A Bayesian Approach for Estimating Causal Effects from Observational Data

Johan Pensar
Dept. of Math. and Stat.
University of Helsinki
johan.pensar@helsinki.fi

Topi Talvitie
Dept. of Computer Science
University of Helsinki
topi.talvitie@helsinki.fi

Antti Hyttinen
HIIT & Dept. of CS
University of Helsinki
antti.hyttinen@helsinki.fi

Mikko Koivisto
Dept. of Computer Science
University of Helsinki
mikko.koivisto@helsinki.fi

Abstract

We present a novel Bayesian method for the challenging task of estimating causal effects from passively observed data when the underlying causal DAG structure is unknown. To rigorously capture the inherent uncertainty associated with the estimate, our method builds a Bayesian posterior distribution of the linear causal effect, by integrating Bayesian linear regression and averaging over DAGs. For computing the exact posterior for all cause-effect variable pairs, we give an algorithm that runs in time $O(3^d d)$ for d variables, being feasible up to 20 variables. We also give a variant that computes the posterior probabilities of all pairwise ancestor relations within the same time complexity, significantly improving the fastest previous algorithm. In simulations, our Bayesian method outperforms previous methods in estimation accuracy, especially for small sample sizes. We further show that our method for effect estimation is well-adapted for detecting strong causal effects markedly deviating from zero, while our variant for computing posteriors of ancestor relations is the method of choice for detecting the mere existence of a causal relation. Finally, we apply our method on observational flow cytometry data, detecting several causal relations that concur with previous findings from experimental data.

Introduction

Understanding the behaviour of a system under the influence of interventions is the ultimate goal of many scientific studies. As a result, the problem of estimating causal effects from empirical data has received a lot of attention in a wide variety of fields. In the most basic setting, we only have a passively observed set of measurements over the variables of interest. In this article, we propose a Bayesian method for estimating causal effects from such data alone, that is, without the often unavailable knowledge of the causal structure of the system.

If the causal structure *is* known, then do-calculus and the ID algorithm can identify the causal effect whenever it can be (non-parametrically) identified, even in the presence of latent confounders (Shpitser and Pearl 2006; Tian and Pearl 2002). When causal relations are restricted to be linear, more effects can be identified through more complicated criteria (Tian 2004; van der Zander and Liskiewicz 2016; Chen, Kumor, and Bareinboim 2017). Moreover, under the assumption of causal sufficiency (i.e., absence of latent confounders), all causal effects can be identified by back-door adjustment according to the known causal structure (Spirtes, Glymour, and Scheines 1993; Pearl 2009; Greenland, Robins, and Pearl 1999).

However, when we do not have extensive knowledge about the system under investigation in form of a causal graph, we also have to make inferences on this graph structure from the available data. From observational data alone, the causal graph can in general only be identified up to its Markov equivalence class. Several works study causal effects in the light of the limited identifiability of causal structures (Entner, Hoyer, and Spirtes 2013; Hyttinen, Eberhardt, and Järvisalo 2015; Perković et al. 2018; Jaber, Zhang, and Bareinboim 2018a; 2018b; Malinsky and Spirtes 2017).

More practical estimation methods are also available, when causal sufficiency can be assumed. Specifically, the IDA algorithm estimates causal effects from linear Gaussian data by backdoor adjustment over the graphs in the equivalence class returned by the PC algorithm (Maathuis, Kalisch, and Bühlmann 2009; Spirtes, Glymour, and Scheines 1993). For a given graph, the causal effect between two variables is estimated by linear regression, where the set of covariates depends on the parents of the cause variable. Consequently, since the parent set of a node typically varies within an equivalence class, a (multi)set of coefficients is returned as the causal effect estimate.

Although a set of coefficients captures some of the uncertainty in the estimated effect, a lot of uncertainty remains unaccounted for. In particular, the accuracy hinges on the structure learning step. In addition to the unoriented edges within a Markov equivalence class, there may be several non-equivalent graphs that fit the finite data set (almost) equally well, yet still yield very different causal effects estimates. Attempting to account for this uncertainty, the original IDA method has been combined with various resampling strategies, which associate the estimates with frequentist measures of confidence (Stekhoven et al. 2012; Taruttis, Spang, and Engelmann 2015).

Here, we introduce a Bayesian approach for causal effect estimation, which employs Bayesian model averaging to fully account for our lack of knowledge of the underlying causal structure. By integrating averaging over graphs and Bayesian linear regression, we produce a Bayesian posterior, which explicitly describes the knowledge and the uncertainty on the causal effect, given the available data. For the computationally heavy averaging over graphs, we give an exact algorithm that runs in time $O(3^d d)$, where d is the number of variables; thus the algorithm scales to moderately high dimensions, without resorting to approximations with uncontrolled accuracy. We show that on realistic-size data sets our approach yields estimates that are more accurate than those produced by previous methods.

When the interest is in the *existence* of a causal relation, rather than in the magnitude of the causal effect, there is a more direct Bayesian solution: compute the posterior probability that the variable of interest is an ancestor of the target variable in the unknown graph. To compute these probabilities for all pairs of variables, Chen, Meng, and Tian (2015) recently gave an algorithm whose running time scales as $O(5^d d^2)$. We give a novel algorithm that improves the running time to $O(3^d d)$; we obtain this algorithm as a variant of our method for averaging over graphs. Armed with our faster algorithm, we also compare empirically this direct approach to that of inferring whether the causal effect is nonzero.

While we in this work focus on exact exponential algorithms, we believe our techniques could be useful also in designing efficient approximate methods. We defer a more thorough discussion of the tradeoff between scalability and controlled error in computations to the end of this paper.

