Discrete Learning Of DAGs Via Backpropagation

Anonymous Author(s)
Affiliation
Address
e-mail

Abstract

Recently continuous relaxations have been proposed in order to learn directed acyclic graphs (DAGs) by backpropagation, instead of combinatorial optimization. However, a number of techniques for fully discrete backpropagation could instead be applied. In this paper, we explore this direction and propose DAG-DB, a framework for learning DAGs by Discrete Backpropagation, based on the architecture of Implicit Maximum Likelihood Estimation (I-MLE). DAG-DB performs competitively using either of two fully discrete backpropagation techniques, I-MLE itself, or straight-through estimation.

1 Introduction

Aim. Directed acyclic graphs (DAGs) occur in a wide range of contexts, including project management, version control systems, evolutionary biology, and Bayesian networks. Bayesian networks have been a particularly popular subject for machine learning, including for the problem considered in this paper, learning a Bayesian network’s DAG from data generated by the Bayesian network. This paper aims to learn DAGs using fully discrete backpropagation, avoiding the relaxation from discrete to continuous variables which is a common part of non-combinatoric DAG learning (for example, 4–16).

Bayesian networks. A Bayesian network associates a DAG with a random variable which has components indexed by the DAG nodes. A directed edge $i \rightarrow j$ between nodes $i$ and $j$ represents a dependency of the random variable’s $j$ component on its $i$ component. Discussion of Bayesian networks and DAGs may be found in textbooks [e.g. 17–19]; to establish our terminology, appendix A gives a very brief outline.

Contribution. Our contribution is to show that backpropagation methods which retain the fully discrete nature of a directed acyclic graph can be used to predict DAGs from data. We avoid the common approach of ‘relaxing’ digraph adjacency matrices from a binary to a real matrix. Instead, we sample discrete variables probabilistically and use methods of backpropagating that do not relax the variables to be continuous. No ethical issues are raised.

2 Related work and background

Learning DAGs from data. Most ways of learning DAGs from data can classed as combinatoric methods [reviewed in 20] or continuous optimisation methods [reviewed in 21]. Combinatoric methods may themselves be labelled as constraint–based, identifying the graph from conditional independence testing, and score–based, searching the space of possible graphs using a score function to evaluate search–results [20]. The PC algorithm [22] is a well–known constraint–based method, with a number of variant algorithms, for example PC-Stable [23] which reduces dependency on arbitrary node ordering. An example of a score–based approach is the Fast Greedy Equivalence Search (FGES)

algorithm \cite{24}, which is an optimised version of the Greedy Equivalence Search (GES) \cite{25} \cite{26}. The development of NOTEARS in 2018 \cite{4} for linearly–generated data signalled the start of widespread development of continuous optimisation approaches, including the GOLEM method \cite{10} which is generally even more successful than NOTEARS for linear data with, for example, Gaussian noise. Many other continuous methods have recently been developed [for example, \cite{5,9,11,14}] and the method presented in this paper, which uses gradient descent, may be viewed as related to these. However, continuous methods generally adopt some approach of ‘relaxing’ the discrete property of an edge being present or absent into a continuous variable, whereas our DAG-DB approach maintains edges as fully discrete, binary objects, even during our method’s training phase. Our approach might be termed a ‘probabilistic relaxation’, and another example of this can be found in the SDI \cite{27} framework, which focuses on contexts where data may be generated by interventions during training, whereas, here, our focus is on pre–generated ‘observational’ data.

Discrete backpropagation. A number of methods have also been developed to allow backpropagation without transforming discrete variables to be continuous in training. We take four such methods as examples. Straight–through estimation (STE) \cite{2,3} essentially ignores the discrete nature of the variables, and, as we will see, can be surprisingly effective. Score–function estimation (SFE) \cite{28} uses the elementary rules of differentiation to estimate the gradient of an expectation by sampling with respect to a probability distribution. ‘Blackbox’ differentiation (BB) \cite{29} is a way of adjusting continuous inputs to a combinatorial solver in order to mimic gradient descent for the discrete variable. Implicit maximum likelihood estimation (I-MLE) \cite{1} incorporates noise into BB to handle discrete random variables. This allows combinatorial solvers to be used for maximum likelihood estimation. I-MLE also makes use of both STE and a seminal result by Domke \cite{30}. We note that SDI \cite{27}, highlighted above, uses a particular Bernoulli–based SFE-like form of backpropagation \cite{31}.

Perturb–and–MAP sampling. We use the technique of perturb–and–MAP (P&M) sampling of random variables \cite{32}. This approximates a parametrized distribution, say \( p(Z; \Theta) \), with a random matrix variable \( Z \) and continuous parameter matrix \( \Theta \), as

\[
p(Z; \Theta) \sim \text{MAP}(\Theta + \tau \Psi),
\]

where each element of the noise matrix \( \Psi \) is i.i.d. sampled from a standard one–dimensional distribution, \( \tau > 0 \) is a temperature, and \( \text{MAP}(\Theta + \tau \Psi) \) denotes the most likely value of \( Z \) with respect to \( p(Z; \Theta + \tau \Psi) \). P&M is useful when sampling \( p(Z; \Theta) \) directly is intractable or expensive, and the MAP solver is cheap.

