On the permutation invariance principle for causal estimands

Jiaqi Tong[1,2] **and Fan Li**[1,2,*]
[1]Department of Biostatistics, Yale University of Public Health, New Haven, Connecticut, USA
[2]Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, CT, USA
**email*: fan.f.li@yale.edu

October 15, 2025

Abstract

In many causal inference problems, multiple action variables share the same causal role, such as mediators, factors, network units, or genotypes, yet lack a natural ordering. To avoid ambiguity in interpretation, causal estimands should remain unchanged under relabeling, an implicit principle we refer to as permutation invariance. We formally characterize this principle, analyze its algebraic and combinatorial structure for verification, and present a class of weighted estimands that are permutation-invariant while capturing interactions of all orders. We further provide guidance on selecting weights that yield residual-free estimands, whose inclusion-exclusion sums capture the maximal effect, and extend our results to ratio effect measures.

Keywords: Causal inference; causal estimands; permutation invariance; combinatorics; causal mediation; factorial experiments

1 Introduction

Causal inference concerns the study of potential outcome contrasts defined with respect to the same target population, and many modern causal inference problems encounter the complexities of multiple variables that share the same type of causal interpretation. Examples include causal mediation with multiple unordered mediators (Xia and Chan, 2022), factorial experiments with multiple factors (Dasgupta et al., 2015), causal inference under interference with multiple network units (Hudgens and Halloran, 2008), and Mendelian randomization with multiple genotypes (instruments) (Hartwig et al., 2017). In each case, one is often interested in causal contrasts that summarize the separate or interactive effects of several variables. When these multiple variables have no inherent ordering, a principal consideration is that causal estimands should be *permutation-invariant*. That is, the definition of a set of causal estimands should not be altered by the relabeling of these variables, such as permuting the order of mediators or swapping genotype labels. The permutation invariance principle has been informally noted in several works (e.g., termed as 'symmetric estimands' in Xia and Chan (2022); Zhao and Ding (2022)) but has not been formally defined or systematically studied. Consequently, uncareful practice may yield label-dependent estimands whose interpretation is contingent upon an arbitrary relabeling, thereby obscuring the meaning of causal effects defined across multiple variables.

This work makes four main contributions. First, we provide a rigorous characterization of the permutation-invariance principle, which has been underexplored in prior literature. Second, we introduce a straightforward procedure to verify whether a given set of estimands satisfies permutation invariance. Third, we develop an interpretable, permutation-invariant, and complete (capturing all forms of interaction) class of weighted estimands that can be applied across various domains of causal inference. Finally, we identify a specific choice of weights that produces unique residual-free estimands whose inclusion-exclusion sum captures the maximal effect, and we discuss extensions to ratio effect measures.

2 What is permutation invariance?

2.1 Introductory examples

We first provide two examples in the context of causal mediation analysis to illustrate the idea of permutation-invariant estimands. Causal mediation analysis concerns the decomposition of a treatment effect into indirect pathways operating through the mediators and a direct pathway that bypasses them (Imai et al., 2010). The primary goal of such an analysis is to quantify these direct and indirect effects to understand the underlying causal mechanisms.

Example 1. *Consider the problem of causal mediation with two mediators that are not causally ordered. Let $X_k(a)$ denote the potential kth ($k = 1, 2$) mediator under condition a ($a = 0, 1$). Let $Y(1, X_1(a), X_2(a^*))$ be the potential outcome under treatment, with the potential mediators taking the values they would naturally have under treatment levels a and a^*. Lange et al. (2014) defined the natural indirect effects through X_1 and X_2 as:*

$$\Delta_{\{X_1\}} = \mu(1, 0) - \mu(0, 0), \quad \Delta_{\{X_2\}} = \mu(1, 1) - \mu(1, 0), \tag{1}$$

where $\mu(a, a^*) = E\{Y(1, X_1(a), X_2(a^*))\}$. Exchanging the labels for the two mediators leads to two new estimands: $\Delta^\dagger_{\{X_1\}} = \mu(0, 1) - \mu(0, 0)$ and $\Delta^\dagger_{\{X_2\}} = \mu(1, 1) - \mu(0, 1)$. Neither $\Delta^\dagger_{\{X_1\}}$ nor $\Delta^\dagger_{\{X_2\}}$ overlap with the original estimands, $\Delta_{\{X_1\}}$ and $\Delta_{\{X_2\}}$. Hence, such estimands are not permutation-invariant.

Example 2. *Following from the Example 1, Xia and Chan (2022) defined the* exit indirect effects *to study causal mediation with unordered mediators:*

$$\Delta_{\{X_1\}} = \mu(1, 1) - \mu(0, 1), \quad \Delta_{\{X_2\}} = \mu(1, 1) - \mu(1, 0). \tag{2}$$

In contrast to (1), (2) are permutation-invariant, because exchanging the labels results in $\Delta^\dagger_{\{X_1\}} = \mu(1, 1) - \mu(1, 0) = \Delta_{\{X_2\}}$ and $\Delta^\dagger_{\{X_2\}} = \mu(1, 1) - \mu(0, 1) = \Delta_{\{X_1\}}$. In other words, under any arbitrary relabeling, the collection of causal estimands remains unchanged up to permutation.

In what follows, we formalize the concept in Examples 1 and 2 through precise algebraic definitions, with applications extending beyond causal mediation.

2.2 General setup and algebraic definition

Suppose one is interested in inferring causal effects with a set of K *action variables*, $\mathcal{X} = \{X_1, \ldots, X_K\}$, but has no prior knowledge of their causal orderings or relationships. In such cases, it is essential to define permutation-invariant causal estimands, which prevents their specification from relying on arbitrary labeling and avoids ambiguity in interpretation. For example, $\mathcal{X}$ could be a set of K unordered mediators for causal mediation (Xia and Chan, 2022), K factors in factorial experiments (Dasgupta et al., 2015), K network units in causal inference under network interference (Hudgens and Halloran, 2008), or K genotypes in Mendelian randomization with pleiotropy (Hartwig et al., 2017). We assume that each action variable has two states, denoted by $X_k(1)$ and $X_k(0)$. Although the notation $X_k(a)$ resembles that of potential outcomes, its interpretation depends on context: it may represent the kth potential mediator under condition a, or simply indicate that the action variable is set to $X_k = a$ (e.g., in factorial experiments). Since our focus is on combinatorial properties, we use the power set $2^{\mathcal{X}}$ (the collection of all subsets of $\mathcal{X}$) to index all 2^K states. For example, $\mathcal{A} = \{X_1\} \in 2^{\mathcal{X}}$ uniquely represents the state where X_1 is set to $X_1(0)$, and all other variables X_k with $k \neq 1$ set to $X_k(1)$.

To proceed, we introduce a few algebraic concepts. We define an equivalence relation, '$\sim$', on $2^{\mathcal{X}}$ such that two subsets $\mathcal{A} \subseteq \mathcal{X}$ and $\mathcal{B} \subseteq \mathcal{X}$ are equivalent, written as $\mathcal{A} \sim \mathcal{B}$, if they have the same cardinality, or $|\mathcal{A}| = |\mathcal{B}|$. This equivalence relation induces equivalence classes on the power set $2^{\mathcal{X}}$. For any subset $\mathcal{A} \in 2^{\mathcal{X}}$, its equivalence class is defined as $[\mathcal{A}] = \{\mathcal{B} \in 2^{\mathcal{X}} : |\mathcal{B}| = |\mathcal{A}|\}$. By construction, each equivalence class can be uniquely indexed by its cardinality, $[q]$, where $0 \leq q \leq K$ (allowing for $|\emptyset| = 0$). The corresponding quotient set, $\mathcal{Q}$, is the set of all such equivalence classes: $\mathcal{Q} := \{[q] : 0 \leq q \leq K\}$.

The causal quantities of interest are typically defined as real-valued functions on $2^{\mathcal{X}}$, denoted as $f : 2^{\mathcal{X}} \to \mathbb{R}$. For example, the expectations of potential outcomes under each state. We then define $\mathbf{v}$ as a vector containing all possible quantities, ordered by $\mathcal{Q}$ or the cardinality of the subsets. Table 1 lists several example functions f for $K = 2$. Certain causal inference settings, such as 2^K factorial experiments (Dasgupta et al., 2015; Zhao and Ding, 2022) allow for weighting. For ease of exposition, we present our generic framework without weighting until Section 3.2.

With the above preliminaries, the set of causal estimands (assumed to have d elements)

Table 1: Examples causal estimands and f when $K = 2$. Here $f(\mathcal{A}) = \mu(a_1, a_2)$ such that $a_k = 0$ if $X_k \in \mathcal{A}$ or explicitly, $f(\emptyset) = \mu(1,1)$, $f(\{X_1\}) = \mu(0,1)$, $f(\{X_2\}) = \mu(1,0)$, $f(\mathcal{X}) = \mu(0,0)$.

Context	X_k	$f(\mathcal{A}) = \mu(a_1, a_2)$ such that $a_k = 0$ if $X_k \in \mathcal{A}$
Mediation with multiple mediators	kth mediator	$\mu(a_1, a_2) = E\{Y(1, X_1(a_1), X_2(a_2))\}$
2^2 factorial experiment	kth factor	$\mu(a_1, a_2) = E\{Y(X_1 = a_1, X_2 = a_2)\}$
Network interference	kth unit's assignment	$\mu(a_1, a_2) = \sum_{k=1}^{2} Y_k(X_1 = a_1, X_2 = a_2)/2$
Mendelian randomization	kth genotype	$\mu(a_1, a_2) = E\{Y(1, X_1 = a_1, X_2 = a_2)\}$

are defined as the contrasts concerning 2^K quantities, formally given by:

$$\mathbf{\Delta} = \mathbf{H}\mathbf{v}, \tag{3}$$

where $\mathbf{H}$ is a $d \times 2^K$ generator matrix of 0's and ± 1's. We first define permutation invariance.

