Bayesian estimation of causal effects from observational categorical data

Vera Kvisgaard[1] **Johan Pensar**[1]

[1]Department of mathematics, University of Oslo , Oslo, Norway

Abstract

We develop a scaleable Bayesian method for the estimation of all pairwise causal effects in a system from observational data, under the assumption that the underlying causal model is an unknown discrete Bayesian network and that there are no latent confounders. Specifically, we build upon the Bayesian IDA (BIDA) and extend this method to the categorical setting. The key-idea is to combine Bayesian estimation of intervention distributions through the so-called backdoor formula with Bayesian model averaging. The main motivation of the method is to inform future experiments about plausible strong relationships, and we demonstrate by numerical experiments that our Bayesian modeling averaging approach can be highly relevant for this task.

1 BACKGROUND

The ambition to learn causal effects from observational data under an unknown causal structure can be approached by combining a structure learning procedure with a method for estimating causal effects given a structure. As a first scalable method in this direction, Maathuis et al. [2009] introduced the IDA algorithm. The IDA algorithm first infers a single CPDAG from the data and subsequently estimates the causal effects consistent with all DAGs in the associated equivalence class [Pearl, 2009]. Providing estimates of the bounds of the causal effects, the algorithm is proven useful for example for the purpose of ranking the pairwise effects. However, it is prone to errors in the single inferred causal structure. Several modifications of the algorithm have later been proposed to better account for the model uncertainty, including frequentist resampling strategies [Stekhoven et al., 2012, Taruttis et al., 2015] and Bayesian approaches [Castelletti and Consonni, 2021, Pensar et al., 2020, Viinikka et al., 2020, Kuipers and Moffa, 2022].

We build upon the Bayesian IDA (BIDA, Pensar et al. [2020]), a procedure that aims to account fully for the model uncertainty. The algorithm builds a posterior over the causal effect of every cause-effect pair in the considered system, by combining Bayesian estimation of the causal effects in a given DAG with Bayesian model averaging over DAGs. The main challenge with the Bayesian approach is computational. Averaging over all possible structures becomes infeasible even for smaller-scale systems, due to the vast model space. As in the original IDA procedure, the BIDA exploits the fact that when causal effects are estimated through the backdoor formula, the estimated effects are common to all DAGs that share the same adjustment set. The target posterior can therefore be consistently estimated by averaging over the unique adjustment sets, rather than unique DAGs. More precisely, let τ_{ij} denote the causal effect of variable X_i on X_j and let G_{ij} be the indices of a valid adjustment set w.r.t X_i, X_j and the graph G. The target posterior $p(\tau_{ij}|D)$ can then be computed as:

$$p(\tau_{ij}|D) = \sum_{G_{ij}} p(\tau_{ij}|G_{ij}, D)p(G_{ij}|D) \qquad (1)$$

where $p(\tau_{ij}|G_{ij})$ is the posterior conditional on the adjustment set being valid for adjustment and $p(G_{ij})$ the posterior probability that G_{ij} is indeed valid with respect to a predefined class of backdoor adjustment sets. In its original form, the BIDA algorithm relies on a dynamical programming routine that computes exact posterior probabilities over parent sets, and it is applicable to systems with up to around 25 nodes.

A key assumption in both the original and modified IDA algorithms is that the underlying model can be assumed to be linear Gaussian, which implies that the causal effects are linear and can conveniently be estimated using linear regression. Here, we assume the data are categorical, and extend the Bayesian IDA to this setting.

