Supplementary material for Network structure learning under uncertain interventions

Federico Castelletti

Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan

and

Stefano Peluso

Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Milan

Sections 1 and 2 contain proofs of Propositions 3.1 and 3.2. In Section 3 we show and discuss Proposition 3.3. Sections 4 and 5 provide details about the proposed MCMC scheme and additional simulated results. Finally, Sections 6 and 7 contain sensitivity analyses to hyperparameter choices and further results for the two real data applications in the main text.

1 Proof of Proposition 3.1 and related discussion

Proposition 3.1. *For* $\mathcal{K}_j(\mathcal{D}) := a_j^{\mathcal{D}}/\boldsymbol{U}_{jj|\prec j\succ}$ *we have*

$$\gamma_j^{(k)}(\boldsymbol{X} \mid \mathcal{A}_j^k)/n^{(k)} \xrightarrow{d} \mathcal{N}\left(\frac{1}{2}\left(\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)} - 1 + \mathrm{tr}\boldsymbol{\Sigma}_{0\prec j\succ}/g - \ln\left(\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)}\right) + C_1\right), \right.$$
$$\left. \left(\phi_{0j}^{(k)}\mathcal{K}_j(\mathcal{D}) - 1\right)^2/(2n^{(k)}) + \mathrm{tr}\boldsymbol{\Sigma}_{0\prec j\succ}^2/(2n^{(k)})\right),$$

where $C_1 = \ln(a_j^{\mathcal{D}}/2) - \psi\left(a_j^{\mathcal{D}}/2\right)$. *Furthermore* $\gamma_j^{(k)}(\boldsymbol{X} \mid \mathcal{A}_j^k) \xrightarrow{a.s.} +\infty$.

Proof. First we can write

$$\begin{aligned}
\tilde{\gamma}_j^{(k)}(\boldsymbol{X},\boldsymbol{\theta}) &= \ln\frac{\pi_k(j)}{1-\pi_k(j)} + \sum_{i=1}^{n^{(k)}} \ln \frac{\varphi\left(x_{ij}^{(k)}|0,\phi_j^{(k)}\right)}{\varphi\left(x_{ij}^{(k)}|-\boldsymbol{L}_{\prec j]}^\top \boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(j)}^{(k)},\sigma_j^2\right)} \\
&= \ln\frac{\pi_k(j)}{1-\pi_k(j)} - \frac{1}{2}(\phi_j^{(k)})^{-1}\sum_{i=1}^{n^{(k)}}(x_{ij}^{(k)})^2 + \frac{n^{(k)}}{2}\ln\frac{\sigma_j^2}{\phi_j^{(k)}} \\
&\quad + \frac{1}{2}\sigma_j^{-2}\sum_{i=1}^{n^{(k)}}(x_{ij}^{(k)} + \boldsymbol{L}_{\prec j]}^\top \boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(j)}^{(k)})^2.
\end{aligned}$$

To derive the quantity of interest $\gamma_j^{(k)}(\boldsymbol{X}\,|\,\mathcal{A}_j^k)$ note that $(\boldsymbol{D},\boldsymbol{L})\,|\,\boldsymbol{X}^{(k)},\mathcal{A}_j^k \overset{d}{=} (\boldsymbol{D},\boldsymbol{L})$ as in (6), and

$$\phi_j^{(k)}\,|\,\boldsymbol{X}^{(k)},\mathcal{A}_j^k \sim \text{I-Ga}\left(a_j^{(k)} + n^{(k)}/2, b_j^{(k)} + \sum_i (x_{ij}^{(k)})^2/2\right).$$

Therefore we can derive $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\mathcal{A}_j^k}\left[1/\sigma_j^2\right] = \mathcal{K}_j(\mathcal{D})$, $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\mathcal{A}_j^k}\left[\ln\sigma_j^2\right] = \ln\left(\boldsymbol{U}_{jj|\prec j\succ}/2\right) - \psi\left(a_j^{\mathcal{D}}/2\right)$, where ψ is the digamma function, and, recalling $\psi(x) \asymp \ln x$ for large x,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\mathcal{A}_j^k}\left[1/\phi_j^{(k)}\right] &= \left[2a_j^{(k)} + n^{(k)}\right] / \left[2b_j^{(k)} + \sum_i (x_{ij}^{(k)})^2\right] \overset{a.s.}{\to} 1/\phi_{0j}^{(k)}, \\
\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\mathcal{A}_j^k}\left[\ln\phi_j^{(k)}\right] &= \ln\left(b_j^{(k)} + \sum_i (x_{ij}^{(k)})^2/2\right) - \psi\left(a_j^{(k)} + n^{(k)}/2\right) \overset{a.s.}{\to} \ln\phi_{0j}^{(k)}, \\
\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\mathcal{A}_j^k}\left[\boldsymbol{L}_{\prec j]}/\sigma_j^2\right] &= -\mathcal{K}_j(\mathcal{D})\boldsymbol{U}_{\prec j\succ}^{-1}\boldsymbol{U}_{\prec j]} = \boldsymbol{0}, \\
\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\mathcal{A}_j^k}\left[\boldsymbol{L}_{\prec j]}\boldsymbol{L}'_{\prec j]}/\sigma_j^2\right] &= \boldsymbol{U}_{\prec j\succ}^{-1} + \mathcal{K}_j(\mathcal{D})\boldsymbol{U}_{\prec j\succ}^{-1}\boldsymbol{U}_{\prec j]}(\boldsymbol{U}_{\prec j\succ}^{-1}\boldsymbol{U}_{\prec j]})^\top = \boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}/g,
\end{aligned}$$

from which, with $\bar{X}_{2j}^{(k)} := \sum_i (x_{ij}^{(k)})^2/n^{(k)}$ and $\bar{\boldsymbol{X}}_{2\prec j\succ}^{(k)} := \sum_i \boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(j)}^{(k)}\left(\boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(j)}^{(k)}\right)^\top/n^{(k)}$,

$$\begin{aligned}
\gamma_j^{(k)}(\boldsymbol{X}\,|\,\mathcal{A}_j^k)/n^{(k)} &\asymp \frac{1}{2}\left[\mathcal{K}_j(\mathcal{D}) - 1/\phi_{0j}^{(k)}\right]\bar{X}_{2j}^{(k)} + \frac{1}{2g}\mathrm{tr}\bar{\boldsymbol{X}}_{2\prec j\succ}^{(k)} - \frac{1}{2}\ln 2\boldsymbol{U}_{jj|\prec j\succ}^{-1}\phi_{0j}^{(k)} - \frac{1}{2}\psi\left(a_j^{\mathcal{D}}/2\right) \\
&\overset{a.s.}{\to} \frac{1}{2}\left(\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)} - 1 + \mathrm{tr}\boldsymbol{\Sigma}_{0\prec j\succ}/g - \ln\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)} + C_1\right). \qquad (1)
\end{aligned}$$

Since $\psi(x) < \ln x$ for all $x > 1/2$, it is true that $C_1 > 0$, since $a_j^{\mathcal{D}} \geq 1$ from the discussion following Equation (6). Also, $\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)} - 1 - \ln\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)} \geq 0$ for all possible values of the terms involved, with equality holding for $\mathcal{K}_j(\mathcal{D})^{-1} = \phi_{0j}^{(k)}$. Then the quantity in (1) is strictly positive, showing $\gamma_j^{(k)}(\boldsymbol{X}\,|\,\mathcal{A}_j^k) \overset{a.s.}{\to} +\infty$.

To prove asymptotic normality, denote vec the matrix vectorization (and vec^{-1} its inverse function), and note that from the above discussion $\gamma_j^{(k)}(\boldsymbol{X}\,|\,\mathcal{A}_j^k)/n^{(k)} \asymp h\left(\bar{X}_{2j}^{(k)}, \mathrm{vec}\,\bar{\boldsymbol{X}}_{2\prec j\succ}^{(k)}\right)$, where $h(x,y) = \left(\mathcal{K}_j(\mathcal{D})x - 1 + \mathrm{tr}\,\mathrm{vec}^{-1}y - \ln\mathcal{K}_j(\mathcal{D})x + C_1\right)/2$, with derivatives computed as $h'_x(x,y) = \partial h(x,y)/\partial x = (\mathcal{K}_j(\mathcal{D}) - 1/x)/2$ and $h'_y(x,y) = \partial h(x,y)/\partial y = \mathrm{vec}\,\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}/2$. Since $\bar{X}_{2j}^{(k)} \overset{d}{\to} \mathcal{N}\left(\phi_{0j}^{(k)}, 2(\phi_{0j}^{(k)})^2/n^{(k)}\right)$ and $\mathrm{vec}\,\bar{\boldsymbol{X}}_{2\prec j\succ}^{(k)} \overset{d}{\to} \mathcal{N}\left(\boldsymbol{\Sigma}_{0\prec j\succ}, 2\left(\boldsymbol{\Sigma}_{0\prec j\succ}\otimes\boldsymbol{\Sigma}_{0\prec j\succ}\right)/n^{(k)}\right)$ (see for

instance Ghazal and Neudecker 2000), with null covariance between $\bar{X}_{2j}^{(k)}$ and $\operatorname{vec} \bar{\boldsymbol{X}}_{2\prec j\succ}^{(k)}$, from the Delta method we have

$$\gamma_j^{(k)}(\boldsymbol{X} \mid \mathcal{A}_j^k)/n^{(k)} \xrightarrow{d} \mathcal{N}\Bigg(h\left(\phi_{0j}^{(k)}, \operatorname{vec} \boldsymbol{\Sigma}_{0\prec j\succ}\right), 2(\phi_{0j}^{(k)})^2/n^{(k)} \cdot h_x'\left(\phi_{0j}^{(k)}, \operatorname{vec} \boldsymbol{\Sigma}_{0\prec j\succ}\right)^2 + 2/n^{(k)} h_y'\left(\phi_{0j}^{(k)}, \operatorname{vec} \boldsymbol{\Sigma}_{0\prec j\succ}\right)' \left(\boldsymbol{\Sigma}_{0\prec j\succ} \otimes \boldsymbol{\Sigma}_{0\prec j\succ}\right) h_y'\left(\phi_{0j}^{(k)}, \operatorname{vec} \boldsymbol{\Sigma}_{0\prec j\succ}\right)\Bigg)$$

corresponding to the distribution in the statement. □

The proposition states that, if $j \in I_k$, i.e. node j is a target under intervention k, this will be detected with sample size large enough and for any given graph $\mathcal{D}$, therefore with a false negative rate eventually zero. The scaled log-odds of the (correct) target classification has asymptotic Gaussian distribution. From the normality and by replacing $\phi_{0j}^{(k)}$ and $\boldsymbol{\Sigma}_{0\prec j\succ}$ by some estimators, we can further provide a reasonable range within which the posterior log-odds is expected to be for a given fixed sample size. Also note that $\phi_{0j}^{(k)} \to 0$ corresponds to the limit case of deterministic (non stochastic) interventions. In our result this scenario clearly coincides with a degenerate normal distribution with infinite mean, since there is no uncertainty in the log-odds of an intervention.

