_2 & \cdots & \alpha_{2r} + \alpha_{Xr}\beta_2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{cX}\beta_p & \alpha_{p1} + \alpha_{X1}\beta_p & \cdots & \alpha_{pr} + \alpha_{Xr}\beta_p
\end{pmatrix}.
\]

By Condition 4(i), \( \|\Gamma_{s}^*\|_{0} = s \). Condition 4(ii) ensures that the minimum of \( \| (\Gamma^* R)_{s} \|_{0} \) is \( s \), where \( R \) is any \( (r+1) \times (r+1) \) matrix with \( R^t R = I \). We prove this by contradiction.

If the minimum is \( s - 1 \) or smaller and it is achieved for some \( R^* \), then there are at least \( p - s + 1 \) zeros in \( (\Gamma^* R^*)_{s} \). If \( \| (\Gamma^* R^*)_{s} \|_{0} = s - 1 \), then the first column of \( \Gamma^* R^* \) has \( p - s + 1 \) zeros. We let \( S = \{ j : (\Gamma^* R^*)_{j1} = 0 \} \), hence \( |S| = p - s + 1 > r + 1 \) by Condition 4(i). Let \( \Gamma_S^* \) be the \( |S| \times (r+1) \) submatrix of \( \Gamma^* \) consisting of rows indexed by \( S \), then \( \Gamma_S^*, R^* \) is not full rank because its first column is all zero. Since the rotation matrix \( R^* \) has rank \( r + 1 \), we have \( \text{rank}(\Gamma_S^*, R^*) = \text{rank}(\Gamma_S^*) \). We then conclude \( \Gamma_S^* \) is not full rank. This violates Condition 4(ii), which implies any submatrix of \( \Gamma^* \) consisting of \( p - s + 1 \) rows has full rank \( r + 1 \). Contradiction!

So, the minimum of \( \| (\Gamma^* R)_{s} \|_{0} \) is \( s \).

Finally, Condition 4(ii) ensures when the minimum \( \ell_0 \)-norm is achieved, it must be that the
p − s 0’s appear at the rows with β_j = 0. Any other combination of p − s rows has full rank, thus the (p − s) × (r + 1) submatrix of Γ* can never have a column of 0’s after rotation.

\[ \Box \]

**S13 Proof of Theorem 2**

*Proof*. In Lemma 1, we have identified the set \( S_0 = \{ j : \beta_j = 0 \} \). Now we focus on \( \{ \ell : \beta_\ell \neq 0 \} \).

Taking conditional expectation of \( W \) on \( (X, Z^\ell) \), we have

\[
E(W \mid X, Z^\ell) = \Gamma_{S_0,-1}^* E(U \mid X, Z^\ell),
\]

where \( \Gamma_{S_0,-1}^* \) is the \((p − s) \times r\) submatrix of \( \Gamma^* \) with row indices in \( S_0 \) and column indices in \( \{2, 3, \ldots, r + 1\} \). By Condition 4(ii), \( \Gamma_{S_0,-1}^* \) has rank \( r \). Therefore, there exists \( W^* = (W_1, \ldots, W_t) \in \mathbb{R}^r \) such that

\[
E(W^* \mid X, Z^\ell) = \alpha_{W^*} E(U \mid X, Z^\ell),
\]

where \( \alpha_{W^*} \) is a full rank \( r \times r \) submatrix of \( \Gamma_{S_0,-1}^* \). Taking conditional expectation of \( Y^{(\ell)} \) on \( (X, Z^\ell) \), we have

\[
E(Y^{(\ell)} \mid X, Z^\ell) = \beta_\ell X + \alpha_\ell^T E(U \mid X, Z^\ell)
= \beta_\ell X + \alpha_\ell^T \alpha_{W^*}^{-1} E(W^* \mid X, Z^\ell).
\]

(S15)

Here \( \alpha_\ell, E(U \mid X, Z^\ell), E(W \mid X, Z^\ell) \) are \( r \)-vectors. By Condition 5, \( X \) is not a linear combination of \( E(W \mid X, Z^\ell) \). So \( X \) is not a linear combination of \( E(W^* \mid X, Z^\ell) \) as \( W^* \) is a sub-vector of \( W \). Therefore, \( \beta_\ell \) can be identified through the regression in (S15).

\[ \Box \]

**S14 Extension of (6)-(7) to include common latent mediators among outcomes**

We show in this section that our model in (6)-(7) can be extended to accommodate common latent mediators among outcomes. Figure S3 provides a causal diagram in which there are common mediators for the effect of \( X \) on \( Y^{(j)}, j = 1, 2, 3, \ldots \). In this case, the assumption that

\[
Y^{(1)} \perp \!\!\!\!\perp Y^{(2)} \perp \!\!\!\!\perp Y^{(3)} \mid (U, X)
\]

(S16)
Figure S3: A causal diagram associated with the parallel-outcome framework when \( p = 3 \) in the presence of unmeasured confounders and mediators.

fails to hold. However, we note that under the causal diagram in Figure S3, we have instead the assumption that

\[
Y^{(1)} \perp \perp Y^{(2)} \perp \perp Y^{(3)} \mid (U, X, M).
\]  

(S17)

Note that (S17) takes the same form as (S16) except that \( U \) is replaced with \( U^* \equiv (U, M) \).

