The interventional Bayesian Gaussian equivalent score for Bayesian causal inference with unknown soft interventions

Jack Kuipers JACK.KUIPERS@BSSE.ETHZ.CH
D-BSSE, ETH Zurich, Basel, Switzerland

Giusi Moffa GIUSI.MOFFA@UNIBAS.CH
Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland

Editors: Biwei Huang and Mathias Drton

Abstract

Describing the causal relations governing a system is a fundamental task in many scientific fields, ideally addressed by experimental studies. However, obtaining data under intervention scenarios may not always be feasible, while discovering causal relations from purely observational data is notoriously challenging. In certain settings, such as genomics, we may have data from heterogeneous study conditions, with soft (partial) interventions only pertaining to a subset of the study variables, whose effects and targets are possibly unknown. Combining data from experimental and observational studies offers the opportunity to leverage both domains and improve the identifiability of causal structures. To this end, we define the interventional BGe score for a mixture of observational and interventional data for linear-Gaussian models, where the targets and effects of intervention may be unknown. Prerogative of our method is that it takes a Bayesian perspective leading to a full characterisation of the posterior distribution of the DAG structures. Given a sample of DAGs, one can also automatically derive full posterior distributions of the intervention effects. Consequently, the method effectively captures the uncertainty both in the structure and the parameter estimates. We additionally demonstrate the performance of the approach both in simulations and data analysis applications. Codes to reproduce the simulations and analyses are publicly available at https://github.com/jackkuipers/iBGe.

Keywords: Graphical models, Bayesian networks, Directed acyclic graphs, Bayesian scores, Structure learning, Causal inference, Interventional data.

1. Introduction

Understanding and predicting the consequences of an action is the ultimate goal of investigation in many scientific disciplines. Questions about the effect of intervention are causal in nature and require understanding the causal relations between the variables under study. Directed acyclic graphs (DAGs) are convenient tools for representing causal mechanisms and help estimate intervention effects (Pearl, 1995; Greenland et al., 1999; Pearl, 2000; Spirtes et al., 2000). Gold standard methods for establishing the effect of an intervention rely on randomised studies. In reality, ethical, financial or practical difficulties may stand in the way of effectively implementing experimental studies. In scenarios where trials are not an option but we have sufficient expert knowledge to draw a causal diagram, one may use 'do' calculus (Pearl, 2000) to evaluate the effect of potential interventions.

In the absence of sufficient prior knowledge, we need strategies that allow us to gain insights into a causal mechanism from observational data. Causal discovery, however, relies on very strict assumptions, especially causal sufficiency, which we also assume here. Furthermore, even under the assumption of no unmeasured confounders, observational data may only ever identify causal

graphical structures up to a Markov equivalence class, also known as essential graphs (EGs; Andersson et al., 1997) or completed partially DAGs (CPDAGs; Chickering, 2002). Methods limited to EGs may not be entirely satisfactory to fully characterise a causal mechanism. To resolve the uncertainty between equivalent DAGs we either need additional assumptions on the structural equations governing the relationships between variables, or we need to perform experiments to generate and collect interventional data.

Since in practice it may only be possible to perform experiments on a subset of the variables, an appealing strategy is combining observational and interventional data to improve the identifiability of causal structures. To extend existing Bayesian methods for structure learning and estimation of intervention effects to deal with a mix of observational and interventional data we need to define a (marginalised likelihood) score which accounts for the mixed nature of the data. Given a suitable score we can use recent methods (Kuipers et al., 2022; Viinikka et al., 2020) to efficiently sample from the posterior distribution of DAGs given the data. The procedure can provide a MAP (Maximum a Posteriori) estimator if of interest, but more importantly, by characterising the posterior distribution of structures it naturally accounts for the uncertainty in the structure learning task.

Early work to combine observational and interventional data for Bayesian structure learning appears in Heckerman (1995) for deterministic interventions, and extended to the more general case of non-deterministic manipulations in Cooper and Yoo (1999). Handling deterministic structural (perfect or hard as per the definition in Eberhardt and Scheines, 2007) interventions is relatively straightforward since the likelihood of the data of the intervened upon variables is simply 1 for the set value and we can ignore them when scoring each corresponding node as a child (Cooper and Yoo, 1999). When an intervened-upon node acts as a parent, the intervention plays no role in the scoring of downstream nodes since the contribution to the score of each node is defined in terms of its conditional probability given the parents. The interventions, however, may disrupt the Gaussianity assumption used for example in the GIES (Greedy Interventional Equivalence Search; Hauser and Bühlmann, 2012); an algorithm developed to perform penalised maximum likelihood-based inference of causal structures from mixed observational and interventional data. An interesting line of recent developments by Castelletti and Peluso (2023) presents a Bayesian approach for Gaussian DAGs where the targets may be unknown but the interventions remain perfectly effective.

In many practical applications, such as genomic studies, imperfect interventions which only partially succeed (*i.e.* soft interventions), or succeed a fraction of the time with a certain probability (*i.e.* stochastic interventions), are not uncommon. For stochastic interventions, one can use a mixture model (Korb et al., 2004) with a certain probability ρ of a successful structural intervention severing all links into the intervened upon node and probability $(1 - \rho)$ that the intervention is not successful and the unperturbed network still describes the data generating mechanism.

Soft interventions, where both the strength and the targets of the intervention may be unknown, constitute a more general and realistic setting. This work tackles this setting for linear-Gaussian continuous data to enable Bayesian inference of effects.

In a discrete data scenario, one can represent interventions as additional nodes in the network (with no parents) and learn the targets by inferring their connections (Eaton and Murphy, 2007). In the soft interventional setting, representing the intervention as an additional parent amounts to modifying the relationship between the intervened upon node and the other parents (differently for each discrete parent state). The BDe score (Heckerman and Geiger, 1995) is fully parametrised (for a given set of parents including interventions) so it automatically includes all interaction terms between the added intervention nodes and the other parents. This framework, therefore, provides

a very general model of how an intervention may affect a node, covering the gamut from hard to soft interventions, though at the cost of increasing the parameter space. Learning the connections between the added intervention nodes and other nodes also allows us to capture any uncertainty in the targets (Eaton and Murphy, 2007). Compared to alternatives, the BDe score has the advantage of enabling a Bayesian approach.

More recent algorithms handling soft interventions, and also extended to deal with continuous data, include the IGSP (Interventional Greedy Sparsest Permutation) algorithm of Wang et al. (2017) for known targets, a hybrid method with the score function defined in terms of conditional independence tests and structure learning with an order-based search, and the UT-IGSP (Unknown Target IGSP) version with unknown targets (Squires et al., 2020). Variational inference approaches to approximate Bayesian inference have also been developed, including with unknown interventions (Hägele et al., 2023), though these do not explicitly define the prior and posterior, nor use analytical marginalisation to speed up inference.

With the current work, we aim to bring the intrinsic flexibility of the discrete setting using the BDe score to scenarios with continuous data, developing efficient marginalised scores through suitably adapting the BGe score (Geiger and Heckerman, 2002). Since we are modelling continuous data, we cannot simply include discrete interventions as additional parents as in the BDe case. However, by leveraging the natural interpretation of interventions as interactions in the discrete setting and extending the strategy to continuous data we can derive a simple and effective interventional BGe (iBGe) score accounting for intervention nodes. The modified score enables scalable and accurate causal inference for continuous observational and interventional data in the presence of soft interventions with possibly unknown targets.

