Approximating counterfactual bounds while fusing observational, biased and randomised data sources

Zaffalon, M., Antonucci, A., Cabañas, R., & Huber, D.

Approximating Counterfactual Bounds while Fusing Observational, Biased and Randomised Data Sources

Marco Zaffalon[a], Alessandro Antonucci[a,*], Rafael Cabañas[b], David Huber[a]

[a]*IDSIA, Lugano (Switzerland)*
[b]*Department of Mathematics, University of Almería, Almería (Spain)*

Abstract

We address the problem of integrating data from multiple, possibly biased, observational and interventional studies, to eventually compute counterfactuals in structural causal models. We start from the case of a single observational dataset affected by a selection bias. We show that the likelihood of the available data has no local maxima. This enables us to use the causal expectation-maximisation scheme to approximate the bounds for partially identifiable counterfactual queries, which are the focus of this paper. We then show how the same approach can address the general case of multiple datasets, no matter whether interventional or observational, biased or unbiased, by remapping it into the former one via graphical transformations. Systematic numerical experiments and a case study on palliative care show the effectiveness of our approach, while hinting at the benefits of fusing heterogeneous data sources to get informative outcomes in case of partial identifiability.

Keywords: Counterfactuals, randomised data, biased data, structural causal models, expectation maximisation.

1. Introduction

Consider the data in Table 1. They represent the results of an artificial case study from Mueller and Pearl (2023), where patients affected by a deadly disease test a drug that might help them. The table reports two independent

---

*Corresponding author

Email addresses: `marco.zaffalon@idsia.ch` (Marco Zaffalon),
`alessandro.antonucci@idsia.ch` (Alessandro Antonucci), `rcabnas@ual.es` (Rafael Cabañas), `david.huber@idsia.ch` (David Huber)

studies carried out on different groups of patients: a randomised control trial (first eight rows) and an observational study (last eight rows). From those data, Mueller and Pearl (2023) compute a typical counterfactual such as the *probability of necessity and sufficiency* (PNS), which is the proportion of people that would recover if treated and die otherwise. They do this analytically by conditioning on gender. For males they obtain that PNS $\in [0.28, 0.49]$ if only the randomised trial is considered, whereas the sharp value PNS $= 0.49$ is obtained when observational data are also taken into account. Analogous results are obtained for females.

Study	Treatment	Gender	Survival	Counts
interventional	do(drug)	female	survived	489
	do(drug)	female	dead	511
	do(drug)	male	survived	490
	do(drug)	male	dead	510
	do(no drug)	female	survived	210
	do(no drug)	female	dead	790
	do(no drug)	male	survived	210
	do(no drug)	male	dead	790
observational	drug	female	survived	378
	drug	female	dead	1022
	drug	male	survived	980
	drug	male	dead	420
	no drug	female	survived	420
	no drug	female	dead	180
	no drug	male	survived	420
	no drug	male	dead	180

Table 1: Data from interventional and observational studies on the potential effects of a drug on patients affected by a deadly disease.

This example illustrates the main traits of the problem we are going to address in this paper: first and foremost that the computation of counterfactuals is often only *partially identifiable*, in the sense that in general we can at best bound them, no matter the amount of data; but also that joining observational and randomised studies (under the guidance of a so-called *structural causal model*) can strengthen the results, i.e., narrow the bounds.

What the original example instead does not take into account is the possi-

ble presence of a *selection bias*, which is the systematic exclusion of a certain subpopulation from the sample. To account for this, we assume here that half of the data (the ones about treated males and untreated females, denoted in grey in the table) from the observational study is missing because of a problem with the communication protocol. The challenge is to produce correct counterfactual bounds for the overall population given the partial view induced by the biased data (in what follows, we call these data under selection bias just 'biased' for brevity).

In this paper we shall give general methods to solve problems as above: that is, the computation of counterfactuals with structural causal models where available data are a mix of observational and interventional (e.g., based on randomised studies) datasets, possibly subject to selection bias. We refer to these data as *heterogeneous*.

Such a setting essentially coincides with the general problem of information fusion as defined by Bareinboim and Pearl (2016). This, and nearly all other works in the literature, have however focused on the special case of identifiable queries, namely, those that can be reduced to probabilistic expressions. The only apparent exception is the recent work by Zhang et al. (2022) that showcases the application of MCMC techniques to solve a problem of information fusion without selection bias. Selection bias in itself has been studied in the causal literature for long (e.g., Cooper, 1995; Pearl, 2012; Bareinboim and Pearl, 2012; and Bareinboim and Tian, 2015) but the unidentifiable case is basically unexplored.

In contrast with the trend in the literature, the present work will entirely focus on the general case of partial identifiability. In particular, after reviewing the necessary background material about structural causal models and Bayesian networks in Section 2, we shall quickly review our EM-based scheme (we call it EMCC) for observational datasets in Section 3. The EMCC is an algorithm that takes a structural causal model in input, along with observational data, and allows us to approximately compute counterfactual inference in the form of a possible *range* of values included in the actual bounds. The EMCC will be extended to the case of biased datasets in Section 4 and to multiple studies is Section 5. We shall present in Section 6 a numerical validation on a synthetic benchmark and a case study on palliative care involving a biased observational dataset. Conclusions and outlooks will be reported in Section 7.[1]

---

[1] Our paper is an extended version of a paper presented at the PGM22 (Zaffalon et al., 2022).

2. Background on Bayesian Networks and Structural Causal Models

A generic variable X is assumed to take values from a set Ω_X. Note that we only consider variables taking values from finite sets. A probability mass function (PMF) over X is denoted as $P(X)$. Given variables Y and X, a conditional probability table (CPT) $P(Y|X)$ is a collection of PMFs over Y indexed by the values of X, i.e., $\{P(Y|X)\}_{x \in \Omega_X}$. If all PMFs in a CPT are *degenerate*, i.e., $\{0,1\}$-valued, we say that also the CPT is degenerate.

A *structural equation* (SE) f associated with variable Y and based on the input variable(s) X, is a surjective function $f : \Omega_X \to \Omega_Y$ that determines the value of Y from that of X. Such an SE induces the degenerate CPT $P(Y|X)$ via $P(y|x) := [\![f(x) = y]\!]$ for each $x \in \Omega_X$ and $y \in \Omega_Y$, where $[\![\cdot]\!]$ denotes the Iverson brackets that take value one if the statement inside the brackets is true and zero otherwise (it is just another way to denote events similarly to what one does via indicator functions).

Consider a joint variable $\boldsymbol{X} := (X_1, \ldots, X_n)$ and a directed acyclic graph $\mathcal{G}$ whose nodes are in a one-to-one correspondence with the variables in $\boldsymbol{X}$ (we use a node in $\mathcal{G}$ and its corresponding variable interchangeably). Given $\mathcal{G}$, a Bayesian network (BN) is a collection of CPTs $\{P(X_i|\text{Pa}_{X_i})\}_{i=1}^n$, where Pa_{X_i} denotes the *parents* of X_i, i.e., the direct predecessors of X_i according to $\mathcal{G}$. A BN induces a joint PMF $P(\boldsymbol{X})$ that factorises as $P(\boldsymbol{x}) = \prod_{i=1}^n P(x_i|\text{pa}_{X_i})$, for each $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$, where $(x_i, \text{pa}_{X_i}) \sim \boldsymbol{x}$, i.e., x_i and pa_{X_i} are the values of X_i and Pa_{X_i} consistent with $\boldsymbol{x}$ for each $i = 1, \ldots, n$.

Now consider two disjoint sets of variables $\boldsymbol{U}$ and $\boldsymbol{V}$, which we respectively refer to as *exogenous* and *endogenous*. A collection of SEs $\{f_V\}_{V \in \boldsymbol{V}}$ such that, for each $V \in \boldsymbol{V}$ the input variables of f_V are in $(\boldsymbol{U}, \boldsymbol{V})$, with at least one of them in $\boldsymbol{U}$, is called a *partially specified* structural causal model (PSCM). It corresponds to the notion of a 'functional causal model' in Pearl (2009, Chapter 1.4). A PSCM M induces the specification of a directed graph $\mathcal{G}$ with nodes in a one-to-one correspondence with the variables in $(\boldsymbol{U}, \boldsymbol{V})$ and such that there is an arc between two variables if and only if the first variable is an input variable for the SE of the second. The exogenous variables are therefore root nodes of $\mathcal{G}$. We focus on *semi-Markovian* PSCMs, i.e., those PSCMs that lead to acyclic graphs (Pearl,

2009). Moreover, if each exogenous variable has exactly one endogenous child, we call the PSCM *Markovian.*[2]

PSCMs assume SEs to be given. However, in the case where endogenous variables take on finitely many values, recent results allow one to specify a PSCM by only giving a causal graph; SEs are automatically defined, without loss of generality (yet at the cost of adding a likely excess of caution to the model), via a *canonical specification.* Let us introduce it in the simple case of Markovian models: we say that SE f_V is canonical if the states of the single (because of Markovianity) exogenous parent U of V index all the deterministic relations between the endogenous parents, denoted here as $\boldsymbol{W}_V$, and V. This requires:

$$|\Omega_U| = |\Omega_V|^{\prod_{W \in \boldsymbol{W}_V} |\Omega_W|}. \tag{1}$$

As an example, in a Markovian model, if V is Boolean and it has a single endogenous parent W, which is also Boolean, because of Equation (1), we need its exogenous parent U to have cardinality four in order to enumerate all the possible deterministic relations between the two Boolean variables W and V. E.g., we might set $f_V(W, U = 0) = 0$, $f_V(W, U = 1) = W$, $f_V(W = U = 2) = \neg W$, and $f_V(W = U = 3) = 1$. A Markovian PSCM whose SEs are all canonical is also called canonical. We refer the reader to the works of Duarte et al. (2023) and Zhang et al. (2022) for a generalisation of such a concept to non-Markovian models.

