Bounding Counterfactuals under Selection Bias

Marco Zaffalon   zaffalon@idsia.ch
IDSIA, Lugano (Switzerland)

Alessandro Antonucci   alessandro@idsia.ch
IDSIA, Lugano (Switzerland)

Rafael Cabañas   rcabanas@ual.es
Department of Mathematics, University of Almería, Almería (Spain)

David Huber   david.huber@idsia.ch
IDSIA, Lugano (Switzerland)

Dario Azzimonti   dario.azzimonti@idsia.ch
IDSIA, Lugano (Switzerland)

Abstract

Causal analysis may be affected by selection bias, which is defined as the systematic exclusion of data from a certain subpopulation. Previous work in this area focused on the derivation of identifiability conditions. We propose instead a first algorithm to address both identifiable and unidentifiable queries. We prove that, in spite of the missingness induced by the selection bias, the likelihood of the available data is unimodal. This enables us to use the causal expectation-maximisation scheme to obtain the values of causal queries in the identifiable case, and to compute bounds otherwise. Experiments demonstrate the approach to be practically viable. Theoretical convergence characterisations are provided.

Keywords: Causal analysis; structural causal models; expectation maximisation; counterfactuals; unidentifiability.

1. Introduction

Table 1 reports the results of a drug study where gender is taken into account. The study involves 700 patients, but the outcomes for treated females and untreated males (grey counts in the table) are absent because of an issue in the communication protocol. This is an example of *selection bias*: a subpopulation is systematically missing from the sample and this makes any direct data analysis unreliable. (In what follows, we call these data just 'biased' for brevity.)

Such situations have been largely studied in the literature (see, e.g., Winship and Mare (1992); Zaffalon and Miranda (2009)), but relatively little work has been devoted to analysing the question from a causal perspective. After the pioneering work of Cooper (1995), the problem was clearly cast within the framework of structural causal models by Pearl (2012). This eventually led to sound and complete graphical and algorithmic conditions for recovering probabilities from biased data (Bareinboim and Pearl, 2012; Bareinboim and Tian, 2015). These contributions can be understood as an extension to biased data of Pearl's do-calculus, which reduces causal to observational queries in the *identifiable* cases. Yet, when the focus is on counterfactuals, most inferences are unidentifiable even without

selection bias, and the problem of bounding such queries under selection bias is basically unexplored.

The goal of this paper is to fill this gap by providing numerical bounds to unidentifiable counterfactual queries under selection bias. To this end, we start from the recently proposed procedure of Zaffalon et al. (2021) designed to compute those bounds by an iterated, so-called, causal EM scheme. That approach exploits the unimodality of the marginal likelihood of the (unbiased) data. Here, we prove that, in spite of the additional missingness induced by the selection of data, the marginal likelihood remains unimodal. This allows us to adopt the causal EM scheme to compute the bounds of unidentifiable queries under selection bias. To the best of our knowledge this is the first technique proposed for such a task.

The paper is organised as follows. In Sect. 2 we define the basic notation and the necessary background material. The causal EM scheme is reviewed in Sect. 3, where an extended characterisation of convergence is provided. Our approach is discussed in Sect. 4. Experiments are reported in Sect. 5 and conclusions in Sect. 6. Proofs are gathered in the appendix.

Treatment (X)	Recovery (Y)	Gender (Z)	#
0	0	0	2
1	0	0	41
0	1	0	114
1	1	0	313
0	0	1	107
1	0	1	109
0	1	1	13
1	1	1	1

Table 1: Data from an observational study involving three Boolean variables (Mueller et al., 2022). True states correspond to treated (X), recovered (Y) and female (Z).

2. Structural Causal Models

We assume the reader to have some familiarity with probability theory and *structural causal models* (SCMs). Let us focus on models with categorical variables. Variable V takes its values from a finite set Ω_V with v denoting its generic value. $P(V)$ is a *probability mass function* (PMF) over V. Notation $P(V|V') := \{P(V|v')\}_{v' \in \Omega_{V'}}$ is used instead for a *conditional probability table* (CPTs), i.e., a collection of conditional PMFs over a variable indexed by the states of another one. We define an SCM M by a directed graph $\mathcal{G}$ whose nodes are in a one-to-one correspondence with both its endogenous variables $\boldsymbol{X} := (X_1, \ldots, X_n)$ and the exogenous ones $\boldsymbol{U} := (U_1, \ldots, U_m)$. We focus on semi-Markovian models, i.e., $\mathcal{G}$ is assumed to be acyclic. The exogenous variables correspond to the root nodes of $\mathcal{G}$. A PMF $P(U)$ is specified for each $U \in \boldsymbol{U}$. A *structural equation* (SE) f_X is instead provided for each $X \in \boldsymbol{X}$; this is a map $f_X : \Omega_{\mathrm{Pa}_X} \to \Omega_X$, where Pa_X are the parents (i.e., the immediate predecessors) of X according to $\mathcal{G}$. To have all the states of X possibly realised, we only consider SEs corresponding to surjective maps. We might also require a *joint surjectivity*, meaning that any $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ can be realised for at least a $\boldsymbol{u} \in \Omega_{\boldsymbol{U}}$. SCM M induces the

following joint PMF:

$$P(\boldsymbol{x}, \boldsymbol{u}) = \prod_{X \in \boldsymbol{X}} P(x|\mathrm{pa}_X) \prod_{U \in \boldsymbol{U}} P(u), \tag{1}$$

for each $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ and $\boldsymbol{u} \in \Omega_{\boldsymbol{U}}$, where the values of the CPTs associated with each X are degenerate, i.e., $P(x|\mathrm{pa}_X) = \delta_{x, f_X(\mathrm{pa}_X)}$. These are *fully specified*, or just *full*, SCMs (FSCMs). If the exogenous PMFs are not provided, we say instead that M is a *partially specified* (or *partial*, PSCMs). In that case, endogenous information is assumed to be available in the form of a dataset $\mathcal{D}$ of observations of $\boldsymbol{X}$. Given $X \in \boldsymbol{X}$ and $x \in \Omega_X$, an *intervention*, denoted as $\mathrm{do}(X = x)$, describes the action of setting X equal to $x \in \Omega_X$. This corresponds to a *surgery* involving the replacement of f_X with a constant map $X = x$, the consequent removal from $\mathcal{G}$ of the arcs pointing to X, and letting $X = x$ be the evidence.

Fig. 1 represents the graph of the SCM M for the study with outcomes in Tab. 1. The model is *Markovian*, i.e., each endogenous variable has a single exogenous parent. We set *conservative* SEs (Zaffalon et al., 2020), i.e., the exogenous states are indexing all the possible (deterministic) functions from the endogenous parents of an endogenous variable to such a variable (see also Zhang et al. (2022) for an extension of this concept). Accordingly, we have $|\Omega_U| = 4$, $|\Omega_W| = 2$, $|\Omega_V| = 16$.
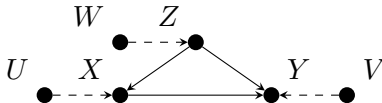

Figure 1: A Markovian SCM.

Causal queries on SCMs involve endogenous variables. Following Pearl's causal hierarchy, we distinguish three kinds of queries. Observational queries, such as $P(X = 1|Y = 1)$, involving the evaluation of marginal or posterior queries for some endogenous variables given an observation of some other ones; interventional queries such as $P(Y = 1|\mathrm{do}(X = 1))$, where the causal effect of an intervention is considered; and counterfactual queries where the same variable is intervened in a state and observed (or intervened) in another one. Classical counterfactual queries for models with Boolean endogenous variables are the probability of necessity PN := $P(Y_{X=0} = 0|X = 1, Y = 1)$, sufficiency PS := $P(Y_{X=1} = 1|X = 0, Y = 0)$ and necessity and sufficiency PNS := $P(Y_{X=0} = 0, Y_{X=1} = 1)$. Subscripts are used to denote interventions related to counterfactuals and, in those queries, $\mathcal{G}$ is so that X topologically precedes Y. E.g., in medical applications PNS can describe the fraction of the population that recovers if cured and does not otherwise. Observational queries can be computed by standard Bayesian network inference tools. In FSCMs, interventional queries can similarly be addressed after standard surgery, while in PSCMs, techniques such as do-calculus allow for a reduction to observational queries in specific cases. Counterfactual queries in FSCMs typically require the application of Bayesian network inference to auxiliary structures like twin networks or counterfactual graphs (Shipster and Pearl, 2007). The problem of addressing counterfactual queries in PSCMs is discussed in the next section.

3

3. Addressing Partial Identifiability (without Selection Bias)

As we have remarked already, the general problem of computing counterfactuals can be reduced to standard Bayesian network inference in FSCMs. The same cannot be done with PSCMs: tools like do-calculus can reduce the causal query to an observational one only in the *identifiable* case. For unidentifiable queries, one needs to compute probability bounds.

As an example consider the PNS for the SCM in Fig. 1 discussed in Sect. 2 with the endogenous data in Tab. 1. This is an unidentifiable query, whose bounds $0 \leq \text{PNS} \leq 0.01458$ have been obtained by Mueller et al. (2022) with some ad-hoc derivation; an exact characterisation of these bounds has been provided by Zaffalon et al. (2020).

More generally speaking, various approaches have recently been proposed to approximate unidentifiable bounds (see, e.g., Zaffalon et al. (2021); Zhang et al. (2022)). Yet, there appears to be no viable technique to compute general counterfactual inference under selection bias. This will be the topic of the next section, which will show how the EMCC of Zaffalon et al. (2021) has a natural extension to problems of selection bias. Before that, we recall the basics of EMCC (Sect. 3.1) and derive improved credible intervals for it (Sect. 3.2).

3.1 EMCC (Zaffalon et al., 2021)

Given a PSCM and a dataset of (unbiased) endogenous observations, Alg. 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 $P_0(u|\boldsymbol{x})$ as a *pseudo-count* for $(u, \boldsymbol{x})$, for each $\boldsymbol{x} \in \mathcal{D}$, $u \in \Omega_U$ and $U \in \boldsymbol{U}$ (E-step). A new estimate $P_1(u) := \sum_{\boldsymbol{x} \in \mathcal{D}} \frac{P_0(u|\boldsymbol{x})}{|\mathcal{D}|}$ is consequently obtained (M-step). The scheme is iterated until convergence. Subroutine `initialisation` (line 1) provides a random initialisation of the exogenous PMFs, ϵ is the threshold to evaluate parameter convergence w.r.t. a metric d (line 3). The FSCM returned by Alg. 1 can indeed be proved compatible with the PSCM in input.