2. The interventional BGe score

2.1. BGe score recap

Consider an n-dimensional vector of random variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$ and a dataset $d = \{\boldsymbol{x}_1, \dots \boldsymbol{x}_N\}$ with N complete observations of the vectors $\boldsymbol{x}_i$. One common strategy for DAG models describing the independence relationships that hold in the data distribution is to define a scoring function. Important results by Geiger and Heckerman (2002) specify the conditions under which the marginal likelihood factorises into components of each node given its parents:

$$\begin{aligned} p(d \mid m^h) = \int p(d \mid \Theta, m^h) p(\Theta \mid m^h) \mathrm{d}\Theta &= \prod_{j=1}^{n} \int p(d^{X_j} \mid d^{\boldsymbol{P}_j}, \theta_j, m^h) p(\theta_j \mid m^h) \mathrm{d}\theta_j \\ &= \prod_{j=1}^{n} \frac{p(d^{X_j \cup \boldsymbol{P}_j} \mid m^h)}{p(d^{\boldsymbol{P}_j} \mid m^h)} \end{aligned} \tag{1}$$

where m^h refers to a model hypothesis for the true distribution of $\boldsymbol{X}$ to be faithful to model m, $d^{\boldsymbol{Y}}$ is the data restricted to the coordinates in $\boldsymbol{Y} \subseteq \boldsymbol{X}$, Θ is the collection of all parameters in the model, $\boldsymbol{P}_j$ are the parent variables of the vertex j and θ_j are the parameters determining the conditional distribution of that node given its parents.

The fundamental assumptions behind the decomposition are

i. *complete model equivalence* where we cannot distinguish between complete DAG models on the basis of observational data;

ii. *regularity* by which there is a one-to-one mapping between parameterisations of two complete models;

iii. *likelihood modularity* by which the local distribution of each variable only depends on its parents in the graph and it is the same for two models where the given variable has the same parents;

iv. *prior modularity* which is the same property for the priors; and

v. *global parameter independence* by which the parameter prior factorises into a product of components, one for each node.

Together these ensure that the factorisation property of Bayesian networks carries over to the marginal likelihood as in Equation (1).

For nominal categorical data, a multinomial Dirichlet prior is required to meet all conditions and it leads to the BDe score (Heckerman and Geiger, 1995). Joint Gaussian data require a normal-Wishart prior to satisfy all assumptions (Geiger and Heckerman, 2002), leading to the BGe score which is the posterior probability of m^h, proportional to the marginal likelihood above and the prior on graphs. Since the score factorises, we focus on a single node X with parents $\boldsymbol{P}$. The likelihood for the data $d^{X\cup\boldsymbol{P}}$ consisting of N observations $(x_i, \boldsymbol{p}_i), i = 1 \dots N$ is

$$p(d^{X\cup\boldsymbol{P}} \mid \boldsymbol{\mu}, W, m^h) \quad = \quad \frac{|W|^{\frac{N}{2}}}{(2\pi)^{\frac{(p+1)N}{2}}} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{N}[\boldsymbol{\mu}-(x_i,\boldsymbol{p}_i)]^{\mathrm{T}} W [\boldsymbol{\mu}-(x_i,\boldsymbol{p}_i)]} \tag{2}$$

following from the assumption of a jointly Gaussian distribution, with W the precision matrix, $\boldsymbol{\mu}$ the mean and p the number of parents.

By placing the conjugate Wishart prior on the full $n \times n$ precision matrix, $\widetilde{W} \sim \mathcal{W}_n(T^{-1}, \alpha_w)$, where $\alpha_w > n-1$ indicates the degrees of freedom and T is the positive definite parametric matrix, and a normal prior on the full mean vector $\widetilde{\boldsymbol{\mu}}$ with mean $\boldsymbol{\nu}$ and precision matrix $\alpha_\mu \widetilde{W}$, with $\alpha_\mu > 0$, the posterior distribution of $\widetilde{W}$ and $\widetilde{\boldsymbol{\mu}}$ are also normal-Wishart with updated parameters

$$\alpha_\mu \to N + \alpha_\mu \,, \qquad \alpha_w \to N + \alpha_w \,, \qquad \boldsymbol{\nu} \to \boldsymbol{\nu}' \,, \qquad T \to R \tag{3}$$

where

$$\boldsymbol{\nu}' = \frac{N\bar{\boldsymbol{x}} + \alpha_\mu \boldsymbol{\nu}}{(N + \alpha_\mu)} \,, \qquad R = T + S_N + \frac{N\alpha_\mu}{(N + \alpha_\mu)} (\bar{\boldsymbol{x}} - \boldsymbol{\nu}) (\bar{\boldsymbol{x}} - \boldsymbol{\nu})^{\mathrm{T}} \tag{4}$$

and

$$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \,, \qquad S_N = \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} \tag{5}$$

as detailed in Geiger and Heckerman (2002); Kuipers et al. (2014). To compute the score for each node X, all that matters is the node itself and its parents with the contribution of

$$\mathrm{BGe}(d, X) = \frac{p(d^{\boldsymbol{Y}} \mid m^h)}{p(d^{\boldsymbol{P}} \mid m^h)} = \left(\frac{\alpha_\mu}{N + \alpha_\mu}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{N+\alpha_w-n+p+1}{2}\right)}{\pi^{\frac{N}{2}} \Gamma\left(\frac{\alpha_w-n+p+1}{2}\right)} \frac{|T_{\boldsymbol{YY}}|^{\frac{\alpha_w-n+p+1}{2}} |R_{\boldsymbol{PP}}|^{\frac{N+\alpha_w-n+p}{2}}}{|T_{\boldsymbol{PP}}|^{\frac{\alpha_w-n+p}{2}} |R_{\boldsymbol{YY}}|^{\frac{N+\alpha_w-n+p+1}{2}}} \tag{6}$$

to the marginal likelihoods, where $\boldsymbol{Y} = X \cup \boldsymbol{P}$, Γ is the Gamma function and $A_{\boldsymbol{YY}}$ means selecting the rows and columns corresponding to $\boldsymbol{Y}$ of a matrix A.

2.2. SEM interpretation

Along with the matrix version, we can reformulate the conditional distribution of an arbitrary node X on its parents $\boldsymbol{P}$ in the Structural Equation Model (SEM) interpretation. If the matrix B stores the edge weights of the DAG then the precision matrix is given by $\widetilde{W} = (1-B)D(1-B)^{\mathrm{T}}$ where D is a diagonal matrix of inverse variances. For the BGe score setting with a normal-Wishart prior on $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{W}$, Viinikka et al. (2020) explore in detail the expression of the posterior distribution of the edge weights ensuing from the SEM reparametrisation and the consequent estimation of the causal effects.

In the absence of intervention, the structural equation at each node takes the form

$$X = \alpha + \boldsymbol{\beta} \cdot \boldsymbol{P} + \epsilon\,, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

where $\boldsymbol{\beta} = B_{\boldsymbol{P},X}$ and $\sigma^2 = D^{-1}_{XX}$. The likelihood for the observed data is simply the 1d Gaussian

$$p(d^X \mid d^{\boldsymbol{P}}, \theta, m^h) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[x_i - \alpha - \boldsymbol{\beta}\cdot\boldsymbol{p}_i]^2} \tag{8}$$

where θ collects the parameters $\alpha, \boldsymbol{\beta}$ and σ. To compute the marginal likelihood and integrate over θ we can avoid the exact mapping to the normal-Wishart space by returning to the last step of Equation (1)

$$\begin{aligned} \int p(d^X \mid d^{\boldsymbol{P}}, \theta, m^h) p(\theta \mid m^h) \mathrm{d}(\theta) &= \int \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[x_i - \alpha - \boldsymbol{\beta}\cdot\boldsymbol{p}_i]^2} p(\theta \mid m^h) \mathrm{d}(\theta) \\ &= \frac{p(d^{X\cup\boldsymbol{P}} \mid m^h)}{p(d^{\boldsymbol{P}} \mid m^h)} = \mathrm{BGe}(d, X) \end{aligned} \tag{9}$$

utilising the simplification afforded by the prior choice and marginal likelihood factorisation.