In a PSCM M we obtain a joint state of $\boldsymbol{V}$ from a (joint) state of $\boldsymbol{U}$ by applying the SEs of M consistently with a topological order for $\mathscr{G}$. A *fully specified* structural causal model (FSCM, see Pearl, 2009, top of p. 69) is just a PSCM M paired with a collection of marginal PMFs, one for each exogenous variable. As SEs induce (degenerate) CPTs, an FSCM defines a BN based on $\mathscr{G}$ whose joint

---

[2] In order to avoid confusion, let us recall that Pearl's original definition of a functional causal model (Pearl, 2009, Section 1.4.1) requires that there is a one-to-one correspondence between endogenous and exogenous variables; that is, each exogenous variable has exactly one endogenous child and vice versa. In such a case, assuming independence of the exogenous variables makes the model Markovian. In this paper, as it is common in much of the literature, we use an equivalent definition that keeps on requiring that endogenous variables have at least one exogenous parent, but it allows exogenous variables to have more than one endogenous child; at the same time it assumes that the exogenous variables are mutually independent. To have Markovianity in this setting, we need to require in addition that exogenous variables have exactly one endogenous child, thus going back to the one-to-one correspondence in the original definition of a functional causal model.

PMF factorises as:

$$P(\boldsymbol{u}, \boldsymbol{v}) = \prod_{U \in \boldsymbol{U}} \theta_u \cdot \prod_{V \in \boldsymbol{V}} P(v|\mathrm{pa}_V), \tag{2}$$

for each $\boldsymbol{u} \in \Omega_{\boldsymbol{U}}$ and $\boldsymbol{v} \in \Omega_{\boldsymbol{V}}$, $(u, v, \mathrm{pa}_V) \sim (\boldsymbol{u}, \boldsymbol{v})$ and where Pa_V are the parents of V according to $\mathcal{G}$ (i.e., the inputs of SE f_V), while θ_u denotes the true but unknown chances for $U = u$, to be considered for each $u \in \Omega_U$ and $U \in \boldsymbol{U}$. Notation $\theta_U := (\theta_u)_{u \in \Omega_U} \in \Delta_U$ is similarly used for the whole PMF, with Δ_U denoting the corresponding probability simplex. We extend this notation to the joint variable $\boldsymbol{U}$ via $\theta_{\boldsymbol{U}} := \times_{U \in \boldsymbol{U}} \theta_U \in \times_{U \in \boldsymbol{U}} \Delta_U =: \Delta_{\boldsymbol{U}}$. We shall instead use $P(u)$ for an estimate of θ_u and likewise for PMFs.

Given the graph $\mathcal{G}$ of a PSCM (or FSCM) M, obtain $\mathcal{G}'$ by removing from $\mathcal{G}$ any arc connecting pairs of endogenous variables. Let $\{\mathcal{G}_c\}_{c \in \mathscr{C}}$ denote the connected components of $\mathcal{G}'$. The *c-components* of M are the elements of the partition $\{\boldsymbol{V}^{(c)}\}_{c \in \mathscr{C}}$ of $\boldsymbol{V}$, where $\boldsymbol{V}^{(c)}$ denotes the endogenous nodes in $\mathcal{G}_c$, for each $c \in \mathscr{C}$ (Tian, 2002). This procedure also induces a partition of $\boldsymbol{U}$, similarly denoted as $\{\boldsymbol{U}^{(c)}\}_{c \in \mathscr{C}}$. Moreover, for each $c \in \mathscr{C}$, let $\boldsymbol{W}^{(c)}$ denote the union of the endogenous parents of the nodes in $\boldsymbol{V}^{(c)}$ and $\boldsymbol{V}^{(c)}$ itself. Finally, for each $V \in \boldsymbol{V}^{(c)}$, obtain $\boldsymbol{W}_V$ by removing from $\boldsymbol{W}^{(c)}$ the nodes topologically following V and V itself (note that in the notation we dropped the index c as this can be implicitly retrieved from V). As an example, Figure 2.a describes the identification of the two c-components of the model in Figure 1. In this case we have $\boldsymbol{V}^{(1)} = \{V_1, V_3\}$, $\boldsymbol{V}^{(2)} = \{V_2\}$, $\boldsymbol{U}^{(1)} = \{U_1\}$, $\boldsymbol{U}^{(2)} = \{U_2\}$, $\boldsymbol{W}_{V_1} = \{V_2\}$, $\boldsymbol{W}_{V_2} = \emptyset$, and $\boldsymbol{W}_{V_3} = \{V_1, V_2\}$.

Tian (2002) also shows that the joint PMF $P(\boldsymbol{V})$ obtained by marginalising the exogenous variables from the joint PMF in Equation (2) is a BN, to be called here *endogenous*, that factorises as follows:

$$P(\boldsymbol{v}) = \prod_{V \in \boldsymbol{V}} P(v|\boldsymbol{w}_V), \tag{3}$$

for each $\boldsymbol{v} \in \Omega_{\boldsymbol{V}}$, with $(v, \boldsymbol{w}_V) \sim \boldsymbol{v}$ for each $V \in \boldsymbol{V}$.

In FSCMs, the CPTs in the right-hand side of Equation (3) can be computed through standard BN inference algorithms by simply regarding the FSCM as a BN. With PSCMs, assuming the availability of a dataset $\mathscr{D}$ of endogenous observations, we might also define an endogenous BN with the same factorisation and whose CPTs are directly assessed from $\mathscr{D}$.

Observational queries in FSCMs can be addressed in the endogenous BN, and the same can be done for PSCMs assuming the availability of the dataset

$\mathscr{D}$ of endogenous observations. To perform causal inference, *interventions* denoted as $do(\cdot)$ should be considered instead. In an FSCM or PSCM M, given $V \in \mathbf{V}$ and $v \in \Omega_V$, $do(V = v)$ simulates a physical action on M forcing V to take the value v. The original SE f_V should be consequently replaced by a constant map $V = v$. Notation M_v is used for such a modified model, whose graph is obtained by removing from $\mathscr{G}$ the arcs entering V, and for which evidence $V = v$ is considered. This operation is often called *surgery*. In an FSCM M, given $V, W \in \mathbf{V}$ and $v \in \Omega_V$, $P(w|do(v))$ denotes the conditional probability of $W = w$ in the post-intervention model, i.e., $P'(w|v)$, where P' is the joint PMF induced by M_v.

A more general setup is provided by *counterfactual* queries, where the same variable may be observed as well as subject to intervention, albeit in distinct *worlds*. In mathematical parlance, if $\mathbf{W}$ are the queried variables, $\mathbf{V}'$ the observed ones and $\mathbf{V}''$ the intervened ones, we write the query by $P(\mathbf{W}_{\mathbf{v}''}|\mathbf{v}')$ with possibly $\mathbf{V}' \cap \mathbf{V}'' \neq \emptyset$. Popular examples of counterfactual queries involving two endogenous Boolean variables X and Y of an FSCM are the *probability of necessity* (PN), i.e., the probability that event Y would not have occurred by disabling X, given that X and Y did in fact occur. This corresponds to $P(Y_{X=0} = 0|X = 1, Y = 1)$. Similarly the probability of *sufficiency* (PS) $P(Y_{X=1} = 1|X = 0, Y = 0)$ and *necessity and sufficiency* (PNS) $P(Y_{X=1} = 1, Y_{X=0} = 0)$ are often considered.

Computing counterfactual queries in an FSCM may be achieved via an auxiliary structure called a *twin network* (Balke and Pearl, 1994). This is simply an FSCM where the original endogenous nodes (and their SEs) have been duplicated, while remaining affected by the same exogenous variables (see, e.g., Figure 2.b). More general and compact structures (e.g., involving more than two copies of the same endogenous node) can be also considered (Shpitser and Pearl, 2007). Computing a counterfactual in the twin network of an FSCM is analogous to what is done with interventional queries provided that interventions and observations are associated with distinct copies of the same variable. If the query involves multiple variables like in the case of the PNS, an auxiliary Boolean child of these variables, that is true if and only if its parent variables are in the queried state, can be added to the model. BN inference eventually allows one to compute the counterfactual query in such an augmented model.

3. Partial Identifiability (with a Single Unbiased Observational Dataset)

In this section we formalise the fundamental notion of partially identifiable query and review the basics of the EM-based scheme proposed by Zaffalon et al.
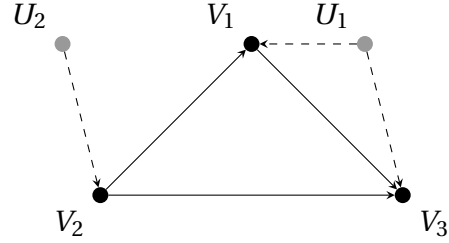
Figure 1: The graph of a structural causal model over three endogenous variables (black nodes) with two exogenous variables (grey nodes). The model is used to describe the data in Table 1 with V_1 corresponding to *Treatment*, V_2 to *Gender* and V_3 to *Survival*. The exogenous variable U_1 acts as a *confounder* for V_1 and V_3, this being sufficient to have results consistent with those of Mueller and Pearl (2023), which are obtained by conditioning on V_2.
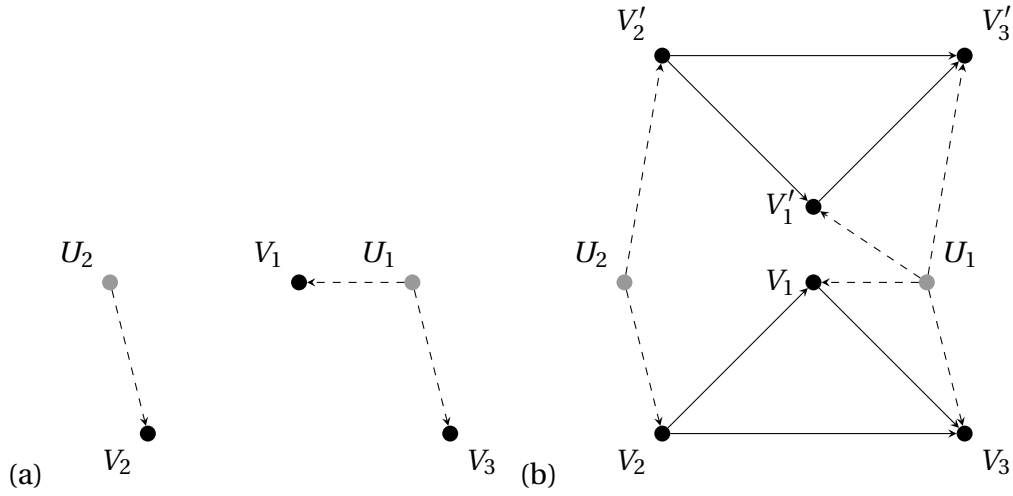

Figure 2: C-components (a) and twin network (b) of the model in Figure 1.

