A Bayesian Approach for Estimating Causal Effects from Observational Data
Paper Supplement

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

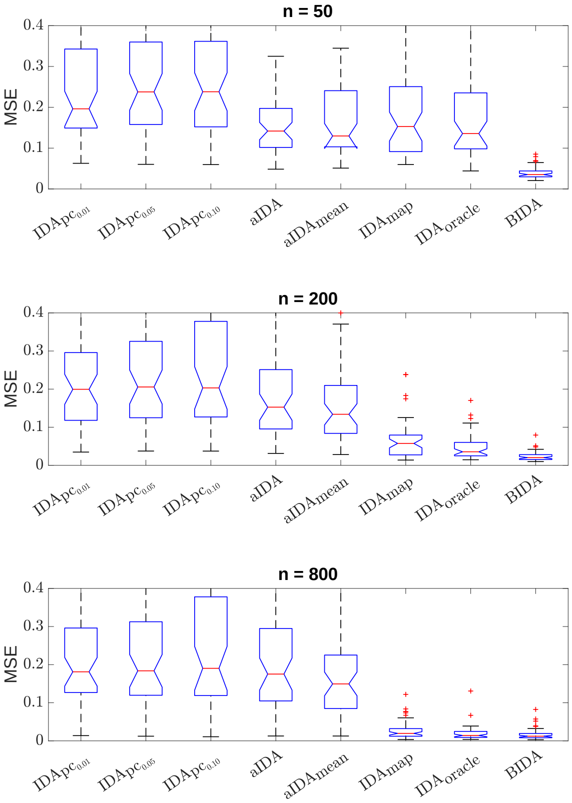

Figure 1: The effect of different p-value theresholds on MSE.

Implementation Details for the Simulations

For the score-based structure learning approaches, we used the fractional marginal likelihood ($\alpha_\Omega = d - 1, n_0 = 1$) together with a uniform graph prior, and the maximum parent set size was set to 6. The R package `pcalg` (Kalisch et al. 2012) was used to estimate the causal effects for the IDA methods, and for running the PC-algorithm (significance level: $\alpha = 0.01$, see Figures S1 and S3 for other values). The MAP structure was inferred by the dynamic programming algorithm in (Silander and Myllymäki 2006), as implemented in the MATLAB package `BDAGL` (available at https://www.cs.ubc.ca/~murphyk/Software/BDAGL/). For aIDA, we used the authors' code under default settings (Taruttis, Spang, and Engelmann 2015).

Further Experiments

Accuracy of the Causal Effect Estimates

Figure 1 shows further simulations for comparing MSE. Different p-value threshold for IDA do not improve MSE.

Discovering Non-zero Causal Effects

Figure 2 shows further results for the section 5.2, comparing the AUC of a precision recall curve.

Overall, the AUC results are in line with the MSE results, with our approach, BIDA, showing the highest accuracy for the smaller sample sizes, and IDA$_{\text{map}}$ and IDA$_{\text{oracle}}$ reaching a similar accuracy only for the largest sample size.

In more detail, the p-value threshold and the different ways of ranking the causal effects have only a slight influence on the accuracy of the methods. Using mean absolute value as the summary of the output gives similar and slightly more accurate result for most methods. In particular for IDA, using the minimum absolute value, as suggested by (Maathuis, Kalisch, and Bühlmann 2009), gives a slightly lower accuracy than using the mean absolute value. For aIDA, the authors' suggestion of ranking the causal effects by the absolute mode gives a slightly higher accuracy than the mean absolute value (Taruttis, Spang, and Engelmann 2015).

Sachs Data

Figure 3 shows the BIDA ranking of the cause-effect pairs for the Sachs data. The horizontal dotted line shows the upper outlier threshold given by Tukey's outlier test; $Q_3 + 1.5 \times IQR$.

Proofs

Proof of Lemma 2

Let $\mathcal{F}_v := \big\{G \in \mathcal{G}(U) : v \text{ has no child in } G\big\}$ for $v \in U$. We have that $\mathcal{G}(U) = \bigcup_{v\in U} \mathcal{F}_v$, since every DAG has at