Computing a causal query in this FSCM gives the exact value of the query in such an identifiable case, which then corresponds to a point of the interval for the unidentifiable case represented by the PSCM (Zaffalon et al., 2021). An inner approximation of such an interval 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.

Algorithm 1 EMCC: given SCM M and dataset $\mathcal{D}$ 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** $d[\{P_{t+1}(U)\}_{U \in \boldsymbol{U}}, \{P_t(U)\}_{U \in \boldsymbol{U}}] > \epsilon$ **do**
4:     **for** $U \in \boldsymbol{U}$ **do**
5:         $P_{t+1}(U) \leftarrow |\mathcal{D}|^{-1} \sum_{\boldsymbol{x} \in \mathcal{D}} P_t(U|\boldsymbol{x})$
6:         $t \leftarrow t + 1$
7:     **end for**
8: **end while**

3.2 A New Credible Interval for Unidentifiable Inference

A characterisation of the EMCC accuracy in terms of credible intervals has been provided by Zaffalon et al. (2021). Let $\rho := \{\pi_i\}_{i=1}^k$ denote the output of k EMCC runs, then $[a, b]$ is the interval induced by ρ, i.e., $a := \min_{i=1}^k \pi_i$ and $b := \max_{i=1}^k \pi_i$. If $[a^*, b^*]$ are the exact probability bounds for a causal query, by construction, we have $a^* \leq a \leq b \leq b^*$. The following equality holds:

$$P\left(a - \varepsilon L \leq a^* \leq b^* \leq b + \varepsilon L \,\middle|\, \rho\right) = \frac{1 + (1 + 2\varepsilon)^{2-k} - 2(1 + \varepsilon)^{2-k}}{(1 - L^{k-2}) - (k - 2)(1 - L)L^{k-2}}, \qquad (2)$$

where $L := (b - a)$ and ϵ can be regarded as the relative error at each extreme of the interval $[a, b]$. Eq. (2) is derived by assuming the k points in ρ to be uniformly distributed over the exact interval $[a^*, b^*]$. Here we obtain a new result based on a weaker assumption.

Theorem 1 *Assume that the EMCC runs in ρ are distributed as a four parameter beta distribution, i.e., $\pi_i \sim Beta(\alpha, \beta, a^*, b^*)$, for each $i = 1, \ldots, k$. The following equality holds:*

$$P\left(a - \varepsilon L \leq a^* \leq b^* \leq b + \varepsilon L \,\middle|\, \rho\right) = \frac{\displaystyle\int_0^{\delta/2} \int_0^{\delta/2} P(x, y; L, \alpha, \beta, k) \,\mathrm{d}x\,\mathrm{d}y}{\displaystyle\int_0^{a+(1-b)} \int_0^{a+(1-b)-y} P(x, y; L, \alpha, \beta, k) \,\mathrm{d}x\,\mathrm{d}y}, \qquad (3)$$

where $P(x, y; L, \alpha, \beta, k)$ is equal to

$$\left(\frac{(L + x)^\alpha {}_2F_1(\alpha, 1 - \beta, \alpha + 1, \frac{L+x}{L+x+y}) - x^\alpha {}_2F_1(\alpha, 1 - \beta, \alpha + 1, \frac{x}{L+x+y})}{\alpha(L + x + y)^\alpha B(\alpha, \beta)}\right)^k, \qquad (4)$$

${}_2F_1$ is the Gaussian (ordinary) hypergeometric function, $B(\alpha, \beta)$ is the beta function evaluated at $\alpha, \beta > 0$, and $\delta := 2\varepsilon L$ is the absolute allowed error, with $\delta \in (0, L)$.

The above theorem is a proper extension of the original EMCC characterisation, as proved by the following result.

Corollary 2 *If the EMCC runs in $\{\pi_i\}_{i=1}^k$ are uniformly distributed in $[a^*, b^*]$, i.e., for each $i = 1, \ldots, k$, $\pi_i \sim Beta(1, 1, a^*, b^*)$, then we recover Eq. (2).*

After any k EMCC runs, we can compute a and b and, given a relative error ε we regard as acceptable, obtain the probability in Eq. (4) by estimating the values of the parameters α and β with a maximum likelihood procedure over the k values collected in ρ. Note that, since a^*, b^* are unknown we also need a way to choose the range of the $Beta$ distribution. Here we proceed by taking $Beta(\alpha, \beta, a - \varepsilon L, b + \varepsilon L)$. If the corresponding probability is sufficiently high, we stop iterating EMCC, otherwise we keep iterating the procedure to collect new points for ρ and achieve greater probabilities.

A specialisation of the bound in Eq. (2) can easily be obtained to decide if the query is identifiable (i.e., $a^* = b^*$), when we obtain the same value of π_i for all $i = 1, \ldots, k$.

Corollary 3 *If $a = b$, i.e., all k runs in ρ are equal, then $P(a^* = b^* | \rho) = 1 + 9/3^k - 8/2^k$.*

This implies that nine equal runs make identifiability probable with 99% confidence.