Definition 1. *A permutation matrix is a square binary matrix that has exactly one entry of* 1 *in each row and each column, with all other entries being* 0*. Any permutation of the elements in $\mathcal{X}$ induces $K+1$ corresponding permutations on the subsets within each equivalence class $[q] \in \mathcal{Q}$, characterized by permutation matrices $\mathbf{P}^{(q)} \in \mathbb{R}^{\binom{K}{q} \times \binom{K}{q}}$ for $0 \leq q \leq K$. A matrix $\mathbf{P}$ of size 2^K is called an induced permutation matrix if it has block diagonal structure with main diagonal of permutation matrices:*

$$\mathbf{P} = diag(\mathbf{P}^{(0)}, \ldots, \mathbf{P}^{(K)}).$$

Definition 2. *A vector of contrasts $\mathbf{\Delta}$ is permutation-invariant if for any induced column permutation matrix $\mathbf{P}_c$, there exists a row permutation matrix $\mathbf{P}_r$ such that $\mathbf{P}_r\mathbf{H} = \mathbf{H}\mathbf{P}_c$.*

Definition 2 states that the causal estimands are permutation-invariant if any induced permutation of the columns of the generator matrix $\mathbf{H}$ is equivalent to a corresponding permutation of its rows. Together with (3), the permutation-invariant estimands should satisfy

$$\mathbf{P}_r\mathbf{\Delta} = \mathbf{H}(\mathbf{P}_c\mathbf{v}).$$

Essentially, the term $\mathbf{P}_c\mathbf{v}$ represents the permutation of $\mathbf{v}$ induced by relabeling the variables in $\mathcal{X}$. In turn, for the vector of estimands $\mathbf{\Delta}$ to be permutation-invariant, it must also be permutable (via a corresponding matrix $\mathbf{P}_r$) to align with the new ordering of the variables.

In what follows, we address three central aspects that arise in permutation-invariant causal estimands. First, we formalize a procedure to check whether a given set of estimands is permutation-invariant. Second, for K action variables, we construct a set of permutation-invariant estimands that is also complete, i.e., capturing all orders of interaction effects. Finally, we discuss conditions under which the aggregation of permutation-invariant and complete estimands captures the maximal effect, and hence are residual-free, based on the inclusion–exclusion principle.

3 Main results

3.1 Algebraic characterization of permutation invariance

Verification of permutation invariance based on Definition 2 may appear nontrivial. Our first result establishes a necessary and sufficient condition for verifying whether a set of estimands is permutation-invariant. To proceed, we leverage the following concepts in algebra and combinatorics (Van Lint and Wilson, 2001). A permutation $\sigma : \mathcal{Y} \to \mathcal{Y}$ is a bijection from a general finite set $\mathcal{Y}$ to itself. Alternatively, σ can also be uniquely expressed as a permutation matrix. Define S_n as the symmetric group for a set with cardinality n, i.e., S_n contains all the permutations on the given set $\mathcal{Y}$ with $|\mathcal{Y}| = n$. A permutation group $\mathcal{P}$ is the subgroup of S_n. A multiset is an extension of a set that allows elements to appear more than once, which is defined by a double: the set of its distinct elements and the multiplicity (or count) of each element. In our example, S_K and S_{2^K} are the symmetric groups on $\mathcal{X}$ and $2^{\mathcal{X}}$, respectively; $\mathcal{P} \subseteq S_{2^K}$ is the permutation group consisting of the right-multiplication of all induced permutation matrices $\mathbf{P}_c$; $\mathcal{R}(\mathbf{H}) = \{(o_1, m_1), \ldots, (o_n, m_n)\}$ (o_i and m_i denote the ith distinct row and its multiplicity) is the multiset of rows of generator matrix $\mathbf{H}$. The permutation group $\mathcal{P}$ induces an action that applies a permutation $\sigma \in \mathcal{P}$ to $\mathcal{R}(\mathbf{H})$ and produces $\mathcal{R}(\mathbf{H}\mathbf{P}_c)$.

Theorem 1. *A vector of contrasts $\boldsymbol{\Delta}$ is permutation-invariant if and only if $\mathcal{R}(\mathbf{H}\mathbf{P}_c) = \mathcal{R}(\mathbf{H})$ for all $\sigma \in \mathcal{P}$.*

By Theorem 1, permutation invariance requires the multiplicity of rows to remain unchanged under any relabeling of the underlying action variables. Therefore, verifying permutation invariance boils down to a counting exercise. We provide an example verification with $K = 2$ below.

Example 3. *The following two generator matrices correspond to the Examples 1 and 2:*

$$\mathbf{H}_1 = \begin{pmatrix} 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 \end{pmatrix}, \quad \mathbf{H}_2 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{pmatrix},$$

where $\mathbf{v} = (\mu(1,1), \mu(0,1), \mu(1,0), \mu(0,0))^\top$*, and* $\mathbf{H}_1$ *and* $\mathbf{H}_2$ *correspond to the estimands defined in Examples 1 and 2, respectively. Then shuffling* X_1 *and* X_2 *for* $\mathbf{H}_2$ *does not change the multiplicity of rows:* $\{((1,-1,0,0),1),((1,0,-1,0),1)\}$*. In contrast, shuffling* X_1 *and* X_2 *for* $\mathbf{H}_1$ *yields a different multiset of rows:* $\{((0,1,0,-1),1),((1,-1,0,0),1)\}$*.*

Theorem 1 also motivates the following algorithm to identify the corresponding row permutation matrix $\mathbf{P}_r$, given an induced column permutation matrix $\mathbf{P}_c$. Its convergence is guaranteed for permutation-invariant contrasts.

Algorithm 1. *Assume that* $\mathbf{H}$ *is permutation-invariant.*

To begin with, compute $\mathbf{H}' = \mathbf{H}\mathbf{P}_c$*. Let* r_i *and* r'_i *denote the* i*th rows of* $\mathbf{H}$ *and* $\mathbf{H}'$*, respectively. For* $i = 1$ *to* $i = d$*, set* $\sigma(i) = j$ *such that* $r_i = r'_j$*, where the choice of* σ *may not be unique. Output* $\mathbf{P}_r$ *that corresponds to the permutation* σ*.*

3.2 A class of permutation-invariant and complete estimands

We next introduce a class of estimands that are both permutation-invariant and *complete*. Here, a set of estimands is said to be complete if it captures all interaction effects concerning the K action variables. For example, the class of estimands in Example 2 is permutation-invariant but not complete, because it only captures first-order effects (defined by varying either X_1 or X_2 while holding the other constant) and thus fails to measure their interaction. To be considered complete, a class of estimands must account for contrasts among all nonempty subsets of $\mathcal{X}$. Definition 3 below presents this class of estimands in full generality for $K \geq 2$.

Definition 3. *With K action variables, an interpretable complete class of estimands can be written as $\boldsymbol{\Delta} = \{\Delta_{\mathcal{Y}} : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}\}$, where*

$$\Delta_{\mathcal{Y}} = \sum_{\mathcal{T} \subseteq \mathcal{Y}^c} w(\mathcal{T}, \mathcal{Y}) \left[\sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z} \cup \mathcal{T}) \right], \tag{4}$$

and $w : \mathcal{D} := \{(\mathcal{T}, \mathcal{Y}) : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}, \mathcal{T} \subseteq \mathcal{Y}^c\} \to \mathbb{R}_{\geq 0}$ ($\bullet^c$ is the set complement) are the normalized weights such that $\sum_{\mathcal{T} \subseteq \mathcal{Y}^c} w(\mathcal{T}, \mathcal{Y}) = 1$ for all $\mathcal{Y}$ and the total number of contrasts is $d = |\boldsymbol{\Delta}| = 2^K - 1$.

For $K = 2$ and when the effect of interest is $\mathcal{Y} = \{X_1\}$, if the weights are set to $w(\emptyset, \{X_1\}) = 1$ and $w(\{X_2\}, \{X_1\}) = 0$ as in the causal mediation context, the estimand (4) becomes $\Delta_{\{X_1\}} = f(\emptyset) - f(\{X_1\})$. Using the notation from Table 1, it follows that $\Delta_{\{X_1\}} = f(\emptyset) - f(\{X_1\}) = \mu(1,1) - \mu(0,1)$, which coincides with $\Delta_{\{X_1\}}$ in Example 2. For the analysis of 2^K factorial experiments, the weighting function is not necessarily a point mass. In this case, $\Delta_{\{X_1\}} = w(\emptyset, \{X_1\})(\mu(1,1) - \mu(0,1)) + w(\{X_2\}, \{X_1\})(\mu(1,0) - \mu(0,0))$. Following Zhao and Ding (2022), the terms $\mu(1,1) - \mu(0,1)$ and $\mu(1,0) - \mu(0,0)$ can be interpreted as the conditional treatment effects of factor X_1 given $X_2 = 1$ and $X_2 = 0$, respectively. In this context, the weights $w(\emptyset, \{X_1\})$ and $w(\{X_2\}, \{X_1\})$ correspond to $\Pr(X_2 = 1)$ and $\Pr(X_2 = 0)$, respectively, and are *compatible* if derived from a common joint distribution of $\mathcal{X}$; see Supplementary Material Remark 2.

The preceding example motivates two distinct viewpoints on the weights. The first arises in factorial designs where weights correspond to action variables and represent probability masses. This perspective implies that the weights must permute consistently with the action variables. For instance, if X_1 and X_2 are interchanged, their weights are swapped accordingly: $w(\emptyset, \{X_1\}) = \Pr(X_2 = 1)$ becomes $w(\emptyset, \{X_2\}) = \Pr(X_1 = 1)$. The second viewpoint applies when such a correspondence is absent, in which case the weights are considered invariant to permutations of the action variables. For example, the estimands $\{\Delta_{\{X_1\}}, \Delta_{\{X_2\}}\}$ with $\Delta_{\{X_1\}} = 1/2(\mu(1,1) - \mu(0,1)) + 1/2(\mu(1,0) - \mu(0,0))$ and $\Delta_{\{X_2\}} = 1/2(\mu(1,1) - \mu(1,0)) + 1/2(\mu(0,1) - \mu(0,0))$ are defined using the fixed weights $1/2$ and only μ is permutable with respect to action variables. We henceforth refer to these two types as *permutable* and *invariant* weights, respectively.

The proposed class of estimands also nests several existing estimands in the causal inference literature. For example, the estimands Δ_i from Xia and Chan (2022) (Example

2) are a special case of the $\boldsymbol{\Delta}$ with invariant weights, recoverable by setting $w(\mathcal{T},\mathcal{Y}) = \mathbf{1}(\mathcal{T}=\emptyset)$ (point mass), where $\mathbf{1}(\bullet)$ is the indicator function. While Xia and Chan (2022) did not provide explicit definitions for $K \geq 3$, $\{\Delta_{\mathcal{Y}} = \sum_{\mathcal{Z}\subseteq\mathcal{Y}}(-1)^{|\mathcal{Z}|}f(\mathcal{Z}) : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}\}$ serve as a natural extension of their estimands for general K. For 2^K factorial experiments, the estimand proposed by Dasgupta et al. (2015) is defined by setting $w(\mathcal{T},\mathcal{Y}) = 2^{-|\mathcal{Y}^c|}$ (equal invariant weights) for all $\mathcal{T}$; the estimand proposed by Zhao and Ding (2022) corresponds to the class $\boldsymbol{\Delta}$ with compatible permutable weights. While the methods in Sections 2.2 and 3.1 are unweighted, Supplementary Material Example 5 ($K = 2$) illustrates how these formulations can be extended to weighted estimands. The following result shows that $\boldsymbol{\Delta}$ is permutation-invariant with permutable weights, and with invariant weights, under additional restrictions.