Note that the mean of the asymptotic distriution increases when $a_j^{\mathcal{D}}$ is large. This is typical of a node with many parents in a large graph, that, coherently to intuition, makes easier the identification of an intervention, since the latter will suppress many dependence relations. With $C_1 \approx 0$ ($a_j^{\mathcal{D}}$ large), we would have that for $\mathcal{K}_j(\mathcal{D})^{-1} = \phi_{0j}^{(k)}$, $\gamma_j^{(k)}(\boldsymbol{X} \mid \mathcal{A}_j^k) \xrightarrow{a.s.} \operatorname{tr}\boldsymbol{\Sigma}_{0\prec j\succ}/(2g)$: with many parents and equal variances with or without intervention, target discrimination is still feasible thanks to the variability of the node parents. Then the extreme case when it is not possible to identify an intervention is when $a_j^{\mathcal{D}}$ large, pre- and post-intervention variances coincide, and either (a) the node has no parents or (b) the parents have a degenerate distribution with null variances.

2 Proof of Proposition 3.2 and related discussion

Proposition 3.2. *For* $\delta_{jk} := \frac{a_j^{(k)}}{b_j^{(k)}}\sigma_{0j}^2 - 1$ *we have*

$$\gamma_j^{(k)}(\boldsymbol{X} \mid \bar{\mathcal{A}}_j^k)/n^{(k)} \xrightarrow{d} \mathcal{N}\Bigg(\frac{1}{2}\left(\frac{a_j^{(k)}}{b_j^{(k)}}\boldsymbol{\Sigma}_{0jj} - 1 - \ln\frac{a_j^{(k)}}{b_j^{(k)}}\sigma_{0j}^2 + C_2\right), \left(\frac{\boldsymbol{\Sigma}_{0jj}}{\sigma_{0j}^2}\delta_{jk}\right)^2/(2n^{(k)}) + \frac{\boldsymbol{\Sigma}_{0jj}^2 - \sigma_{0j}^4}{\sigma_{0j}^4}\left[\delta_{jk}/n^{(k)} + \left(\frac{5\boldsymbol{\Sigma}_{0jj} - 3\sigma_{0j}^2}{\boldsymbol{\Sigma}_{0jj} + \sigma_{0j}^2}\right)/(2n^{(k)})\right]\Bigg),$$

where $C_2 = \ln\left(a_j^{(k)}\right) - \psi\left(a_j^{(k)}\right)$. *Furthermore* $\gamma_j^{(k)}(\boldsymbol{X} \mid \bar{\mathcal{A}}_j^k) \xrightarrow{a.s.} +\infty$.

Proof. From Ben-David et al. (2015) we know that

$$\begin{aligned}
\sigma_j^2 \,|\, \boldsymbol{X}^{(k)} &\sim \text{I-Ga}\left(\frac{1}{2}\left(a_j^{\mathcal{D}} + n^{(k)}\right), \frac{1}{2}\boldsymbol{U}^{(k)}_{jj|\prec j\succ}\right), \\
\boldsymbol{L}_{\prec j\,]} \,|\, \sigma_j^2, \boldsymbol{X}^{(k)} &\sim \mathcal{N}_{|\text{pa}_{\mathcal{D}}(j)|}\left(-\left(\boldsymbol{U}^{(k)}_{\prec j\succ}\right)^{-1}\boldsymbol{U}^{(k)}_{\prec j\,]}, \sigma_j^2\left(\boldsymbol{U}^{(k)}_{\prec j\succ}\right)^{-1}\right),
\end{aligned}$$

where $\boldsymbol{U}^{(k)} := \boldsymbol{U} + (\boldsymbol{X}^{(k)})'\boldsymbol{X}^{(k)}$, and we have that $\boldsymbol{\Phi}^{(k)}|\boldsymbol{X}^{(k)}, \bar{\mathcal{A}}_j^k \overset{d}{=} \boldsymbol{\Phi}^{(k)}$ as in (7) in the main text. Therefore we can first compute $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[1/\phi_j^{(k)}\right] = a_j^{(k)}/b_j^{(k)}$ and $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\ln \phi_j^{(k)}\right] = \ln b_j^{(k)} - \psi\left(a_j^{(k)}\right)$ Furthermore, with $\mathcal{K}_j^{(k)}(\mathcal{D}) := (a_j^{\mathcal{D}} + n^{(k)})/\boldsymbol{U}^{(k)}_{jj|\prec j\succ}$, we can derive $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[1/\sigma_j^2\right] = \mathcal{K}_j^{(k)}(\mathcal{D}) \overset{a.s.}{\to} \sigma_{0j}^{-1}$, and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\ln \sigma_j^2\right] &= \ln\left(\boldsymbol{U}^{(k)}_{jj|\prec j\succ}/2\right) - \psi\left(\frac{1}{2}\left(a_j^{\mathcal{D}} + n^{(k)}\right)\right) \overset{a.s.}{\to} \ln \sigma_{0j}^2, \\
\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\boldsymbol{L}_{\prec j\,]}/\sigma_j^2\right] &= -\mathcal{K}_j^{(k)}(\mathcal{D})\left(\boldsymbol{U}^{(k)}_{\prec j\succ}\right)^{-1}\boldsymbol{U}^{(k)}_{\prec j\,]} \overset{a.s.}{\to} -\sigma_{0j}^{-2}\boldsymbol{\Sigma}^{-1}_{0\prec j\succ}\boldsymbol{\Sigma}_{0\prec j\,]}, \\
\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\boldsymbol{L}_{\prec j\,]}\boldsymbol{L}'_{\prec j\,]}/\sigma_j^2\right] &= \left(\boldsymbol{U}^{(k)}_{\prec j\succ}\right)^{-1} + \mathcal{K}_j^{(k)}(\mathcal{D})\left(\boldsymbol{U}^{(k)}_{\prec j\succ}\right)^{-1}\boldsymbol{U}^{(k)}_{\prec j\,]}\boldsymbol{U}^{(k)\prime}_{\prec j\,]}\left(\boldsymbol{U}^{(k)}_{\prec j\succ}\right)^{-1} \\
&\overset{a.s.}{\to} \sigma_{0j}^{-2}\boldsymbol{\Sigma}^{-1}_{0\prec j\succ}\boldsymbol{\Sigma}_{0\prec j\,]}\boldsymbol{\Sigma}'_{0\prec j\,]}\boldsymbol{\Sigma}^{-1}_{0\prec j\succ}
\end{aligned}$$

Then, with $\bar{X}_{2j}^{(k)} := \sum_i (x_{ij}^{(k)})^2/n^{(k)} \overset{a.s.}{\to} \boldsymbol{\Sigma}_{0jj}$, $\text{vec}\,\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ} := \text{vec}\sum_i \boldsymbol{x}^{(k)}_{i,\text{pa}_{\mathcal{D}}(j)}\left(\boldsymbol{x}^{(k)}_{i,\text{pa}_{\mathcal{D}}(j)}\right)^{\top}/n^{(k)} \overset{a.s.}{\to} \text{vec}\,\boldsymbol{\Sigma}_{0\prec j\succ}$ and $\bar{\boldsymbol{X}}^{(k)}_{2[j\succ} := \sum_{i=1}^{n^{(k)}} x_{ij}^{(k)}\left(\boldsymbol{x}^{(k)}_{i,\text{pa}_{\mathcal{D}}(j)}\right)^{\top}/n^{(k)} \overset{a.s.}{\to} \boldsymbol{\Sigma}_{0\,[j\succ}$,