Maximum DAG. Our random variable \( Z \) will represent the adjacency matrix of a directed graph. To find a DAG from that digraph, we will use an approximate solution to the ‘maximum DAG’ problem. This considers a weighted digraph (each edge is assigned a real number) and the challenge is to remove edges from the digraph to form a DAG, whilst maximising the sum of the remaining edge–weights. We found it most efficient, and necessary as the number of digraph nodes rose toward 100, to use an approximate maximum DAG solver, ‘Greedy Feedback Arc Set’ (GFAS) \cite{33}, our Python implementation being adapted from (more optimised) Java code \cite{34,35}.

\[
\begin{aligned}
\Theta &\sim p(Z; \Theta), \\
Z &\rightarrow x \odot Z \\
D_{\text{pred}} &\rightarrow \ell = \text{MSE}(x, \tilde{x}) + r(Z) \\
&\rightarrow L
\end{aligned}
\]
3 Method

Framework. Figure 1 gives an overview of the DAG-DB framework. Suppose we wish to learn a DAG with $d$ nodes from a dataset $X \in \mathbb{R}^{n \times d}$ of $n$ data points $x \in \mathbb{R}^d$. Let $\mathcal{R} \subseteq \mathbb{R}^{d \times d}$ and $\mathcal{B} \subseteq \{0, 1\}^{d \times d}$ be the sets of zero–diagonal, respectively real and binary, matrices. A learnable vector $\Theta \in \mathcal{R}$ parametrizes a exponential family distribution $p(Z; \Theta)$,

$$ p(Z; \Theta) = \exp \left( (Z, \Theta)_{PE} / \tau - \mathcal{A}(\Theta) \right), $$

where $Z \in \mathcal{Z} \subseteq \mathcal{B}$ is a discrete matrix, $\tau > 0$ is a temperature, and $\mathcal{A}(\Theta)$ normalizes $p(Z; \Theta)$ to sum to one. Recognising the matrix form of $\Theta$ and $Z$, eq. (2) uses the Frobenius scalar product,

$$ \langle Z, \Theta \rangle_{PE} := \text{tr}(Z^T \Theta) = \sum_{ij} Z_{ij} \theta_{ij}. $$

The matrix $Z$ is interpreted as the adjacency matrix of a directed graph with, because of its zero diagonal, no self–loops. In the forward pass, samples $Z_s^{(s)}$, $s = 1, ..., S$, from $p(Z; \Theta)$ are taken using the P&M sampling outlined in section 2. For $Z = B$, the MAP solver sets matrix elements of $\mathcal{A}(\Theta)$ to be one if $\Theta > 0$ and zero otherwise. We found empirically that it is best to use the standard logistic distribution to generate the noise $\Phi$ in eq. (1). This can also be theoretically justified as a good choice because of the binary nature of $Z$’s matrix elements.

As is common for continuous optimisation methods, DAG learning proceeds by solving the problem of predicting values of $x$ components $x_i$ from its values $x_i$ at ‘parent’ nodes with edges $i \rightarrow j$. As set out in appendix C this is ensured by the ‘graphification’ operation $x \circ Z$ in the context of the linear map $f$. Elsewhere, this technique usually employs a weighted adjacency matrix with real elements [see, e.g., 4, 8, 10]; but we use the binary adjacency matrix $Z$ without such a relaxation. To the extent we relax the problem, this is by treating the digraph probabilistically via the distribution $p(Z; \Theta)$, which, in practice, becomes concentrated around the most likely adjacency matrix.

The parameters $\Theta$ and $\Phi$ are trained by feeding data points $x \in X$ into the model batchwise. The mean–squared error between $\hat{x}$ and $x$ is added to a regularizing function $r(Z)$ to give a loss $L$ associated with the sample $Z$. The empirical mean of these losses, over the samples $Z$ and the batch members $x$, gives the batch loss $L$. Our learnt parameters $\Phi$ and $\Theta$ are then updated by backpropagation: while for $\Phi$ this is standard backpropagation, a technique needs to be chosen for discrete backpropagation from $Z$ to $\Theta$. We found I-MLE [1] and straight–through estimation [2, 3] to be useful techniques for such backpropagation; we had less success here with score–function estimation [28] and ‘blackbox’ differentiation [29]. Appendix B gives details of I-MLE and straight–through estimation as used in DAG-DB.

After the final epoch of training, the GFAS algorithm is applied. We take the single most probable directed graph $Z = \text{MAP}(\Theta)$, and then find its maximum DAG when weighted by $\Theta$. Tests for 30 nodes and fewer, for which we could use an exact maximum DAG solver we built using MiniZinc [36, 37], suggested that the solutions of our solver were in practice exact, or close to exact, for 30 nodes and fewer, for which we could use an exact maximum DAG solver we built using MiniZinc [36, 37], suggested that the solutions of our solver were in practice exact, or close to exact, for 30 nodes and fewer. For a maximum size $M$, for all but one of our experiments, we use GFAS only at evaluation after the final epoch.

Regularization. DAG-DB can employ both functional and constraint regularization. Functional regularization is based on that of NOTEARS [4], and provided by the function

$$ r(Z) = \rho_{\text{DAG}} r_{\text{DAG}}(Z) + \rho_p r_p(Z), $$

where $\rho_{\text{DAG}}, \rho_p > 0$ are strength coefficients. For the DAG-regularizer $r_{\text{DAG}}$, we use the binary version of NOTEARS’ DAG regularizer,

$$ r_{\text{DAG}}(Z) = \| \text{tr}(\exp Z) - d \|^2, $$

which has been shown [4] prop. 2] to vanish if and only if $Z$ is the adjacency matrix for a DAG. The second regularizer, $r_p$, is an $L^1$–regularizer promoting sparsity by summing $Z$’s matrix elements.