Theorem 2. *For all $1 \leq q \leq K$, the subclass $\{\Delta_{\mathcal{Y}} : \mathcal{Y} \in [q]\}$ with permutable weights is permutation-invariant and but incomplete because it only captures q-th order interaction effects. With invariant weights, this subclass retains the same properties if Supplementary Material condition 1 (also see Supplementary Material Remark 1) for the invariant weights additionally holds. In contrast, the full class $\{\Delta_{\mathcal{Y}} : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}\}$ is both permutation-invariant and complete.*

By Theorem 2, the estimand classes in Dasgupta et al. (2015) and Xia and Chan (2022) (with invariant weights) and in Zhao and Ding (2022) (with permutable weights) are all permutation-invariant and complete. To further enhance interpretation of $\Delta_{\mathcal{Y}}$, we provide an alternative characterization for $\Delta_{\mathcal{Y}}$ below.

Proposition 1. *For any $X_k \in \mathcal{Y}$, the estimand given by (4) can alternatively be expressed as,*

$$\Delta_{\mathcal{Y}} = \sum_{\mathcal{T}\subseteq\mathcal{Y}^c} w(\mathcal{T},\mathcal{Y}) \left[\sum_{\mathcal{Z}\subseteq\mathcal{Y}\setminus\{X_k\}} (-1)^{|\mathcal{Z}|}\big(f(\mathcal{Z}\cup\mathcal{T}) - f(\mathcal{Z}\cup\{X_k\}\cup\mathcal{T})\big) \right].$$

Consider the interaction effect where $\mathcal{Y} = \{X_1, X_2\}$. Proposition 1 defines this estimand as $\Delta_{\{X_1,X_2\}} = (\mu(1,1) - \mu(0,1)) - (\mu(1,0) - \mu(0,0))$. This estimand is interpreted as the extent to which the effect of variable X_1 is modified by variable X_2. Similarly, permutation invariance implies that the interaction can also be interpreted as the extent to which the effect of X_2 is modified by X_1. More generally, Proposition 1 shows that $\Delta_{\mathcal{Y}}$ quantifies the extent to which the effect of one variable in $\mathcal{Y}$, X_k, is modified by the interaction of the remaining $|\mathcal{Y}|-1$ variables, with the choice of X_k being arbitrary within $\mathcal{Y}$ due to permutation invariance.

3.3 Residual-free estimands

The permutation-variant and complete class of estimands in Definition 3 depends on complex weights. For permutable weights, these are inherently determined by the action variables and are not subject to manipulation. In contrast, invariant weights can be adjusted. To guide their practical selection, we introduce the *residual-free* property.

Definition 4. *The class of estimands* $\boldsymbol{\Delta}$ *in Definition 3 is said to be residual-free if its inclusion-exclusion sum* $\sum_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{Y}|+1} \Delta_{\mathcal{Y}}$ *equals the maximum effect, defined as* $f(\emptyset) - f(\mathcal{X})$*; that is, the contrast between setting all variables to the treated state versus the control state.*

Here, the name '*inclusion–exclusion*' is motivated by its connection to the *Möbius inversion* in combinatorics (Rota, 1964), as detailed in the proof of Theorem 3. According to Definition 4, the inclusion–exclusion sum of a residual-free class of estimands is required to exactly capture the largest possible causal effect among the action variables—the maximal effect. The maximal effect is of broad interest across different areas of causal inference. In causal mediation analysis, it corresponds to the natural indirect effect (Tchetgen and Shpitser, 2012); in causal network analysis, it is particularly valuable because it can be identified and estimated under minimal structural assumptions (Hudgens and Halloran, 2008). Alternatively, the equation $f(\emptyset) - f(\mathcal{X}) = \sum_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{Y}|+1} \Delta_{\mathcal{Y}}$ can be interpreted as a natural decomposition of estimands, with immediate applications to causal mediation analysis. Moreover, for the class that are not residual-free, we define the residual effect as the difference between the combinatorial sum and the maximal effect,

$$\text{res} = \sum_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{Y}|+1} \Delta_{\mathcal{Y}} - (f(\emptyset) - f(\mathcal{X})),$$

which captures the remaining effect of inclusion-exclusion aggregation that cannot be explained by the maximal effect. An explicit expression for the residual effect applicable to the non–residual-free class of estimands is provided in the Supplementary Material Lemma

1. An immediate next question is how the residual-free class is identified through the weights defined in Definition 3. The result below provides this characterization.

Theorem 3. *For the class of estimands $\boldsymbol{\Delta}$ in Definition 3 with invariant weights, then res* $= 0$ *if and only if* $w(\mathcal{T}, \mathcal{Y}) = \mathbf{1}(\mathcal{T} = \emptyset)$ *and* $\Delta_{\mathcal{Y}} = \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z})$. *That is, the residual-free class exists and is uniquely identified by point-mass weights.*

Theorem 3 states that the unique residual-free estimand class is obtained by setting X_k to its treated state $X_k(1)$ for all X_k not included in the subset under comparison, i.e., $X_k \notin \mathcal{Y}$. Based on Theorem 3, the estimand class proposed by Xia and Chan (2022) for the $K = 2$ case is residual-free, and the class $\{\Delta_{\mathcal{Y}} = \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z}) : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}\}$ extends their formulation to arbitrary K. In practice, we recommend choosing weights according to their natural probabilistic context, which yields permutable weights. When such a context is unclear for invariant weights, one may simply use the default permutation-invariant, complete, and residual-free class of estimands: $\{\Delta_{\mathcal{Y}} = \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z}) : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}\}$.

4 Extensions to ratio estimands

Causal inference with binary outcomes is not uncommon, and the risk ratio and odds ratio are typical alternative effect measures. The proposed permutation-invariant, complete, and residual-free estimands $\boldsymbol{\Delta}$ generalize to ratio measures. Specifically, the log risk ratio and log odds ratio are obtained by replacing the building block f with the compositions $h \circ f = \log(f)$ and $h \circ f = \log(f/(1-f))$, respectively. We illustrate applications of the previous results to the risk ratio and odds ratio in the following two corollaries.

Corollary 1. *Consider replacing the mapping* f *with* $h \circ f$, *where* $h(x) = \log(x)$. *The class of estimands that are permutation-invariant, complete, and residual-free is given by* $\{\Delta_{\mathcal{Y}} : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}\}$, *where* $\Delta_{\mathcal{Y}} = \Pi_{\mathcal{Z} \subseteq \mathcal{Y}} f(\mathcal{Z})^{(-1)^{|\mathcal{Z}|}}$ *and its inclusion-exclusion product satisfies* $\Pi_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} \Delta_{\mathcal{Y}}^{(-1)^{|\mathcal{Y}|+1}} = f(\emptyset)/f(\mathcal{X})$.

Corollary 2. *Consider replacing the mapping* f *with* $h \circ f$, *where* $h(x) = \log(x/(1-x))$. *The class of estimands that are permutation-invariant, complete, and residual-free is given by* $\{\Delta_{\mathcal{Y}} : \emptyset \neq \mathcal{Y} \in 2^{\mathcal{X}}\}$, *where*

$$\Delta_{\mathcal{Y}} = \Pi_{\mathcal{Z} \subseteq \mathcal{Y}} \left(\frac{f(\mathcal{Z})}{1 - f(\mathcal{Z})} \right)^{(-1)^{|\mathcal{Z}|}},$$

and its inclusion-exclusion product satisfies $\Pi_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}}\Delta_{\mathcal{Y}}^{(-1)^{|\mathcal{Y}|+1}} = \{f(\emptyset)(1-f(\mathcal{X}))\}/\{f(\mathcal{X})(1-f(\emptyset))\}$.

Corollaries 1 and 2 show that, algebraically, applying a logarithmic transformation converts the additive structure of estimands on the difference scale into a multiplicative structure on the ratio scale. To illustrate, we revisit the Example 2 with a focus on ratio scale.

Example 4. *The exit indirect effects on the risk ratio scale are defined as* $\Delta_{\{X_1\}} = \mu(1,1)/\mu(0,1)$, $\Delta_{\{X_2\}} = \mu(1,1)/\mu(1,0)$, *and* $\Delta_{\{X_1,X_2\}} = \{\mu(0,0)\mu(1,1)\}/\{\mu(0,1)\mu(1,0)\}$. *The multiplicative inclusion-exclusion decomposition is given by* $\Delta_{\{X_1\}}\Delta_{\{X_2\}}/\Delta_{\{X_1,X_2\}} = \mu(1,1)/\mu(0,0)$. *Similarly, the exit indirect effects on the odds ratio are* $\Delta_{\{X_1\}} = \{\mu(1,1)(1-\mu(0,1))\}/\{\mu(0,1)(1-\mu(1,1))\}$, $\Delta_{\{X_2\}} = \{\mu(1,1)(1-\mu(1,0))\}/\{\mu(1,0)(1-\mu(1,1))\}$, *and* $\Delta_{\{1,2\}} = \mu(0,0)\mu(1,1)(1-\mu(0,1))(1-\mu(1,0))/\{\mu(0,1)\mu(1,0)(1-\mu(0,0))(1-\mu(1,1))\}$. *The multiplicative inclusion-exclusion decomposition is given by* $\Delta_{\{X_1\}}\Delta_{\{X_2\}}/\Delta_{\{X_1,X_2\}} = \{\mu(1,1)(1-\mu(0,0))\}/\{\mu(0,0)(1-\mu(1,1))\}$.

5 Conclusion

The contributions of our work are fourfold. First, we formalize the permutation-invariance principle, a notion that has been insufficiently clarified in existing literature. Second, we provide a simple procedure for verifying the permutation invariance of a given set of estimands. Third, we formally characterize an interpretable, permutation-invariant, and complete class of weighted estimands applicable across diverse areas of causal inference. Finally, we identify the choice of weights that yields residual-free estimands whose combinatorial sum captures the maximal effect, and discuss extensions to ratio measures.

For simplicity, we demonstrate the proposed framework for the concrete cases $K = 2$ and $K = 3$ in the Supplementary Material Appendix B. While this paper focuses on action variables with only two states (treated or control) as a foundational step, and a future direction is to generalize to multiple states. This would allow applicability to factorial experiments with multiple categorical factors and to causal mediation with multiple unordered mediators and an ordinal treatment.

Acknowledgement

This work is supported by the Patient-Centered Outcomes Research Institute® (PCORI® Award ME-2023C1-31350). The statements are solely the responsibility of the authors and do not necessarily represent the views of PCORI®, its Board of Governors or Methodology Committee.