2.3. Soft interventions as interactions

To consider soft interventions we return to the idea of Eaton and Murphy (2007) of including them as additional parent nodes. In the discrete setting, the scoring function automatically accounts for interactions between the intervention node and the other parents. For the continuous Gaussian case, we wish to mimic the same structure, and define the model in an analogous way. In particular, given a mixture of observational and interventional data, we can view the intervention as another binary parent node I and include an interaction term in the SEM

$$X = \alpha + \boldsymbol{\beta} \cdot \boldsymbol{P} + \widetilde{\alpha}_I I + \widetilde{\boldsymbol{\beta}}_I \cdot \boldsymbol{P} I + \epsilon(I) \tag{10}$$

If we re-parameterise the regression coefficients ($\boldsymbol{\beta}_I = \widetilde{\boldsymbol{\beta}}_I + \boldsymbol{\beta}$, $\alpha_I = \widetilde{\alpha}_I + \alpha$) we can rewrite as

$$X = \begin{cases} \alpha + \boldsymbol{\beta} \cdot \boldsymbol{P} + \epsilon & \text{for } I = 0 \\ \alpha_I + \boldsymbol{\beta}_I \cdot \boldsymbol{P} + \epsilon_I & \text{for } I = 1 \end{cases} \tag{11}$$

The above SEM representation implies that the conditional likelihoods of node X in the observed data d and intervened data d_I takes the form

$$p(\widetilde{d}^X \mid \widetilde{d}^{\boldsymbol{P}}, \theta, m^h) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[x_i - \alpha - \boldsymbol{\beta}\cdot\boldsymbol{p}_i]^2} \frac{1}{(2\pi\sigma_I^2)^{\frac{N_I}{2}}} \mathrm{e}^{-\frac{1}{2\sigma_I^2}\sum_{i=N+1}^{N+N_I}[x_i - \alpha_I - \boldsymbol{\beta}_I\cdot\boldsymbol{p}_i]^2} \tag{12}$$

In the case that the intervention may change all the parameters in the SEM determining node X we can easily define the marginal likelihood contribution to the interventional BGe score

$$\begin{aligned}
\text{iBGe}(\widetilde{d}, X) &= \int p(\widetilde{d}^X \mid \widetilde{d}^{\boldsymbol{P}}, \theta, \theta_I m^h) p(\theta, \theta_I \mid m^h) \mathrm{d}(\theta) \\
&= \int \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[x_i - \alpha - \boldsymbol{\beta}\cdot\boldsymbol{p}_i]^2} p(\theta \mid m^h) \mathrm{d}(\theta) \\
&\quad \times \int \frac{1}{(2\pi\sigma_I^2)^{\frac{N_I}{2}}} \mathrm{e}^{-\frac{1}{2\sigma_I^2}\sum_{i=N+1}^{N+N_I}[x_i - \alpha_I - \boldsymbol{\beta}_I\cdot\boldsymbol{p}_i]^2} p(\theta_I \mid m^h) \mathrm{d}(\theta_I) \\
&= \text{BGe}(d, X) \times \text{BGe}(d_I, X) \qquad (13)
\end{aligned}$$

by applying Equation (9) to each term in Equation (12).

2.4. Several interventions

A dataset may consist of observations from many different experimental conditions and several of them may affect the relationship between a node X and its parents $\boldsymbol{P}$. It is convenient to distinguish between interventions and experimental conditions which may consist of several interventions at the same time. For example, an intervention could be a gene perturbation where some molecular agent targets the expression of a gene, while several such agents may be added together in a particular experimental setting. Since the effects of multiple soft interventions may not be simply additive we include potential interactions between the agents. Amongst a set of m potential interventions $\{I_1, \ldots, I_m\}$ denote with I_X the subset which are connected to the node X along with the other observational parents $\boldsymbol{P}$. Each combination of states of I_X corresponds to a different experimental condition which we can equivalently represent by a categorical variable E with a local SEM as

$$X = \begin{cases} \alpha_0 + \boldsymbol{\beta}_0 \cdot \boldsymbol{P} + \epsilon_0 & \text{for } E = 0 \\ \alpha_1 + \boldsymbol{\beta}_1 \cdot \boldsymbol{P} + \epsilon_1 & \text{for } E = 1 \\ \ldots & \\ \alpha_K + \boldsymbol{\beta}_K \cdot \boldsymbol{P} + \epsilon_K & \text{for } E = K \end{cases} \qquad (14)$$

where there are $(K+1)$ distinct conditions. By including potential interactions between the effects of interventions this setting is a simple extension of the approach in Section 2.3. By defining $\widetilde{d}$ as the entirety of the data, and d_k the subset of the data for which the experiment condition corresponds to category k, the marginal likelihood contribution to the interventional BGe score directly follows as

$$\text{iBGe}(\widetilde{d}, X) = \prod_{k=0}^{K} \text{BGe}(d_k, X) \qquad (15)$$

For the full iBGe score, we multiply the marginal likelihoods above for each node X and include the prior on graphical structures.

2.5. Equivalence classes

Under perfect intervention, Hauser and Bühlmann (2012) characterised the space of Markov-equivalent DAGs. One way of expressing the Markov equivalence condition for a class of DAGs is to require

them to be Markov equivalent (achieving the same score) for each intervention and its associated data (Theorem 10, Hauser and Bühlmann, 2012). Since the BGe score satisfies equivalence (Geiger and Heckerman, 2002), the iBGe score of Equation (15) inherits the property. More recently, Yang et al. (2018) proved that the same characterisation also holds for soft interventions, as long as purely observational data [the $k = 0$ case in Equation (15)] exist. Consequently, the iBGe score will respect equivalence under the same conditions. Moreover, by representing the interventions as additional nodes connected to their targets in the networks, the equivalence class under interventions is defined by the usual graphical criteria of having the same skeleton and v-structures in the extended network (Yang et al., 2018).

2.6. Unknown targets

The formula in Equation (15) allows us to score a DAG when the targets of each intervention are known. To extend to the case where the targets are unknown we also need to infer the edges between the intervention and the observation nodes (akin to the discrete case, Eaton and Murphy, 2007). By implementing the interventional BGe score into a Bayesian sampling approach (Kuipers and Moffa, 2017; Kuipers et al., 2022) we can learn and sample the structure as well as the targets of the interventions. One peculiarity is that the intervention nodes are fixed by the experimental setting and have essentially undergone hard interventions meaning that they have no parents in the network and can only affect downstream observables. The relevant size for inference is the internal structure of the DAG excluding the intervention nodes, which is the number n of observed nodes.

2.7. Causal effect estimation

Bayesian approaches to structure learning capture the uncertainty in the network structure by characterising their posterior distribution. Furthermore, we can use an ensemble of structures drawn from the posterior to perform downstream analyses and quantify for example the uncertainty in the estimation of intervention effects. From each structure in the sample and given a prior distribution on its parameters we can evaluate (or sample) intervention effects and derive their posterior distributions (Moffa et al., 2017; Kuipers et al., 2019; Moffa et al., 2023). The iBGe therefore also opens this possibility for mixed observational and interventional data in the linear-Gaussian setting.

When simulating intervention effects on a given network, it is meaningful to think in terms of hard interventions, since we can easily use them to derive effects under different scenarios. At each node, only the data generated under the natural state (without direct intervention) informs the computation of the estimates of edge coefficients for the unperturbed network. These in turn enter the quantification of causal effects propagating downstream of the intervened-upon node. The intervention data on the other hand helps determine the structural relations in the structure learning or sampling phase of the algorithm. For a given network, Viinikka et al. (2020) derived an explicit expression (for the case of purely observational data) for the posterior distribution of the edge coefficients (the $\boldsymbol{\beta}$ in the SEMs of Section 2.2), which we apply to the natural state data.