The other form of regularization is to constrain $Z$ to within a proper subset $Z \subseteq B$. We optionally introduce a maximum size of digraph, constraining $Z$ to have no more than a given maximum number $M$ of non–zero entries (edges). It is here that P&M sampling becomes particularly helpful: with no such constraint, we could have used Bernoulli sampling instead of P&M; but, with a maximum size, there is no tractable direct sampling method. For a maximum size $M$, the MAP solver identifies the $M$ biggest elements of $\Theta$, but drops any that are not positive. Matrix elements of $\text{MAP}(\Theta)$ corresponding to the resulting index set have value one, and the remainder have value zero.
Table 1: Metrics for selected models for the Sachs observational dataset of 853 data points. Best metric scores are in bold. See the text and appendices for discussion.

<table>
<thead>
<tr>
<th>Model</th>
<th>SHDc</th>
<th>precc</th>
<th>recc</th>
<th>pred. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>STE_84 MAX_SIZE = 23</td>
<td>20</td>
<td>0.158</td>
<td>0.176</td>
<td>19</td>
</tr>
<tr>
<td>STE_84 MAX_SIZE = 8</td>
<td>15</td>
<td>0.600</td>
<td>0.176</td>
<td>5</td>
</tr>
<tr>
<td>IMLE_None mean</td>
<td>12.7</td>
<td>0.869</td>
<td>0.255</td>
<td>5</td>
</tr>
<tr>
<td>IMLE_None median</td>
<td>13</td>
<td>1.000</td>
<td>0.235</td>
<td>5</td>
</tr>
<tr>
<td>GOLEM</td>
<td>11</td>
<td>1.000</td>
<td>0.353</td>
<td>6</td>
</tr>
<tr>
<td>NOTEARS</td>
<td>11</td>
<td>0.467</td>
<td>0.412</td>
<td>15</td>
</tr>
<tr>
<td>FGES</td>
<td>11</td>
<td>0.750</td>
<td>0.353</td>
<td>8</td>
</tr>
<tr>
<td>PC</td>
<td>11</td>
<td>0.750</td>
<td>0.353</td>
<td>8</td>
</tr>
</tbody>
</table>

4 Experiments

Overview. As detailed in appendix D.3, we identified useful hyperparameters settings, STE_84, using STE with a maximum size 84, and IMLE_None, using I-MLE with no maximum size constraint. We then conducted ablation experiments, and tests on linearly-generated synthetic data, and on a real dataset. The main metric we use is SHDc, often normalised to nSHDc, complemented by a precision measure prec and, sometimes also by a recall measure rec. Appendix D.2 gives more details on these metrics, as part of supplementary material on the experiments in appendix D.

Ablation experiments. We performed further ablation experiments to identify the relative role of the regularizers, with results shown in appendix D.4. We found that DAG and, the appropriate setting or not of a maximum size regularizer, both have a notable good effect. However, the presence of a sparsity regularizer sp, although almost always helpful, was only very relevant for STE_84, and then only when the maximum size constraint was also ablated.

Synthetic data experiments. We generated random Erdős–Rényi ‘ERk’ and Barabási—Albert ‘SFk’ DAGs and further generated data from these using a Gaussian equal–variance linear additive noise model. The number of edges possible in these synthetic graphs varies considerably, so the maximum size limit for STE_84 was adjusted in proportion to the expected size. We tested our methods against those highlighted in section 2 on sets of 24 random graphs, with results shown in fig. 2. We can see that the DAG–DB methods usually outperform the combinatoric approaches, and they are themselves outperformed by the other continuous methods. Of our methods, STE_84 performs better, with the maximum size adjustments noted above, while IMLE_None’s performance is fairly close, except for ER4 and SF4 with d = 100 nodes, where its metrics worsen noticeably.

Real data experiments. We performed experiments on the Sachs cellular biochemistry dataset, representable as a DAG with 11 nodes and 17 edges [38]. Performance noted previously [4,10] suggests that linear models perform reasonably on Sachs’ 853–data point set of purely observational data. Table 1 compares our results against other methods. For STE_84, the table also indicates maximum size constraints which have to be reset, by a proportioning method (23) and with reference to typical, relatively successful, sizes predicted by other methods (8). We can see that IMLE_None is much more successful than either STE_84 setting, and, to account for a stochastic element in DAG–DB results, we therefore show, for I-MLE only, the results over 24 runs on the same Sachs data. We can see from the table that IMLE_None’s performance is quite close to that of GOLEM and NOTEARS which are the best of the comparison methods. Standard deviations are discussed in appendix D.7.

