Journal of Machine Learning Research 17 (2016) 1-53 Submitted 4/15; Revised 8/15; Published 4/16

Causal Inference through a Witness Protection Program

Ricardo Silva RICARDO@STATS.UCL.AC.UK
Department of Statistical Science and CSML
University College London
London WC1E 6BT, UK

Robin Evans EVANS@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford OX1 3TG, UK

Editor: Kevin Murphy

Abstract

One of the most fundamental problems in causal inference is the estimation of a causal effect when treatment and outcome are confounded. This is difficult in an observational study, because one has no direct evidence that all confounders have been adjusted for. We introduce a novel approach for estimating causal effects that exploits observational conditional independencies to suggest "weak" paths in an unknown causal graph. The widely used faithfulness condition of Spirtes et al. is relaxed to allow for varying degrees of "path cancellations" that imply conditional independencies but do not rule out the existence of confounding causal paths. The output is a posterior distribution over bounds on the average causal effect via a linear programming approach and Bayesian inference. We claim this approach should be used in regular practice as a complement to other tools in observational studies.

Keywords: Causal inference, instrumental variables, Bayesian inference, linear programming

1. Contribution

We provide a new methodology for obtaining bounds on the average causal effect (ACE) of a treatment variable X on an outcome variable Y. We introduce methods for binary models and for linear continuous models. For binary variables, the ACE is defined as

$$E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] = P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0)), \quad (1)$$

where $do(\cdot)$ is the operator of Pearl (2000), denoting distributions where a set of variables has been intervened on by an external agent.

In this paper, we assume the reader is familiar with the concept of causal graphs, the basics of the *do* operator, and the basics of causal discovery algorithms such as the PC algorithm of Spirtes et al. (2000). We provide a short summary for context in Section 2.

The ACE is in general not identifiable from observational data. We obtain upper and lower bounds on the ACE by exploiting a set of covariates, which we assume are not affected by X or Y as justified by temporal ordering or other background assumptions. Such covariate sets are often found in real-world problems, and form the basis of many of the

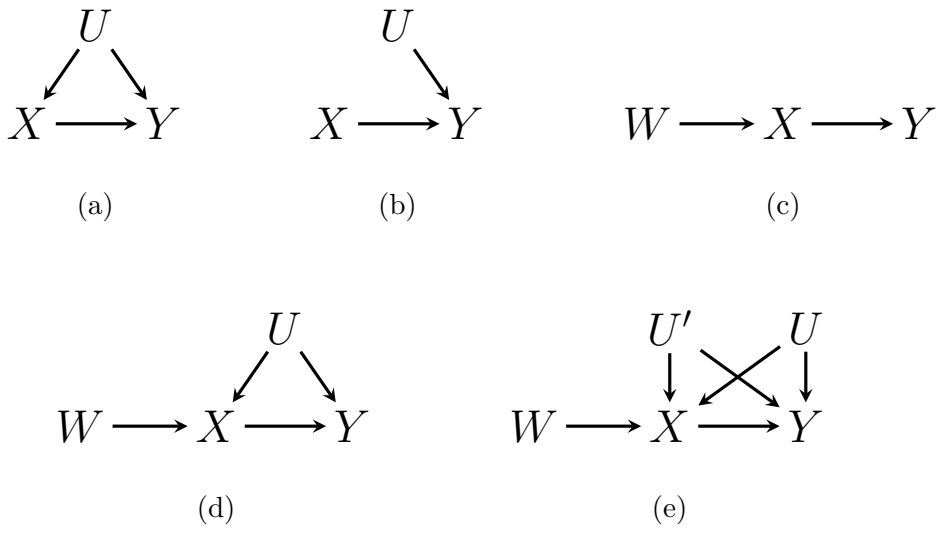


Figure 1: (a) A generic causal graph where X and Y are confounded by some U. (b) The same system in (a) where X is intervened upon by an external agent. (c) A system where W and Y are independent given X. (d) A system where it is possible to use faithfulness to discover that U is sufficient to block all back-door paths between X and Y. (e) Here, U itself is not sufficient.

observational studies done in practice (Rosenbaum, 2002a). However, it is not obvious how to obtain the ACE as a function of the covariates. Our contribution modifies the results of Entner et al. (2013), who exploit conditional independence constraints to obtain point estimates of the ACE but rely on assumptions that might be unstable with finite sample sizes. Our modification provides a different interpretation of their search procedure, which we use to generate candidate *instrumental variables* (Manski, 2007). The linear programming approach of Dawid (2003), inspired by Balke and Pearl (1997) and further refined by Ramsahai (2012), is then modified to generate bounds on the ACE by introducing constraints on some causal paths, motivated as relaxations of Entner et al. (2013). The new setup can be computationally expensive, so we introduce further relaxations to the linear program to generate novel symbolic bounds, and a fast algorithm that sidesteps the full linear programming optimization.

In Section 2, we discuss the background of the problem. In Section 3 we provide an overview of the methodology, which is divided into several subcomponents and described through Sections 4–8. Section 9 contains experiments with synthetic and real data.

2. Background: Instrumental Variables, Witnesses and Admissible Sets

Assuming X is a potential cause of Y, but not the opposite, a cartoon of the possibly complex real-world causal system containing X and Y is shown in Figure 1(a). U represents the universe of common causes of X and Y. In control and policy-making problems, we

would like to know what will happen to the system when the distribution of X is overridden by some external agent (e.g., a doctor, a robot or an economist). The resulting modified system is depicted in Figure 1(b), and represents the family of distributions indexed by $do(X = x)$: the graph in (a) has undergone a "surgery" that removes incoming edges to X. Spirtes et al. (2000) provide an account of the first graphical methods exploiting this idea, which are related to the overriding of structural equations proposed by Haavelmo (1943). Notice that if U is observed in the data set, then we can obtain the distribution $P(Y = y \,|\, do(X = x))$ by simply calculating $\sum_u P(Y = y \,|\, X = x, U = u)P(U = u)$ (Spirtes et al., 2000). This was popularized by Pearl (2000) as the *back-door adjustment*. In general $P(Y = y \,|\, do(X = x))$ can be substantially different from $P(Y = y \,|\, X = x)$.

The ACE can usually be estimated via a trial in which X is randomized: this is equivalent to estimating the conditional distribution of Y given X under data generated as in Figure 1(b). In contrast, in an *observational study* (Rosenbaum, 2002a) we obtain data generated by the system in Figure 1(a). If one believes all relevant confounders U have been recorded in the data then the back-door adjustment can be used, though such completeness is uncommon. By postulating knowledge of the causal graph relating components of U, one can infer whether a measured subset of the causes of X and Y is enough (Pearl, 2000; VanderWeele and Shpitser, 2011; Pearl, 2009). Without knowledge of the causal graph, assumptions such as *faithfulness* (Spirtes et al., 2000) are used to infer it.

The faithfulness assumption states that a conditional independence constraint in the observed distribution exists if and only if a corresponding structural independence exists in the underlying causal graph. For instance, observing the independence $W \perp\!\!\!\perp Y \,|\, X$, and assuming faithfulness and the causal order, we can infer the causal graph Figure 1(c); in all the other graphs this conditional independence in not implied. We deduce that no unmeasured confounders between X and Y exist. This simple procedure for identifying chains $W \to X \to Y$ is useful in exploratory data analysis (Chen et al., 2007; Cooper, 1997), where a large number of possible causal relations $X \to Y$ are unquantified but can be screened off using observational data before experiments are performed. The purpose of using faithfulness is to be able to sometimes identify such quantities.

Entner et al. (2013) generalize the discovery of chain models to situations where a non-empty set of covariates is necessary to block all back-doors. Suppose $\mathbf{W}$ is a set of covariates which are known not to be effects of either X or Y, and we want to find an *admissible set* contained in $\mathbf{W}$: a set of observed variables which we can use for back-door adjustment to obtain $P(Y = y \,|\, do(X = x))$. Entner et al.'s "Rule 1" states the following:

Rule 1*: If there exists a variable $W \in \mathbf{W}$ and a set $\mathbf{Z} \subseteq \mathbf{W} \backslash \{W\}$ such that*

$$\textit{(i)} \quad W \not\perp\!\!\!\perp Y \,|\, \mathbf{Z} \qquad\qquad \textit{(ii)} \quad W \perp\!\!\!\perp Y \,|\, \mathbf{Z} \cup \{X\},$$

then infer that $\mathbf{Z}$ is an admissible set.

Entner et al. (2013) also provide ways of identifying zero effects with a "Rule 2." For simplicity of presentation, for now we assume that the effect of interest was already identified as non-zero. Section 7 discusses the case of zero effects.

A point estimate of the ACE can then be found using $\mathbf{Z}$. Given that $(W, \mathbf{Z})$ satisfies Rule 1, we call W a *witness* for the admissible set $\mathbf{Z}$. The model in Figure 1(c) can be

identified with Rule 1, where W is the witness and $\mathbf{Z} = \emptyset$. In this case, for binary models a so-called "naïve" functional $P(Y = 1 \,|\, X = 1) - P(Y = 1 \,|\, X = 0)$ will provide the correct ACE. If U is observable in Figure 1(d), then it can be identified as an admissible set for witness W. Notice that in Figure 1(a), taking U as a scalar, it is not possible to find a witness since there are no remaining variables. Also, if in Figure 1(e) our covariate set $\mathbf{W}$ is $\{W, U\}$, then no witness can be found since U' cannot be blocked. Hence, it is possible for a procedure based on Rule 1 to answer "I don't know," even when a back-door adjustment would be possible *if* one knew the causal graph. However, using the faithfulness assumption alone one cannot do better: Rule 1 is complete for non-zero effects if no further information is available (Entner et al., 2013).

Despite its appeal, the faithfulness assumption is not without difficulties. Even if unfaithful distributions can be ruled out as pathological under seemingly reasonable conditions (Meek, 1995), distributions which lie close to (but not on) an unfaithful model may in practice be indistinguishable from distributions within that unfaithful model at finite sample sizes.

To appreciate these complications, consider the structure in Figure 1(d) with U unobservable and the remaining (observable) variables binary. Here W is randomized but X is not, and we would like to know the ACE of X on Y[1]. W is sometimes known as an *instrumental variable* (IV), and we call Figure 1(d) the *standard IV structure* (SIV); the distinctive features here being the constraints $W \perp\!\!\!\perp U$ and $W \perp\!\!\!\perp Y \mid \{X, U\}$, statements which include latent variables. If this structure is known, optimal bounds

$$\mathcal{L}_{SIV} \leq E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] \leq \mathcal{U}_{SIV}$$

can be obtained without further assumptions, and estimated using only observational data over the binary variables W, X and Y (Balke and Pearl, 1997). This structure cannot be found using faithfulness, as the only independence constraints involve a latent variable. However, there exist distributions faithful to the IV structure but which at finite sample sizes may appear to satisfy the Markov property for the structure $W \to X \to Y$; in practice this can occur at any finite sample size (Robins et al., 2003). The true average causal effect may lie anywhere in the interval $[\mathcal{L}_{SIV}, \mathcal{U}_{SIV}]$, which can be rather wide even when $W \perp\!\!\!\perp Y \,|\, X$, as shown by the following result:

Proposition 1 *If $W \perp\!\!\!\perp Y \,|\, X$ and the model follows the causal structure of the standard IV graph, then $\mathcal{U}_{SIV} - \mathcal{L}_{SIV} = 1 - |P(X = 1 \,|W = 1) - P(X = 1 \,|\, W = 0)|$.*

All proofs in this manuscript are given in Appendix A. For a fixed joint distribution $P(W, X, Y)$, the length of such an interval cannot be further minimized (Balke and Pearl, 1997). Notice that the length of the interval will depend on how strongly associated W and X are: $W = X$ implies $\mathcal{U}_{IV} - \mathcal{L}_{IV} = 0$ as expected, since this is the scenario of a perfect intervention. The scenario where $W \perp\!\!\!\perp X$ is analogous to not having any instrumental variable, and the length of the corresponding interval is 1.

1. A classical example is in non-compliance: suppose W is the assignment of a patient to either drug or placebo, X is whether the patient actually took the medicine or not, and Y is a measure of health status. The doctor controls W but not X. This problem is discussed by Pearl (2000) and Dawid (2003).

Thus, the true ACE may differ considerably from the "naïve" functional supported by Enter et al.'s Rule 1, appropriate for the simpler structure $W \to X \to Y$ but not for the standard IV structure. While we emphasize that this is a worst-case scenario analysis and by itself should not rule out faithfulness as a useful assumption, it is desirable to provide a method that gives greater control over violations of faithfulness.

3. Outline

In **Section 4**, we introduce the main algorithm, the *Witness Protection Program*. The core idea is (i) to *invert the usage of Entner et al.'s Rule 1*, so that pairs $(W, \mathbf{Z})$ should provide an instrumental variable bounding method instead of a back-door adjustment; (ii) express violations of faithfulness as *bounded violations of local independence*; (iii) find bounds on the ACE using *a linear programming formulation*. Unless stated otherwise, it is assumed that all observed variables are binary.

A simplified version of the algorithm is shown in Algorithm 1. This version assumes we know the distribution of the observed variables, $P(\mathbf{W}, X, Y)$, which simplifies the exposition of the method. The loops in Steps 2 and 3 are a search for pairs $(W, \mathbf{Z})$ of witness-admissible sets that satisfy Enter et al.'s Rule 1, done by verifying independence constraints in the given joint distribution. If we assumed faithfulness, the job would be complete: we either obtain an empty set or the true ACE. This is essentially the contribution of Entner et al. (2013).

However, we assume that faithfulness need not hold, and that all variables can be connected to each other in the causal graph, including a set U of hidden common causes of X and Y. At the same time, we cannot allow for arbitrary violations of faithfulness, as the presence of hidden common causes leads to only very weak constraints on the ACE. Instead, we allow for the expression of a subset of possible violations, expressed as "weak edges" on the fully connected causal graph of $W, \mathbf{Z}, X, Y$ and U. The meaning of a "weak edge" is given in detail in Section 4, and it is fully defined by a set of hyperparameters $\aleph$ that needs to be provided to the algorithm, also explained in Section 4. Given $\aleph$, a generalization of the linear programming problem for instrumental variables described by Ramsahai (2012) can be used to find tight lower and upper bounds on the ACE. As the approach provides each witness a degree of protection against faithfulness violations, using a linear program, we call this framework the *Witness Protection Program* (WPP).

Thus, *this procedure unifies back-door adjustments and (a generalization of) instrumental variable approaches in a single framework, while not requiring knowing the true causal graph and relying on assumptions weaker than faithfulness.* This is the main message of the paper.

The output of the algorithm provides a set of lower/upper bounds on the ACE. If one could assume that $\aleph$ is conservative (that is, the actual edges are "weaker" than the ones implied by the set of causal models $(W, X, Y, \mathbf{Z}, U)$ compatible with $\aleph$), then a tight interval containing the ACE will be given by the largest lower bound and the smallest upper bound. However, there are several practical issues that need to be solved, the main ones being: (i) we do not know $P(\mathbf{W}, X, Y)$ and hence it needs to estimated from data; (ii) once statistical errors are introduced, it is not clear how to combine the different constraints implied by the algorithm; (iii) the computational cost of the procedure can be high, particularly if

input : A distribution $P(\mathbf{W}, X, Y)$;
A set of relaxation parameters $\aleph$;
Covariate index set $\mathbf{W}$ and cause-effect indices X and Y;
output: A set of quadruplets $(W, \mathbf{Z}, \mathcal{L}_{W\mathbf{Z}}, \mathcal{U}_{W\mathbf{Z}})$, where $(W, \mathbf{Z})$ is a witness-admissible set pair and $(\mathcal{L}_{W\mathbf{Z}}, \mathcal{U}_{W\mathbf{Z}})$ are a lower and upper bound on the ACE, respectively;

$\mathcal{R} \leftarrow \emptyset$;
for *each* $W \in \mathbf{W}$ **do**
  for *every admissible set* $\mathbf{Z} \subseteq \mathbf{W} \backslash \{W\}$ *identified by* W **do**
    $(\mathcal{L}_{W\mathbf{Z}}, \mathcal{U}_{W\mathbf{Z}}) \leftarrow$ bounds on the ACE as given by $P(W, X, Y, \mathbf{Z})$ and $\aleph$;
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{(W, \mathbf{Z}, \mathcal{L}_{W\mathbf{Z}}, \mathcal{U}_{W\mathbf{Z}})\}$;
  end
end
return $\mathcal{R}$

Algorithm 1: A simplified Witness Protection Program algorithm, assuming the observable distribution $P(\mathbf{W}, X, Y)$ is known.

uncertainty estimates are required; (iv) we would like to have some results for continuous data; (v) the set of hyperparameters $\aleph$ needs to be chosen somehow, and some objective criterion to choose them is important in practice.

Section 4.2 addresses points (i) and (ii) using a Bayesian approach. This requires a likelihood function. Since the latent variable model that includes U is not identifiable, we work directly on the marginal observable distribution under the constraints implied by the linear program. Independence constraints can be tested using Bayesian model selection, but optionally can be ignored in the linear programming step to provide a more stringent test of feasibility of $\aleph$, as the feasible region for the ACE might be empty if the tested independence does not fit the data well enough even if it passes the test. An interpretation of this usage of independence tests is given in **Section 4.3**. We criticize naïve uses of Bayesian inference for latent variable models in **Section 4.4**.

A convenient implication of using Bayesian inference is that credible intervals for the ACE bounds can be computed in a conceptually simple way, using Monte Carlo methods. However, numerically running a linear program for each sample is expensive. A fully analytical solution to the linear program is not known, but a solution to a relaxed version of it can be found in a much cheaper and more numerically stable iterative algorithm (compared to a black-box solver) given in **Section 5**. This addresses point (iii), but bounds are looser than those obtained with a numerical solver as a consequence.

Point (iv) is partially addressed by **Section 6**, where we derive bounding methods for linear models. This complements Entner et al. (2012), which relies on non-Gaussianity and faithfulness using conditions weaker than Rule 1. Conceptually the method can be adapted to discrete non-binary data without major modifications, although presentation gets considerably more complicated. Treating continuous $\mathbf{W}$ is not a theoretical problem (at least by discretizing each W on demand while keeping $\mathbf{Z}$ continuous), although different estimation techniques and parametric assumptions would be required. Likewise, it is theoretically

possible to get bounds on the cumulative distribution function of Y by dichotomizing it at different levels $Y \leq y$, but we will not further discuss these generalizations in this paper.

Section 7 is a final note concerning the main procedure, where we discuss the possibility of exploiting Enter et al.'s Rule 2 for detecting zero effects. Although we do not further analyze this modification in our experiments, this section provides further insights on how WPP is related to sensitivity analysis methods for observational studies previously found in the literature.

Finally, **Section 8** is an extensive discussion on point (v), the choice of ℵ and the need to deal with possibly incoherent bounds, which also relates back to point (ii). While this discussion is orthogonal to the main algorithm, which takes ℵ as a given and it is guaranteed to be at least as conservative as Entner et al. (2013), it is a important practical issue. This section also complements the discussion started in **Section 7** on the relation between WPP and sensitivity analysis methods.

4. The Witness Protection Program

Let $(W, \mathbf{Z})$ be any pair found by a search procedure that decides when Rule 1 holds. W will play the role of an instrumental variable, instead of being discarded. Conditional on $\mathbf{Z}$, the lack of an edge $W \to Y$ can be justified by faithfulness (as $W \perp\!\!\!\perp Y \,|\, \{X, \mathbf{Z}\}$). For the same reason, there should not be any (conditional) dependence between W and a possible unmeasured common parent[2] U of X and Y. Hence, $W \perp\!\!\!\perp U$ and $W \perp\!\!\!\perp Y \mid \{U, X\}$ hold given $\mathbf{Z}$. A standard IV bounding procedure such as (Balke and Pearl, 1997) can then be used conditional on each individual value $\mathbf{z}$ of $\mathbf{Z}$, then averaged over $P(\mathbf{Z})$. That is, we can independently obtain lower and upper bounds $\{\mathcal{L}(\mathbf{z}),\mathcal{U}(\mathbf{z})\}$ for each value $\mathbf{z}$, and bound the ACE by

$$\sum_{\mathbf{z}} \mathcal{L}(\mathbf{z})P(\mathbf{Z}=\mathbf{z}) \leq E[Y \mid do(X=1)] - E[Y \mid do(X=0)] \leq \sum_{\mathbf{z}} \mathcal{U}(\mathbf{z})P(\mathbf{Z}=\mathbf{z}), \qquad (2)$$

since $E[Y \mid do(X=1)] - E[Y \mid do(X=0)] = \sum_{\mathbf{z}}(E[Y \mid do(X=1), \mathbf{Z}=\mathbf{z}] - E[Y \mid do(X=0), \mathbf{Z}=\mathbf{z}])P(\mathbf{Z}=\mathbf{z})$.

Under the assumption of faithfulness and the satisfiability of Rule 1, the calculation of the above interval is redundant, as Rule 1 allows the direct use of the back-door adjustment using $\mathbf{Z}$. Our goal is to not enforce faithfulness, but use Rule 1 as a motivation to allow for a subset of violations of faithfulness, but not arbitrary violations.

In what follows, assume $\mathbf{Z}$ is set to a particular value $\mathbf{z}$ and all references to distributions are implicitly assumed to be defined conditioned on the event $\mathbf{Z}=\mathbf{z}$. That is, for simplicity of notation, we will neither represent nor condition on $\mathbf{Z}$ explicitly. The causal ordering where X and Y cannot precede any other variable is also assumed, as well as the causal ordering between X and Y.

Consider a standard parameterization of a directed acyclic graph (DAG) model, not necessarily causal, in terms of conditional probability tables (CPTs): let $\theta^V_{v.\mathbf{p}}$ represent $P(V = v \mid Par(V) = \mathbf{p})$ where $V \in \{W, X, Y, U\}$ denotes both a random variable and a vertex in the corresponding DAG; $Par(V)$ is the corresponding set of parents of V.

2. In this manuscript, we will sometimes refer to U as a *set* of common parents, although we do not change our notation to bold face to reflect that.

Faithfulness violations occur when independence constraints among observables are not *structural*, but due to "path cancellations." This means that parameter values are arranged so that $W \perp\!\!\!\perp Y \mid X$ holds, but paths connecting W and U, or W and Y, may exist so that either $W \not\perp\!\!\!\perp U$ or $W \not\perp\!\!\!\perp Y \mid \{U, X\}$. In this situation, some combination of the following should hold true:

$$\begin{array}{rcl} P(Y = y \mid X = x, W = w, U = u) & \neq & P(Y = y \mid X = x, U = u) \\ P(Y = y \mid X = x, W = w, U = u) & \neq & P(Y = y \mid X = x, W = w) \\ P(X = x \mid W = w, U = u) & \neq & P(X = x \mid W = w) \\ P(U = u \mid W = w) & \neq & P(U = u), \end{array} \tag{3}$$

for some $\{w, x, y, u\}$ in the sample space of $P(W, X, Y, U)$.

For instance, if the second and third statements above are true, this implies the existence of an active path into X and Y via U, conditional on W[3], such as $X \leftarrow U \rightarrow Y$. If the first statement is true, this corresponds to an active path between W and Y that is not blocked by $\{X, U\}$. If the fourth statement is true, U and W are marginally dependent, with a corresponding active path. Notice that some combinations are still compatible with a model where $W \perp\!\!\!\perp U$ and $W \perp\!\!\!\perp Y \mid \{U, X\}$ hold: if the second statement in (3) is false, this does not mean that U is necessarily a common parent of X and Y. Such a family of models is observationally equivalent[4] to one where U is independent of all variables.