4. Addressing 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 $\mathbf{X}$ and false otherwise. We consider deterministic selectors, i.e., $S := g(\mathbf{X})$ with $g : \Omega_{\mathbf{X}} \to \{0, 1\}$. E.g., the selector for the biased data in Tab. 1 is an exclusive disjunction between treatment and gender, i.e., $S := X \oplus Z$. S can be embedded in SCM M as a common child of the related endogenous variables, with g being the SE of S, thus acting as an additional endogenous variable with endogenous parents only.

Given a dataset $\mathcal{D}$ of observations of $\mathbf{X}$, S partitions $\mathcal{D}$ in $\mathcal{D}_0$ where $S = 0$ and the endogenous observations are missing, and $\mathcal{D}_1$ where $S = 1$ and the observations remain as in $\mathcal{D}$. We use notation $\mathcal{D}_S := \mathcal{D}_0 \cup \mathcal{D}_1$ and, for cardinalities, $d_0 := |\mathcal{D}_0|$, $d_1 := |\mathcal{D}_1|$, with $d_0 + d_1 = d := |\mathcal{D}| = |\mathcal{D}_S|$. Note $\mathcal{D}_0$ contains d_0 identical records. An example is in Fig. 2.

Z	X	Y	S	#
0	0	0	0	2
0	0	1	0	114
0	1	0	1	41
0	1	1	1	313
1	0	0	1	107
1	0	1	1	13
1	1	0	0	109
1	1	1	0	1

$\mathcal{D}$

W	U	V	Z	X	Y	S	#
*	*	*	0	1	0	1	41
*	*	*	0	1	1	1	313
*	*	*	1	0	0	1	107
*	*	*	1	0	1	1	13

$\mathcal{D}_1$

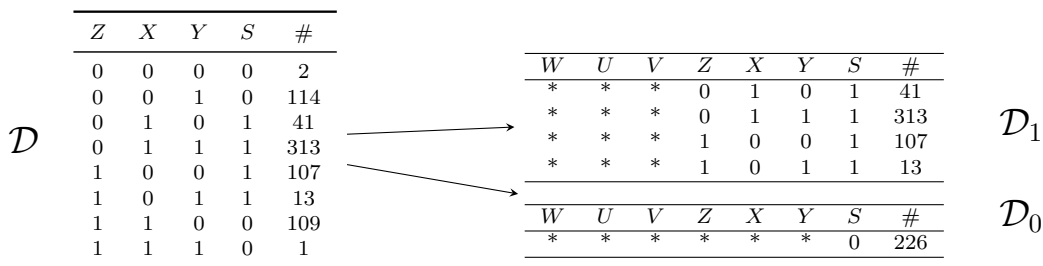| W	U	V	Z	X	Y	S	#
*	*	*	*	*	*	0	226

$\mathcal{D}_0$

Figure 2: Selecting the data in Tab. 1 with $S := X \oplus Z$.

The marginal log-likelihood LL of $\mathcal{D}_S$ according to M is:

$$LL := \log P(\mathcal{D}_S) = \log P(\mathcal{D}_0) + \log P(\mathcal{D}_1) = d_0 \log P(S = 0) + \sum_{\mathbf{x} \in \mathcal{D}_1} \log P(\mathbf{x}) \,. \qquad (5)$$

As Eq. (5) is not the likelihood of a dataset of complete endogenous observations for an SCM, the characterisation provided by Zaffalon et al. (2021) to justify the application of EMCC does not directly hold.

Yet, Eq. (5) contains the sum of two log-likelihoods for disjoint datasets. This leads us to define a simple upper bound LL^* for LL by using maximum-likelihood estimates separately for $\mathcal{D}_0$ and $\mathcal{D}_1$, i.e.,

$$LL^* := d_0 \log \hat{P}(S = 0) + \sum_{\mathbf{x} \in \mathcal{D}_1} \log \hat{P}(\mathbf{x}) \,, \qquad (6)$$

where $\hat{P}(S = 0) := d_0/d$ and, for each $\mathbf{x} \in \mathcal{D}_1$, $\hat{P}(\mathbf{x})$ is obtained from the counts in $\mathcal{D}_1$ by means of the endogenous factorisation induced by the *c-components* of M as in Eqs. (2) and (6) of Zaffalon et al. (2021). As an explicit function of the exogenous PMFs, the log-likelihood of $\mathcal{D}_S$ is instead:

$$LL[\{P(U)\}_{U \in \mathbf{U}}] := d_0 \log \sum_{\mathbf{x} \in \Omega_{\mathbf{X}}^{S=0}} \sum_{\mathbf{u} \in \Omega_{\mathbf{U}}} \delta_{f(\mathbf{u}), \mathbf{x}} P(\mathbf{u}) + \sum_{\mathbf{x} \in \mathcal{D}_1} \log \sum_{\mathbf{u} \in \Omega_{\mathbf{U}}} \delta_{f(\mathbf{u}), \mathbf{x}} P(\mathbf{u}) \,, \qquad (7)$$

where $\Omega_{\mathbf{X}}^{(S=0)} := g^{-1}(S = 0)$ and $f : \Omega_{\mathbf{U}} \to \Omega_{\mathbf{X}}$ is the *joint* SE we build from $\{f_X\}_{X \in \mathbf{X}}$. Remember that we should only be after FSCMs that are compatible with the given PSCM