Combining DAG sampling with conditional parameter sampling given a structure we build the full posterior distributions of causal effects and obtain a Monte Carlo estimate of hard (perfect) effects through the network. In case we wish to predict the effect of known soft interventions, in a linear setting, we can obtain their distribution by a simple weighted combination of the hard effects.

2.8. Software implementation

To use our iBGe approach we interfaced the interventional BGe score with the **BiDAG** package (Suter et al., 2023) which implements a state-of-the-art hybrid method for structure learning and Bayesian sampling (Kuipers et al., 2022) and which offers good performance for continuous observational data in benchmarking studies (Rios et al., 2021). Along with defining the iBGe score, in the software implementation, we treat the interventions as background nodes since they may have no parents in the network. For estimating causal effects, we also interfaced the implementation with the **Bestie** package (`https://CRAN.R-project.org/package=Bestie`). Code for computing the iBGe score, as well as for reproducing the simulations and the real data analysis is hosted at `https://github.com/jackkuipers/iBGe`.

3. Simulation benchmarking

As a proof of concept, we first tested the performance of the BGe score in the well-understood case of hard interventions, with the results discussed in Appendix A. Here, we focus on the performance of the iBGe score in the more realistic case where both the targets of intervention, as well as the exact magnitude of their effects are unknown. As a relevant competitor handling unknown and soft interventions, we include UT-IGSP (Squires et al., 2020), an order-based greedy search with constraint-based tests. We exclude the recent Bayesian approach of Castelletti and Peluso (2023) since it assumes hard interventions and its structure-based scheme does not scale to larger networks.

3.1. Simulation setting

The simulation setup included data with $n = 100$ observed variables and $m = 10$ different interventions. The graphical structure amongst the observed nodes was sampled as a random DAG with the default options of the **pcalg** package (Kalisch et al., 2012) with an expected number of parents per node set to 2. The number of targets of each intervention was sampled from a Poisson with rate parameter 1 shifted by 1, while the targets themselves were sampled uniformly from the 100 observed nodes. The edge weights in the DAG were sampled uniformly in the range $[0.25, 1]$. The interventions were non-overlapping and their effect on the target nodes was to shift their mean by an amount sampled from a standard normal and to damp the effects of the other parents by multiplying their edge weights by a uniformly sampled number from the interval $[0.1, 1]$. The data was generated from the SEM in topological order, where each node is given by the linear combination of its parents in the DAG (possibly modified by the interventions) and with standard normal noise added on top. This choice is for simplicity, but we do not impose or assume equal variance in the modelling or inference. Although the iBGe score and other score-equivalent approaches are insensitive to the scale of the data, other metrics may not be (Reisach et al., 2021) so we standardise the data by default.

For each of the 10 interventions, we generated $N_I = (5, 10, 20)$ observations under that condition, and then added purely observational data to achieve $N = 400$ observations in total. In addition, we examine a large data setting where the number of each type of observation is multiplied by 10. To capture the sampling variability we repeated the simulation of each setting 100 times.

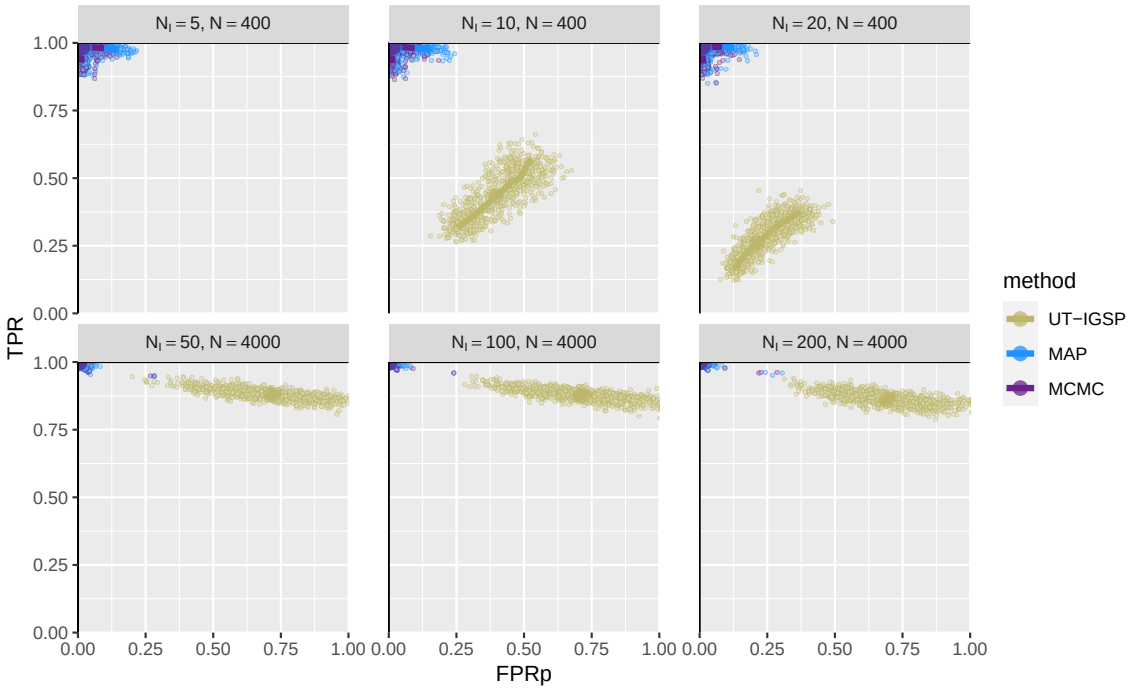


Figure 1: Comparison of the iBGe score derived MAP and MCMC consensus network to UT-IGSP. Each point is a single repetition for a single parameter value, while the thicker lines show the average behaviour for each value. The larger dot is placed at $10^5\alpha = 1 = 10\alpha_\mu$.

3.2. Simulation results

To evaluate performance we use ROC-like curves comparing the inferred and data-generating DAG after mapping to the equivalence class. Details of the measure used are in Appendix B.

Employing the iBGe score for the iterative MAP search with the **BiDAG** package, and building a consensus graph through posterior thresholding (with threshold 0.5) from a sample of DAGs using its MCMC scheme achieves very high performance (Figure 1). The constraint-based UT-IGSP algorithm appears to perform quite poorly. For the lowest number of samples per intervention of $N_I = 5$ with a total of $N = 400$ samples, the algorithm runs into numerical errors and cannot complete when it tries to build and test the covariance matrix in each experimental condition. With twice as many samples per intervention ($N_I = 10$), UT-IGSP typically fails to find half the true edges in the graph, and this gets worse with more interventional samples as this setting has fewer observational samples (with the fixed total of 400). Increasing the sample sizes by a factor of 10 (Figure 1, bottom row), drastically improves the number of true edges found by UT-IGSP, but also leads to large numbers of false positives. When comparing the bottom row of Figure 1 with the top row it is apparent that the UT-IGSP even with much larger sample sizes achieves worse performance than the iBGe methods proposed here do at the smaller sample size. The very good performance of the iBGe score at $N = 400$ becomes near perfect at the larger sample size of $N = 4000$ (Figure 2).