When translating the conditions (3) into parameters $\{\theta^V_{v.\mathbf{p}}\}$, we need to define parent sets for each vertex, which only need to respect the partial causal ordering being assumed; similarly to the Prediction Algorithm of Spirtes et al. (2000), we do not need to fully specify a causal model in order to identify some of its interventional distributions. In our conditional probability table (CPT) factorization, we define $Par(X) = \{W, U\}$ and $Par(Y) = \{W, X, U\}$. The joint distribution of $\{W, U\}$ can be factorized arbitrarily: the causal directionality among W, U (and $\mathbf{Z}$) is not relevant to the only interventional distribution of interest, $do(X)$. In the next subsection, we refine the parameterization of our model by introducing *redundancies*: we provide a parameterization for the latent variable model $P(W, X, Y, U)$, the interventional distribution $P(W, Y, U \mid do(X))$ and the corresponding (latent-free) marginals $P(W, X, Y)$, $P(W, Y \mid do(X))$. These parameters cannot vary fully independently of each other. It is this fact that will allow us to bound the ACE using only $P(W, X, Y)$.

4.1 Encoding Faithfulness Relaxations with Linear Constraints

We define a *relaxation of faithfulness* as any set of assumptions that allows the relations in (3) to be true, but not necessarily in an *arbitrary* way: this means that while the left-hand and right-hand sides of each entry of (3) are indeed different, their difference is bounded by either the absolute difference or by ratios. Without such restrictions, (3) will only imply uninteresting bounds, as discussed in our presentation of Proposition 1.

Consider the following parameterization of the distribution of $\{W, X, Y, U\}$ under the observational and interventional regimes, and their respective marginals obtained by in-

3. That is, a path that d-connects X and Y and includes U, conditional on W; it is "into" X (and Y) because the edge linking X to the path points to X. See Spirtes et al. (2000) and Pearl (2000) for formal definitions and more examples.
4. Meaning a family of models where $P(W, X, Y)$ satisfies the same constraints.

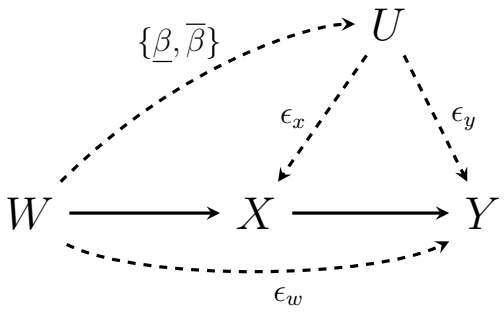


Figure 2: A visual depiction of the family of assumptions introduced in our framework. Dashed edges correspond to conditional dependencies that are constrained according to free parameters, displayed along each corresponding edge. This is motivated by observing $W \perp\!\!\!\perp Y \mid X$.

tegrating U away[5]. Again we condition everywhere on a particular value $\mathbf{z}$ of $\mathbf{Z}$ but, for simplicity of presentation, we suppress this from our notation, since it is not crucial to the developments in this section:

$$\begin{aligned}
\zeta^{\star}_{yx.w} &\equiv P(Y=y, X=x \,|\, W=w, U) \\
\zeta_{yx.w} &\equiv \textstyle\sum_U P(Y=y, X=x \,|\, W=w, U) P(U \,|\, W=w) \\
&= P(Y=y, X=x \,|\, W=w)
\end{aligned}$$

$$\begin{aligned}
\eta^{\star}_{xw} &\equiv P(Y=1 \,|\, X=x, W=w, U) \\
\eta_{xw} &\equiv \textstyle\sum_U P(Y=1 \,|\, X=x, W=w, U) P(U \,|\, W=w) \\
&= P(Y=1 \,|\, do(X=x), W=w)
\end{aligned}$$

$$\begin{aligned}
\delta^{\star}_{w} &\equiv P(X=1 \,|\, W=w, U) \\
\delta_{w} &\equiv \textstyle\sum_U P(X=x \,|\, W=w, U) P(U \,|\, W=w) \\
&= P(X=1 \,|\, W=w).
\end{aligned}$$

Under this encoding, the ACE is given by

$$\eta_{11} P(W=1) + \eta_{10} P(W=0) - \eta_{01} P(W=1) - \eta_{00} P(W=0). \tag{4}$$

Notice that we do not explicitly parameterize the marginal of U, for reasons that will become clear later.

We introduce the following assumptions, as illustrated by Figure 2:

$$|\eta^{\star}_{x1} - \eta^{\star}_{x0}| \leq \epsilon_w \tag{5}$$

$$|\eta^{\star}_{xw} - P(Y=1 \,|\, X=x, W=w)| \leq \epsilon_y \tag{6}$$

$$|\delta^{\star}_{w} - P(X=1 \,|\, W=w)| \leq \epsilon_x \tag{7}$$

$$\underline{\beta} P(U) \leq P(U \,|\, W=w) \leq \overline{\beta} P(U). \tag{8}$$

5. Notice from the development in this section that U is not necessarily a scalar, nor discrete.

Setting $\epsilon_w = 0$, $\underline{\beta} = \overline{\beta} = 1$ recovers the standard IV structure. Further assuming $\epsilon_y = \epsilon_x = 0$ recovers the chain structure $W \to X \to Y$. Under this parameterization in the case $\epsilon_y = \epsilon_x = 1$, $\underline{\beta} = \overline{\beta} = 1$, Ramsahai (2012), extending Dawid (2003), used linear programming to obtain bounds on the ACE. We will briefly describe the four main steps of the framework of Dawid (2003), and refer to the cited papers for more details of their implementation.

For now, assume that $\zeta_{yx.w}$ and $P(W = w)$ are known constants—that is, treat $P(W, X, Y)$ as known. This assumption will be dropped later. Dawid's formulation of a bounding procedure for the ACE is as follows.

Step 1 *Notice that parameters $\{\eta^\star_{xw}\}$ take values in a 4-dimensional polytope. Find the extreme points of this polytope. Do the same for $\{\delta^\star_w\}$.*

In particular, for $\epsilon_w = \epsilon_y = 1$, the polytope of feasible values for the four dimensional vector $(\eta^\star_{00}, \eta^\star_{01}, \eta^\star_{10}, \eta^\star_{11})$ is the unit hypercube $[0, 1]^4$, a polytope with a total of 16 vertices $(0, 0, 0, 0), (0, 0, 0, 1), \ldots (1, 1, 1, 1)$. Dawid (2003) covered the case $\epsilon_w = 0$, where a two-dimensional vector $\{\eta^\star_x\}$ replaces $\{\eta^\star_{xw}\}$. In Ramsahai (2012), the case $0 \leq \epsilon_w < 1$ is also covered: some of the corners in $[0, 1]^4$ disappear and are replaced by others. The case where $\epsilon_w = \epsilon_x = \epsilon_y = 1$ is vacuous, in the sense that the consecutive steps cannot infer non-trivial constraints on the ACE.

Step 2 *Find the extreme points of the joint space $\{\zeta^\star_{yx.w}\} \times \{\eta^\star_{xw}\}$ by mapping them from the extreme points of $\{\delta^\star_w\} \times \{\eta^\star_{xw}\}$, since $\zeta^\star_{yx.w} = (\delta^\star_w)^x (1 - \delta^\star_w)^{(1-x)} \eta^\star_{xw}$.*

The extreme points of the joint space $\{\delta^\star_w\} \times \{\eta^\star_{xw}\}$ are just the combination of the extreme points of each space. Some combinations $\delta^\star_w \times \eta^\star_{xw}$ map to the same $\zeta^\star_{yx.w}$, while the mapping from a given $\delta^\star_w \times \eta^\star_{xw}$ to $\eta^\star_{xw}$ is just the trivial projection. At this stage, we obtain all the extreme points of the polytope $\{\zeta^\star_{yx.w}\} \times \{\eta^\star_{xw}\}$ that are entailed by the factorization of $P(W, X, Y, U)$ and our constraints.

Step 3 *Using the extreme points of the joint space $\{\zeta^\star_{yx.w}\} \times \{\eta^\star_{xw}\}$, find the dual polytope of this space in terms of linear inequalities. Points in this polytope are convex combinations of $\{\zeta^\star_{yx.w}\} \times \{\eta^\star_{xw}\}$, shown by Dawid (2003) to correspond to the marginalizations over some $P(U)$, and each $P(U)$ corresponds to some point in the polytope. This results in constraints over $\{\zeta_{yx.w}\} \times \{\eta_{xw}\}$.*

This is the core step in Dawid (2003): points in the polytope $\{\zeta_{yx.w}\} \times \{\eta_{xw}\}$ correspond to different marginalizations of U according to different $P(U)$. Describing the polytope in terms of inequalities provides all feasible distributions that result from marginalizing U according to some $P(U)$. Because we included both $\zeta^\star_{yx.w}$ and $\eta^\star_{xw}$ in the same space, this will tie together $P(Y, X \mid W)$ and $P(Y \mid do(X), W)$.

Step 4 *Finally, maximize/minimize (4) with respect to $\{\eta_{xw}\}$ subject to the constraints found in Step 3 to obtain upper/lower bounds on the ACE.*

Allowing for the case where $\epsilon_x < 1$ or $\epsilon_y < 1$ is just a matter of changing the first step, where box constraints are set on each individual parameter as a function of the known $P(Y = y, X = x \,|\, W = w)$, prior to the mapping in Step 2. The resulting constraints are now implicitly non-linear in $P(Y = y, X = x \,|\, W = w)$, but at this stage this does not matter as the distribution of the observables is treated as a constant. That is, each resulting constraint in Step 3 is a linear function of $\{\eta_{xw}\}$ and a multilinear function on $\{\{\zeta_{yx.w}\}, \epsilon_x, \epsilon_y, \epsilon_w, \overline{\beta}, \underline{\beta}, P(W)\}$, as discussed in Section 5. Within the objective function (4), the only decision variables are $\{\eta_{xw}\}$, and hence Step 4 still sets up a linear programming problem even if there are multiplicative interactions between $\{\zeta_{yx.w}\}$ and other parameters.

To allow for the case $\underline{\beta} < 1 < \overline{\beta}$, we substitute every occurrence of $\zeta_{yx.w}$ due to the dualization in Step 3 above[6] by $\kappa_{yx.w} \equiv \sum_U \zeta^\star_{yx.w} P(U)$; notice the difference between $\kappa_{yx.w}$ and $\zeta_{yx.w}$. Likewise, we substitute every occurrence of η_{xw} in the constraints by $\omega_{xw} \equiv \sum_U \eta^\star_{xw} P(U)$. Instead of plugging in constants for the values of $\kappa_{yx.w}$ and turning the crank of a linear programming solver, we treat $\{\kappa_{yx.w}\}$ (and $\{\omega_{xw}\}$) as unknowns, linking them to observables and η_{xw} by the constraints

$$\begin{array}{rcl} \eta_{xw}/\overline{\beta} \leq & \omega_{xw} & \leq \min(1, \eta_{xw}/\underline{\beta}), \\ \zeta_{yx.w}/\overline{\beta} \leq & \kappa_{yx.w} & \leq \zeta_{yx.w}/\underline{\beta}, \end{array} \tag{9}$$

$$\sum_{yx} \kappa_{yx.w} = 1. \tag{10}$$

Finally, the steps requiring finding extreme points and converting between representations of a polytope can be easily implemented using a package such as Polymake[7] or the SCDD package[8] for R. Once bounds are obtained for each particular value of $\mathbf{Z}$, Equation (2) is used to obtain the unconditional bounds assuming $P(\mathbf{Z})$ is known.

In Section 8, we provide some guidance on how to choose the free parameters of the relaxation. However, it is relevant to point out that any choice of $\epsilon_w \geq 0, \epsilon_y \geq 0, \epsilon_x \geq 0, 0 \leq \underline{\beta} \leq 1 \leq \overline{\beta}$ is *guaranteed to provide bounds that are at least as conservative* as the back-door adjusted point estimator of Entner et al. (2013), which is always covered by the bounds. Background knowledge, after a user is suggested a witness and admissible set, can also be used to set relaxation parameters.

So far, the linear programming formulated through Steps 1–4 assumes one has already identified an appropriate witness W and admissible set $\mathbf{Z}$, and that the joint distribution $P(W, X, Y, \mathbf{Z})$ is known. In the next section, we discuss how this procedure is integrated with statistical inference for $P(W, X, Y, \mathbf{Z})$ and the search procedure of Entner et al. (2013).

6. Notice the subtlety: the values of $P(y, x \mid w)$ appear within the extreme points of $\{\zeta^\star_{yx.w}\} \times \{\eta^\star_{xw}\}$, but here we are only concerned about the symbols $\zeta_{yx.w}$ emerging from convex combinations of $\zeta^\star_{yx.w}$. In the original formulation of Dawid (2003), $\kappa_{yx.w} = P(y, x \mid w)$ is satisfied, because $P(U) = P(U \mid W)$ is assumed, but in our case in general this will not be true. Hence, the need for a different symbol. Ramsahai (2012) deals with the $P(U) \neq P(U \mid W)$ relaxation in a different way by conditioning on each value of W, but his ACE intervals always include zero.
7. `http://www.poymake.org`
8. `http://cran.r-project.org/`

input : A binary data matrix $\mathcal{D}$;
A set of relaxation parameters $\aleph$;
A covariate index set $\mathbf{W}$ and cause-effect indices X and Y;
output: A set of triplets $(W, \mathbf{Z}, \mathcal{B})$, where $(W, \mathbf{Z})$ is a witness-admissible set pair contained in $\mathbf{W}$ and $\mathcal{B}$ is a distribution over lower/upper bounds on the ACE implied by the pair

$\mathcal{R} \leftarrow \emptyset$;
for *each* $W \in \mathbf{W}$ **do**
  for *every admissible set* $\mathbf{Z} \subseteq \mathbf{W} \backslash \{W\}$ *identified by* W *given* $\mathcal{D}$ **do**
    $\mathcal{B} \leftarrow$ posterior over lower/upper bounds on the ACE as given by $(W, \mathbf{Z}, X, Y, \mathcal{D}, \aleph)$;
    if *there is no evidence in* $\mathcal{B}$ *to falsify the* $(W, \mathbf{Z}, \aleph)$ *model* **then**
      $\mathcal{R} \leftarrow \mathcal{R} \cup \{(W, \mathbf{Z}, \mathcal{B})\}$;
    end
  end
end
return $\mathcal{R}$

Algorithm 2: The outline of the Witness Protection Program algorithm.

4.2 Bayesian Learning and Result Summarization

In the previous section, we treated (the conditional) $\zeta_{yx.w}$ and $P(W = w)$ as known. A common practice is to replace them by plug-in estimators (and in the case of a non-empty admissible set $\mathbf{Z}$, an estimate of $P(\mathbf{Z})$ is also necessary). Such models can also be falsified, as the constraints generated are typically only supported by a strict subset of the probability simplex. In principle, one could fit parameters without constraints, and test the model by a direct check of satisfiability of the inequalities using the plug-in values. However, this does not take into account the uncertainty in the estimation. For the standard IV model, Ramsahai and Lauritzen (2011) discuss a proper way of testing such models in a frequentist sense.

Our models can be considerably more complicated. Recall that constraints will depend on the extreme points of the $\{\zeta^{\star}_{yx.w}\}$ parameters. As implied by (6) and (7), extreme points will be functions of $\zeta_{yx.w}$. Writing the constraints fully in terms of the observed distribution will reveal non-linear relationships. We approach the problem in a Bayesian way. We will assume first the dimensionality of $\mathbf{Z}$ is modest (say, 10 or less), as this is the case in most applications of faithfulness to causal discovery. We parameterize $\zeta^{\mathbf{z}}_{yxw} \equiv P(Y = y, X = x, W = w \mid \mathbf{Z} = \mathbf{z})$ as a full $2 \times 2 \times 2$ contingency table[9]. In the context of the linear programming problem of the previous section, for a given $\mathbf{z}$, we have $\zeta_{yx.w} = \zeta_{yxw}/P(W = w)$, $P(W = w) = \sum_{yx} \zeta_{yxw}$.

Given that the dimensionality of the problem is modest, we assign to each three-variate distribution $P(Y, X, W \mid \mathbf{Z} = \mathbf{z})$ an independent Dirichlet prior for every possible assignment of $\mathbf{Z}$, constrained by the inequalities implied by the model (and renormalized accordingly).

9. That is, we allow for dependence between W and Y given $\{X, \mathbf{Z}\}$, interpreting the decision of independence used in Rule 1 as being only an indicator of approximate independence.

The posterior is also a 8-dimensional constrained Dirichlet distribution, where we use rejection sampling to obtain a posterior sample by proposing from the unconstrained Dirichlet. A Dirichlet prior is assigned to $P(\mathbf{Z})$. Using a sample from the posterior of $P(\mathbf{Z})$ and a sample (for each possible value $\mathbf{z}$) from the posterior of $P(Y, X, W \mid \mathbf{Z} = \mathbf{z})$, we obtain a sample upper and lower bound for the ACE by just running the linear program for each sample of $\{\eta^{\mathbf{z}}_{yxw}\}$ and $\{P(\mathbf{Z} = \mathbf{z})\}$.

The full algorithm is shown in Algorithm 2, where $\aleph \equiv \{\epsilon_w, \epsilon_x, \epsilon_y, \underline{\beta}, \overline{\beta}\}$. The search procedure is left unspecified, as different existing approaches can be plugged into this step. See Entner et al. (2013) for a discussion. In Section 9 we deal with small dimensional problems only, using the brute-force approach of performing an exhaustive search for $\mathbf{Z}$. In practice, brute-force can be still valuable by using a method such as discrete PCA (Buntine and Jakulin, 2004) to reduce $\mathbf{W}\backslash\{W\}$ to a small set of binary variables. To decide whether the premises in Rule 1 hold, we merely perform Bayesian model selection with the BDeu score (Buntine, 1991) between the full graph $\{W \to X, W \to Y, X \to Y\}$ (conditional on $\mathbf{Z}$) and the graph with the edge $W \to Y$ removed.

Step 5 in Algorithm 2 is a "falsification test." Since the data might provide a bad fit to the constraints entailed by the model[10], we opt not to accept every pair $(W, \mathbf{Z})$ that passes Rule 1. One possibility is to calculate the posterior distribution of the model where constraints are enforced, and compare it against the posteriors of the saturated model given by the unconstrained contingency table. This requires another prior over the constraint hypothesis and the calculation of the corresponding marginal likelihoods. As an alternative approach, we adopt the pragmatic rule of thumb suggested by Richardson et al. (2011): sample M samples from the $\{\zeta^{\mathbf{z}}_{yxw}\}$ posterior given the *unconstrained* model, and check the proportion of values that are rejected. If more than 95% of them are rejected, we take this as an indication that the proposed model provides a bad fit and reject the given choice of $(W, \mathbf{Z})$.

The final result provides a set of posterior distributions over bounds, possibly contradictory, which should be summarized as appropriate. One possibility is to check for the union of all intervals or, as a simpler alternative, report the lowest of the lower bound estimates and the highest of the upper bound estimates using a point estimate for each bound:

1. for each $(W, \mathbf{Z})$ in $\mathcal{R}$, calculate the posterior expected value of the lower and upper bounds[11];

2. report the interval $\mathcal{L} \leq ACE \leq \mathcal{U}$ where $\mathcal{L}$ is the minimum of the lower bounds and $\mathcal{U}$ the maximum of the upper bounds.

In our experiments, we use a different summary. As we calculate the log-marginal posterior M_1, M_2, M_3, M_4 for the hypotheses $W \not\perp\!\!\!\perp Y \mid \mathbf{Z}$, $W \perp\!\!\!\perp Y \mid \mathbf{Z}$, $W \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{X\}$, $W \not\perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{X\}$, respectively, we use the score

$$(M_1 - M_2) + (M_3 - M_4) \tag{11}$$

10. This is a result of not enforcing $W \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{X\}$ in $\eta^{\mathbf{z}}_{yxw}$.
11. Alternatively to using the expected posterior estimator for the lower/upper bounds, one can, for instance, report the 0.025 quantile of the marginal lower bound distribution and the 0.975 quantile of the marginal upper bound distribution. Notice, however, this does not give a 0.95 credible interval over ACE intervals as the lower bound and the upper bound are dependent in the posterior.

to assess the quality of the bounds obtained with the corresponding witness-admissible set pair. $M_1 - M_2$ and $M_3 - M_4$ are the log-posterior odds for the models required by the premises of Rule 1 against the respective alternatives. Just reporting the posterior of each $(W, \mathbf{Z})$ model to rank witness-admissible set pairs would not be entirely appropriate, as we are comparing models for different random variables. Adding the log-posteriors instead of a different combination is done for the sake of simplicity, contrasted to the idea of comparing the posterior of $W \to X \to Y$ against the other seven combinations of edges involving $\{W, X, Y\}$.

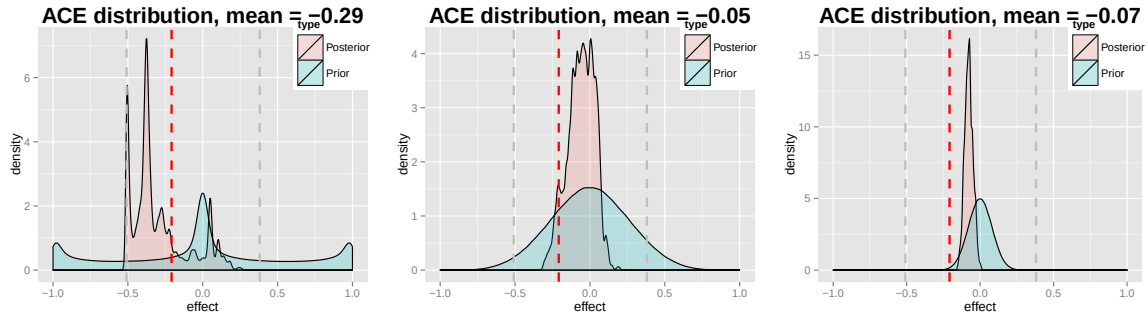
Given the score, we then report the corresponding top-scoring interval and evaluation metric based on this criterion. The reason for reporting a single representative interval is our belief that it is more productive to accommodate most of the (possibly large) discrepancies among estimated ACE intervals in the selection of $\aleph$, as done in Section 8. By selecting a single baseline with a unique lower/upper bound pair, it is simpler to communicate uncertainty, as we can then provide credible intervals or full distributions for the selected lower/upper bounds[12].

4.3 A Note on Weak Dependencies

As we briefly mentioned in the previous section, our parameterization $\{\zeta^{\mathbf{z}}_{yxw}\}$ does not enforce the independence condition $W \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{X\}$ required by Rule 1. Our general goal is to let WPP accept "near independencies," in which the meaning of the symbol $\perp\!\!\!\perp$ in practice means weak dependence[13]. We do not define what a weak dependence should mean, except for the general guideline that some agreed measure of conditional association should be "small." Our pragmatic view on WPP is that Rule 1, when supported by weak dependencies, should be used as a motivation for the constraints in Section 4.1. That is, one makes the assumption that "weak dependencies are not generated by arbitrary near-path cancellations," reflecting the belief that very weak associations should correspond to weak direct causal effects (and, where this is unacceptable, WPP should either be adapted to exclude relevant cases, or not be used). At the same time, users of WPP do not need to accept this view, as the method does not change under the usual interpretation of $\perp\!\!\!\perp$.