Preliminaries

A *directed acyclic graph (DAG)* $G = (V, E)$ consists of a node set V and an edge set $E \subseteq V \times V$ that contains no directed cycles. If (i, j) is an edge from i to j in G, then i is called a *parent* of j. We denote by G_j the set of parents of node j in G. If there is a directed path from i to j, then i is called an *ancestor* of j and j a *descendant* of i.

In a (probabilistic) *DAG model*, we take V as an index set $\{1, \ldots, d\}$ and associate each node $i \in V$ with a random variable x_i. The model asserts that each variable x_i is conditionally independent of its non-descendants given its parental variables x_{G_i}, enabling a factorization of the joint distribution over the variables:

$$f(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i \,|\, x_{G_i}).$$

We assume here that the conditional distributions are linear Gaussians (Geiger and Heckerman 1994):

$$f(x_i \,|\, x_{G_i}) = \mathrm{N}(\beta_0 + \beta^\top x_{G_i}; \sigma_i^2)\,. \qquad (1)$$

As a result, the joint distribution of $(x_1, \ldots, x_d)$ is a d-dimensional normal distribution. Without loss of generality, we assume that the distribution is zero-centered, that is, $\beta_0 = 0$ in (1).

In this work, we assume that the DAG has a causal interpretation, and we are interested in estimating the causal (intervention) effect of a variable x_i on another variable x_j for any $i, j \in V$. In the considered model space, the causal effect can be estimated from the intervention distribution $f\big(x_j \,|\, \mathrm{do}(x_i = u)\big)$, where the do-operator represents that x_i is set to the value $u \in \mathbb{R}$ by an external intervention, such that the rest of the system is left unaltered (Pearl 2009).

For a given DAG, the intervention distribution can be calculated using a technique known as back-door adjustment (Spirtes, Glymour, and Scheines 1993; Pearl 2009; Greenland, Robins, and Pearl 1999). In particular, if $j \notin G_i$, then G_i satisfies the back-door criterion for the above case and the intervention distribution can be calculated from the pre-intervention (i.e. observational) distribution:

$$f\big(x_j \mid \mathrm{do}(x_i = u)\big) = \int f(x_j \mid x_i = u, x_{G_i}) f(x_{G_i})\, dx_{G_i}.$$

If $j \in G_i$, this is simply reduced to the marginal distribution:

$$f\big(x_j \,|\, \mathrm{do}(x_i = u)\big) = f(x_j).$$

Furthermore, the causal relationship can be quantified by

$$\theta_{ij} = \frac{\partial}{\partial x} \mathbb{E}\left(x_j \,|\, \mathrm{do}(x_i = x)\right) |_{x=u} \;. \qquad (2)$$

While the function in (2) in general depends on the value u, it turns out to be a constant single value for Gaussian DAG models (Pearl 2009). For a given Gaussian DAG model, θ_{ij} can be calculated either from the joint distribution or from the edge weights in (1) using the method of path coefficients (Wright 1934).

The problem considered in this work is estimating the causal effect parameter θ_{ij} in (2) from a set of data, D, without any prior knowledge about the causal DAG. Moreover, we assume that the data set is observational (non-interventional) and has been generated from an underlying Gaussian DAG model. We also consider the related problem of computing the posterior probability of an ancestral relation, that is, the existence of any directed path from a node to another.

A Bayesian Posterior for Causal Effects

In the absence of a known causal structure, a causal effect estimator involves a considerable amount of uncertainty. This uncertainty is primarily related to the unknown DAG, but also to the estimated effect for a given DAG (Maathuis, Kalisch, and Bühlmann 2009). In this section, we present a Bayesian approach which explicitly accounts for all the uncertainty involved in the estimation procedure.

A Bayesian Posterior When the DAG Is Known

For a given DAG, a consistent estimator for the causal effect θ_{ij} is obtained by solving the linear model

$$x_j = \beta_i x_i + \beta_{G_i}^\top x_{G_i} + \epsilon\,, \quad \epsilon \sim \mathrm{N}(0, \sigma_j^2)\,, \qquad (3)$$

and reading off the estimate of regression coefficient β_i. Using this technique, the causal effect estimator is

$$\hat{\theta}_{ij} := \begin{cases} \hat{\beta}_i & \text{if } j \notin G_i\,, \\ 0 & \text{if } j \in G_i\,, \end{cases} \qquad (4)$$

where $\hat{\beta}_i$ is the estimated regression coefficient in the parent-adjusted linear model (3). Note that the estimator only requires local information about the DAG: the parents of node i (Maathuis, Kalisch, and Bühlmann 2009).

We assume here a specific form of prior distribution over the model parameters, and derive the Bayesian posterior distribution analytically. More specifically, for the parameters in the linear model in (3), here denoted by $\beta = (\beta_i, \beta_{G_i})$ and $\sigma^2 = \sigma_j^2$, we assume a prior distribution

$$f(\beta, \sigma^2) = f(\beta \mid \sigma^2) f(\sigma^2)$$

where

$$\begin{aligned} \beta \mid \sigma^2 &\sim \mathrm{N}(m_0, \sigma^2 \Lambda_0^{-1}) \,, \\ \sigma^2 &\sim \text{Inv-Gamma}(a_0, b_0) \,, \end{aligned} \tag{5}$$

with hyperparameters m_0, Λ_0, a_0 and b_0. Under this so-called conjugate prior, the posterior distribution will be of the same form, and the hyperparameters are updated according to:

$$\begin{aligned} m_n &= (\mathbf{X}^\top \mathbf{X} + \Lambda_0)^{-1} (\mathbf{X}^\top \mathbf{y} + \Lambda_0 m_0) \,, \\ \Lambda_n &= \mathbf{X}^\top \mathbf{X} + \Lambda_0 \,, \\ a_n &= a_0 + n/2 \,, \\ b_n &= b_0 + (\mathbf{y}^\top \mathbf{y} + m_0^\top \Lambda_0 m_0 - m_n^\top \Lambda_n m_n)/2 \,, \end{aligned}$$

where $\mathbf{X}$ represents a matrix of n observations over variables (x_i, x_{G_i}), and $\mathbf{y}$ denotes a corresponding vector of observations of variable x_j. Moreover, after marginalizing out σ^2, the joint distribution over the regression coefficients follow a multivariate t-distribution:

$$\beta \mid D \sim t_{2a_n}(m_n, \frac{b_n}{a_n} \Lambda_n^{-1}) \,.$$

As a result, a Bayesian estimator of θ_{ij} under a given DAG is conveniently given by

$$f(\theta_{ij} \mid D, G_i) = \begin{cases} t_{2a_n}(\mu_i, \Sigma_{i,i}) & \text{if } j \notin G_i \,, \\ 0 & \text{if } j \in G_i \,, \end{cases} \tag{6}$$

where μ_i is the location parameter corresponding to node i, and $\Sigma_{i,i}$ is the diagonal value in the shape matrix corresponding to node i.

A Bayesian Posterior When the DAG Is Unknown

The estimators in (4) and (6) assume that the causal structure, or DAG, is known. This is usually not the case in practice, and the DAG needs to be inferred from the data. However, as a DAG can be inferred only up to its equivalence class from observational data, one can only hope to identify a collection of causal effect estimates, one for each DAG in the equivalence class. The IDA algorithm straightforwardly implements this inference by learning a CPDAG from the data, under which a (multi)set Θ_{ij} of possible values of θ_{ij} is estimated (Maathuis, Kalisch, and Bühlmann 2009).

In general, and especially for small data sets, it is difficult to single out one specific CPDAG with high confidence. Therefore, rather than fixing a single CPDAG, we employ Bayesian model averaging (BMA) over the set of DAGs. BMA provides a principled mechanism for converting uncertainty about the model, in our case the causal DAG, into uncertainty about a parameter of interest (Hoeting et al. 1999).

Using BMA, we thus obtain a general posterior causal effect distribution that is not tied down to a specific causal structure. Since the estimator in (6) is the same for all graphs that agree with G_i (parents of node i), averaging over graphs boils down to averaging over the parent sets of node i:

$$f(\theta_{ij} \,|\, D) = \sum_{G_i \subseteq V \setminus \{i\}} f(\theta_{ij} \,|\, D, G_i) p(G_i \,|\, D) \,, \tag{7}$$

where $f(\theta_{ij} \,|\, D, G_i)$ is the posterior of the causal effect θ_{ij} for a given parent set G_i, as defined in (6), and $p(G_i \,|\, D)$ is the posterior probability of the parent set.[1] This key observation renders the posterior computationally feasible for moderately large d, as shown in the following section. The posterior distribution (7) is a mixture distribution, where each parent-specific component is either a t-distribution or a point mass at zero, and its corresponding weight is the posterior probability of that particular parent set.

In terms of asymptotic properties, the estimator in (7) is clearly consistent in the following sense: Let G^* be a DAG and F a faithful linear Gaussian distribution, i.e., F has no other independencies than those entailed by G^*. Then, as the size of the data drawn from F tends to infinity, the posterior of θ_{ij} converges to a discrete distribution whose support is precisely the set of possible values for the true causal effect, Θ_{ij}. The mild assumptions we need for this convergence result are that the structure prior $p(G)$ is everywhere positive and that the marginal likelihood $p(D \,|\, G)$ is *consistent*, i.e., maximized by G^* with probability that tends to 1. The latter property holds under standard proper parameter priors (Geiger and Heckerman 2002), but also under the objective Bayesian scheme (Consonni and Rocca 2012).

Moreover, we have that the multiplicities in Θ_{ij} match the corresponding (limiting) posterior probabilities, provided that the posterior is *score equivalent*, i.e., assigns the same probability for equivalent DAGs. Again, this holds under the mentioned parameters priors if we additionally use a structure prior $p(G)$ that assigns the same probability for equivalent DAGs (e.g., a uniform prior over DAGs).

Exact Computation of the Parent Set and Ancestor Relation Probabilities

The main computational challenge related to the posterior (7) is the calculation of the parent set posterior probabilities. To formulate the problem more precisely, we assume that the prior $p(G)$ is modular, i.e., it is a product of node-wise weights $q_v(G_v)$ (which are generally not proportional to the priors $p(G_v)$ they imply), and the parameters associated with each node (and incoming edges) are independent given the graph G. Under these standard assumptions (Koller and Friedman 2009, pp. 804–806), the posterior probability that G_i is the parent set of i can be written as

$$p(G_i \,|\, D) = p(D)^{-1} \sum_{G : G_i} \prod_{v \in V} w_v(G_v) \,, \tag{8}$$

$$\text{with } \; w_v(G_v) := p(D_v \,|\, D_{G_v}, G_v) q_v(G_v) \,,$$

where the sum is over all DAGs G on V with G_i as the parents of i. If one wants to ensure score equivalence of the posterior,

[1] When the value of G_i is unspecified, G_i denotes the random variable that corresponds to the parent set of i; however, when G_i is given a value, like in the summation (7), G_i is interpreted as the value of the random variable, whose identity is clear in the context.

the local marginal likelihood terms $p(D_v \mid D_{G_v}, G_v)$ need to be specified under a particular class of parameter priors (Geiger and Heckerman 2002).