(2021) to address queries of this kind.[3] Such an EM procedure considers the case of a single, purely observational and unbiased, dataset. An extension to the case of (single and observational) biased datasets is in Section 4, while the fully general case of multiple, observational and interventional, unbiased or biased, datasets is addressed in Section 5.

As already noticed in the previous section, the general problem of computing counterfactuals can be reduced to standard BN inference in FSCMs. Yet exogenous variables are typically latent; their marginal PMFs being unavailable leaves us with PSCMs.

Say that we have a PSCM M over variables $(\boldsymbol{U}, \boldsymbol{V})$ together with a dataset $\mathscr{D}$ of endogenous observations. This permits quantifying the endogenous BN from which any observational query can be computed. The do-calculus (Pearl, 2009) instead permits reducing *identifiable* interventional queries on PSCMs to observational ones, which are possibly then computed in the endogenous BN. The remaining queries are called here *partially* identifiable.

To start approaching partially identifiable queries in the PSCM M from $\mathscr{D}$, we first compute the endogenous BN inducing the endogenous joint PMF $P(\boldsymbol{V})$. To be consistent with the endogenous PMF, the exogenous PMF $\theta_{\boldsymbol{U}}$ of an FSCM should satisfy:

$$\sum_{\boldsymbol{u} \in \Omega_{\boldsymbol{U}}} \left[\prod_{U \in \boldsymbol{U}} \theta_u \cdot \prod_{V \in \boldsymbol{V}} P(v|\mathrm{pa}_V) \right] = P(\boldsymbol{v}), \tag{4}$$

for each $\boldsymbol{v} \in \Omega_{\boldsymbol{V}}$, with $(u, v, \mathrm{pa}_V) \sim (\boldsymbol{u}, \boldsymbol{v})$ and $\theta_u \sim \theta_{\boldsymbol{U}}$. If such a $\theta_{\boldsymbol{U}}$ exists, we say that the endogenous BN is *M-compatible* with the original PSCM (Zaffalon et al., 2021). In general there are multiple FSCMs satisfying Equation (4), that is, there is a set $\Theta_{\boldsymbol{U}} \subseteq \Delta_{\boldsymbol{U}}$ of elements $\theta_{\boldsymbol{U}}$ that satisfies Equation (4). To compute a partially identifiable causal query in PSCMs, we should therefore consider all $\theta_{\boldsymbol{U}} \in \Theta_{\boldsymbol{U}}$ and compute the query for each of them. The result of the partially identifiable query is summarised by the lower and upper bounds obtained considering the outcomes of all the (M-)compatible FSCMs.

To practically pursue this goal, it is useful to consider the log-likelihood of $\mathscr{D}$ from a generic FSCM based on M, i.e.,

$$LL(\theta_{\boldsymbol{U}}) := \sum_{\boldsymbol{v} \in \mathscr{D}} \log \sum_{\boldsymbol{u} \in \Omega_{\boldsymbol{U}}} \left[\prod_{U \in \boldsymbol{U}} \theta_u \cdot \prod_{V \in \boldsymbol{V}} P(v|\mathrm{pa}_V) \right], \tag{5}$$

---

[3]An extended version of that conference paper is under review (Zaffalon et al., 2023).

with $(u, v, \mathrm{pa}_V) \sim (\boldsymbol{u}, \boldsymbol{v})$ and $\theta_u \sim \theta_{\boldsymbol{u}}$. Interestingly, this function has a single global maximum and achieving such a maximum is an equivalent condition to identify the (M-)compatible FSCMs, as stated by the following result.

Theorem 1 (Zaffalon et al., 2021, Theorem 1). *If $\Theta_{\boldsymbol{U}} \neq \emptyset$, the log-likelihood in Equation* (5) *has no local maxima and achieving its global maximum at $\theta_{\boldsymbol{U}}$ is an equivalent condition for $\theta_{\boldsymbol{U}} \in \Theta_{\boldsymbol{U}}$. If $\Theta_{\boldsymbol{U}} = \emptyset$, instead, the log-likelihood takes only values strictly smaller than the global maximum.*

The proof of the above result is in Zaffalon et al. (2021), and also in Zaffalon et al. (2023). For the sake of a self-contained presentation, we also report a proof in Appendix A.

The lack of local maxima allows us to employ an *expectation-maximisation* (EM) scheme, iterated with different random initialisations, to sample points from $\Theta_{\boldsymbol{U}}$. Given a PSCM M and a dataset of (unbiased) endogenous observations $\mathscr{D}$, Algorithm 1 yields an FSCM compatible with the PSCM (provided that there is one): such an FSCM is obtained by adding PMFs to the exogenous variables of the PSCM in such a way that the resulting model can generate the distribution of the available data.

In practice, given an initialisation $\{P_0(U)\}_{U \in \boldsymbol{U}}$, the EM algorithm consists in regarding the posterior probability $P_0(u|\boldsymbol{v})$ as a pseudo-count for $(u, \boldsymbol{v})$, for each $\boldsymbol{v} \in \mathscr{D}$, $u \in \Omega_U$ and $U \in \boldsymbol{U}$ (E-step). A new estimate $P_1(u) := \sum_{\boldsymbol{v} \in \mathscr{D}} \frac{P_0(u|\boldsymbol{v})}{|\mathscr{D}|}$ is consequently obtained (M-step). The scheme is iterated until convergence. Subroutine `initialisation` provides a random initialisation of the exogenous PMFs (line 1). The iterative procedure stops when the log-likelihood achieves a maximum (line 3).

Note that the numerical value of the global maximum of the log-likelihood can be proved to coincide with the log-likelihood of the dataset $\mathscr{D}$ from the endogenous BN whose parameters are quantified with the frequencies in $\mathscr{D}$ itself (Zaffalon et al., 2021). Such a value can be used to detect the two special cases in which the EMCC does not converge to the global maximum. The first is the convergence to a saddle point: this is an unstable point and a global maximum replacing that value can be easily obtained by a small perturbation in the initialisation. On the other side, if there is no convergence to the global maximum irrespectively of the initialisation, then the endogenous BN is not M-compatible. As discussed by Zaffalon et al. (2021), this is due to a wrong modelling or to insufficient data and makes inferences unreliable irrespectively of the procedure adopted to compute them.

Computing a causal query in the FSCM returned by a single EMCC run gives the exact value of the query in the identifiable case, which then corresponds to a point inside the bounds that characterise the partially identifiable case represented by the PSCM. An approximation of those bounds is eventually achieved by iterating the EM scheme above with variable initialisation values. This corresponds to sampling the space of FSCMs compatible with the given PSCM (the overall number of EM runs is denoted by r), and collecting the wanted counterfactual values across all these models. We call 'range' the approximation to the actual bounds that is delimited by the lower and upper values of the r outcomes above. We say that the range is an 'inner' approximation to stress that it lies between the actual lower and upper bounds. Approximations that instead encompass the actual bounds are called 'outer'.

Algorithm 1 EMCC: given SCM M and dataset $\mathscr{D}$ returns $\{P(U)\}_{U \in \boldsymbol{U}}$.

1: $\{P_0(U)\}_{U \in \boldsymbol{U}} \leftarrow \texttt{initialisation}(M)$
2: $t \leftarrow 0$
3: **while** $LL(\{P_{t+1}(U)\}_{U \in \boldsymbol{U}}) > LL(\{P_t(U)\}_{U \in \boldsymbol{U}})$ **do**
4:     **for** $U \in \boldsymbol{U}$ **do**
5:         $P_{t+1}(U) \leftarrow |\mathscr{D}|^{-1} \sum_{\boldsymbol{v} \in \mathscr{D}} P_t(U|\boldsymbol{v})$
6:         $t \leftarrow t+1$
7:     **end for**
8: **end while**

A demonstrative example of our EM procedure is reported here below. Here and in the other examples we consider *canonical* SEs (see Section 2). Nevertheless, all the methods we present in this paper can be applied to models with SEs of any kind, no matter whether canonical or not.

Example 1. *Consider the observational study associated with the last eight rows in Table 1 and a PSCM M based on the causal graph in Figure 1 with a canonical specification of the SEs. This requires $|\Omega_{U_1}| = 64$ and $|\Omega_{U_2}| = 2$. We are interested in a counterfactual query corresponding to the PNS for V_1 (i.e.,* Treatment*) as subject of the intervention and V_3 (i.e.,* Survival*) as queried variable. Each run of Algorithm 1 corresponds to an FSCM. The range of the corresponding PNS after $r = 300$ runs and a rounding to the second digit is given by* $[0.00, 0.43]$. *This is the same range we obtain by applying the mapping to credal networks of Zaffalon*

et al. (2020), and then running an exact credal network inference algorithm.[4] *In the other examples discussed in the rest of the paper we consider, with different setups, the same counterfactual query computed with the same number of runs and rounding strategy. The script to reproduce all these simulations is available together with the material released to reproduce the experiments (see Section 6).*

Other approaches have been proposed to approximate partially identifiable bounds (Zaffalon et al., 2020; Duarte et al., 2023; Zhang et al., 2022). Yet, these are focused on the case of unbiased data, and no viable technique to compute general counterfactual inference under selection bias seems to be available.[5] This is the focus of the next section, which shows how the EMCC of Zaffalon et al. (2021) has a natural extension to problems of selection bias.