$$\begin{aligned}
\bar{\gamma}_j^{(k)}(\boldsymbol{X}|\bar{\mathcal{A}}_j^k)/n^{(k)} &\asymp \frac{1}{2}\bar{X}_{2j}^{(k)}\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[1/\phi_j^{(k)}\right] - \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\ln\sigma_j^2\right] - \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\ln\phi_j^{(k)}\right]\right) \\
&\quad -\frac{1}{2}\bar{X}_{2j}^{(k)}\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[1/\sigma_j^2\right] - \frac{1}{2}\text{tr}\left\{\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ}\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\boldsymbol{L}_{\prec j\,]}\boldsymbol{L}'_{\prec j\,]}/\sigma_j^2\right]\right\} \\
&\quad -\bar{\boldsymbol{X}}^{(k)}_{2[j\succ}\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[\boldsymbol{L}_{\prec j\,]}/\sigma_j^2\right] \\
&\overset{a.s.}{\to} \frac{1}{2}\frac{a_j^{(k)}}{b_j^{(k)}}\boldsymbol{\Sigma}_{0jj} - \frac{1}{2}\left(\ln\sigma_{0j}^2 - \ln b_j^{(k)} + \psi\left(a_j^{(k)}\right)\right) - \frac{1}{2}\frac{\boldsymbol{\Sigma}_{0jj}}{\sigma_{0j}^2} \\
&\quad -\frac{1}{2}\sigma_{0j}^{-2}\text{tr}\left\{\boldsymbol{\Sigma}_{0\prec j\,]}\boldsymbol{\Sigma}'_{0\prec j\,]}\boldsymbol{\Sigma}^{-1}_{0\prec j\succ}\right\} + \sigma_{0j}^{-2}\boldsymbol{\Sigma}'_{0\prec j\,]}\boldsymbol{\Sigma}^{-1}_{0\prec j\succ}\boldsymbol{\Sigma}_{0\prec j\,]} \\
&= \frac{1}{2}\left(\frac{a_j^{(k)}}{b_j^{(k)}}\boldsymbol{\Sigma}_{0jj} - 1 - \ln\frac{a_j^{(k)}}{b_j^{(k)}}\sigma_{0j}^2 + C_2\right), \qquad (2)
\end{aligned}$$

where we have used $\boldsymbol{\Sigma}'_{0\prec j\,]}\boldsymbol{\Sigma}^{-1}_{0\prec j\succ}\boldsymbol{\Sigma}_{0\prec j\,]} = \boldsymbol{\Sigma}_{0jj} - \sigma_{0j}^2$. Since $\psi(x) < \ln x$ for all $x > 1/2$, it is true that $C_2 > 0$, since $a_j^{(k)} \geq 1$ from the discussion following Equation (7) in the main text. Also, $a_j^{(k)}/b_j^{(k)}\boldsymbol{\Sigma}_{0jj} - 1 - \ln a_j^{(k)}/b_j^{(k)}\sigma_{0j}^2 \geq 0$ for all possible values of the terms involved, with equality holding for $a_j^{(k)}/b_j^{(k)} = \boldsymbol{\Sigma}_{0jj} = \sigma_{0j}^2$. Then (2) is strictly positive, showing $\bar{\gamma}_j^{(k)}(\boldsymbol{X}|\bar{\mathcal{A}}_j^k) \overset{a.s.}{\to} +\infty$.

To show the asymptotic distribution, first note that we can write the quantity of interest as $\bar{\gamma}_j^{(k)}(\boldsymbol{X}|\bar{\mathcal{A}}_j^k)/n^{(k)} = \bar{h}\left(\bar{X}_{2j}^{(k)}, \text{vec}\,\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ}, \bar{\boldsymbol{X}}^{(k)}_{2\prec j]}\right)$, where vec is the matrix vectorization (and

vec^{-1} its inverse function) and $\bar{\boldsymbol{X}}^{(k)}_{2\prec j]} = (\bar{\boldsymbol{X}}^{(k)}_{2[j\succ})^\top$, and where the function $\bar{h}$ is defined as

$$\begin{aligned}
\bar{h}(x,y,z) &= \frac{1}{2}\left(\frac{a_j^{(k)}}{b_j^{(k)}} - u(x,y,z)^{-1}\right)x - \frac{1}{2}\mathrm{tr}\left(\frac{zz'\omega(y)^{-1}}{u(x,y,z)}\right) \\
&\quad + \frac{1}{u(x,y,z)} z'w(y)^{-1}z - \frac{1}{2}\left(\ln u(x,y,z) - \ln b_j^{(k)} + \psi\left(a_j^{(k)}\right)\right) \\
&= \frac{1}{2}\left(\frac{a_j^{(k)}}{b_j^{(k)}}x - 1 - \ln\frac{a_j^{(k)}}{b_j^{(k)}}u(x,y,z) + C_2\right), \qquad (3)
\end{aligned}$$

with $u(x,y,z) = x - z^\top\omega(y)^{-1}z$ and $\omega(y) = \mathrm{vec}^{-1}(y)$. The partial derivatives of $\bar{h}$ can be recovered as $\bar{h}'_x(x,y,z) = \partial\bar{h}(x,y,z)/\partial x = (a_j^{(k)}/b_j^{(k)} - 1/u(x,y,z))/2$, $\bar{h}'_y(x,y,z) = \partial\bar{h}(x,y,z)/\partial y = -\mathrm{vec}(\omega(y)^{-1}zz^\top\omega(y)^{-1}/u(x,y,z))/2$ and $\bar{h}'_z(x,y,z) = \partial\bar{h}(x,y,z)/\partial z = \omega(y)^{-1}z/u(x,y,z)$. It is then easy to see, using $\boldsymbol{L}_{0\prec j]} = -\boldsymbol{\Sigma}^{-1}_{0\prec j\succ}\boldsymbol{\Sigma}_{0\prec j]}$, that

$$\nabla\bar{g}\left(\bar{X}^{(k)}_{2j}, \mathrm{vec}\,\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ}, \bar{\boldsymbol{X}}^{(k)}_{2\prec j]}\right)^\top \overset{a.s.}{\to} \frac{1}{2\sigma^2_{0j}}\left(\frac{a_j^{(k)}}{b_j^{(k)}}\sigma^2_{0j} - 1,\ -\mathrm{vec}\left\{\boldsymbol{L}_{0\prec j]}\boldsymbol{L}_{0[j\succ}\right\}^\top, -2\boldsymbol{L}_{0[j\succ}\right)^\top \qquad (4)$$

Furthermore, from the properties of higher-order moments of the multivariate Gaussian distribution, see for instance Ghazal and Neudecker (2000), we know that

$$\begin{aligned}
\bar{X}^{(k)}_{2j} &\overset{d}{\to} \mathcal{N}_1\left(\boldsymbol{\Sigma}_{0jj}, \frac{2}{n^{(k)}}\boldsymbol{\Sigma}^2_{0jj}\right), \\
\mathrm{vec}\,\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ} &\overset{d}{\to} \mathcal{N}_{|\mathrm{pa}_{\mathcal{D}}(j)|^2}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_{0\prec j\succ}, \frac{2}{n^{(k)}}\boldsymbol{\Sigma}_{0\prec j\succ}\otimes\boldsymbol{\Sigma}_{0\prec j\succ}\right), \qquad (5) \\
\bar{\boldsymbol{X}}^{(k)}_{2\prec j]} &\overset{d}{\to} \mathcal{N}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\left(\boldsymbol{\Sigma}_{0\prec j]}, \frac{1}{n^{(k)}}\left[\boldsymbol{\Sigma}_{0jj}\boldsymbol{\Sigma}_{0\prec j\succ} + \boldsymbol{\Sigma}_{0\prec j]}\boldsymbol{\Sigma}_{0[j\succ}\right]\right),
\end{aligned}$$

where $\otimes$ is the Kronecker product. The cross-covariance matrices can be derived as follows:

$$\begin{aligned}
\mathbb{C}\left(\bar{X}^{(k)}_{2j}, \mathrm{vec}\,\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ}\right) &= \frac{1}{n^{(k)}}\mathbb{C}\left(\left(x^{(k)}_{ij}\right)^2, \mathrm{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^\top\right\}\right) \\
&= \frac{1}{n^{(k)}}\mathbb{C}\left(\mathbb{E}\left[\left(x^{(k)}_{ij}\right)^2\middle|\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}\right], \mathrm{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^\top\right\}\right) \\
&= \frac{1}{n^{(k)}}\mathbb{C}\left(\boldsymbol{L}_{0[j\succ}\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^\top\boldsymbol{L}_{0\prec j]}, \mathrm{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^\top\right\}\right) \\
&= \frac{1}{n^{(k)}}\mathbb{C}\left(\left(\boldsymbol{L}_{0[j\succ}\otimes\boldsymbol{L}_{0[j\succ}\right)\mathrm{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^\top\right\}, \mathrm{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^\top\right\}\right) \\
&= \frac{1}{n^{(k)}}\left(\boldsymbol{L}_{0[j\succ}\otimes\boldsymbol{L}_{0[j\succ}\right)\mathbb{V}\left(\mathrm{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^\top\right\}\right) \\
&= \frac{2}{n^{(k)}}\left(\boldsymbol{L}_{0[j\succ}\boldsymbol{\Sigma}_{0\prec j\succ}\right)\otimes\left(\boldsymbol{L}_{0[j\succ}\boldsymbol{\Sigma}_{0\prec j\succ}\right) = \frac{2}{n^{(k)}}\boldsymbol{\Sigma}_{0[j\succ}\otimes\boldsymbol{\Sigma}_{0[j\succ}, \qquad (6)
\end{aligned}$$

where we have used $\mathrm{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top\otimes\boldsymbol{A})\,\mathrm{vec}\,\boldsymbol{B}$ and $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = (\boldsymbol{AC})\otimes(\boldsymbol{BD})$ for

conformable matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$. Similarly, we see that

$$\begin{aligned}
\mathbb{C}\left(\bar{\boldsymbol{X}}^{(k)}_{2\prec j]}, \operatorname{vec}\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ}\right) &= \frac{1}{n^{(k)}}\mathbb{C}\left(x^{(k)}_{ij}\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}, \operatorname{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\right\}\right) \\
&= -\frac{1}{n^{(k)}}\mathbb{C}\left(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\boldsymbol{L}_{0\prec j], \operatorname{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\right\}\right) \\
&= -\frac{1}{n^{(k)}}\mathbb{C}\left(\left(\boldsymbol{L}_{0[j\succ}\otimes\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\right)\operatorname{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\right\},\right. \\
&\qquad\left.\operatorname{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\right\}\right) \\
&= -\frac{2}{n^{(k)}}\left(\boldsymbol{L}_{0[j\succ}\boldsymbol{\Sigma}_{0\prec j\succ}\right)\otimes\boldsymbol{\Sigma}_{0\prec j\succ} = \frac{2}{n^{(k)}}\boldsymbol{\Sigma}_{0[j\succ}\otimes\boldsymbol{\Sigma}_{0\prec j\succ}, \qquad (7)
\end{aligned}$$