(also called M-*compatible*), as mentioned in Sect. 3.1. In the present case, M-compatibility corresponds to imposing the following constraints:

$$\sum_{\boldsymbol{x} \in \Omega_{\boldsymbol{X}}^{S=0}} \sum_{\boldsymbol{u} \in \Omega_{\boldsymbol{U}}} \delta_{f(\boldsymbol{u}),\boldsymbol{x}} P(\boldsymbol{u}) = \hat{P}(S=0) \,, \tag{8}$$

and, for each $\boldsymbol{x} \in \mathcal{D}_1$,

$$\sum_{\boldsymbol{u} \in \Omega_{\boldsymbol{U}}} \delta_{f(\boldsymbol{u}),\boldsymbol{x}} P(\boldsymbol{u}) = \hat{P}(\boldsymbol{x}) \,. \tag{9}$$

These compatibility conditions should be self-explanatory in that the models we look for should be capable of generating the distribution of the data, besides being an extension of the given PSCM. It can anyway be shown that they are just the form that the notion of M-compatibility, given in Zaffalon et al. (2021), takes when it is applied to the present context (we omit the proof).

Checking such a compatibility is equivalent to a simple evaluation of the function in Eq. (7) as stated by the following result.

Theorem 4 *As a function of $\{P(U)\}_{U \in \boldsymbol{U}}$, the log-likelihood in Eq. (7) has no local maxima and a global maximum equal to the value LL^* in Eq. (6). Such a maximum is achieved if and only if the M-compatibility constraints in Eqs. (8) and (9) are satisfied.*

As a consequence of Thm. 4, the application of Alg. 1 to PSCM M and data $\mathcal{D}_S$, returns only FSCMs whose log-likelihood takes its global maximum LL^* because of the EM properties and hence satisfies the compatibility constraints. Causal queries computed with those FSCMs are consequently inner points of the exact interval for the partially identifiable query. The more points the better the approximation.

Note that as the data in $\mathcal{D}_0$ are d_0 instances of the same observation $S = 0$, when coping with selection bias, line 5 of Alg. 1 rewrites as:

$$P_{t+1}(U) \leftarrow \frac{d_0 P_t(U|S=0) + \sum_{\boldsymbol{x} \in \mathcal{D}_1} P_t(U|\boldsymbol{x})}{(d_0 + d_1)} \,. \tag{10}$$

As a demonstrative example, Fig. 3 displays the outputs of EMCC for the SCM M described in Sect. 2 in the presence of a selection bias on the dataset in Tab. 1. Different selection mechanisms, including the one discussed in Fig. 2 that corresponds to $P(S = 1) = 0.68$, are considered. The size of the PNS interval for the unselected case $P(S = 1) = 1$, i.e., $[0, 0.1458]$, rapidly increases if fewer data are selected. We also observe that: (i) with no data at all, the PNS interval is very large $[0, 0.9564]$ but not vacuous (this means that SEs alone provide some information about the interval); (ii) if records are removed incrementally, the bounds increase in a *monotone* way (these empirical facts have been also observed in all the experiments considered in Sect. 5); (iii) the intervals are quite large as a consequence of the simple graphical structure considered.

Note that in this limit, we are basically ignoring the data in $\mathcal{D}_1$. If this is the case, the log-likelihood is proportional to d_0 and this implies that EMCC gives the same results irrespectively of the size of the (completely missing) dataset. Let us remark, with reference to Eq. (10), that the execution of EMCC under selection bias requires the cardinality d_0 of $\mathcal{D}_0$. Yet, as the data in $\mathcal{D}_0$ are unavailable by definition, d_0 might be unavailable too. If
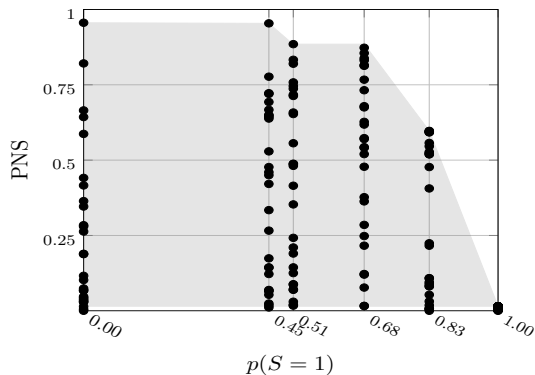
Figure 3: PNS bounds for different selection levels (x-axis). Exact bounds (grey) induced by $r = 30$ EMCC runs and points associated with each run (black) are depicted.

this is the case, it is in principle possible to estimate d_0 from $P(S = 0)$ as $d_0/d_1 \simeq P(S = 0)/P(S = 1)$. If also $P(S = 0)$ is not available, a conservative approach may consist in using an upper bound for $P(S = 0)$. If this is not possible either, one can still use the (very) conservative approach of taking the limit $P(S = 0) \to 1$.