We also ran a method IMLE_None_Tr, which, unlike our other DAG–DB methods, applies GFAS to make Z the adjacency matrix of a DAG, in training. In light of the associated compute time, we did not undertake a full hyperparameter search for IMLE_None, and instead only vary a Domke–related hyperparameter to which I-MLE can be sensitive. Generally digraph training performs marginally better on SHD, and often markedly better on prec and rec. This may be because it was not practical to optimise hyperparameters for DAG training. The results can be found in appendix D.7.
Figure 2: Tests of eight methods against selected graph type (rows) and numbers of nodes \(d\) (columns). The two panels show \(\text{nSHD}_c\) (left) for which low values are good, and \(\text{prec}_c\) (right) for which high values are good. Vertical axis scales vary by panel, and, in the left–hand panel, also by row. The ‘box–and–whiskers’ indicate spread over 24 test DAGs (see appendix D.5 for details).

5 Conclusion and future work

For linear data, DAG-DB mostly outperforms combinatoric methods tested here, but is itself out–performed by the continuous methods. The DAG-DB framework should adapt to data generated by non–linear models and, potentially, to discrete data, or to interventional causal models [18]. It would also be interesting to see if the SFE-like backpropagation [31] used in SDI [27] would work for DAG-DB. Separately, one of the I-MLE paper’s experiments [1] built a variational autoencoder (VAE), and a similar experiment could be performed with a latent space of DAGs. A VAE might learn a single DAG for a given dataset, and also learn values for each data point on that DAG’s nodes. This might be interpretable as learning a discrete hierarchy representing characteristics of the dataset.

References


Checklist

1. For all authors...
   - Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
A Notation and terminology

Notation. We write vectors in bold, like $x$. Matrices are in bold capitals, like $Z$ or $\Theta$. Random variables, like $U$, are in non–bold capitals even if they yield vectors. Graph–related objects appear in sans serif, as in $D$ or $j$. Non–standard sets are in calligraphic capitals, such as $\mathcal{R}$. Components of vectors and elements of matrices retain their bold typeface and their case, for example, $z_i$ or $\Theta_{ij}$.

Graph terminology. To fix terminology, a directed graph, or digraph, $G = (V, E)$ comprises a set of nodes $V$ and a set of edges $E \subseteq V \times V$. An edge $e = (i, j) \in E$ may be represented as either $i \rightarrow j$ or $j \leftarrow i$. We do not allow self–loops $i \rightarrow i$. A directed cycle is an aligned sequence of edges $C = i_0 \rightarrow i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_{c-1} \rightarrow i_0$, for some positive integer, $c$.

A directed acyclic graph (DAG) is a digraph with no directed cycles. A partially–directed graph (PDAG) is a partially–directed graph with no directed cycles: cycles including undirected edges are allowed [see [19] p. 82].

A Bayesian network, $M = (D, U)$ consists of a DAG $D = (V, E)$ and a vector of random variables indexed by the nodes $U = (U_j)_{j \in V}$, such that,

$$p(U) = \prod_{j \in V} p(U_j \mid \{U_i : i \rightarrow j \in E\}), \quad (6)$$

so the probability for $U_j$ at a node $j$ only directly depends on the values of $U$ at the j’s ‘parent’ nodes.
B Discrete backpropagation

**Sampling.** In the forward pass, via eq. (1), DAG-DB generates i.i.d. P&M samples $Z^{(s)} = \text{MAP}(\Theta + \tau \Psi^{(s)})$, $s = 1, \ldots, S$, of the binary adjacency matrix, where $\tau$ is the exponential family distribution temperature from eq. (2). For each matrix element $ij$, the noise samples $\Psi^{(s)}_{ij}$ are generated by the standard logistic distribution. DAG-DB then uses a method such as STE or I-MLE to backpropagate the gradient of resulting empirical loss function, $L = S^{-1} \sum_{s=1}^{S} \ell(Z^{(s)})$, from the samples $Z^{(s)}$ to $\Theta$.

Blackbox estimation (BB) \(^{29}\) can be considered as a special case of I-MLE: with no P&M sampling, it considers a purely deterministic function from $\Theta$ to $Z = \text{MAP}(\Theta)$ \(^{11}\). BB’s inability to deliver good results suggests that the sampling element of DAG-DB plays a useful role. Tracking $\Theta$ components shows that $p(Z; \Theta)$ becomes more concentrated around its MAP value as training develops: an initial, broad distribution helps exploration, and, as training progresses, the distribution concentrates around the predicted digraph $Z$.

The maximum size constraint show the versatility of P&M. With no maximum size constraint, P&M logistic noise sampling is the same as independent Bernoulli sampling for each (non–diagonal) component $Z_{ij}$ [see, for example, \(^{32}\)]. With such a constraint, Bernoulli sampling is not possible because the $Z_{ij}$ components are no longer independent.

**Straight–through estimation (STE).** The STE approximation for backpropagation is given by

$$\nabla_{\Theta} L \approx \tau^{-1} \sum_{s=1}^{S} \nabla Z^{(s)} \ell,$$

(7)

This is as in the original treatments of STE \(^{2,3}\), with the inclusion of a factor to account for the temperature.