where we have used the relation $\operatorname{vec}(\boldsymbol{AB}) = (\boldsymbol{B}^{\top}\otimes\boldsymbol{I})\operatorname{vec}\boldsymbol{A}$, and

$$\begin{aligned}
\mathbb{C}\left(\bar{X}^{(k)}_{2j}, \bar{\boldsymbol{X}}^{(k)}_{2\prec j]}\right) &= \frac{1}{n^{(k)}}\mathbb{C}\left(\left(x^{(k)}_{ij}\right)^2, x^{(k)}_{ij}\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}\right) \\
&= \frac{1}{n^{(k)}}\mathbb{E}\left[\mathbb{C}\left(\left(x^{(k)}_{ij}\right)^2, x^{(k)}_{ij}\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}\middle|\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}\right)\right] \\
&\quad+\frac{1}{n^{(k)}}\mathbb{C}\left(\mathbb{E}\left[\left(x^{(k)}_{ij}\right)^2\middle|\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}\right], \boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}\mathbb{E}\left[x^{(k)}_{ij}\middle|\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}\right]\right) \\
&= \frac{1}{n^{(k)}}\left\{-2\sigma^2_{0j}\boldsymbol{L}_{0[j\succ}\mathbb{E}\left[\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\right]+\right. \\
&\qquad\left.+\,\mathbb{C}\left(\boldsymbol{L}_{0[j\succ}\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\boldsymbol{L}_{0\prec j], -\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\boldsymbol{L}_{0\prec j]}\right)\right\} \\
&= \frac{1}{n^{(k)}}\left\{2\sigma^2_{0j}\boldsymbol{\Sigma}_{0[j\succ}-\mathbb{C}\left(\left(\boldsymbol{L}_{0[j\succ}\otimes\boldsymbol{L}_{0[j\succ}\right)\operatorname{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\right\},\right.\right. \\
&\qquad\left.\left.\left(\boldsymbol{L}_{0[j\succ}\otimes\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\right)\operatorname{vec}\left\{\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)}(\boldsymbol{x}^{(k)}_{i,\mathrm{pa}_{\mathcal{D}}(j)})^{\top}\right\}\right)\right\} \\
&= \frac{2}{n^{(k)}}\left\{\sigma^2_{0j}\boldsymbol{\Sigma}_{0[j\succ}-\left(\boldsymbol{L}_{0[j\succ}\otimes\boldsymbol{L}_{0[j\succ}\right)\left(\boldsymbol{\Sigma}_{0\prec j\succ}\otimes\boldsymbol{\Sigma}_{0\prec j\succ}\right)\left(\boldsymbol{L}_{0\prec j]}\otimes\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\right)\right\} \\
&= \frac{2}{n^{(k)}}\left\{\sigma^2_{0j}\boldsymbol{\Sigma}_{0[j\succ}-\left(\boldsymbol{L}_{0[j\succ}\boldsymbol{\Sigma}_{0\prec j\succ}\boldsymbol{L}_{0\prec j]}\right)\left(\boldsymbol{L}_{0[j\succ}\boldsymbol{\Sigma}_{0\prec j\succ}\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\right)\right\} \\
&= \frac{2}{n^{(k)}}\boldsymbol{\Sigma}_{0jj}\boldsymbol{\Sigma}_{0[j\succ}, \qquad (8)
\end{aligned}$$

using $(\boldsymbol{A}\otimes\boldsymbol{B})^{\top} = \boldsymbol{A}^{\top}\otimes\boldsymbol{B}^{\top}$ and $\boldsymbol{\Sigma}_{0jj}-\sigma^2_{0j} = \boldsymbol{L}_{0[j\succ}\boldsymbol{\Sigma}_{0\prec j\succ}\boldsymbol{L}_{0\prec j]}$. Putting together the results in (3)-(8), by Delta method that we have that

$$\begin{aligned}
\bar{h}\left(\bar{X}^{(k)}_{2j}, \operatorname{vec}\bar{\boldsymbol{X}}^{(k)}_{2\prec j\succ}, \bar{\boldsymbol{X}}^{(k)}_{2\prec j]}\right) \xrightarrow{d} \mathcal{N}\Big(&\bar{h}\left(\boldsymbol{\Sigma}_{0jj}, \operatorname{vec}\boldsymbol{\Sigma}_{0\prec j\succ}, \boldsymbol{\Sigma}_{0\prec j]}\right), \\
&\frac{2}{n^{(k)}}\nabla\bar{h}\left(\boldsymbol{\Sigma}_{0jj}, \operatorname{vec}\boldsymbol{\Sigma}_{0\prec j\succ}, \boldsymbol{\Sigma}_{0\prec j]}\right)^{\top}\Lambda\nabla\bar{h}\left(\boldsymbol{\Sigma}_{0jj}, \operatorname{vec}\boldsymbol{\Sigma}_{0\prec j\succ}, \boldsymbol{\Sigma}_{0\prec j]}\right)\Big),
\end{aligned}$$

where

$$\Lambda = \begin{pmatrix}
\boldsymbol{\Sigma}^2_{0jj} & \boldsymbol{\Sigma}_{0[j\succ}\otimes\boldsymbol{\Sigma}_{0[j\succ} & \boldsymbol{\Sigma}_{0[j\succ}\otimes\boldsymbol{\Sigma}_{0jj} \\
\boldsymbol{\Sigma}_{0\prec j]}\otimes\boldsymbol{\Sigma}_{0\prec j]} & \boldsymbol{\Sigma}_{0\prec j\succ}\otimes\boldsymbol{\Sigma}_{0\prec j\succ} & \boldsymbol{\Sigma}_{0\prec j]}\otimes\boldsymbol{\Sigma}_{0\prec j\succ} \\
\boldsymbol{\Sigma}_{0\prec j]}\otimes\boldsymbol{\Sigma}_{0jj} & \boldsymbol{\Sigma}_{0[j\succ}\otimes\boldsymbol{\Sigma}_{0\prec j\succ} & \frac{1}{2}\left(\boldsymbol{\Sigma}_{0\prec j\succ}\boldsymbol{\Sigma}_{0jj}+\boldsymbol{\Sigma}_{0\prec j]}\boldsymbol{\Sigma}_{0[j\succ}\right)
\end{pmatrix}$$

and this distribution corresponds to the stated result. □

Proposition 3.2 tells that a true negative case of no intervention, i.e. $j \notin I_k$, will be eventually detected with sample size large enough. Note that the mean of the asymptotic distribution is closer to zero when node j is independent from any other node, that is $|\mathrm{pa}_{\mathcal{D}}(j)| = \emptyset$, and therefore $\boldsymbol{\Sigma}_{0jj} = \sigma_{0j}^2$. Intuitively, it is more difficult to understand the absence of an intervention since there are no parent-child relations that are removed by the intervention on node j, which makes the intervention effect less apparent. In this special case, the second addend in the asymptotic variance disappears, and the whole asymptotic distribution becomes very similar, *mutatis mutandis*, to the one recovered in Proposition 3.1.

3 Graph Consistency

In the current section we prove model selection consistency of the true DAG observational equivalence class; the latter, combined with consistent estimation of targets, allows to identify the group-specific intervention graphs. In the following lemma and proposition, we first extend the conjugacy result on the DAG-Wishart prior of Ben-David et al. (2015) to interventional Gaussian multivariate data from multiple groups; then we prove, following Cao et al. (2019) and Peluso and Consonni (2020), its Bayes factor and posterior ratio consistency outside $[\mathcal{D}_0]$, the equivalence class of the true DAG, and its asymptotic compatibility within $[\mathcal{D}_0]$. Let $\boldsymbol{P}^k$ be a $q \times q$ diagonal matrix such that $\boldsymbol{P}^k_{jj} = 1$ if $j \notin I_k$, that is node j is not intervened in group k.