SUPPLEMENTARY MATERIAL

Summary

This supplementary material is organized as follows. Appendix A contains the proofs of all Theorems, Propositions, and Corollaries. Appendix B presents illustrative representations of the proposed methods for the cases $K = 2$ and $K = 3$.

Appendix A: Proofs of Theorems, Propositions, and Corollaries

We state three key lemmas useful for the subsequent proofs. The first lemma gives an alternative representation of $\Delta_{\mathcal{Y}}$ and an explicit characterization of the residual effect res.

Lemma 1. *For general weights and any $\mathcal{Y}$, $\Delta_{\mathcal{Y}}$ includes all terms of the form $f(\mathcal{Z})$ with $\mathcal{Z} \subseteq \mathcal{X}$. More precisely, $\Delta_{\mathcal{Y}}$ can be alternatively expressed as*

$$\Delta_{\mathcal{Y}} = \sum_{\mathcal{Z} \subseteq \mathcal{X}} (-1)^{|\mathcal{Z} \cap \mathcal{Y}|} w(\mathcal{Z} \backslash \mathcal{Y}, \mathcal{Y}) f(\mathcal{Z}). \tag{5}$$

Consequently, the residual effect can be explicitly written as

$$res = \sum_{\mathcal{Z} \subseteq \mathcal{X}} coef(\mathcal{Z}) f(\mathcal{Z}),$$

where the coefficients for $f(\mathcal{Z})$ are given by

$$coef(\mathcal{Z}) = \begin{cases} \sum_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{Y}| + |\mathcal{Z} \cap \mathcal{Y}| + 1} w(\mathcal{Z} \backslash \mathcal{Y}, \mathcal{Y}), & \mathcal{Z} \neq \emptyset, \mathcal{X}, \\ \sum_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{Y}| + 1} w(\emptyset, \mathcal{Y}) - 1, & \mathcal{Z} = \emptyset, \\ 1 - \sum_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} w(\mathcal{Y}^c, \mathcal{Y}), & \mathcal{Z} = \mathcal{X}. \end{cases} \tag{6}$$

Proof. We first show (5). By the De Morgan's Laws, any subset $\mathcal{Z} \subseteq \mathcal{X}$ can be written as $\mathcal{Z} = \mathcal{T} \cup \mathcal{Z}'$, where $\mathcal{Z}' = \mathcal{Z} \cap \mathcal{Y} \subseteq \mathcal{Y}$ and $\mathcal{T} = \mathcal{Z} \cap \mathcal{Y}^c = \mathcal{Z}\backslash\mathcal{Y} \subseteq \mathcal{Y}^c$, which further implies the right-hand-side of (5) equals to

$$\begin{aligned}
&\sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\mathcal{Y}|} w(\mathcal{Z}\backslash\mathcal{Y}, \mathcal{Y}) f(\mathcal{Z}) \\
&= \sum_{\mathcal{T}\subseteq\mathcal{Y}^c} \sum_{\mathcal{Z}'\subseteq\mathcal{Y}} (-1)^{|(\mathcal{T}\cup\mathcal{Z}')\cap\mathcal{Y}|} w((\mathcal{T}\cup\mathcal{Z}')\backslash\mathcal{Y}, \mathcal{Y}) f(\mathcal{T}\cup\mathcal{Z}') \\
&= \sum_{\mathcal{T}\subseteq\mathcal{Y}^c} \sum_{\mathcal{Z}'\subseteq\mathcal{Y}} (-1)^{|\mathcal{Z}'|} w(\mathcal{T}, \mathcal{Y}) f(\mathcal{T}\cup\mathcal{Z}') \\
&= \sum_{\mathcal{T}\subseteq\mathcal{Y}^c} w(\mathcal{T}, \mathcal{Y}) \left[\sum_{\mathcal{Z}'\subseteq\mathcal{Y}} (-1)^{|\mathcal{Z}'|} f(\mathcal{Z}'\cup\mathcal{T}) \right] = \text{(Equation (4), main manuscript)}.
\end{aligned}$$

Next, we show (6). By definition, we have that

$$\begin{aligned}
\text{res} &= \sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}|+1} \Delta_{\mathcal{Y}} - [f(\emptyset) - f(\mathcal{X})] \\
&= \sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}|+1} \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\mathcal{Y}|} w(\mathcal{Z}\backslash\mathcal{Y}, \mathcal{Y}) f(\mathcal{Z}) - [f(\emptyset) - f(\mathcal{X})] \\
&= \sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}|+|\mathcal{Z}\cap\mathcal{Y}|+1} w(\mathcal{Z}\backslash\mathcal{Y}, \mathcal{Y}) f(\mathcal{Z}) - [f(\emptyset) - f(\mathcal{X})] \\
&= \sum_{\mathcal{Z}\subseteq\mathcal{X}} \text{coef}^*(\mathcal{Z}) f(\mathcal{Z}) - [f(\emptyset) - f(\mathcal{X})],
\end{aligned}$$

where $\text{coef}^*(\mathcal{Z}) = \sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}|+|\mathcal{Z}\cap\mathcal{Y}|+1} w(\mathcal{Z}\backslash\mathcal{Y}, \mathcal{Y})$ and the last equality follows from the fact that the summation is exchangeable over a rectangle. We conclude by letting $\text{coef}(\mathcal{Z}) = \text{coef}^*(\mathcal{Z})$ for $\mathcal{Z} \neq \emptyset, \mathcal{X}$; and $\text{coef}(\emptyset) = \text{coef}^*(\emptyset) - 1$ and $\text{coef}(\mathcal{X}) = \text{coef}^*(\mathcal{X}) + 1$. □

Another useful combinatorial lemma is the *weighted inclusion–exclusion principle (w-IEP)*, which follows from *Möbius inversion theory* in combinatorics. Here, the inclusion-exclusion weights differ from the probability mass weights w introduced in the main manuscript. We state it without proof, as it is a standard result found in standard combinatorics texts, such as Rota (1964); Van Lint and Wilson (2001).

Lemma 2 (Weighted Inclusion–Exclusion Principle)**.** *Suppose $\lambda : \mathcal{X} \to \mathbb{R}$ is a inclusion-exclusion weight function that assigns a weight to each element in $\mathcal{X}$. With a slight abuse of notation, define $\lambda(\mathcal{A}) = \sum_{X\in\mathcal{A}} \lambda(X)$; that is, the total weight of a subset $\mathcal{A}$ is the sum of the weights of its elements. Then for any collection of subsets $\mathcal{G}_s \subseteq \mathcal{X}, 1 \leq s \leq S$, the* *w-IEP is*

$$\lambda(\cap_{s=1}^{S}\mathcal{G}_s) = \lambda(\mathcal{X}) + \sum_{s=1}^{S}(-1)^s \sum_{i_1<\ldots<i_s} \lambda(\mathcal{G}_{i_1}^c \cap \ldots \cap \mathcal{G}_{i_s}^c),$$

where $\bullet^c$ *is the set complement.*

The third lemma establishes a simple property: applying any element-wise permutation to the elements of the power set defines a surjective mapping.

Lemma 3. *For any permutation* $\sigma \in S_K$ *and any subset* $\mathcal{A} \subseteq \mathcal{X}$*, we define* $\sigma^*(\mathcal{A}) := \{\sigma(X) : X \in \mathcal{A}\}$*, which is well-defined because* σ *is a bijective map. Then* $\sigma^* : 2^{\mathcal{X}} \to 2^{\mathcal{X}}$ *is a bijective map.*

Proof. The injectivity of σ^* follows directly from the bijectivity of σ. We establish surjectivity. For any $\mathcal{B} \in 2^{\mathcal{X}}$, we construct a well-defined set using inverse of σ (inverse function exists as σ is bijective), $\mathcal{A} = \{\sigma^{-1}(X') : X' \in \mathcal{B}\}$. It is enough to show that $\mathcal{A} \in 2^{\mathcal{X}}$, i.e., for any $X \in \mathcal{A}$, we have that $X \in \mathcal{X}$. This holds because $X = \sigma^{-1}(X') \in \mathcal{X}$, and σ is a bijective. □

Proof of Theorem 1

Proof. For simplicity, we denote the generator matrix after permuting action variables as $\mathbf{H}' = \mathbf{H}\mathbf{P}_c$. We first show the necessity, i.e., forward direction ($\Rightarrow$). By definition, the class of permutation-invariant estimands $\mathbf{\Delta}$ satisfy that for any $\sigma \in \mathcal{P}$ and the corresponding induced permutation matrix $\mathbf{P}_c$, there exists a row permutation matrix $\mathbf{P}_r$ such that $\mathbf{P}_r\mathbf{H} = \mathbf{H}'$, which implies that $\mathcal{R}(\mathbf{P}_r\mathbf{H}) = \mathcal{R}(\mathbf{H}')$. On the other hand, the multiset of rows is invariant under row permutation, i.e., $\mathcal{R}(\mathbf{P}_r\mathbf{H}) = \mathcal{R}(\mathbf{H})$. Therefore, we obtain $\mathcal{R}(\mathbf{H}') = \mathcal{R}(\mathbf{P}_r\mathbf{H}) = \mathcal{R}(\mathbf{H})$.

We next prove sufficiency (the backward direction, $\Leftarrow$). It suffices to construct a row permutation matrix $\mathbf{P}_r$ such that $\mathbf{P}_r\mathbf{H} = \mathbf{H}'$. The procedure is as follows. Since the multisets of rows of $\mathbf{H}$ and $\mathbf{H}'$ coincide, we denote their common multiset by (M, η), where $M = \{m_1, \ldots, m_l\}$ is the set of l distinct rows and $\eta : M \to \mathbb{R}_+$ denotes the multiplicity function. Let d be the total number of rows of $\mathbf{H}$ or $\mathbf{H}'$. For ease of exposition, we also define $\{r_1, \ldots, r_d\}$ and $\{r'_1, \ldots, r'_d\}$ as the multisets of rows of $\mathbf{H}$ and $\mathbf{H}'$, respectively, where r_i and r'_i denote the ith rows of $\mathbf{H}$ and $\mathbf{H}'$, respectively. Further, we define two maps $\mu : \{1, \ldots, d\} \to \{m_1, \ldots, m_l\}$ and $\mu' : \{1, \ldots, d\} \to \{m_1, \ldots, m_l\}$, where $\mu(i) = m_s$ if $r_i = m_s$ and $\mu'(i) = m_s$ if $r'_i = m_s$, for $i \in \{1, \ldots, d\}$ and $s \in \{1, \ldots, l\}$. Of note, μ and μ' are surjective but not injective. For each m_s (where $s \in \{1, \ldots, l\}$), we define $I_s = \mu^{-1}(m_s)$ as the set of indices for all rows in the matrix $\mathbf{H}$ that are equal to m_s, and similarly, $J_s = \mu'^{-1}(m_s)$ as the corresponding index set for the matrix $\mathbf{H}'$, where $\bullet^{-1}$ denotes the preimage. By assumption, these two sets have the same cardinality, such that $|I_s| = |J_s| = \eta(m_s)$. The fact that I_s and J_s have the same cardinality implies that for all $s \in \{1, \ldots, l\}$, there exist a bijection between I_s and J_s and we denote it as $\gamma_s : I_s \to J_s$. We set the permutation from $\mathbf{H}$ to $\mathbf{H}'$ as $\sigma : \{1, \ldots, d\} \to \{1, \ldots, d\}$ such that $\sigma(i) = \gamma_s(i)$ and $\mu(i) = m_s$, and find $\mathbf{P}_r$ that uniquely characterizes the permutation σ. Lastly, we show that σ is a permutation. First, σ is well-defined, as $\mu(i)$ maps to a unique distinct row m_s. Second, the bijectivity of σ follows from the bijectivity of γ_s all $s \in \{1, \ldots, l\}$. In fact, the above procedure not only establishes sufficiency but also yields an explicit algorithm for constructing the required row permutation matrix $\mathbf{P}_r$, thereby proving the convergence of Algorithm 1 in the main manuscript. □