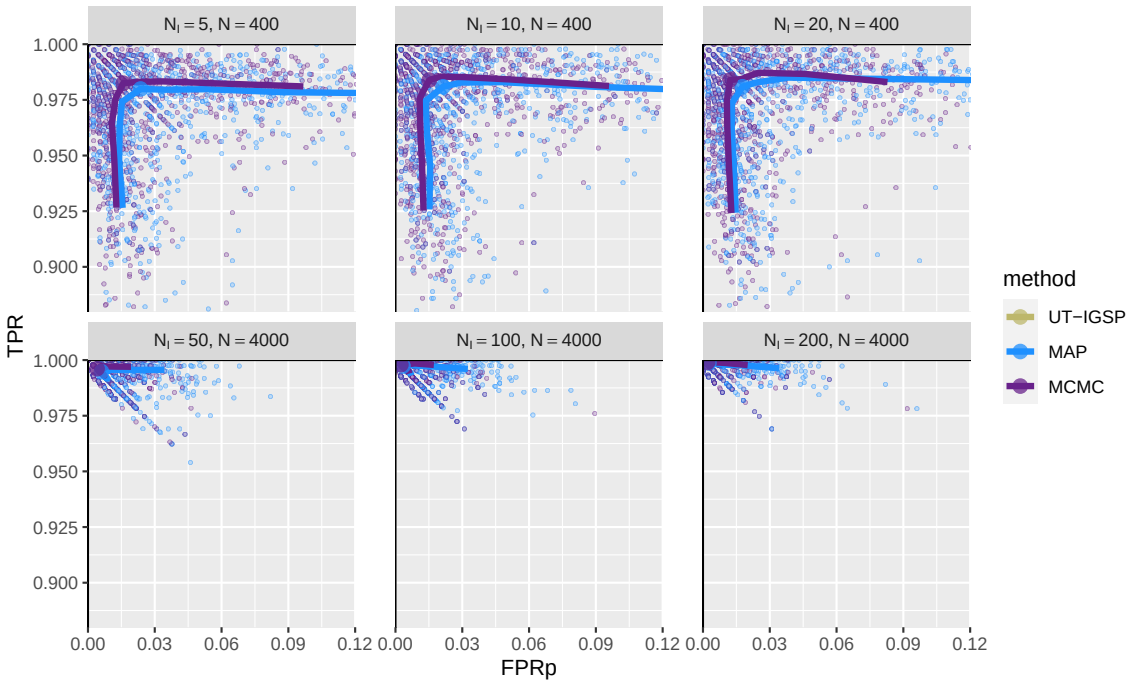


Figure 2: We zoom into the top left of Figure 1 to examine how closely our MAP and MCMC methods with the iBGe score approach (0,1). At the smaller sample sizes, the differences are typically a few edges with a slight advantage from the Bayesian model averaging with the MCMC scheme. At the larger sample size, we often have perfect performance.

4. Biological perturbation data

To compare the iBGe approach to alternatives on real data we consider the commonly used dataset of Sachs et al. (2005). For each T-cell in multivariate flow cytometry experiments, the amount of 11 phosphorylated molecules was measured via fluorescent readouts. The experiment aimed to quantify the causal relationship between these 11 nodes in the signalling pathway. The experiments were repeated under 9 experimental conditions, of which we use the first 7 (following Wang et al., 2017; Squires et al., 2020). These all contain the T-cell activator Anti-CD3/CD28 which is considered the underlying observational condition. Experiments 2–7 contain an additional agent which for experiments 3–7 directly targets a measured signalling node. The raw data is log-transformed but since the batch and experimental condition are the same and no further information on the experimental design is included in the dataset, we do not perform batch correction as would be standard.

We compare applying our iBGe score to the UT-IGSP (Squires et al., 2020) approach in terms of network recovery compared to the canonical network depicted in Sachs et al. (2005), both with and without their missing (dashed) edges. Since there are so many observations (5,846 in total) the prior parameters of the iBGe score make little difference, therefore to obtain a ROC-like curve we vary a penalisation on edges to induce networks of different densities instead. The results (Figure 3) demonstrate a clear advantage of the iBGe approach over the constraint-based UT-IGSP.

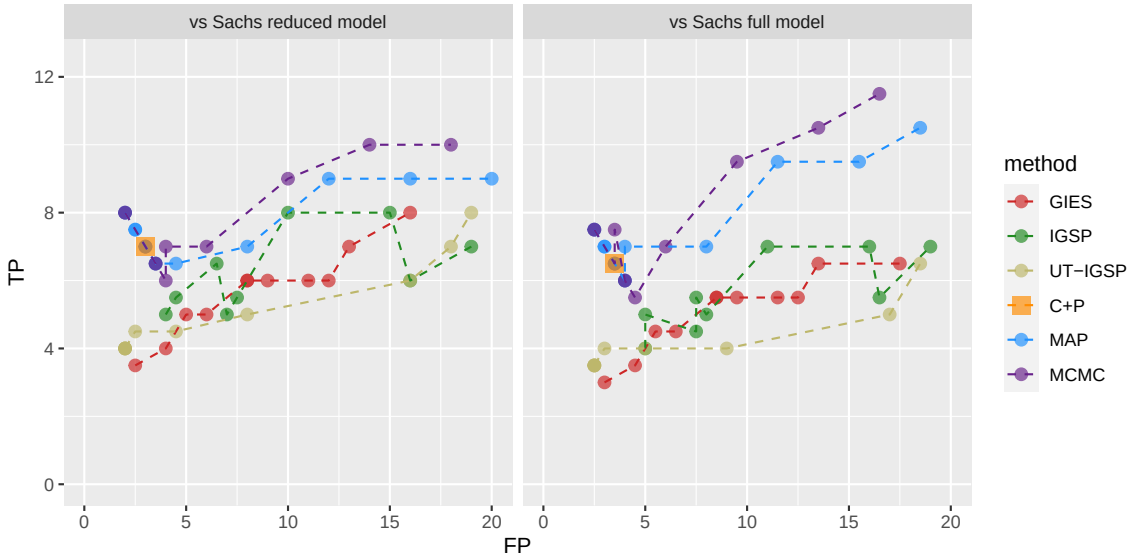
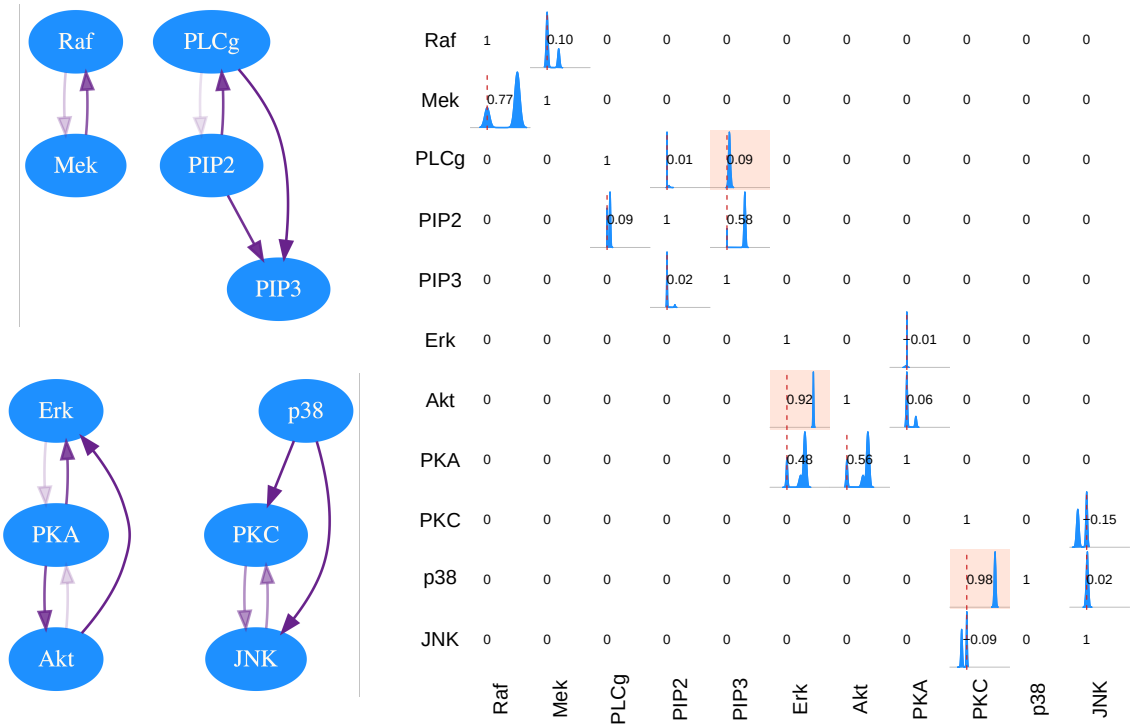


Figure 3: Performance on reconstructing the network of Sachs et al. (2005). We compare the iBGe score derived MAP and MCMC consensus network with UT-IGSP, IGSP, GIES and the consensus network of Castelletti and Peluso (2023), labelled C+P, in terms of the number of TPs and FPs defined as in Appendix B. We compare to the canonical network of Sachs et al. (2005) both without their missing edges (reduced model) and with (full model).