Motivated by this, we shall now show that under the linear structural equation model described in Section 4.1, we can apply our estimation procedures, and still obtain the total effects of \( X \) on \( Y^{(j)} \). Specifically, suppose we have a latent confounder \( U \in \mathbb{R}^r \) and a latent mediator \( M \in \mathbb{R}^s \). Similar to Section 4.1, we assume the following linear structural equation models:

\[
X = \alpha_X^T U + \epsilon_X, \tag{S18}
\]

\[
M = \beta_M X + \epsilon_M, \tag{S19}
\]

\[
Y^{(j)} = \alpha_j^T U + \beta_j^* X + \gamma_j^T M + \epsilon_j. \tag{S20}
\]

Define \( \beta_j^* = \beta_j + \gamma_j^T \beta_M \), which characterizes the total effect of \( X \) on \( Y^{(j)} \). By plugging (S18) and (S19) into (S20), we have

\[
Y^{(j)} = (\alpha_j + \alpha_X^T \beta_j^*) U + \gamma_j^T \epsilon_M + \beta_j^* \epsilon_X + \epsilon_j.
\]

This has a similar form as equation (8) in the paper, i.e., the above model is also an orthogonal factor model with latent confounders \( U \) and \( \epsilon_M \) of dimension \( r + s \). By assuming similar conditions as Conditions 3-5 in our main paper but replacing \( \beta_j \) with \( \beta_j^* \), we may still identify and estimate the causal parameters \( \beta_j^* \).
S15 Lemmas

Lemma S1. Under Assumptions 1, 3, if for all $x$ and $j = 1, 2, 3$, $U \not\perp Y^{(j)} \mid X = x$, and $U, Y^{(2)}, Y^{(3)}$ are binary, then for all $x$, $P(Y^{(2)}, Y^{(3)} \mid x)$ is of full rank.

Lemma S2. Let $\lambda_0$ be an algebraically simple eigenvalue of a matrix $M_0 \in M_k(\mathbb{R})$, where $M_k(\mathbb{R})$ is the set of matrices of size $k \times k$ with entries in $\mathbb{R}$. Then there exists an open neighborhood $\mathcal{M}$ of $M_0$ in $M_k(\mathbb{R})$, and two analytic functions

$$M \mapsto \Lambda(M), \quad M \mapsto X(M)$$

(S21)

over $\mathcal{M}$, such that

- $\Lambda(M)$ is an eigenvalue of $M$.
- $X(M)$ is an eigenvector, associated with $\Lambda(M)$.
- $\Lambda(M_0) = \lambda_0$.
- $X(M_0) = X_0$, where $X_0$ is a normalized eigenvector pertaining to $\lambda_0$ with $1_k^T X_0 = 1$; here $1_k$ denotes a $k$-vector whose elements are all 1’s.

Lemma S3. If $f$ is a continuous function, and $f(x + y) = f(x) + f(y)$, then $f(x) = cx$, where $c \in \mathbb{R}$ is a constant.

Lemma S4. For $G = \mathcal{L}^1, \mathcal{L}_{\text{bnd}}^1$, the adjoint of $T_{f_{a,b}(a,b)}$, denoted by $T_{f_{a,b}(a,b)}^*$, is an operator from $G^*(A)$ to $G^*(B)$ with the form

$$(T_{f_{a,b}(a,b)}^* g^*) (b) = \int f(a,b) g^*(a) \, da.$$  

(S22)

Lemma S5. If $T_{f_{a,b}(a,b)}$ and $T_{f_{a,b}(a,b)}$ are injective, then $T_{f_{a,b}(a,b)}^{-1}$ is densely defined over $G(A)$ for $G = \mathcal{L}^1, \mathcal{L}_{\text{bnd}}^1$.

Lemma S6. If $k(a, b; \theta) = f_{a|b}(a \mid b; \theta) = \psi(b; \theta) h(a; \theta) \exp\{\eta(b; \theta)^T \lambda(a; \theta)\}$ satisfies: (i) $\psi(b; \theta) > 0$, (ii) $\eta(b; \theta)$ is one-to-one in $b$, (iii) the range of $\lambda(a; \theta)$ contains a k-dimensional open set, where $k$ is the dimension of $\eta(b; \theta)$. Then $k(a, b; \theta)$ is complete in $a$. 

22
S16 Proof of Lemmas

S16.1 Proof of Lemma S1

Proof. For \( j = 2, 3 \), \( P(Y^{(j)} \mid U, x) \) is full rank if and only if

\[
\frac{\Pr(Y^{(j)} = 1 \mid U = 1, x)}{\Pr(Y^{(j)} = 2 \mid U = 1, x)} \neq \frac{\Pr(Y^{(j)} = 1 \mid U = 2, x)}{\Pr(Y^{(j)} = 2 \mid U = 2, x)},
\]

which holds if and only if \( U \not\perp Y^{(j)} \mid X = x \). By \( U \not\perp Y^{(j)} \mid X = x \) for \( j = 1, 2, 3 \), we also have that \( U \) is not a constant variable so that \( \Pr(U = u) > 0 \) for \( u = 1, 2 \), implying that \( P_D(U \mid x) \) is full rank. Equation (3) then implies that \( P(Y^{(2)}, Y^{(3)} \mid x) \) is full rank.

\( \square \)

S16.2 Proof of Lemma S2

Proof. This is a minor modification of the proof in Serre (2010, p.90, Theorem 5.3).

Let \( X_0 \) be an eigenvector of \( M_0 \) associated with \( \lambda_0 \). We normalize \( X_0 \) such that \( 1_k^T X_0 = 1 \). It is easy to see that \( \lambda_0 \) is also a simple eigenvalue of \( M_0^T \). Due to Proposition 3.15 in Serre (2010), an eigenvector \( Y_0 \) of \( M_0^T \) associated with \( \lambda_0 \) satisfies \( Y_0^T X_0 \neq 0 \). We also normalize \( Y_0 \) in such a way that \( 1_k^T Y_0 = 1 \), and denote \( a = Y_0^T X_0 \neq 0 \).