Finally, notice that as in principle S might be a common child of all the variables in $\boldsymbol{X}$, 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 $\boldsymbol{X}$ of bounded cardinality. This limit might be bypassed by the circuital approach of Darwiche (2021).

5. Experiments

Benchmark Generation We assume the endogenous variables of our synthetic SCMs to be Boolean. We randomly generate a directed acyclic graph over $\boldsymbol{X}$ with maximum indegree smaller than or equal to n_p. This graph is augmented with an exogenous parent for each endogenous node to obtain a Markovian model. The model might become non-Markovian by merging together q (distinct) random pairs of exogenous nodes. In these models we specify *conservative* SEs (Zaffalon et al., 2021, Sect. 2), thus obtaining a PSCM for each graph. For each $U \in \boldsymbol{U}$, we sample a PMF $P(U)$ to obtain an FSCM M'. Model M' is used to sample d complete observations of both $\boldsymbol{X}$ and $\boldsymbol{U}$. We denote by $\mathcal{D}$ the corresponding dataset of endogenous observations. Only M-compatible datasets are considered (Zaffalon et al., 2021, Sect. 5). An iterative procedure is applied to the exogenous states that are removed until it is possible to do that while preserving such a compatibility. Given dataset $\mathcal{D}$ and the PSCM M, we consider three endogenous variables X, Y, Z so that X is a root node in $\mathcal{G}$, Y is a leaf and Z is an internal node. $P(S = 1|X, Y, Z)$ is set to one on some of the eight states of (X, Y, Z) and zero elsewhere. We estimate $P(S = 1)$ from the instances of $\mathcal{D}$ that are selected by that particular mechanism. Our benchmark includes 115 SCMs generated by the above procedure with $n_p = 3$, $|\boldsymbol{X}| \in \{4, 5, 6, 7, 8\}$, $|\mathcal{D}| = 1000$ and, for non-Markovian models, $q \in \{2, \ldots, |\boldsymbol{X}|/2\}$.

Results The proposed adaptation of EMCC to cope with selection bias has been implemented within the CREDICI library (Cabañas et al., 2020) and the code to reproduce models and experiments is freely available.[1] Given SCM M, the biased dataset $\mathcal{D}_1$, and the number d_0 of unavailable records, we compute the PNSs $\{\pi_i\}_{i=1}^r$ in the twin networks induced by the FSCMs returned by r EMCC runs. This induces interval $[a_r, b_r]$, where $a_r := \min_{i=1}^r \pi_i$ and $b_r := \max_{i=1}^r \pi_i$. The ground-truth interval $[a^*, b^*]$ for the same setup is obtained as the limit of a large (namely $r_{\max} = 80$) number or runs (cf. Thm. 1). For the unbiased case, we use the method in Zaffalon et al. (2020) to evaluate the interval $[a^*, b^*]$. The quality of the (inner) approximation of $[a^*, b^*]$ provided by $[a_r, b_r]$ can be described by the root mean square error of the difference between the upper bounds and that involving the lower bounds. We normalised that by the size $b^* - a^*$ of the true interval in order to properly compare results for intervals of different sizes. Such a *relative* RMSE, i.e.,

$$\mathrm{RRMSE}_r := \sqrt{\frac{(a_r - a^*)^2 + (b_r - b^*)^2}{2(b^* - a^*)^2}}, \tag{11}$$

is computed for each experiment. Boxplots are finally used to visualise the aggregated results. Fig. 4(a) depicts such errors for increasing numbers of runs, separately for Markovian (360 experiments) and non-Markovian (306 experiments) models. As expected, more EMCC runs give smaller errors and this happens both for Markovian and non-Markovian models. With few runs ($r = 30$) we obtain mean relative errors smaller than 8% for Markovian models and 5% for non-Markovian ones in an average time of 7.6 seconds per run. Such a good accuracy is consistent with the results in Sect. 3.2 and the probabilities returned by Eq. (3) for different values of ε appear to be consistent with the relative RMSEs we get. The difference between Markovian and non-Markovian models appears to be due to the fact that $n_{\boldsymbol{U}} := \prod_{U \in \boldsymbol{U}} |\Omega_U|$ is smaller for non-Markovian cases. Those differences are not detected when comparing models w.r.t. different treewidth or interval sizes. In Fig. 4(b) we compare the EMCC intervals ($r = 30$) with the intervals obtained with the corresponding unbiased data. The same metric as in Eq. (11) is considered, with the interval under selection bias playing the role of the ground truth, and results grouped by different ranges for $P(S = 1)$. As expected higher selection-bias levels make the intervals larger, and the behaviour seems to be monotone.

6. Conclusions

We have proposed a first algorithm to numerically bound counterfactual queries under selection bias. The algorithm is approximate, being based on the EM, and yields consistently accurate results for a relatively low number of EM runs. As future work, we would like to extend the support to non-deterministic selection mechanisms, while using tractable circuits for the modelling of the deterministic case.