**Implicit maximum likelihood estimation (I-MLE).** I-MLE \(^{11}\) uses a procedure due to Domke \(^{30}\) to set target distribution parameters, which, for DAG-DB, are, for $s = 1, \ldots, S$,

$$\Theta^{(s)} = \Theta - \lambda \nabla Z^{(s)} \ell,$$

(8)

where $\lambda$ arises from the Domke procedure. In I-MLE, $\lambda$ is treated a hyperparameter, with the approximation of the loss gradient with respect to $\Theta$ then being

$$\nabla_{\Theta} L \approx \frac{1}{\lambda \tau S} \sum_{s=1}^{S} \left[ \text{MAP}(\Theta + \tau \Psi^{(s)}) - \text{MAP}(\Theta^{(s)} + \tau \Psi^{(s)}) \right].$$

(9)

C Details of graphification

Each data point is combined with every sample $Z = Z^{(s)}$ by the ‘graphification’

$$(x \circ Z)_{ij} := x; Z_{ij}.$$  

(10)

The resulting matrix is then fed into a function $f_{\Phi}$, where the $\Phi$ are learnable parameters. In principle $f_{\Phi}$ could be a multi–layer neural network, for predictions on non–linear Bayesian networks, but, for simplicity, we confine attention to a linear function with no bias, having $\Phi \in \mathbb{R}$ as real–matrix parameters, and

$$[f_{\Phi}(M)]_{ij} := \sum_{i} \Phi_{ij} M_{ij}, \quad \text{(with no sum over j)},$$

(11)

for a $d \times d$ matrix $M$. Given the digraph $Z$, define the parents of node $j$ as $\text{pa}_{Z}(j) := \{i : Z_{ij} = 1\}$. Equations (10) and (11) then together imply a prediction for the value of $x$ at node $j$ from the values of $x$ its parent nodes,

$$\tilde{x}_{j} := [f_{\Phi}(x \circ Z)]_{j} = \sum_{i \in \text{pa}_{Z}(j)} x; \Phi_{ij}, \quad \text{(with no sum over j)}.$$  

(12)

This ensures that, in prediction, a node $j$’s value will only be a function of values associated with its parent nodes.
D Details of experiments

D.1 Synthetic DAGs and data

Summary. Synthetic data is often used for learning hyperparameters and testing as it can be easily and cheaply generated. We follow the approach in the NOTEARS paper [4], also adopted by GOLEM [10] and many other papers. The approach is to generate a random DAG, based on an underlying distribution for randomly-generating graphs or digraphs, and then to randomly-generate an appropriate linear additive noise model (LANM) on that DAG. The LANM is then used to create the synthetic dataset, the n x d matrix \(X\). To generate synthetic DAGs and data, we used open-source code from the GOLEM paper [39], itself derived from NOTEARS [40].

Generating DAGs. The NOTEARS and GOLEM papers consider two types of randomly-generated DAGs: Erdős–Rényi ER\(k\) DAGs, where the number of edges is binomially distributed, and in expectation, there are \(dk\) edges; and Barabási–Albert SF\(k\) DAGs, where the number of edges is fixed, at \(k(d - (k + 1)/2)\), with some nodes being preferred in terms of having more edges. We use both these types in our experiments.

Generating data. We use a standard Gaussian equal-variance linear additive noise model. Such a model, with variance \(\sigma^2 \in \mathbb{R}^+\), is a Bayesian network \(M = (D, U)\) with values \(u \sim U\) having each node-component \(u_j \in \mathbb{R}\), and satisfying

\[
    u_j = \left( \sum_{i \in \text{pa}(j)} \Phi_{ij} u_i \right) + \nu_j, \quad \nu_j \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.,}
\]

for some weights matrix \(\Phi\). The NOTEARS/GOLEM implementation, which we use, creates such a model by sampling \(\Phi_{ij}\) i.i.d. uniformly randomly from \([-2, -0.5] \cup [0.5, 2]\).

Potential limitations. There have been criticisms of this approach. In 2008, it was noted that a similar approach to DAG generation does not generate DAGs uniformly [41], s. 3.3], and this issue also affects the current DAG-generation method. It has also, more recently, been pointed out that linear additive noise models offer clues to prediction methods which may be undesirable. These clues are tendencies for variances and co-variances between nodes to increase as we follow the DAG’s edges [42]. Nonetheless these approaches remain the standard synthetic data benchmark and we retain their use, supplemented by our real data experiment, which may mitigate these concerns.

D.2 Metrics

CPDAGs. Different DAGs may represent the same Bayesian network, being, in general, impossible to distinguish simply by observation — that is, by i.i.d. sampling of the Bayesian network’s random variable. The Markov Equivalence Class (MEC) of a DAG \(D\) is formed by the DAGs \(D'\) for which there exists a Bayesian network which may be represented by both \(D\) and \(D'\). A MEC may be represented by the DAG’s class partially-directed acyclic graph (CPDAG) in which some of the DAG’s edges become undirected. All DAGs in a MEC have the same skeleton: that is, the same edges, ignoring direction. Any MEC can be represented by a CPDAG: undirected edges indicate that the direction may differ between DAGs within the MEC.

Identifiability. For some particular Bayesian networks, however, the exact DAG can be identified. Gaussian equal-variance linear noise models have this property. For real data, such as the Sachs data we experiment on, there is no such guarantee. For consistency, we will therefore use metrics based on CPDAGs. More discussion of identifiability can be found in Peters, Janzing & Schölkopf’s textbook [19], sec. 7.1.2–4).

