On Testability of the Front-Door Model via Verma Constraints: Appendix

Rohit Bhattacharya[1] **Razieh Nabi**[2]

[1]Department of Computer Science, Williams College, Williamstown, Massachusetts, USA
[2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA

The supplementary materials are organized as follows. Appendix A provides additional commentary on assumptions used in the testability criterion, while Appendix B lists all ADMGs that satisfy the criterion. Appendices C, D, E, and F provide additional notes on non-parametric Verma tests, models with additional baseline covariates, the nested Markov factorization, and the simulated datasets respectively. Appendix G contains proofs of the main results.

A ON THE TRUE MEDIATION AND RELEVANCE ASSUMPTIONS

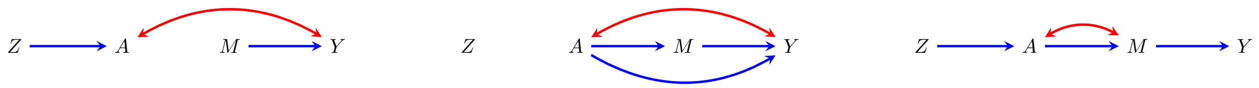


Figure 1: (a) A model that does not satisfy M being a mediator between A and Y; (b) A model that does not satisfy $Z \not\perp\!\!\!\perp A$; (c) A model that does not satisfy $Z \not\perp\!\!\!\perp Y|A, M$.

Assumptions (A1) (true mediation) and (A2) (relevance of the anchor) help ensure that any Verma constraint detected between the anchor Z and outcome Y is "non-trivial," i.e., is not implied by an ordinary independence constraint. We first show why this is important by way of example, and then show that no ordinary constraint exists between Z and Y in distributions satisfying (A1) and (A2). Before proceeding, we also briefly note that (A1) is a relatively weak assumption in that it relies on some minimal background knowledge about how the treatment affects the outcome, while (A2) is testable.

Consider the ADMG in Fig. 1(a). Here, M does not satisfy assumption (A1). It can be checked via m-separation that any distribution that nested factorizes with respect to this ADMG would encode the ordinary independence $Z \perp\!\!\!\perp Y$. Trivially, this constraint would also hold in the post-intervention distribution $p(Z, A, M, Y)/p(M|A, Z)$. Fig. 1(b) demonstrates why accidentally testing for trivial Verma constraints such as this one may lead to incorrect conclusions about identifiability of the target $\mathbb{E}[Y \,|\, \text{do}(a)]$. In Fig. 1(b), we have that $Z \perp\!\!\!\perp A$, so assumption (A2) is not satisfied. There exist no paths between Z and Y, leading to a trivial Verma constraint implied by the ordinary independence $Z \perp\!\!\!\perp Y$. Relying on a test of this trivial constraint would also mislead the analyst into thinking $\mathbb{E}[Y \,|\, \text{do}(a)]$ is identified, when in fact, the front-door conditions are not met due to the $A \to Y$ edge. Finally, consider the ADMG in Fig. 1(c). Distributions factorizing according to this graph satisfy $Z \perp\!\!\!\perp Y \mid A, M$, which violates assumption (A2). These distributions will also exhibit the same trivial Verma constraint, and here too, relying on such a test would lead to incorrect conclusions on identifiability. The following lemma confirms that under the given assumptions such trivial constraints do not arise.

Lemma 1. *Distributions $p(Z, A, M, Y)$ satisfying assumptions (A1) and (A2) do not entail any ordinary independence constraints between Z and Y.*

Proof. Given (A1), $p(Z, A, M, Y)$ must nested factorize wrt an ADMG, say $\mathcal{G}$, containing the path $A \to M \to Y$. Given $Z \not\perp\!\!\!\perp A$ and Z is not a causal consequence of A due to (A2), $\mathcal{G}$ also contains either $Z \to A$ or $Z \leftrightarrow A$ or both. This implies the existence of an unblocked path $Z \to A \to M \to Y$ and/or a similar one that starts with $Z \leftrightarrow A$. These paths become