When provided with the targets, IGSP (Wang et al., 2017) and GIES (Hauser and Bühlmann, 2012) perform very similarly to each other, marginally better than UT-IGSP apart from near the origin, but still distinctly worse than the iBGe score. The consensus network of the Bayesian approach of Castelletti and Peluso (2023) which handles hard interventions with unknown targets sits amongst the iBGe results (those without any edge penalisation). The iBGe however covers the more general and complex cases of unknown targets with soft interventions.

The canonical network considered in Wang et al. (2017); Squires et al. (2020) has an edge mistakenly reversed. Using their ground truth network simply makes all results correspondingly worse, but this does not affect the comparative performance of Figure 3. We focused on the relative performance of different generalised approaches for continuous data, but the absolute performance of all is moderately low with a minimum SHD of 11 for the iBGe approaches, 13 for Castelletti and Peluso (2023), 15 for UT-IGSP and 16 for IGSP and GIES for the reduced network with 17 edges in total. The original network of Sachs et al. (2005) was created by discretising the data, and network recovery for this data can be quite variable, possibly due to unreliability in the ground truth network, unmeasured experimental confounding, non-linearities and skew (Ramsey and Andrews, 2018).

Along with allowing us to learn the graphical structure, and its uncertainty (Figure 4a), the iBGe score allows us to further estimate the causal effects from each network (Section 2.7). Through Bayesian model averaging over the sampled graphs and parameters, we can then derive a sample approximation of the posterior distribution of causal effects (Figure 4b). By doing so we can characterise both the structural relationships between variables as well as the magnitude of the effect that intervening on one variable may have on another.



(a) (b)

Figure 4: Summary of posterior distributions with the iBGe score on the Sachs data: (a) The posterior DAG distribution with the edge opacity corresponding to the posterior probability of the edge presence. (b) The posterior causal effect distribution with dashed red lines at no effect and effects whose 95% credible interval excludes 0 highlighted in peach.

5. Conclusions

Starting from the BGe score (Geiger and Heckerman, 2002) for graphical models with purely observational data we developed a Bayesian scoring metric for a mix of observational and interventional data. In particular, we define the score as allowing interventions to be soft and unknown. Accordingly, interventions are not necessarily structural and may affect the strength of a relationship, while the targets may be unknown. For discrete data, we may view soft interventions (Eaton and Murphy, 2007) as interactions, and we developed a model for continuous data by taking the analogy over to the continuous case. By further leveraging the connections between the SEM and the matrix parametrisation of the BGe score, we could define the interventional BGe (iBGe) score as a natural combination of BGe scores over experimental conditions. The novel framework covers the case of soft interventions while handling uncertainty in the targeting by also learning the connections between the interventions and the observational nodes.

The iBGe score we derived is simple to compute and include in score-based algorithms, allowing their easy adaptation to mixed data with unknown and soft interventions. The highlight of the iBGe score, however, is that it is Bayesian so it can enter as a target in MCMC schemes and Bayesian approaches for model averaging over DAGs, such as order (Friedman and Koller, 2003)

or partition MCMC (Kuipers and Moffa, 2017). We employ these by interfacing the score with a hybrid approach (Kuipers et al., 2022; Suter et al., 2023). The iBGe approach, especially combined with such state-of-the-art hybrid inference (Kuipers et al., 2022), outperforms current alternatives like UT-IGSP (Squires et al., 2020) in simulation studies and on real data.

The Bayesian approach to causal structure learning accomplishes some important analysis tasks: quantifying the uncertainty in the network structure, characterising the uncertainty in the parameter distributions and automatically propagating both into the downstream analyses of intervention effects (Moffa et al., 2017; Kuipers et al., 2019; Moffa et al., 2023). As with the BGe score for purely observational data (Viinikka et al., 2020), the iBGe score now enables the same Bayesian inference of causal effects for mixed observational and interventional data.

The biological perturbation data of Sachs et al. (2005), displays non-Gaussian skewness and possible non-linearity, breaking the underlying linear-Gaussian assumptions of the BGe and iBGe scores. Although constraint-based methods can relatively easily change their conditional independence tests, building a marginalisable likelihood like the BGe suitable for Bayesian analyses in the presence of non-linearity and non-Gaussianity is more challenging, though sampling-based approaches have been developed using Gaussian processes (Friedman and Nachman, 2000; Giudice et al., 2023, 2024). For these and similar scores developed for non-Gaussian observational data, however, we can expect that the approach developed here will allow for a direct extension to handle data with soft unknown interventions.

References

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25:505–541, 1997.

Federico Castelletti and Stefano Peluso. Network structure learning under uncertain interventions. *Journal of the American Statistical Association*, 118:2117–2128, 2023.

David M. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Conference on Uncertainty in Artificial Intelligence*, pages 116–125, 1999.

Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics*, pages 107–114, 2007.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74:981–995, 2007.

Nir Friedman and Daphne Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003.

Nir Friedman and Iftach Nachman. Gaussian process networks. In *Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 211–219, 2000.

Dan Geiger and David Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Annals of Statistics*, 30:1412–1440, 2002.

Enrico Giudice, Jack Kuipers, and Giusi Moffa. A Bayesian take on Gaussian process networks. *Advances in Neural Information Processing Systems*, 36:56602–56614, 2023.

Enrico Giudice, Jack Kuipers, and Giusi Moffa. Bayesian causal inference with Gaussian process networks. *arXiv:2402.00623*, 2024.

Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48, 1999.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. BaCaDi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 1411–1436, 2023.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13: 2409–2464, 2012.

David Heckerman. A Bayesian approach to learning causal networks. In *Eleventh Conference on Conference on Uncertainty in Artificial Intelligence*, pages 285–295, 1995.

David Heckerman and Dan Geiger. Learning Bayesian networks: A unification for discrete and Gaussian domains. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 274–284, 1995.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47:1–26, 2012.

Kevin B. Korb, Lucas R. Hope, Ann E. Nicholson, and Karl Axnick. Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*, pages 322–331. Springer, 2004.

Jack Kuipers and Giusi Moffa. Partition MCMC for inference on acyclic digraphs. *Journal of the American Statistical Association*, 12:282–299, 2017.

Jack Kuipers, Giusi Moffa, and David Heckerman. Addendum on the scoring of Gaussian directed acyclic graphical models. *Annals of Statistics*, 42:1689–1691, 2014.

Jack Kuipers, Giusi Moffa, Elizabeth Kuipers, Daniel Freeman, and Paul Bebbington. Links between psychotic and neurotic symptoms in the general population: An analysis of longitudinal British national survey data using directed acyclic graphs. *Psychological Medicine*, 49:388–395, 2019.

Jack Kuipers, Polina Suter, and Giusi Moffa. Efficient sampling and structure learning of Bayesian networks. *Journal of Computational and Graphical Statistics*, 31:639–650, 2022.