Lemma 3.1. *Let $\mathcal{D}_0$ be the true DAG. Assume $(\boldsymbol{D}, \boldsymbol{L}) \,|\, \mathcal{D}$ follows a DAG-Wishart distribution with hyperparameters $\boldsymbol{U}$ and $\boldsymbol{a}^{\mathcal{D}}$ and consider the likelihood function $f\big(\boldsymbol{X} \,|\, \boldsymbol{\theta}, I_1, \ldots, I_K, \mathcal{D}\big)$ defined as in (5) of the main text. A posteriori we have that $(\boldsymbol{D}, \boldsymbol{L}) \,|\, \mathcal{D}, \boldsymbol{X}, I_1, \ldots, I_K$ is also DAG-Wishart with hyperparameters $n\tilde{\boldsymbol{S}}$ and $\boldsymbol{a} + \boldsymbol{n}^*$, where $\tilde{\boldsymbol{S}} := \boldsymbol{U}/n + \sum_k \boldsymbol{P}^k \boldsymbol{X}^{(k)\top} \boldsymbol{X}^{(k)} \boldsymbol{P}^k / n$ and $\boldsymbol{n}^* := \sum_k \mathrm{diag}\{\boldsymbol{P}^k\} n^{(k)}$.*

Proof. The DAG-Wishart (prior) distribution has density

$$\frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}^{\mathcal{D}})} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top\right)\boldsymbol{U}\right)\right\} \prod_{j \in V} \boldsymbol{D}_{jj}^{-\frac{a_j^{\mathcal{D}}}{2}}$$

for all $(\boldsymbol{D}, \boldsymbol{L})$ in the Cholesky space, with $\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}^{\mathcal{D}})$ being the normalizing constant. In the posterior distribution, we can write the likelihood $f\big(\boldsymbol{X} \,|\, \boldsymbol{\theta}, I_1, \ldots, I_K, \mathcal{D}\big)$ as proportional to

$$\begin{aligned}
\prod_{k=1}^{K} p(\boldsymbol{X}^{(k)} \,|\, \boldsymbol{D}, \boldsymbol{L}, I_1, \ldots, I_K,) \;&\propto\; \prod_{k=1}^{K} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top\right)\boldsymbol{P}^k\boldsymbol{S}^{(k)}\boldsymbol{P}^k\right)\right\} \prod_{j \in V} \boldsymbol{D}_{jj}^{-\frac{\boldsymbol{P}^k_{jj} n^{(k)}}{2}} \\
&= \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top\right)\sum_k \boldsymbol{P}^k\boldsymbol{S}^{(k)}\boldsymbol{P}^k\right)\right\} \prod_{j \in V} \boldsymbol{D}_{jj}^{-\sum_k \frac{\boldsymbol{P}^k_{jj} n^{(k)}}{2}},
\end{aligned}$$

with $\boldsymbol{S}^{(k)} = (\boldsymbol{X}^{(k)})^\top \boldsymbol{X}^{(k)}$, where we omitted the interventional terms in the likelihood not dependent on $\boldsymbol{D}$ and $\boldsymbol{L}$. The DAG-Wishart (posterior) distribution of $(\boldsymbol{D}, \boldsymbol{L}) \,|\, \mathcal{D}, \boldsymbol{X}, I_1, \ldots, I_K$ follows immediately from Bayes theorem by combining prior and likelihood. □

We have *Bayes factor consistency* if, for all $\mathcal{D} \neq \mathcal{D}_0$, the Bayes factor

$$BF_{\mathcal{D},\mathcal{D}_0} = \frac{m(\boldsymbol{X} \,|\, \mathcal{D}, I_1, \ldots, I_K)}{m(\boldsymbol{X} \,|\, \mathcal{D}_0, I_1, \ldots, I_K)} \xrightarrow{\bar{P}} 0,$$

whenever $\mathcal{D}_0$ is the true DAG generating $\boldsymbol{X}$, where $\xrightarrow{\bar{P}}$ denotes convergence in probability, $\bar{P}$ is the probability measure under the true DAG $\mathcal{D}_0$, and $m(\boldsymbol{X} \,|\, \mathcal{D}, I_1, \ldots, I_k)$ is the marginal (or integrated) likelihood. Define also $\tilde{\boldsymbol{n}} = n - \boldsymbol{n}^*$. On the other hand, we have *posterior ratio consistency* if, with $\mathcal{D}_0$ being the true DAG, it holds that, as $n \to \infty$,

$$\max_{\mathcal{D} \neq \mathcal{D}_0} \frac{p(\mathcal{D} \,|\, \boldsymbol{X}, I_1, \ldots, I_K)}{p(\mathcal{D}_0 \,|\, \boldsymbol{X}, I_1, \ldots, I_K)} = \max_{\mathcal{D} \neq \mathcal{D}_0} BF_{\mathcal{D},\mathcal{D}_0}(\boldsymbol{X} \,|\, I_1, \ldots, I_k) \frac{p(\mathcal{D})}{p(\mathcal{D}_0)} \xrightarrow{\bar{P}} 0. \tag{9}$$

Proposition 3.3. *Let $\mathcal{D}_0$ be the true DAG. Assume $(\boldsymbol{D}, \boldsymbol{L}) \,|\, \mathcal{D}$ follows a DAG-Wishart distribution with hyperparameters $\boldsymbol{U}$ and $\boldsymbol{a}^{\mathcal{D}}$ as in Equation* (8) *of the main text, and consider the likelihood function $f\big(\boldsymbol{X} \,|\, \boldsymbol{\theta}, I_1, \ldots, I_K, \mathcal{D}\big)$ defined as in* (5) *of the main text. If (a) $a_j^{\mathcal{D}} = a + |\mathrm{pa}_{\mathcal{D}}(j)| - q + 1$, (b) $\tilde{n}_j = o(n_j^*)$ for all $j \in V$, and (c) for all $j \neq l \in V$ there exists a k such that $j \notin I_k$ and $l \notin I_k$ hold, then as $n \to \infty$,*

$$\begin{aligned} &i) \quad \max_{\mathcal{D} \notin [\mathcal{D}_0]} \frac{p(\mathcal{D} \,|\, \boldsymbol{X}, I_1, \ldots, I_K)}{p(\mathcal{D}_0 \,|\, \boldsymbol{X}, I_1, \ldots, I_K)} \xrightarrow{\bar{P}} 0, \\ &ii) \quad \frac{p(\mathcal{D} \,|\, \boldsymbol{X}, I_1, \ldots, I_K)}{p(\mathcal{D}_0 \,|\, \boldsymbol{X}, I_1, \ldots, I_K)} \xrightarrow{\bar{P}} \frac{p(\mathcal{D})}{p(\mathcal{D}_0)} \quad \textit{for all } \mathcal{D} \in [\mathcal{D}_0]. \end{aligned}$$

Proof. From the conjugacy of the DAG-Wishart distribution (Lemma 3.1), we can write the marginal likelihood of DAG $\mathcal{D}$ as

$$m(\boldsymbol{X} \,|\, \mathcal{D}, I_1, \ldots, I_K) = (2\pi)^{-n/2} \mathcal{Z}_{\mathcal{D}} \left(n\tilde{\boldsymbol{S}}, \boldsymbol{a}^{\mathcal{D}} + \boldsymbol{n}^* \right) / \mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}^{\mathcal{D}}).$$

Accordingly, the Bayes factor between any two DAGs $\mathcal{D}$ and $\mathcal{D}_0$ is

$$BF_{\mathcal{D},\mathcal{D}_0}(\boldsymbol{X} \,|\, I_1, \ldots, I_K) = \frac{\mathcal{Z}_{\mathcal{D}}(n\tilde{\boldsymbol{S}}, \boldsymbol{a}^{\mathcal{D}} + \boldsymbol{n}^*)}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}^{\mathcal{D}})} \frac{\mathcal{Z}_{\mathcal{D}_0}(\boldsymbol{U}, \boldsymbol{a}^{\mathcal{D}_0})}{\mathcal{Z}_{\mathcal{D}_0}(n\tilde{\boldsymbol{S}}, \boldsymbol{a}^{\mathcal{D}} + \boldsymbol{n}^*)}.$$

Since $\tilde{n}_j = o(n_j^*)$ for all $j \in V$, then $(a + n_j)/(a + n_j^*) = (a + n_j^* + \tilde{n}_j)/(a + n_j^*) \to 1$. Furthermore,

$$\begin{aligned} \sum_{k=1}^{K} \boldsymbol{P}^k \boldsymbol{X}^{(k)\top} \boldsymbol{X}^{(k)} \boldsymbol{P}^k &= \left(\sum_{k=1}^{K} \sum_{i=1}^{n^{(k)}} \boldsymbol{x}_{ij}^{(k)} \boldsymbol{x}_{il}^{(k)} \mathbb{1}_{\{j \notin \boldsymbol{I}_k\}} \mathbb{1}_{\{l \notin \boldsymbol{I}_k\}} \right)_{jl} \\ &\asymp \left(\sum_{k=1}^{K} \sum_{i=1}^{n^{(k)}} \boldsymbol{x}_{ij}^{(k)} \boldsymbol{x}_{il}^{(k)} \right)_{jl} = \sum_{k=1}^{K} \boldsymbol{X}^{(k)\top} \boldsymbol{X}^{(k)} = \boldsymbol{X}^\top \boldsymbol{X}, \end{aligned}$$

where the asymptotic equivalence in probability holds from the assumption (c) that for all nodes there exists a setting (group) k where both nodes are not intervened. Then, from Continuity Mapping theorem

$$BF_{\mathcal{D},\mathcal{D}_0}(\boldsymbol{X} \,|\, I_1, \ldots, I_K) \asymp \frac{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U} + \boldsymbol{X}'\boldsymbol{X}, \boldsymbol{a}^{\mathcal{D}} + n)}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}^{\mathcal{D}})} \frac{\mathcal{Z}_{\mathcal{D}_0}(\boldsymbol{U}, \boldsymbol{a}^{\mathcal{D}_0})}{\mathcal{Z}_{\mathcal{D}_0}(\boldsymbol{U} + \boldsymbol{X}'\boldsymbol{X}, \boldsymbol{a}^{\mathcal{D}} + n)},$$

in the same form provided by Cao et al. (2019). Result i) comes therefore from Cao et al. (2019, Theorem 4.1), whilst result ii) comes from assumption b) together with Peluso and Consonni (2020, Proposition 3.5). □

Proposition 3.3 shows that posterior ratio and Bayes factor consistency under DAG-Wishart prior holds *outside* the Markov equivalence class of the true generating DAG $\mathcal{D}_0$. On the other hand, the posterior ratio tends to the prior ratio (Bayes factor equal to one) within the true equivalence class. This result is coherent with Peluso and Consonni (2020), in the context of a single-group observational data. Assumption (a) guarantees *compatibility* within equivalence classes (Geiger and Heckerman, 2002; Peluso and Consonni, 2020), to have Bayes factors between equivalent graphs always equal to one. The removal of this assumption does not break posterior consistency outside $[\mathcal{D}_0]$, but the posterior ratio within $[\mathcal{D}_0]$ would converge to some value dependent on the true parameters, losing compatibility. With assumption (b) we require that, for each node, the number of non-interventional observations for node j will eventually dominate the number of intervened records for the node. Finally, the last assumption (c) intuitively says that, for each pair of nodes, we can estimate their dependence if there exists at least a group in which these two nodes are not intervened.