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

blocked when we condition on A and M. However, since by assumption (A2), we have that $Z \not\perp\!\!\!\perp Y | A, M$ at least one of the following paths consisting of colliders exist in $\mathcal{G}$ (we use the notation $\circ\!\!\rightarrow$ below to mean a directed or bidirected edge):

- The edge $Z \circ\!\!\rightarrow Y$ exists in $\mathcal{G}$ (trivially a collider path.)
- A path $Z \circ\!\!\rightarrow A \leftrightarrow Y$ exists in $\mathcal{G}$ where A is a collider.
- A path $Z \circ\!\!\rightarrow M \leftrightarrow Y$ exists in $\mathcal{G}$ where M is a collider.
- A path $Z \circ\!\!\rightarrow A \leftrightarrow M \leftrightarrow Y$ and/or $Z \circ\!\!\rightarrow M \leftrightarrow A \leftrightarrow Y$ exists in $\mathcal{G}$ where both A and M are colliders.

Note that conditioning on a descendant of a collider also opens the collider; however, given just these 4 variables, this situation arises only when we condition on M with A being the collider (a case which is already covered above.) An inducing path between two variables P and Q is defined as a path (comprised of any combination of directed and bidirected edges) between two variables such that every intermediate node is a collider and a causal ancestor of either P or Q. If such a path exists, then it is known that P and Q are not m-separable given any subset of other variables, i.e., there is no ordinary independence constraint between them [Verma and Pearl, 1990]. Here, any of the above paths would form inducing paths ($Z \circ\!\!\rightarrow Y$ is trivially an inducing path), as by assumption (A1) both A and M are causal ancestors of Y. Hence, there are no ordinary conditional independence constraints implied in distributions $p(Z, A, M, Y)$ that satisfy assumptions (A1-A3). □

B EXHAUSTIVE LIST OF ADMGS SATISFYING THE TESTABLE CRITERION

Fig. 2 is an exhaustive list of all ADMGs that satisfy the Verma restriction in Theorem 1. Importantly, all of these ADMGs also satisfy the front-door criteria (or its extensions) that permit identification of $\mathbb{E}[Y | \operatorname{do}(a)]$. The ADMGs in Fig. 2(a-c) are ones where the anchor Z also satisfies the instrumental variable conditions. A version of Fig. 2 appears in Shpitser et al. [2014] as the conjectured list of ADMGs that satisfy a Verma restriction which distinguishes it from a supermodel where the $Z \rightarrow Y$ edge is present. This conjecture was based on empirical comparisons of Bayesian Information Criterion scores of all 4 variable ADMGs using a parameterization of the nested Markov model for discrete data [Evans and Richardson, 2014].

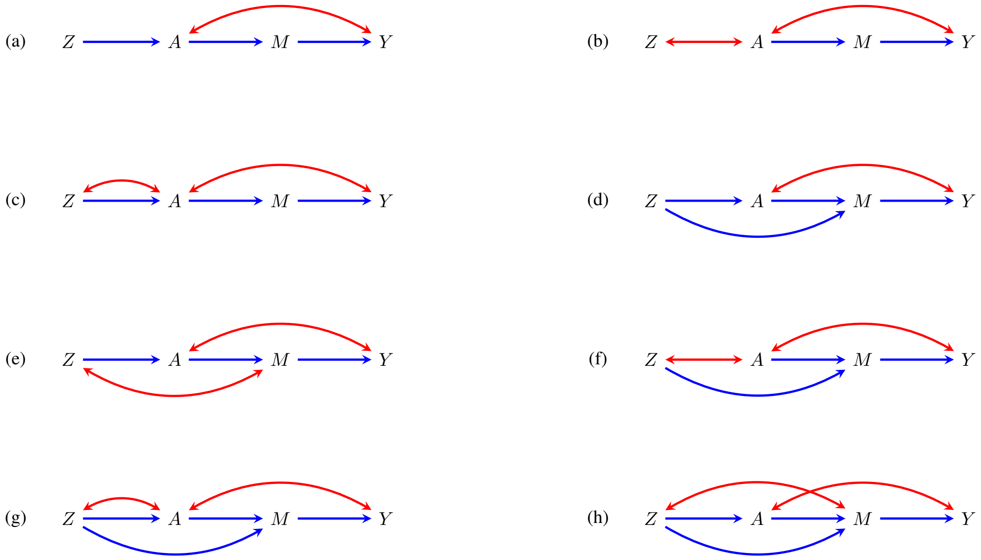


Figure 2: An exhaustive list of all ADMGs that satisfy the Verma constraint $Z \perp\!\!\!\perp Y$ in $p(Z, A, M, Y)/p(M|A, Z)$.