Define a polynomial function \( F \) over \( M_k(\mathbb{R}) \times \mathbb{R} \times \mathbb{R}^k \), with values in \( \mathbb{R} \times \mathbb{R}^k \), by

\[
F(M, \lambda, x) = (1_k^T x - 1, Mx - \lambda x).
\]

We have \( F(M_0, \lambda_0, X_0) = (0, 0) \). The differential of \( F \) with respect to \( (\lambda, x) \), at the base point \( (M_0, \lambda_0, X_0) \), is the linear map

\[
(\mu, y) \mapsto (1_k^T y, (M_0 - \lambda_0)y - \mu X_0).
\]

We first show that \( \delta \) is one-to-one. Let \( (\mu, y) \) be such that \( \delta(\mu, y) = (0, 0) \). Then \( \mu = \frac{1}{a} Y_0^T X_0 = \frac{1}{a} Y_0^T (M_0 - \lambda_0)y = \frac{1}{a} 0^T y = 0 \). Therefore, \( (M_0 - \lambda_0)y - \mu X_0 = 0 \) implies that \( (M_0 - \lambda_0)y = 0 \).

Inasmuch as \( \lambda_0 \) is simple, \( y \) is colinear to \( X_0 \); now the fact that \( 1_k^T y = 0 \) and \( 1_k^T X_0 = 1 \) yields \( y = 0 \).

Because \( \delta \) is a one-to-one endomorphism of \( \mathbb{R} \times \mathbb{R}^k \), it is an isomorphism. We may then apply the implicit function theorem to \( F \): there exists neighborhoods \( \mathcal{M}, \mathcal{V}, \) and \( \mathcal{W} \) and analytic functions \( (\Lambda, X) : \mathcal{M} \rightarrow \mathcal{V} \) such that

\[
\left( \begin{array}{c}
M, \lambda, x \in \mathcal{W} \\
F(M, \lambda, x) = (0, 0)
\end{array} \right) \iff \left( \begin{array}{c}
M \in \mathcal{M} \\
(\lambda, x) = (\Lambda(M), X(M))
\end{array} \right).
\]
Notice that $F = 0$ implies that $(\lambda, x)$ is an eigenpair of $M$ and $1_k^T x = 1$. Therefore, there exists an open neighborhood $\mathcal{M}$ of $M_0$ in $M_k(\mathbb{R})$, and two analytic functions $M \mapsto \Lambda(M)$, $M \mapsto X(M)$ over $\mathcal{M}$, such that

- $\Lambda(M)$ is an eigenvalue of $M$, $X(M)$ is an eigenvector associated with $\Lambda(M)$ normalized in a way such that $1_k^T X(M) = 1$.
- $\Lambda(M_0) = \lambda_0$, $X(M_0) = X_0$, $1_k^T X_0 = 1$.

$\square$

S16.3 Proof of Lemma S3

Proof. Let $y = 0$, we have $f(x) = f(x) + f(0)$, therefore $f(0) = 0$.

Let $y = x$, we have $f(2x) = 2f(x)$. By induction, we know $f(nx) = nf(x)$ for any positive integer $n$.

Let $y = -x$, we have $f(x) + f(-x) = f(x - x) = 0$, thus $f(-x) = -f(x)$. So $f(nx) = nf(x)$ for any integer $n$.

Since $f(x) = f(\frac{x}{2} + \frac{x}{2}) = f(\frac{x}{2}) + f(\frac{x}{2}) = 2f(\frac{x}{2})$, we conclude $f(\frac{x}{2}) = \frac{1}{2} f(x)$. By induction, we have $f(\frac{x}{n}) = \frac{1}{n} f(x)$ for any positive integer $n$. By the fact $f(-x) = -f(x)$, we have $f(\frac{x}{n}) = \frac{1}{n} f(x)$ for any integer $n$.

Then for any rational number $\frac{m}{n}$, where $m, n$ are integers, $f(\frac{m}{n}) = f(m \frac{1}{n}) = mf(\frac{1}{n}) = \frac{m}{n} f(1)$. Let $f(1) = c$, then $f(x) = cx$ for $x \in \mathbb{Q}$.

Let $g(x) = cx$ for $x \in \mathbb{R}$, then $f(x) = g(x)$ for $x \in \mathbb{Q}$. For any $x \in \mathbb{R}$, there exists a sequence of rational numbers $\{r_n\}$ such that $x = \lim_{n \to \infty} r_n$. Therefore,

$$f(x) = \lim_{n \to \infty} f(r_n) = \lim_{n \to \infty} g(r_n) = g(x),$$

where the first and third equality is by the continuity of $f$ and $g$.

So $f(x) = g(x) = cx$ for $x \in \mathbb{R}$.

$\square$
S16.4 Proof of Lemma S4

Proof. The adjoint $T^*$ of an operator $T \in \mathcal{L}(X, Y)$ is defined via
\[ \langle Tx, y^* \rangle = \langle x, T^*y^* \rangle \quad \text{or} \quad y^*T(x) = T^*y^*(x) \]
with $T^*: Y^* \to X^*$ and $y^* \in Y^*$. The duality is given by $\langle r, s \rangle = \int r(x)s(x) \, dx$ for $r \in \mathcal{G}(X)$ and $s \in \mathcal{G}^*(X)$ in our case. We can check that
\[
\langle T_{f_{a,b}}(a,b)g, h^* \rangle = \int \int f(a,b)g(b) \, dh^*(a) \, da
= \int g(b) \int f(a,b)h^*(a) \, da \, db
= \langle g, T_{f_{a,b}}^*(a,b)h^* \rangle
\]
where $g \in \mathcal{G}(B)$ and $h^* \in \mathcal{G}^*(A)$, so (S22) is correct. \qed

S16.5 Proof of Lemma S5

Proof. The operator $T_{f_{b|a}(b,a)}$ is defined via
\[
(T_{f_{b|a}(b,a)}g)(b) = \int f(b \mid a)g(a) \, da = \int \frac{f(a,b)}{f(a)}g(a) \, da = \int f(a,b)\frac{g(a)}{f(a)} \, da.
\]
By the injectivity assumption, if $(T_{f_{b|a}(b,a)}g)(a) = 0$, then $g(a) = 0$.