However, it is worthwhile to point out that computational gains can be obtained by using a parameterization that encodes the independence: that is, if we change our parameterization to enforce the independence constraint $W \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{X\}$, then there is no need to perform the check in line 5 of Algorithm 2, as the model is compatible with the (conditional on $\mathbf{Z}$) chain $W \to X \to Y$ regardless of $\aleph$. One can then generate full posterior distribution bounds only for those witness-admissible sets for which uncertainty estimates are required. The validity of point estimates of a bound is guaranteed by running a single linear programming on a point estimate of the distribution implied by the Bayesian net-

12. One should not confuse credible intervals with ACE intervals, as these are two separate concepts: each lower or upper bound is a function of the unknown $P(W, X, Y, \mathbf{Z})$ and needs to be estimated. There is posterior uncertainty over each lower/upper bound as in any problem where a functional of a distribution needs to be estimated. So the posterior distribution and the corresponding *credible intervals* over the *ACE intervals* are perfectly well-defined as in any standard Bayesian inference problem.
13. The procedure that decides conditional independencies in Section 4.2 is a method for testing exact independencies, although the prior on the independence assumption regulates how strong the evidence in the data should be for independence to be accepted.



(a) (b) (c)

Figure 3: Posterior over the ACE obtained by three different priors conditioned on a synthetic data set of size 1,000,000. Posterior computed by running 1,000,000 iterations of Gibbs sampling. The (independent) priors for $\theta^Y_{1.xu}$ and $\theta^X_{x.wu}$ are Beta (α, α), while θ^U_u is given a Dirichlet $(\alpha, \alpha, \alpha, \alpha)$. We set $\alpha = 0.1, 1, 10$ for the cases shown in (a), (b) and (c), respectively. Vertical red line shows the true ACE, while the population IV bounds are shown with gray lines. As the prior gets less informative (moving from (c) to (a)), the erratic shape of the posterior distribution also shows the effect of bad Gibbs sampling mixing. Even with a very large data set, the concentration of the posterior is highly dependent on the concentration of the prior.

work $W \to X \to Y$ (for every instance of $\mathbf{Z}$), as no further constraints in the observable distribution will exist. That is, if one wants to enforce the independence constraints, Line 4 of Algorithm 2 can be modified to directly generate just point estimates of the bounds for any witness-admissible set pair where a full posterior distribution is not required, and the falsification test in Step 5 (with the costly polytope construction) can also be skipped.

4.4 A Note on Unidentifiability

An alternative to bounding the ACE or using back-door adjustments is to put priors directly on the latent variable model for $\{W, X, Y, U\}$. Using the standard IV model as an example, we can define parameters $\theta^Y_{y.xu} \equiv P(Y = y \mid X = x, U = u)$, $\theta^X_{x.wu} \equiv P(X = x \mid W = w, U = u)$ and $\theta^U_u \equiv P(U = u)$, on which priors are imposed[14]. No complicated procedure for generating constraints in the observable marginal is necessary, and the approach provides point estimates of the ACE instead of bounds.

This sounds too good to be true, and indeed it is: results strongly depend on the prior, regardless of sample size. To illustrate this, consider a simulation from a standard IV model (Figure 1(c)), with $\mathbf{Z} = \emptyset$ and U an unobservable discrete variable of 4 levels. We generated a model by setting $P(W = w) = 0.5$ and sampling parameters $\theta^Y_{1.xu}$ and $\theta^X_{1.wu}$ from the uniform $[0, 1]$ distribution, while the 4-dimensional vector θ^U_u comes from a

14. $P(W = w)$ is not necessary, as the standard IV bounds (Balke and Pearl, 1997) do not depend on it.

Dirichlet $(1, 1, 1, 1)$. The resulting model had an ACE of -0.20, with a wide IV interval $[-0.50, 0.38]$ as given by the method of Balke and Pearl (1997). Narrower intervals can only be obtained by making more assumptions: there is no free lunch. However, as in this case where WPP cannot identify any witness, one might put priors on the latent variable model to get a point estimate, such as the posterior expected value of the ACE.

To illustrate the pitfalls of this approach, we perform Bayesian inference by putting priors directly on the CPT parameters of the latent variable model, assuming we know the correct number of levels for U. Figure 3 shows some results with a few different choices of priors. The sample size is large enough so that the posterior is essentially entirely within the population bounds and the estimation of $P(W, X, Y, Z)$ is itself nearly exact. The posterior over the ACE covers a much narrower area than the IV interval, and its behavior is erratic.

This is not to say that informative priors on a latent variable model cannot produce important results. For instance, Steenland and Greenland (2004) discuss how empirical priors on smoking habits among blue-collar workers were used in their epidemiological question: the causal effect of the occupational hazard of silica exposure on lung cancer incidence among industrial sand workers. Smoking is a confounding factor given the evidence that smoking and occupation are associated. The issue was that smoking was unrecorded among the workers, and so priors on the latent variable relationship to the observables were necessary. Notice, however, that this informative prior is essentially a way of performing a back-door adjustment when the adjustment set $\mathbf{Z}$ and treatment-outcome pair $\{X, Y\}$ are not simultaneously measured within the same subjects. When latent variables are "unknown unknowns," a prior on $P(Y \mid X, U)$ may be hard to justify. Richardson et al. (2011) discuss more issues on priors over latent variable models as a way of obtaining ACE point estimates, one alternative being the separation of identifiable and unidentifiable parameters to make transparent the effect of prior (mis)specification.

5. Algebraic Bounds and the Back-substitution Algorithm

Posterior sampling is expensive within the context of Bayesian WPP: constructing the dual polytope for possibly millions of instantiations of the problem is time consuming, even if each problem is small. Moreover, the numerical procedure described in Section 4 does not provide any insight on how the different free parameters $\{\epsilon_w, \epsilon_x, \epsilon_y, \underline{\beta}, \overline{\beta}\}$ interact to produce bounds, unlike the analytical bounds available in the standard IV case. Ramsahai (2012) derives analytical bounds under (5) given a *fixed*, *numerical* value of ϵ_w. We know of no previous analytical bounds as an algebraic function of ϵ_w.

5.1 Algebraic Bounds

We derive a set of bounds, whose validity are proved by three theorems. The first theorem derives separate upper and lower bounds on ω_{xw} using all the assumptions except Equation (5); this means constraints which do not link distributions under different values of $W = w$. The second theorem derives linear constraints on $\{\omega_{xw}\}$ using (5) and more elementary constraints. Our final result will construct less straightforward bounds, again using Equation (5) as the main assumption. As before, assume we are implicitly conditioning on some $\mathbf{Z} = \mathbf{z}$ everywhere.

We introduce the notation

$$\begin{array}{rcl} L^{YU}_{xw} & \equiv & \max(P(Y=1|X=x,W=w)-\epsilon_y,0) \\ U^{YU}_{xw} & \equiv & \min(P(Y=1|X=x,W=w)+\epsilon_y,1) \\ L^{XU}_{w} & \equiv & \max(P(X=1|W=w)-\epsilon_x,0) \\ U^{XU}_{w} & \equiv & \min(P(X=1|W=w)+\epsilon_x,1) \end{array}$$

and define $\underline{L} \equiv \min\{L^{YU}_{xw}\}, \overline{U} \equiv \max\{U^{YU}_{xw}\}$. Moreover, some further redundant notation is used to simplify the description of the constraints:

$$\begin{array}{rcl} \delta^{\star}_{1.w} & \equiv & \delta^{\star}_{w} \\ \delta^{\star}_{0.w} & \equiv & 1-\delta^{\star}_{w} \\ L^{XU}_{11} & \equiv & L^{XU}_{1} \\ L^{XU}_{01} & \equiv & 1-U^{XU}_{1} \\ U^{XU}_{11} & \equiv & U^{XU}_{1} \\ U^{XU}_{01} & \equiv & 1-L^{XU}_{1} \end{array}$$

and, following Ramsahai (2012), for any $x \in \{0,1\}$, we define x' as the complementary binary value (i.e. $x' = 1-x$). The same convention applies to pairs $\{w, w'\}$. Finally, define $\chi_{x.w} \equiv \sum_U P(X=x \mid W=w,U)P(U) = \kappa_{1x.w} + \kappa_{0x.w}$.

Theorem 2 *The following constraints are entailed by the assumptions expressed in Equations (6), (7) and (8):*

$$\omega_{xw} \leq \min \begin{cases} \kappa_{1x.w} + U^{YU}_{xw}(\kappa_{0x'.w} + \kappa_{1x'.w}) \\ \kappa_{1x.w}/L^{XU}_{xw} \\ 1-\kappa_{0x.w}/U^{XU}_{xw} \end{cases} \tag{12}$$

$$\omega_{xw} \geq \max \begin{cases} \kappa_{1x.w} + L^{YU}_{xw}(\kappa_{0x'.w} + \kappa_{1x'.w}) \\ \kappa_{1x.w}/U^{XU}_{xw} \\ 1-\kappa_{0x.w}/L^{XU}_{xw} \end{cases} \tag{13}$$

Theorem 3 *The following constraints are entailed by the assumptions expressed in Equations (5), (6), (7) and (8):*

$$\omega_{xw} \leq \min \begin{cases} (\kappa_{1x.w'} + \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/L^{XU}_{xw'} \\ 1-(\kappa_{0x.w'} - \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/U^{XU}_{xw'} \end{cases} \tag{14}$$

$$\omega_{xw} \geq \max \begin{cases} (\kappa_{1x.w'} - \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/U^{XU}_{xw'} \\ 1-(\kappa_{0x.w'} + \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/L^{XU}_{xw'} \end{cases} \tag{15}$$

$$\begin{array}{rcl} \omega_{xw} - \omega_{xw'}U^{XU}_{x'w} & \leq & \kappa_{1x.w} + \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w}) \\ \omega_{xw} - \omega_{xw'}L^{XU}_{x'w} & \geq & \kappa_{1x.w} - \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w}) \\ \omega_{xw} - \omega_{xw'}U^{XU}_{x'w} & \geq & 1-\kappa_{0x.w} - U^{XU}_{x'w} - \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w}) \\ \omega_{xw} - \omega_{xw'}L^{XU}_{x'w} & \leq & 1-\kappa_{0x.w} - L^{XU}_{x'w} + \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w}) \\ \omega_{xw} - \omega_{xw'} & \leq & \epsilon_w \\ \omega_{xw} - \omega_{xw'} & \geq & -\epsilon_w \end{array} \tag{16}$$

Theorem 4 *The following constraints are entailed by the assumptions expressed in Equations (5), (6), (7) and (8):*

$$\omega_{xw} \leq \min \begin{cases} \kappa_{1x'.w'} + \kappa_{1x.w'} + \kappa_{1x.w} - \kappa_{1x'.w} + \chi_{x'w}(\overline{U} + \underline{L} + 2\epsilon_w) - \underline{L} \\ \kappa_{1x'.w} + \kappa_{1x.w} + \kappa_{1x.w'} - \kappa_{1x'.w'} + 2\chi_{x'w}\epsilon_w + \chi_{x'w'}(\overline{U} + \underline{L}) - \underline{L} \end{cases} \tag{17}$$

$$\omega_{xw} \geq \max \begin{cases} -\kappa_{1x'.w'} + \kappa_{1x.w'} + \kappa_{1x'.w} + \kappa_{1x.w} + \chi_{x'w'}(\overline{U} + \underline{L}) - 2\epsilon_w\chi_{x'w} - \overline{U} \\ -\kappa_{1x'.w} + \kappa_{1x.w} + \kappa_{1x'.w'} + \kappa_{1x.w'} - \chi_{x'w}(2\epsilon_w - \overline{U} - \underline{L}) - \overline{U} \end{cases} \tag{18}$$

$$\begin{array}{rcl} \omega_{xw} + \omega_{x'w} - \omega_{x'w'} & \geq & \kappa_{1x'.w} + \kappa_{1x.w} - \kappa_{1x'.w'} + \kappa_{1x.w'} - \chi_{xw'}(\overline{U} + \underline{L} + 2\epsilon_w) + \underline{L} \\ \omega_{xw} + \omega_{x'w'} - \omega_{x'w} & \geq & \kappa_{1x'.w'} + \kappa_{1x.w'} - \kappa_{1x'.w} + \kappa_{1x.w} - 2\chi_{xw'}\epsilon_w - \chi_{xw}(\overline{U} + \underline{L}) + \underline{L} \\ \omega_{xw} + \omega_{x'w'} - \omega_{x'w} & \leq & -\kappa_{1x'.w} + \kappa_{1x.w} + \kappa_{1x'.w'} + \kappa_{1x.w'} - \chi_{xw}(\overline{U} + \underline{L}) + 2\epsilon_w\chi_{xw'} + \overline{U} \\ \omega_{xw} + \omega_{x'w} - \omega_{x'w'} & \leq & -\kappa_{1x'.w'} + \kappa_{1x.w'} + \kappa_{1x'.w} + \kappa_{1x.w} + \chi_{xw'}(2\epsilon_w - \overline{U} - \underline{L}) + \overline{U} \end{array} \tag{19}$$

Although at first such relations seem considerably more complex than those given by Ramsahai (2012), on closer inspection they illustrate qualitative aspects of our free parameters. For instance, consider

$$\omega_{xw} \geq \kappa_{1x.w} + L^{YU}_{xw}(\kappa_{0x'.w} + \kappa_{1x'.w}),$$

one of the instances of (13). If $\epsilon_y = 1$ and $\underline{\beta} = \overline{\beta} = 1$, then $L^{YU}_{xw} = 0$ and this relation collapses to $\eta_{xw} \geq \zeta_{1x.w}$, one of the original relations found by Balke and Pearl (1997) for the standard IV model. Decreasing ϵ_y will linearly increase L^{YU}_{xw} only after $\epsilon_y \leq P(Y = 1 \mid X = x, W = w)$, tightening the corresponding lower bound given by this equation.

Consider now

$$\omega_{xw} \leq 1 - (\kappa_{0x.w'} - \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/U^{XU}_{xw'}.$$

If also $\epsilon_w = 0$ and $\epsilon_x = 1$, from this inequality it follows that $\eta_{xw} \leq 1 - \zeta_{0x.w'}$. This is another of the standard IV inequalities (Balke and Pearl, 1997).

Equation (5) implies $|\omega_{x'w} - \omega_{x'w'}| \leq \epsilon_w$, and as such by setting $\epsilon_w = 0$ we have that

$$\omega_{xw} + \omega_{x'w} - \omega_{x'w'} \geq \kappa_{1x'.w} + \kappa_{1x.w} - \kappa_{1x'.w'} + \kappa_{1x.w'} - \chi_{xw'}(\overline{U} + \underline{L} + 2\epsilon_w) + \underline{L} \tag{20}$$

implies $\eta_{xw} \geq \eta_{1x.w} + \eta_{1x.w'} - \eta_{1x'.w'} - \eta_{0x.w'}$, one of the most complex relationships in (Balke and Pearl, 1997). Further geometric intuition about the structure of the binary standard IV model is given by Richardson and Robins (2010).

These bounds are not tight, in the sense that we opt not to fully exploit all possible algebraic combinations for some results, such as (20): there we use $\underline{L} \leq \eta^{\star}_{xw} \leq \overline{U}$ and $0 \leq \delta^{\star}_{w} \leq 1$ instead of all possible combinations resulting from (6) and (7). The proof idea in Appendix A can be further refined, at the expense of clarity. Because our derivation is a further relaxation, our final bounds are more conservative (that is, looser).

5.2 Efficient Optimization and Falsification Tests

Besides providing insight into the structure of the problem, the algebraic bounds give an efficient way of checking whether a proposed parameter vector $\{\zeta_{yxw}\}$ is valid in Step 5 of Algorithm 2, as well as finding the ACE bounds: we can now use back-substitution on the symbolic set of constraints to find box constraints $\mathcal{L}_{xw} \leq \omega_{xw} \leq \mathcal{U}_{xw}$. The proposed parameter will be rejected whenever an upper bound is smaller than a lower bound, and (4) can be trivially optimized conditioning only on the box constraints—this is yet another relaxation, added on top of the ones used to generate the algebraic inequalities. We initialize by intersecting all algebraic box constraints (of which (12) and (14) are examples); next we refine these by scanning relations $\pm\omega_{xw} - a\omega_{xw'} \leq c$ (the family given by (16)) in lexicographical order, and tightening the bounds of ω_{xw} using the current upper and lower bounds on $\omega_{xw'}$ where possible. We then identify constraints $\mathcal{L}_{xww'} \leq \omega_{xw} - \omega_{xw'} \leq \mathcal{U}_{xww'}$ starting from $-\epsilon_w \leq \omega_{xw} - \omega_{xw'} \leq \epsilon_w$ and the existing bounds, and plug them into relations $\pm\omega_{xw} + \omega_{x'w} - \omega_{x'w'} \leq c$ (as exemplified by (20)) to get refined bounds on ω_{xw} as functions of $(\mathcal{L}_{x'ww'}, \mathcal{U}_{x'ww'})$. We iterate this until convergence, which is guaranteed since lower/upper bounds never decrease/increase at any iteration. This back-substitution of inequalities follows the spirit of message-passing, in the sense that we iteratively update quantities of interest (intervals bounding the decision variables of the linear program) based on a small subset of other quantities, and it can be orders of magnitude more efficient than the fully numerical solution, while not increasing the width of the intervals by too much. In Section 9, we provide evidence for this claim. The back-substitution method is used in our experiments, combined with the fully numerical linear programming approach as explained in Section 9. The full method is given in Algorithm 3.

6. Linear Models

Entner et al. (2012) present a variant of their methodology for linear non-Gaussian models. The main difference is that in this case no witness variable W is necessary: it is possible to validate $\mathbf{Z}$ as an admissible set from a regression of X on $\mathbf{Z}$ and Y on $\{X, \mathbf{Z}\}$. Faithfulness is not necessary under some non-Gaussianity assumptions, although not all of these assumptions are testable without faithfulness and assumptions of parameter stability are still necessary for constraints other than independence constraints.

In this section, we adapt WPP to linear models. Vanishing partial correlations are used in the premise of Rule 1 instead of independence constraints, even if variables are non-Gaussian. The computation of the ACE bounds is vastly simplified. It complements Entner et al. (2012) in the sense that, although the method does not provide point estimates of the ACE and it might fail to identify some admissible sets, it does not require either faithfulness or non-Gaussianity[15]. Only second-order moments are necessary in the construction of the bound, although nonparametric linear models for testing partial correlations or sampling from the posterior distribution of covariance matrices might be necessary.

15. Even if variables are clearly non-Gaussian, the residuals of their regression on the admissible set might be close to Gaussian—this is after all the motivation for Gaussian likelihoods in most regression models, parametric or not.

input : Distributions $\{\zeta_{yx.w}\}$ and $\{P(W=w)\}$;
output: Lower and upper bounds $(\mathcal{L}_{xw}, \mathcal{U}_{xw})$ for every ω_{xw}

Find tightest lower and upper bounds $(\mathcal{L}_{xw}, \mathcal{U}_{xw})$ for each ω_{xw} using inequalities (12), (13) (14), (15), (17) and (18);
Let $\mathcal{L}_{xw}^{\epsilon_w}$ and $\mathcal{U}_{xw}^{\epsilon_w}$ be lower/upper bounds of $\omega_{xw} - \omega_{xw'}$;
for *each pair* $(x, w) \in \{0,1\}^2$ **do**
    $\mathcal{L}_{xw}^{\epsilon_w} \leftarrow -\epsilon_w$;
    $\mathcal{U}_{xw}^{\epsilon_w} \leftarrow \epsilon_w$;
end
while *TRUE* **do**
    for *each relation* $\omega_{xw} - b \times \omega_{xw'} \leq c$ *in (16)* **do**
        $\mathcal{U}_{xw}^{\epsilon_w} \leftarrow \min\{\mathcal{U}_{xw}^{\epsilon_w}, (b-1)\mathcal{L}_{xw} + c\}$
    end
    for *each relation* $\omega_{xw} - b \times \omega_{xw'} \geq c$ *in (16)* **do**
        $\mathcal{L}_{xw}^{\epsilon_w} \leftarrow \max\{\mathcal{L}_{xw}^{\epsilon_w}, (b-1)\mathcal{U}_{xw} + c\}$
    end
    for *each relation* $\omega_{xw} + \omega_{x'w} - \omega_{x'w'} \leq c$ *in (19)* **do**
        $\mathcal{U}_{xw} \leftarrow \min\{\mathcal{U}_{xw}, c - \mathcal{L}_{xw'}^{\epsilon_w}\}$
    end
    for *each relation* $\omega_{xw} - (\omega_{x'w} - \omega_{x'w'}) \leq c$ *in (19)* **do**
        $\mathcal{U}_{xw} \leftarrow \min\{\mathcal{U}_{xw}, c + \mathcal{U}_{xw'}^{\epsilon_w}\}$
    end
    for *each relation* $\omega_{xw} + \omega_{x'w} - \omega_{x'w'} \geq c$ *in (19)* **do**
        $\mathcal{U}_{xw} \leftarrow \max\{\mathcal{U}_{xw}, c - \mathcal{U}_{xw'}^{\epsilon_w}\}$
    end
    for *each relation* $\omega_{xw} - (\omega_{x'w} - \omega_{x'w'}) \geq c$ *in (19)* **do**
        $\mathcal{U}_{xw} \leftarrow \max\{\mathcal{U}_{xw}, c + \mathcal{L}_{xw'}^{\epsilon_w}\}$
    end
    if *no changes in* $\{(\mathcal{L}_{xw}, \mathcal{U}_{xw})\}$ **then**
        break
    end
end
return $(\mathcal{L}_{xw}, \mathcal{U}_{xw})$ *for each* $(x, w) \in \{0,1\}^2$

Algorithm 3: The iterative back-substitution procedure for bounding $\mathcal{L}_{xw} \leq \omega_{xw} \leq \mathcal{U}_{xw}$ for all combinations of x and w in $\{0,1\}^2$.

6.1 A Bounding Procedure for Linear Models

Consider for now the linear model case with an empty admissible set $\mathbf{Z}$:

$$\begin{array}{rcl} X & = & aW + U_x \\ Y & = & bX + cW + U_y, \end{array} \tag{21}$$

where $\{W, U_x, U_y\}$ are assumed to be zero mean variables, and $\{U_x, U_y\}$ are unobservable. The case with non-empty $\mathbf{Z}$ is analogous and discussed in the next section. The ACE is

given by b. We denote as s_{ww}, s_{wx}, s_{wy}, ... the corresponding variances/covariances of $\{W, U_x, U_y\}$. Moreover, let the variances of $\{W, U_x, U_y\}$ be set such that each element of $\{W, X, Y\}$ has unit variance, and denote as $\rho_{wx}, \rho_{wy}, \rho_{xy}$ the corresponding correlations of $\{W, X, Y\}$. Notice that $s_{ww} = 1$, and no assumptions about Gaussianity are being made. As before, we assume for now that $\rho_{wx}, \rho_{wy}, \rho_{xy}$ are known constants, and we would like to bound b as a function of this observable correlation matrix. The *implied* correlation matrix of model (21) needs to match the observable correlation matrix:

$$\rho_{wx} = a + s_{wx} \tag{22}$$
$$\rho_{wy} = b\rho_{wx} + c + s_{wy} \tag{23}$$
$$\rho_{xx} = 1 = a^2 + 2as_{wx} + s_{xx} \tag{24}$$
$$\rho_{xy} = b + c\rho_{wx} + as_{wy} + s_{xy} \tag{25}$$
$$\rho_{yy} = 1 = b^2 + 2bc\rho_{wx} + c^2 + s_{yy} + 2[b(as_{wy} + s_{xy}) + cs_{wy}], \tag{26}$$

where the above identities follow directly from (21). The feasible values of the parameters are given by the intersection of the above and

$$-\epsilon_c \leq c \leq \epsilon_c \tag{27}$$

$$-\epsilon_{wx} \leq s_{wx} \leq \epsilon_{wx}, \quad -\epsilon_{wy} \leq s_{wy} \leq \epsilon_{wy}, \quad -\epsilon_{xy} \leq s_{xy} \leq \epsilon_{xy} \tag{28}$$
$$0 \leq s_{xx} \leq 1, \quad 0 \leq s_{yy} \leq 1. \tag{29}$$

We ignore the positive semidefiniteness requirement of the covariance matrix of $\{W, U_X, U_Y\}$ for simplicity.