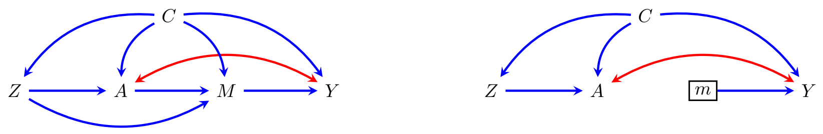
C ADDITIONAL DETAILS ON NON-PARAMETRIC TESTS

In this section we provide additional details on how a non-parametric Verma test can be performed using our methods. Assume we are given a data set $\mathcal{S}_n$ with n i.i.d samples of data, and any appropriate non-parametric test $\tau(Z, Y, M)$ that can be used to test an ordinary conditional independence $Z \perp\!\!\!\perp Y \mid M$. The idea for a non-parametric Verma test is simple: Verma constraints resemble ordinary conditional independencies albeit in identified post-intervention distributions. Thus, if we are able to generate a pseudo-dataset $\mathcal{S}^p_{(\cdot)}$ that resembles the post-intervention distribution in which the desired dormant conditional independence holds, then $\tau(Z, Y, M)$ can easily be applied to $\mathcal{S}^p_{(\cdot)}$ rather than $\mathcal{S}_n$. More concretely, to use the dual weights for a non-parametric Verma test we would proceed as follows:

1. Compute $q^d(M|A, Z) \equiv p(M|A, Z)/p(M|A = a, Z)$ for each row of data $i = 1, \dots, n$ using suitable machine learning methods, e.g., kernel regressions or random forests.
2. Create a pseudo-dataset $\mathcal{S}^p_{n/2}$ that mimics the post-intervention distribution $p(Z, M, Y \mid \text{do}(a))$ by drawing $n/2$ samples with replacement from $\mathcal{S}_n$, where each row i has an (unnormalized) probability of being sampled given by the dual weights $1/q^d(M = m_i|A = a_i, Z = z_i)$.
3. Apply $\tau(Z, Y, M)$ to $\mathcal{S}^p_{n/2}$ with some pre-specified significance level α.

Note that some of the statistical inefficiency in our tests come from resampling only half the data from the original dataset when performing the Verma test. This is the simplest sampling scheme proposed by Thams et al. [2021], but the authors have also proposed more complex adaptive schemes that may improve performance; see their paper for details. A non-parametric test using primal weights would proceed in exactly the same way except we would fit the appropriate models to generate the primal weights $1/\widetilde{q}(A \mid Y, Z, M)$.

D MODELS WITH BASELINE COVARIATES



(a) $p(C, Z, A, M, Y)$

(b) $p(C, Z, A, Y \mid \text{do}(m))$

Figure 3: (a) A modification of the model in Fig. 1(b) to include baseline covariates C; (b) CADMG showing that the Verma constraint still exists as a conditional independence $Z \perp\!\!\!\perp Y \mid C$ in $p(C, Z, A, Y \mid \text{do}(m))$.

In this section we make explicit how our framework extends to settings with baseline covariates. Consider the ADMG in Fig. 3(a) that is a modification of Fig. 1(b) to include baseline covariates C that point to all other variables Z, A, M, Y. We show that the inclusion of additional covariates requires only small modifications to the underlying theory in the main paper.