We further note that assumption (b) on the asymptotic dominance of purely observational data over interventional ones is particularly convenient since it permits to reconduct the problem of graph consistency to the one in Cao et al. (2019) and in Peluso and Consonni (2020) for the comparison, respectively, of (observationally) non-equivalent and equivalent graphs. In this way we guarantee the identification of the correct observational equivalence class and, together with the correct identification of target nodes, we are able to consistently estimate the interventional DAG, up to relations of interventional equivalences. We conjecture that the removal of this assumption, substituted by $n_j^* = \alpha n$ for some $\alpha \in (0, 1)$, that is observational and interventional sample sizes diverging at the same speed, can lead to the more precise result of graph posterior consistency by breaking, within the observational equivalence class of the true DAG, the compatibility among graphs that are not anymore equivalent after the interventions. We refer the reader to Hauser and Bühlmann (2012) for the formal definition of interventional Markov equivalence. Moreover, by Chickering (1995, Theorem 2), there exists a sequence of obervationally Markov equivalent DAGs differing only for the reversal of a *covered edge*, that is for an edge (u, v) of $\mathcal{D}$ for which $\mathrm{pa}_v(\mathcal{D}) = \mathrm{pa}_u(\mathcal{D}) \cup \{u\}$. A comparison between $\mathcal{D}$ and $\mathcal{D}_0$, differing only by one covered edge reversal, could show, when an intervention occurs on u or on v and breaks the equivalence, that compatibility is replaced by posterior ratio consistency.

4 Sampler for posterior inference

4.1 Marginal likelihood derivation

In this section we provide further details on the MCMC scheme adopted for posterior inference on DAGs and intervention targets. Our algorithm is based on a collapsed Metropolis-Hastings sampler (Metropolis et al., 1953), where the DAG parameters $\boldsymbol{\theta} = \left\{\boldsymbol{D}, \boldsymbol{L}, \boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(K)}\right\}$ are integrated out. Accordingly, we first focus on the integrated likelihood

$$m(\boldsymbol{X} \,|\, I_1, \dots, I_K, \mathcal{D}) = \int\limits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} f\big(\boldsymbol{X} \,|\, \boldsymbol{\theta}, I_1, \dots, I_K, \mathcal{D}\big)\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{10}$$

To integrate out all parameters in $\boldsymbol{\theta}$, we exploit the structure of the prior $p(\boldsymbol{\theta})$ and we first re-write Equation (5) in the main text as

$$f\big(\boldsymbol{X} \,|\, \boldsymbol{\theta}, I_1, \dots, I_K, \mathcal{D}\big) = \prod_{j=1}^{q} \left\{ \varphi_{n_j^*}\big(\boldsymbol{X}_j^* \,|\, -\boldsymbol{X}^*_{\mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{L}_{\prec j]}, \sigma_j^2 \boldsymbol{I}_{n_j^*}\big) \prod_{k: j \in I_k} \varphi_{n^{(k)}}\big(\boldsymbol{X}_j^{(k)} \,|\, \boldsymbol{0}, \phi_j^{(k)} \boldsymbol{I}_{n_j^{(k)}}\big) \right\},$$

where $\boldsymbol{X}_j^*$ is the $(n_j^*, 1)$ vector collecting all the observations $\boldsymbol{X}_j^{(k)}$ such that $j \notin I_k$ while $\boldsymbol{X}^*_{\mathrm{pa}_{\mathcal{D}}(j)}$ is the corresponding $n_j^* \times |\mathrm{pa}_{\mathcal{D}}(j)|$ matrix of parent-nodes observations. Accordingly, $n_j^* = \sum_{k: j \notin I_k} n^{(k)}$. Now notice that, because of prior parameter independence, the likelihood admits the same factorization of (8) (see our main text). Therefore, the integrated likelihood can be obtained as $m\big(\boldsymbol{X} \,|\, I_1, \dots, I_K, \mathcal{D}\big)$

$$\begin{aligned} &= \prod_{j=1}^{q} \Bigg\{ \int\limits_0^{\infty} \int\limits_{\mathbb{R}^{|\mathrm{pa}_{\mathcal{D}}(j)|}} \left\{ \varphi_{n_j^*}\big(\boldsymbol{X}_j^* \,|\, -\boldsymbol{X}^*_{\mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{L}_{\prec j]}, \sigma_j^2 \boldsymbol{I}_{n_j^*}\big) \cdot p\big(\boldsymbol{L}_{\prec j]} \,|\, \sigma_j^2\big) p(\sigma_j^2)\, d\boldsymbol{L}_{\prec j]}\, d\sigma_j^2 \right\} \\ &\qquad \cdot \prod_{k: j \in I_k} \int\limits_0^{\infty} \left\{ \varphi_{n^{(k)}}\big(\boldsymbol{X}_j^{(k)} \,|\, \boldsymbol{0}, \phi_j^{(k)} \boldsymbol{I}_{n_j^{(k)}}\big) \cdot p\big(\phi_j^{(k)}\big)\, d\phi_j^{(k)} \right\} \Bigg\}. \end{aligned} \tag{11}$$

Also, because of conjugacy each integral in the product is available in closed-form as the ratio of prior and posterior normalizing constants. Therefore we obtain

$$m\big(\boldsymbol{X} \,|\, I_1, \dots, I_K, \mathcal{D}\big) = \prod_{j=1}^{q} \left\{ m(\boldsymbol{X}_j^* \,|\, \boldsymbol{X}^*_{\mathrm{pa}_{\mathcal{D}}(j)}, \mathcal{D}) \cdot \prod_{k: j \in I_k} m(\boldsymbol{X}_j^{(k)}) \right\}, \tag{12}$$

where

$$\begin{aligned} m\big(\boldsymbol{X}_j^* \,|\, \boldsymbol{X}^*_{\mathrm{pa}_{\mathcal{D}}(j)}, \mathcal{D}\big) &= (2\pi)^{-\frac{n_j^*}{2}} \cdot \frac{\big|g \boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\big|^{1/2}}{\big|g \boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|} + \boldsymbol{S}^*_{\mathrm{pa}_{\mathcal{D}}(j)}\big|^{1/2}} \cdot \frac{\Gamma\left(\frac{a_j^{\mathcal{D}}}{2} + \frac{n_j^*}{2}\right)}{\Gamma\left(\frac{a_j^{\mathcal{D}}}{2}\right)} \cdot \frac{\left(\frac{g}{2}\right)^{a_j^{\mathcal{D}}/2}}{\left(\frac{g + r_j^*}{2}\right)^{(a_j^{\mathcal{D}} + n_j^*)/2}}, \\ m\big(\boldsymbol{X}_j^{(k)}\big) &= (2\pi)^{-\frac{n_j^{(k)}}{2}} \cdot \frac{\Gamma\left(\frac{a_j^{(k)}}{2} + \frac{n_j^{(k)}}{2}\right)}{\Gamma\left(\frac{a_j^{(k)}}{2}\right)} \cdot \frac{\left(\frac{g}{2}\right)^{a_j^{(k)}/2}}{\left(\frac{g + s_j^{(k)}}{2}\right)^{(a_j^{(k)} + n_j^{(k)})/2}}, \end{aligned}$$

and

$$
\begin{aligned}
\boldsymbol{S}^*_{\mathrm{pa}_{\mathcal{D}}(j)} &= \boldsymbol{X}^{*\top}_{\mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{X}^*_{\mathrm{pa}_{\mathcal{D}}(j)}, \\
r^*_j &= \boldsymbol{X}^{*\top}_j \boldsymbol{X}^*_j - \boldsymbol{X}^{*\top}_j \boldsymbol{X}^*_{\mathrm{pa}_{\mathcal{D}}(j)} \left(g \boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|} + \boldsymbol{S}^*_{\mathrm{pa}_{\mathcal{D}}(j)} \right)^{-1} \boldsymbol{X}^{*\top}_{\mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{X}^*_j, \\
s^{(k)}_j &= \boldsymbol{X}^{(k)\top}_j \boldsymbol{X}^{(k)}_j.
\end{aligned}
$$

We then construct a collapsed Metropolis-Hastings sampler (Metropolis et al., 1953) to approximate

$$
p(I_1, \ldots, I_K, \mathcal{D} \,|\, \boldsymbol{X}) \propto m\big(\boldsymbol{X} \,|\, I_1, \ldots, I_K, \mathcal{D}\big) \cdot p(I_1, \ldots, I_K)\, p(\mathcal{D}), \tag{13}
$$

the posterior distribution of DAGs and intervention targets. The output is a collection of DAGs $\{\mathcal{D}^{(s)}\}_{s=1}^S$ and targets $\{I_1^{(s)}, \ldots, I_K^{(s)}\}_{s=1}^S$ approximately sampled from (13), where S is the number of finally kept MCMC iterations. We summarize our MCMC scheme in Algorithm 1 and provide details about the update of DAGs and targets in the following sections.