The set of constraints can be simplified as follows:

Theorem 5 *Assume*[16] $\rho_{wy} = \rho_{wx}\rho_{xy}$. *If an assignment of values to* $\{a, b, c, s_{wx}, s_{xx}, s_{xy}, s_{xx}\}$ *satisfies (27)-(29), then it satisfies (22)-(26) if and only it satisfies the following:*

$$\rho_{wy} = b\rho_{wx} + c + s_{wy} \tag{30}$$

$$\begin{cases} \rho_{xy} - \epsilon_{xy} - \mathcal{U}_a s_{wy} \leq b + c\rho_{wx} \leq \rho_{xy} + \epsilon_{xy} - \mathcal{L}_a s_{wy}, & \text{if } s_{wy} \geq 0 \\ \rho_{xy} - \epsilon_{xy} - \mathcal{L}_a s_{wy} \leq b + c\rho_{wx} \leq \rho_{xy} + \epsilon_{xy} - \mathcal{U}_a s_{wy}, & \text{if } s_{wy} < 0 \end{cases} \tag{31}$$

$$b^2 + 2bc\rho_{wx} + c^2 - 2(b\rho_{xy} + c\rho_{wy}) \leq 0 \tag{32}$$

where $\mathcal{L}_a \equiv \max(\min(0, 2\rho_{wx}), \rho_{wx} - \epsilon_{wx})$ *and* $\mathcal{U}_a \equiv \min(\max(0, 2\rho_{wx}), \rho_{wx} + \epsilon_{wx})$.

This means that optimizing b subject to constraints (27)−(32) is a convex program on $\{b, c, s_{wy}\}$ (conditioned on the sign of s_{wy}). Notice that, because of the assumption $\rho_{wy} = \rho_{wx}\rho_{xy}$, the system is always satisfiable. It can nevertheless rule out some the possible values of b (e.g. $b = \rho_{xy} = \rho_{wy}/\rho_{wx}$ if $\epsilon_{xy} = \epsilon_{wy}$ or $\epsilon_c = \epsilon_{wy} = 0$). Given $\{\rho_{wx}, \rho_{xy}\}$ (and setting $\rho_{wy} = \rho_{wx}\rho_{xy}$), we can find an upper bound for the ACE by maximizing b under the constraints (27)−(32) and $0 \leq s_{wy} \leq \epsilon_{wy}$, followed by maximization under the condition $-\epsilon_{wy} \leq s_{wy} \leq 0$. The upper bound is the maximum of the two conditional maxima. The lower bound is derived in an analogous way.

16. This assumption can be dropped, but the proof of Theorem 5 gets more complicated.

6.2 Algorithm for Gaussian Copula Models

One general model family in which vanishing partial correlations are closely connected to independence is the Gaussian copula (Elidan, 2013; Nelsen, 2007). Consider the following generative model:

$$\begin{array}{rcl} \mathbf{V}^\star & \sim & \mathcal{N}(0, R) \\ V_i & = & F_i^{-1}(\phi(V_i^\star)), i = 1, 2, \ldots, p, \end{array} \tag{33}$$

where $\mathbf{V}^\star$ is a p-dimensional random vector generated according to the Gaussian with $p \times p$ correlation matrix R, $F_i(\cdot)$ is some arbitrary cumulative distribution function (CDF), and $\phi(\cdot)$ is standard Gaussian CDF. In continuous distributions, $F_i(\cdot)$ is invertible, and Markov properties of $\mathbf{V}^\star$ are preserved in the distribution of $\mathbf{V}$. See Harris and Drton (2013) for a discussion of Gaussian copula models in the context of causal inference, in particular for structure learning using the PC algorithm (Spirtes et al., 2000). Causal structure and effects are defined for $\mathbf{V}^*$ as in a typical linear causal system. Conditional independencies can be tested by copula-based measures, such as Spearman's rank correlation, by testing for the corresponding vanishing partial correlations. Given a target treatment $X \in \mathbf{V}$ and outcome $Y \in \mathbf{V}$, we are interested in bounding the ACE of $X^\star$ in $Y^\star$, the Gaussian variables underlying the possibly non-Gaussian X and Y.

For simplicity, we search for admissible sets $\mathbf{Z}$ with corresponding witness W using a Gaussian copula correlation matrix estimate $\hat{R}$. The unobserved data for $\mathbf{V}^\star$ is then for simplicity assumed to have zero empirical mean and empirical covariance matrix $\hat{R}$. We score models entailing independence of some V_i and V_j given $\mathbf{V}_Z$ by scoring two Gaussian networks, $G_1 \equiv \{\mathbf{V}_Z^\star \to V_i^\star, \mathbf{V}_Z^\star \to V_j^\star\}$ against $G_2 \equiv \{\mathbf{V}_Z^\star \to V_i^\star, \mathbf{V}_Z^\star \to V_j^\star, V_i^\star \to V_j^\star\}$. This is analogous to the binary case, where here we use the corrected BGe score (Kuipers et al., 2014). Notice this test is approximate, as $\hat{R}$ is used as a surrogate for the empirical covariance matrix of the unobserved data $\mathbf{V}^\star$, which is required by BGe[17].

For a given $(W, \mathbf{Z})$ accepted by Rule 1, we calculate the empirical residual (rank) correlation matrix obtained by regressing $W^\star$ on $\mathbf{Z}^\star$, $X^\star$ on $W^\star$ and $\mathbf{Z}^\star$, and $Y^\star$ on $X^\star$ and $\mathbf{Z}^\star$, so that the partial (residual) correlation of W and Y given X and $\mathbf{Z}$ is zero. Regression of subsets of $\mathbf{V}^\star$ on other subsets is done by standard regression using $\hat{R}$: let $\{\hat{\sigma}_{ww.\mathbf{z}}, \hat{\sigma}_{xx.\mathbf{z}}, \hat{\sigma}_{yy.\mathbf{z}}\}$ be the residual variances of the regression of $\{W^\star, X^\star, Y^\star\}$ on $\mathbf{Z}^\star$ as defined by $\hat{R}$. The resulting residual covariance matrix is scaled to unit variance, and the method in Section 6.1 is used to generate scaled bounds $\mathcal{L}_{b_s} \leq b_{standardized} \leq \mathcal{U}_{b_s}$, which are converted in bounds on the ACE in the original scale as $[\sqrt{\hat{\sigma}_{xx}\hat{\sigma}_{yy}}\mathcal{L}_{b_s}, \sqrt{\hat{\sigma}_{xx}\hat{\sigma}_{yy}}\mathcal{U}_{b_s}]$.

The algorithm is basically the same as Algorithm 2, except we report only point estimates for the bounds instead of posteriors, and no falsification step is necessary (Step 5 of Algorithm 2) as the model cannot be falsified given the accepted conditional independence.

17. Alternatively, one could test for the corresponding vanishing partial correlations in the empirical Spearman rank correlation matrix, as suggested by Harris and Drton (2013), at a particular significance level α. However, this only provides p-values, which are not ideal to sort witness/admissible sets by a score, as p-values measure only the surprise of seeing the observed data under a constraint. This does not measure strength of dependence nor a posterior over models. A fully Bayesian version of this approach is conceptually simple, although nonparametric modeling of $\{F_i(\cdot)\}$ (the so-called nonparanormal model) might require Markov chain Monte Carlo methods and computing marginal likelihoods is computationally very intensive.

7. Zero Effects

Under faithfulness, the premise of Rule 1 will not be true in a system where X is not a cause of Y. The result is that no conclusion about the ACE can be made, but identifiability can still be achieved by other means. Rule 2 (Entner et al., 2013) covers all identifiable cases where X is not a cause of Y:

Rule 2a: *If there exists a set* $\mathbf{Z} \subseteq \mathbf{W}$ *such that* $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$*, then infer an ACE of zero.*

Rule 2b: *If there exists a variable* $W \in \mathbf{W}$ *and a set* $\mathbf{Z} \subseteq \mathbf{W} \backslash \{W\}$ *such that:*

$$\textit{(i)} \quad W \not\perp\!\!\!\perp X \,|\, \mathbf{Z} \qquad\qquad \textit{(ii)} \quad W \perp\!\!\!\perp Y \,|\, \mathbf{Z},$$

then infer an ACE of zero.

Rule 2a is a direct application of faithfulness, while Rule 2b essentially corresponds to the "unshielded collider" check in the FCI algorithm of Spirtes et al. (2000). Figure 4 illustrates the paths that can be weakened under Rules 2a and 2b, excluding any a priori restrictions on $X \to Y$, since this is the relation that we want to bound given conditions on other paths. It is clear that for 2a not much of interest can be said beyond this: any association that cancels the causal effect of X and Y should be due to a corresponding association generated by unmeasured confounders. If such a strong contribution due to confounders does not exist, then we should not expect the ACE to be strong[18].

Rule 2b seems more interesting. However, as suggested by Figure 4(b), we are forcing *fewer* structural constraints in the linear program compared to Rule 1, as there is nothing from Rule 2b that motivates weak confounding effects. The reason for that is that Rule 2b concerns the removal of $X \to Y$, which corresponds to the effect we want to bound as a consequence of assumptions elsewhere (instead of assuming a priori, say, $|ACE| \leq \epsilon_\emptyset$ for some new hyperparameter $\epsilon_\emptyset$). One possibility is that for a pair $(W, \mathbf{Z})$ that satisfies Rule 2b, we perform the standard WPP bounding with $\epsilon_x = \epsilon_y = 1$ and, if desired, the added constraint $|ACE| \leq \epsilon_\emptyset$ to be assumed given the firing of Rule 2b.

Another possibility is to exploit Rule 2 to learn something about the possible effects of W on Y: in this case, we condition on constraints $|\eta^\star_{0w} - \eta^\star_{1w}| \leq \epsilon_\emptyset$ to derive bounds on the direct effect of W on Y (Cai et al., 2008). In the context of our ACE problem, it might suggest information about ϵ_w that can be reused in another suggested pair $(W, \mathbf{Z}')$, but this will require further assumptions or tests, as the differences between $\mathbf{Z}$ and $\mathbf{Z}'$ will make this transfer of information not trivial. For the rest of the paper, we will ignore the use of Rule 2 for simplicity. In the next section, however, we will consider the implications of having different pairs of witness/admissible set as a way of learning information about our hyperparameters.

18. This is *not* to say that such an observation has little scientific value. A similar statement is that a strong association between X and Y should be indicative of some causal effect, in the absence of a set of confounders that could fully explain this association. Simple as this is, this type of reasoning has long been explored in observational studies (Cornfield et al., 1959), and it is essentially what is behind Rosenbaum's sensitivity analysis methods (Rosenbaum, 2002a). Our point is that the linear programming approach for this setup is trivial.

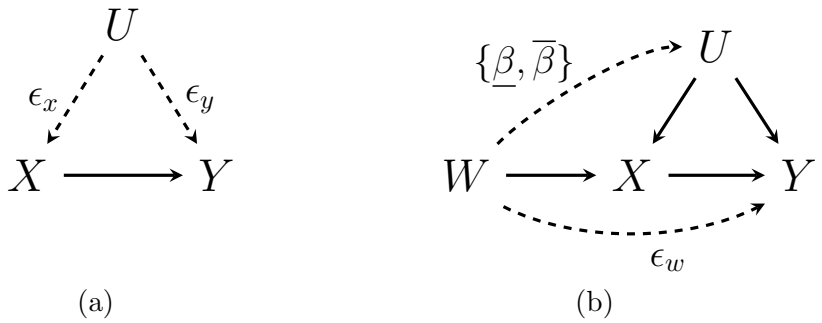


Figure 4: An illustration of conditional dependencies to be weakened under the acceptance of Rule 2. (a) Unmeasured confounding between X and Y is considered when these two variables are (conditionally) independent (given a possibly non-empty set $\mathbf{Z}$). (b) (Conditional) observable independence of W and Y is used to suggest that W and Y have bounded dependence conditioned on U, as well as weak dependence between W and U. Notice no weakening of effects $\{U \to X, U \to Y\}$.

8. Choosing Relaxation Parameters

The free parameters $\aleph \equiv \{\epsilon_w, \epsilon_x, \epsilon_y, \underline{\beta}, \overline{\beta}\}$ do not have a unique, clear-cut, domain-free procedure by which they can be calibrated. However, as we briefly discussed in Section 4, it is useful to state explicitly the following simple guarantee of WPP:

Corollary 6 *Given $W \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ and $W \perp\!\!\!\perp Y \mid \{X, \mathbf{Z}\}$, the* WPP *population bounds on the ACE will always include the back-door adjusted population ACE based on $\mathbf{Z}$.*

Proof The proof follows directly by plugging in the quantities $\epsilon_w = \epsilon_y = \epsilon_x = 0$, $\underline{\beta} = \overline{\beta} = 1$, into the analytical bounds of Section 5.1, which will give the tightest bounds on the ACE (generalized to accommodate a background set $\mathbf{Z}$): a single point, which is also the functional obtained by the back-door adjustment. ■

The implication is that, regardless of the choice of free parameters, the result is guaranteed to be more conservative than the one obtained using the faithfulness assumption. This does not mean that a judicious choice of relaxation parameters is of secondary importance. Ideally, domain knowledge should be used: given a witness and admissible set, an expert decides which relaxations are reasonable. This is domain dependent, and might not be easier than choosing an admissible set from background knowledge. As an alternative, this section covers more generic methods for choosing relaxation parameters. Two main approaches are discussed:

- $\aleph$ is deduced by the outcome of a sensitivity analysis procedure; given a particular interval length L, we derive a quantification of faithfulness violations (represented by $\aleph$) required to generate causal models compatible with the observational data and an interval of length L containing the ACE. This is covered in Section 8.1;

- exploit the multiplicity of solutions (pairs of candidate witness/admissible sets) usually provided by Rule 1 to learn about the extent of possible faithfulness violations. Combine the multiple solutions with constraints or prior distributions for $\aleph$ to obtain estimates of the relaxation parameters. This is covered in Section 8.2.

8.1 Choice by Grid Search Conditioned on Acceptable Information Loss

One pragmatic default method is to first ask how wide an ACE interval can be so that the result is still useful for the goals of the analysis (e.g., sorting possible control variables X as candidates for a lab experiment based on lower bounds on the ACE). Let L be the interval width the analyst is willing to accept. Set $\epsilon_w = \epsilon_x = \epsilon_y = k_\epsilon$ and $\underline{\beta} = c$, $\overline{\beta} = 1/c$, for some pair (k_ϵ, c) such that $0 \leq k_\epsilon < 1$, $0 < c \leq 1$, and let (k_ϵ, c) vary over a grid of values. For each witness/admissible set candidate pair, pick the (k_ϵ, c) choice(s) entailing interval(s) of length closest to L. In case of more than one solution, summarize them by a criterion such the union of the intervals.

This methodology provides an explicit trade-off between length of the interval and tightness of assumptions. Notice that, starting from the back-door adjusted point estimator of Entner et al. (2013), it is not obvious how the trade-off could be obtained: that is, how to build an interval around the back-door point estimate that can be interpreted as bounds under an acceptable amount of information loss. WPP provides a principled way of building such an interval, with the resulting assumptions on $\aleph$ being explicitly revealed as a by-product. If the analyst believes that the resulting values of $\aleph$ are not strict enough, and no substantive knowledge exists that allows particular parameters to be tightened up, then one either has to concede that wider intervals are necessary or to find other means of identifying the ACE without the faithfulness assumption[19].

In the experiments in Section 9.2, we define a parameter space of $k_\epsilon \in \{0.05, 0.10, \ldots, 0.30\}$ and $c \in \{0.9, 1\}$. More than one interval of approximately the same width is identified. For instance, the configurations $(k_\epsilon = 0.25, c = 1)$ and $(k_\epsilon = 0.05, c = 0.9)$ both produced intervals of approximately length 0.30.

8.2 Linking Selection on the Observables to Selection on the Unobservables

Observational studies cannot be carried out without making assumptions that are untestable given the data at hand. There will always be degrees of freedom that must be chosen, even if such choices are open to criticism. The game is to provide a language to express assumptions in as transparent a manner as possible. Our view on priors for the latent variable model (Section 4.4) is that such prior knowledge is far too difficult to justify when the interpretation of U is unclear. Moreover, putting a prior on a parameter such as $P(Y = 1 \,|X = x, W = w, U = u)$ so that this prior is bounded by the constraint $|P(Y = 1 \,|X = x, W = w, U = u) - P(Y = 1 \mid X = w, W = w)| \leq \epsilon_w$ has no clear advantage over

19. That is, expert knowledge should of course still be invoked to decide whether the resulting relaxation is plausible or not (and hence, whether the resulting interval is believable), although communication by sensitivity analysis might facilitate discussion and criticism of the study. In his rejoinder to the discussion of (Rosenbaum, 2002b), Rosenbaum points out that the sensitivity analysis procedure just states the logical outcome of the structural assumptions: the deviation of (say) $P(Y = 1 \mid X = x, W = w)$ from $P(Y = 1 \mid X = x, W = w, U = u)$, required to explain the given magnitude of variation of plausible ACEs, is not imposed a priori by expert knowledge, but deduced.

WPP: a specification of the shape of this prior is still necessary and may have undesirable side effects; it has no computational advantages, as constraints will have to be dealt with now within a Markov chain Monte Carlo procedure; it provides no insight on how constraints are related to one another (Section 5); it still suggests a point estimate that should not be trusted lightly, and posterior bounds which cannot be interpreted as data-driven bounds; and it still requires a choice of ϵ_w.

That is not to say that subjective priors on the relationship between U and the observables cannot be exploited, but the level of abstraction at which they need to be specified should have advantages when compared to the latent variable model approach. For instance, Altonji et al. (2005) introduced a framework to deal with violations of the IV assumptions (in the context of linear models). Their main idea is to linearly decompose the (observational) dependence of W and $\mathbf{Z}$, and the (causal) dependence of Y and $\mathbf{Z}$, as two signal-plus-noise decompositions, and assume that dependence among the signals allows one to infer the dependence among the noise terms. In this linear case, the dependence among noise terms gives the association between W and Y through unmeasured confounders. The constraint given by the assumption can then be used to infer bounds on the (differential) ACE. The details are not straightforward, but the justification for the assumption is indirectly derived by assuming $\mathbf{Z}$ is chosen by a sampling mechanism that picks covariates from the space of confounders U, so that $|\mathbf{Z}|$ and $|U|$ are large. The core idea is that the dependence between the covariates which are observed (i.e. $\mathbf{Z}$) and the other variables (W, X, Y) should tell us something about the impact of the unmeasured confounders. Their method is presented for linear models only, and the justification requires a very large $|\mathbf{Z}|$.

We introduce a very different method inspired by the same general principle, but exploiting the special structure of our procedure. Instead of relying on linearity and a fixed set of covariates, consider the following postulate: the variability of back-door adjusted ACE estimators based on different admissible sets, as implied by Rule 1, should provide some information about the extent of faithfulness violations in the given domain. In what follows, let $\aleph$ be simplified so that $\aleph \equiv \{\epsilon_w, \epsilon_{xy}, \beta\}$, where $\epsilon_{xy} \equiv \epsilon_x = \epsilon_y$ and $\beta \equiv \overline{\beta} = 1/\underline{\beta}$. The task then is to choose the three parameters in this set.

8.2.1 Method 1: Tightest ACE Coverage

Let $\{(W_1, \mathbf{Z}_1), \ldots, (W_k, \mathbf{Z}_k)\}$ be the set of all pairs found by WPP. Let A_i be the ACE calculated by the back-door adjustment on $\mathbf{Z}_i$, which will be the true ACE if faithfulness holds (we assume for now the joint distribution of the observables is known, so no statistical uncertainty plays a role yet). Let $(\mathcal{L}_i(\aleph), \mathcal{U}_i(\aleph))$ be the corresponding lower bound and upper bound implied by $(W_i, \mathbf{Z}_i)$ and $\aleph$. Finally, let $i^\star \in \{1, 2, \ldots, k\}$ be the index of a witness/admissible set that will be our reference pair[20] to output the final bounds on the ACE, once we choose $\aleph$.

The idea is simple: minimize $(\epsilon_w, \epsilon_{xy}, \beta)$ subject to $A_j \in [\mathcal{L}_{i^\star}(\aleph), \mathcal{U}_{i^\star}(\aleph)]$ for $1 \leq j \leq k$. This is a multi-objective minimization problem, of which we can return the Pareto frontier. Because this is a small dimensional problem in which high precision is not needed, a simple grid search will suffice, as performed in Section 9. Given that the Pareto frontier is likely to contain multiple points, we can report all intervals implied by each possible choice of $\aleph$.

20. The score in Equation (11) is used to pick $i^\star$.

Alternatively, we can provide a summary of the resulting bounds, such as the union of the intervals.

The rationale for this is as follows: if faithfulness is true, then all A_i will collapse to the same value, which implies that all bounds will collapse to a single point. Differences among A_i are the result of faithfulness violations, and we explain the contradictions via our $\aleph$ parameters. Contradictions in constraints entailed by faithfulness have been exploited before to achieve robust causal inference, as in the Conservative PC algorithm of Ramsey et al. (2006). To the best of our knowledge, we provide here the first algorithm for accommodating faithfulness contradictions in a space of constraints other than conditional independence constraints among observables.

For the real case where the observable joint distribution needs to be estimated from data, one simple alternative is just to use empirical estimates of the unconstrained joint. In one sense, this provides a conservative choice of $\aleph$, as one could modify the constraints in the minimization of $(\epsilon_w, \epsilon_{xy}, \beta)$ to require instead a less stringent criterion: that (say) the 95% credible interval for each A_j overlaps with credible intervals for $[\mathcal{L}_{i^\star}(\aleph), \mathcal{U}_{i^\star}(\aleph)]$. Credible intervals can be obtained as a function of the posterior distribution of the parameters of the joint of $\{X, Y, W_1, \ldots, W_k\} \cup \bigcup_{i=1}^{k} \mathbf{Z}_i$, where a prior over the multivariate binary distributions is subject to the WPP constraints for $(W_i, \mathbf{Z}_i)$ and the independence constraints for all other pairs, at each candidate value of $\aleph$. This is very costly, and better sampling procedures than the off-the-shelf rejection sampler will be necessary. We adopt the simple conservative approach with plug-in estimators instead.

8.2.2 Method 2: Bayesian Learning of Relaxation Parameters

A criticism of the tightest ACE coverage method is that it does not take into account the size of k: for $k = 1$, it will return $\epsilon_w = \epsilon_{xy} = 0$, for instance. Judgment is necessary on whether k is large enough in order to trust the results of this analysis. Alternatively, one may cast the problem of learning $\aleph$ as yet another Bayesian learning problem, with $\{(W_1, \mathbf{Z}_1), \ldots, (W_k, \mathbf{Z}_k)\}$ providing evidence for $\aleph$ according to some reasonable definition of "likelihood function." In what follows, again we assume the joint distribution of the observables is given, so that the back-door ACE functionals $A_1, A_2, \ldots, A_k$ given by Rule 1 are observable. As in the previous section, in our implementation we just plug-in the empirical distribution of the observables, but more sophisticated approaches accounting for the uncertainty in this estimate can in principle be constructed.

The principle is: if we allow for many ways in which faithfulness violations might be detected by contradictory results, but contradictions are not found, then this should be evidence that faithfulness violations do not exist. If contradictions are "small" (i.e., ACEs implied by different back-door adjustments are close), faithfulness violations should be small. Our uncertainty should decrease as more opportunities for contradictions are allowed. In particular, we want the posterior to converge to the single values $\epsilon_w = 0, \epsilon_{xy} = 0$ and $\beta = 1$ as the number of witness/admissible set pairs increase and they agree on the same value.