First, notice that there is still no ordinary conditional independence between Z and Y implied by Fig. 3(a). Further, the post-intervention distribution $p(C, Z, A, Y \mid \text{do}(m))$ is identified as $p(C, Z, A, Y, M = m)/p(M = m|A, Z, C)$. This post-intervention distribution factorizes according to the CADMG shown in Fig. 3(b) from which we can read off the dormant conditional independence $Z \perp\!\!\!\perp Y \mid C$ in $p(C, Z, A, Y \mid \text{do}(m))$. That is, the Verma constraint between the anchor and the outcome is now a *conditional* rather than marginal independence in the post-intervention distribution, and the distribution itself is obtained using a propensity score for M that includes C in the conditioning set. The identifying functional for the post-intervention distribution where we intervene on A is also very similar: $p(C, Z, M, Y \mid \text{do}(a)) = p(C, Z) \times p(M|A = a, Z, C) \times \sum_A p(A|C, Z) \times p(Y|Z, A, M, C)$.

Based on the above observations, the modifications to the theory are as follows. Assumptions (A1) and (A3) remain the same, however, the relevance assumption (A2) now includes C in the conditioning set (similar to the relevance assumption in conditional IV models.) That is, (A2) is modified to read: Z is a covariate that is *not* a causal consequence of A such that $Z \not\perp\!\!\!\perp A \mid C$ and $Z \not\perp\!\!\!\perp Y \mid A, M, C$. As mentioned earlier, the Verma constraint that allows us to test that A has no bidirected path to its children is then the dormant *conditional* independence $Z \perp\!\!\!\perp Y \mid C$ in $p(Z, A, M, Y, C)/p(M|A, Z, C)$. The tests proposed in Section 5 are modified appropriately by defining $\widetilde{q}(A|Y, Z, M, C)$ and $q^d(M|A, Z, C)$ as:

$$\widetilde{q}(A \mid Y, Z, M, C) \equiv \frac{p(A \mid Z, C) \times p(Y \mid A, M, Z, C)}{\sum_A p(A \mid Z, C) \times p(Y \mid A, M, Z, C)}$$

$$q^d(M \mid A, Z, C) \equiv \frac{p(M \mid A, Z, C)}{p(M \mid A = a, Z, C)}.$$

The primal and dual weights are then used to test $Z \perp\!\!\!\perp Y \mid M, C$ in $p(C, Z, A, M, Y)/\widetilde{q}(A|Y, Z, M, C)$ and $Z \perp\!\!\!\perp Y \mid M, C$ in $p(C, Z, A, M, Y)/q^d(M|A, Z, C)$. The same weights can be re-used for causal effect estimation as before.

Hence, the proposed framework does not exclude the possibility of including additional baseline covariates. Similar to conditional IV models, the inclusion of such covariates can be beneficial from an identification standpoint by reducing the possibility of unmeasured confounding on the causal path from $A \rightarrow M \rightarrow Y$. The addition of these covariates may also yield additional statistical efficiency in the test and downstream causal effect estimation.

E NESTED MARKOV FACTORIZATION OF ADMGS

The nested Markov factorization of $p(V)$ with respect to an ADMG $\mathcal{G}(V)$ is defined using conditional ADMGs (CADMGs) derived from $\mathcal{G}(V)$ and kernel objects derived from $p(V)$ via a *fixing* operation [Richardson et al., 2017]. The fixing operation can be causally interpreted as an application of the g-formula on a single variable to a given graph-kernel pair to obtain another graph-kernel pair corresponding to a conditional ADMG and a post-intervention distribution that factorizes according to it. We define the relevant concepts below.

Conditional ADMG (CADMG) – A CADMG $\mathcal{G}(V, W)$ is an ADMG whose vertices can be partitioned into random variables V and fixed/intervened variables W. Only outgoing directed edges may be adjacent to variables in W. For any random variable $V_i \in V$ in a CADMG $\mathcal{G}(V, W)$, the usual definitions of genealogical relations, such as parents and descendants, extend naturally by allowing for the inclusion of fixed variables into these sets. However, bidirected connected components a.k.a districts of a CADMG $\mathcal{G}(V, W)$ are only defined for elements of V.

Kernel – A kernel $q_V(V \mid W)$ is a mapping from values in W to normalized densities over V. That is, a kernel acts in most respects like an ordinary conditional distribution. In particular, given a kernel $q_V(V \mid W)$ and a subset $X \subseteq V$, conditioning and marginalization are defined in the usual way as $q_X(X \mid W) \equiv \sum_{V \setminus X} q_V(V \mid W)$ and $q_V(V \setminus X \mid X, W) \equiv q_V(V \mid W)/q_V(X \mid W)$.