4. Partial Identifiability under Selection Bias

To model the effect of selection bias, we define a *selector S*, i.e., a Boolean variable that is true for selected states of $\boldsymbol{V}$ and false otherwise. We consider deterministic selectors, i.e., $S := g(\boldsymbol{V})$ with $g : \Omega_{\boldsymbol{V}} \rightarrow \{0, 1\}$. E.g., for the observational data in Table 1, if we regard *drug* and *female* as the first states of the variables *Gender* and *Treatment*, the selector is true if and only if the two variables are in the same state. S can be embedded in a PSCM M as a common child of the endogenous variables, with g being the SE of S, thus acting as an additional endogenous variable with endogenous parents only. Notation M^S is used to denote such an augmented PSCM obtained from M.

---

[4]As discussed in Section 2, here and in the following, we compute PNS values in the twin network associated with the FSCM returned by each EMCC run.

[5]This is also the case of the formulae proposed by Tian and Pearl (2000) to bound PNS (but also PN and PS) queries by observational and interventional ones. Yet, at least in general, only outer approximations are returned by this approach. As a simple example, consider a PSCM over *Treatment* (V_1) and *Survival* (V_3), with the first endogenous variable being a parent of the second one and the common exogenous parent U_1. Setting $|\Omega_{U_1}| = 8$ allows to implement a canonical specification of the SEs of both V_1 and V_2. With the observational data in Table 1, such a canonical PSCM induce PNS bounds equal to $[0.00, 0.43]$. These values coincide with the bounds of Tian and Pearl (2000), which are therefore tight. Removing from Ω_U the state $U = \tilde{u}$, such that $f_{V_1}(\tilde{u}) = 1$ and $f_{V_3}(\tilde{u}, V_1) = V_1$ does not affect the bounds of Tian and Pearl (2000). Yet, the actual bounds of such a non-canonical PSCM become $[0.00.0.09]$. The code to reproduce this simple result can be found in the companion repository of the paper (see Section 6).

Selector S induces a partition among the records of $\mathscr{D}$. The endogenous values are assumed to become missing on the records such that $S = 0$, while remaining available if $S = 1$. Let us denote the corresponding datasets as $\mathscr{D}_{S=0}$ and $\mathscr{D}_{S=1}$. In terms of cardinalities, if $N_{S=0} := |\mathscr{D}_{S=0}|$ and $N_{S=1} := |\mathscr{D}_{S=1}|$, then $N_{S=0} + N_{S=1} = N =: |\mathscr{D}|$. Note that $\mathscr{D}_{S=0}$ contains $N_{S=0}$ identical records. Figure 3 depicts an example of the partition for the biased observational data of Table 1.

V_1	V_2	V_3	S	#
0	0	0	1	378
0	0	1	1	1,022
0	1	0	0	980
0	1	1	0	420
1	0	0	0	420
1	0	1	0	180
1	1	0	1	420
1	1	1	1	180

U_1	U_2	V_1	V_2	V_3	S	#
*	*	0	1	0	1	378
*	*	0	1	1	1	1,022
*	*	1	0	0	1	420
*	*	1	0	1	1	180

U_1	U_2	V_1	V_2	V_3	S	#
*	*	*	*	*	0	2,000

Figure 3: Selecting the data of the observational study in Table 1.

To address partially identifiable queries under selection bias, let us first reformulate the notion of M-compatibility for M^S. On the selected states the compatibility condition is like in the unbiased case, and we consequently require Equation (4) to be satisfied for each $\boldsymbol{v}$ such that $g(\boldsymbol{v}) = 1$, where the joint endogenous probability on the right-hand side is computed by an endogenous BN whose CPTs are obtained from the frequencies in $\mathscr{D}_{S=1}$. For the unselected states, as the only endogenous observation is $S = 0$, we require instead:

$$\sum_{\boldsymbol{v}:g(\boldsymbol{v})=0} \sum_{\boldsymbol{u}\in\Omega_{\boldsymbol{U}}} \left[\prod_{U\in\boldsymbol{U}} \theta_u \cdot \prod_{V\in\boldsymbol{V}} [\![f_V(\mathrm{pa}_V) - v]\!] \right] = P(S = 0), \qquad (6)$$

where, like for the endogenous joint probabilities, also the marginal probabilities of the selector are obtained from the (cardinalities of the) data, e.g., $P(S = 0) \simeq N_{S=0}/N$. A discussion on how to proceed if these cardinalities or probabilities are not available is provided at the end of this section.

The task of detecting the set $\Theta_{\boldsymbol{U}}$ of elements $\theta_{\boldsymbol{U}}$ satisfying the above M-compatibility conditions can be reduced to a log-likelihood maximisation as in the unbiased case. The log-likelihood of $\mathscr{D}_S := \mathscr{D}_{S=0} \cup \mathscr{D}_{S=1}$ from an FSCM

based on M^S can be written as:

$$LL(\theta_{\boldsymbol{U}}) := LL_{S=0}(\theta_{\boldsymbol{U}}) + LL_{S=1}(\theta_{\boldsymbol{U}}), \tag{7}$$

with:

$$LL_{S=0}(\theta_{\boldsymbol{U}}) \;\; := \;\; N_{S=0} \log \sum_{\boldsymbol{v}:g(\boldsymbol{v})=0} \sum_{\boldsymbol{u} \in \Omega_{\boldsymbol{U}}} \left[\prod_{U \in \boldsymbol{U}} \theta_u \cdot \prod_{V \in \boldsymbol{V}} [\![f_V(\mathrm{pa}_V) - v]\!] \right], \tag{8}$$

$$LL_{S=1}(\theta_{\boldsymbol{U}}) \;\; := \;\; \sum_{\boldsymbol{v} \in \mathscr{D}_1} \log \left[\prod_{U \in \boldsymbol{U}} \theta_u \cdot \prod_{V \in \boldsymbol{V}} [\![f(\mathrm{pa}_V) - v]\!] \right]. \tag{9}$$

The following result, whose proof is in Appendix A, provides a generalisation of Theorem 1 to the case of biased data.

Theorem 2. *For $\Theta_{\boldsymbol{U}} \neq \emptyset$, the log-likelihood in Equation (7) has no local maxima and achieving its global maximum at $\theta_{\boldsymbol{U}}$ is an equivalent condition for $\theta_{\boldsymbol{U}} \in \Theta_{\boldsymbol{U}}$. For $\Theta_{\boldsymbol{U}} = \emptyset$, instead, the log-likelihood takes only values strictly smaller than the global maximum.*

As a consequence of Theorem 2, we can apply Algorithm 1 to PSCM M^S and dataset $\mathscr{D}_S$ to obtain FSCMs whose log-likelihood takes its global maximum and hence satisfies the compatibility constraints. Causal queries computed with those FSCMs are consequently inside the bounds for the partially identifiable query. The more points the better the approximation.

Note that the data in $\mathscr{D}_{S=0}$ are $N_{S=0}$ instances of the same observation $S = 0$. Thus, when coping with biased data, line 5 of Algorithm 1 rewrites as:

$$P_{t+1}(U) \leftarrow \frac{N_{S=0} P_t(U|S=0) + \sum_{\boldsymbol{v} \in \mathscr{D}_1} P_t(U|\boldsymbol{v})}{N_{S=0} + N_{S=1}}. \tag{10}$$

Equation (10) shows the divergent effect of two datasets $\mathscr{D}_{S=0}$ and $\mathscr{D}_{S=1}$ on the exogenous chances. While the selected states push the chances towards their unbiased values, the unselected ones act as a noise term only forcing the chances to be zero on the exogenous states generating the selected endogenous states. Overall, compared to the unbiased case, we expect this weighted average to induce wider bounds on the counterfactual queries. An example is reported here below.

Example 2. *Consider the same setup as in Example 1. Here we compute the PNS for the observational data in Table 1 with a selector preventing the data*

14

associated with the grey rows from being available. These are half of the data, which means $P(S = 1) = 0.5$. With the two datasets induced by such a selector (see Figure 3), the modified version of Algorithm 1 for biased data returns the range $[0.00, 0.73]$. Unlike the unbiased case considered by Example 1, a comparison with the exact method cannot be performed in this case. Figure 4 depicts the PNS range (as determined by the points) for different selection mechanisms corresponding to different values of $P(S = 1)$. The range size for the unbiased case increases for decreasing values of $P(S = 1)$, this reflecting the fact that we focus on selection mechanisms obtained by removing records incrementally.
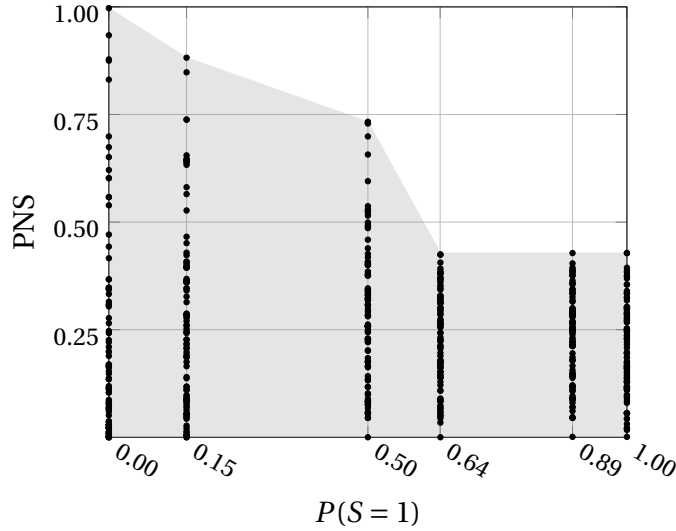

Figure 4: PNS ranges (grey) for different selection levels (x-axis) induced by $r = 300$ EMCC runs (black points denote the result of each run).