Structural Hamming distance (SHD). Suppose now that we have a true DAG \(D_{\text{true}} = (V, E_{\text{true}})\) and a predicted DAG \(D_{\text{pred}} = (V, E_{\text{pred}})\), on the same set of nodes \(V\), but with potentially differing edge sets \(E_{\text{true}}\) and \(E_{\text{pred}}\). A given unordered pair of nodes \(\{i, j\}\)’s join status will be one of the following: unjoined, joined \(i \rightarrow j\), or joined \(i \leftarrow j\). If we also allow undirected edges, as in CPDAGs, then this adds a fourth possibility \(i \leftrightarrow j\). SHD is the count of the ordered pairs \(i, j\) which differ in
terms of join status between $D_{\text{true}}$ and $D_{\text{pred}}$. The class structural Hamming distance $\text{SHD}_c$ is the corresponding metric for the CPDAGs associated with $D_{\text{true}}$ and $D_{\text{pred}}$. Dividing by the number of nodes gives the normalized class structural Hamming distance $n\text{SHD}_c$. $n\text{SHD}$ is found in, for example, the NOTEARS [4] and GOLEM [10] papers, and $(n)\text{SHD}_c$ in the GOLEM paper. Note that if we predicted an empty DAG $D_{\text{pred}}$, with no edges, we would have $n\text{SHD}_c = |E_{\text{true}}|/|V|$, because a CPDAG has the same skeleton, and so the same number of edges, as any DAG it represents.

**Precision and recall.** In the current context, precision is defined as

$$\text{prec}(D_{\text{true}}, D_{\text{pred}}) = \frac{|E_{\text{pred}} \cap E_{\text{true}}|}{\max(1, |E_{\text{pred}}|)},$$

(14)

 treating the edges $i \rightarrow j$ and $i \leftarrow j$ as distinct. Similarly, recall is defined as

$$\text{rec}(D_{\text{true}}, D_{\text{pred}}) = \frac{|E_{\text{pred}} \cap E_{\text{true}}|}{\max(1, |E_{\text{true}}|)}.$$  

(15)

In graph literature [e.g. 4, 10], precision may be replaced by the false discovery rate, which is one minus the precision, and recall may be termed the true positive rate. As for SHD, we may also use the CPDAGs associated with $D_{\text{true}}$ and $D_{\text{pred}}$ to define the class precision $\text{prec}_c$ and class recall $\text{rec}_c$.

In this extension to CPDAGs, we regard $i \rightarrow j$ as a further distinct type of edge between $i$ and $j$.

**D.3 Hyperparameters**

**Hyperparameter settings.** To choose hyperparameters, we sought to optimise on a randomly-generated set of six synthetic Erdős–Rényi ‘ER2’ DAGs with $d = 30$ nodes and a mean of 60 edges. On each of these graphs, a synthetic dataset was created using a Gaussian equal–variance linear additive noise model. We assessed the viability of I-MLE, STE, SFE and BB as the discrete backpropagation method, finding I-MLE and STE to be the most promising. In exploring hyperparameter settings, we used a combination of Optuna [43, 44], a well–known package for Bayesian optimisation of hyperparameters, and manual adjustment, for example taking an Optuna–generated set of hyperparameters and varying the standard probability distribution used as noise for P&M. We also performed a grid search over hyperparameters, which was less successful in identifying optimal settings. We identified two settings as particularly promising, STE, using STE with a maximum size constraint of 84 edges, and IMLE, using I-MLE with no maximum size constraint.

We chose synthetic ER2 DAGs with 30 nodes, as lying roughly in the middle of the range of types and node–numbers considered here, and in the NOTEARS and GOLEM papers. Note that, following appendix D.2, a DAG with no edges would score an expected $n\text{SHD}_c = 2$ when compared with a “true” ER2 DAG, suggesting a cut–off maximum $n\text{SHD}_c$. With BB, we did not find any settings which delivered $n\text{SHD}_c < 2$; whilst for SFE, we only managed $n\text{SHD}_c$ slightly less than 2. Table 2 shows the best hyperparameters for DAG-DB we found for STE and IMLE.

**D.4 Ablation of regularizers**

We can see from tables 3 and 4 that, for both STE and IMLE, all ablations worsen the metrics, although, for IMLE, the effect of ablating $\tau_{\text{qp}}$ is very minor, as might be expected from table 2, which shows that its coefficient $\rho_{\text{qp}}$ is rather small.

**D.5 Variation by DAG**

The standard box–and–whisker plots of fig. 2 indicate variation over 24 test DAGs. Each box covers from the first to the third quartile, with an interior line indicating the median. A white dot indicates the mean. The whiskers each extend to cover all points within a further 1.5 times the inter–quartile range, whilst any points outside that extended range are marked individually.