Proof of Theorem 2

Proof. We first focus on the permutable weights. For any permutation $\sigma \in S_K$ and any subset $\mathcal{A} \subseteq \mathcal{X}$, we define $\sigma^*(\mathcal{A}) := \{\sigma(X) : X \in \mathcal{A}\}$. It suffices to show that, for any permutation σ and any $\Delta_{\mathcal{Y}}$ with $\mathcal{Y} \subseteq [q]$ for $1 \leq q \leq K$, the quantity $\Delta_{\sigma^*(\mathcal{Y})}$ coincides with the estimand, $\Delta^{\sigma}_{\mathcal{Y}}$, obtained by applying σ to both the building block f and the weight $w(\mathcal{T}, \mathcal{Y})$. We employ the alternative characterization of $\Delta_{\mathcal{Y}}$ provided in Lemma 1. That is, we must show that $\Delta_{\sigma(\mathcal{Y})} = \Delta^{\sigma}_{\mathcal{Y}}$, where

$$\Delta_{\sigma^*(\mathcal{Y})} = \sum_{\mathcal{Z} \subseteq \mathcal{X}} (-1)^{|\mathcal{Z} \cap \sigma^*(\mathcal{Y})|} w(\mathcal{Z} \backslash \sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) f(\mathcal{Z}),$$

$$\Delta^{\sigma}_{\mathcal{Y}} = \sum_{\mathcal{Z} \subseteq \mathcal{X}} (-1)^{|\mathcal{Z} \cap \mathcal{Y}|} w(\sigma^*(\mathcal{Z} \backslash \mathcal{Y}), \sigma^*(\mathcal{Y})) f(\sigma^*(\mathcal{Z})).$$

Since σ is a bijective map, the cardinality is preserved under mapping σ^*, i.e., $|\sigma^*(\mathcal{A})| = |\mathcal{A}|$ for any $\mathcal{A} \subseteq \mathcal{X}$, and that σ^* is distributive with respect to set intersection, i.e., $\sigma^*(\mathcal{A} \cap \mathcal{B}) =$

$\sigma^*(\mathcal{A}) \cap \sigma^*(\mathcal{B})$ for any $\mathcal{A}, \mathcal{B} \subseteq \mathcal{X}$. Thus, we have that (i)

$$\begin{aligned}\sigma^*(\mathcal{Z}\backslash\mathcal{Y}) =& \sigma^*(\mathcal{Z} \cap \mathcal{Y}^c) \\ =& \sigma^*(\mathcal{Z}) \cap \sigma^*(\mathcal{Y}^c) \\ =& \sigma^*(\mathcal{Z}) \cap \sigma^*(\mathcal{Y})^c \\ =& \sigma^*(\mathcal{Z})\backslash\sigma^*(\mathcal{Y}) \end{aligned} \tag{7}$$

and (ii)

$$\begin{aligned}|\sigma^{*-1}(\mathcal{Z}) \cap \mathcal{Y}| =& |\sigma^*(\sigma^{*-1}(\mathcal{Z}) \cap \mathcal{Y})| \\ =& |\sigma^*(\sigma^{*-1}(\mathcal{Z})) \cap \sigma^*(\mathcal{Y}))| \\ =& |\mathcal{Z} \cap \sigma^*(\mathcal{Y})|. \end{aligned} \tag{8}$$

Finally, we have that

$$\begin{aligned}&\Delta_{\sigma^*(\mathcal{Y})} - \Delta^\sigma_{\mathcal{Y}} \\ =& \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) f(\mathcal{Z}) - \sum_{\mathcal{Z}^*\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}^*\cap\mathcal{Y}|} w(\sigma^*(\mathcal{Z}^*)\backslash\sigma^*(\mathcal{Y})), \sigma^*(\mathcal{Y})) f(\sigma^*(\mathcal{Z}^*)) \\ =& \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) f(\mathcal{Z}) - \sum_{\sigma^{*-1}(\mathcal{Z})\subseteq\mathcal{X}} (-1)^{|\sigma^{*-1}(\mathcal{Z})\cap\mathcal{Y}|} w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y})), \sigma^*(\mathcal{Y})) f(\mathcal{Z}) \\ =& \sum_{\mathcal{Z}\subseteq\mathcal{X}} ((-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} - (-1)^{|\sigma^{*-1}(\mathcal{Z})\cap\mathcal{Y}|}) w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) f(\mathcal{Z}) \\ =& 0, \end{aligned}$$

where (i) the first equality follows from (7); (ii) the second equality follows by setting $\mathcal{Z} = \sigma^*(\mathcal{Z}^*)$ and noting that $\mathcal{Z}^* = \sigma^{*-1}(\mathcal{Z})$, by Lemma 3; (iii) the second to last equality follows from the fact that Lemma 3 implies $\sigma^{*-1}(\mathcal{Z}) \subseteq \mathcal{X}$ if and only if $\mathcal{Z} \subseteq \mathcal{X}$; and (iv) the last equality is because of (8).

For invariant weights, we note that the permutation only applies to $f(\mathcal{Z})$ rather than the weights, i.e.,

$$\Delta^\sigma_{\mathcal{Y}} = \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\mathcal{Y}|} w(\mathcal{Z}\backslash\mathcal{Y}, \mathcal{Y}) f(\sigma^*(\mathcal{Z})).$$

By the similar reasoning, we have that

$$
\begin{aligned}
&\Delta_{\sigma^*(\mathcal{Y})} - \Delta^\sigma_{\mathcal{Y}} \\
=& \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) f(\mathcal{Z}) - \sum_{\mathcal{Z}^*\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}^*\cap\mathcal{Y}|} w(\mathcal{Z}^*\backslash\mathcal{Y}), \mathcal{Y}) f(\sigma^*(\mathcal{Z}^*)) \\
=& \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) f(\mathcal{Z}) - \sum_{\sigma^{*-1}(\mathcal{Z})\subseteq\mathcal{X}} (-1)^{|\sigma^{*-1}(\mathcal{Z})\cap\mathcal{Y}|} w(\sigma^{*-1}(\mathcal{Z})\backslash\mathcal{Y}, \mathcal{Y}) f(\mathcal{Z}) \\
=& \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) f(\mathcal{Z}) - \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} w(\sigma^{*-1}(\mathcal{Z})\backslash\mathcal{Y}, \mathcal{Y}) f(\mathcal{Z}) \\
=& \sum_{\mathcal{Z}\subseteq\mathcal{X}} (-1)^{|\mathcal{Z}\cap\sigma^*(\mathcal{Y})|} (w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) - w(\sigma^{*-1}(\mathcal{Z})\backslash\mathcal{Y}, \mathcal{Y})) f(\mathcal{Z}).
\end{aligned}
$$

We conclude if the following condition holds.

Condition 1. *For all $\mathcal{Z}$, $\mathcal{Y}$, and σ^*, the following condition holds*

$$w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) = w(\sigma^{*-1}(\mathcal{Z})\backslash\mathcal{Y}, \mathcal{Y}).$$

Finally, we comment on the implications of the condition $w(\mathcal{Z}\backslash\sigma^*(\mathcal{Y}), \sigma^*(\mathcal{Y})) = w(\sigma^{*-1}(\mathcal{Z})\backslash\mathcal{Y}, \mathcal{Y})$ for the degrees of freedom of the invariant weights, as below.

Remark 1. The total degrees of freedom for the invariant weights are given by df $= (K-1)K/2$. To see this, the weighted class of estimands given in Definition 3 is permutation-invariant only if $w(\mathcal{T}, \mathcal{Y})$ only depends on $|\mathcal{T}|$ and $|\mathcal{Y}|$, as the equivalence class $[q]$ is closed under permutation. Therefore, the total degrees of freedom can be computed as

$$\text{df} = \sum_{q=1}^{K-1} (K-q+1) - (K-1) = \frac{(K-1)K}{2}.$$

□

Proof of Theorem 3

Proof. When $w(\mathcal{T}, \mathcal{Y}) = \mathbf{1}(\mathcal{T} = \emptyset)$, the proposed estimands reduce to

$$\Delta_{\mathcal{Y}} = \sum_{\mathcal{Z}\subseteq\mathcal{Y}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z}). \tag{9}$$

We first show sufficiency (backward direction, $\Leftarrow$), which uses a combinatorial technique, namely, w-IEP in the Lemma 2. Without loss of generality, we index the collection of subsets $\mathcal{G}_s$ in Lemma 2 by elements $X \in \mathcal{X}$, written as $\mathcal{G}_X, X \in \mathcal{X}$. We define the weight function on $2^{\mathcal{X}}$ as follows. First, select a collection of subsets $\mathcal{G}_X$ such that $\bigcap_{X\in\mathcal{Y}} \mathcal{G}_X^c$ are all distinct across $\mathcal{Y} \subseteq \mathcal{X}$. Second, define λ as the solution to the following system of linear equations:

$$\lambda((\cup_{X\in\mathcal{Z}}\mathcal{G}_X)^c) = f(\mathcal{Z}), \tag{10}$$

where the solution exits and is unique because the number of equations match the number of unknowns (both equal 2^K) and the column space is of full rank by construction. Applying w-IEP yields that

$$\Delta_{\mathcal{Y}} = \lambda(\cap_{X\in\mathcal{Y}}\mathcal{G}_X), \tag{11}$$

$$\lambda(\cup_{X\in\mathcal{X}}\mathcal{G}_X) = \sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}|+1}\Delta_{\mathcal{Y}}. \tag{12}$$