We start by defining

$$d_w \equiv \max_{i,x,\mathbf{z}} |P(Y = 1 \mid X = x, W_i = 1, \mathbf{Z}_i = \mathbf{z}, U) - P(Y = 1 \mid X = x, W_i = 0, \mathbf{Z}_i = \mathbf{z}, U)|,$$

for $i = 1, 2, \ldots, k$. An analogous definition is given for d_{xy} and d_β. Next, we define the "likelihood" function for d_w under "data set" $\{A_1, A_2, \ldots A_k\}$ as follows;

$$P(A_1, A_2, \ldots, A_k \mid d_w, d_{xy}, d_\beta) = \prod_{i=1}^{k} p_{U[\mathcal{L}(d_w,d_{xy},d_\beta),\mathcal{U}(d_w,d_{xy},d_\beta)]}(A_i), \tag{34}$$

where $p_{U[a,b]}(\cdot)$ is the uniform distribution in $[a, b]$. Functions $\mathcal{L}(d_w, d_{xy}, d_\beta), \mathcal{U}(d_w, d_{xy}, d_\beta)$ are the respective lower bound and upper bound implied by the WPP constraints parameterized by $\{d_w, d_{xy}, d_\beta\}$, and the (given) joint distribution of the observables. A uniform (discrete) prior for $\{d_w, d_{xy}, d_\beta\}$ is given over a pre-defined grid of values for these parameters[21]. We then choose a set of $\{\epsilon_w, \epsilon_{xy}, \beta\}$ as the high posterior density region defined by sorting all $\{d_w, d_{xy}, d_\beta\}$ in decreasing value of posterior mass, picking the minimum set that adds up to at least 95% of posterior mass. We summarize the implied set of bounds as necessary, see Section 9.

There is no reason why a uniform prior and the uniform likelihood (34) should be the only choices. Our motivation is that the chosen likelihood function penalizes parameters that imply wide intervals, while remaining agnostic about the position of each ACE within bounds. More importantly, the penalization increases as k increases, making the posterior more peaked. It however forces all intervals of equal length to be distinguished based on the prior only. Priors matter in applied work, but in our experiments we choose the uniform prior for its simplicity. We leave the discussion of other choices of likelihood and priors for future work.

A criticism of Equation (34) is that the pairs in set $\{(W_1, \mathbf{Z}_1), \ldots, (W_k, \mathbf{Z}_k)\}$ might have much overlap (in the sense that a same witness may appear in many pairs, and the intersection among $\{\mathbf{Z}_i\}$ may be large). As such, the multiplication in Equation (34) provides overconfident posteriors, as pairs are considered to be independent pieces of information for the relaxation parameters. More free parameters accounting for the dependence of $\{A_i\}$ given $\{d_w, d_{xy}, d_\beta\}$ should be added. However, while we remove a class of irrelevant pairs (any $(W_j, \mathbf{Z}_j)$ such that there is some $i \neq j$ where $\mathbf{Z}_i \subset \mathbf{Z}_j$ and $W_i = W_j$), in this work we ignore more complex adjustments for simplicity.

9. Experiments

In this section, we start with a comparison of the back-substitution algorithm of Section 5.2 against the fully numerical procedure, which generates constraints using standard algorithms for changing between polytope representations. We then perform studies with synthetic data, comparing different back-door estimation algorithms against WPP. Finally, we perform studies with real data sets.

21. See Section 9. Using a pre-defined discretization simplifies computation, as no MCMC is required and we do not need high precision in estimating relaxation parameters. A continuous space would also imply challenges to the MCMC approach, as the posterior can be flat in some regions where different parameter settings imply intervals of same length $\mathcal{U}(d_w, d_{xy}, d_\beta) - \mathcal{L}(d_w, d_{xy}, d_\beta)$.

9.1 Empirical Investigation of the Back-substitution Algorithm

We compare the back-substitution algorithm introduced in Section 5.2 with the fully numerical algorithm. Comparison is done in two ways: (i) computational cost, as measured by the wallclock time taken to generate 100 samples by rejection sampling; (ii) width of the generated intervals. As discussed in Section 5.2, bounds obtained by the back-substitution algorithm are at least as wide as in the numerical algorithm, barring rounding problems[22].

We ran two batches of 1000 trials each, varying the level of the relaxation parameters. In the first batch, we set $\epsilon_x = \epsilon_y = \epsilon_w = 0.2$, and $\underline{\beta} = 0.9$, $\overline{\beta} = 1.1$. In the second batch, we change parameters so that $\underline{\beta} = \overline{\beta} = 1$. Experiments were run on a Intel Xeon E5-1650 at 3.20Ghz. Models were simulated according the the structure $W \to X \to Y$, sampling each conditional distribution of a vertex being equal to 1 given its parent from the uniform $(0, 1)$ distribution. The numerical procedure of converting extreme points to linear inequalities was done using the package RCDD, a R wrapper for the *cddlib* by Komei Fukuda. Inference is done by rejection sampling, requiring 100 samples per trial. We fix the number of iterations of the back-substitution method to 4, which is more than enough to achieve convergence. All code was written in R.

For the first batch, the average time difference between the fully numerical method and the back-substitution algorithm was 1 second, standard deviation (s.d.) 0.34. The ratio between times had a mean of 203 (s.d. 82). Even with a more specialized implementation of the polytope dualization step[23], two orders of magnitude of difference seem hard to remove by better coding. Concerning interval widths, the mean difference was 0.15 (s.d. 0.06), meaning that the back-substitution on average has intervals where the upper bound minus the lower bound difference is 0.15 units more than the numerical method, under this choice of relaxation parameters and averaged over problems generated according to our simulation scheme. There is a correlation between the width difference and the interval width given by the numerical method the gap, implying that differences tend to be larger when bounds are looser: the gap between methods was as small as 0.04 for a fully numerical interval of width 0.19, and as large as 0.23 for a fully numerical interval of width 0.49. For the case where $\overline{\beta} = \underline{\beta} = 1$, the average time difference was 0.92 (s.d. of 0.24), ratio of 152 (s.d. 54.3), interval width difference of 0.09 (s.d. 0.03); The gap was as small as 0.005 for a fully numerical interval of width 0.09, and as large as 0.17 for a fully numerical interval of with 0.23.

9.2 Synthetic Studies

We describe a set of synthetic studies for binary data where, for procedures that estimate ACE intervals, we assess the trade-off between its correctness (that is, how far from the true ACEs the intervals are, for a suitable definition of distance) and its informativeness (how wide the intervals are).

22. We did not use rational arithmetic in the polytope generator in order to speed it up; consequently, about 1% of the time we observed numerical problems. Those were excluded from the statistics reported in this section.
23. One advantage of the analytical bounds, as used by the back substitution method, is that it is easy to express them as matrix operations over all Monte Carlo samples, while the polytope construction requires iterations over the samples.

In the synthetic study setup, we compare our method against NE1 and NE2, two naïve point estimators defined by back-door adjustment on the whole of set of available covariates $\mathbf{W}$ and on the empty set, respectively. The former is widely used in practice, even when there is no causal basis for doing so (Pearl, 2009). The point estimator of Entner et al. (2013), based solely on the faithfulness assumption, is also assessed.

We generate problems where conditioning on the whole set $\mathbf{W}$ is guaranteed to give incorrect estimates. In detail: we generate graphs where $\mathbf{W} \equiv \{Z_1, Z_2, \ldots, Z_8\}$. Four independent latent variables $L_1, \ldots, L_4$ are added as parents of each $\{Z_5, \ldots, Z_8\}$; L_1 is also a parent of X, and L_2 a parent of Y. L_3 and L_4 are each randomly assigned to be a parent of either X or Y, but not both. $\{Z_5, \ldots, Z_8\}$ have no other parents. The graph over $Z_1, \ldots, Z_4$ is chosen by adding edges uniformly at random according to a fixed topological order. As a consequence, using the full set $\mathbf{W}$ for back-door adjustment is always incorrect, as at least four paths $X \leftarrow L_1 \rightarrow Z_i \leftarrow L_2 \rightarrow Y$ are active for $i = 5, 6, 7, 8$. The conditional probabilities of a vertex given its parents are generated by a logistic regression model with pairwise interactions, where parameters are sampled according to a zero mean Gaussian with standard deviation 20 / number of parents. Parameter values are also further bounded, so that if the generated value if greater than 0.975 or less than 0.025, it is resampled uniformly in $[0.950, 0.975]$ or $[0.025, 0.050]$, respectively.

We analyze two variations: in the first, it is guaranteed that at least one valid pair witness-admissible set exists; in the second, all latent variables in the graph are set also as common parents also of X and Y, so no valid witness exists. We divide each variation into two subcases: in the first, "hard" subcase, parameters are chosen (by rejection sampling, proposing from the model described in the previous paragraph) so that NE1 has a bias of at least 0.1 in the population; in the second, no such a selection is enforced, and as such our exchangeable parameter sampling scheme makes the problem relatively easy. We summarize each WPP interval by the posterior expected value of the lower and upper bounds. In general WPP returns more than one bound: we select the upper/lower bound corresponding to the $(W, \mathbf{Z})$ pair which maximizes the score described at the end of Section 4.2. A BDeu prior with an equivalent sample size of 10 was used.

Our main evaluation metric for an estimate is the Euclidean distance (henceforth, "error") between the true ACE and the closed point in the given estimate, whether the estimate is a point or an interval. For methods that provide point estimates (NE1, NE2, and faithfulness), this means just the absolute value of the difference between the true ACE and the estimated ACE. For WPP, the error of the interval $[\mathcal{L}, \mathcal{U}]$ is zero if the true ACE lies in this interval. We report *error average* and *error tail mass at 0.1*, the latter meaning the proportion of cases where the error exceeds 0.1. Moreover, the faithfulness estimator is defined by averaging over all estimated ACEs as given by the accepted admissible sets in each problem.

As discussed in Section 8.1, WPP can be understood as providing a trade-off between information loss and accuracy. For instance, while the trivial interval $[-1, 1]$ will always have zero error, it is not an interesting solution. We assess the trade-off by running simulations at different levels of k_ϵ, where $\epsilon_w = \epsilon_y = \epsilon_x = k_\epsilon$. We also have two configurations for $\{\underline{\beta}, \overline{\beta}\}$: we set them at either $\underline{\beta} = \overline{\beta} = 1$ or $\underline{\beta} = 0.9, \overline{\beta} = 1.1$.

For the cases where no witness exists, Entner's Rule 1 should theoretically report no solution. Entner et al. (2013) used stringent thresholds for deciding when the two conditions

of Rule 1 held, we refer to that paper for an evaluation on how well Rule 1 can be correctly activated under the more conservative setup. Instead we take a more relaxed approach, using a uniform prior on the hypothesis of independence. As such, due to the nature of our parameter randomization, more often than not it will propose at least one witness. That is, for the problems where no exact solution exists, we assess how sensitive the methods are given conclusions taken from "approximate independencies" instead of exact ones.

The analytical bounds are combined with the numerical procedure as follows. We use the analytical bounds to test each proposed model using the rejection sampling criterion. Under this scheme, we calculate the posterior expected value of the contingency table and, using this single point, calculate the bounds using the fully numerical method. This is not guaranteed to work: the point estimator using the analytical bounds might lie outside the polytope given by the full set of constraints. If this situation is detected, we revert to calculating the bounds using the analytical method. The gains in interval length reduction using the full numerical method are relatively modest (e.g., at $k_\epsilon = 0.20$, the average interval width reduced from 0.30 to 0.24) but depending on the application they might make a sizeable difference.

We simulate 100 data sets for each one of the four cases (hard case/easy case, with theoretical solution/without theoretical solution), 5000 points per data set, 1000 Monte Carlo samples per decision. Results for the point estimators (NE1, NE2, faithfulness) are obtained using the population contingency tables. Results are summarized in Table 1. The first observation is that at very low levels of k_ϵ we increase the ability to reject all witness candidates: this is due mostly not because Rule 1 never fires, but because the falsification rule of WPP (which does not enforce independence constraints) rejects the proposed witnesses found by Rule 1. The trade-off set by WPP is quite stable, where larger intervals are indeed associated with smaller error. The point estimates vary in quality, being particularly bad in the situation where no witness should theoretically exist. The set-up where $\underline{\beta} = 0.9, \overline{\beta} = 1.1$ is especially uninformative compared to $\underline{\beta} = \overline{\beta} = 1$. At $k_\epsilon = 0.2$, we obtain interval widths around 0.50. As Manski (2007) emphasizes, this is the price for making fewer assumptions. Even there, they typically cover only about 25% of the interval $[-1, 1]$ of *a priori* possibilities for the ACE.

9.2.1 Selection of Relaxation Parameters

We performed an automated choice of relaxation parameters applying the methods in Section 8.2 to the same synthetic data sets. For each data set and each parameter choice method, we obtain a set B of intervals defined by a lower/upper bound. We summarize B in two ways: the *tightest* bound, meaning we choose the narrowest interval in B; the *loosest* bound, defined as the interval where the lower (upper) bound is the smallest lower (largest upper) bound in B. We then report results for each of the four synthetic case scenarios and each of the two methods: the Tightest ACE Coverage (TAC) method from Section 8.2.1 and the high posterior density (HPD) method of Section 8.2.2. Each parameter ϵ_w and $\epsilon_{xy} = \epsilon_x = \epsilon_y$ was allowed to assume values in the discretized grid $\{0.01, 0.05, 0.10, \ldots, 0.50\}$. Parameter $\beta = \overline{\beta} = 1/\underline{\beta}$ was allowed to take values in $\{1, 1.05, \ldots, 1.20\}$. Results are summarized in Table 2.

Hard, Solvable: NE1 = $(0.12, 1.00)$, NE2 = $(0.02, 0.03)$									
k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.74	0.03	0.05	0.02	0.05	0.05	0.00	0.00	0.34
0.10	0.94	0.04	0.05	0.01	0.01	0.11	0.00	0.00	0.41
0.15	0.99	0.04	0.05	0.01	0.02	0.16	0.00	0.00	0.46
0.20	1.00	0.05	0.05	0.01	0.01	0.24	0.00	0.00	0.53
0.25	1.00	0.05	0.07	0.00	0.00	0.32	0.00	0.00	0.60
0.30	1.00	0.05	0.10	0.00	0.00	0.41	0.00	0.00	0.69
Easy, Solvable: NE1 = $(0.01, 0.01)$, NE2 = $(0.07, 0.24)$									
k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.81	0.03	0.02	0.02	0.04	0.04	0.00	0.01	0.34
0.10	0.99	0.02	0.02	0.01	0.02	0.09	0.00	0.00	0.40
0.15	1.00	0.02	0.01	0.00	0.00	0.17	0.00	0.00	0.46
0.20	1.00	0.02	0.01	0.00	0.00	0.24	0.00	0.00	0.54
0.25	1.00	0.02	0.01	0.00	0.00	0.32	0.00	0.00	0.61
0.30	1.00	0.02	0.01	0.00	0.00	0.41	0.00	0.00	0.67
Hard, Not Solvable: NE1 = $(0.16, 1.00)$, NE2 = $(0.20, 0.88)$									
k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.67	0.20	0.90	0.17	0.76	0.06	0.04	0.14	0.32
0.10	0.91	0.19	0.91	0.13	0.63	0.10	0.02	0.07	0.39
0.15	0.97	0.19	0.92	0.10	0.41	0.18	0.01	0.03	0.45
0.20	0.99	0.19	0.95	0.07	0.25	0.24	0.01	0.01	0.51
0.25	1.00	0.19	0.96	0.03	0.13	0.31	0.00	0.00	0.58
0.30	1.00	0.19	0.96	0.02	0.06	0.39	0.00	0.00	0.66
Easy, Not Solvable: NE1 = $(0.09, 0.32)$, NE2 = $(0.14, 0.56)$									
k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.68	0.13	0.51	0.10	0.37	0.05	0.02	0.07	0.33
0.10	0.97	0.12	0.53	0.08	0.28	0.10	0.01	0.05	0.39
0.15	1.00	0.12	0.52	0.05	0.17	0.16	0.01	0.03	0.46
0.20	1.00	0.12	0.53	0.03	0.08	0.23	0.01	0.03	0.52
0.25	1.00	0.12	0.48	0.02	0.05	0.31	0.00	0.02	0.59
0.30	1.00	0.12	0.48	0.01	0.04	0.39	0.00	0.01	0.65

Table 1: Summary of the outcome of the synthetic studies. Columns labeled WPP1 refer to results obtained for $\underline{\beta} = \overline{\beta} = 1$, while WPP2 refers to the case $\underline{\beta} = 0.9, \overline{\beta} = 1.1$. The first column is the level in which we set the remaining parameters, $\epsilon_x = \epsilon_y = \epsilon_w = k_\epsilon$. The second column is the frequency by which a WPP solution has been found among 100 runs. For each particular method (NE1, NE2, Faithfulness and WPP) we report the pair (error average, error tail mass at 0.1), as explained in the main text. The Faithfulness estimator is the back-door adjustment obtained by using as the admissible set the same set found by WPP1. Averages are taken only over the cases where a witness-admissible set pair has been found. The columns following each WPP results are the median width of the respective WPP interval across the 100 runs.

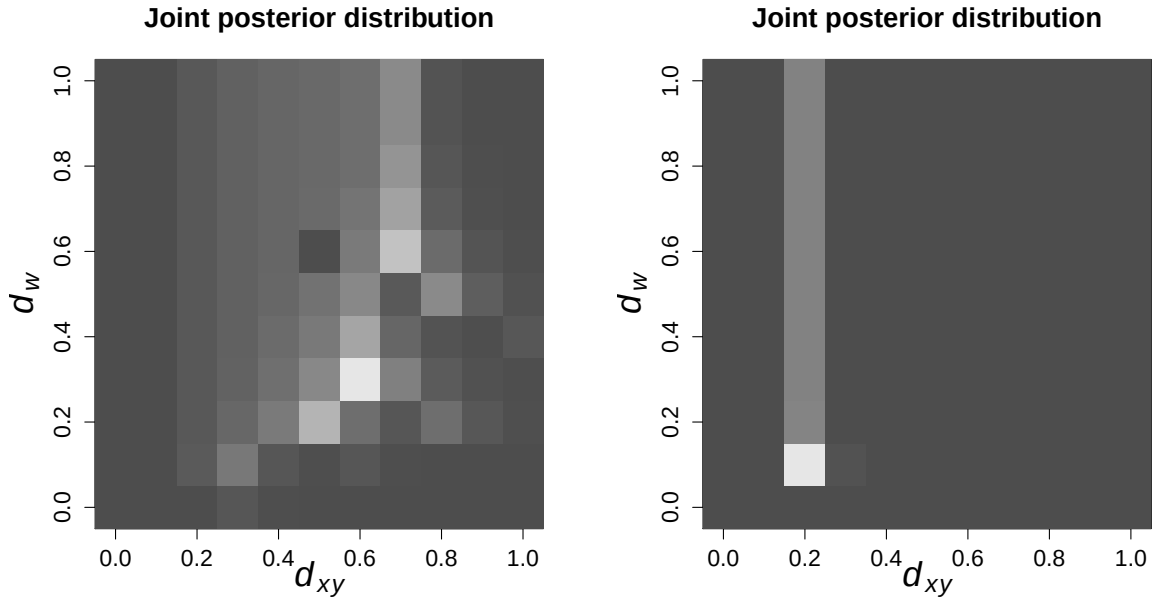
	Tightest				Loosest			
Case	TAC		HPD		TAC		HPD	
	error	width	error	width	error	width	error	width
Hard, Solvable	0.004	0.18	0.004	0.18	0.002	0.24	0.00009	0.40
Easy, Solvable	0.002	0.13	0.002	0.13	0.001	0.20	0.002	0.26
Hard, Not Solvable	0.12	0.14	0.12	0.14	0.10	0.20	0.07	0.33
Easy, Not Solvable	0.07	0.14	0.07	0.14	0.05	0.20	0.04	0.35

Table 2: Applying the criteria for choosing relaxation parameters from Section 8.2 to the four synthetic case scenarios. "Error" is the average error, as formalized for Table 1. "Width" is the average width over all 100 subcases of the respective study. "Tightest" and "Loosest" are the two criteria for summarizing a set of intervals, as explained in the main text.

Comparing it against Table 1, results seem to be slightly worse than WPP1 at the same interval width, but without making prior assumptions on β. Compared to WPP2, overall widths are much smaller. The HPD method agrees with TAC on the tightest interval, as our choice of prior will always imply a posterior mode on the TAC solution. The loosest interval for HPD will always be larger or equal to the loosest in TAC, as the 95% posterior mass that generates intervals will include the Pareto frontier and possibly many other candidates. In our simulations, the reduction in error for HPD with the loosest bound came with a non-trivial increase on the length of the corresponding intervals. While we do not explicitly advocate one method over another, the HPD method can be used to classify problems as harder than others by assessing how much of the posterior mass of hyperparameters is not on the Pareto frontier. In Figure 5, we visualize the marginal posterior distribution of $\{d_{xy}, d_w\}$ for two synthetic problems in the easy/solvable case, where in one problem the tightest interval failed to cover the true ACE, while in the other the ACE was correctly accounted for.

9.3 Influenza Study

Our empirical study concerns the effect of influenza vaccination on a patient being later on hospitalized with chest problems. $X = 1$ means the patient got a flu shot, $Y = 1$ indicates the patient was hospitalized. A negative ACE therefore suggests a desirable vaccine. The study was originally discussed by McDonald et al. (1992). Shots were not randomized, but doctors were randomly assigned to receive a reminder letter to encourage their patients to be inoculated, an event recorded as binary variable *GRP*. This suggests the standard IV model in Figure 1(d), with $W = GRP$ and U unobservable. That is, W and U are independent because W is randomized, and there are reasonable justifications to believe the lack of a direct effect of letter randomization on patient hospitalization. Richardson et al. (2011) and Hirano et al. (2000) provide further discussion.



(a) (b)

Figure 5: Marginal posterior distribution for $\{d_{xy}, d_w\}$ in two problems instances. Darker values represent smaller probabilities. In instance (a), the length of the tightest interval was 0.32 and did not contain the true ACE (but the error was still < 0.01). In instance (b), the length of the tightest interval was 0.11 and did contain the true ACE. Parameter d_w does not seem to be as influential conditional on d_{xy}, and the uniform prior allows for little variability in the d_w posterior away from the mode.

From this randomization, it is possible to directly estimate the ACE[24] of W on Y: -0.01. This is called *intention-to-treat* (ITT) analysis (Rothman et al., 2008), as it is based on the treatment assigned by randomization and not on the variable of interest (X), which is not randomized. While the ITT can be used for policy making, the ACE of X on Y would be a more relevant result, as it reveals features of the vaccine that are not dependent on the encouragement design. X and Y can be confounded, as X is not controlled. For instance, the patient choice of going to be vaccinated might be caused by her general health status, which will be a factor for hospitalization in the future.

The data contains records of $2,681$ patients, with some demographic indicators (age, sex and race) and some historical medical data (for instance, whether the patient is diabetic). A total of 9 covariates is available. Using the bounds of Balke and Pearl (1997) and observed

24. Notice that while the ACE might be small, this does not mean that in another scale, such as odd-ratios, the results do not reveal an important effect. This depends on the domain.

frequencies produces an interval of $[-0.23, 0.64]$ for the ACE. WPP could *not* validate *GRP* as a witness for any admissible set.