We emphasize that the results of this section hold with any priors that admit modularity and parameter independence and yield local weights $w_v(G_v)$ that can be efficiently computed, e.g., using a closed-form expression. Thus we also cover common settings where some or all variables are discrete.

We will give an algorithm that, when slightly modified, also computes the ancestor posterior probabilities between all pairs of distinct nodes i, j, given by

$$p(i \rightsquigarrow j \mid D) = p(D)^{-1} \sum_{G: i \rightsquigarrow j} \prod_{v \in V} w_v(G_v)\,, \tag{9}$$

where G runs over all DAGs on V such that there is a directed path from i to j in G.

Theorem 1. *The $2^{d-1}d$ parent set probabilities and the $d(d-1)$ ancestor relation probabilities can be computed in time $O(3^d d)$ and space $O(2^d d)$, given the $2^{d-1}d$ weights $w_v(G_v)$ as input.*

The algorithm, given in the next subsection, is inspired by an algorithm of Tian and He (2009), which computes the posterior probability of any fixed subgraph in time $O(3^d)$. Running it for all possible parent sets would be expensive, however. Tian and He also gave a variant that computes all edge posterior probabilities in time $O(3^d d)$ by reusing intermediate results. We observe that a similar trick enables handling all the exponentially many parent sets with essentially no computational overhead. We will also see that adding the constraint $i \rightsquigarrow j$ does not lead to computational complications; our algorithm substantially improves upon a previous, $O(5^d d^2)$-time algorithm (Chen, Meng, and Tian 2015).

We note that related dynamic programming algorithms have been given for finding a maximum-a-posteriori DAG (Silander and Myllymäki 2006; Yuan and Malone 2013) and for Bayesian model averaging under so-called order-modular priors (Koivisto and Sood 2004; Koivisto 2006). The modular prior is often preferred as it supports, e.g., the uniform distribution over DAGs, whereas the order-modular prior favors DAGs that admit a larger number of topological sorts.

The Algorithm—Proof of Theorem 1

For convenience, for two distinct nodes $i, j \in V$ and a node set $S \subseteq V \setminus \{i\}$ denote the unnormalized posteriors by

$$W_i(S) := \sum_{G: G_i = S} \prod_{v \in V} w_v(G_v) \quad \text{and}$$
$$W_{i,j} := \sum_{G: i \rightsquigarrow j} \prod_{v \in V} w_v(G_v)\,,$$

where G runs over all DAGs on V under the shown constraints. To prove Theorem 1, we show that these numbers can be computed within the claimed resources; specifically, the normalizing constant $p(D)$ is obtained by summing up the total weights $W_i(S)$ over $S \subseteq V \setminus \{i\}$ for any fixed i.

We present our forward–backward algorithm first for the parent set probabilities. The term "forward–backward" was used in a similar context by Koivisto (2006) and it originates from a popular algorithm for computing the posterior marginals of all hidden state variables in a hidden Markov model. We use a similar idea: we obtain each quantity $W_i(S)$ by combining some "forward weights" of DAGs on non-descendants of i and some "backward weights" of DAGs on descendants of i. To enable such a decomposition, we need to fix the set of non-descendants U of node i by summing over all possibilities:

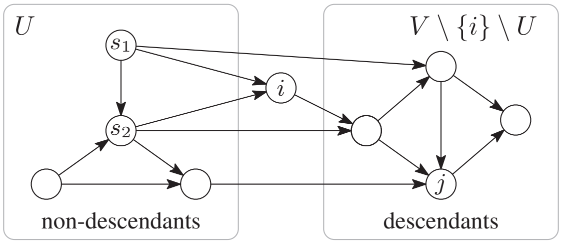

Figure 1: The forward–backward decomposition with a node i and a set of its non-descendants U. Shown is also an example of a DAG that is compatible with the decomposition and the constraints for $W_i\big(\{s_1, s_2\}\big)$ and $W_{i,j}$.

$$W_i(S) = \sum_{S \subseteq U \subseteq V \setminus \{i\}} f(U)\; w_i(S)\; b_i(V \setminus \{i\} \setminus U)\,, \tag{10}$$

where $f(U)$ is the total weight of all DAGs on U and $b_i(T)$ is the total weight of all combinations of parents sets $G_v \subseteq V \setminus \{v\}$ for the remaining nodes $v \in T$ such that there are no directed cycles and each $v \in T$ is a descendant of i. More formally,

$$f(U) := \sum_{G \in \mathcal{G}(U)} \prod_{v \in U} w_v(G_v) \quad \text{and}$$
$$b_i(T) := \sum_{G \in \mathcal{G}(i,T,V)} \prod_{v \in T} w_v(G_v)\,,$$

where $\mathcal{G}(U)$ is the set of all DAGs on U and $\mathcal{G}(i, T, V)$ is the set of tuples $(G_v)_{v \in T}$ such that

- $G_v \subseteq V \setminus \{v\}$ for each $v \in T$,
- the directed graph $\big(T, \bigcup_{v \in T} \{uv : u \in G_v \cap T\}\big)$ is acyclic,
- every G_v intersects $T \cup \{i\}$;

by the last condition and acyclicity, each $v \in T$ is a descendant of i. To justify the product rule in Eq. (10), observe that combining $G_i = S$ with any members of $\mathcal{G}(U)$ and $\mathcal{G}(i, V \setminus \{i\} \setminus U, V)$ results in a DAG on V with U as the set of non-descendants of i. See Fig. 1 for an illustration.