In other words, (i) (11) implies that the interaction effect $\Delta_{\mathcal{Y}}$ among action variables in $\mathcal{Y}$ can be interpreted as the total weight of the interactions of the subsets $\mathcal{G}_X$ corresponding to those variables, and (ii) (12) implies that the inclusion-exclusion sum of all interaction effects equals the total weight across the union of all subsets. Finally, we compute $\lambda(\cup_{X\in\mathcal{X}}\mathcal{G}_X)$. Setting $\mathcal{Z} = \emptyset$ in the equation (10) yields

$$f(\emptyset) = \lambda((\cup_{X\in\emptyset}\mathcal{G}_X)^c) = \lambda(\mathcal{X}). \tag{13}$$

Similarly, setting $\mathcal{Z} = \mathcal{X}$ in the equation (10) yields

$$\begin{aligned} f(\mathcal{X}) =& \lambda((\cup_{X\in\mathcal{X}}\mathcal{G}_X)^c) \\ =& \lambda(\mathcal{X}) - \lambda(\cup_{X\in\mathcal{X}}\mathcal{G}_X) \quad \text{(definition of } \lambda) \\ =& \lambda(\mathcal{X}) - \sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}|+1}\Delta_{\mathcal{Y}} \quad \text{(Equation (12))}. \end{aligned} \tag{14}$$

Therefore, subtracting (12) from (14) yields

$$\sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}|+1}\Delta_{\mathcal{Y}} = f(\emptyset) - f(\mathcal{X}).$$

Second, we establish necessity (forward direction, $\Rightarrow$), or equivalently, uniqueness of class of estimands consisting of (9). In particular, we prove a stronger result in which uniqueness does not require the Condition 1. The objective is to show that, for all $\mathcal{Y}$, $w(\emptyset, \mathcal{Y}) = 1$ and $w(\mathcal{T}, \mathcal{Y}) = 0$ for $\mathcal{T} \neq \emptyset$. By the normalizing conditions, it suffices to show that for all $\mathcal{Y}$,

$w(\mathcal{T},\mathcal{Y}) = 0$ for $\mathcal{T} \neq \emptyset$, as $w(\emptyset,\mathcal{Y}) = 1$ follows immediately. We prove by the backward induction on $m = |\mathcal{T}| + |\mathcal{Y}| \leq K$ for $\mathcal{T} \subseteq \mathcal{Y}^c$ and $\emptyset \neq \mathcal{Y} \subset \mathcal{X}$.

For base case $m = K$, we must have $\mathcal{T} = \mathcal{Y}^c$, and then the objective is to show $w(\mathcal{Y}^c,\mathcal{Y}) = 0$ for all $\emptyset \neq \mathcal{Y} \neq \mathcal{X}$. To see this, the condition $\text{coef}(\mathcal{X}) = 0$ in Lemma 1 implies that $\sum_{\emptyset\neq\mathcal{Y}\subseteq\mathcal{X}} w(\mathcal{Y}^c,\mathcal{Y}) = 1$. Then $\sum_{\emptyset\neq\mathcal{Y}\subset\mathcal{X}} w(\mathcal{Y}^c,\mathcal{Y}) + w(\emptyset,\mathcal{X}) = 1$. Since the weights are normalized, we have that $w(\emptyset,\mathcal{X}) = 1$, and consequently, $\sum_{\emptyset\neq\mathcal{Y}\subset\mathcal{X}} w(\mathcal{Y}^c,\mathcal{Y}) = 0$. By nonnegativity of the weights, we conclude $w(\mathcal{Y}^c,\mathcal{Y}) = 0$ for all $\mathcal{Y} \neq \mathcal{X}$.

The inductive hypothesis is that provided that $\mathcal{T} \subseteq \mathcal{Y}^c$ and $\emptyset \neq \mathcal{Y} \subset \mathcal{X}$, $w(\mathcal{T},\mathcal{Y}) = 0$, whenever $m < |\mathcal{T}| + |\mathcal{Y}| \leq K$.

The inductive step is to show that, for $\mathcal{T} \subseteq \mathcal{Y}^c$ and $\emptyset \neq \mathcal{Y} \subset \mathcal{X}$, $w(\mathcal{T},\mathcal{Y}) = 0$ holds when $|\mathcal{T}| + |\mathcal{Y}| = m$. Define $\mathcal{Z} = \mathcal{T} \cup \mathcal{Y}$. Then $|\mathcal{Z}| = m$, because $\mathcal{T}$ and $\mathcal{Y}$ are disjoint by definition. Of note, $\mathcal{Z}$ is neither $\emptyset$ or $\mathcal{X}$ for $1 \leq m \leq K-1$, because $\mathcal{Y} \neq \emptyset$. By the condition $\text{coef}(\mathcal{Z}) = 0$ in Lemma 1, we have

$$\sum_{\emptyset\neq\mathcal{Y}'\subseteq\mathcal{X}} (-1)^{|\mathcal{Y}'|+|\mathcal{Z}\cap\mathcal{Y}'|+1} w(\mathcal{Z}\backslash\mathcal{Y}',\mathcal{Y}') = 0. \tag{15}$$

By the inclusion-exclusion principle for the cardinality, it follows that $|\mathcal{Z}\backslash\mathcal{Y}'| + |\mathcal{Y}'| = |\mathcal{Z}\cup\mathcal{Y}'| = |\mathcal{Z}| + |\mathcal{Y}'| - |\mathcal{Z}\cap\mathcal{Y}'| = m + |\mathcal{Y}'| - |\mathcal{Z}\cap\mathcal{Y}'|$. For the case when $|\mathcal{Y}'| - |\mathcal{Z}\cap\mathcal{Y}'| \geq 1$, we have that $|\mathcal{Z}\backslash\mathcal{Y}'| + |\mathcal{Y}'| \geq m+1$, which by the induction hypothesis, $w(\mathcal{Z}\backslash\mathcal{Y}',\mathcal{Y}') = 0$ for all $\mathcal{Y}' \neq \mathcal{X}$. Therefore, the condition in Equation (15) is simplified to

$$\sum_{\emptyset\neq\mathcal{Y}'\subseteq\mathcal{Z}} (-1)^{2|\mathcal{Y}'|+1} w(\mathcal{Z}\backslash\mathcal{Y}',\mathcal{Y}') = 0. \tag{16}$$

In Equation (16), the exponent is odd, which implies that $\sum_{\emptyset\neq\mathcal{Y}'\subseteq\mathcal{Z}} w(\mathcal{Z}\backslash\mathcal{Y}',\mathcal{Y}') = 0$. Since the weights are nonnegative, $w(\mathcal{Z}\backslash\mathcal{Y}',\mathcal{Y}') = 0$ for all $\emptyset \neq \mathcal{Y}' \subseteq \mathcal{Z}$. We conclude if for any $\mathcal{T} \subseteq \mathcal{Y}^c$ and $\emptyset \neq \mathcal{Y} \subset \mathcal{X}$, we can find $\mathcal{Y}'$ such that $(\mathcal{Z}\backslash\mathcal{Y}',\mathcal{Y}') = (\mathcal{T},\mathcal{Y})$. To see this, we note that setting $\mathcal{Y}' = \mathcal{Y}$ (valid since $\mathcal{Y} \subseteq \mathcal{Z} = \mathcal{T}\cup\mathcal{Y}$) implies $w(\mathcal{Z}\backslash\mathcal{Y},\mathcal{Y}) = w((\mathcal{T}\cup\mathcal{Y})\backslash\mathcal{Y},\mathcal{Y}) = w(\mathcal{T},\mathcal{Y}) = 0$, as $\mathcal{T} \subseteq \mathcal{Y}^c$.

$\square$

Proof of Proposition 1

Proof. Without loss of generality, we focus on $w(\mathcal{T}, \mathcal{Y}) = \mathbf{1}(\mathcal{T} = \emptyset)$. That is, the objective is to show that

$$\sum_{\mathcal{Z} \subseteq \mathcal{Y} \setminus \{X_k\}} (-1)^{|\mathcal{Z}|}(f(\mathcal{Z}) - f(\mathcal{Z} \cup \{X_k\})) = \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z}). \tag{17}$$

The left-hand side can be written as

$$\begin{aligned}
\text{LHS} &= \sum_{\mathcal{Z} \subseteq \mathcal{Y} \setminus \{X_k\}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z}) - \sum_{\mathcal{Z} \subseteq \mathcal{Y} \setminus \{X_k\}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z} \cup \{X_k\}) \\
&= \sum_{\mathcal{Z} \subseteq \mathcal{Y}, \{X_k\} \notin \mathcal{Z}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z}) + \sum_{\mathcal{Z}' \subseteq \mathcal{Y}, \{X_k\} \in \mathcal{Z}'} (-1)^{|\mathcal{Z}'|} f(\mathcal{Z}') \\
&= \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} f(\mathcal{Z}),
\end{aligned}$$

where the second equality follows from the change of varible $\mathcal{Z}' = \mathcal{Z} \cup \{X_k\}$. $\square$

Proof of Corollaries 1 and 2

Proof. We provide a unified proof for Corollaries 1 and 2. Specifically, suppose the function transformation $\log \circ g \circ f$ is applied to the building block f. The proposed estimands on the log scale then become

$$\begin{aligned}
\log(\Delta_{\mathcal{Y}}) &= \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} \log(g \circ f(\mathcal{Z})) \\
&= \log \left[\prod_{\mathcal{Z} \subseteq \mathcal{Y}} (g \circ f(\mathcal{Z}))^{(-1)^{|\mathcal{Z}|}} \right].
\end{aligned}$$

By Theorem 3, the inclusion-exclusion sum of $\log(\Delta_{\mathcal{Y}})$ is given by

$$\sum_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{Y}|+1} \log(\Delta_{\mathcal{Y}}) = \log(g \circ f(\emptyset)) - \log(g \circ f(\mathcal{X})),$$

which can be simplified into

$$\log \left(\prod_{\emptyset \neq \mathcal{Y} \subseteq \mathcal{X}} \Delta_{\mathcal{Y}}^{(-1)^{|\mathcal{Y}|+1}} \right) = \log \left(\frac{g \circ f(\emptyset)}{g \circ f(\mathcal{X})} \right).$$

The Corollary 1 is obtained by choosing $g(x) = x$, whereas the Corollary 2 is obtained by setting $g(x) = x/(1-x)$. $\square$

Appendix B

While the methods proposed in the main manuscript (Sections 2.2 and 3.1) are unweighted, the following illustrative example with $K = 2$ demonstrates how these formulations can be adapted for weighted estimands.