Giusi Moffa, Gennaro Catone, Jack Kuipers, Elizabeth Kuipers, Daniel Freeman, Steven Marwaha, Belinda R. Lennox, Matthew R. Broome, and Paul Bebbington. Using directed acyclic graphs in epidemiological research in psychosis: An analysis of the role of bullying in psychosis. *Schizophrenia Bulletin*, 43:1273–1279, 2017.

Giusi Moffa, Jack Kuipers, Giuseppe Carrà, Cristina Crocamo, Elizabeth Kuipers, Matthias Angermeyer, Traolach Brugha, Mondher Toumi, and Paul Bebbington. Longitudinal symptomatic interactions in long-standing schizophrenia: A novel five-point analysis based on directed acyclic graphs. *Psychological Medicine*, 53:1371–1378, 2023.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–688, 1995.

Judea Pearl. *Causality: models, reasoning and inference*. MIT press, 2000.

Joseph Ramsey and Bryan Andrews. FASK with interventional knowledge recovers edges from the Sachs model. *arXiv:1805.03108*, 2018.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

Felix L. Rios, Giusi Moffa, and Jack Kuipers. Benchpress: a scalable and platform-independent workflow for benchmarking structure learning algorithms for graphical models. *arXiv:2107.03863*, 2021.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048, 2020.

Polina Suter, Jack Kuipers, Giusi Moffa, and Niko Beerenwinkel. Bayesian structure learning and sampling of Bayesian networks with the R package BiDAG. *Journal of Statistical Software*, 105: 1–31, 2023.

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.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30, 2017.

Karren D. Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

Appendix A. Simulation study with perfect interventions

For hard interventions, if in addition to a set of N observations of $\boldsymbol{X}$ we also have N_I observations obtained after perfectly intervening on node X and setting it to some normally sampled value stored in data d_I, the likelihood for the full data $\widetilde{d} = (d, d_I)$ for node X given its parents is

$$p(\widetilde{d}^X \mid \widetilde{d}^{\boldsymbol{P}}, \boldsymbol{\mu}, W, m^h) = p(d^X \mid d^{\boldsymbol{P}}, \boldsymbol{\mu}, W, m^h) \prod_{i=N+1}^{N+N_I} f(x_i) \tag{16}$$

where $f(x)$ is the interventional distribution (Hauser and Bühlmann, 2012). When intervening, we break the connection between X and its parents so that the data d_I only contributes a constant factor to the likelihood and a constant term to the log-likelihood. Consequently, there is no bearing on the relative score of different DAGs. Removing the constant term corresponding to the interventional data, reduces to the setting and result for the BGe score for the observational data d alone. For scoring a node X we simply remove all data where X has undergone a deterministic hard intervention (Cooper and Yoo, 1999) and then compute the BGe score as usual. Different nodes may undergo intervention on different occasions and the score for each node given its parents is only based on the data where that node has been observed under conditions without interventions.

With perfect interventions, we can additionally compare to GIES (Hauser and Bühlmann, 2012) and IGSP (Wang et al., 2017), the precursor of UT-IGSP (Squires et al., 2020) with known targets. We follow the same simulation strategy as in the main text, select 10 nodes randomly to be targets, and fix a fraction of the data $\rho = (0, 0.01, 0.03, 0.1, 0.3, 1)$ to be interventional. This covers the range from fully observational to fully interventional. Amongst the interventional data, we randomly select the target for each observation. By default we did not standardise the data, since then IGSP failed to perform. In the comparison, we use the following range of penalisation parameters for GIES

$$\lambda = (0.607, 0.847, 1.18, 1.65, 2.3, 3.21, 4.48, 6.26, 8.74) \tag{17}$$

The constraint-based algorithm of IGSP performs relatively poorly in this setting (Figure S1), especially as it seems to require more observational data, often returning the empty DAG when there is too much interventional data. The iBGe score and GIES are more robust, with a clear strong advantage to using the iBGe score over GIES in terms of performance (Figure S2).

Timewise (Figure S3), GIES is much faster than the sampling-based inference schemes of **BiDAG**, in line with results for purely observational data (Kuipers et al., 2022), but at the cost of worse performance. IGSP is slower still, but this may depend heavily on the implementation as the python-based UT-IGSP runs notably faster.

For completeness, we include the results when we standardised the data (Figure S4) to remove the possibility of artificially using the scale of the data to improve performance (Reisach et al., 2021). As expected, the performance of the iBGe score and GIES are relatively unchanged. However, IGSP only returns the empty DAG for $\rho > 0$ and fails to learn meaningful DAGs.

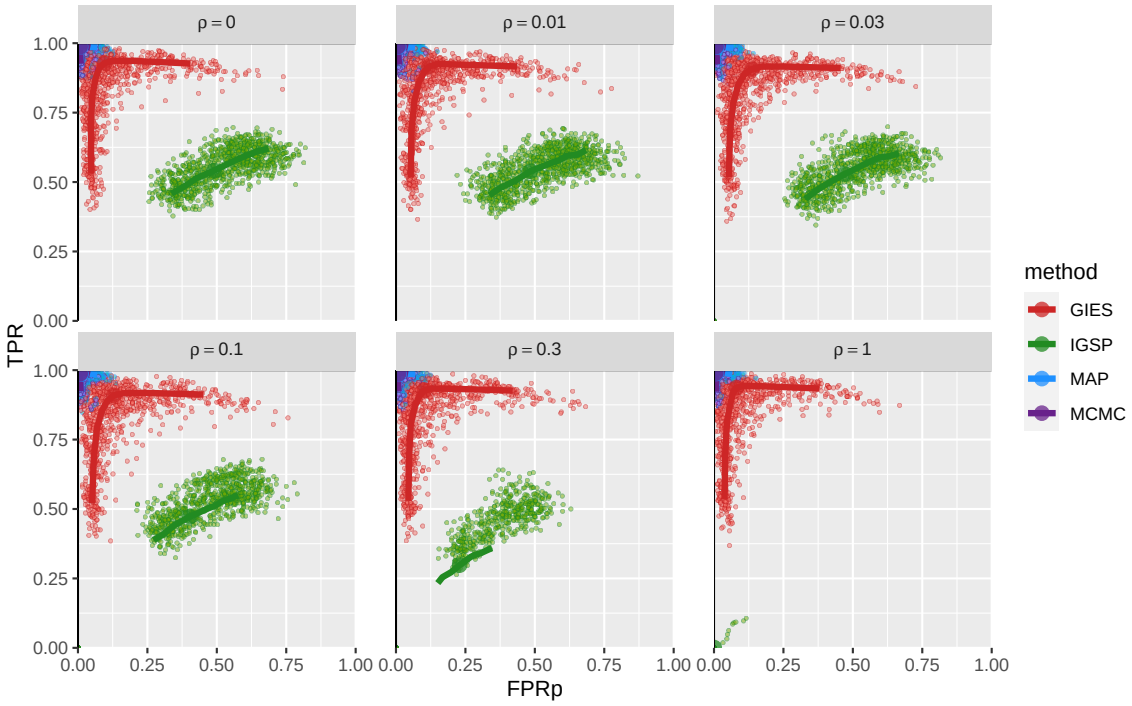


Figure S1: With perfect and known interventions, a comparison of the iBGe score derived MAP and MCMC consensus networks to IGSP and GIES, as the fraction of interventional data ρ increases. Each point is a single repetition for a single parameter value, while the thicker lines show the average behaviour for each parameter value with the larger dot placed at $10^5\alpha = 1 = 10\alpha_\mu$, and $\lambda = 2.3$. IGSP often returns the empty DAG at (0,0) with more limited amounts of purely observational data at larger ρ, leading to the divergence between its cloud of dots and the average line.