2 BAYESIAN IDA FOR CATEGORICAL VARIABLES

Consider a set of categorical variables $X_1, \ldots, X_n$ each with r_i possible outcomes and assume their joint distribution $P(x_1, \ldots, x_n|\theta, G)$ can be modeled by a causal Bayesian network, such that the parameters $\theta = \{\theta_i\}_{i=1}^n$ is a set of conditional probability tables θ_i, one for each node i. Our goal then is to estimate the marginal intervention probabilities

$$\pi_{x_j|x_i} = P(X_j = x_j|do(X_i = x_i), \theta)$$

and the associated causal effects τ_{ij} for every cause-effect pair (i, j), given data sampled from P with both θ and G unknown. To quantify the causal effect between categorical variables, we use the Jensen-Shannon-divergence between intervention distributions [Lin, 1991, Griffiths et al., 2015]:

$$\tau_{ij} = \frac{1}{r_i} \sum_{x_i} \sum_{x_j} \pi_{x_j|x_i} \log \frac{\pi_{x_j|x_i}}{\frac{1}{r_i} \sum_{x_i} \pi_{x_j|x_i}}.$$

We develop a Bayesian estimator for the target posterior $p(\tau_{ij}|D)$ by first considering the setting where we are given a causal DAG G, then the setting where G is not known.

2.1 POSTERIOR UNDER KNOWN DAG

If the causal DAG is known and there are no unobserved confounders, interventional distributions can be estimated through the backdoor formula. Suppose $X_{G_{ij}}$ is a set of valid adjustment variables. The backdoor formula then states that [Pearl, 2009],

$$\pi_{x_j|x_i} = \begin{cases} P(x_j) & \text{if } j \in G_{ij} \\ P(x_j|x_i, x_{G_{ij}}, \theta)P(x_i|x_{G_{ij}}, \theta) & \text{otherwise} \end{cases}$$

The right-hand side of the formula includes only the observational distribution, and can consistently from non-experimental data. We estimate the involved marginal distributions directly, and thereby avoid inference in the full model, which is typically computationally demanding in the categorical setting. In particular, we assume a conjugate Dirichlet prior over the parameters of the relevant marginal distributions. Each row in the interventional probability table $\pi_{ij} = \{\{\pi_{x_j|x_i}\}_{x_j=1}^{r_j}\}_{x_i=1}^{r_i}$ is then a linear combination of Dirichlet vectors, unless the adjustment set is empty. Moreover, the row-vectors will generally be dependent. While there are no closed-form expression for the exact posterior distribution $p(\pi_{ij}|G_{ij}, D)$ and $p(\tau_{ij}|G_{ij}, D)$, they can easily be sampled from through Monte Carlo sampling.

2.2 POSTERIOR UNDER AN UNKNOWN DAG

Under model uncertainty, the target posterior $p(\tau_{ij}|D)$ can be computed by equation (1), given the local estimator for the conditional posterior $p(\tau_{ij}|G_{ij}, D)$ outlined in the previous section. To scale up the procedure, we compute the posterior probability $P(G_{ij}|D)$ of G_{ij} being a valid adjustment set by the means of MCMC. In particular, we employ the partition-MCMC scheme [Kuipers et al., 2022] and sample a sequence of DAGs from the graph posterior. The posterior probability $P(G_{ij}|D)$ can then be approximated simply by counting the number of sampled DAGs in which G_{ij} is indeed a valid adjustment set with respect to a specific class of adjustment sets (e.g. a parent set or o-set). Where the current exact BIDA is limited to the use of parent sets for adjustment, this MCMC approach also allows us to consider various non-local backdoor adjustment sets identifiable through graphical criteria. Specifically, we consider the o-set [Henckel et al., 2022, Witte et al., 2020] as well as the minimal o-set and minimal parent set. The minimal sets are obtained by removing the variables that do not close any backdoor paths from the respective adjustment set.

3 NUMERICAL EXPERIMENTS

We implemented the proposed BIDA method in `R` and evaluated its performance in a simulation experiment, on data sampled from 9 discrete Bayesian networks of various sizes from the bnlearn repository. The results show that the use of non-local adjustment sets improves the accuracy of the point-estimates over parent sets, also when the true DAG is not known (Figure 1). Compared with two variants of the IDA method, the original procedure as well as the optimal IDA [Witte et al., 2020], our proposed method is overall the most accurate both in terms of point-estimates (Figure 2) and in terms of discovering strong effects (Figure 3). Still, there are some particular networks where both the absolute and relative performance of our method is rather poor. The poor results are explained by the MCMC-chain that has trouble mixing given data from these networks.