Equation (10) gives us a way to compute all $W_i(S)$ in time $O(3^d d)$, provided that the functions f and b_i have been precomputed. Indeed, for each i, exactly 3^{d-1} pairs S, U satisfy $S \subseteq U \subseteq V \setminus \{i\}$.

Before addressing the precomputation, let us consider the unnormalized ancestor relation probabilities $W_{i,j}$. Again, we partition the DAGs according to the set U of non-descendants of node i. However, instead of having a fixed parent set

$S \subseteq U$, we now sum over all possible parent sets $S \subseteq U$ and require that node j is a descendant of i, i.e., j is not among the non-descendants of i. We get that

$$W_{i,j} = \sum_{j \notin U \subseteq V \setminus \{i\}} f(U)\, \hat{w}_i(U)\, b_i(V \setminus \{i\} \setminus U)\,, \quad (11)$$

$$\text{with } \hat{w}_i(U) := \sum_{S \subseteq U} w_i(S)\,.$$

For each i, the function $\hat{w}_i$ is called the *zeta transform* of w_i and is straightforward to compute with $O(3^d)$ additions, or faster, in time $O(2^d d)$, using *fast zeta transform* (e.g., Koivisto 2006). Thus, once the functions f and b_i have been precomputed, all $W_{i,j}$ can be computed in time $O(2^d d^2)$.

The following recurrence equations are the key for efficient computation of the functions f and b_i; for proofs see the Supplement and also Tian and He (2009).

Lemma 2. *We have that* $f(\emptyset) = 1$ *and for any nonempty* U,

$$f(U) = \sum_{\emptyset \subset I \subseteq U} (-1)^{|I|-1} f(U \setminus I) \prod_{v \in I} \hat{w}_v(U \setminus I)\,. \quad (12)$$

Lemma 3. *We have that for all* $i \in V$ *and* $T \subseteq V \setminus \{i\}$,

$$b_i(T) = \sum_{I \subseteq T} (-1)^{|I|} g(T \setminus I) \prod_{v \in I} \hat{w}_v(V \setminus \{i\} \setminus T)\,, \quad (13)$$

where $g(\emptyset) = 1$ *and for any nonempty* T,

$$g(T) = \sum_{\emptyset \subset I \subseteq T} (-1)^{|I|-1} g(T \setminus I) \prod_{v \in I} \hat{w}_v(V \setminus T)\,. \quad (14)$$

The computational steps are summarized in Algorithm 1. Each of the five steps in Algorithm 1 requires $O(3^d d)$ additions and multiplications with a storage for $O(2^d d)$ numbers. The least straightforward is perhaps step 3, which can be implemented, e.g., as follows: for each $i \in V$ and $T \subseteq V \setminus \{i\}$, visit the subsets I of T in non-decreasing order by size, storing the product $\prod_{v \in I} \hat{w}_v(V \setminus \{i\} \setminus T)$ for each visited I (but reusing the memory for different i and T). We remark that, like step 1, also step 4 can in fact be implemented in $O(2^d d^2)$ time using fast (upward) zeta transform (Koivisto 2006).

Algorithm 1 Computing the unnormalized parent set and ancestor relation probabilities.

```
1: Compute the zeta transform $\hat{w}_v$ of the local weight function $w_v$ for each $v \in V$.
2: Compute the forward function $f$ and the auxilliary function $g$ using Eqs. (12) and (14).
3: Compute the backward function $b_i$ for each $i \in V$ using Eq. (13).
4: Compute the weight $W_i(S)$ for each $i \in V$ and $S \subseteq V \setminus \{i\}$ using Eq. (10).
5: Compute the weight $W_{i,j}$ for each pair $i, j \in V$ using Eq. (11).
```

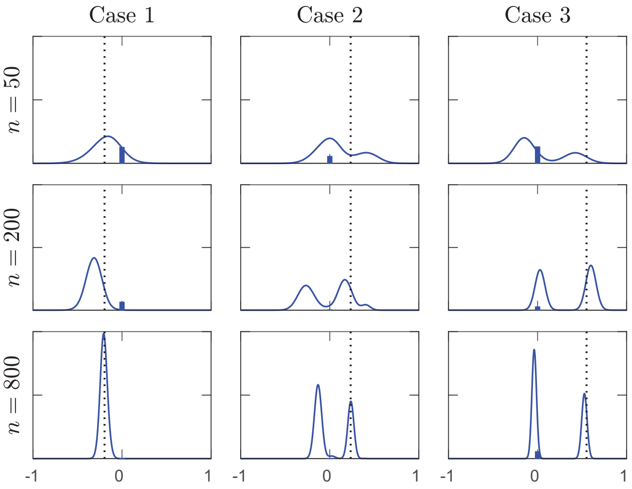

Figure 2: Posterior densities for three selected cause–effect pairs: the relative height of the bar located at zero represents the probability mass of the zero component and the true effects are shown by the vertical dotted lines.

Experiments

We have implemented Algorithm 1 in C++ and the rest of the method in R.[2] For a data set on 20 variables, the computations take about 25 minutes on a modern laptop computer (single thread, Intel Core i7-6600U, 2.60 GHz).

For calculating the parent set and ancestor relation probabilities, we limited the parent set size to 6 and used the fractional marginal likelihood (Consonni and Rocca 2012), with $\alpha_\Omega = d - 1$ and $n_0 = 1$, together with a uniform graph prior. The hyperparameters in the prior for the linear model (5) were set as follows: $m_0 = 0$, $\Lambda_0 = I$, and $a_0 = b_0 = 1$.