The operator $T_{f_{a,b}(a,b)}$ is defined via $(T_{f_{a,b}(a,b)}g)(a) = \int f(a,b)g(b) \, db$.

By Lemma S4, the adjoint of $T_{f_{a,b}(a,b)}$, the operator $T_{f_{a,b}(a,b)}^*$ from $\mathcal{G}^*(A)$ to $\mathcal{G}^*(B)$, is given by $(T_{f_{a,b}(a,b)}^*g^*)(b) = \int f(a,b)g^*(a) \, da$.

For any $h^* \in \mathcal{G}^*(A)$, let $\ell(a) = h^*(a)f(a)$ so that $\ell \in \mathcal{G}^2(A)$, then
\[
(T_{f_{b|a}(b,a)}\ell)(b) = \int f(b \mid a)\ell(a) \, da = \int f(a,b)\frac{\ell(a)}{f(a)} \, da = \int f(a,b)h^*(a) = 0
\]
implies $\ell = 0$. Since $f(a) \neq 0$, we must have $h^* = 0$. So the injectivity of $T_{f_{a,b}(a,b)}^*$ is proved.

By Lemma VI.2.8 in Dunford and Schwartz (1988), $\mathcal{R}(T_{f_{a,b}(a,b)})$ is the orthogonal complement of $\mathcal{N}(T_{f_{a,b}(a,b)}^*)$, the null space of $T_{f_{a,b}(a,b)}^*$. Since $T_{f_{a,b}(a,b)}^*$ is injective, $\mathcal{N}(T_{f_{a,b}(a,b)}^*) = \{0\}$. Hence, $\mathcal{R}(T_{f_{a,b}(a,b)}) = \mathcal{G}(A)$. Since $T_{f_{a,b}(a,b)}$ is injective by assumption, we have that $T_{f_{a,b}(a,b)}^{-1}$ is densely defined over $\mathcal{G}(A)$ for $\mathcal{G} = L^1, L^1_{\text{bnd}}$. \qed
S16.6 Proof of Lemma S6

Proof. For brevity of notation, we suppress $\theta$ in the following proof. Suppose that $\int g(b)k(a, b) \, db = 0$, which, in this setting, is

\[ h(a) \int \tilde{g}(b) \exp \{ \eta(b)^T \lambda(a) \} \, db = 0, \quad (S23) \]

where $\tilde{g}(b) = g(b)\psi(b)$. Since the mapping $b \mapsto \eta(b)$ is one-to-one, let $t = \eta(b)$ or equivalently $b = \eta^{-1}(t)$. Then (S23) becomes

\[ h(a) \int \tilde{g}\{\eta^{-1}(t)\}[\eta\{\eta^{-1}(t)\}]^{-1} \exp \{ \lambda(a)^T t \} \, dt = 0, \quad (S24) \]

where $\dot{\eta}(b) = \partial\eta(b)/\partial b$ and $[\dot{\eta}\{\eta^{-1}(t)\}]^{-1}$ is the Jacobian matrix. By making a translation of the parameter space one can assume without loss of generality that the range of $\lambda(a)$ contains the rectangle $I = [-m, m]^k$. Let $\tilde{g}\{\eta^{-1}(t)\}[\eta\{\eta^{-1}(t)\}]^{-1} = r(t) = r^+(t) - r^-(t)$ be such that

\[ \int r(t) \exp \{ \lambda(a)^T t \} \, dt = 0 \text{ for all } \lambda(a) \in I. \]

Then

\[ \int r^+(t) \exp \{ \lambda(a)^T t \} \, dt = \int r^-(t) \exp \{ \lambda(a)^T t \} \, dt. \]

Hence in particular let $\lambda(a) = 0$,

\[ \int r^+(t) \, dt = \int r^-(t) \, dt. \]

Dividing $r$ by a constant, one can take the common value of these two integrals to be 1, so that

\[ dP^+(t) = r^+(t) \, dt, \quad dP^-(t) = r^-(t) \, dt \]

are probability measures, and

\[ \int \exp \{ \lambda(a)^T t \} \, dP^+(t) = \int \exp \{ \lambda(a)^T t \} \, dP^-(t) \]

for all $\lambda(a) \in I$. From another point of view, these integrals can also be regarded as functions of the complex variables $\lambda_j(a) = \xi_j + i\delta_j z, j = 1, \ldots, k$, with real parts strictly between $-m$ and $m$. By Theorem 2.7.1 in Lehmann and Romano (2005), they are analytic functions of $\lambda_j(a)$ in the strip $R_j : -m < \xi_j < m, -\infty < \delta_j < \infty$ of the complex plane. For $\lambda_2(a), \ldots, \lambda_k(a)$ fixed, real, and between $-m$ and $m$, the equality of the integrals holds on the line segment $\{(\xi_1, \delta_1) : -m < \xi_1 < m, \delta_1 = 0\}$ and can therefore be extended to the strip $R_1$, in which the integrals are analytic. By induction the equality can be extended to the complex region $\{\lambda_1(a), \ldots, \lambda_k(a) : (\xi_j, \delta_j) \in R_j \text{ for } j = 1, \ldots, k\}$. It follows in particular that for all real
\( (\delta_1, \ldots, \delta_k), \)
\[
\int e^{i\sum \delta_j t_j} dP^+(t) = \int e^{i\sum \delta_j t_j} dP^-(t).
\]
These integrals are the characteristic functions of the distributions \( P^+ \) and \( P^- \) respectively, and by the uniqueness theorem for characteristic functions, the two distributions \( P^+ \) and \( P^- \) coincide. From the definition of these distributions it then follows \( r^+(t) = r^-(t) \) almost everywhere, and hence that \( r(t) = 0 \) almost everywhere. Since \( \tilde{\lambda} \{ \lambda_1^{-1}(t) \}^{-1} \) is not zero, it can hold only if \( \tilde{g}(b) \) is zero almost everywhere. Since \( \psi(b) \) is not zero, \( g(b) \) is zero almost everywhere. \( \square \)

S17 Proof of Theorem S1

*Proof.* The proof can be decomposed into four main steps.