Note that in the limit $P(S = 1) \to 0$, we are basically ignoring the data in $\mathscr{D}_{S=1}$. If this is the case, the log-likelihood is proportional to $N_{S=0}$ and this implies that EMCC gives the same results irrespectively of the size of the (completely missing) dataset. In those cases our EMCC trivially returns the exogenous PMFs sampled for initialisation. This explains, for instance, the vacuous result obtained in this limit in the left side of Figure 4.

Let us also remark, with reference to Equation (10), that the execution of EMCC under selection bias requires the cardinality $N_{S=0}$ of $\mathscr{D}_{S=0}$. This corresponds to the common case of an observational process making unavailable the output of the measure, but not the fact that the measurement (attempt) was

performed. We expect therefore that an investigator will have in many cases an idea of the size of the neglected part of the population; or that they will be able to estimate $N_{S=0}$ from $P(S=0)$ as $N_{S=0}/N_{S=1} \simeq P(S=0)/P(S=1)$; or they could estimate an upper bound for $P(S=0)$. These estimates can follow even simply by first estimating the sample size before the selection; time and resources constraints or domain knowledge are usually enough to bound such a size. In the worst possible case, one could use the (very) conservative approach of taking the limit $P(S=0) \to 1$. This is not recommended because it corresponds to a very extreme measure (e.g., like in Figure 4).

Finally notice that as in principle S might be a common child of all the variables in $\boldsymbol{V}$, implementing SE g as a CPT might lead to an exponential blow-up, this making the inference of $P(U|S=0)$ required by EMCC intractable. In practice, as in Bareinboim and Pearl (2012) or in our example, the selector might only depend on a subset of $\boldsymbol{V}$ of bounded cardinality. Circuital approach like the one recently proposed by Darwiche (2022) might be considered to bypass such a limit.

5. Coping with Multiple Datasets

In the previous sections we addressed the problem of partial identifiability in causal queries when coping with a purely observational dataset, no matter whether unbiased (Section 3) or biased (Section 4). The challenge of this section is to extend our approach to the case of multiple datasets in a fully general setup mixing observational and interventional data. We initially assume the data unbiased just for the sake of presentation. The extension to the biased case is discussed in the last part of the section.

Say that our domain of interest is described by a PSCM M over $(\boldsymbol{U}, \boldsymbol{V})$. Endogenous observations from *d independent*, observational or interventional, studies are available. Denote as $\mathscr{D}^k$ the dataset of size $N^k := |\mathscr{D}^k|$ with the results of the k-th study, for each $k \in \{1, \dots, d\}$. The endogenous variables $\boldsymbol{V}_I^k \subseteq \boldsymbol{V}$ are those subject to interventions in $\mathscr{D}^k$, while for observational studies we simply set $\boldsymbol{V}_I^k := \emptyset$. Without any lack of generality, we assume that the intervened variables are always the same across the records of a same dataset. If this is not the case, we just split the dataset into smaller groups homogeneous w.r.t. the set of intervened variables. The states induced by the interventions inside each dataset might instead be different (e.g., a clinical study where we intervene by giving or not giving the drug to a patient). For $\mathscr{D}^k$, we denote these states as

16

$\tilde{\Omega}_{V_I^k} \subseteq \Omega_{V_I^k}$, while $\tilde{\Omega}_{V_I^k} := \emptyset$ for an observational $\mathscr{D}^k$. We further clarify such a notation by means of the following example.

Example 3. *Let $\mathscr{D}^1$ denote the interventional dataset in Table 1 and $\mathscr{D}^2$ the (unbiased) observational one. We have V_I^1 = {Treatment}, $\tilde{\Omega}_{V_I^1}$ = {drug, no drug}, $V_I^2 = \emptyset$ and $\tilde{\Omega}_{V_I^2} = \emptyset$.*

We merge the d datasets into a single one, say $\mathscr{D} := \cup_{k=1}^d \mathscr{D}^k$, with cardinality $N := \sum_{k=1}^d N^k$. We introduce a new variable W to index all the interventional states in $\mathscr{D}$. An additional state of W, denoted as $w_\emptyset$, describes the case of no intervention. Overall we have $\Omega_W := \cup_{k=1}^d \tilde{\Omega}_{V_I^k} \cup \{w_\emptyset\}$. We consequently augment $\mathscr{D}$ with a column of observations of W. The values of W for records originally belonging to the interventional datasets are those corresponding to the intervened states, while the state $w_\emptyset$ is assigned to the records from the observational datasets. We eventually denote as $\mathscr{D}'$ the dataset of N (complete) observations of (V, W) obtained in this way. Let us demonstrate the above merging procedure by means of an example.

Example 4. *Consider the setup in Example 3 with $\mathscr{D} := \mathscr{D}^1 \cup \mathscr{D}^2$. The index variable W is such that $\Omega_W := \{\text{drug}, \text{no drug}, w_\emptyset\}$. The corresponding augmented dataset $\mathscr{D}'$ is in Table 2.*

Given M and $\mathscr{D}'$, let us build a, so-called *auxiliary*, PSCM M' as follows. Add variable W to M as an endogenous auxiliary variable corresponding to an additional parent of all the endogenous variables in V. The SE f'_V associated with V in M' is defined as:

$$f'_V(\text{pa}_V, W = w) := \begin{cases} v_w & \text{if} \quad V \in V_I^{k(w)}, \\ f_V(\text{pa}_V) & \text{otherwise}, \end{cases} \tag{11}$$

for each $\text{pa}_V \in \Omega_{\text{Pa}_V}$, $w \in \Omega_W$, and $V \in V$, where f_V is the SE of V in M, and, $k(w)$ is the index of the interventional dataset associated with w, for each $w \neq w_\emptyset$, and $v_w \in \Omega_V$ is the state of V appearing in w. In practice, Equation (11) implements the surgery on V required by an intervention if W takes a value corresponding to an intervention involving also V, while leaving the original SE of M otherwise. For a proper PSCM specification, as our definition requires each endogenous variable to have at least an exogenous parent, we finally add to M' an exogenous variable U_W defined as a unique parent of W, with $\Omega_{U_W} := \Omega_W$ and the identical map as SE for W. As an example, Figure 5 depicts the graph of

17

Treatment	Gender	Survival	W	Counts
drug	female	survived	drug	489
drug	female	dead	drug	511
drug	male	survived	drug	490
drug	male	dead	drug	510
no drug	female	survived	no drug	210
no drug	female	dead	no drug	790
no drug	male	survived	no drug	210
no drug	male	dead	no drug	790
drug	female	survived	$w_\emptyset$	378
drug	female	dead	$w_\emptyset$	1022
drug	male	survived	$w_\emptyset$	980
drug	male	dead	$w_\emptyset$	420
no drug	female	survived	$w_\emptyset$	420
no drug	female	dead	$w_\emptyset$	180
no drug	male	survived	$w_\emptyset$	420
no drug	male	dead	$w_\emptyset$	180

Table 2: A merged version of the two datasets in Table 1 with the index variable W.
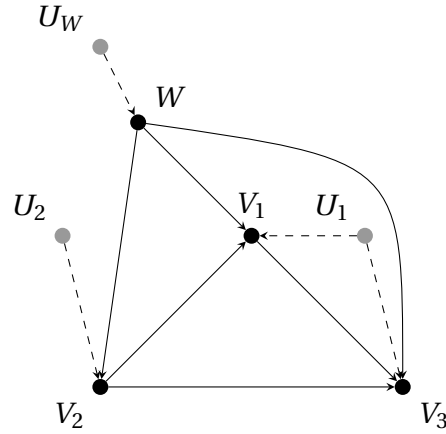

Figure 5: The auxiliary model M' obtained from the PSCM M in Example 1 and the merged dataset in Table 2.

the auxiliary PSCM M' obtained from the PSCM M discussed in Example 1 and the dataset $\mathscr{D}'$ in Table 2.

The auxiliary PSCM M' embeds all the possible interventions included in the d datasets. Thanks to the index variable W, which keeps track of the different interventions, we can regard $\mathscr{D}'$ as a purely observational dataset for the endogenous variables of M'. By Theorem 1, we can therefore apply the EMCC scheme to address non-identifiable queries as in Section 3 even when coping with the generalised setup discussed in this section.

In particular, the global maximum of the log-likelihood achieved by the EM scheme should correspond to the log-likelihood assigned to the dataset by an endogenous BN whose parameters have been trained from the dataset itself. As W is a common parent of all the endogenous variables, the endogenous graph of M' is just the endogenous graph of M augmented by W, which acts as a common parent of all the other endogenous variables. As a first example of a partially identifiable query based on heterogeneous data let us consider the following example, where the data integration induce more informative ranges.

Example 5. *The PNS query for the (unbiased) observational data in Table 1 yields range* $[0.00, 0.43]$. *By also considering the interventional data, the EMCC applied to the model in Figure 5 and the dataset in Table 2 returns instead the range* $[0.32, 0.42]$.

So far we only considered the case of unbiased datasets. In the presence of a selection bias, even if only on some datasets, we should first add the selector variable S to all the d datasets, with its value being constantly true on the unbiased ones. The construction of the auxiliary PSCM M' from M is analogous, provided that the auxiliary variable W becomes also a parent of S, with the SE of S in M' implementing the different selection functions depending on the particular value of W. This requires S to be a common child of all the endogenous variables. A more compact approach consists in considering only the union of the endogenous variables involved in the different biases for the different datasets. Other approaches involving auxiliary variables might be also considered. Note also that we might need different states of W to model the same intervention (or lack of intervention) in different datasets. We do not explicitly formalise this point just for the sake of light notation. Once this is done, Algorithm 1 can be executed on M' with $\mathscr{D}'$ as in the unbiased case. An example is reported here below.