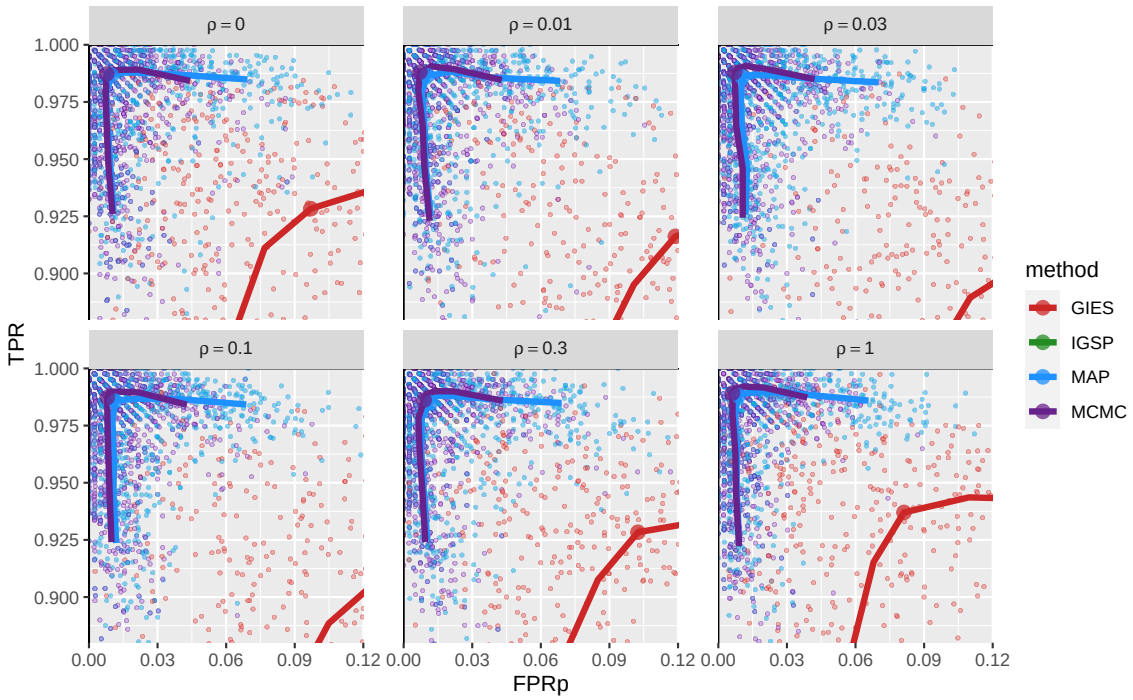


Figure S2: Zoom into the top left of Figure S1 to better compare the iBGe score input into the MAP and MCMC inference schemes to GIES.

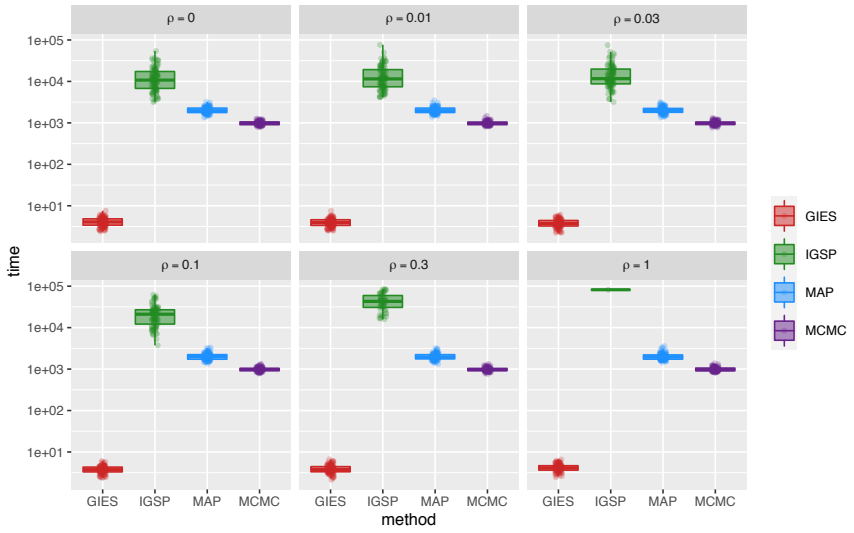


Figure S3: The time taken to learn a network with GIES, IGSP, and the MAP and MCMC consensus networks running with the iBGe score for known perfect interventions. The MCMC scheme requires the MAP steps to be run first and its times are the additional time for the sampling.

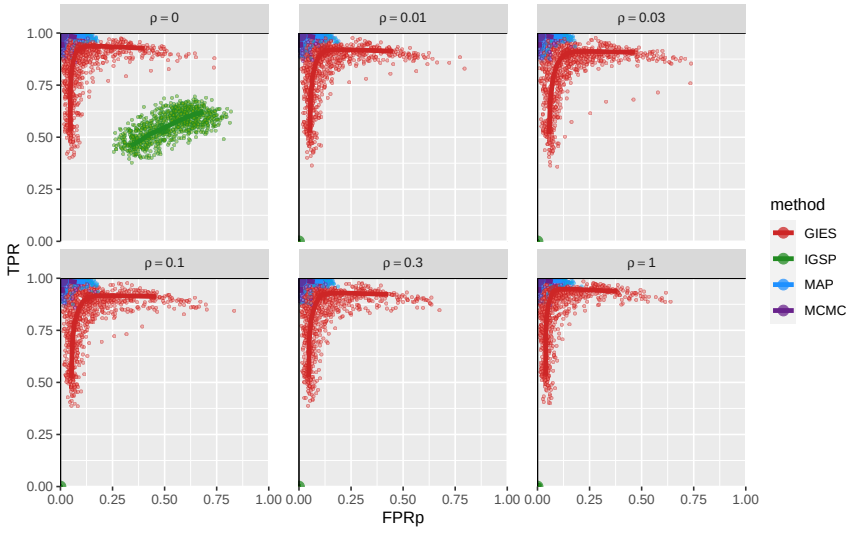


Figure S4: Comparison for known perfect interventions, as in Figure S1, but where the data has been standardised. This highlights that IGSP fails in this setting while the iBGe score (MAP and MCMC) and GIES are robust.

Appendix B. Performance measure

As a measure of performance, we compared the inferred DAG to the data-generating DAG, after mapping both to the equivalence-class space. Although the inference schemes are unaware of the actual targets, they are used for deriving equivalence classes for the performance evaluation (Hauser and Bühlmann, 2012). We compute the number of TP edges (directed edges in the same direction in inferred and model graphs, or undirected edges in both) and the number of FP edges (directed or undirected edges in the inferred graph not in the model graph). Edge directions which do not match (wrong direction or undirected) between the inferred and model graphs count as $\frac{1}{2}$ to FP and $\frac{1}{2}$ to FN so that the structural Hamming distance (SHD) is

$$\mathrm{SHD} = \mathrm{FN} + \mathrm{FP} = \mathrm{P} - \mathrm{TP} + \mathrm{FP} \tag{18}$$

and it coincides with the Manhattan distance from $(0, \mathrm{P})$ in a TP vs FP plot. Since the total number of edges is random, we scale by P and plot the TPR against FPRp = $\frac{\mathrm{FP}}{\mathrm{P}}$. The performance further depends on algorithmic parameters, so we create ROC-like curves by varying the significance level α of the independence tests in UT-IGSP and by varying the prior parameter α_μ in the iBGe score while keeping $\alpha_w = \alpha_\mu + n + 1$. For the plots, we used the following values

$$10^5\alpha = (0.00248, 0.0111, 0.0498, 0.223, 1, 1.65, 2.72, 4.48, 7.39) = 10\alpha_\mu \tag{19}$$