We evaluated the performance of our method by three empirical studies. The first study uses simulated data to examine the behaviour of our posterior and, in particular, the accuracy of the resulting estimates as compared to those of different IDA variants. In the second study, we assess the accuracy of our method as well as ancestor relation probabilities in terms of causal effect discovery. The third study demonstrates the applicability of our approach to real-world data.

Accuracy of the Causal Effect Estimates

We generated 50 random DAGs over $d = 20$ nodes with an expected neighbourhood size of four. The edge weights were sampled uniformly from $[-2, 2]$ and the the error terms variances were sampled uniformly from $[0.5, 1.5]$. For each model, we generated three data sets with increasing sample size, $n = 50, 200, 800$. The data were standardized to zero mean and unit variance (cf. Maathuis, Kalisch, and Bühlmann 2009, Assump. B). Analogously, the true causal effects were calculated from the corresponding standardized models.

[2] The code package and the supplementary material are available at https://github.com/jopensar/BIDA.

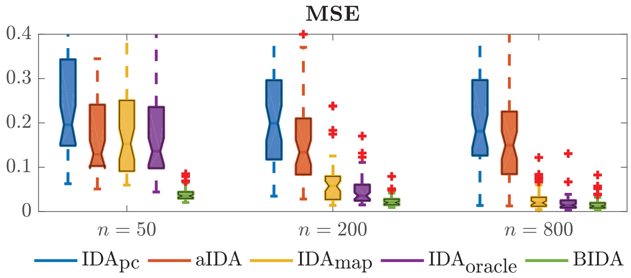

Figure 3: MSE of the different estimators in the simulation study.

In Figure 2, the posteriors are shown for three different cases in which the true effect is non-zero.In each case, we see that the posterior density mass converges around certain values as the sample size is increased. As explained earlier, these points will equal the true support, Θ_{ij}, as $n \to \infty$. In case 1, the causal effect is identifiable (unimodal density). In cases 2–3, the causal effect is non-identifiable (bimodal density). Still, in case 2, most of the density mass is shifted away from the immediate region around zero, indicating that the causal effect is non-zero.

We compared the accuracy of BIDA against several variants of the original IDA method. These variants differ by the way the CPDAG is obtained: IDA$_{\text{pc}}$ uses a CPDAG estimated by the PC algorithm (the original method), IDA$_{\text{map}}$ uses a maximum-a-posteriori CPDAG, and IDA$_{\text{oracle}}$ uses the true CPDAG (which is usually unavailable). In addition to these three IDA variants, we also included aIDA (Taruttis, Spang, and Engelmann 2015), which uses the PC-based IDA in combination with a resampling strategy to build a density over the causal effect. For implementation details and parameter settings, see the supplement.

We measured the accuracy of the methods by the mean squared error (MSE) between the true causal effects and the corresponding point estimates provided by the methods. To summarize the output of each method into a single point estimate, we calculated the mean of the output. Figure 3 shows that BIDA clearly outperformed its competitors in accuracy, especially for small sample sizes. As the sample size was increased, IDA$_{\text{map}}$ and IDA$_{\text{oracle}}$ approached a similar level of accuracy as BIDA. The PC-based IDA methods were clearly struggling in this experiment.

Discovering Non-zero Causal Effects

In previous works, the main target of the competing IDA-based methods has been to rank the causal effects in a system (Maathuis et al. 2010; Stekhoven et al. 2012; Taruttis, Spang, and Engelmann 2015). In line with this, we also compared how well the methods performed in terms of discovering non-zero causal effects, using the same experimental setup as in the previous section and ranking the effects by the mean (or minimum) absolute value. Since the interest in this experiment was solely the existence of a causal effect, we included our variant method for calculating ancestor relation probabilities (ARP), for which the effects were ranked by probability in a descending order. The performance was evaluated by the area under the precision-recall curve (AUC).

Again, BIDA outperformed the IDA-based methods (Fig. 2 in the supplement). Therefore, we focus our attention on comparing BIDA against ARP. Figure 4 (left) shows the results when a true positive was defined as a non-zero true effect, i.e., the existence of an ancestral path in the true model. In this setup, ARP is clearly more accurate than BIDA. However, when confining the set of true positives to only strong causal effects, here defined as $|\theta_{ij}| \geq 0.2$, BIDA outperformed ARP for the larger sample sizes (Fig. 4, right).

In summary, our variant for computing ancestor relation posteriors (ARP) is better at discovering non-zero causal relations (no matter the magnitude of the effect), while BIDA is better at discovering strong causal relations where the magnitude of the effect deviates markedly from zero and there is an evident response to the intervention.

Sachs Data

As an example of a possible application for our method, we consider the flow cytometry data (Sachs et al. 2005). The data collection contains abundance measurements of 11 biomolecules under various perturbation conditions. We focused on a data set of 853 observations, obtained under general stimulatory conditions (anti-CD3/CD28) and considered as passively observed. After log-transformation and standardization, we let BIDA rank the causal effects by their mean absolute value (Fig. S3). We examined the 13 highest ranked cause-effect pairs further (Table 1), as these were shown to deviate significantly from the rest using Tukey's outlier test (Fig. 3 in the supplement).