Example 6. *In the same setup of Example 5 consider also the selection bias preventing the grey rows of the observational study in Table 1 from being available. To compute the PNS in this case we add the Boolean selector S to all the records and obtain a merged dataset $\mathscr{D}'$ as in Table 2 with an additional column associated with S, which is always one apart from the four grey rows where it is zero. Figure 6 depicts the corresponding auxiliary model M'. Given $\mathscr{D}'$ and M', the EMCC returns the range $[0.27, 0.53]$ for PNS (with the observational data only we had $[0.00, 0.73]$).*
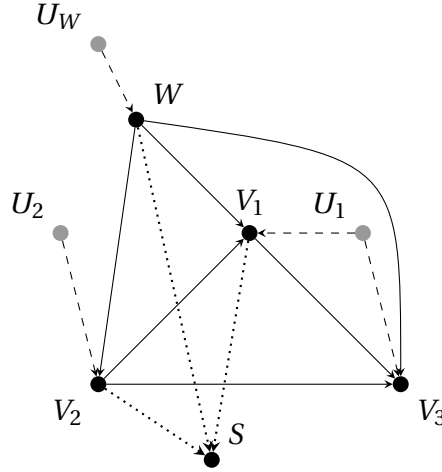

Figure 6: The auxiliary model M' obtained from the PSCM M in Example 1 and the merged dataset in Table 2 in the presence of a selection bias.

As a final remark, notice that so far we implicitly assumed the different datasets to be generated by the same exogenous chances. Yet, this might not always be the case: consider for instance two clinical studies in different countries where the body mass index has very different distributions over the population. Different chances for the different datasets can be simply introduced to cope with such an extended setup. To model this in the auxiliary PSCM M'', we should set the index variable W to be also a parent (or the parent of an auxiliary parent) of the exogenous variables having different chances. Although this is not respecting the standard definition of PSCMs (cf. Section 2), W is always observed in $\mathscr{D}'$ and therefore acts as a selector for the exogenous PMF to be updated by the EMCC. Such an index should be also specified in the query of interest. The above setup is demonstrated by means of the following example.

Example 7. *In the same setup of Example 6 assume that the two studies in Table 1 refer to two different exogenous chances associated with the endogenous variable V_2 (i.e.,* Gender). *Let us modify the auxiliary PSCM M' in Figure 6 in order to take into account this difference. An arc directly connecting W and U_2 would not work as the (three) states of W distinguish between the two interventional states in $\mathscr{D}^1$, which in our assumptions refer to the same exogenous chance. This can be solved by a coarsening variable $W' := i(W)$, specified as a child of W and returning the index of the dataset we refer to, i.e., $i(W = w_\emptyset) = 1$ and $i(W = \mathrm{drug}) = i(W = \mathrm{no\,drug}) = 2$. Variable W' is eventually specified as a parent of U_2 with $P(U_2|W' = 1)$ and $P(U_2|W' = 2)$ modelling the expectations of the exogenous chances in the two datasets. The PNS range we obtain in this case is $[0.20, 0.54]$ for the chances associated with the observational dataset, this corresponding to a looser range than $[0.27, 0.53]$, which we obtained in Example 6 by forcing the chances to be equal.*

6. Empirical Validation

To evaluate the potential of our approach to the computation of partially identifiable queries with heterogeneous data sources, we report and discuss the results of extensive tests based on a benchmark of synthetic data and causal models (Section 6.1), and a real-world application of counterfactual analysis with biased data (Section 6.2).

Algorithm 1 was already implemented within the CREDICI library, a Java tool for causal inference (Cabañas et al., 2020).[6] We enhance the library with an implementation of the procedures presented in Sections 4 and 5, thus allowing to compute counterfactuals with heterogeneous data. To the best of our knowledge this is the first tool for causal inference in such a general setting. The code to reproduce the experiments is available in a dedicated repository.[7]

6.1. Synthetic Data

We use the Erdös-Rényi scheme to sample directed acyclic graphs. If the sampled graph is such that there are non-root nodes without root parents, we add to the graph a new root parent node for each one of these nodes. Parentless nodes are then regarded as exogenous variables, and the other ones as Boolean endogenous ones. SEs and exogenous cardinalities are obtained by sampling

---

[6] github.com/idsia/credici.
[7] github.com/IDSIA-papers/2023-IJAR-bias-hybrid.

(without replacement) the deterministic relations between each endogenous variable and its endogenous parents, letting the states of the exogenous parents index the relations with $|\Omega_U| \leq 64$ for each $U \in \boldsymbol{U}$. Note that this typically induces a non-canonical specification of the corresponding SE.

From such a PSCM M we sample the *ground-truth* chances, thus obtaining FSCM M^*. Overall we generate 220 models with $|\boldsymbol{V}|$ ranging from 5 to 17.

For each model we select three endogenous variables to be called, respectively, input, target, and covariate. Following a topological order, we set as input the first variable with an exogenous confounder. The target is a leaf such that there is a directed (endogenous) path connecting the input to the target, and the covariate is a variable belonging to that path. We sample from M^* three datasets of endogenous observations, namely: (i) an observational dataset $\mathscr{D}_O$; (ii) an interventional dataset $\mathscr{D}_I$, with interventions on the input and an equal number of positive and negative values for the intervened variable; (iii) another interventional dataset $\mathscr{D}_{IB}$ with interventions on the covariate and a selector based on the values of input, covariate and target. To avoid very strong or very weak biases, the selectors are designed in order to have $0.25 \leq P(S = 1) \leq 0.75$. For each model, the three datasets have the same size, with values ranging from $1{,}000$ to $5{,}500$ over the different models in order to ensure M-compatibility. Note that for the biased dataset, the size is intended to be the one before the selection process.

We evaluate the PNS for the input and the target from the PSCM M. This is done by adding the different datasets following two different orderings. In the first we start from $\mathscr{D}_{IB}$, then we add $\mathscr{D}_O$, and finally also $\mathscr{D}_I$. In the second we start from $\mathscr{D}_I$, then we add $\mathscr{D}_O$, and finally also $\mathscr{D}_{IB}$. In both cases the EMCC is executed with $r = 300$ runs. Such a sequential approach is considered here only to evaluate the impact of the different datasets. In a real scenario all the available data would be used from the start.

Before commenting on the results, we use the credibility intervals derived by Zaffalon et al. (2022) to evaluate the quality of the inner approximation given by the EMCC-delivered range with the selected number of runs. We compute the probability of a relative error (at each extreme of the range) smaller than $\epsilon = 0.1$ in our simulations and obtain a median confidence equal to 0.999 and a lower quartile equal to 0.921. Those high values support the accuracy of our inference scheme, and allow to reasonably approximate the ground-truth intervals with the EMCC values.

Concerning the results with the heterogeneous data, we notice that adding a new dataset typically induces a shrink in the interval. To summarise the

results of our experiments we therefore compute the *relative shrink* defined as one minus the ratio between the size of the PNS range with the additional dataset(s) and the PNS range with the initial dataset. Figure 7 depicts the corresponding boxplots. On the left we consider the model with the same chances shared over the different datasets (called *global*), while for the *local* models on the right we let the chances associated with the covariate to be different for each dataset, and we run the PNS query with the chances based on $\mathscr{D}_{IB}$ for the first ordering, and $\mathscr{D}_I$ for second.
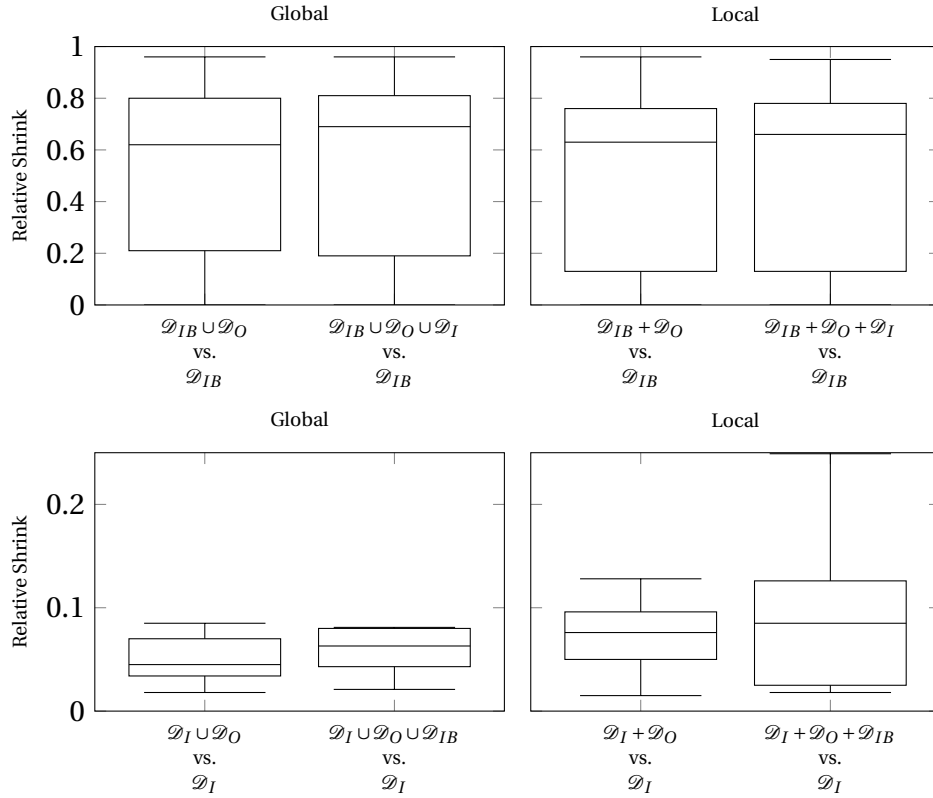

Figure 7: Relative shrinks induced by integration of additional datasets.

When starting with biased data (plots on the top of Figure 7) we observe noticeable shrinks w.r.t. the initial PNS ranges, these being above 60% for both global and local models. This advocates the benefits of developing tools for counterfactual inference with heterogeneous data. Such a shrink is achieved by adding a single dataset to the initial one. The effect of adding the second

dataset is considerably weaker (about 5% for both class of models), but still noticeable. Low shrinks are also obtained when we start with the unbiased interventional data (plots on the bottom). This supports the intuition that the unbiased interventional data are the most informative for causal analysis. Yet, we still observe a non-negligible advantage in adding also observational or biased data.