Fixing operation for a single variable – Fixability of a single variable and the graphical and probabilistic operations of fixing it are defined as follows:

- *Fixability* – A vertex $V_i \in V$ is said to be *fixable* in $\mathcal{G}(V, W)$ if $V_i \rightarrow \ldots \rightarrow V_j$ and $V_i \leftrightarrow \ldots \leftrightarrow V_j$ do not both exist in $\mathcal{G}$, for any $V_j \in V \setminus V_i$.
- *Graphical operation of fixing* – The graphical operation of fixing V_i, denoted by $\phi_{V_i}(\mathcal{G})$, yields a new CADMG $\mathcal{G}(V \setminus V_i, W \cup V_i)$ where all edges with arrowheads into V_i are removed, and V_i is fixed to a particular value v_i. All other edges in $\mathcal{G}$ are kept the same.
- *Probabilistic operation of fixing* – Given a kernel $q_V(V \mid W)$, the associated CADMG $\mathcal{G}(V, W)$, and $V_i \in V$, the corresponding probabilistic operation of fixing V_i, denoted by $\phi_{V_i}(q_V; \mathcal{G})$, yields a new kernel defined as follows:

$$\phi_{V_i}(q_V; \mathcal{G}) \equiv q_{V \setminus V_i}(V \setminus V_i \mid W \cup V_i) \equiv \frac{q_V(V \mid W)}{q_V(V_i \mid \text{mb}_{\mathcal{G}}(V_i), W)}, \tag{1}$$

  where $\text{mb}_{\mathcal{G}}(V_i)$ denotes the Markov blanket of V_i, which consists of all vertices in the district of V_i and the parents of the district of V_i (excluding V_i itself.)

Fixing operation for a set of variables – The above definitions can be extended to a set of vertices S as follows:

- *Fixability* – A set $S \subseteq V$ is said to be fixable if there exists an ordering $(S_1, \ldots, S_p)$ such that S_1 is fixable in $\mathcal{G}$, S_2 is fixable in $\phi_{S_1}(\mathcal{G})$, and so on. Such an ordering is said to form a valid fixing sequence for S and we denote it by σ_S.
- *Graphical operation* – It is known that applying any two valid fixing sequences on S yield the same CADMG, so we denote this by $\phi_S(\mathcal{G}(V, W))$.
- *Probabilistic operation* – $\phi_{\sigma_S}(q_V; \mathcal{G})$ is defined via the usual function composition to yield operators that fix all elements in S in the order given by σ_S.

Intrinsic set – A set D is said to be *intrinsic* in $\mathcal{G}(V)$ if $V \setminus D$ is fixable in $\mathcal{G}$[1] and $\phi_{V \setminus D}(\mathcal{G})$ contains a single district.

Finally, a distribution $p(V)$ is said to satisfy the nested Markov factorization wrt an ADMG $\mathcal{G}(V)$ if there exists a set of kernels $q_D(D \mid \text{pa}_{\mathcal{G}}(D))$, one for every D intrinsic in $\mathcal{G}(V)$, s.t. for every fixable set S and every valid fixing sequence σ_S,

$$\phi_{\sigma_S}(p(V); \mathcal{G}) = \prod_{D \in \mathcal{D}(\phi_S(\mathcal{G}))} q_D(D \mid \text{pa}_{\mathcal{G}}(D)), \tag{2}$$

where $\mathcal{D}(\phi_S(\mathcal{G}))$ denotes the set of all districts in the CADMG $\phi_S(\mathcal{G})$. The nested Markov factorization asserts that every kernel that can be derived via a valid sequence of fixing satisfies the district factorization with respect to the CADMG obtained by this sequence, and each of the kernels appearing in the factorization corresponds to intrinsic sets.

[1] That is, from a causal perspective, $q_D(D \mid \text{pa}_{\mathcal{G}}(D)))$ is identified from $p(V)$ via sequential applications of the g-formula.