Example 5. *For permutable weights, one specifies all possible weight(w)-by-estimand(f) interaction in* $\mathbf{v}$*,*

$$\begin{aligned}\mathbf{v} =& (w(\emptyset, \{X_1\})f(\emptyset), w(\emptyset, \{X_1\})f(\{X_1\}), w(\{X_2\}, \{X_1\})f(\{X_2\}), w(\{X_2\}, \{X_1\})f(\{X_1, X_2\}),\\ & w(\emptyset, \{X_2\})f(\emptyset), w(\emptyset, \{X_2\})f(\{X_2\}), w(\{X_1\}, \{X_2\})f(\{X_1\}), w(\{X_1\}, \{X_2\})f(\{X_1, X_2\}),\\ & w(\emptyset, \{X_1, X_2\})f(\emptyset), w(\emptyset, \{X_1, X_2\})f(\{X_1\}), w(\emptyset, \{X_1, X_2\})f(\{X_2\}), w(\emptyset, \{X_1, X_2\})f(\{X_1, X_2\}))^\top.\end{aligned}$$

Then one can express the estimand contrasts in Definition 3 as the matrix form in the main manuscript Equation (3):

$$\boldsymbol{\Delta} = (\Delta_{\{X_1\}}, \Delta_{\{X_2\}}, \Delta_{\{X_1, X_2\}})^\top = \mathbf{H}\mathbf{v} = \begin{pmatrix} \mathbf{e}_1^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{e}_1^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{e}_2^\top \end{pmatrix} \mathbf{v},$$

where $\mathbf{e}_1 = (1, -1, 1, -1)^\top$ *and* $\mathbf{e}_2 = (1, -1, -1, 1)^\top$*. For invariant weights, the basis* $\mathbf{v}$ *remains the same and the elements of the generator matrix may take values in* $[-1, 1]$*. For example, with equal invariant weights,*

$$\boldsymbol{\Delta} = (\Delta_{\{X_1\}}, \Delta_{\{X_2\}}, \Delta_{\{X_1, X_2\}})^\top = \mathbf{H}\mathbf{v} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 1 & -1 & -1 & 1 \end{pmatrix} \mathbf{v},$$

Once $\mathbf{H}$ *is specified, our previous results apply. For example, Theorem 1 holds:* $\mathcal{R}(\mathbf{H})$ *remains unchanged, as exchanging the labels is equivalent to permuting the first and second rows in* $\mathbf{H}$*.*

We illustrate the proposed framework in the concrete cases $K = 2$ and $K = 3$ in the following two examples. For simplicity, we recycle the notation introduced in Example 2 of the main manuscript. For notational convenience, we use $\{k_1, k_2\}$ as a shorthand for the set of action variables $\{X_{k_1}, X_{k_2}\}$. In factorial experiments, the weights can be derived from a common joint distribution for $\mathcal{X}$, as shown below.

Remark 2. A class of weight functions $w(\mathcal{T}, \mathcal{Y})$ is said to be compatible across $\mathcal{Y}$ if there exists a probability mass function $\widetilde{w} : 2^{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ ($\sum_{\mathcal{A}\subseteq\mathcal{X}} \widetilde{w}(\mathcal{A}) = 1$) and $w(\mathcal{T}, \mathcal{Y})$ satisfies $w(\mathcal{T}, \mathcal{Y}) = \sum_{\mathcal{Z}\subseteq\mathcal{Y}} \widetilde{w}(\mathcal{Z} \cup \mathcal{T})$.

Example 6 ($K = 2$)**.** *We define the following class of unweighted estimands*

$$\begin{aligned}
\Delta_{\{1\}} &= \mu(1,1) - \mu(0,1),\\
\Delta_{\{2\}} &= \mu(1,1) - \mu(1,0),\\
\Delta_{\{1,2\}} &= \mu(1,1) - \mu(0,1) - \mu(1,0) + \mu(0,0).
\end{aligned}$$

They satisfy the inclusion–exclusion decomposition $\Delta_{\{1\}} + \Delta_{\{2\}} - \Delta_{\{1,2\}} = \mu(1,1) - \mu(0,0)$. *Further, the weighted variant is given by*

$$\begin{aligned}
\Delta_{\{1\}} &= w(\emptyset, \{1\})(\mu(1,1) - \mu(0,1)) + w(\{2\}, \{1\})(\mu(1,0) - \mu(0,0)),\\
\Delta_{\{2\}} &= w(\emptyset, \{2\})(\mu(1,1) - \mu(1,0)) + w(\{1\}, \{2\})(\mu(0,1) - \mu(0,0)),\\
\Delta_{\{1,2\}} &= \mu(1,1) - \mu(0,1) - \mu(1,0) + \mu(0,0),
\end{aligned}$$

where the weights are nonnegative and satisfy

$$\begin{aligned}
w(\emptyset, \{1\}) + w(\{2\}, \{1\}) &= 1,\\
w(\emptyset, \{2\}) + w(\{1\}, \{2\}) &= 1.
\end{aligned}$$

For the compatible permutable weights, we mean that there exists a probability mass $\widetilde{w}$ *for a bivariate binary random vector and all the compatible weights are generated from this common probability mass through*

$$\begin{aligned}
w(\emptyset, \{1\}) &= \widetilde{w}(0,1) + \widetilde{w}(1,1),\\
w(\{2\}, \{1\}) &= \widetilde{w}(0,0) + \widetilde{w}(1,0),\\
w(\emptyset, \{2\}) &= \widetilde{w}(1,1) + \widetilde{w}(1,0),\\
w(\{1\}, \{2\}) &= \widetilde{w}(0,1) + \widetilde{w}(0,0),
\end{aligned}$$

where the probability mass $\widetilde{w}$ *is normalized via* $\widetilde{w}(0,0) + \widetilde{w}(0,1) + \widetilde{w}(1,0) + \widetilde{w}(1,1) = 1$. *For invariant weights, the additional restrictions required to guarantee permutation invariance are*

$$w(\emptyset, \{1\}) = w(\emptyset, \{2\}), \quad w(\{2\}, \{1\}) = w(\{1\}, \{2\}).$$

To confirm the permutation invariance property, when the labels of X_1 and X_2 are swapped, we have $\Delta^{\dagger}_{\{1\}} = \Delta_{\{2\}}$, $\Delta^{\dagger}_{\{2\}} = \Delta_{\{1\}}$, and $\Delta^{\dagger}_{\{1,2\}} = \Delta_{\{1,2\}}$. The residual effect for the weighted estimands, $res = \Delta_{\{1\}} + \Delta_{\{2\}} - \Delta_{\{1,2\}} - (\mu(1,1) - \mu(0,0))$, can be written as

$$res = coef(\emptyset)\mu(1,1) + coef(\{1\})\mu(0,1) + coef(\{2\})\mu(1,0) + coef(\{1,2\})\mu(0,0),$$

where

$$\begin{aligned}
coef(\emptyset) &= w(\emptyset, \{1\}) + w(\emptyset, \{2\}) - 2, \\
coef(\{1\}) &= -w(\emptyset, \{1\}) + w(\{1\}, \{2\}) + 1 \\
coef(\{2\}) &= w(\{2\}, \{1\}) - w(\emptyset, \{2\}) + 1, \\
coef(\{1,2\}) &= -\left(w(\{2\}, \{1\}) + w(\{1\}, \{2\})\right).
\end{aligned}$$

Example 7 ($K = 3$)**.** *We draw a third index to denote the state of the third action variable of interest. That is, we define the following unweighted class of estimands*

$$\begin{aligned}
\Delta_{\{1\}} &= \mu(1,1,1) - \mu(0,1,1), \\
\Delta_{\{2\}} &= \mu(1,1,1) - \mu(1,0,1), \\
\Delta_{\{3\}} &= \mu(1,1,1) - \mu(1,1,0), \\
\Delta_{\{1,2\}} &= \mu(1,1,1) - \mu(0,1,1) - \mu(1,0,1) + \mu(0,0,1), \\
\Delta_{\{1,3\}} &= \mu(1,1,1) - \mu(0,1,1) - \mu(1,1,0) + \mu(0,1,0), \\
\Delta_{\{2,3\}} &= \mu(1,1,1) - \mu(1,0,1) - \mu(1,1,0) + \mu(1,0,0), \\
\Delta_{\{1,2,3\}} &= \mu(1,1,1) - \mu(0,1,1) - \mu(1,0,1) - \mu(1,1,0) + \mu(0,0,1) + \mu(0,1,0) + \mu(1,0,0) - \mu(0,0,0).
\end{aligned}$$

They satisfy the inclusion–exclusion decomposition

$$\mu(1,1,1) - \mu(0,0,0) = \sum_{j=1}^{3} \Delta_{\{j\}} - \Delta_{\{1,2\}} - \Delta_{\{1,3\}} - \Delta_{\{2,3\}} + \Delta_{\{1,2,3\}}.$$

Further, the weighted variant is given by

$$
\begin{aligned}
\Delta_{\{1\}} =& w(\emptyset,\{1\})(\mu(1,1,1)-\mu(0,1,1)) + w(\{2\},\{1\})(\mu(1,0,1)-\mu(0,0,1)) + \\
& w(\{3\},\{1\})(\mu(1,1,0)-\mu(0,1,0)) + w(\{2,3\},\{1\})(\mu(1,0,0)-\mu(0,0,0)), \\
\Delta_{\{2\}} =& w(\emptyset,\{2\})(\mu(1,1,1)-\mu(1,0,1)) + w(\{1\},\{2\})(\mu(0,1,1)-\mu(0,0,1)) + \\
& w(\{3\},\{2\})(\mu(1,1,0)-\mu(1,0,0)) + w(\{1,3\},\{2\})(\mu(0,1,0)-\mu(0,0,0)), \\
\Delta_{\{3\}} =& w(\emptyset,\{3\})(\mu(1,1,1)-\mu(1,1,0)) + w(\{1\},\{3\})(\mu(0,1,1)-\mu(0,1,0)) + \\
& w(\{2\},\{3\})(\mu(1,0,1)-\mu(1,0,0)) + w(\{1,2\},\{3\})(\mu(0,0,1)-\mu(0,0,0)), \\
\Delta_{\{1,2\}} =& w(\emptyset,\{1,2\})(\mu(1,1,1)-\mu(0,1,1)-\mu(1,0,1)+\mu(0,0,1)) + \\
& w(\{3\},\{1,2\})(\mu(1,1,0)-\mu(0,1,0)-\mu(1,0,0)+\mu(0,0,0)), \\
\Delta_{\{1,3\}} =& w(\emptyset,\{1,3\})(\mu(1,1,1)-\mu(0,1,1)-\mu(1,1,0)+\mu(0,1,0)) + \\
& w(\{2\},\{1,3\})(\mu(1,0,1)-\mu(0,0,1)-\mu(1,0,0)+\mu(0,0,0)), \\
\Delta_{\{2,3\}} =& w(\emptyset,\{2,3\})(\mu(1,1,1)-\mu(1,0,1)-\mu(1,1,0)+\mu(1,0,0)) + \\
& w(\{1\},\{2,3\})(\mu(0,1,1)-\mu(0,0,1)-\mu(0,1,0)+\mu(0,0,0)), \\
\Delta_{\{1,2,3\}} =& \mu(1,1,1)-\mu(0,1,1)-\mu(1,0,1)-\mu(1,1,0)+\mu(0,0,1)+\mu(0,1,0)+\mu(1,0,0)-\mu(0,0,0),
\end{aligned}
$$

where the weights are nonnegative and normalized such that

$$
\begin{aligned}
& w(\emptyset,\{1\}) + w(\{2\},\{1\}) + w(\{3\},\{1\}) + w(\{2,3\},\{1\}) = 1, \\
& w(\emptyset,\{2\}) + w(\{1\},\{2\}) + w(\{3\},\{2\}) + w(\{1,3\},\{2\}) = 1, \\
& w(\emptyset,\{3\}) + w(\{1\},\{3\}) + w(\{2\},\{3\}) + w(\{1,2\},\{3\}) = 1, \\
& w(\emptyset,\{1,2\}) + w(\{3\},\{1,2\}) = 1, \quad w(\emptyset,\{1,3\}) + w(\{2\},\{1,3\}) = 1, \quad w(\emptyset,\{2,3\}) + w(\{1\},\{2,3\}) = 1.
\end{aligned}
$$