Step 1: We show that a known operator constructed from observed distributions has a form of eigenvalue-eigenfunction decomposition, where the eigenvalues and eigenfunctions are conditional densities that involve unobserved variables.

Step 2: We prove the eigenvalue-eigenfunction decomposition is unique, in the sense that the eigenvalues and corresponding eigenspaces are identified.

Step 3: We show that the eigenspace corresponding to each eigenvalue is one dimensional. And the eigenfunction which forms the eigenspace is uniquely determined.

Step 4: We uncover the correspondence between each eigen-pair and a value of the unobserved confounder. Then we restore the mean potential outcomes.

Before providing detailed proofs for each step, we introduce some relevant math definitions and theorem.

- An integral operator denoted by \( T_{k(a,b)} \) is an operator mapping \( g \in \mathcal{G}(B) \) to \( T_{k(a,b)}g \in \mathcal{G}(A) \), defined as \( (T_{k(a,b)}g)(a) = \int_B k(a,b)g(b) \, db \), where \( k(a,b) \) is called the kernel of the operator.
• The *spectrum* of a bounded linear operator $T$ is the set of all $\lambda \in \mathbb{C}$ for which the operator $T - \lambda I$ does not have an inverse that is a bounded linear operator.

• A *quasi-nilpotent* operator is an operator whose spectrum is $\{0\}$.

• An operator $T$ is said to *commute* with another operator $A$ if $NA = AN$.

• A *projection-valued measure* assigns a projection operator to each set in a field. Here the field is the Borel $\sigma$-field. A projection $P$ is a linear operator such that $P^2 = P$. Just like a real-valued measure, we can do integration with respect to a projection-valued measure to get a new operator.

• (Theorem XV.4.5 in Dunford and Schwartz (1988)) If a bounded operator $T$ can be written as $T = A + N$, where $A$ is an operator of the form $A = \int_{\sigma} \lambda P(d\lambda)$ with $P$ a projection-valued measure supported on the spectrum $\sigma$, a subset of the complex plane, and $N$ is a quasi-nilpotent operator commuting with $A$, then there is a unique decomposition of $T$.

Step 1. From the parallel-outcome assumption, for fixed $(y^{(1)}, x)$, we have

\[
T(Y^{(2)}, Y^{(3)} | x) = \int f(y^{(1)}, y^{(2)}, y^{(3)} | x) \, \mu(dx),
\]

From the above equations, by Tonelli’s theorem we have

\[
T_{f(Y^{(2)}, Y^{(3)} | x)} = T_{f(Y^{(2)} | U, x)} \Delta_{x} T_{f(Y^{(3)} | U, x)}, \quad (S25)
\]

\[
T_{f(y^{(1)}, Y^{(2)}, Y^{(3)} | x)} = T_{f(Y^{(2)} | U, x)} \Delta_{y^{(1)}(x)} \Delta_{x} T_{f(Y^{(3)} | U, x)}, \quad (S26)
\]

which are exactly (S1) and (S2) in the main paper. The left-hand sides of (S25) and (S26) are known operators because the kernels are observed distributions.

Since $T_{f(Y^{(2)} | U, x)}$ is injective, by Condition S2, we can rewrite (S25) as

\[
T^{-1}_{f(Y^{(2)} | U, x)} T_{f(Y^{(2)}, Y^{(3)} | x)} = \Delta_{x} T_{f(Y^{(3)} | U, x)}^{T}. \quad (S27)
\]

Plug (S27) into (S26), we have

\[
T_{f(y^{(1)}, Y^{(2)}, Y^{(3)} | x)} = T_{f(Y^{(2)} | U, x)} \Delta_{y^{(1)}(x)} \Delta_{x} T_{f(Y^{(3)} | U, x)} T^{-1}_{f(Y^{(2)} | U, x)} T_{f(Y^{(2)}, Y^{(3)} | x)}^{T}. \quad (S28)
\]

Since $T_{f(Y^{(3)} | U, x)}$ is injective, by Condition S2, we can apply $T^{-1}_{f(Y^{(2)}, Y^{(3)} | x)}$ from the right on each side of (S28),

\[
T_{f(y^{(1)}, Y^{(2)}, Y^{(3)} | x)} T^{-1}_{f(Y^{(2)}, Y^{(3)} | x)} = T_{f(Y^{(2)} | U, x)} \Delta_{y^{(1)}(x)} \Delta_{x} T_{f(Y^{(3)} | U, x)}^{T}. \quad (S29)
\]
As $T_{f(Y|U,x)}^{-1}$ is densely defined over $L^1_{bnd}(Y_2)$ by Condition S2 and Lemma S5, the operator equivalence (S29) can be extended to the whole domain space $L^1_{bnd}(Y_2)$ using the standard extension procedure for linear operators. The left-hand side of (S29) is a known operator since we can observe the joint distribution of $(X, Y^{(1)}, Y^{(2)}, Y^{(3)})$. The right-hand side is the spectral decomposition in the form of an eigenvalue-eigenfunction decomposition. The eigenvalues are elements from the multiplication operator, i.e., $f(y^{(1)} | U, x)$ for a given $(y^{(1)}, x)$ and for all $u$. The eigenfunctions are the kernels of the integral operators $T_{f(Y|U,x)}$, i.e., $f(Y^{(2)} | U, x)$ for a given $x$, and for all $(y^{(2)}, u)$. To see this, we note that $\int_{U} f(y^{(2)} | u, x) \delta(u - u_0) \, du = f(y^{(2)} | u_0, x)$, where $\delta(\cdot)$ is the Dirac delta function. For any $u_0 \in U$,

$$T_{f(Y|U,x)} \Delta f(y^{(1)} | U, x) T_{f(Y|U,x)}^{-1} f(y^{(2)} | u_0, x)$$

$$= T_{f(Y|U,x)} \Delta f(y^{(1)} | U, x) \delta(u - u_0)$$

$$= T_{f(Y|U,x)} f(y^{(1)} | u, x) \delta(u - u_0)$$

$$= \int_{U} f(y^{(2)} | u, x) f(y^{(1)} | u, x) \delta(u - u_0) \, du$$

$$= f(y^{(1)} | u_0, x) f(y^{(2)} | u_0, x).$$

Therefore $f(Y^{(2)} | u_0, x)$ is an eigenfunction corresponding to the eigenvalue $f(y^{(1)} | u_0, x)$.