4.2 Update of $\mathcal{D}$

The full conditional distribution of $\mathcal{D}$ is $p(\mathcal{D} \,|\, I_1, \ldots, I_K, \boldsymbol{X}) \propto m(\boldsymbol{X} \,|\, I_1, \ldots I_K, \mathcal{D})\, p(\mathcal{D})$. Update of DAG $\mathcal{D}$ can be performed through a Metropolis Hastings step, where, given the current DAG, a new DAG $\widetilde{\mathcal{D}}$ is proposed from $q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})$ and accepted with probability

$$
\alpha = \min \left\{ 1; \frac{m\big(\boldsymbol{X} \,|\, I_1, \ldots, I_K, \widetilde{\mathcal{D}}\big)}{m\big(\boldsymbol{X} \,|\, I_1, \ldots, I_K, \mathcal{D}\big)} \cdot \frac{p(\widetilde{\mathcal{D}})}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \,|\, \widetilde{\mathcal{D}})}{q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})} \right\}. \tag{14}
$$

The proposal distribution $q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})$ follows the lines of Chickering (2002). We consider three types of operators that locally modify the input DAG $\mathcal{D}$: insert a directed edge (InsertD $u \to v$ for short), delete a directed edge (DeleteD $u \to v$) and reverse a directed edge (ReverseD $u \to v$). For a given $\mathcal{D}$ we then construct the set of valid operators $\mathcal{O}_{\mathcal{D}}$, that is operators whose resulting graph is a DAG. A DAG $\widetilde{\mathcal{D}}$ is then called a *direct successor* of $\mathcal{D}$ if it can be reached by applying an operator in $\mathcal{O}_{\mathcal{D}}$ to $\mathcal{D}$. Given the current $\mathcal{D}$ we propose $\widetilde{\mathcal{D}}$ by uniformly sampling an element in $\mathcal{O}_{\mathcal{D}}$ and applying it to $\mathcal{D}$. Since there is a one-to-one correspondence between each operator and resulting DAG, the probability of transition is $q(\widetilde{\mathcal{D}} \,|\, \mathcal{D}) = 1/|\mathcal{O}_{\mathcal{D}}|$, for each $\widetilde{\mathcal{D}}$ direct successor of $\mathcal{D}$.

4.3 Update of $I_1, \ldots, I_K$

Conditionally on DAG $\mathcal{D}$ we update the K targets $I_1, \ldots, I_K$ (equivalently, the indicator vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K$) sequentially; see also Section 2.4 in the main text. For a given k, the full conditional of I_k is $p(I_k \,|\, \{I_s\}_{s \neq k}, \mathcal{D}, \boldsymbol{X}) \propto m(\boldsymbol{X} \,|\, I_1, \ldots, I_K, \mathcal{D}) \cdot p(I_1, \ldots, I_K)$ Again, update of I_k conditionally on $\{I_s\}_{s \neq k}$ and DAG $\mathcal{D}$ can be performed through a Metropolis Hastings step, where a

new target $\widetilde{I}_k$ proposed from $q(\widetilde{I}_k \,|\, I_k)$ is accepted with probability

$$\beta_k = \min\left\{1; \frac{m\big(\boldsymbol{X} \,|\, \widetilde{I}_k, \{I_s\}_{s\neq k}, \mathcal{D}\big)}{m\big(\boldsymbol{X} \,|\, I_k, \{I_s\}_{s\neq k}, \mathcal{D}\big)} \cdot \frac{p(\widetilde{I}_k)}{p(I_k)} \cdot \frac{q(I_k \,|\, \widetilde{I}_k)}{q(\widetilde{I}_k \,|\, I_k)}\right\}. \tag{15}$$

Starting from a given target I_k, a candidate target $\widetilde{I}_k$ is obtained by randomly drawing a node $j \in \{1, \ldots, q\}$ which is included in I_k if $j \notin I_k$, otherwise removed. The structure of this proposal is such that $q(I_k \,|\, \widetilde{I}_k)/q(\widetilde{I}_k \,|\, I_k) = 1$ for any I_k and $\widetilde{I}_k$ differing in one j, and equal to 0 otherwise.

Algorithm 1: MCMC scheme to sample from the posterior (13)

Input: K datasets $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$

Output: S samples from the posterior (13)

Initialize $\mathcal{D}^{(0)}$, e.g. the empty DAG, and the targets $I_1, \ldots, I_K$, e.g. $I_k = \emptyset$ for $k = 1, \ldots, K$;

for $s = 1, \ldots, S$ **do**

  Sample $\widetilde{\mathcal{D}}$ from $q(\widetilde{\mathcal{D}} \,|\, \mathcal{D}^{(s-1)})$ and set $\mathcal{D}^{(s)} = \widetilde{\mathcal{D}}$ with probability α in (14),

  otherwise $\mathcal{D}^{(s)} = \mathcal{D}^{(s-1)}$;

  for $k = 1, \ldots, K$ **do**

    Propose $\widetilde{I}_k$ from $q(\widetilde{I}_k \,|\, I_k)$ and set $I_k^{(s)} = \widetilde{I}_k$ with probability β_k in (15),

    otherwise $I_k^{(s)} = I_k^{(s-1)}$;

  end

end

4.4 Computational and convergence issues

We first investigate the computational time of our method as a function of the number of variables q and sample size $n^{(k)}$. Figure 1 summarizes the computational time (averaged over 12 repetitions) *per iteration* for $q \in \{5, 10, 20, 50, 100, 200\}$ and for sample sizes $n^{(k)} \in \{10, 20, 50, 100, 200, 500, 1000\}$. Each dotted line in the left (right) panel of the figure corresponds to a fixed value of q ($n^{(k)}$); moreover, higher curves are associated to higher values of q and $n^{(k)}$ respectively. The behavior of all curves suggests a polynomial dependence of the computational time from both q and $n^{(k)}$. Results were obtained on a PC Intel(R) Core(TM) i7-8550U 1,80 GHz.

Computational challenges are related the huge dimension of the graph space. More precisely, our proposal distribution for DAG moves in the MCMC is *uniform* over the set of direct successors of the current DAG $\mathcal{D}$, a set consisting of all those DAGs obtained from $\mathcal{D}$ by applying singe-edge removals, insertions or reversals. This kind of proposal results in a DAG acceptance rate which is relatively small, especially when the posterior over DAGs is highly concentrated (typically when n is large), meaning that many of the proposed DAGs are discarded by the MCMC. As a consequence, convergence to the target posterior over DAGs can require a higher number of iterations. One possible solution not investigated in the current manuscript

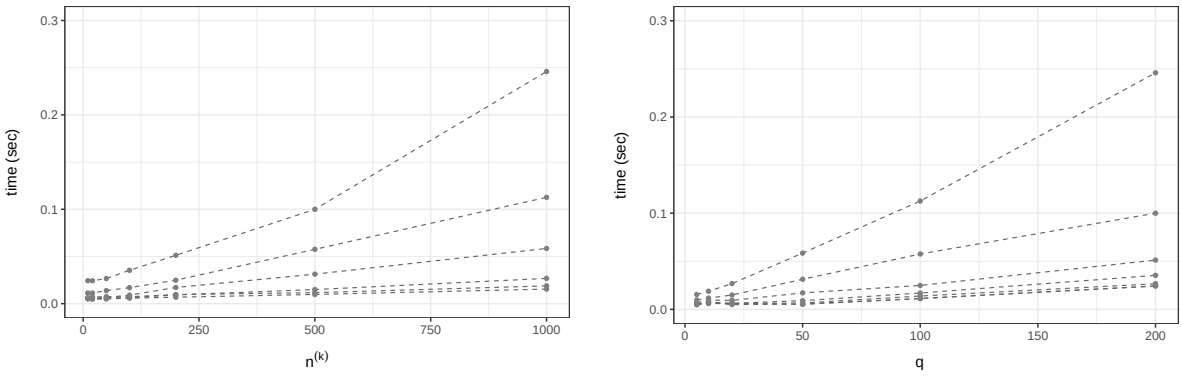

Figure 1: Computational time (in seconds) *per iteration*, as a function of the sample size $n^{(k)}$ (left panel, for increasing values of q) and as a function of q (right panel, for increasing values of $n^{(k)}$), averaged over 12 simulated multiple datasets.

is represented by Feature Inclusion Stochastic Search (FINCS) methods, based on the notion of (marginal) probability of inclusion for a *potential* edge; see for instance Scott and Carvalho (2008) and Altomare et al. (2013) who adopt FINCS for undirected graphs and DAGs with fixed ordering of the nodes respectively. In brief, the intuition is that insertion of edges with a higher posterior probability of inclusion (computed up to time t) is more likely to be accepted. Accordingly, at each step t of the MCMC, moves are proposed with different probabilities, each reflecting the probability of (non) inclusion of those edges which are involved in the move.

In a similar way, alternative choices of the proposal distribution for target nodes can speed up MCMC convergence. More precisely, in the current version of our MCMC scheme, starting from a given target I_k, a candidate target $\widetilde{I}_k$ is obtained by randomly drawing a node $j \in \{1, \ldots, q\}$ which is included in I_k if $j \notin I_k$, otherwise removed from I_k if j already belongs to I_k. The structure of this proposal is such that $q(I_k \,|\, \widetilde{I}_k)/q(\widetilde{I}_k \,|\, I_k) = 1$ for any I_k and $\widetilde{I}_k$ differing in one j, and equal to 0 otherwise. Accordingly, our proposal modifies one target at a time. Alternatively, one can randomly choose a node $j \in \{1, \ldots, q\}$ and propose a new allocation of the node in the K datasets. More precisely, for $j = 1, \ldots, q$, let $\boldsymbol{\xi}_j = (\xi_j(1), \ldots, \xi_j(K))^\top$ be a K-th vector of indicators such that for each $k = 1, \ldots, K$, $\xi_j(k) = 1$ if and only if $j \in I_k$. Similarly as before, we can then sample $j \in \{1, \ldots, q\}$ and propose a new configuration of vector $\boldsymbol{\xi}_j$ given the current one. This kind of move applied for the *simultaneous* update of a set of nodes can result in an improved MCMC mixing and convergence to the posterior of DAGs and targets.