4 CONCLUSION

We have extended the Bayesian IDA to the categorical setting. Additionally, using MCMC methods in the space of DAGs, we have both scaled up the method and improved its accuracy by replacing parent sets with non-local adjustment sets. A good approximation of the posterior over adjustment sets is crucial for the method, and delving deeper into the structure learning part of the method is a natural way to improve upon the current version. To further scale up the method, some approximation of the conditional posterior distribution is required, as sampling from the exact posterior distribution is computationally costly.

References

Federico Castelletti and Guido Consonni. Bayesian inference of causal effects from observational data in Gaussian graphical models. *Biometrics*, 77(1):136–149, March 2021. ISSN 15410420. doi: 10.1111/BIOM.13281.

Paul E. Griffiths, Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight. Measuring Causal Specificity. *Philosophy of Science*, 82(4):529–555, October 2015. ISSN 0031-8248, 1539-767X. doi: 10.1086/682914.

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical Criteria for Efficient Total Effect Estimation Via Adjustment in Causal Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, April 2022. ISSN 1369-7412. doi: 10.1111/rssb.12451.

Jack Kuipers and Giusi Moffa. The interventional Bayesian Gaussian equivalent score for Bayesian causal inference with unknown soft interventions, May 2022.

Jack Kuipers, Polina Suter, and Giusi Moffa. Efficient Sampling and Structure Learning of Bayesian Networks. *Journal of Computational and Graphical Statistics*, 31(3):639–650, July 2022. ISSN 1061-8600. doi: 10.1080/10618600.2021.2020127.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. ISSN 1557-9654. doi: 10.1109/18.61115.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009. ISSN 0090-5364. doi: 10.1214/09-AOS685.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. ISSN 19357516. doi: 10.1214/09-SS057.

Johan Pensar, Topi Talvitie, Antti Hyttinen, and Mikko Koivisto. A Bayesian Approach for Estimating Causal Effects from Observational Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5395–5402, 2020. ISSN 2374-3468. doi: 10.1609/AAAI.V34I04.5988.

Daniel J. Stekhoven, Izabel Moraes, Gardar Sveinbjörnsson, Lars Hennig, Marloes H. Maathuis, and Peter Bühlmann. Causal stability ranking. *Bioinformatics*, 28(21):2819–2823, November 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts523.

Franziska Taruttis, Rainer Spang, and Julia C. Engelmann. A statistical approach to virtual cellular experiments: Improved causal discovery using accumulation IDA (aIDA). *Bioinformatics*, 31(23):3807–3814, December 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv461.

Jussi Viinikka, Antti Hyttinen, Johan Pensar, and Mikko Koivisto. Towards scalable Bayesian learning of causal DAGs. *Advances in Neural Information Processing Systems*, 33:6584–6594, 2020. ISSN 10495258.

Janine Witte, Leonard Henckel, Marloes H Maathuis, and Vanessa Didelez. On Efficient Adjustment in Causal Graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.