Finally let us also report a dedicated analysis of the effect of the missingness induced by a selection bias with a single observational dataset as in Section 4. We consider different selectors with no restrictions on the values of $P(S = 1)$. For each dataset we compare the PNS ranges with the narrower ones obtained by removing the bias. The quality of the EMCC ranges as based on the credible intervals is analogous to the one observed with the heterogeneous data. As a descriptor we use the difference between the two lower values and we normalise it with the difference obtained when the biased range is $[0, 1]$. We call this descriptor *normalised bias effect*. We similarly proceed for the upper values. The aggregated results are grouped w.r.t. a discretisation of $P(S = 1)$ and the corresponding boxplots displayed in Figure 8. As expected with less missing records the ranges become more informative and closer to their unbiased values.
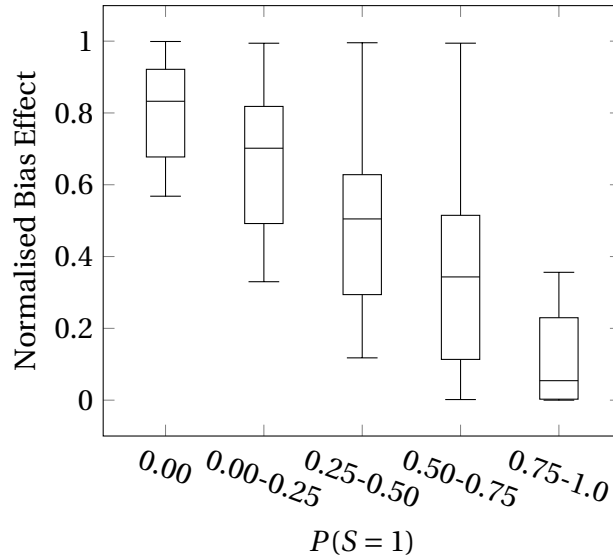

Figure 8: Difference between the ranges of biased and unbiased data w.r.t. $P(S = 1)$.

6.2. A Counterfactual Analysis in Palliative Care with Biased Data

Figure 9 represents the causal graph used for a study on the preferences of terminally ill cancer patients regarding the place they want to spend their final moments: home or hospital. Although a majority of patients prefer to pass away at home, most of them end up dying in institutional settings. The study focuses on exploring the interventions that healthcare professionals can take to increase the chance of patients dying at home. All the variables in the graph are intended to be endogenous and the graph corresponds to the BN proposed by Kern et al. (2020) reduced to the subset of variables for which data were available—variables have been binarised too. Due to ethical reasons, the original data of the study cannot be used and we sampled instead a dataset $\mathscr{D}$ of $1,000$ observations.
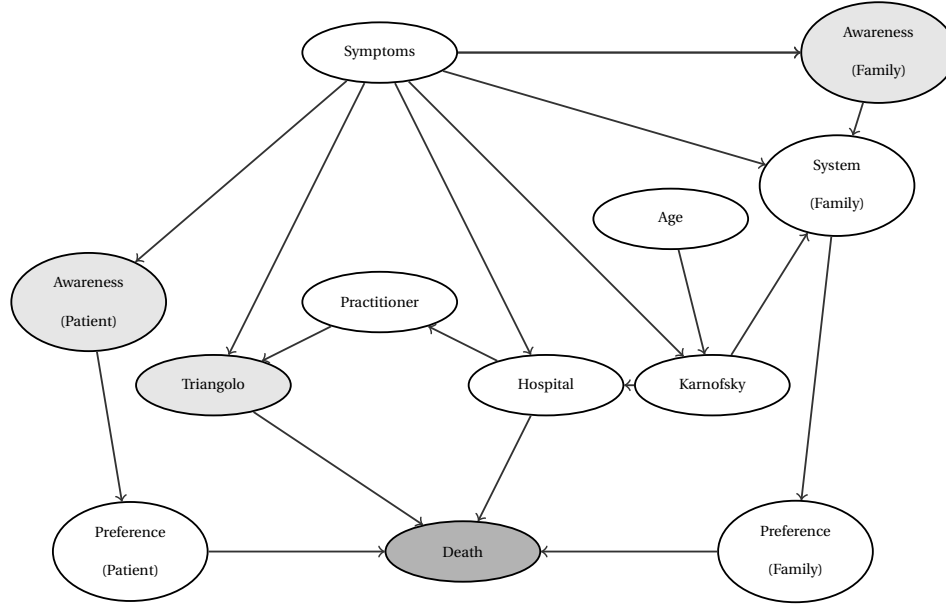

Figure 9: The graph of the model used to study preferences about the place of death in cancer patients.

One can intervene on three variables in the network (light grey nodes in Figure 9): the patient's and the family's *awareness* of death (which involves communication with the doctors); and home assistance (provided by the *Triangolo* association). The model is taken without exogenous confounders as a consequence of the fact that all the potential confounders have been explicitly represented in the causal graph. In practice we consider a *Markovian* model

where each endogenous variable has a separate and unique exogenous parent, that is not reported in Figure 9 just for the sake of readability. As for SEs, we stick to the *canonical* (see Section 2) representation since we want to be least committal w.r.t. the true underlying mechanisms. Such a conservative modelling assumption can induce high cardinality for the exogenous variables. In particular because of Equation (1), the exogenous parent U of variable *Death* has cardinality $|\Omega_U| = 2^{16}$. Yet, this does not prevent EMCC from running and, in order to evaluate the relative importance of those three variables, we compute the corresponding PNS ranges having those variables as input and *Death* as target node. With the unbiased dataset we obtain:

$$
\begin{aligned}
\text{Triangolo} &\rightarrow [0.27, 0.35], \\
\text{Patient Awareness} &\rightarrow [0.03, 0.11], \\
\text{Family Awareness} &\rightarrow [0.04, 0.11].
\end{aligned}
$$

The lower value associated to Triangolo being greater than the upper values for the two other options, shows that home assistance dominates them and hence it is the key factor determining the death place.

The above result has been presented by Zaffalon et al. (2023). Here we investigate whether or not such an evidence in favour of home assistance holds even in the presence of a selection bias. In particular we want to model the fact that studies of this kind are typically biased towards patients progressed to severe conditions. We consequently consider selectors based on *Karnofsky* (that is an index to track disease progressions) and *Symptoms*. For both these binary variables we consider the *bad-condition* states, i.e., low Karnofksy index and serious symptoms. We consequently consider a weak bias removing only the records of patients such that both variables are in the bad-condition state, and a strong bias such that even a single variable in a bad-condition state makes the datum unavailable. The results are:

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad\text{weak / strong} \\
\text{Triangolo} &\rightarrow [0.29, 0.35] \,/\, [0.22, 0.43], \\
\text{Patient Awareness} &\rightarrow [0.02, 0.11] \,/\, [0.02, 0.19], \\
\text{Family Awareness} &\rightarrow [0.04, 0.12] \,/\, [0.00, 0.19].
\end{aligned}
$$

In practice even in the presence of a selection bias, no matter whether weak or strong, it is possible to draw informative conclusions from data and keep advocating the importance of home assistance.

26

Let us finally note that the fact that the lower value of the PNS range for Triangolo with a weak bias is slightly higher than the corresponding lower value in the unbiased case reflects the fact that, with fewer complete data, the range provided by the EMCC with the current number of runs might be stronger than the one obtained with the complete data.

7. Conclusions and Outlook

We have presented an EMCC algorithm that learns the parameters of a given partially specified structural causal model from a mix of observational and interventional, as well as possibly biased, data. This setting is essentially that of information fusion put forward by Bareinboim and Pearl (2016). In this setting, we appear to be the first to deliver an algorithm with this type of generality and to make the code public.

A few remarks may be useful to clarify some main aspects of our approach:

- Our algorithm aims at reconstructing the uncertainty about the latent variables. It does this in an approximate way, by delivering a set of fully specified SCMs contained in all those that are compatible with the partially specified one. Once this is done, one can compute any counterfactual by iterating its computation over the FSCMs above and aggregating their results so as to yield a range of values for unidentifiable queries that approximates the actual bounds from inside.

- Empirical results support the quality of the approximation and confirm the increased informativeness of ranges obtained by merging multiple studies. Note that more efficient, or accurate, computation would be possible in principle if the aim was to compute a specific counterfactual from the beginning, because searching the space of compatible FSCMs could be tailored to such a goal; but this is not the purpose of our work, which aims at remaining agnostic w.r.t. the follow-up computations, and hence general.

- The PSCM is an input for our algorithm. We stress that the requirement that SEs are given (they are part of the PSCM) is not as stringent as it might seem, given that we can produce the needed SEs by a preprocessing step if the actual ones are not available: this is possible thanks to recent work (Duarte et al., 2023; Zhang et al., 2022) that has introduced

27

a *canonical* specification of the SEs. It can be understood as a least-committal specification that, loosely speaking, can be used without loss of generality (the implication however is that the output will tend to be less informative than the case where the actual SEs are given). In this sense, our work is therefore as general as the works that do not assume the SEs to be given.

As for future work, it would be interesting to make a quantitative comparison of our EMCC with the MCMC approach recently put forward by Zhang et al. (2022) (a qualitative comparison is available in Zaffalon et al. 2023, Appendix B). Also, it would be important to take advantage of the recent circuital compilation of the EMCC algorithms presented in Huber et al. (2023) to handle bigger models; and it would definitely be useful to extend the basic EMCC to continuous data.