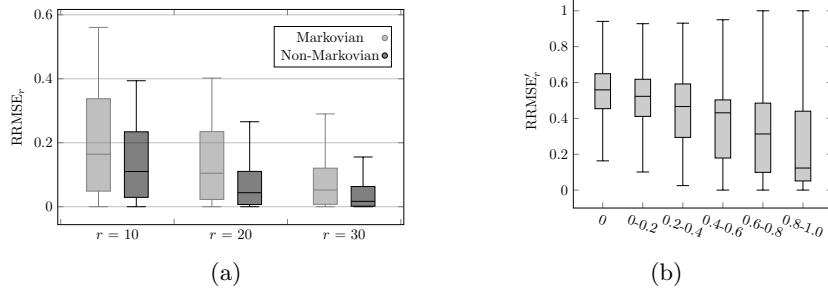1. `https://github.com/IDSIA-papers/2022-PGM-selection`.

Figure 4: (a) EMCC accuracy for different numbers of runs. (b) Difference between the bounds of biased and unbiased data w.r.t. $P(S = 1)$ (x-axis).

Proofs

Proof [of Thm. 1] Consider the l.h.s. of Eq. (3). The corresponding joint density is:

$$P\left(\Delta_a \le \frac{\delta}{2}, \Delta_b \le \frac{\delta}{2}, \rho\right) = \gamma \int_0^{\frac{\delta}{2}} \int_0^{\frac{\delta}{2}} P(\rho|\Delta_a = x, \Delta_b = y)\mathrm{d}x\mathrm{d}y, \qquad (12)$$

where a uniform prior is considered and $\Delta_a := (a - a^*)$ and $\Delta_b := (b^* - b)$. As $\pi_i \sim$ Beta$(\alpha, \beta, a^*, b^*)$, $P(\pi_i \in [a, b]|\Delta_a = x, \Delta_b = y)$ is $P(\pi_i \in [a, b]|a^* = a - x, b^* = b + y)$ and:

$$\int_a^b \frac{(\rho - a + x)^{\alpha-1}(b + y - \rho)^{\beta-1}}{(b - a + y + x)^{\alpha+\beta-1}B(\alpha, \beta)}\mathrm{d}\rho. \qquad (13)$$

Then we have $P(\rho|\Delta_a = x, \Delta_b = y)$ equal to:

$$\left(\frac{(b - a + x)^\alpha {}_2F_1(\alpha, 1 - \beta, \alpha + 1, \frac{b-a+x}{b-a+x+y}) - x^\alpha {}_2F_1(\alpha, 1 - \beta, \alpha + 1, \frac{x}{b-a+x+y})}{\alpha(b - a + x + y)^\alpha B(\alpha, \beta)}\right)^k. \qquad (14)$$

The joint $P(\rho, \Delta_a \le \frac{\delta}{2}, \Delta_b \le \frac{\delta}{2})$ can thus be obtained by the following integral:

$$\gamma \int_0^{\delta/2} \int_0^{\delta/2} \left(\frac{(b - a + x)^\alpha {}_2F_1(\alpha, 1 - \beta, \alpha + 1, \frac{b-a+x}{b-a+x+y}) - x^\alpha {}_2F_1(\alpha, 1 - \beta, \alpha + 1, \frac{x}{b-a+x+y})}{\alpha(b - a + x + y)^\alpha B(\alpha, \beta)}\right)^k \mathrm{d}x\mathrm{d}y. \qquad (15)$$

The marginal distribution for ρ can also be obtained by solving the following integral:

$$P(\rho) = \gamma \int_0^{a+(1-b)} \int_0^{a+(1-b)-y} P(\rho|\Delta_a = x, \Delta_b = y)\mathrm{d}x\mathrm{d}y.$$

The l.h.s. of Eq. (3) is just the ratio between $P(\rho, \Delta_a \le \frac{\delta}{2}, \Delta_b \le \frac{\delta}{2})$ and $P(\rho)$. ∎

Proof [of Cor. 2] For $\alpha = \beta = 1$, ${}_2F_1(\alpha, 1 - \beta, \alpha + 1, \mu) = 1$, $B(\alpha, \beta) = 1$ and hence:

$$P(\rho|\Delta_a = x, \Delta_b = y) = \left(\frac{b - a + x - x}{b - a + x + y}\right)^n = \left(\frac{b - a}{b - a + x + y}\right)^k. \qquad (16)$$

Eq. (2) is finally obtained by computing the integrals as in Zaffalon et al. (2021). ∎

Proof [of Cor. 3] If $a^* = b^*$ then we have that $a \to b$, i.e., $L \to 0$. Moreover we assume here that $\delta \to 0$ and $\epsilon \to 1$. We assume that the distribution of the outputs of the EMCC iterations is uniformly distributed in $[a^*, b^*]$. $P(a = b|\rho)$ is obtained from Eq. (2) in the limit $L \to 0$ and $\epsilon \to 1$ and the thesis follows trivially. ∎

Proof [of Thm. 4] Let us first define a variable $Y := h(\boldsymbol{X})$ where:

$$h(\boldsymbol{x}) := \begin{cases} \boldsymbol{x} & \text{if } g(\boldsymbol{x}) = 1, \\ * & \text{otherwise} . \end{cases} \tag{17}$$