A SELECTED SIMULATION RESULTS

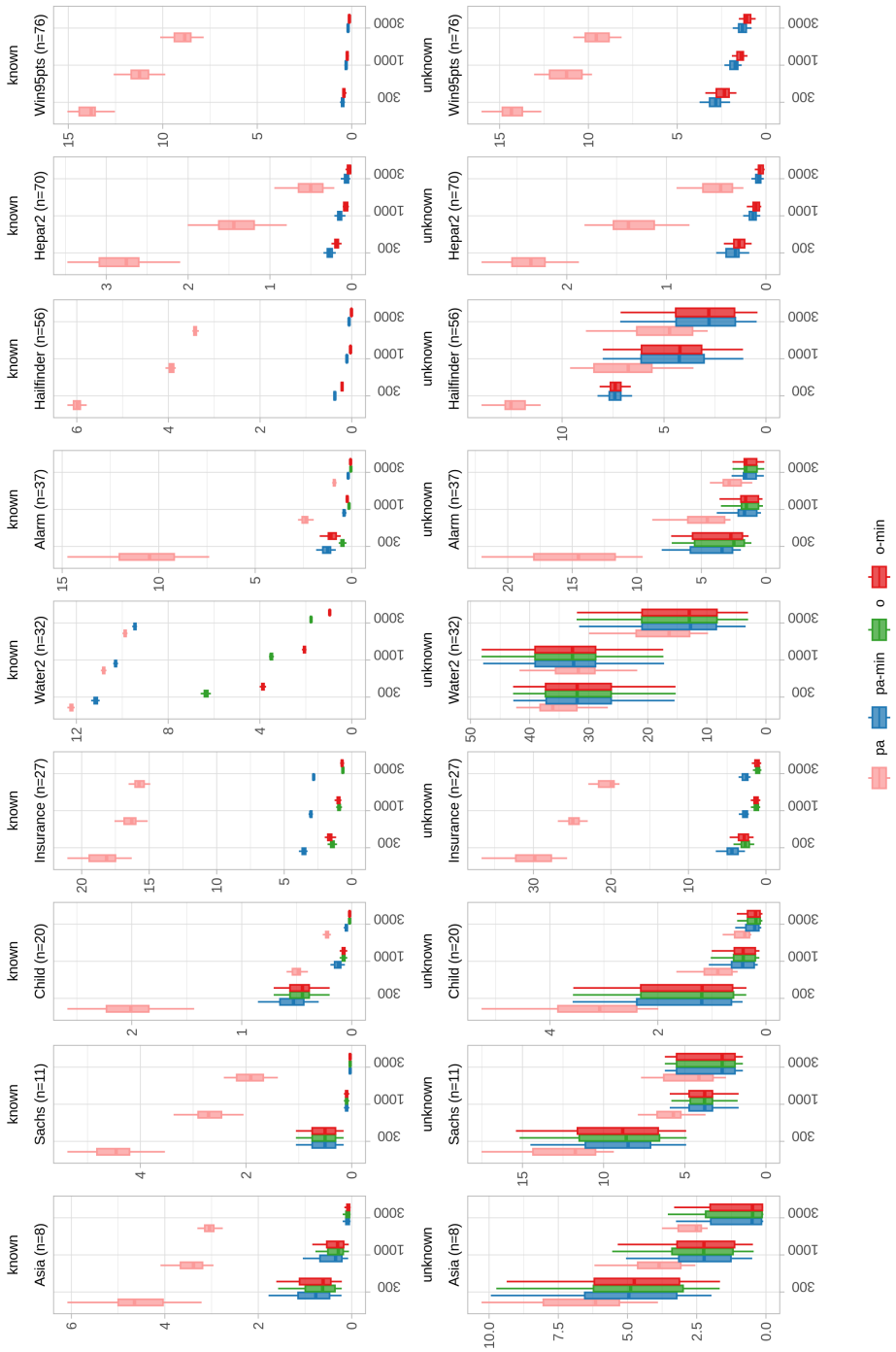

Figure 1: The point-estimates of our proposed BIDA method using different adjustment sets for adjustment: parent sets (pa), minimal parent sets (pa-min), o-sets (o) or minimal o-sets (o-min), as measured by the mean squared errors (MSE) between the true and estimated causal effects, τ. Each subfigure corresponds to one network and the boxplot shows for each sample size the distribution over 30 independently sampled data sets. Outliers are excluded. In the upper row the DAG is assumed known, in the bottom unknown.