Instead, when forbidding *GRP* to be included in an admissible set (since the theory says *GRP* cannot be a common direct cause of vaccination and hospitalization), WPP selected as the highest-scoring pair the witness *DM* (patient had history of diabetes prior to vaccination) with admissible set composed of *AGE* (dichotomized as "60 or less years old," and "above 60") and *SEX*. Choosing, as an illustration, $\epsilon_w = \epsilon_y = \epsilon_x = 0.2$ and $\underline{\beta} = 0.9$, $\overline{\beta} = 1.1$, we obtain the posterior expected interval $[-0.10, 0.17]$. This does *not* mean the vaccine is more likely to be bad (positive ACE) than good: the posterior distribution is over bounds, not over points, being completely agnostic about the distribution within the bounds. Notice that even though we allow for full dependence between all of our variables, the bounds are stricter than in the standard IV model due to the weakening of hidden confounder effects postulated by observing conditional independencies. It is also interesting that two demographic variables ended up being chosen by Rule 1, instead of other indicators of past diseases.

When allowing *GRP* to be included in an admissible set, the pair (*DM*, {*AGE*, *SEX*}) is now ranked second among all pairs that satisfy Rule 1, with the first place being given by *RENAL* as the witness (history of renal complications), with the admissible set being *GRP, COPD* (history of pulmonary disease), and *SEX*. In this case, the expected posterior interval was approximately the same, $[-0.08, 0.16]$. It is worthwhile to mention that, even though this pair scored highest by our criterion that measures the posterior probability distribution of each premise of Rule 1, it is clear that the fit of this model is not as good as the one with *DM* as the witness, as measured by the much larger proportion of rejected samples when generating the posterior distribution. This suggests future work on how to rank such models.

In Figure 6 we show a scatter plot of the posterior distribution over lower and upper bounds on the influenza vaccination, where *DM* is the witness. In Figure 7(a) and (b) we show kernel density estimators based on the Monte Carlo samples for the cases where *DM* and *RENAL* are the witnesses, respectively. While the witnesses were tested using the analytical bounds, the final set of samples shown here were generated with the fully numerical optimization procedure, which is quite expensive.

We also analyze how to select $\aleph = \{\epsilon_w, \epsilon_{xy} = \epsilon_x = \epsilon_y, \beta = \overline{\beta} = 1/\underline{\beta}\}$ using the Tightest ACE Coverage (TAC) method of Section 8.2.1. The motivation is that this is a domain with overall weak dependencies among variables. From one point of view, this is bad as instruments will be weak and generate wide intervals (as suggested by Proposition 1). From another perspective, this suggests that the effect of hidden confounders may also be weak.

A total of 48 witness/admissible sets were proposed by WPP via Rule 1. The TAC Pareto frontier, using the same parameter space as in Section 9.2.1, included only two possibilities, $\epsilon_{xy} = 0.05, \epsilon_w = 0.01, \beta = 1$ and $\epsilon_{xy} = 0.01, \epsilon_w = 0.01, \beta = 1.05$. Using the empirical distribution as an estimator of the joint of the observables, the respective ACE intervals were $[-0.01, 0.01]$ and $[-0.02, 0.02]$. Although the sign of the ACE is not determined from the data, the WPP procedure suggests that the magnitude of the ACE is no greater than 0.02, which by itself is of interest.

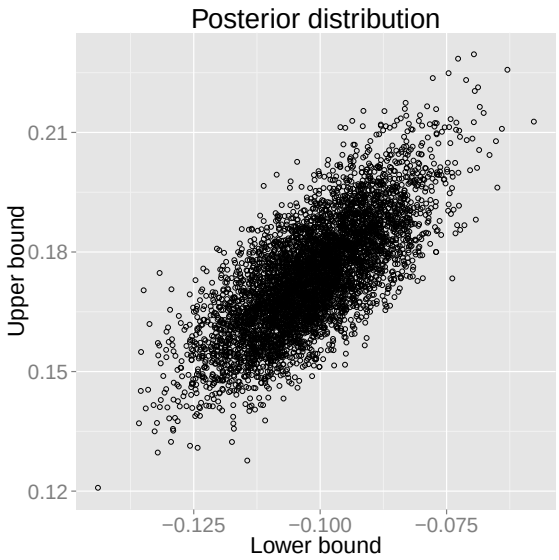


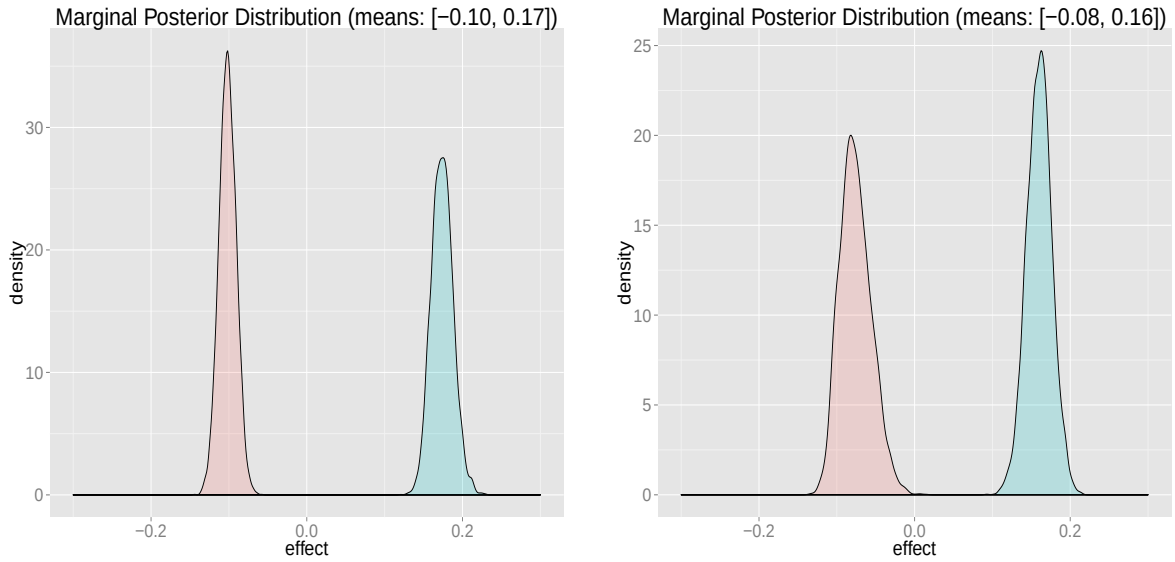
Figure 6: Scatterplot of the joint posterior distribution of lower bounds and upper bounds, Pearson correlation coefficient of 0.71.

9.4 Linear Models

In this section, we assess the usage of the method in the linear case. Independently, we also introduce a complementary way of summarizing the outcome of a WPP analysis.

We first check the performance of the method with a small synthetic study. We generate models following the same pattern of the "hard/solvable" case of Section 9.2, the ACE being the coefficient of $X^\star$ in the equation for $Y^\star$. Each conditional model for a variable $V_i^\star$ given its parents is generated by sampling its coefficients from independent standard Gaussians, sampling the variance of the error term from a uniform $[0, 0.5]$, then rescaling the coefficients such that the marginal variance of $V_i^\star$ is 1. Observable data is then generated by transforming each $V_i^\star$ to follow a gamma distribution with mean and variance equal to 2. We generated 100 data sets with a sample size of 1000 each. We perform experiments[25] setting all hyperparameters $\epsilon_c = \epsilon_{wx} = \epsilon_{wy} = \epsilon_{xy} = 0.2$ and $\epsilon_c = \epsilon_{wx} = \epsilon_{wy} = \epsilon_{xy} = 0.1$. Estimates of the Gaussian copula correlation matrices are obtained using function huge.npn from the R package huge to transform the data, of which we compute the empirical correlation matrix. We obtained average errors of 0.04 for the method with parameters set at 0.2, and 0.07 for parameters set at 0.1. The average length of the proposed intervals were 0.5 and 0.26, respectively. For comparison, the population error for the two naïve estimators was 0.23 and 0.18.

25. The test for conditional independencies is done with the corrected BGe score (Kuipers et al., 2014) as discussed in Section 6.2. The hyperparameters are a prior of 0.5 for the independence constraint hypothesis, and a inverse Wishart prior with $\upsilon \equiv p + 2$ degrees of freedom and a scale matrix given by the $p \times p$ identity matrix multiplied by υ, where p is the number of variables in the test.



(a) (b)

Figure 7: In (a), the marginal densities for the lower bound (red) and upper bound (blue) on the ACE, smoothed kernel density estimates based on 5000 Monte Carlo samples. Bounds were derived using DM as the witness. In (b), a similar plot using $RENAL$ as the witness.

We performed an empirical study with the 1976 Panel Study of Income Dynamics. The study uses data from 1975, assessing incoming of couples in 1976. Our outcome variable Y is the wife's reported wage at the time of the 1976 interview, and the treatment X is the number of years of the wife's previous labor market experience. The data was discussed by Mroz (1987) and can be obtained from the R package AER (Kleiber and Zeileis, 2008). Covariate set $\mathbf{W}$ includes a combination of discrete and continuous variables, such as husband's wages, number of children, and whether the wife went to college. We infer a Gaussian copula correlation matrix using the extended rank likelihood method of Hoff (2007) with the R package SBGCOP, which can deal with discrete and continuous variables but requires expensive MCMC sampling. Notice that conditional independencies among the discrete observable elements of $\mathbf{V} \equiv \{X, Y\} \cup \mathbf{W}$ do not follow from conditional independencies among the unobservable Gaussian variables $\mathbf{V}^\star$. We nevertheless test Rule 1 among $\mathbf{V}^\star$ using the estimated copula correlation matrix and the relatively high prior of 0.5 for the hypothesis of independence, for any given independence assessment. Sample size is 753, with 17 covariates[26]. We then make the (strong) assumption that work experience in 1975 is not a cause of any other variable in the covariate set.

26. We removed two covariates from the original data set: the indicator of economical participation, which is a deterministic function of other covariates; and the estimated wage of the wife in 1975, which for that year was not available directly via self-report. In order to speed up the search algorithm, for each witness candidate W, the space of variables to test for an admissible set is composed of the 10 covariates mostly

Setting all relaxation parameters $\epsilon_c = \epsilon_{wx} = \epsilon_{wy} = \epsilon_{xy}$ to 0.1, we obtain in Figure 8(a) all corresponding intervals, with black dots representing the corresponding estimated ACE for the chosen admissible set. An explanation of all variables can be found in the documentation of package AER (Kleiber and Zeileis, 2008). Recall that the units here are given in the latent Gaussian space, where each $V_i^\star$ is a non-linear transformation of the corresponding V_i, as explained in Section 6.2. This analysis reveals two clear clusters of behavior, which internally show little variability but are very different from one another, even accounting for a violation of 0.1. This illustrates possible ways of communicating the output of a WPP analysis so that issues with assumptions and data can be raised.

In this case, the two clusters of intervals differ in one variable in the admissible set: variable *HOURS* is present in cases where the intervals are closer to zero. This variable measures the number of work hours of the wife in 1975, and is partially embedded in the definition of the experience level measured at 1975. By removing all admissible sets that include the *HOURS* variable, we obtain the summary given as Figure 8(b). This type of visualization step can be used to flag major contradictions that cannot be easily explained by allowing mild violations of faithfulness, but which might suggest problematic measurements to be reconsidered in the analysis.

10. Conclusion

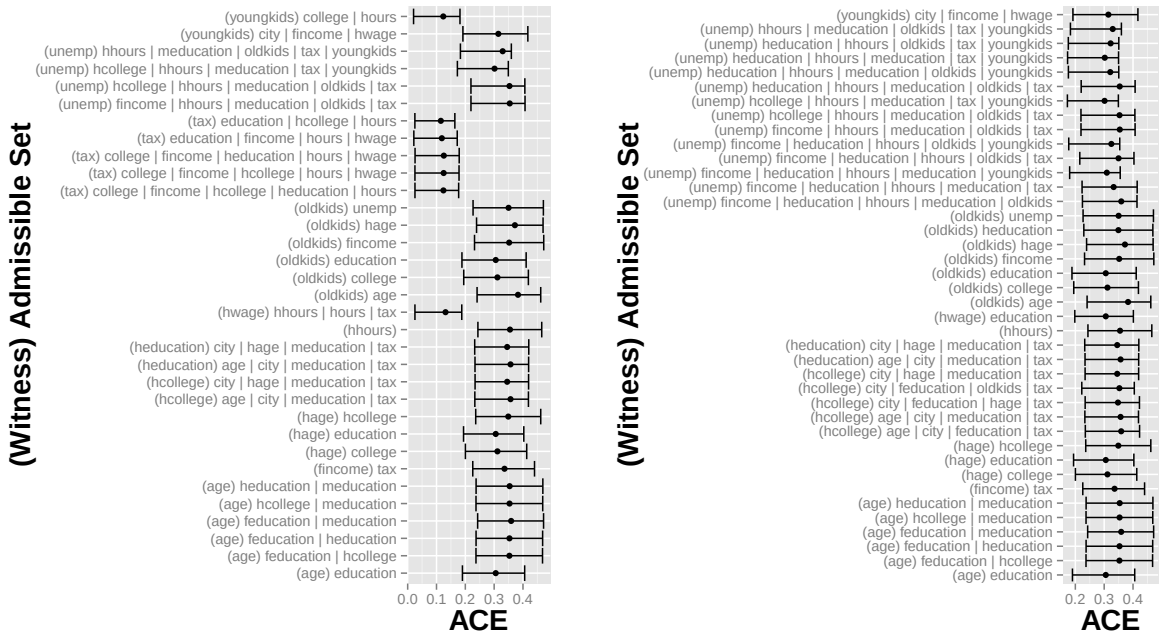
Our model provides a novel compromise between point estimators given by the faithfulness assumption and bounds based on instrumental variables. We believe such an approach should become a standard item in the toolbox of methodologies for observational studies, as it provides means to draw conclusions from a complementary set of assumptions. Ongoing updates of software for WPP is provided as part of the R package CausalFX, available at the Comprehensive R Network[27] and GitHub[28]. A snapshot of the code used in this paper is available at `http://www.homepages.ucl.ac.uk/~ucgtrbd/wpp`.

In particular, unlike Bayesian approaches that put priors directly on the parameters of the unidentifiable latent variable model $P(Y, X, W, U \mid \mathbf{Z})$, the constrained Dirichlet prior on the observed distribution does not suffer from massive sensitivity to the choice of hyperparameters. When a strongly informative prior is lacking, WPP keeps inference more honest by focusing on bounds. While it is tempting to look for an alternative that will provide a point estimate of the ACE, it is also important to have a method that trades-off information for fewer assumptions. WPP provides a framework to express such assumptions.

The brute-force search used in the implementation of Rule 1 can be substituted by other combinatorial search procedures and dimensionality reduction methods. Entner et al. (2013) provide alternatives by borrowing ideas from the PC algorithm, for instance. Package CausalFX implements the idea discussed briefly in Section 9.4, where for each witness candidate W we pre-select a small set of candidates from $\mathbf{W}\backslash\{W\}$ and perform a brute-force search for admissible sets within this candidate set only. Pre-selection in CausalFX 1.0 is done by first sorting all $Z \in \mathbf{W}\backslash\{W\}$ according to the empirical mutual information

strongly associated with W, measured by the absolute value of the corresponding copula correlation matrix entry.

27. `https://cran.r-project.org/web/packages/CausalFX/index.html`
28. `https://github.com/rbas2015/CausalFX`



(a) (b)

Figure 8: The diagrams depict some ACE intervals obtained for the linear model of the impact of work experience up to 1975 of a married woman into her salary in 1976. On the y-axis, we show the witness in brackets, followed by all variables in the admissible set; the x-axis shows the point estimates of the interval for the ACE using $\epsilon_c = \epsilon_{wx} = \epsilon_{wy} = \epsilon_{xy} = 0.1$. Black dots are the corresponding point estimates of the ACE using the back-door method. All variable names are explained in the documentation of package AER (Kleiber and Zeileis, 2008). In (a), we allow all other recorded variables into the covariate set $\mathbf{W}$ from which witnesses and admissible sets are generated. In (b), we remove $HOURS$ from the pool of possible covariates.

of Z and W given X and then picking the top K candidates, in descending value of mutual information (the heuristic being that we should look first at paths $W \to X \leftarrow Z$ that are "strong"). K is chosen such that enumerating 2^K candidate admissible sets is possible within the available computer resources. Although this restricted search procedure might miss some admissible sets, it has the advantage of avoiding sensitivity to propagation of statistical mistakes that creates difficulties for the PC algorithm and similar methods.

We emphasize that the credible intervals obtained by the procedure are conditioned on the search results, discarding uncertainty coming from the choices of witnesses, admissible sets and relaxation parameters. Ideally, uncertainty concerning the outcome of the Rule 1 search should also be taken into account. An approach analogous to (Friedman and Koller, 2003) is necessary, which we leave as future work.

As further future work, we will look at a generalization of the procedure beyond relaxations of chain structures $W \to X \to Y$. Much of the machinery here developed, including Entner et al.'s Rules, can be adapted to the case where causal ordering is unknown: starting from the algorithm of Mani et al. (2006) to search for "Y-structures," it is possible to generalize Rule 1 to setups where we have an outcome variable Y that needs to be controlled, but where there is no covariate X known not to be a cause of other covariates. Mooij and Cremers (2015) investigate the robustness of the faithfulness condition in this setup. Finally, the techniques used to derive the symbolic bounds in Section 5 may prove useful in a more general context, and complement other methods to find subsets of useful constraints such as the graphical approach of Evans (2012).

Acknowledgments

We thank McDonald, Hiu and Tierney for their flu vaccine data and the anonymous reviewers for their many suggestions to improve our paper. Much of this work was done while RS was hosted by the Department of Statistics at the University of Oxford. Parts of this work were previously published in the 2014 Neural Information Processing Systems Conference (Silva and Evans, 2014). Section 6, 7 and 8 are completely new, and all remaining sections have new material added, including the appendix.

Appendix A. Proofs

In this Appendix, we prove the results mentioned in the main text.

A.1 Basic Results

We divide the proofs in four main sections. The first section provides the basic methods, including how classical results in instrumental variable bounding can be rederived. The second and third sections are proofs for the most complex types of bounds. Finally, the fourth section covers the linear continuous case.

Proof of Proposition 1 In the standard IV case, simple analytical bounds are known for $P(Y = y \mid do(X = x))$ (Balke and Pearl, 1997; Dawid, 2003):

$$\eta_0 \leq \min \begin{cases} 1 - \zeta_{00.0} \\ 1 - \zeta_{00.1} \\ \zeta_{01.0} + \zeta_{10.0} + \zeta_{10.1} + \zeta_{11.1} \\ \zeta_{10.0} + \zeta_{11.0} + \zeta_{01.1} + \zeta_{10.1} \end{cases} \qquad \eta_0 \geq \max \begin{cases} \zeta_{10.1} \\ \zeta_{10.0} \\ \zeta_{10.0} + \zeta_{11.0} - \zeta_{00.1} - \zeta_{11.1} \\ -\zeta_{00.0} - \zeta_{11.0} + \zeta_{10.1} + \zeta_{11.1} \end{cases}$$

$$\eta_1 \leq \min \begin{cases} 1 - \zeta_{01.1} \\ 1 - \zeta_{01.0} \\ \zeta_{10.0} + \zeta_{11.0} + \zeta_{00.1} + \zeta_{11.1} \\ \zeta_{00.0} + \zeta_{11.0} + \zeta_{10.1} + \zeta_{11.1} \end{cases} \qquad \eta_1 \geq \max \begin{cases} \zeta_{11.1} \\ \zeta_{11.0} \\ -\zeta_{01.0} - \zeta_{10.0} + \zeta_{10.1} + \zeta_{11.1} \\ \zeta_{10.0} + \zeta_{11.0} - \zeta_{01.1} - \zeta_{10.1} \end{cases}$$

where $\eta_x \equiv P(Y = 1 \mid do(X = x))$ and $\zeta_{yx.w} \equiv P(Y = y, X = x \mid W = w)$. Define also $\alpha_x \equiv P(Y = 1 \mid X = x)$ and $\beta_w \equiv P(X = 1 \mid W = w)$ so that

$$\zeta_{yx.w} = \alpha_x^{I(y=1)}(1-\alpha_x)^{I(y=0)}\beta_w^{I(x=1)}(1-\beta_w)^{I(x=0)}, \tag{35}$$

where $I(\cdot)$ is the indicator function returning 1 or 0 depending on whether its argument is true or false, respectively.

Assume for now that $\beta_1 \geq \beta_0$, that is, $P(X = 1 \mid W = 1) \geq P(X = 1 \mid W = 0)$. We will first show that $1-\zeta_{00.0} \leq \min\{1-\zeta_{00.1}, \zeta_{01.0}+\zeta_{10.0}+\zeta_{10.1}+\zeta_{11.1}, \zeta_{10.0}+\zeta_{11.0}+\zeta_{01.1}+\zeta_{10.1}\}$.

That $1-\zeta_{00.0} \leq 1-\zeta_{00.1}$ follows directly from the relationship (35) and the assumptions $W \perp\!\!\!\perp Y \mid X$ and $\beta_1 \geq \beta_0$: $(1-\zeta_{00.0}) - (1-\zeta_{00.1}) = -(1-\alpha_0)(1-\beta_0) + (1-\alpha_0)(1-\beta_1) = (1-\alpha_0)(\beta_0-\beta_1) \leq 0$.

Now consider $(1-\zeta_{00.0}) - (\zeta_{01.0}+\zeta_{10.0}+\zeta_{10.1}+\zeta_{11.1})$. This is equal to

$$\begin{aligned} &= (1-(1-\alpha_0)(1-\beta_0)) - ((1-\alpha_1)\beta_0 + \alpha_0(1-\beta_0) + \alpha_0(1-\beta_1) + \alpha_1\beta_1) \\ &= (\beta_0 + \alpha_0(1-\beta_0)) - (\beta_0 - \alpha_1\beta_0 + \alpha_0(1-\beta_0) + \alpha_0 - \alpha_0\beta_1 + \alpha_1\beta_1) \\ &= \alpha_1(\beta_0-\beta_1) - \alpha_0(1-\beta_1) \leq 0 \end{aligned}$$

Analogously, we can show that $1-\zeta_{00.0} \leq \zeta_{10.0}+\zeta_{11.0}-\zeta_{01.1}-\zeta_{10.1}$. Tedious but analogous manipulations lead to the overall conclusion

$$1-\zeta_{00.0} = \min\begin{cases} 1-\zeta_{00.0} \\ 1-\zeta_{00.1} \\ \zeta_{01.0}+\zeta_{10.0}+\zeta_{10.1}+\zeta_{11.1} \\ \zeta_{10.0}+\zeta_{11.0}+\zeta_{01.1}+\zeta_{10.1} \end{cases} \qquad \zeta_{10.0} = \max\begin{cases} \zeta_{10.1} \\ \zeta_{10.0} \\ \zeta_{10.0}+\zeta_{11.0}-\zeta_{00.1}-\zeta_{11.1} \\ -\zeta_{00.0}-\zeta_{11.0}+\zeta_{10.1}+\zeta_{11.1} \end{cases}$$

$$1-\zeta_{01.1} = \min\begin{cases} 1-\zeta_{01.1} \\ 1-\zeta_{01.0} \\ \zeta_{10.0}+\zeta_{11.0}+\zeta_{00.1}+\zeta_{11.1} \\ \zeta_{00.0}+\zeta_{11.0}+\zeta_{10.1}+\zeta_{11.1} \end{cases} \qquad \zeta_{11.1} = \max\begin{cases} \zeta_{11.1} \\ \zeta_{11.0} \\ -\zeta_{01.0}-\zeta_{10.0}+\zeta_{10.1}+\zeta_{11.1} \\ \zeta_{10.0}+\zeta_{11.0}-\zeta_{01.1}-\zeta_{10.1} \end{cases}$$

The upper bound on the ACE $\eta_1 - \eta_0$ is obtained by subtracting the lower bound on η_0 from the upper bound on η_1. That is, $\eta_1 - \eta_0 \leq (1-\zeta_{01.1}) - \zeta_{10.0} = \mathcal{U}_{SIV}$. Similarly, $\eta_1 - \eta_0 \geq \zeta_{11.1} - (1-\zeta_{00.0}) = \mathcal{L}_{SIV}$. It follows that $\mathcal{U}_{SIV} - \mathcal{L}_{SIV} = 1 - (P(X = 1 \mid W = 1) - P(X = 1 \mid W = 0))$.