Although Dirac delta function is not a function, we make use of it to simplify the notation and to illustrate the idea. This is based on the fact that $f(y^{(2)} | u_0, x) = \lim_{n \to \infty} (T_{f(Y|U,x)} g_{n,u_0})(y^{(2)})$, where $g_{n,u_0} = n1(|u - u_0| \leq n^{-1})$, a sequence of absolutely integrable and bounded functions.

We have shown that the operator equivalence (S29) can be extended to the whole domain space $L^1_{bnd}(Y_2)$. Thus $f(y^{(2)} | u_0, x)$ belongs to the extended domain of $T_{f(Y|U,x)}^{-1}$, and the derivation in (S30) can be seen as a limiting process.

Step 2. We will then show the uniqueness of the decomposition in (S29). In this step, we follow similar proof technique used in Hu and Schennach (2008), and make use of Theorem XV.4.5 in Dunford and Schwartz (1988). In our case, this theorem applies with $N = 0$ and $\sigma \subset \mathbb{R}$. It’s easy to see $N$ is quasi-nilpotent and it commutes with any operator. The spectrum $\sigma$ of $A$ is the range of $f(y^{(1)} | U, x)$ as a function of $U$. Since $f(y^{(1)} | U, x)$ is bounded by Condition S1, the operator $T$ is bounded as required by the theorem. As for $A$, the projection-valued measure $P$ assigned to any subset $\Lambda$ of $\mathbb{R}$ in our case is

$$P(\Lambda) = T_{f(Y|U,x)} I_{\Lambda} T_{f(Y|U,x)}^{-1},$$

(S31)
where the operator $I_\lambda$ is defined as
\[(I_\lambda g)(u) = 1\{f(Y^{(1)} \mid U, X) \in \Lambda\}(u)g(u).\] (S32)

Alternatively, we can define $P(\Lambda)$ by introducing the subspace
\[S(\Lambda) = \text{span}\{f(Y^{(2)} \mid u, x) : u \text{ such that } f(y^{(1)} \mid u, x) \in \Lambda\}\]
for any subset $\Lambda$ of the spectrum $\sigma$. The projection $P(\Lambda)$ is then determined by specifying its range $S(\Lambda)$ and its null space $S(\sigma \setminus \Lambda)$.

Now we will show $\int_\sigma \lambda P(d\lambda) = T_f(Y^{(2)} | U, x) \Delta f(y^{(1)} | U, x) T_f^{-1}(Y^{(2)} | U, x)$. We can write
\[
\int_\sigma \lambda P(d\lambda) = \int_\sigma \lambda \left(\frac{d}{d\lambda} P((-\infty, \lambda])\right) d\lambda
\]
where the second equality holds by (S31). To show that the middle term in parentheses is equivalent to $\Delta f(y^{(1)} | U, x)$, we let it act on a function $g(u)$.
\[
\left(\int_\sigma \lambda \frac{dI_{(-\infty, \lambda]}}{d\lambda} d\lambda \right)(u) = \int_\sigma \lambda \frac{d}{d\lambda} 1\{f(Y^{(1)} \mid U, X)(u) \in [-\infty, \lambda]\} g(u) d\lambda
\]
\[= \int_\sigma \lambda \frac{d}{d\lambda} 1\{f(Y^{(1)} \mid U, X)(u) - \lambda \leq 0\} g(u) d\lambda
\]
\[= \int_\sigma \lambda \delta \{f(Y^{(1)} \mid U, X)(u) - \lambda\} g(u) d\lambda
\]
\[= \int_\sigma \lambda \delta \{\lambda - f(Y^{(1)} \mid U, X)(u)\} g(u) d\lambda
\]
\[= f(Y^{(1)} \mid U, X)(u) g(u)
\]
\[= (\Delta f(y^{(1)} | U, x)) g(u).\]

The first equality holds by (S32). For the third and fourth equalities, we use the fact that the generalized differential of a step function $1(\lambda \leq 0)$ is a Dirac delta $\delta(\lambda)$, and $\delta(\lambda) = \delta(-\lambda)$.

The second last equation is by the fact that $\int \delta(\lambda - \lambda_0) h(\lambda) d\lambda = h(\lambda_0)$ for any function $h(\lambda)$ continuous at $\lambda = \lambda_0$ and here, for $h(\lambda) = \lambda$.