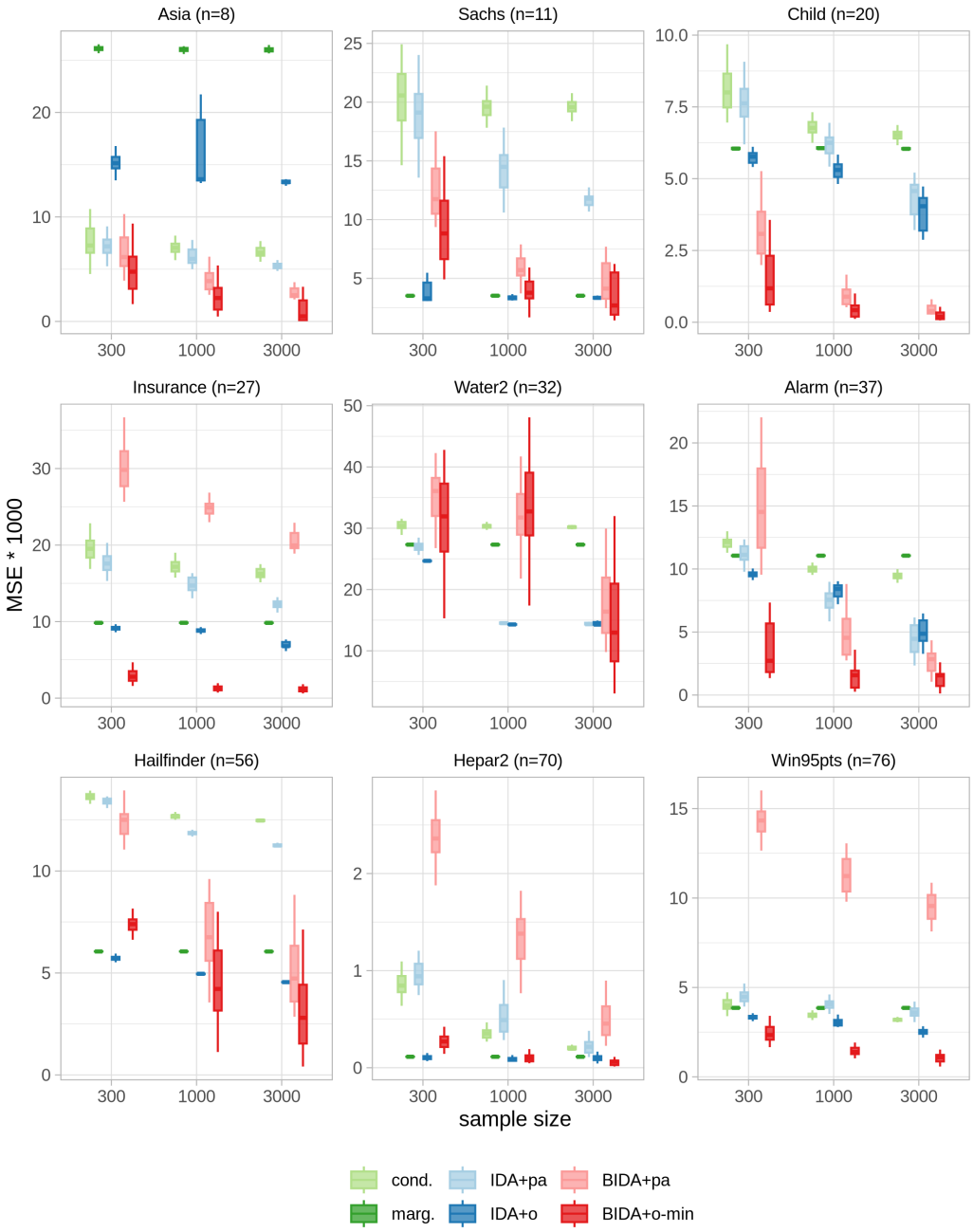

Figure 2: The accuracy of BIDA using parent sets (BIDA+pa) or minimal o-sets (BIDA+o-min) for adjustment, as measured by the mean squared errors (MSE) between the true and estimated causal effects, τ. Compared to the estimated marginal probabilities (marg.), conditional probabilities (cond.), the original IDA (IDA+pa) and the optimal IDA (IDA+o). Each subfigure corresponds to one network and the boxplot shows for each sample size the distribution over 30 independently sampled data sets. Outliers are excluded.