F DETAILS ON SIMULATED DATA

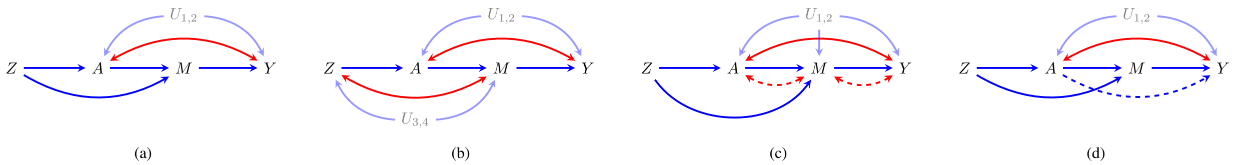


Figure 4: Data for the simulations are generated according to: (a, b) Two hidden variable causal DAGs and the corresponding ADMGs that satisfy the front-door assumptions; (c, d) Two hidden variable causal DAGs and the corresponding ADMGs that do not satisfy the front-door assumptions due to additional confounding and a direct effect of A on Y respectively.

The causal graphs used to perform experiments in Task (i) of Section 7 are shown in Fig. 4 – underlying hidden variables are depicted with lower opacity to highlight the structure of the ADMG. The first two graphs are ones in which A has no bidirected path to its children and the Verma constraint $Z \perp\!\!\!\perp Y$ in $p(Z, A, Y \mid \text{do}(m))$ holds; the latter two are ones in which there is additional confounding or violation of the exclusion restriction, which also coincide with the absence of any non-parametric equality constraint between Z and Y. We first describe generating data according to the hidden variable causal model in Fig. 4(a) whose observed margin would nested factorize wrt to the ADMG shown in the same figure. Each coefficient $\beta_{(\cdot)}$ appearing in any of the equations below is generated randomly from a uniform distribution Uniform$(1, 2)$.

$$\begin{aligned}
&U_1 = \text{Uniform}(-1, 1)\\
&p(U_2 = 1) = \text{expit}(0.5)\\
&p(Z = 1) = \text{expit}(0.5)\\
&p(A = 1 \mid Z, U_1, U_2) = \text{expit}(-0.5 + \beta_{ZA} Z - \beta_{U_1 A} U_1 + \beta_{U_2 A} U_2)\\
&p(M = 1 \mid Z, A) = \text{expit}(-0.5 - \beta_{ZM} Z + \beta_{AM} A)\\
&Y \mid U_1, U_2, M = \beta_{U_1 Y} U_1 + \beta_{U_2 Y} U_2 - \beta_{MY} M + \text{Normal}(0, 1).
\end{aligned}$$

Similarly, data generation for Fig. 4 (b) is performed as follows:

$$\begin{aligned}
&U_1 = \text{Uniform}(-1, 1)\\
&U_3 = \text{Uniform}(-1, 1)\\
&p(U_2 = 1) = \text{expit}(0.5)\\
&p(U_4 = 1) = \text{expit}(0.5)\\
&p(Z = 1 \mid U_3, U_4) = \text{expit}(0.5 + \beta_{U_3 Z} U_3 - \beta_{U_4 Z} U_4)\\
&p(A = 1 \mid Z, U_1, U_2) = \text{expit}(-0.5 + \beta_{ZA} Z - \beta_{U_1 A} U_1 + \beta_{U_2 A} U_2)\\
&p(M = 1 \mid U_3, U_4, A) = \text{expit}(-0.5 - \beta_{U_3 M} U_3 + \beta_{U_4 M} U_4 + \beta_{AM} A)\\
&Y \mid U_1, U_2, M = \beta_{U_1 Y} U_1 + \beta_{U_2 Y} U_2 - \beta_{MY} M + \text{Normal}(0, 1).
\end{aligned}$$

Data generation for Fig. 4(c) is similar to Fig. 4(a) except the equation for M is modified to include $U_{1,2}$ as follows,

$$p(M = 1 \mid Z, A, U_1, U_2) = \text{expit}(-0.5 - \beta_{ZM} Z + \beta_{AM} A + \beta_{U_1 M} U_1 - \beta_{U_2 M} U_2).$$