The states of Y are those in $\Omega_Y := \Omega_{\boldsymbol{X}}^{S=1} \cup \{*\}$ with $\Omega_{\boldsymbol{X}}^{S=1} := g^{-1}(S = 1)$. As $\boldsymbol{X} = f(\boldsymbol{U})$, $Y = l(\boldsymbol{U})$ with $l := h \circ f$. This allows to obtain from M a PSCM M' with the same exogenous variables, and Y as the only endogenous with SE is l. Note that l satisfies surjectivity and M' is therefore a PSCM s.t. Y is a common child of all the nodes associated with the variables in $\boldsymbol{U}$. We similarly obtain from $\mathcal{D}_S$ a dataset $\mathcal{D}'$ of d complete observations of Y as follows: for each (missing) observation of $\boldsymbol{X}$ in $\mathcal{D}_0$ (whose number is d_0) we add an observation $Y = *$ to $\mathcal{D}'$ (remember that $*$ is a proper state of Y); while the observations of $\boldsymbol{X}$ in $\mathcal{D}_1$ are directly added to $\mathcal{D}'$, where they are regarded as observations of Y instead of $\boldsymbol{X}$. Denote the latter set as $\mathcal{D}'_1$. The marginal log-likelihood of $\mathcal{D}'$ for M' is:

$$LL' := \log P(\mathcal{D}') = \sum_{y \in \mathcal{D}'} P(y) = d_0 \log P(Y = *) + \sum_{\boldsymbol{x} \in \mathcal{D}'_1} \log P(Y = \boldsymbol{x}) . \tag{18}$$

As $Y = *$ if and only if $S = 0$ and, by construction, $\mathcal{D}_1 = \mathcal{D}'_1$, writing the log-likelihood of $\mathcal{D}'$ as a function of the exogenous PMFs gives:

$$LL[\{P(U)\}_{U \in \boldsymbol{U}}] = d_0 \log P(S = 0) + \sum_{\boldsymbol{x} \in \mathcal{D}_1} \log \sum_{\boldsymbol{u} \in \Omega_{\boldsymbol{U}}} \delta_{l(\boldsymbol{u}), \boldsymbol{x}} P(\boldsymbol{u}) . \tag{19}$$

Yet, $\boldsymbol{x} \in \mathcal{D}_1$ means $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}^{S=1}$ and hence $g(\boldsymbol{x}) = 1$. On these values h in Eq. (17) acts as the identical map, and $l(\boldsymbol{u}) = h \circ f(\boldsymbol{u}) = f(\boldsymbol{u})$. This proves that functions in Eqs. (19) and (7) coincide, and hence that the log-likelihood of $\mathcal{D}_S$ for M coincides with the log-likelihood of $\mathcal{D}'$ for SCM M'. Applying Thm. 1 in Zaffalon et al. (2021) to such a log-likelihood proves that it has no local maxima and a single global maximum value; that is the first claim.

To prove the equivalence between the M-compatibility constraints and the log-likelihood achieving its (global) maximum value LL^*, we start from the direct implication. This easily follows by putting Eqs. (8) and (9) in Eq. (7) as this gives Eq. (6). For the converse, assume *ad absurdum* an incompatible specification giving value LL^*. The two terms in the r.h.s. of Eq. (7) cannot coincide with those in the r.h.s. of Eq. (6) otherwise the model would be compatible. One of them should be greater than the corresponding term in LL^*: but this is impossible as both terms of LL^* have been obtained by independent maximisation of the log-likelihood functions associated with $\mathcal{D}_0$ resp. $\mathcal{D}_1$. This proves the equivalence between the compatibility constraints and the fact that the log-likelihood is equal to LL^*. It also proves that LL^* is actually the global maximum of the log-likelihood, since incompatible models return values smaller than LL^*, while compatible ones match LL^* exactly. ∎

Acknowledgments

This research was partially funded by MCIN/AEI/10.13039/501100011033 with FEDER funds for the project PID2019-106758GB-C32, and also by Junta de Andalucía grant P20-00091. Finally we would like to thank the "María Zambrano" grant (RR_C_2021_01) from the Spanish Ministry of Universities and funded with NextGenerationEU funds.

References

E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Proceedings of AISTATS*, 2012.

E. Bareinboim and J. Tian. Recovering causal effects from selection bias. In *Proceedings of AAAI*, 2015.

R. Cabañas, A. Antonucci, D. Huber, and M. Zaffalon. CREDICI: a Java library for causal inference by credal networks. In *Proceedings of PGM*, volume 138 of *PMLR*, pages 597–600, 2020.

G. Cooper. Causal discovery from data in the presence of selection bias. In *Proceedings of AISTATS*, 1995.

A. Darwiche. Causal inference using tractable circuits. In *WHY-21 @ NeurIPS*, 2021. URL https://why21.causalai.net/papers/WHY21_51.pdf.

S. Mueller, A. Li, and J. Pearl. Causes of effects: Learning individual responses from population data. In *Proceedings of IJCAI*, 2022.

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

I. Shipster and J. Pearl. What counterfactuals can be tested. In *Proceedings of UAI*, 2007.

C. Winship and R. D. Mare. Models for sample selection bias. *Annual review of sociology*, 18(1):327–350, 1992.

M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821, 2009.

M. Zaffalon, A. Antonucci, and R. Cabañas. Structural causal models are (solvable by) credal networks. In *Proceedings of PGM*, volume 138 of *PMLR*, pages 581–592, 2020.

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

J. Zhang, J. Tian, and E. Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of ICML*, 2022.