For the compatible permutable weights, we mean that there exists a probability mass $\widetilde{w}$ for a trivariate binary random vector and all the compatible weights are generated from this common probability mass through

$$
\begin{aligned}
&w(\emptyset,\{1\})=\widetilde{w}(0,1,1)+\widetilde{w}(1,1,1), \quad w(\{2\},\{1\})=\widetilde{w}(0,0,1)+\widetilde{w}(1,0,1),\\
&w(\{3\},\{1\})=\widetilde{w}(0,1,0)+\widetilde{w}(1,1,0), \quad w(\{2,3\},\{1\})=\widetilde{w}(0,0,0)+\widetilde{w}(1,0,0),\\
&w(\emptyset,\{2\})=\widetilde{w}(1,1,1)+\widetilde{w}(1,0,1), \quad w(\{1\},\{2\})=\widetilde{w}(0,1,1)+\widetilde{w}(0,0,1),\\
&w(\{3\},\{2\})=\widetilde{w}(1,1,0)+\widetilde{w}(1,0,0), \quad w(\{1,3\},\{2\})=\widetilde{w}(0,1,0)+\widetilde{w}(0,0,0),\\
&w(\emptyset,\{3\})=\widetilde{w}(1,1,0)+\widetilde{w}(1,1,1), \quad w(\{1\},\{3\})=\widetilde{w}(0,1,0)+\widetilde{w}(0,1,1),\\
&w(\{2\},\{3\})=\widetilde{w}(1,0,0)+\widetilde{w}(1,0,1), \quad w(\{1,2\},\{3\})=\widetilde{w}(0,0,0)+\widetilde{w}(0,0,1),\\
&w(\emptyset,\{1,2\})=\widetilde{w}(0,0,1)+\widetilde{w}(0,1,1)+\widetilde{w}(1,0,1)+\widetilde{w}(1,1,1),\\
&w(\{3\},\{1,2\})=\widetilde{w}(0,0,0)+\widetilde{w}(0,1,0)+\widetilde{w}(1,0,0)+\widetilde{w}(1,1,0),\\
&w(\emptyset,\{1,3\})=\widetilde{w}(0,1,0)+\widetilde{w}(0,1,1)+\widetilde{w}(1,1,0)+\widetilde{w}(1,1,1),\\
&w(\{2\},\{1,3\})=\widetilde{w}(0,0,0)+\widetilde{w}(0,0,1)+\widetilde{w}(1,0,0)+\widetilde{w}(1,0,1),\\
&w(\emptyset,\{2,3\})=\widetilde{w}(1,0,0)+\widetilde{w}(1,0,1)+\widetilde{w}(1,1,0)+\widetilde{w}(1,1,1),\\
&w(\{1\},\{2,3\})=\widetilde{w}(0,0,0)+\widetilde{w}(0,0,1)+\widetilde{w}(0,1,0)+\widetilde{w}(0,1,1),
\end{aligned}
$$

where the probability mass $\widetilde{w}$ is normalized via

$$\widetilde{w}(0,0,0)+\widetilde{w}(1,0,0)+\widetilde{w}(0,1,0)+\widetilde{w}(0,0,1)+\widetilde{w}(1,1,0)+\widetilde{w}(1,0,1)+\widetilde{w}(0,1,1)+\widetilde{w}(1,1,1)=1.$$

For invariant weights, the additional restrictions required to guarantee permutation invariance are

$$
\begin{aligned}
&w(\emptyset,\{1\})=w(\emptyset,\{2\})=w(\emptyset,\{3\}),\\
&w(\{2\},\{1\})=w(\{1\},\{2\})=w(\{1\},\{3\})=w(\{3\},\{1\})=(\{2\},\{3\})=(\{3\},\{2\}),\\
&w(\{2,3\},\{1\})=w(\{1,3\},\{2\})=w(\{1,2\},\{3\}),\\
&w(\emptyset,\{1,2\})=w(\emptyset,\{1,3\})=w(\emptyset,\{2,3\}),\\
&w(\{3\},\{1,2\})=w(\{2\},\{1,3\})=w(\{1\},\{2,3\}).
\end{aligned}
$$

To verify the permutation invariance property, consider an example label permutation σ with $\sigma(1)=2$, $\sigma(2)=3$, and $\sigma(3)=1$. In this case, we have

$$
\begin{aligned}
&\Delta^{\dagger}_{\{1\}}=\Delta_{\{2\}}, \quad \Delta^{\dagger}_{\{2\}}=\Delta_{\{3\}}, \quad \Delta^{\dagger}_{\{3\}}=\Delta_{\{1\}},\\
&\Delta^{\dagger}_{\{1,2\}}=\Delta_{\{2,3\}}, \quad \Delta^{\dagger}_{\{1,3\}}=\Delta_{\{1,2\}}, \quad \Delta^{\dagger}_{\{2,3\}}=\Delta_{\{1,3\}}, \quad \Delta^{\dagger}_{\{1,2,3\}}=\Delta_{\{1,2,3\}}.
\end{aligned}
$$

The residual effect for the weighted estimands, res $= \sum_{j=1}^{3} \Delta_{\{j\}} - \Delta_{\{1,2\}} - \Delta_{\{1,3\}} - \Delta_{\{2,3\}} + \Delta_{\{1,2,3\}} - (\mu(1,1,1) - \mu(0,0,0))$, can be written as

$$\begin{aligned}\text{res} =&\text{coef}(\emptyset)\mu(1,1,1) + \text{coef}(\{1\})\mu(0,1,1) + \text{coef}(\{2\})\mu(1,0,1) + \text{coef}(\{3\})\mu(1,1,0)+\\ &\text{coef}(\{1,2\})\mu(0,0,1) + \text{coef}(\{1,3\})\mu(0,1,0) + \text{coef}(\{2,3\})\mu(1,0,0) + \text{coef}(\{1,2,3\})\mu(0,0,0),\end{aligned}$$

where

$$\begin{aligned}\text{coef}(\emptyset) =& w(\emptyset,\{1\}) + w(\emptyset,\{2\}) + w(\emptyset,\{3\}) - w(\emptyset,\{1,2\}) - w(\emptyset,\{1,3\}) - w(\emptyset,\{2,3\}),\\ \text{coef}(\{1\}) =& -w(\emptyset,\{1\}) + w(\{1\},\{2\}) + w(\{1\},\{3\}) + w(\emptyset,\{1,2\})+\\ & w(\emptyset,\{1,3\}) - w(\{1\},\{2,3\}) - 1,\\ \text{coef}(\{2\}) =& -w(\emptyset,\{2\}) + w(\{2\},\{1\}) + w(\{2\},\{3\}) + w(\emptyset,\{1,2\})+\\ & w(\emptyset,\{2,3\}) - w(\{2\},\{1,3\}) - 1,\\ \text{coef}(\{3\}) =& -w(\emptyset,\{3\}) + w(\{3\},\{2\}) + w(\{3\},\{1\}) + w(\emptyset,\{2,3\})+\\ & w(\emptyset,\{1,3\}) - w(\{3\},\{1,2\}) - 1,\\ \text{coef}(\{1,2\}) =& -w(\{2\},\{1\}) - w(\{1\},\{2\}) + w(\{1,2\},\{3\}) - w(\emptyset,\{1,2\})+\\ & w(\{2\},\{1,3\}) + w(\{1\},\{2,3\}) + 1,\\ \text{coef}(\{1,3\}) =& -w(\{3\},\{1\}) - w(\{1\},\{3\}) + w(\{1,3\},\{2\}) - w(\emptyset,\{1,3\})+\\ & w(\{3\},\{1,2\}) + w(\{1\},\{2,3\}) + 1,\\ \text{coef}(\{2,3\}) =& -w(\{2\},\{3\}) - w(\{3\},\{2\}) + w(\{2,3\},\{1\}) - w(\emptyset,\{2,3\})+\\ & w(\{2\},\{1,3\}) + w(\{3\},\{1,2\}) + 1,\\ \text{coef}(\{1,2,3\}) =& -(w(\{2,3\},\{1\}) + w(\{1,3\},\{2\}) + w(\{1,2\},\{3\}) + w(\{3\},\{1,2\})+\\ & w(\{2\},\{1,3\}) + w(\{1\},\{2,3\}))\,.\end{aligned}$$

References

Dasgupta, T., N. S. Pillai, and D. B. Rubin (2015). Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology 77*(4), 727–753.

Hartwig, F. P., G. Davey Smith, and J. Bowden (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology 46*(6), 1985–1998.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Imai, K., L. Keele, and T. Yamamoto (2010, February). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Lange, T., M. Rasmussen, and L. C. Thygesen (2014). Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology 179*(4), 513–518.

Rota, G.-C. (1964). On the foundations of combinatorial theory: I. theory of möbius functions. In *Classic Papers in Combinatorics*, pp. 332–360. Springer.

Tchetgen, E. J. T. and I. Shpitser (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *The Annals of Statistics 40*(3), 1816.

Van Lint, J. H. and R. M. Wilson (2001). *A course in combinatorics*. Cambridge university press.

Xia, F. and K. C. G. Chan (2022). Decomposition, identification and multiply robust estimation of natural mediation effects with multiple mediators. *Biometrika 109*(4), 1085–1100.

Zhao, A. and P. Ding (2022). Regression-based causal inference with factorial experiments: estimands, model specifications and design-based properties. *Biometrika 109*(3), 799–815.