Step 3. Condition S3 ensures that the eigenspace corresponding to each eigenvalue is one dimensional. To show this, we utilize the fact that the operator $T_f(Y^{(2)} | U, x)$ that defines the eigenfunctions does not depend on $Y^{(1)}$, while the eigenvalues $f(y^{(1)} \mid U, x)$ do. Let $D(y^{(1)}, u) = \{u' : f(y^{(1)} \mid u', x) = f(y^{(1)} \mid u, x)\}$, the set of values of $U$ that index eigenfunctions sharing the same eigenvalue $f(y^{(1)} \mid u, x)$. Any linear combination of functions $f(Y^{(2)} \mid u', x)$ for $u' \in D(y^{(1)}, u)$ could be an eigenfunction of $T_f(y^{(1)}, Y^{(2)}, Y^{(3)} | x) T_f^{-1}(Y^{(2)}, Y^{(3)} | x)$. If $v(u) \equiv
\( \cap_{y(1) \in \mathcal{Y}(1)} \text{span}\{ f(Y^{(2)} | u', x) | u' \in D(y^{(1)}, u) \} \) is not one dimensional, it must have at least two eigenfunctions, say \( f(Y^{(2)} | u, x) \) and \( f(Y^{(2)} | u', x) \). Thus, \( \cap_{y(1) \in \mathcal{Y}} D(y^{(1)}, u) \) must at least contain \( u \) and \( u' \). By the definition of \( D(y^{(1)}, u) \), it must hold that \( f(y^{(1)} | u, x) = f(y^{(1)} | u', x) \) for all \( y^{(1)} \in \mathcal{Y}_1 \). This violates Condition S3. So we conclude \( v(u) \) is one dimensional, and it uniquely determines the eigenfunction \( f(Y^{(2)} | u, x) \) after normalization to integrate to 1 by the requirement that \( \int_{\mathcal{Y}} f(y^{(2)} | u, x) \, dy^{(2)} = 1 \).

Step 4. For \( (X, Y^{(1)}) = (x, y^{(1)}) \), we have a set of eigenvalues \( f(y^{(1)} | U, x) \) and corresponding eigenfunctions \( f(Y^{(2)} | U, x) \) indexed by \( U \). The ideal situation would be that we could connect each eigen-pair with a value of \( U \). However this is not possible in our case, where only \( x \) is known. For each pair \( (u, x) \), we can act a known functional \( M \), such as taking expectation, on the eigenfunction \( f(Y^{(2)} | u, x) \) to derive some information on \( (u, x) \), e.g., \( h_1(u) + h_2(x) \) assuming \( g \) is the identity function if Condition S4 holds. We shall not assume that \( h_2 \) is known, since this is to some degree assuming the effect of \( X \) on \( Y^{(2)} \) is known. For example, if \( M \) is the mean and \( Y^{(2)} = \alpha U + \beta X + \epsilon \) where \( E(\epsilon) = 0 \), then \( h_2(x) = \beta x \) and \( \beta \) is the causal effect of \( X \) on \( Y^{(2)} \). With \( h_2 \) unknown, we can not derive \( h_1(u) \) for each \( (u, x) \) pair. Nevertheless, if we know \( h_1(u) + h_2(x) \) for a fixed \( u \) and different \( x_1, x_2 \), then we can identify \( h_2(x_2) - h_2(x_1) \). The following gives more details.

By Condition S4, assuming \( h_1(u) \) is bounded from below,
\[
\inf_{u \in U} g[M\{ f(Y^{(2)} | u, x) \}] = \inf_{u \in U} h_1(u) + h_2(x).
\]
We take the supremum if \( h_1(u) \) is not bounded from below but bounded from above. Then for \( x_0 \) and \( x_1 \), we can derive
\[
h_2(x_1) - h_2(x_0) = \inf_u g[M\{ f(Y^{(2)} | u, x_1) \}] - \inf_u g[M\{ f(Y^{(2)} | u, x_0) \}].
\]
Since the right-hand side is known, we can determine \( h_2(x) \) up to a constant. Let \( \tilde{U} = h_1(U) + h_2(x_0), \tilde{h}_2(x) = h_2(x) - h_2(x_0), \) then \( g[M\{ f(Y^{(2)} | U, X) \}] = h_1(U) + h_2(X) = \tilde{U} + \tilde{h}_2(X) \). Now that \( \tilde{h}_2(X) \) is known, we can thus derive \( \tilde{U} \), which is a one-to-one mapping of \( U \). That is to say, for \( (X, Y^{(1)}) = (x, y^{(1)}) \), we have a set of eigenvalues \( f(y^{(1)} | U, x) \) and corresponding eigenfunctions \( f(Y^{(2)} | U, x) \) indexed by known \( \tilde{U} \) instead of unknown \( U \). In other words, for a pair of eigenvalue and eigenfunction, \( f(y^{(1)} | u, x) \) and \( f(Y^{(2)} | u, x) \), although we do not know which \( u \) they are associated to, we can relate them to \( \tilde{u} = h_1(u) + h_2(x_0) \) which is known. Hence we know \( f(Y^{(2)} | \tilde{U}, x) \) and \( f(y^{(1)} | \tilde{U}, x) \) rather than \( f(Y^{(2)} | U, x) \) and \( f(y^{(1)} | U, x) \). Since
Since entries of \( P \) are estimated by sample means, each element of \( \overline{P}(Y^{(2)},Y^{(3)} \mid x) \) and \( \overline{P}(y^{(1)},Y^{(2)},Y^{(3)} \mid x) \) is \( \sqrt{n} \)-consistent. As the matrix dimension is fixed, entrywise convergence is equivalent to convergence in Frobenius norm, i.e., \( \| \overline{P}(Y^{(2)},Y^{(3)} \mid x) - P(Y^{(2)},Y^{(3)} \mid x) \|_F = O_p(n^{-1/2}) \) and \( \| \overline{P}(y^{(1)},Y^{(2)},Y^{(3)} \mid x) - P(y^{(1)},Y^{(2)},Y^{(3)} \mid x) \|_F = O_p(n^{-1/2}) \).