Data generation for Fig. 4(d) is also similar to Fig. 4(a) except the equation for Y is modified to include A as follows,

$$Y \mid U_1, U_2, M, A = \beta_{U_1 Y} U_1 + \beta_{U_2 Y} U_2 - \beta_{MY} M - \beta_{AY} A + \text{Normal}(0, 1).$$

For non-linear data generating processes, the equations for M are modified in all the graphs in Fig. 4 to add an interaction term between Z and A. Since we are using the dual weights estimated via machine learning techniques to perform the front-door test and compute the causal effect, it suffices to make this relation non-linear – non-linearity in other portions of the data generating process would not affect the computation of the dual weights due to variational independence.

Finally, for experiments comparing IV and front-door estimation, the graph used to generate data in a manner that satisfies the front-door but not IV assumptions is Fig. 4(a), and so the data generating mechanism remains the same. To generate data for a graph satisfying both assumptions, we use a subgraph where the $Z \to M$ is absent corresponding to dropping Z from the equation for M as follows: $p(M = 1 \mid A) = \text{expit}(-1 + \beta_{AM} A)$.

G PROOFS

Theorem 1:

Proof. Per Lemma 1 if $p(Z, A, M, Y)$ satisfies assumptions (A1-A3), the Verma constraint is non-trivial in the following sense: There is no ordinary independence constraint between Z and Y in the observed data distribution and the independence $Z \perp\!\!\!\perp Y$ arises only in the post-intervention distribution $p(Z, A, Y | \operatorname{do}(m))$. Under Verma faithfulness, distributions that satisfy this constraint must nested Markov factorize wrt an ADMG that permits identification of $p(Z, A, Y | \operatorname{do}(m))$, and where Z and Y are m-separated in the resulting CADMG obtained by removal of edges into M. Clearly Z and Y are not adjacent (via a directed or bidirected edge) in any such ADMG $\mathcal{G}$. We now show that the existence of any structure such that A has a bidirected path to one of its children would contradict the aforementioned conditions.

From results in Tian and Pearl [2002], we know that $p(Z, A, Y \mid \operatorname{do}(m))$ is identified from an observed data distribution nested Markov relative to $\mathcal{G}$ if and only if there is no bidirected path from M to Y in $\mathcal{G}$. Given the existence of $A \rightarrow M \rightarrow Y$ due to assumption (A1), the presence of $M \leftrightarrow Y$ in $\mathcal{G}$ then contradicts identifiability of $p(Z, A, Y | \operatorname{do}(m))$. From the discussion in Appendix A, the relevance assumption is not satisfied if only $A \leftrightarrow M$ exists in $\mathcal{G}$ (see Fig. 1(c) for an example.) Hence, $A \leftrightarrow M$ must be paired with either $M \leftrightarrow Y$, which we know contradicts identifiability, or $A \leftrightarrow Y$. When paired with the latter we get the bidirected path $M \leftrightarrow A \leftrightarrow Y$ which also contradicts identifiability of $p(Z, A, Y | \operatorname{do}(m))$. Hence, in ADMGs encoding the Verma constraint, the only bidirected edge permitted between A, M, and Y is $A \leftrightarrow Y$. In fact, $A \leftrightarrow Y$ must be present to satisfy assumption (A2). This is why the edge is compelled in all ADMGs in the pattern Fig. 3(a).

The presence of $A \rightarrow Y$ in $\mathcal{G}$ does not necessarily contradict identifiability of $p(Z, A, Y | \operatorname{do}(m))$. However, it contradicts m-separability between Z and Y in the resulting CADMG due to the presence of the chain $Z \rightarrow A \rightarrow Y$ which must be blocked by conditioning on A, but doing so opens a collider path $Z \circ\!\!\rightarrow A \leftrightarrow Y$ ($\circ\!\!\rightarrow$ depicts a directed edge or bidirected edge; at least one of these must be present in $\mathcal{G}$ due to the relevance assumption A2.) So far we have shown that if the Verma constraint holds, the corresponding ADMG $\mathcal{G}$ is one in which (i) Z and Y are not adjacent, (ii) the only bidirected edge present between the variables A, M, and Y is $A \leftrightarrow Y$, and (iii) $A \rightarrow Y$ is not present.