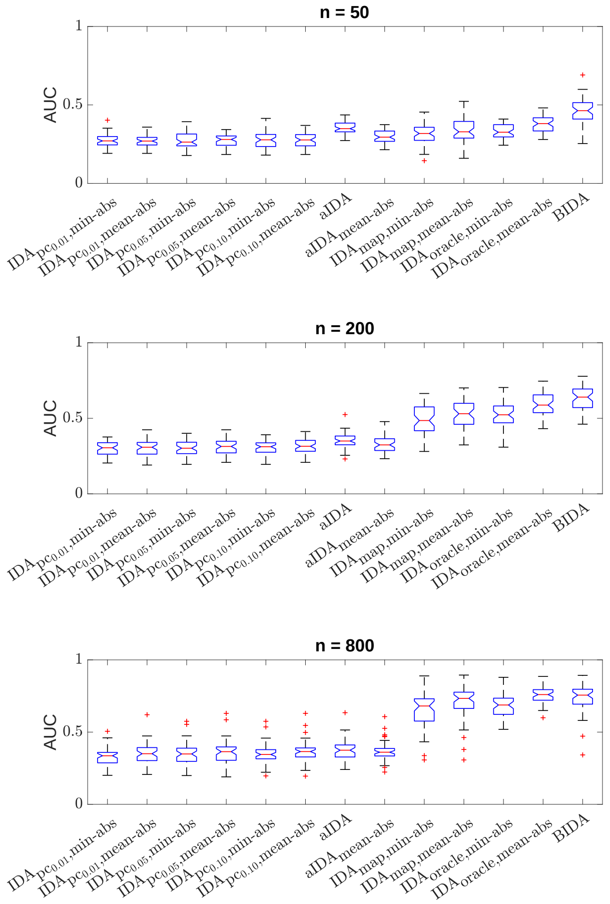

Figure 2: The effect of different summaries used in the ranking of causal effects.

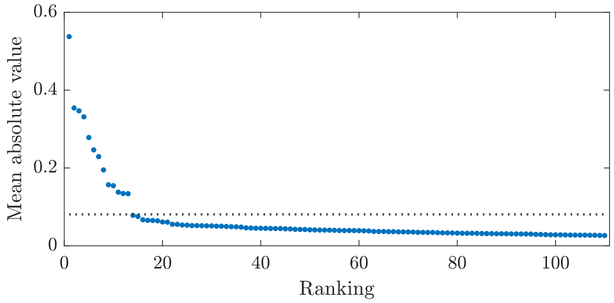

Figure 3: The BIDA ranking of the cause-effect pairs in the Sachs data.

least one sink node. Thus, by the principle of inclusion and exclusion,

$$f(U) = \sum_{\emptyset\subset I\subseteq U} (-1)^{|I|-1} \sum_{G\in\mathcal{F}_I} \prod_{v\in U} w_v(G_v) = \sum_{\emptyset\subset I\subseteq U} (-1)^{|I|-1} \sum_{G\in\mathcal{G}(U\setminus I)} \prod_{i\in U\setminus I} w_v(G_v) \prod_{v\in I} \hat{w}_v(U\setminus I)\,,$$

where $\mathcal{F}_I = \bigcap_{v\in I} \mathcal{F}_v$. Here we used the fact the nodes in I can choose any subset of parents from $U\setminus I$ independently of each other, while on the remaining nodes $U\setminus I$ we can have any DAG. Finally, apply the definition of $f(U\setminus I)$ to arrive at the claimed formula.

Proof of Lemma 3

We show first that, for a nonempty T,

$$g(T) = \sum_{G\in\mathcal{G}(T,V)} \prod_{v\in T} w_v(G_v)\,,$$

where $\mathcal{G}(T,V)$ consists of all $(G_v)_{v\in T}$ such that $G_v \subseteq V$ for each $v \in T$ and the induced subgraph $(G_v \cap T)_{v\in T}$ is acyclic. To this end it suffices to take the above formula as the definition of g and show that then g satisfies the recurrence given in Lemma 3.

Let $\mathcal{G}_v := \{G \in \mathcal{G}(T,V) : v \text{ has no parent in } T \text{ in } G\}$ for $v \in T$. We have that $\mathcal{G}(T,V) = \bigcup_{v\in T} \mathcal{G}_v$, since every DAG on T has at least one source node in T. Thus,

$$g(T) = \sum_{\emptyset\subset I\subseteq T} (-1)^{|I|-1} \sum_{G\in\mathcal{G}_I} \prod_{v\in T} w_v(G_v) = \sum_{\emptyset\subset I\subseteq T} (-1)^{|I|-1} \sum_{G\in\mathcal{G}(T\setminus I,V)} \prod_{v\in T\setminus I} w_v(G_v) \prod_{v\in I} \hat{w}_v(V\setminus T)\,,$$

where $\mathcal{G}_I = \bigcap_{v\in I} \mathcal{G}_v$. Here we used the fact the nodes in I can choose any subset of parents from $V\setminus T$ independently of each other, while on the remaining nodes $T\setminus I$ we can have any tuple of parent sets from $\mathcal{G}(T\setminus I, V)$. Finally, apply the (new) definition of $g(T\setminus I)$.

It remains to prove the formula for $b_i(T)$. Again, we apply the principle of inclusion and exclusion. Observe that $\mathcal{G}(i,T,V)$ consists of all members $(G_v)_{v\in T}$ in $\mathcal{G}(T,V)$ that satisfy the additional constraint that each G_v intersects $T \cup \{i\}$. Thus, letting $\mathcal{B}_v$ consist of all $G \in \mathcal{G}(T,V)$ such that G_v contains *no* node from $T\cup\{i\}$, i.e., $G_v \subseteq V\setminus\{i\}\setminus T$, we have that $\mathcal{G}(i,T,V) = \mathcal{G}(T,V) \setminus \bigcup_{v\in T} \mathcal{B}_v$. We get that

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

We see that this equals the claimed formula, recalling the convention that an empty product equals 1.

References

Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M. H.; and Bühlmann, P. 2012. Causal inference using graphical models with the R package pcalg. *J. Statist. Software* 47(11):1–26.

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

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

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.