We also considered whether 24 DAGs was enough to give a reasonable sample, experimenting via a test set of 240 ER2 DAGs with 30 nodes. We found that taking a 100 000 random 24 DAG sub–samples of the 240 gave standard errors for $n\text{SHD}_c$ of 0.1 for STE and 0.15 for IMLE, and of 0.03 for $\text{prec}_c$ with either method. We then examined how much results could vary in repeated prediction on the same 24 DAGs. We made predictions ten times for a constant set of 24 DAGs,
## Table 2: Hyperparameters for DAG-DB.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>STE_84</th>
<th>IMLE_None</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>Total number of data points in data matrix ( X )</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>( \text{Number of epochs training} )</td>
<td></td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>( \text{Whether to shuffle batches for each new epoch} )</td>
<td>True</td>
<td>True</td>
<td></td>
</tr>
<tr>
<td>( \text{Batch size} )</td>
<td>16</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>( S )</td>
<td>Number of P&amp;M samples</td>
<td>10</td>
<td>47</td>
</tr>
<tr>
<td>( \tau )</td>
<td>Temperature in the exponential family distribution of eq. (2)</td>
<td>( 1.771 \times 10^{-1} )</td>
<td>( 8.786 \times 10^{-1} )</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>I-MLE Domke hyperparameter</td>
<td>n/a</td>
<td>2.714 \times 10^4</td>
</tr>
<tr>
<td>( M )</td>
<td>Width of initial uniform distribution for each ( \Theta ) component. (Note that ( \Phi ) is initialised as for torch.nn.Linear layers)</td>
<td>( 2.169 \times 10^{-1} )</td>
<td>( 1.137 \times 10^{-4} )</td>
</tr>
<tr>
<td>( \rho_{\text{DAG}} )</td>
<td>Coefficient for the ( r_{\text{DAG}} ) regularizer</td>
<td>( 4.101 \times 10^{-1} )</td>
<td>( 1.575 \times 10^{-1} )</td>
</tr>
<tr>
<td>( \rho_{\Phi} )</td>
<td>Coefficient for the ( r_{\Phi} ) regularizer</td>
<td>( 1.023 \times 10^{-2} )</td>
<td>( 1.208 \times 10^{-3} )</td>
</tr>
</tbody>
</table>

finding variation from minimum to maximum for \( n_{\text{SHD}_c} \) of 0.03 for STE_84 and 0.05 for IMLE_None, and of 0.01 for \( \text{prec}_c \) with either method. These results suggest that 24 DAGs may be a sufficient sample size for testing our methods.

### D.6 Methods used for comparison

We follow the methods used for comparison in the GOLEM paper [10], with one exception. The combinatoric methods we use are FGES [24] (formerly known as FGS [45]) and PC [22]. The GOLEM paper used the Conservative PC algorithm [46] as its version of PC, however, this produces results in a form more general than a CPDAG, complicating comparisons. We therefore use the PC-Stable [23] variant, which, like the original PC algorithm, returns a CPDAG. The continuous methods we compare with are NOTEARS [4] and GOLEM [10]. For NOTEARS, as was done in the GOLEM paper, we employ the NOTEARS-L1 variant with the settings given as defaults for linear SEMs in the NOTEARS code repository. GOLEM provides two variants: GOLEM-EV, suitable for data generated with equal variance, and GOLEM-NV, suitable for non–equal variance data. Accordingly, we use GOLEM-EV for our synthetic experiments and GOLEM-NV for our real data experiment.

To run experiments for these methods, we used py-causal [47] for FGES and PC, and the NOTEARS [40] and GOLEM [39] repos. Licences for these packages, and for the code we adapted for GFAS, are shown in table 5.
Table 3: Effects of ablation for STE_84, with crosses indicating ablation of the relevant regularizer.

<table>
<thead>
<tr>
<th>MAX_SIZE = 84</th>
<th>r_DAG</th>
<th>r_sp</th>
<th>nSHD_c</th>
<th>prec_c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>3.901</td>
<td>0.282</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.022</td>
<td>0.724</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>2.068</td>
<td>0.481</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.708</td>
<td>0.552</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.406</td>
<td>0.565</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.286</td>
<td>0.643</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.919</td>
<td>0.719</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.732</td>
<td>0.814</td>
</tr>
</tbody>
</table>

Table 4: Effects of ablation for IMLE_None. Crosses are as in table 3, but note that, in this table, MAX_SIZE = None is marked by ✓, while ‘ablation’ to MAX_SIZE = 66, which was found to be a useful maximum size for I-MLE, is marked by X.

<table>
<thead>
<tr>
<th>MAX_SIZE = None</th>
<th>r_DAG</th>
<th>r_sp</th>
<th>nSHD_c</th>
<th>prec_c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.425</td>
<td>0.584</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.411</td>
<td>0.589</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.268</td>
<td>0.637</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.293</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.293</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.311</td>
<td>0.619</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.064</td>
<td>0.733</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>1.058</td>
<td>0.736</td>
</tr>
</tbody>
</table>

Table 5: Licences for comparison methods and GFAS.

<table>
<thead>
<tr>
<th>Code</th>
<th>Ref.</th>
<th>Link to licence description</th>
</tr>
</thead>
<tbody>
<tr>
<td>py-causal</td>
<td>47</td>
<td>GNU Lesser General Public License v2.1+</td>
</tr>
<tr>
<td>NOTEARS</td>
<td>40</td>
<td>Apache License 2.0</td>
</tr>
<tr>
<td>GOLEM</td>
<td>39</td>
<td>Apache License 2.0</td>
</tr>
<tr>
<td>GFAS</td>
<td>35</td>
<td>BSD License</td>
</tr>
</tbody>
</table>

D.7 Supplementary results on Sachs experiments

See table 6 for results from I-MLE, with training using digraphs, IMLE_None, and with training using DAGs, IMLE_None_Tr. Table 7 shows standard deviations, over 24 runs, for these methods. The varied hyperparameter, $\lambda$, is the Domke hyperparameter discussed in appendix B. We varied this as, in I-MLE, results can be particularly sensitive to $\lambda$. Table 7's $\lambda = 27.14$ rows for IMLE_None are the standard deviations corresponding to table 1. GOLEM and NOTEARS results have very small standard deviations, whilst FGES and PC are deterministic.