Finally, assuming $\beta_1 \leq \beta_0$ gives by symmetry the interval width $1 - (P(X = 1 \mid W = 0) - P(X = 1 \mid W = 1))$, implying the width in the general case is given by $1 - |P(X = 1 \mid W = 1) - P(X = 1 \mid W = 0)|$. ■

Now we will prove the main theorems stated in Section 5. To facilitate reading, we repeat here the notation used in the description of the constraints with a few additions, as well as the identities mapping different parameter spaces and the corresponding assumptions exploited in the derivation.

We start with the basic notation,

$$
\begin{array}{rcl}
\zeta^{\star}_{yx.w} & \equiv & P(Y=y, X=x \mid W=w, U) \\
\zeta_{yx.w} & \equiv & \sum_U P(Y=y, X=x \mid W=w, U) P(U \mid W=w) \\
 & = & P(Y=y, X=x \mid W=w) \\
\kappa_{yx.w} & \equiv & \sum_U P(Y=y, X=x \mid W=w, U) P(U)
\end{array}
$$

$$
\begin{array}{rcl}
\eta^{\star}_{xw} & \equiv & P(Y=1 \mid X=x, W=w, U) \\
\eta_{xw} & \equiv & \sum_U P(Y=1 \mid X=x, W=w, U) P(U \mid W=w) \\
 & = & P(Y=1 \mid do(X=x), W=w) \\
\omega_{xw} & \equiv & \sum_U P(Y=1 \mid X=x, W=w, U) P(U)
\end{array}
$$

$$
\begin{array}{rcl}
\delta^{\star}_{w} & \equiv & P(X=1 \mid W=w, U) \\
\delta_{w} & \equiv & \sum_U P(X=1 \mid W=w, U) P(U \mid W) = P(X=1 \mid W=w) \\
 & = & \zeta_{11.w} + \zeta_{01.w} \\
\chi_{x.w} & \equiv & \sum_U P(X=x \mid W=w, U) P(U) \\
 & = & \kappa_{1x.w} + \kappa_{0x.w}
\end{array}
$$

The explicit relationship between parameters describing the latent variable model is:

$$
\begin{array}{rcl}
\zeta^{\star}_{00.0} & = & (1-\eta^{\star}_{00})(1-\delta^{\star}_{0}) \\
\zeta^{\star}_{01.0} & = & (1-\eta^{\star}_{10})\delta^{\star}_{0} \\
\zeta^{\star}_{10.0} & = & \eta^{\star}_{00}(1-\delta^{\star}_{0}) \\
\zeta^{\star}_{11.0} & = & \eta^{\star}_{10}\delta^{\star}_{0} \\
\zeta^{\star}_{00.1} & = & (1-\eta^{\star}_{01})(1-\delta^{\star}_{1}) \\
\zeta^{\star}_{01.1} & = & (1-\eta^{\star}_{11})\delta^{\star}_{1} \\
\zeta^{\star}_{10.1} & = & \eta^{\star}_{01}(1-\delta^{\star}_{1}) \\
\zeta^{\star}_{11.1} & = & \eta^{\star}_{11}\delta^{\star}_{1}
\end{array}
$$

All upper bound constants $U^{\cdot U}_{\cdot\cdot}$ are assumed to be positive. For $L^{\cdot U}_{\cdot\cdot} = 0$, $c \geq 0$, all ratios $c/L^{\cdot U}_{\cdot\cdot}$ are defined to be positive infinite.

In what follows, we define "the standard IV model" as the one which obeys exogeneity of W and exclusion restriction—that is, the model following the directed acyclic graph $\{W \to X \to Y, X \leftarrow U \to Y\}$. All variables are binary, and the goal is to bound the average causal effect (ACE) of X on Y given a non-descendant W and a possible (set of) confounder(s) U of X and Y.

Proof of Theorem 2 Start with the relationship between η_{xw} and its upper bound:

$$
\begin{array}{rcll}
\eta^{\star}_{xw} & \leq & U^{YU}_{xw} & \text{(Multiply both sides by } \delta^{\star}_{x'.w}\text{)} \\
\eta^{\star}_{xw}(1-(1-\delta^{\star}_{x'.w})) & \leq & U^{YU}_{xw}\delta^{\star}_{x'.w} & \text{(Marginalize over } P(U)\text{)} \\
\omega_{xw} - \kappa_{1x.w} & \leq & U^{YU}_{xw}\chi_{x'.w} & \\
\omega_{xw} & \leq & \kappa_{1x.w} + U^{YU}_{xw}(\kappa_{0x'.w} + \kappa_{1x'.w}) &
\end{array}
$$

and an analogous series of steps gives $\omega_{xw} \geq \kappa_{1x.w} + L^{YU}_{xw}(\kappa_{0x'.w} + \kappa_{1x'.w})$. Notice such bounds above will depend on how tight ϵ_y is. As an illustration of its implications, consider

the derived identity $\zeta^\star_{0x.w} = (1-\eta^\star_{xw})\delta^\star_{x.w} \Rightarrow 1-\eta^\star_{xw} = \zeta^\star_{0x.w}/\delta^\star_{x.w} \Rightarrow 1-\eta^\star_{xw} \geq \zeta^\star_{0x.w} \Rightarrow \eta^\star_{xw} \leq 1-\zeta^\star_{0x.w} = \zeta^\star_{0x.w} + \zeta^\star_{0x'.w} + \zeta^\star_{1x'.w} \Rightarrow \omega_{xw} \leq \kappa_{0x.w} + \kappa_{0x'.w} + \kappa_{1x'.w}$.

It follows from $U^{YU}_{xw} \leq 1$ that that the derived bound $\omega_{xw} \leq \kappa_{1x.w} + U^{YU}_{xw}(\kappa_{0x'.w} + \kappa_{1x'.w})$ is at least as tight as the one obtained via $\eta^\star_{xw} \leq 1-\zeta^\star_{0x.w}$. Notice also that the standard IV bound $\eta_{xw} \leq 1-\zeta_{0x.w}$ (Balke and Pearl, 1997; Dawid, 2003) is a special case for $\epsilon_y = 0$, $\underline{\beta} = \overline{\beta} = 1$.

For the next bounds, consider

$$\begin{array}{rcll} \delta^\star_{x.w} & \leq & U^{XU}_{xw} & \\ \eta^\star_{xw}\delta^\star_{x.w} & \leq & U^{XU}_{xw}\eta^\star_{xw} & \text{(Marginalize over } P(U)) \\ \kappa_{1x.w} & \leq & U^{XU}_{xw}\omega_{xw} & \\ \omega_{xw} & \geq & \kappa_{1x.w}/U^{XU}_{xw} & \end{array}$$

where the bound $\omega_{xw} \leq \kappa_{1x.w}/L^{XU}_{xw}$ can be obtained analogously. The corresponding bound for the standard IV model (with possible direct effect $W \to Y$) is $\eta_{xw} \geq \zeta_{1x.w}$, obtained again by choosing $\epsilon_x = 1$, $\underline{\beta} = \overline{\beta} = 1$. The corresponding bound $\omega_{xw} \geq \kappa_{1x.w}$ is a looser bound for $U^{XU}_{xw} < 1$. Notice that if $L^{XU}_{xw} = 0$, the upper bound is defined as infinite.

Finally, the last bounds are similar to the initial ones, but as a function of ϵ_x instead of ϵ_y:

$$\begin{array}{rcll} \delta^\star_{x.w} & \leq & U^{XU}_{xw} & \\ (1-\eta^\star_{xw})\delta^\star_{x.w} & \leq & U^{XU}_{xw}(1-\eta^\star_{xw}) & \text{(Marginalize over } P(U)) \\ \kappa_{0x.w} & \leq & U^{XU}_{xw}(1-\omega_{xw}) & \\ \omega_{xw} & \leq & 1-\kappa_{0x.w}/U^{XU}_{xw} & \end{array}$$

The lower bound $\omega_{xw} \geq 1-\kappa_{0x.w}/L^{XU}_{xw}$ is obtained analogously, and implied to be minus infinite if $L^{XU}_{xw} = 0$. ■

Proof of Theorem 3 We start with the following derivation,

$$\begin{array}{rcll} \eta^\star_{xw'} - \eta^\star_{xw} & \leq & \epsilon_w & \\ \eta^\star_{xw'}\delta^\star_{x.w'} - \eta^\star_{xw}\delta^\star_{x.w'} & \leq & \epsilon_w\delta^\star_{x.w'} & (\text{Use } -U^{XU}_{xw'} \leq -\delta^\star_{x.w'}) \\ \eta^\star_{xw'}\delta^\star_{x.w'} - \eta^\star_{xw}U^{XU}_{xw'} & \leq & \epsilon_w\delta^\star_{x.w'} & (\text{Marginalize over } P(U)) \\ \kappa_{1x.w'} - \omega_{xw}U^{XI}_{xw} & \leq & \epsilon_w\chi_{x.w'} & \\ \omega_{xw} & \geq & (\kappa_{1x.w'} - \epsilon_w\chi_{x.w'})/U^{XU}_{xw'} & \\ \omega_{xw} & \geq & (\kappa_{1x.w'} - \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/U^{XU}_{xw'} & \end{array}$$

Analogously, starting from $\eta^\star_{xw'} - \eta^\star_{xw} \geq \epsilon_w$, we obtain $\omega_{xw} \leq (\kappa_{1x.w'} + \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/L^{XU}_{xw'}$. Notice that for the special case ϵ_w and $U^{XU}_{xw'} = 1$, we obtain the corresponding lower bound $\omega_{xw} \geq \kappa_{1x.w'}$ that relates ω and κ across different values of W.

The result corresponding to the upper bound $\eta_{xw} \leq 1-\zeta_{0x.w'}$ can be obtained as follows:

$$\begin{array}{rcll} \eta^\star_{xw'} - \eta^\star_{xw} & \geq & -\epsilon_w & \\ 1 + \eta^\star_{xw'} - 1 - \eta^\star_{xw} & \geq & -\epsilon_w & \\ (1-\eta^\star_{xw}) - (1-\eta^\star_{xw'}) & \geq & -\epsilon_w & \\ (1-\eta^\star_{xw})\delta^\star_{x.w'} - (1-\eta^\star_{xw'})\delta^\star_{x.w'} & \geq & -\epsilon_w\delta^\star_{x.w'} & \\ (1-\eta^\star_{xw})U^{XU}_{xw'} - (1-\eta^\star_{xw'})\delta^\star_{x.w'} & \geq & -\epsilon_w\delta^\star_{x.w'} & (\text{Marginalize over } P(U)) \\ (1-\omega_{xw})U^{XU}_{xw'} - \kappa_{0x.w'} & \geq & -\epsilon_w\chi_{x.w'} & \\ \omega_{xw} & \leq & 1-(\kappa_{0x.w'} - \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/U^{XU}_{xw'} & \end{array}$$

with the corresponding lower bound (non-trivial for $L^{XU}_{xw'} > 0$) given by $\omega^\star_{xw} \geq 1 - (\kappa_{0x.w'} + \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/L^{XU}_{xw'}$.

The final block of relationships can be derived as follows:

$$\begin{array}{rcll}
\eta^\star_{xw} - \eta^\star_{xw'} & \leq & \epsilon_w & \\
\eta^\star_{xw}\delta^\star_{x'.w} - \eta^\star_{xw'}\delta^\star_{x'.w} & \leq & \epsilon_w\delta^\star_{x'.w} & \\
\eta^\star_{xw}(1-(1-\delta^\star_{x'.w})) - \eta^\star_{xw'}\delta^\star_{x'.w} & \leq & \epsilon_w\delta^\star_{x'.w} & (\text{Use } -U^{XU}_{x'w} \leq -\delta^\star_{x'.w}) \\
\eta^\star_{xw} - \eta^\star_{xw}(1-\delta^\star_{x'.w}) - \eta^\star_{xw'}U^{XU}_{x'w} & \leq & \epsilon_w\delta^\star_{x'.w} & (\text{Marginalize over } P(U)) \\
\omega_{xw} - \kappa_{1x.w} - \omega_{xw'}U^{XU}_{x'w} & \leq & \epsilon_w\chi_{x'.w} & \\
\omega_{xw} - \omega_{xw'}U^{XU}_{x'w} & \leq & \kappa_{1x.w} + \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w}) &
\end{array}$$

with the lower bound $\omega_{xw} - \omega_{xw'}L^{XU}_{x'w} \geq \kappa_{1x.w} - \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w})$ derived analogously. Moreover,

$$\begin{array}{rcl}
\eta^\star_{xw'} - \eta^\star_{xw} & \leq & \epsilon_w \\
(1-\eta^\star_{xw})\delta^\star_{x'.w} - (1-\eta^\star_{xw'})\delta^\star_{x'.w} & \leq & \epsilon_w\delta^\star_{x'.w} \\
(1-\eta^\star_{xw})(1-(1-\delta^\star_{x'.w})) - (1-\eta^\star_{xw'})U^{XU}_{x'w} & \leq & \epsilon_w\delta^\star_{x'.w} \\
1 - \omega_{xw} - \kappa_{0x.w} - (1-\omega_{xw'})U^{XU}_{x'w} & \leq & \epsilon_w\chi_{x'.w} \\
\omega_{xw} - \omega_{xw'}U^{XU}_{x'w} & \geq & 1 - \kappa_{0x.w} - U^{XU}_{x'w} - \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w})
\end{array}$$

and the corresponding $\omega_{xw} - \omega_{xw'}L^{XU}_{x'w} \leq 1 - \kappa_{0x.w} - L^{XU}_{x'w} + \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w})$. The last two relationships follow immediately from the definition of ϵ_w. ■

Our constraints found so far collapse to some of the constraints found in the standard IV models (Balke and Pearl, 1997; Dawid, 2003) given $\epsilon_w = 0$, $\underline{\beta} = \overline{\beta} = 1$. Namely,

$$\begin{array}{rcl}
\eta_{xw} & \leq & 1 - \zeta_{0x.w} \\
\eta_{xw} & \leq & 1 - \zeta_{0x.w'} \\
\eta_{xw} & \geq & \zeta_{1x.w} \\
\eta_{xw} & \geq & \zeta_{1x.w'}
\end{array}$$

However, none of the constraints so far found counterparts in the following:

$$\begin{array}{rcl}
\eta_{xw} & \leq & \zeta_{0x.w} + \zeta_{1x.w} + \zeta_{1x.w'} + \zeta_{1x'.w'} \\
\eta_{xw} & \leq & \zeta_{0x.w'} + \zeta_{1x.w'} + \zeta_{1x.w} + \zeta_{1x'.w} \\
\eta_{xw} & \geq & \zeta_{1x.w} + \zeta_{1x'.w} - \zeta_{0x.w'} - \zeta_{1x'.w'} \\
\eta_{xw} & \geq & \zeta_{1x.w'} + \zeta_{1x'.w'} - \zeta_{0x.w} - \zeta_{1x'.w}
\end{array}$$

These constraints have the distinctive property of being functions of both $P(Y = x, X = x \mid W = w)$ and $P(Y = x, X = x \mid W = w')$, simultaneously. So far, we have only used the basic identities and constraints, without attempting at deriving constraints that are not a direct application of such identities. In the framework of (Dawid, 2003; Ramsahai, 2012), it is clear that general linear combinations of functions of $\{\delta^\star_{x.w}\eta^\star_{1x.w}, \delta^\star_{x.w}, \eta^\star_{1x.w}\}$ can generate constraints on observable quantities $\zeta_{yx.w}$ and causal quantities of interest, η_{xw}. We need to encompass these possibilities in such a way that we get a framework for generating symbolic constraints as a function of $\{\epsilon_w, \epsilon_y, \epsilon_x, \underline{\beta}, \overline{\beta}\}$.

One of the difficulties on exploiting a black-box polytope package for that is due to the structure of the process, which exploits the constraints in Section 4 by first finding the extreme points of the feasible region of $\{\delta^\star_w\}$, $\{\eta^\star_{xw}\}$. If we use the constraints

$$\begin{gathered}|\eta^\star_{x1} - \eta^\star_{x0'}| \leq \epsilon_w \\ 0 \leq \eta^\star_{xw} \leq 1\end{gathered}$$

then assuming $0 < \epsilon_w < 1$, we always obtain the following six extreme points,

$$\begin{gathered}(0,0) \\ (0,\epsilon_w) \\ (\epsilon_w,0) \\ (1-\epsilon_w,1) \\ (1,1-\epsilon_w) \\ (1,1)\end{gathered}$$

In general, however, once we introduce constraints $L^{YU}_{xw} \leq \eta^\star_{xw} \leq U^{XU}_{xw}$, the number of extreme points will vary. Moreover, when multiplied with the extreme points of the space $\delta^\star_1 \times \delta^\star_0$, the resulting extreme points of $\zeta^\star_{yx.w}$ might be included or excluded of the polytope depending on the relationship among $\{\epsilon_w, \epsilon_x, \epsilon_y\}$ and the observable $P(Y, X \mid W)$. Numerically, this is not a problem (barring numerical instabilities, which do occur with a nontrivial frequency). Algebraically, this makes the problem considerably complicated[29]. Instead, in what follows we will define a simpler framework that will not give tight constraints, but will shed light on the relationship between constraints, observable probabilities and the ϵ parameters. This will also be useful to scale up the full Witness Protection Program, as discussed in the main paper.

A.2 Methodology for Cross-W Constraints

Consider the standard IV model again, i.e., where W is exogenous with no direct effect on Y. So far, we have not replicated anything such as e.g. $\eta_1 \leq \zeta_{00.0} + \zeta_{11.0} + \zeta_{10.1} + \zeta_{11.1}$. We call this a "cross-W" constraint, as it relates observables under different values of $W \in \{0,1\}$. These are important when considering weakening the effect $W \to Y$. The recipe for deriving them will be as follows. Consider the template

$$\delta^\star_0 f_1(\eta^\star_0, \eta^\star_1) + \delta^\star_1 f_2(\eta^\star_0, \eta^\star_1) + f_3(\eta^\star_0, \eta^\star_1) \geq 0 \tag{36}$$

such that $f_i(\cdot,\cdot)$ are linear. Linearity is imposed so that this function will correspond to a linear function of $\{\zeta^\star, \eta^\star, \delta^\star\}$, of which expectations will give observed probabilities or interventional probabilities.

We will require that evaluating this expression at each of the four extreme points of the joint space $(\delta^\star_0, \delta^\star_1) \in \{0,1\}^2$ will translate into one of the basic constraints $1 - \eta^\star_i \geq 0$ or $\eta^\star_i \geq 0$, $i \in \{0,1\}$. This implies any combination of $\{\delta^\star_0, \delta^\star_1, \eta^\star_0, \eta^\star_1\}$ will satisfy (36) (more on that later).

29. As a counterpart, imagine we defined a polytope through the matrix inequality $A\mathbf{x} \leq \mathbf{b}$. If we want to obtain its extreme point representation as an algebraic function of the entries of matrix A and vector $\mathbf{b}$, this will be a complicated problem since we cannot assume we know the magnitudes and signs of the entries.

Given a choice of basic constraint (say, $\eta_1^\star \geq 0$), and setting $\delta_0^\star = \delta_1^\star = 0$, this immediately identifies $f_3(\cdot,\cdot)$. We assign the constraint corresponding to $\delta_0^\star = \delta_1^\star = 1$ with the "complementary constraint" for η_1 (in this case, $\eta_1^\star \leq 1$). This leaves two choices for assigning the remaining constraints.

Why do we associate the $\delta_0^\star = \delta_1^\star = 1$ case with the complementary constraint? Let us parameterize each function as $f_i(\eta_0^\star, \eta_1^\star) \equiv a_i\eta_0^\star + b_i\eta_1^\star + c_i$. Let $a_3 = q$, where either $q = 1$ (case $\eta_0^\star \geq 0$) or $q = -1$ (case $1 - \eta_0^\star \geq 0$). Without loss of generality, assume case $(\delta_0^\star = 1, \delta_1^\star = 0)$ is associated with the complementary constraint where the coefficient of $\eta_0^\star$ should be $-q$. For the other two cases, the coefficient of $\eta_0^\star$ should be 0 by construction. We get the system

$$\begin{aligned} a_3 &= q \\ a_1 + a_3 &= -q \\ a_2 + a_3 &= 0 \\ a_1 + a_2 + a_3 &= 0 \end{aligned}$$

This system has no solution. Assume instead $\delta_0^\star = \delta_1^\star = 1$ is associated with the complementary constraint where the coefficient of $\eta_0^\star$ should be $-q$. The system now is:

$$\begin{aligned} a_3 &= q \\ a_1 + a_3 &= 0 \\ a_2 + a_3 &= 0 \\ a_1 + a_2 + a_3 &= -q \end{aligned}$$

This system always have the solution $a_1 = a_2 = -q$. We do have freedom with b_1, b_2, b_3, which means we can choose to allocate the remaining two cases in two different ways.