Acknowledgements

Appendix A. Proofs

Proof of Theorem 1. Let us first consider the case $\Theta_{\boldsymbol{U}} \neq \emptyset$. Take $\{P(U)\}_{U \in \boldsymbol{U}} \in \Theta_{\boldsymbol{U}}$. Equation (4) is equivalent to have the following equation:

$$\sum_{\boldsymbol{u}^c \in \Omega_{\boldsymbol{U}}^c} \left[\prod_{U \in \boldsymbol{U}^c} P(u) \cdot \prod_{V \in \boldsymbol{V}^c} P(v|\mathrm{pa}(V)) \right] = \prod_{V \in \boldsymbol{V}^c} P(v|\boldsymbol{w}_V), \qquad (\text{A.1})$$

satisfied for each $c \in \mathscr{C}$. Such a result has been proved by Tian (2002) and Equation (A.1) corresponds to Equation (4.41) in that work. We similarly decompose Equation (5) w.r.t. the c-components:

$$LL(\{P(U)\}_{U \in \boldsymbol{U}}) = \sum_{c \in \mathscr{C}} \sum_{\boldsymbol{w}^c} n(\boldsymbol{w}^c) \cdot \log \sum_{\boldsymbol{u}^c \in \Omega_{\boldsymbol{u}^c}} \left[\prod_{U \in \boldsymbol{U}^c} P(u) \cdot \prod_{V \in \boldsymbol{V}^c} P(v|\mathrm{pa}_V) \right]. \quad (\text{A.2})$$

Putting Equation (A.1) in Equation (A.2) we obtain:

$$LL(\{P(U)\}_{U \in \boldsymbol{U}}) = \sum_{c \in \mathscr{C}} \sum_{\boldsymbol{w}^c} n(\boldsymbol{w}^c) \sum_{V \in \boldsymbol{V}^c} \log P(v|\boldsymbol{w}_V). \tag{A.3}$$

The r.h.s. of Equation (A.3) exhibits the decomposable structure of the multi-nomial likelihood of $\mathscr{D}$ from the endogenous BN. Such a concave function has a unique global maximum, say λ^*, achieved where the relative frequencies are attained by the BN parameters, i.e.,

$$P(v|\boldsymbol{w}_V) := \frac{n(v, \boldsymbol{w}_V)}{n(\boldsymbol{w}_V)}, \tag{A.4}$$

and no local maxima (see, e.g., Koller and Friedman, 2009). This proves the sufficient condition. To prove the necessary condition, consider again Equation (A.1), which is equivalent to Equation (4). Assume, *ad absurdum*, that there is a $\{P(U)\}_{U \in \boldsymbol{U}} \not\in \Theta_{\boldsymbol{U}}$ such that the log-likelihood in Equation (5) attains its global maximum. This means that Equation (A.1) should be violated for at least a value of $\boldsymbol{W}^c$ in a c-component. In this case, putting Equation (A.1) in Equation (A.2) produces a value smaller than the global maximum λ^*, as the maximum is achieved if and only if the values in Equation (A.4) are used. The same deduction can be applied to any $\{P(U)\}_{U \in \boldsymbol{U}}$ when $\Theta_{\boldsymbol{U}} = \emptyset$. $\qquad\square$

Proof of Theorem 2. Let us first define the following function:

$$h(\boldsymbol{v}) := \begin{cases} \boldsymbol{v} & \text{if } g(\boldsymbol{v}) = 1, \\ * & \text{otherwise}. \end{cases} \tag{A.5}$$

Variable $T := h(\boldsymbol{V})$ takes its values from $\Omega_T := \Omega_{\boldsymbol{V}}^{S=1} \cup \{*\}$ where:

$$\Omega_{\boldsymbol{V}}^{S=1} := \left\{ \boldsymbol{v} \in \Omega_{\boldsymbol{V}} : g(\boldsymbol{v}) = 1 \right\}. \tag{A.6}$$

Let $f := \Omega_{\boldsymbol{U}} \to \Omega_{\boldsymbol{V}}$ denote the joint representation of the SEs $\{f_V\}_{V \in \boldsymbol{V}}$. From f and h we can build a map $l : \Omega_{\boldsymbol{U}} \to \Omega_T$ by composition, i.e., $l := h \circ f$. This allows to obtain from M a PSCM M' with the same exogenous variables of M, and T being the only endogenous variable with l as SE. Note that l satisfies surjectivity and M' is therefore a proper PSCM such that T is a common child of all the nodes associated with the variables in $\boldsymbol{U}$. We similarly obtain from $\mathscr{D}_S$ a dataset $\mathscr{D}'$ of N complete observations of T as follows: for each (missing) observation of $\boldsymbol{V}$ in $\mathscr{D}_{S=0}$ (whose number is $N_{S=0}$) we add an observation $T = *$

to $\mathscr{D}'$ (remember that $*$ is a proper state of T); while the observations of $\boldsymbol{V}$ in $\mathscr{D}_{S=1}$ are directly added to $\mathscr{D}'$ and regarded as observations of T instead of $\boldsymbol{V}$.

The log-likelihood of $\mathscr{D}'$ from M' can be computed as in Equation (5), i.e.,

$$LL'(\theta_{\boldsymbol{U}}) := \sum_{t \in \mathscr{D}'} \log \sum_{\boldsymbol{u} \in \Omega_{\boldsymbol{U}}} \left[\prod_{U \in \boldsymbol{U}} \theta_u \cdot [\![l(\boldsymbol{u}) - t]\!] \right]. \tag{A.7}$$

Let us show that Equation (A.7) coincides with Equation (7). Because of Equation (A.5), the elements of $\mathscr{D}'$ such that $T = *$ are those corresponding to $S = 0$, their number being therefore $N_{S=0}$. Their contribution to the log-likelihood is therefore the term in Equation (8). For the other elements of $\mathscr{D}'$, function h in Equation (A.5) acts as the identical map and hence $l(\boldsymbol{u}) = f(\boldsymbol{u})$. The corresponding contribution to the log-likelihood is the one in Equation (9). This proves that the log-likelihood of a biased dataset in Equation (5) can be expressed as the log-likelihood of an unbiased dataset (namely $\mathscr{D}'$) from a proper PSCM (that is M'). This allows to apply Theorem 1 and hence to prove the thesis. $\qquad\square$

References

Balke, A., Pearl, J., 1994. Counterfactual probabilities: Computational methods, bounds and applications, in: De Mantaras, R.L., Poole, D. (Eds.), Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc.. p. 4654.

Bareinboim, E., Pearl, J., 2012. Controlling selection bias in causal inference, in: Lawrence, N.D., Girolami, M. (Eds.), Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR. pp. 100–108.

Bareinboim, E., Pearl, J., 2016. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences 113, 7345–7352.

Bareinboim, E., Tian, J., 2015. Recovering causal effects from selection bias, in: Koenig, S., Blai Bonet, B. (Eds.), Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI Press.

Cabañas, R., Antonucci, A., Huber, D., Zaffalon, M., 2020. CREDICI: a Java library for causal inference by credal networks, in: Yaeger, M., Nielsen, T.D. (Eds.), Proceedings of the tenth International Conference on Probabilistic Graphical Models. JMLR.org. volume 138 of *PMLR*, pp. 597–600.

Cooper, G.F., 1995. Causal discovery from data in the presence of selection bias, in: Fisher, D., Lenz, H.J. (Eds.), Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, PMLR. pp. 140–150.

Darwiche, A., 2022. Causal inference using tractable circuits. URL: `arxiv.org/abs/2202.02891`.

Duarte, G., Finkelsteinm, N., Knox, D., Mummolo, J., Shpitser, I., 2023. An automated approach to causal inference in discrete settings. Journal of the American Statistical Association .

Huber, D., Chen, Y., Antonucci, A., Darwiche, A., Zaffalon, M., 2023. Tractable bounding of counterfactual queries by knowledge compilation. URL: `https://tractable-probabilistic-modeling.github.io/tpm2023/papers`. Sixth Workshop on Tractable Probabilistic Modeling @ UAI 2023.

Kern, H., Corani, G., Huber, D., Vermes, N., Zaffalon, M., Varini, M., Wenzel, C., Fringer, A., 2020. Impact on place of death in cancer patients: a causal exploration in southern Switzerland. BMC Palliative Care 19, 1–10.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT.

Mueller, S., Pearl, J., 2023. Personalized decision making–a conceptual introduction. Journal of Causal Inference 11, 1–13.

Pearl, J., 2009. Causality. Cambridge University Press.

Pearl, J., 2012. A solution to a class of selection-bias problems. Technical Report R-407. UCLA Cognitive Systems Laboratory.

Shpitser, I., Pearl, J., 2007. What counterfactuals can be tested, in: Parr, R., van der Gaag, L. (Eds.), Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, AUAI Press. p. 352359.

Tian, J., 2002. Studies in Causal Reasoning and Learning. Ph.D. thesis. UCLA.

Tian, J., Pearl, J., 2000. Probabilities of causation: Bounds and identification. Annals of Mathematics and Artificial Intelligence 28, 287–313.

Zaffalon, M., Antonucci, A., Cabañas, R., 2020. Structural causal models are (solvable by) credal networks, in: Jaeger, M., Nielsen, T.D. (Eds.), Proceedings of the tenth International Conference on Probabilistic Graphical Models. JMLR.org. volume 138 of *PMLR*, pp. 581–592.

Zaffalon, M., Antonucci, A., Cabañas, R., 2021. Causal expectation-maximisation. Why-21@NeurIPS URL: `https://why21.causalai.net/papers/WHY21_52.pdf`.

Zaffalon, M., Antonucci, A., Cabañas, R., Huber, D., Azzimonti, D., 2022. Bounding counterfactuals under selection bias, in: Salmerón, A., Rumí, R. (Eds.), Proceedings of the eleventh International Conference on Probabilistic Graphical Models. JMLR.org. volume 186 of *PMLR*, pp. 289–300.

Zaffalon, M., Antonucci, A., Cabañas, R., Huber, D., Azzimonti, D., 2023. On the efficient bounding of counterfactual queries (submitted).

Zhang, J., Tian, J., Bareinboim, E., 2022. Partial counterfactual identification from observational and experimental data, in: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (Eds.), Proceedings of the Thirty-Ninth International Conference on Machine Learning, JMLR.org. pp. 26548–26558.