As noted, DAG training is time consuming. Running time for each single prediction increases from about 0h15m with IMLE_None to around 1h15m with IMLE_None_Tr.

D.8 Computational resources

Computations were done on a combination of a laptop with a GeForce RTX 3080 Mobile 16GB GPU and a computer cluster. Prediction for a single ER2 graph on 30 nodes (roughly the mid–range of the graphs shown in fig. 2) took around 0h12 on the laptop, for the longest–running of all the methods in fig. 2 IMLE_None, giving around 72 hours to compile that element of the figure. The corresponding STE_84 computation took around a quarter of that time, and the other methods we compared with were quicker, giving a total time to compute fig. 2 of the order of 100 hours. Hyperparameter exploration, ablation experiments and Sachs experiments are estimated to add another 200 hours compute, while the experiments on variation described in appendix D.5 took around 120 hours.
Table 6: As for table I but comparing variants with and without DAGs in training for a range of values of $\lambda$. For each $\lambda$, the first row is the mean and the second row is the median. $\lambda = 27.14$ is the default setting for IMLE_None.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>SHD_c</th>
<th>prec_c</th>
<th>rec_c</th>
<th>pred. size</th>
<th>SHD_c</th>
<th>prec_c</th>
<th>rec_c</th>
<th>pred. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>12.0</td>
<td>1.000</td>
<td>0.294</td>
<td>5.0</td>
<td>13.3</td>
<td>0.778</td>
<td>0.314</td>
<td>7.0</td>
</tr>
<tr>
<td>12</td>
<td>1.000</td>
<td>0.294</td>
<td>5</td>
<td>13.0</td>
<td>0.750</td>
<td>0.294</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>12.5</td>
<td>0.950</td>
<td>0.267</td>
<td>4.8</td>
<td>12.3</td>
<td>0.950</td>
<td>0.275</td>
<td>4.9</td>
</tr>
<tr>
<td>12</td>
<td>1.000</td>
<td>0.294</td>
<td>5</td>
<td>12</td>
<td>1.000</td>
<td>0.294</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>12.3</td>
<td>0.823</td>
<td>0.275</td>
<td>5.7</td>
<td>13.6</td>
<td>0.584</td>
<td>0.208</td>
<td>6.1</td>
</tr>
<tr>
<td>12.5</td>
<td>1.000</td>
<td>0.265</td>
<td>6</td>
<td>14</td>
<td>0.536</td>
<td>0.206</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>12.1</td>
<td>0.894</td>
<td>0.287</td>
<td>5.5</td>
<td>13.2</td>
<td>0.520</td>
<td>0.292</td>
<td>9.7</td>
</tr>
<tr>
<td>12</td>
<td>1.000</td>
<td>0.294</td>
<td>6</td>
<td>13</td>
<td>0.500</td>
<td>0.294</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>27.14</td>
<td>12.7</td>
<td>0.869</td>
<td>0.255</td>
<td>5.0</td>
<td>13.8</td>
<td>0.475</td>
<td>0.287</td>
<td>10.5</td>
</tr>
<tr>
<td>13</td>
<td>1.000</td>
<td>0.235</td>
<td>5</td>
<td>14</td>
<td>0.455</td>
<td>0.294</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>100.0</td>
<td>12.1</td>
<td>0.894</td>
<td>0.287</td>
<td>5.5</td>
<td>14.0</td>
<td>0.451</td>
<td>0.279</td>
<td>10.8</td>
</tr>
<tr>
<td>12</td>
<td>1.000</td>
<td>0.294</td>
<td>5</td>
<td>14</td>
<td>0.455</td>
<td>0.294</td>
<td>10.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: As for table I but showing standard deviations over the 24 runs.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>SHD_c</th>
<th>prec_c</th>
<th>rec_c</th>
<th>pred. size</th>
<th>SHD_c</th>
<th>prec_c</th>
<th>rec_c</th>
<th>pred. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.0</td>
<td>1.0</td>
<td>0.113</td>
<td>0.028</td>
<td>1.2</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.135</td>
<td>0.046</td>
<td>0.5</td>
<td>1.0</td>
<td>0.135</td>
<td>0.059</td>
<td>0.8</td>
</tr>
<tr>
<td>1.0</td>
<td>1.3</td>
<td>0.198</td>
<td>0.077</td>
<td>0.8</td>
<td>1.1</td>
<td>1.175</td>
<td>0.057</td>
<td>0.9</td>
</tr>
<tr>
<td>10.0</td>
<td>1.3</td>
<td>0.169</td>
<td>0.076</td>
<td>1.0</td>
<td>0.6</td>
<td>0.061</td>
<td>0.012</td>
<td>1.2</td>
</tr>
<tr>
<td>27.14</td>
<td>1.2</td>
<td>0.189</td>
<td>0.073</td>
<td>0.9</td>
<td>0.8</td>
<td>0.074</td>
<td>0.020</td>
<td>1.5</td>
</tr>
<tr>
<td>100.0</td>
<td>1.2</td>
<td>0.186</td>
<td>0.073</td>
<td>0.6</td>
<td>0.9</td>
<td>0.076</td>
<td>0.031</td>
<td>1.8</td>
</tr>
</tbody>
</table>

*aOur data reports confirm that the zero standard deviation results for IMLE_None are indeed distinct runs.*