When \( X \) is categorical, the determinant \( \det \{ P(Y^{(2)},Y^{(3)} \mid x) \} \) is bounded away from zero by Condition 1. Since matrix inversion is an analytic function, we have \( \| \overline{P}(Y^{(2)},Y^{(3)} \mid x)^{-1} - P(Y^{(2)},Y^{(3)} \mid x)^{-1} \|_F = O_p(n^{-1/2}) \). First we will show \( \overline{P}(y^{(1)},Y^{(2)},Y^{(3)} \mid x)\overline{P}(Y^{(2)},Y^{(3)} \mid x)^{-1} \overline{P}(Y^{(2)},Y^{(3)} \mid x)^{-1} = O_p(n^{-1/2}) \).

\[ f(Y^{(2)} \mid X) \text{ is observable, we are able to identify } f(\tilde{U} \mid X) \text{ from } f(Y^{(2)} \mid \tilde{U},X) = \int f(Y^{(2)} \mid \tilde{U},X) \, d\tilde{U}, \text{ or equivalently } f(Y^{(2)} \mid X) = T_{f(Y^{(2)} \mid \tilde{U},x)} f(\tilde{U} \mid X), \text{ if } T_{f(Y^{(2)} \mid \tilde{U},x)} \text{ is injective. Actually, we can show injectivity of } T_{f(Y^{(2)} \mid \tilde{U},x)} \text{ implies injectivity of } T_{f(Y^{(2)} \mid \tilde{U},x)}. \]

\[ \text{Proof.} \text{ We only prove the first part of Theorem S2. The second part follows from standard M-estimation theory. We shall use Lemma S2 in our proof.} \]

\[ \text{For a square matrix } A \in M_k(\mathbb{R}), \text{ define the Frobenius norm as } \| A \|_F = (\sum_{i,j=1}^{k} a_{ij}^2)^{1/2}. \]

\[ \text{Since entries of } P(Y^{(2)},Y^{(3)} \mid x) \text{ and } P(y^{(1)},Y^{(2)},Y^{(3)} \mid x) \text{ are estimated by sample means, each element of } \overline{P}(Y^{(2)},Y^{(3)} \mid x) \text{ and } \overline{P}(y^{(1)},Y^{(2)},Y^{(3)} \mid x) \text{ is } \sqrt{n} \text{-consistent. As the matrix dimension is fixed, entrywise convergence is equivalent to convergence in Frobenius norm, i.e., } \| \overline{P}(Y^{(2)},Y^{(3)} \mid x) - P(Y^{(2)},Y^{(3)} \mid x) \|_F = O_p(n^{-1/2}) \text{ and } \| \overline{P}(y^{(1)},Y^{(2)},Y^{(3)} \mid x) - P(y^{(1)},Y^{(2)},Y^{(3)} \mid x) \|_F = O_p(n^{-1/2}). \]

\[ \text{When } X \text{ is categorical, the determinant } \det \{ P(Y^{(2)},Y^{(3)} \mid x) \} \text{ is bounded away from zero by Condition 1. Since matrix inversion is an analytic function, we have } \| \overline{P}(Y^{(2)},Y^{(3)} \mid x)^{-1} - P(Y^{(2)},Y^{(3)} \mid x)^{-1} \|_F = O_p(n^{-1/2}). \]
$x$ is a $\sqrt{n}$-consistent estimator under the Frobenius norm. In particular, one has

$$\|\hat{P}(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x) \hat{P}(Y^{(2)}, Y^{(3)} \mid x)^{-1} - P(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)P(Y^{(2)}, Y^{(3)} \mid x)^{-1}\|_F$$

$$= \|\hat{P}(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x) \hat{P}(Y^{(2)}, Y^{(3)} \mid x)^{-1} - P(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)P(Y^{(2)}, Y^{(3)} \mid x)^{-1}\|_F$$

$$= \|P(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)\|_F \|\hat{P}(Y^{(2)}, Y^{(3)} \mid x)^{-1} - P(Y^{(2)}, Y^{(3)} \mid x)^{-1}\|_F$$

$$= O_p(n^{-1/2}) \|\hat{P}(Y^{(2)}, Y^{(3)} \mid x)^{-1}\|_F + \|P(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)\|_F O_p(n^{-1/2})$$

$$= O_p(n^{-1/2}) + O(1)O_p(n^{-1/2})$$

$$= O_p(n^{-1/2}).$$

Let $M_0$ be such a matrix in Lemma S2, then one can find a closed neighborhood of $M_0$ over which $\Lambda(M)$ and $X(M)$ satisfy the Lipschitz condition. Therefore, the $n^{-1/2}$ convergence rate of $\tilde{P}_D(y^{(1)} \mid U, x)$ and $\tilde{P}(Y^{(2)} \mid U, x)$ can be inferred from the $n^{-1/2}$ convergence rate of $\tilde{P}(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)\tilde{P}(Y^{(2)}, Y^{(3)} \mid x)^{-1}$. The same applies to $\tilde{P}(Y^{(3)} \mid U, x)$ due to symmetry. The $n^{-1/2}$ convergence rate of $\tilde{P}_D(U \mid x)$ in (S7) can be verified using the technique in proving the $\sqrt{n}$-consistency of $\hat{P}(y^{(1)}, Y^{(2)}, Y^{(3)} \mid x)\hat{P}(Y^{(2)}, Y^{(3)} \mid x)^{-1}$. According to (S8), one can obtain $\hat{p}(u) - p(u) = O_p(n^{-1/2})$ by noticing that $\hat{p}(u \mid x)$ and $\hat{p}(u \mid x)$ are both root-n consistent.
References

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Dunford, N. and Schwartz, J. T. (1988). *Linear operators. 3: Spectral operators*. Wiley Interscience, New York.

Feng, M., Mitchell, J. J., Pang, J.-S., Shen, X., and Wachter, A. (2018). Complementarity formulations of $\ell_0$-norm optimization. *Pacific Journal of Optimization*, 14(2):273–305.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer texts in statistics. Springer, New York, 3rd edition.

Mattner, L. (1993). Some incomplete but boundedly complete location families. *The Annals of Statistics*, 21(4).

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.

Serre, D. (2010). *Matrices: Theory and Applications*. Springer, New York.

Yang, S., Wang, L., and Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika*, 106(4):875–888.