The only remaining possibility for the existence of a bidirected path from A to its child M then is if we have *both* $Z \leftrightarrow A$ and $Z \leftrightarrow M$ in $\mathcal{G}$. This too produces a contradiction, as these edges complete another bidirected path $M \leftrightarrow Z \leftrightarrow A \leftrightarrow Y$ resulting in non-identifiability of $p(Z, A, Y | \operatorname{do}(m))$. Hence, the distribution factorizes according to some ADMG $\mathcal{G}$ that has no bidirected path from A to any of its children (and the choice of valid ADMGs are given by the pattern in Fig. 3(a).)

□

Lemma 1:

Proof. In any valid ADMG derived from pattern 3(a), we know that A does not have any bidirected path to any of its children. Let D_A represent the district of A in any such ADMG. Per results in Tian and Pearl [2002], the post-intervention distribution $p(V \setminus A | \operatorname{do}(a))$ is identified as

$$p(V) \times \frac{\sum_A q_{D_A}(D_A \mid \operatorname{pa}_{\mathcal{G}}(D_A))}{q_{D_A}(D_A \mid \operatorname{pa}_{\mathcal{G}}(D_A))}.$$

In our case, this implies $p(Z, M, Y | \operatorname{do}(a)) = p(Z, A, M, Y)/\tilde{q}(A|Z, M, Y)|_{A=a}$, where $\tilde{q}(.) \equiv \frac{q_{D_A}(D_A|\operatorname{pa}_{\mathcal{G}}(D_A))}{\sum_A q_{D_A}(D_A|\operatorname{pa}_{\mathcal{G}}(D_A))}$. Hence, to show that the post-intervention distribution is identified by the same functional in any ADMG in Fig. 3(a), it suffices to show that $\frac{q_{D_A}(D_A|\operatorname{pa}_{\mathcal{G}}(D_A))}{\sum_A q_{D_A}(D_A|\operatorname{pa}_{\mathcal{G}}(D_A))}$ is the same in all such ADMGs. There are only two cases: the first where $D_A = \{Z, A, Y\}$ and second where $D_A = \{Z, Y\}$ (M cannot be in D_A by the pre-condition that A has no bidirected path to its children.) In the first case, $q_{D_A}(D_A \mid \operatorname{pa}_{\mathcal{G}}(D_A)) \equiv p(Z, A, Y \mid \operatorname{do}(m)) = p(Z) \times p(A \mid Z) \times p(Y \mid A, M, Z)$. Therefore,

$$\frac{q_{D_A}(D_A \mid \operatorname{pa}_{\mathcal{G}}(D_A))}{\sum_A q_{D_A}(D_A \mid \operatorname{pa}_{\mathcal{G}}(D))} = \frac{p(Z) \times p(A \mid Z) \times p(Y \mid A, M, Z)}{\sum_A p(Z) \times p(A \mid Z) \times p(Y \mid A, M, Z)} = \frac{p(A \mid Z) \times p(Y \mid A, M, Z)}{\sum_A p(A \mid Z) \times p(Y \mid A, M, Z)}.$$

In the second case when $D_A = \{A, Y\}$ we have already seen in Section 3 that $q_{D_A}(D_A \mid \operatorname{pa}_{\mathcal{G}}(D_A)) = p(A \mid Z) \times p(Y \mid A, M, Z)$. The result immediately follows that in both scenarios the conditional kernels are the same. That is, regardless of the specific edges incorporated from the pattern in Fig. 3(a), the functional for $p(Z, M, Y \mid \operatorname{do}(a))$ and hence $\mathbb{E}[Y \mid \operatorname{do}(a)]$ remains the same. □

References

Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1452–1482, 2014.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017. Working paper.

Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Nikolaj Thams, Sorawit Saengkyongam, Niklas Pfister, and Jonas Peters. Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*, 2021.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 567–573. American Association for Artificial Intelligence, 2002.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, 1990.