To assess the inferred cause–effect pairs, we first compared our results to the consensus network (Sachs et al. 2005). In total, 7 of the inferred pairs are supported by the consensus network. However, since there is much uncertainty about the true nature of consensus network, we also compared our results with those of a recent causal discovery method, called invariant causal prediction (ICP) (Peters, Bühlmann, and Meinshausen 2016; Meinshausen et al. 2016). In contrast to BIDA, ICP and hiddenICP (allows for hidden variables) are designed for interventional data and were given 7 interventional data sets from the Sachs collection, in addition to the observational data set. In total, 5/7 and 9/15 of the effects discovered by ICP and hiddenICP, respectively, were included among the discovered cause-effect pairs (Table 1), suggesting that many of the effects inferred by BIDA are supported by the interventional data.

Concluding Remarks

We presented a new, Bayesian method for estimating causal effects from passively observed data. To fully account for the uncertainty due to limited data, we integrated Bayesian linear regression and Bayesian model averaging (BMA) over graph structures. Our empirical results confirmed our hypothesis that the full Bayesian approach is superior to its rivals: the resulting posterior distributions of causal effects yield more accurate point estimates and improved accuracy in detecting causal relations, as compared to the state of the

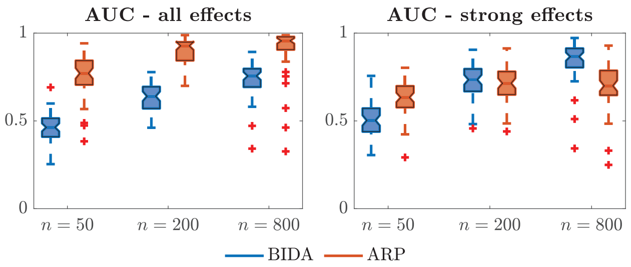

Figure 4: AUC for BIDA and our variant (ARP) in the simulation study when non-zero (left) and strong causal effects (right) in the true model were defined as true positives.

Table 1: Top cause-effect pairs discovered by BIDA for the Sachs data. A check mark means that the relation was included in the Sachs consensus network, detected by ICP, or detected by hiddenICP, respectively.

Cause	Effect	Mean abs.	Sachs	ICP	hiddenICP
AKT	ERK	0.54		✓	✓
PKC	P38	0.35	✓		✓
RAF	MEK	0.35	✓		✓
MEK	RAF	0.33			✓
ERK	AKT	0.28		✓	✓
AKT	PKA	0.25			
P38	PKC	0.23			✓
PIP3	PIP2	0.19	✓	✓	✓
ERK	PKA	0.16			
PIP2	PIP3	0.15	✓		
PKA	AKT	0.14	✓		✓
PKA	ERK	0.13	✓	✓	
PKC	JNK	0.13	✓	✓	✓

art. In particular, the comparison to estimates based on a single structure—found by greedy (Spirtes, Glymour, and Scheines 1993) or exact algorithms (Yuan and Malone 2013; Silander and Myllymäki 2006; Barlett and Cussens 2013)—showed that BMA is crucial for achieving the improved accuracy. The rankings of causal relations inferred from passively observed flow cytometry data concur to those previously inferred from experimental data, demonstrating the applicability of the method to real data.

Admittedly, BMA presents a computational challenge. We gave an *exact* algorithm that is able to compute the posterior of causal effects for data sets of realistic size, with up to around 20 variables. This scaling is similar to the scaling of other exact exponential algorithms for BMA over structures in graphical models, a topic of active research (see, e.g., Koivisto and Sood 2004; Tian and He 2009; Tian, He, and Ram 2010; Parviainen and Koivisto 2011; Chen and Tian 2014; Kangas, Niinimäki, and Koivisto 2015; Talvitie and Koivisto 2019). With a small modification to our approach for computing parent set posteriors, we obtained a variant that is currently the fastest known algorithm for computing ancestor relation posterior probabilities (Chen, Meng, and Tian 2015). It is worth noting that these two new algorithms apply to virtually any local model, e.g., to discrete and nonlinear models. An open question is whether *approximate* methods for BMA, e.g., based on Markov chain Monte Carlo (see Niinimäki, Parviainen, and Koivisto 2016, Kuipers and Moffa 2017, and references therein) or other approaches (Liao et al. 2019), can be employed to scale up the Bayesian approach, yet preserving the statistical efficiency now achieved with exact computation. Exact algorithms can be indispensable, however, when it is important to know whether the results of the computations are correct or within some tolerated error at least.

Finally, our present study suggests that causal effect estimation over more general model spaces—e.g., cycles, latent confounders, and nonlinear causal relations—would also likely benefit from a similar Bayesian approach. In this direction, Moffa et al. (2017) recently reported on a case study with a psychiatric data set over nine binary variables (allowing neither cycles nor latent confounders), demonstrating the value of Bayesian analysis in applications. How the modeling and computational issues are best addressed for such more general model spaces, is an intriguing question for future work.

Acknowledgements

This work was supported by the Academy of Finland, grants 295673 (A.H.) and 316771 (M.K.).

References

Barlett, M., and Cussens, J. 2013. Advances in Bayesian network learning using integer programming. In *Proc. UAI*.

Chen, Y., and Tian, J. 2014. Finding the k-best equivalence classes of Bayesian network structures for model averaging. In *Proc. AAAI*, 2431–2438. AAAI Press.

Chen, B.; Kumor, D.; and Bareinboim, E. 2017. Identification and model testing in linear structural equation models using auxiliary variables. In *Proc. ICML*.

Chen, Y.; Meng, L.; and Tian, J. 2015. Exact Bayesian learning of ancestor relations in Bayesian networks. In *Proc. AISTATS*.

Consonni, G., and Rocca, L. L. 2012. Objective Bayes factor for Gaussian directed acyclic graphical models. *Scand. J. Stat.* 39(4):743–756.