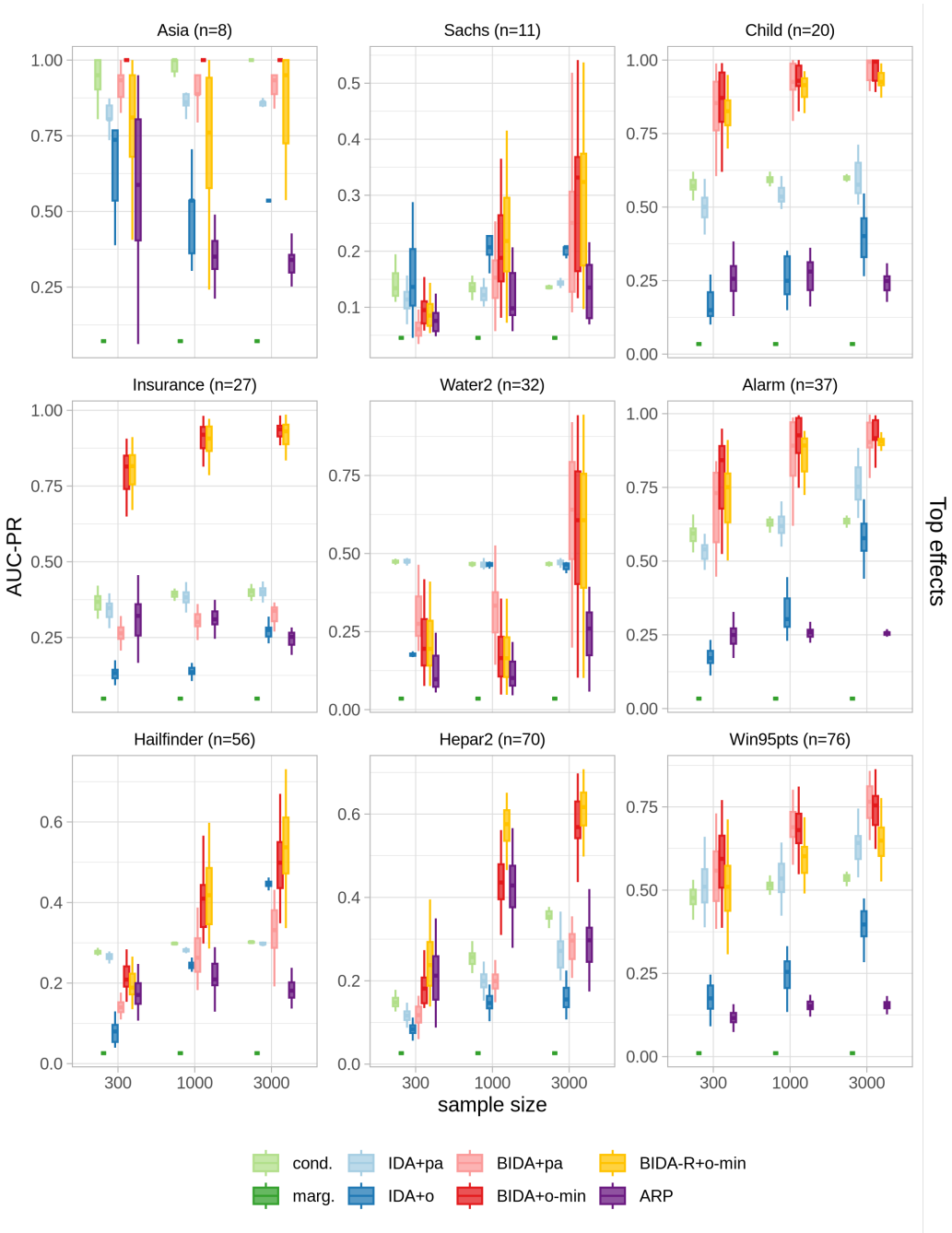

Figure 3: The accuracy in predicting the strongest effects (the top 20 percentile of positive effects in each network) using the mean values (BIDA+o-min) and mean ranks (BIDA-R+o-min) from the BIDA posterior, as measured by the area under the precision-recall curve (AUC-PR). Compared to the estimated marginal probabilities (marg.), conditional probabilities (cond.), the original IDA (IDA+pa), the optimal IDA (IDA+o) and the ancestor relation probabilities (ARP). Each subfigure corresponds to one network and the boxplot shows for each sample size the distribution over 30 independently sampled data sets. Outliers are excluded.