Lemma 7 *Consider the constraints derived by the above procedure. Then any choice of* $(\delta_0^\star, \delta_1^\star, \eta_0^\star, \eta_1^\star) \in [0,1]^4$ *will satisfy these constraints.*

Proof Without loss of generality, let $f_3(\eta_0^\star, \eta_1^\star) = q\eta_0^\star + (1-q)/2$, $q \in \{-1, 1\}$. That is, $a_3 = q, b_3 = 0, c_3 = (1-q)/2$. This implies $a_1 = a_2 = -q$ (as above). Associating $(\delta_0^\star = 1, \delta_1^\star = 0)$ with $\eta_1^\star \geq 0$ gives $\{b_1 = 1, c_1 = (q-1)/2\}$ and consequently associating $(\delta_0^\star = 0, \delta_0^\star = 1)$ with $1 - \eta_1^\star \geq 0$ implies $\{b_2 = -1, c_2 = (1+q)/2\}$. Plugging this into the expression $\delta_0^\star f_1(\eta_0^\star, \eta_1^\star) + \delta_1^\star f_2(\eta_0^\star, \eta_1^\star) + f_3(\eta_0^\star, \eta_1^\star)$ we get

$$\begin{aligned} &= \delta_0^\star(-q\eta_0^\star + \eta_1^\star + (q-1)/2) + \delta_1^\star(-q\eta_0^\star - \eta_1^\star + (1+q)/2) + q\eta_0^\star + (1-q)/2 \\ &= \eta_0^\star(q - (\delta_0^\star + \delta_1^\star)q) + \eta_1^\star(\delta_0^\star - \delta_1^\star) + \delta_0^\star(q-1)/2 + \delta_1^\star(1+q)/2 + (1-q)/2 \\ &= \eta_0^\star(q - (\delta_0^\star + \delta_1^\star)q) + \eta_1^\star(\delta_0^\star - \delta_1^\star) + (-q + (\delta_0^\star + \delta_1^\star)q)/2 + (\delta_1^\star - \delta_0^\star + 1)/2 \\ \\ &= q((\delta_1^\star + \delta_0^\star) - 1)(1 - 2\eta_0^\star)/2 + ((\delta_1^\star - \delta_0^\star)(1 - 2\eta_1^\star) + 1)/2 \\ &= (\delta_1^\star + \delta_0^\star - 1)s/2 + (\delta_1^\star - \delta_0^\star)t/2 + 1/2 \end{aligned}$$

where $s = q(1 - 2\eta_0^\star) \in [-1, 1]$ and $t = (1 - 2\eta_1^\star) \in [-1, 1]$. Then evaluating at the four extreme points $s, t \in \{-1, +1\}$ we get $\delta_0, \delta_1, 1 - \delta_0, 1 - \delta_1$, all of which are non-negative. ■

The procedure derives 8 bounds (4 cases that we get by associating f_3 with either $\eta_x \geq 0$ or $1 - \eta_x \geq 0$. For each of these cases, 2 subcases what we get by assigning $(\delta_0^\star = 1, \delta_1^\star = 0)$

with either $\eta_{x'} \geq 0$ or $1 - \eta_{x'} \geq 0$). Now, for an illustration of one case:

Deriving a constraint for the standard IV model, example: $f_3(\eta_0^\star, \eta_1^\star) \equiv \eta_0^\star \geq 0$

Associate $\eta_1^\star \geq 0$ with assignment $(\delta_0^\star = 1, \delta_1^\star = 0)$ (implying we associate $\eta_1^\star \leq 1$ with assignment $(\delta_0^\star = 0, \delta_1^\star = 1)$ and $\eta_0^\star \leq 1$ with $(\delta_0^\star = 1, \delta_1^\star = 1)$). This uniquely gives $f_1(\eta_0^\star, \eta_1^\star) = \eta_1^\star - \eta_0^\star$, $f_2(\eta_0^\star, \eta_1^\star) = -\eta_1^\star - \eta_0^\star + 1$. The resulting expression is

$$\delta_0^\star(\eta_1^\star - \eta_0^\star) + \delta_1^\star(-\eta_1^\star - \eta_0^\star + 1) + \eta_0^\star \geq 0$$

from which we can verify that the assignment $(\delta_0^\star = 1, \delta_1^\star = 1)$ gives $\eta_0^\star \leq 1$. Now, we need to take the expectation of the above with respect to U to obtain observables ζ and causal distributions η. However, first we need some rearrangement so that we match $\eta_0^\star$ with corresponding $(1 - \delta_w^\star)$ and so on.

$$\begin{aligned}
\eta_1^\star(\delta_0^\star - \delta_1^\star) + \eta_0^\star(1 - \delta_0^\star - \delta_1^\star) + \delta_1^\star &\geq 0 \\
\eta_1^\star(\delta_0^\star - \delta_1^\star) + \eta_0^\star((1 - \delta_0^\star) + (1 - \delta_1^\star) - 1) + \delta_1^\star &\geq 0 \\
\zeta_{11.0}^\star - \zeta_{11.1}^\star + \zeta_{10.0}^\star + \zeta_{10.1}^\star - \eta_0^\star + \zeta_{01.1}^\star + \zeta_{11.1}^\star &\geq 0
\end{aligned}$$

Taking expectations and rearranging it, we have

$$\eta_0 \leq \zeta_{11.0} + \zeta_{10.0} + \zeta_{10.1} + \zeta_{01.1}$$

rediscovering one of the IV bounds for η_0. Choosing to associate $\eta_1^\star \geq 0$ with assignment $(\delta_0^\star = 0, \delta_1^\star = 1)$ will give instead

$$\eta_0 \leq \zeta_{11.1} + \zeta_{10.1} + \zeta_{10.0} + \zeta_{01.0}$$

Basically the effect of one of the two choices within any case is to switch $\zeta_{yx.w}$ with $\zeta_{yx.w'}$. ■

A.3 Deriving Cross-W Constraints

What is left is a generalization of that under the condition $|\eta_{xw} - \eta_{xw'}| \leq \epsilon_w$, $w \neq w'$, instead of $\eta_{xw} = \eta_{xw'}$. In this situation, we exploit the constraint $\underline{L} \leq \eta_{xw}^\star \leq \overline{U}$ instead of $0 \leq \eta_{xw}^\star \leq 1$ or $L_{xw}^{YU} \leq \eta_{xw}^\star \leq U_{xw}^{YU}$, where $\underline{L} \equiv \min\{L_{xw}^{YU}\}, \overline{U} \equiv \max\{U_{xw}^{YU}\}$. Using $L_{xw}^{YU} \leq \eta_{xw}^\star \leq U_{xw}^{YU}$ complicates things considerably. Also, we will not derive here the analogue proof of Lemma 1 for the case where $(\eta_0^\star, \eta_1^\star) \in [\underline{L}, \overline{U}]^2$, as it is analogous but with a more complicated notation.

Proof of Theorem 4 We demonstrate this through two special cases.
General Model, Special Case 1: $f_3(\eta_{0w}^\star, \eta_{1w}^\star) \equiv \eta_{xw}^\star - \underline{L} \geq 0$

There are two modifications. First, we perform the same associations as before, but with respect to $\underline{L} \leq \eta_{xw}^\star \leq \overline{U}$ instead of $0 \leq \eta_x^\star \leq 1$. Second, before we take expectations, we swap some of the $\eta_{xw}^\star$ with $\eta_{xw'}^\star$ up to some error ϵ_w.

Following the same sequence as in the example for the IV model, we get the resulting expression (where $x' \equiv \{0, 1\} \backslash x$):

$$\delta_w^\star(\eta_{x'w}^\star - \eta_{xw}^\star) + \delta_{w'}^\star(-\eta_{x'w}^\star - \eta_{xw}^\star + \overline{U} + \underline{L}) + \eta_{xw}^\star - \underline{L} \geq 0$$

from which we can verify that the assignment $(\delta^\star_w = 1, \delta^\star_{w'} = 1)$ gives $\overline{U} - \eta^\star_{xw} \geq 0$. Now, we need to take the expectation of the above with respect to U to obtain "observables" κ and causal effects ω. However, the difficulty now is that terms $\eta^\star_{xw}\delta^\star_{w'}$ and $\eta^\star_{xw'}\delta^\star_w$ have no observable counterpart under expectation. We get around this transforming $\eta^\star_{xw'}\delta^\star_w$ into $\eta^\star_{xw}\delta^\star_w$ (and $\eta^\star_{xw}\delta^\star_{w'}$ into $\eta^\star_{xw'}\delta^\star_{w'}$) by adding the corresponding correction $-\eta^\star_{xw} \leq -\eta^\star_{xw'} + \epsilon_w$:

$$\begin{array}{rcl}
\delta^\star_w(\eta^\star_{x'w} - \eta^\star_{xw}) + \delta^\star_{w'}(-\eta^\star_{x'w} - \eta^\star_{xw} + \overline{U} + \underline{L}) + \eta^\star_{xw} - \underline{L} & \geq & 0 \\
\delta^\star_w(\eta^\star_{x'w} - \eta^\star_{xw}) + \delta^\star_{w'}(-\eta^\star_{x'w'} + \epsilon_w - \eta^\star_{xw'} + \epsilon_w + \overline{U} + \underline{L}) + \eta^\star_{xw} - \underline{L} & \geq & 0 \\
\eta^\star_{x'w}\delta^\star_w + \eta^\star_{xw}(1 - \delta^\star_w) - \eta_{x'w'}\delta^\star_{w'} - \eta_{xw'}\delta^\star_{w'} + \delta^\star_{w'}(\overline{U} + \underline{L} + 2\epsilon_w) - \underline{L} & \geq & 0
\end{array}$$

Now, the case for $x = 1$ gives

$$\begin{array}{rcl}
\eta^\star_{0w}\delta^\star_w + \eta^\star_{1w}(1 - \delta^\star_w) - \eta_{0w'}\delta^\star_{w'} - \eta_{1w'}\delta^\star_{w'} + \ldots & \geq & 0 \\
\eta^\star_{0w}(1 - (1 - \delta^\star_w)) + \eta^\star_{1w}(1 - \delta^\star_w) - \eta^\star_{0w'}(1 - (1 - \delta^\star_{w'})) - \eta^\star_{1w'}\delta^\star_{w'} + \ldots & \geq & 0
\end{array}$$

Taking the expectations:

$$\omega_{0w} - \kappa_{10.w} + \omega_{1w} - \kappa_{11.w} - \omega_{0w'} + \kappa_{10.w'} - \kappa_{11.w'} + \chi_{w'}(\overline{U} + \underline{L} + 2\epsilon_w) - \underline{L} \geq 0 \tag{37}$$

Notice that for $\underline{\beta} = \overline{\beta} = 1$, $\underline{L} = 0$, $\overline{U} = 1$, $\epsilon_w = 0$, this implies $\eta_{xw} = \eta_{xw'}$ and this collapses to

$$\eta_{0w} - \zeta_{10.w} + \eta_{1w} - \zeta_{11.w} - \eta_{0w'} + \zeta_{10.w'} - \zeta_{11.w'} + \delta_{w'} \geq 0$$

$$\eta_{1w} \geq \zeta_{10.w} + \zeta_{11.w} - \zeta_{10.w'} - \zeta_{01.w'}$$

which is one of the lower bounds one obtains under the standard IV model.

The case for $x = 0$ is analogous and gives

$$\omega_{0w'} \leq \kappa_{11.w} + \kappa_{10.w} + \kappa_{10.w'} - \kappa_{11.w'} + \chi_{w'}(\overline{U} + \underline{L} + 2\epsilon_w) - \underline{L} \tag{38}$$

The next subcase is when we exchange the assignment of $(\delta^\star_w, \delta^\star_{w'})$ to other constraints. We obtain the following inequality:

$$\delta^\star_{w'}(\eta^\star_{x'w} - \eta^\star_{xw}) + \delta^\star_w(-\eta^\star_{x'w} - \eta^\star_{xw} + \overline{U} + \underline{L}) + \eta^\star_{xw} - \underline{L} \geq 0$$

which from an analogous sequence of steps leads to

$$\begin{array}{rcl}
\delta^\star_{w'}(\eta^\star_{x'w} - \eta^\star_{xw}) + \delta^\star_w(-\eta^\star_{x'w} - \eta^\star_{xw} + \overline{U} + \underline{L}) + \eta^\star_{xw} - \underline{L} & \geq & 0 \\
\delta^\star_{w'}(\eta^\star_{x'w'} + \epsilon_w - \eta^\star_{xw'} + \epsilon_w) + \delta^\star_w(-\eta^\star_{x'w} - \eta^\star_{xw} + \overline{U} + \underline{L}) + \eta^\star_{xw} - \underline{L} & \geq & 0 \\
\eta^\star_{x'w'}\delta^\star_{w'} - \eta^\star_{xw'}\delta^\star_{w'} + 2\delta^\star_{w'}\epsilon_w - \eta^\star_{x'w}\delta^\star_w + \eta^\star_{xw}(1 - \delta^\star_w) + \delta^\star_w(\overline{U} + \underline{L}) - \underline{L} & \geq & 0
\end{array}$$

For $x = 1$,

$$\begin{array}{rcl}
\eta^\star_{0w'}\delta^\star_{w'} - \eta^\star_{1w'}\delta^\star_{w'} + \eta^\star_{0w}\delta^\star_w + \eta^\star_{1w}(1 - \delta^\star_w) + \ldots & \geq & 0 \\
\eta^\star_{0w'}(1 - (1 - \delta^\star_{w'})) - \eta^\star_{1w'}\delta^\star_{w'} - \eta^\star_{0w}(1 - (1 - \delta^\star_w)) + \eta^\star_{1w}(1 - \delta^\star_w) + \ldots & \geq & 0
\end{array}$$

Taking expectations,

$$\omega_{0w'} - \kappa_{10.w'} - \kappa_{11.w'} - \omega_{0w} + \kappa_{10.w} + \omega_{1w} - \kappa_{11.w} + 2\chi_{w'}\epsilon_w + \chi_w(\overline{U} + \underline{L}) - \underline{L} \geq 0 \tag{39}$$

For $x = 0$,

$$\begin{aligned} \eta^\star_{1w'}\delta^\star_{w'} - \eta^\star_{0w'}\delta^\star_{w'} + \eta^\star_{1w}\delta^\star_w + \eta^\star_{0w}(1-\delta^\star_w) + \ldots &\geq 0 \\ \eta^\star_{1w'}\delta^\star_{w'} - \eta^\star_{0w'}(1-(1-\delta^\star_{w'})) - \eta^\star_{1w}\delta^\star_w + \eta^\star_{0w}(1-\delta^\star_w) + \ldots &\geq 0 \\ \kappa_{11.w'} - \omega_{0w'} + \kappa_{10.w'} - \kappa_{11.w} + \kappa_{10.w} + 2\chi_{w'}\epsilon_w + \chi_w(\overline{U}+\underline{L}) - \underline{L} &\geq 0 \end{aligned}$$

$$\omega_{0w'} \leq \kappa_{11.w'} + \kappa_{10.w'} - \kappa_{11.w} + \kappa_{10.w} + 2\chi_{w'}\epsilon_w + \chi_w(\overline{U}+\underline{L}) - \underline{L} \tag{40}$$

General Model, Special Case 2: $f_3(\eta^\star_{0w}, \eta^\star_{1w}) \equiv \overline{U} - \eta^\star_{xw} \geq 0$

Associate $\eta^\star_{x'w} \geq \underline{L}$ with assignment $(\delta^\star_w = 1, \delta^\star_{w'} = 0)$ (implying we associate $\eta^\star_{x'w} \leq \overline{U}$ with assignment $(\delta^\star_w = 0, \delta^\star_{w'} = 1)$ and $\eta^\star_{xw} \geq \underline{L}$ with $(\delta^\star_w = 1, \delta^\star_{w'} = 1)$). The resulting expression is

$$\delta^\star_w(\eta^\star_{x'w} + \eta^\star_{xw} - \overline{U} - \underline{L}) + \delta^\star_{w'}(-\eta^\star_{x'w} + \eta^\star_{xw}) + \overline{U} - \eta^\star_{xw} \geq 0$$

Following the same line of reasoning as before, we get this for $x = 1$:

$$\omega_{0w} - \omega_{0w'} - \omega_{1w} - \kappa_{10.w} + \kappa_{11.w} + \kappa_{10.w'} + \kappa_{11.w'} - \chi_w(\overline{U}+\underline{L}) + 2\epsilon_w\chi_{w'} + \overline{U} \geq 0 \tag{41}$$

We get this for $x = 0$:

$$\omega_{0w'} \geq -\kappa_{11.w} + \kappa_{10.w} + \kappa_{11.w'} + \kappa_{10.w'} + \chi_w(\overline{U}+\underline{L}) - 2\epsilon_w\chi_{w'} - \overline{U} \tag{42}$$

With the complementary assignment, we start with the relationship

$$\delta^\star_{w'}(\eta^\star_{x'w} + \eta^\star_{xw} - \overline{U} - \underline{L}) + \delta^\star_w(-\eta^\star_{x'w} + \eta^\star_{xw}) + \overline{U} - \eta^\star_{xw} \geq 0$$

For $x = 1$,

$$\omega_{0w'} - \omega_{0w} - \omega_{1w} - \kappa_{10.w'} + \kappa_{11.w'} + \kappa_{10.w} + \kappa_{11.w} + \chi_{w'}(2\epsilon_w - \overline{U} - \underline{L}) + \overline{U} \geq 0 \tag{43}$$

For $x = 0$,

$$\omega_{0w'} \geq -\kappa_{11.w'} + \kappa_{10.w'} + \kappa_{11.w} + \kappa_{10.w} - \chi_{w'}(2\epsilon_w - \overline{U} - \underline{L}) - \overline{U} \tag{44}$$

Notice that the bounds obtained are asymmetric in x, i.e., we derive different bounds for ω_{0w} and ω_{1w}. Symmetry is readily obtained by the same derivation where $\delta^\star_w$ is interpreted as $P(X = 0 \mid W = w, U)$ and x is swapped with x'. ■

A.4 Linear Case

Our final proof refers to results introduced in Section 6.

Proof of Theorem 5 Variable s_{xx} appears only in Equation (24) and the inequalities $0 \leq s_{xx} \leq 1$. The intersection of these relationships is satisfiable if and only if $0 \leq a^2 + 2as_{wx} \leq 1$ is satisfiable. Moreover, s_{wx} appears only in Equation (22). Solving this equation for s_{wx} and plugging it in $0 \leq a^2 + 2as_{wx} \leq 1$, we obtain $0 \leq -a^2 + 2a\rho_{wx} \leq 1$. The quadratic

expression for a achieves a unique maximum at $a^\star = \rho_{wx}$, implying $-a^{\star 2}+2a^\star\rho_{wx} = \rho_{wx}^2 \leq 1$. We can then drop the inequality $-a^2 + 2a\rho_{wx} \leq 1$ as this is always satisfied. The resulting interval is $a^2 - 2a\rho_{wx} \leq 0$, and the set of values of a satisfying it is either the interval $[2\rho_{wx}, 0]$ or $[0, 2\rho_{wx}]$ depending on the sign of ρ_{wx}. This can be written as

$$\min(0, 2\rho_{wx}) \leq a \leq \max(0, 2\rho_{wx}) \tag{45}$$

The intersection of Equation (22) and $-\epsilon_{wx} \leq s_{wx} \leq \epsilon_{wx}$ is satisfiable only if $\rho_{wx}-\epsilon_{wx} \leq a \leq \rho_{wx} + \epsilon_{wx}$. Combining this interval with (45), we obtain the inequality

$$\max(\min(0, 2\rho_{wx}), \rho_{wx} - \epsilon_{wx}) \leq a \leq \min(\max(0, 2\rho_{wx}), \rho_{wx} + \epsilon_{wx}) \tag{46}$$

which $\mathcal{L}_a$ and $\mathcal{U}_a$ being defined as the lower and upper bounds, respectively, in the interval above.

Since a now only appears in (46) and Equation (25), and assuming $s_{wy} \geq 0$, the intersection of the two equations is satisfiable if and only if

$$\rho_{xy} - \mathcal{U}_a s_{wy} \leq b + c\rho_{wx} + s_{xy} \leq \rho_{xy} - \mathcal{L}_a s_{wy} \tag{47}$$

Equation (31) follows from s_{xy} not appearing anywhere else but in the relationship $-\epsilon_{xy} \leq s_{xy} \leq \epsilon_{xy}$, and also considering the case $s_{wy} < 0$.

Equation (26) and the relationship $0 \leq s_{yy} \leq 1$ is satisfiable if and only if

$$0 \leq b^2 + 2bc\rho_{wx} + c^2 + 2[b(as_{wy} + s_{xy}) + cs_{wy}] \leq 1 \tag{48}$$

is satisfiable. From Equation (25) we have $as_{wy} + s_{xy} = \rho_{xy} - b - c\rho_{wx}$ and from Equation (23) we have $s_{wy} = \rho_{wy} - b\rho_{wx} - c$. Making these substitutions into (48), we get

$$0 \leq -b^2 - 2bc\rho_{wx} - c^2 + 2(b\rho_{xy} + c\rho_{wy}) \leq 1.$$

The quadratic function in (b, c) has a unique maximum. Assuming for now c is unconstrained, taking the derivatives of $-b^2 - 2bc\rho_{wx} - c^2 + 2(b\rho_{xy} + c\rho_{wy})$ with respect to b and c, and setting them to zero, we obtain the stationary point

$$b^\star = \frac{\rho_{xy} - \rho_{wx}\rho_{wy}}{1 - \rho_{wx}^2}, \quad c^\star = \frac{\rho_{wy} - \rho_{wx}\rho_{xy}}{1 - \rho_{wx}^2}, \tag{49}$$

and by assumption it follows that $c^\star = 0$. Plugging $c^\star$ in $-b^2 - 2bc\rho_{wx} - c^2 + 2(b\rho_{xy} + c\rho_{wy})$, we get $b(2\rho_{xy} - b) \leq \rho_{xy}^2 \leq 1$. So, it is sufficient to satisfy

$$b^2 + 2bc\rho_{wx} + c^2 - 2(b\rho_{xy} + c\rho_{wy}) \leq 0. \tag{50}$$

■

References

J. Altonji, T. Elder, and C. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113:151–184, 2005.

A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92:1171–1176, 1997.

W. Buntine. Theory refinement on Bayesian networks. *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence (UAI 1991)*, pages 52–60, 1991.

W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. *Proceedings of 20th Conference on Uncertainty in Artificial Intelligence (UAI 2004)*, pages 59–66, 2004.

Z. Cai, M. Kuroki, J. Pearl, and J. Tian. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64:695–701, 2008.

L. Chen, F. Emmert-Streib, and J. D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8:R219, 2007.

G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 2, 1997.

J. Cornfield, W. Haenszel, E. Hammond, A. Lilienfeld, M. Shimkin, and E. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203, 1959.

A.P. Dawid. Causal inference using influence diagrams: the problem of partial compliance. In P.J. Green, N.L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 45–65. Oxford University Press, 2003.

G. Elidan. Copulas in Machine Learning. *Copulae in Mathematical and Quantitative Finance, Lecture Notes in Statistics*, pages 39–60, 2013.

D. Entner, P.O. Hoyer, and P. Spirtes. Statistical test for consistent estimation of causal effects in linear non-Gaussian models. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, pages 364–372, 2012.

D. Entner, P. Hoyer, and P. Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, pages 256–264, 2013.

R. Evans. Graphical methods for inequality constraints in marginalized DAGs. *Proceedings of the 22nd Workshop on Machine Learning and Signal Processing*, 2012.

N. Friedman and D. Koller. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning Journal*, 50:95–126, 2003.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

N. Harris and M. Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14:3365–3383, 2013.

K. Hirano, G. Imbens, D. Rubin, and X.-H. Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biometrics*, 1:69–88, 2000.

P. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1:265–283, 2007.

C. Kleiber and A. Zeileis. *Applied Econometrics with R*. Springer-Verlag, 2008.

J. Kuipers, G. Moffa, and D. Heckerman. Addendum on the scoring of Gaussian directed acyclic graphical models. *Annals of Statistics*, 42:1689–1691, 2014.

S. Mani, G. Cooper, and P. Spirtes. A theoretical study of Y structures for causal discovery. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI2006)*, pages 314–323, 2006.

C. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2007.

C. McDonald, S. Hiu, and W. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD Computing*, 9:304–312, 1992.

C. Meek. Strong completeness and faithfulness in Bayesian networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 411–418, 1995.

J. Mooij and J. Cremers. An empirical study of one of the simplest causal prediction algorithms. *Proceedings of the Advances in Causal Inference Workshop, UAI 2015*, 2015.

T. Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55:765–799, 1987.

R. Nelsen. *An Introduction to Copulas*. Springer-Verlag, 2007.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

J. Pearl. Myth, confusion, and science in causal analysis. *UCLA Cognitive Systems Laboratory, Technical Report (R-348)*, 2009.

R. Ramsahai. Causal bounds and observable constraints for non-deterministic models. *Journal of Machine Learning Research*, 13:829–848, 2012.

R. Ramsahai and S. Lauritzen. Likelihood analysis of the binary instrumental variable model. *Biometrika*, 98:987–994, 2011.

J. Ramsey, P. Spirtes, and J. Zhang. Adjacency-faithfulness and conservative causal inference. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 401–408, 2006.

T. Richardson and J. Robins. Analysis of the binary instrumental variable model. In R. Dechter, H. Geffner, and J.Y. Halpern, editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 415–444. College Publications, 2010.

T. Richardson, R. Evans, and J. Robins. Transparent parameterizations of models for potential outcomes. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 9*, pages 569–610. Oxford University Press, 2011.

J. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.

P. Rosenbaum. *Observational Studies*. Springer-Verlag, 2002a.

P. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002b.

K. Rothman, S. Greenland, and T. Lash. *Modern Epidemiology*. Wolters Kluwer, 2008.

R. Silva and R. Evans. Causal inference through a witness protection program. *Advances in Neural Information Processing Systems*, 27:298–306, 2014.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Cambridge University Press, 2000.

K. Steenland and S. Greenland. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology*, 160:384–392, 2004.

T. VanderWeele and I. Shpitser. A new criterion for confounder selection. *Biometrics*, 64: 1406–1413, 2011.