Entner, D.; Hoyer, P.; and Spirtes, P. 2013. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proc. AISTATS*.

Geiger, D., and Heckerman, D. 1994. Learning Gaussian networks. In *Proc. UAI*.

Geiger, D., and Heckerman, D. 2002. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* 30(5):1412–1440.

Greenland, S.; Robins, J. M.; and Pearl, J. 1999. Confounding and collapsibility in causal inference. *Statist. Sci.* 14(1):29–46.

Hoeting, J. A.; Madigan, D.; Raftery, A. E.; and Volinsky, C. T. 1999. Bayesian model averaging: A tutorial. *Statist. Sci.* 14(4):382–417.

Hyttinen, A.; Eberhardt, F.; and Järvisalo, M. 2015. Do-calculus when the true graph is unknown. In *Proc. UAI*.

Jaber, A.; Zhang, J.; and Bareinboim, E. 2018a. Causal identification under Markov equivalence. In *Proc. UAI*, 978–987.

Jaber, A.; Zhang, J.; and Bareinboim, E. 2018b. A graphical criterion for effect identification in equivalence classes of causal diagrams. In *Proc. IJCAI*.

Kangas, K.; Niinimäki, T. M.; and Koivisto, M. 2015. Averaging of decomposable graphs by dynamic programming and sampling. In *Proc. UAI*, 415–424.

Koivisto, M., and Sood, K. 2004. Exact Bayesian structure discovery in Bayesian networks. *J. Mach. Learn. Res.* 5:549–573.

Koivisto, M. 2006. Advances in exact Bayesian structure discovery in Bayesian networks. In *Proc. UAI*.

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

Kuipers, J., and Moffa, G. 2017. Partition MCMC for inference on acyclic digraphs. *J. Amer. Statist. Assoc.* 112:282–299.

Liao, Z. A.; Sharma, C.; Cussens, J.; and van Beek, P. 2019. Finding all Bayesian network structures within a factor of optimal. In *Proc. AAAI*, 7892–7899.

Maathuis, M. H.; Colombo, D.; Kalisch, M.; and Bühlmann, P. 2010. Predicting causal effects in large-scale systems from observational data. *Nature Methods* 7:247–248.

Maathuis, M. H.; Kalisch, M.; and Bühlmann, P. 2009. Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* 37(6A):3133–3164.

Malinsky, D., and Spirtes, P. 2017. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *Internat. J. Approx. Reason.* 88:371–384.

Meinshausen, N.; Hauser, A.; Mooij, J. M.; Peters, J.; Versteeg, P.; and Bühlmann, P. 2016. Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. USA* 113(27):7361–7368.

Moffa, G.; Catone, G.; Kuipers, J.; Kuipers, E.; Freeman, D.; Marwaha, S.; Lennox, B. R.; Broome, M. R.; and Bebbington, P. 2017. Using directed acyclic graphs in epidemiological research in psychosis: An analysis of the role of bullying in psychosis. *Schizophrenia Bulletin* 43(6):1273–1279.

Niinimäki, T.; Parviainen, P.; and Koivisto, M. 2016. Structure discovery in Bayesian networks by sampling partial orders. *J. Mach. Learn. Res.* 17:1–47.

Parviainen, P., and Koivisto, M. 2011. Ancestor relations in the presence of unobserved variables. In *Proc. ECML/PKDD (2)*, volume 6912 of *Lecture Notes in Computer Science*, 581–596. Springer.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Perković, E.; Textor, J.; Kalisch, M.; and Maathuis, M. H. 2018. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *J. Mach. Learn. Res.* 18(220):1–62.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 78(5):947–1012.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721): 523–529.

Shpitser, I., and Pearl, J. 2006. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proc. AAAI*, 1219–1226. AAAI Press.

Silander, T., and Myllymäki, P. 2006. A simple approach for finding the globally optimal Bayesian network structure. In *Proc. UAI*.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. Springer-Verlag. (2nd ed. MIT Press 2000).

Stekhoven, D. J.; Sveinbjörnsson, G.; Moraes, I.; Hennig, L.; Maathuis, M. H.; and Bühlmann, P. 2012. Causal stability ranking. *Bioinformatics* 28(21):2819–2823.

Talvitie, T., and Koivisto, M. 2019. Counting and sampling Markov equivalent directed acyclic graphs. In *Proc. AAAI*, 7984–7991.

Taruttis, F.; Spang, R.; and Engelmann, J. C. 2015. A statistical approach to virtual cellular experiments: improved causal discovery using accumulation IDA (aIDA). *Bioinformatics* 31(23):3807–3814.

Tian, J., and He, R. 2009. Computing posterior probabilities of structural features in Bayesian networks. In *Proc. UAI*.

Tian, J., and Pearl, J. 2002. A general identification condition for causal effects. In *Proc. AAAI/IAAI*, 567–573. AAAI Press / The MIT Press.

Tian, J.; He, R.; and Ram, L. 2010. Bayesian model averaging using the k-best Bayesian network structures. In *Proc. UAI*, 589–597. AUAI Press.

Tian, J. 2004. Identifying linear causal effects. In *Proc. AAAI*, 104–111. AAAI Press / The MIT Press.

van der Zander, B., and Liskiewicz, M. 2016. On searching for generalized instrumental variables. In *Proc. AISTATS*.

Wright, S. 1934. The method of path coefficients. *Ann. Math. Statist.* 5(3):161–215.

Yuan, C., and Malone, B. M. 2013. Learning optimal Bayesian networks: A shortest path perspective. *J. Artif. Intell. Res.* 48:23–65.