Finally, we assess convergence of the MCMC algorithm by investigating how selected features of the model space behave across iterations, and by evaluating the agreement between independent MCMC chains with different starting points. To this end, we considered some pilot simulations for each setting defined by $q = 20$ and $q = 40$ variables for a number of iterations $S = 25000$ and $S = 50000$ respectively. Specifically, under each simulation, we ran two inde-

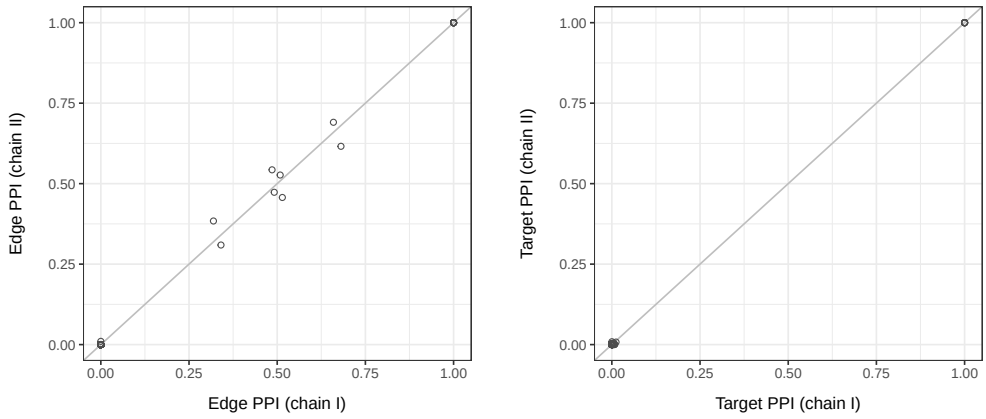

Figure 2: Simulations. Scatter plots of estimated marginal posterior probabilities of edge inclusion (left panel) and marginal posterior probabilities of intervention (target inclusion, right panel) obtained from two independent MCMC chains, chain I and chain II.

pendent chains of same length, with random starting points. We then compare the estimated marginal posterior probabilities of edge inclusion obtained from the two chains as well as the marginal posterior probabilities of intervention (target inclusion); see also Equations (13) and (14) in the main text. As an instance, Figure 2 shows the scatter plots of the estimated posterior probabilities of edge inclusion and intervention obtained from the two chains for one pilot simulation with $q = 20$. By visual inspection, we see that the agreement between the chains is highly satisfactory, since points are clustered around the main diagonal of the plot.

5 More simulated scenarios

We present extensive simulation studies to evaluate the performance of our methodology and comparisons with alternative approaches. We first include the complete results for the simulation study introduced in the main text, with all interventional datasets. Next, we consider additional simulation scenarios including both observational *and* interventional data. Under both cases, we fix the number of groups (interventions) $K = 4$, while we vary the number of variables $q \in \{20, 40\}$ and the number of observations for each group $k = 1, \ldots, K$, $n^{(k)} \in \{10, 20, 50, 100, 200, 500\}$. Under each scenario defined by $(q, n^{(k)})$, we perform 40 simulations, each corresponding to a true DAG $\mathcal{D}$, family of targets $\{I_1, \ldots, I_K\}$ and resulting in a (multiple with $K = 4$ groups) dataset. Specifically, we first randomly generate a topologically ordered DAG $\mathcal{D}$ with probability of edge inclusion $p_{edge} = 2/q$. Generation of targets $I_1, \ldots, I_K$ is specific to each of the two following settings and therefore detailed in the next. Given DAG $\mathcal{D}$ and family of targets $\{I_1, \ldots, I_K\}$, parameters $\boldsymbol{D}$, $\boldsymbol{L}$ and $\boldsymbol{\Phi}^{(k)} = \left\{\phi_j^{(k)}, j \in I_k\right\}$ are generated by fixing $\boldsymbol{D} = \boldsymbol{I}_q$, and $\phi_j^{(k)} = 0.1$ for each $j \in I_k$ and $k = 1, \ldots, K$; non zero elements of $\boldsymbol{L}$ are

instead uniformly chosen in the interval $[-1, -0.1] \cup [0.1, 1]$; see also Equation (4) of the main text.

In the first case with all interventional datasets, a family of intervention targets $I_1, \ldots, I_K$ is generated under two scenarios resembling different degrees of "sparsity" in the targets. Scenario *Sparse* is characterized by a moderate number of interventions, with each target I_k obtained by drawing without replacement $s \in \{2, 4\}$ nodes, respectively for $q = 20$ or $q = 40$. On the other hand, in Scenario *Diffuse* we assign each node to one of the K targets. As a consequence, each variable is involved in one of the K interventions, with an overall larger number of simultaneous interventions (sizes of the targets I_k). Next, for each $k = 1 \ldots, K$, $n^{(k)}$ i.i.d. interventional data collected in the $n^{(k)} \times q$ data matrix $\boldsymbol{X}^{(k)}$ are generated as in Equation (4) of our paper. Each dataset is therefore a collection of K interventional data matrices $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$.

In the following we present results relative to our Bayesian method (*Bayes*) and some benchmark methods. In particular we include the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm of Squires et al. (2020), a non-parametric strategy for target estimation and structure learning based on permutation tests that we implement at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1% and IGSP 0.001% respectively), as recommended in the original paper. Other significance levels are not included since result in worse performances. We also include Algorithm 1 of He and Geng (2016), a method for DAG learning from interventional data with unknown targets, that builds on the PC algorithm of Spirtes et al. (2000). We implement it at significance level 0.05; results obtained at different significance levels and not included led to worse performances. As a further benchmark, we also construct a baseline node-wise regression approach, by adapting to our interventional setting the two-stage adaptive lasso method of Han et al. (2016): we first recover a baseline DAG structure by applying the adaptive lasso to the union of the K data matrices; next, we apply adaptive lasso to each dataset $\boldsymbol{X}^{(k)}$ separately, and estimate intervention targets by comparing overall and group-specific DAGs; we call this benchmark *Node-wise*. Finally, we include the Greedy Interventional Equivalence Search (GIES) method of Hauser and Bühlmann (2012), a search-and-score method based on maximum likelihood estimation, which provides an estimate of the graph representing the interventional Markov equivalence class of the true DAG. GIES was developed for known intervention targets, that we input, as in an *oracle* setting, using the *true* intervened nodes. We implement GIES using the Extended Bayesian Information Criterion (EBIC) with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5 and GIES 1 respectively) as also recommended in Foygel and Drton (2010). All methods can be adopted for DAG structure learning, whilst only IGSP and *Node-wise* can perform target estimation. Finally, notice that IGSP requires $n^{(k)} > q$, so that results of IGPS for $n^{(k)} \in \{10, 20\}$ are missing.

All methods can be adopted for DAG structure learning, whilst only IGSP 0.1, IGSP 0.001% and *Node-wise* can perform target estimation; see the main text for more details. Moreover,

while all the benchmarks directly output single estimates of DAGs (and targets), we adapt our Bayesian method to provide point estimates as described in our paper (Section 5.2). Finally, notice that IGSP requires $n^{(k)} > q$, so that results of IGSP for $n^{(k)} \in \{10, 20\}$ are missing.

A summary for $q = 20$ and $q = 40$ is reported in the box-plots of Figures 3 and 4. This reports the distribution of FPR and FNR constructed across the $N = 40$ simulations for three methods under evaluation and increasing sample sizes $n^{(k)}$ under *Sparse* and *Diffuse* Scenarios. We notice that, coherently with the theoretical results of Section 4 in the main text, for our method both sources of error vanish as sample size increases. This tendency is more evident for FNR that rapidly goes to zero already at moderate sample sizes, e.g. $n^{(k)} = 20$. It is clear the outperformance of our proposal, relative to the benchmarks, with *Node-wise* performing equally well only in terms of FNR.

We then evaluate the performance of each methodology in recovering the DAG structure. We compare each DAG estimate $\widehat{\mathcal{D}}$ with the true DAG $\mathcal{D}$, by measuring the Structural Hamming Distance (SHD, Tsamardinos et al. 2006) between the two graphs; see also the main paper for details. Results are summarized in the box-plots of Figures 9 and 10, where each plot reports the distribution of SHD across the $N = 40$ simulated datasets for the various methods and increasing sample sizes $n^{(k)} \in \{10, \ldots, 500\}$ under *Sparse* and *Diffuse* Scenarios. It is clear the tendency of a better and better recovery of the true graphical structure as we increase the amount of available data, and an overall better performance relative to all the benchmarks. The only exception is GIES 0.5 which was however implemented with input the true targets and outperforms our Bayesian method in few settings characterized by small sample sizes, where indeed target identification was more difficult for our method. However, it performs worse than *Bayes* as $n^{(k)}$ increases, especially under Scenario *Sparse*

In the second case, we repeat all simulation settings above, but we impose $I_1 = \emptyset$, implying no interventions for the first data matrix $\boldsymbol{X}^{(1)}$, which therefore consists of *observational* data. Targets I_2, I_3, I_4 are again generated under two different scenarios, *Sparse* and *Diffuse*. We present results relative to both target estimation (Figures 7 and 8) and DAG structure learning (Figures 9 and 10). We note no substantial differences relative to the case with only interventional data.

6 Sensitivity analyses

For the two real data problems presented in the main text (Section 6), we perform sensitivity analyses with respect to the hyperparameter η, which represents the prior probability of edge inclusion in the DAG. Therefore, tuning of η can regulate the inclusion of specific links in the graph space. Since interventions modify the DAG structure through edge removals, choice of η may affect intervention targets' identification. Accordingly, we evaluate the impact of η

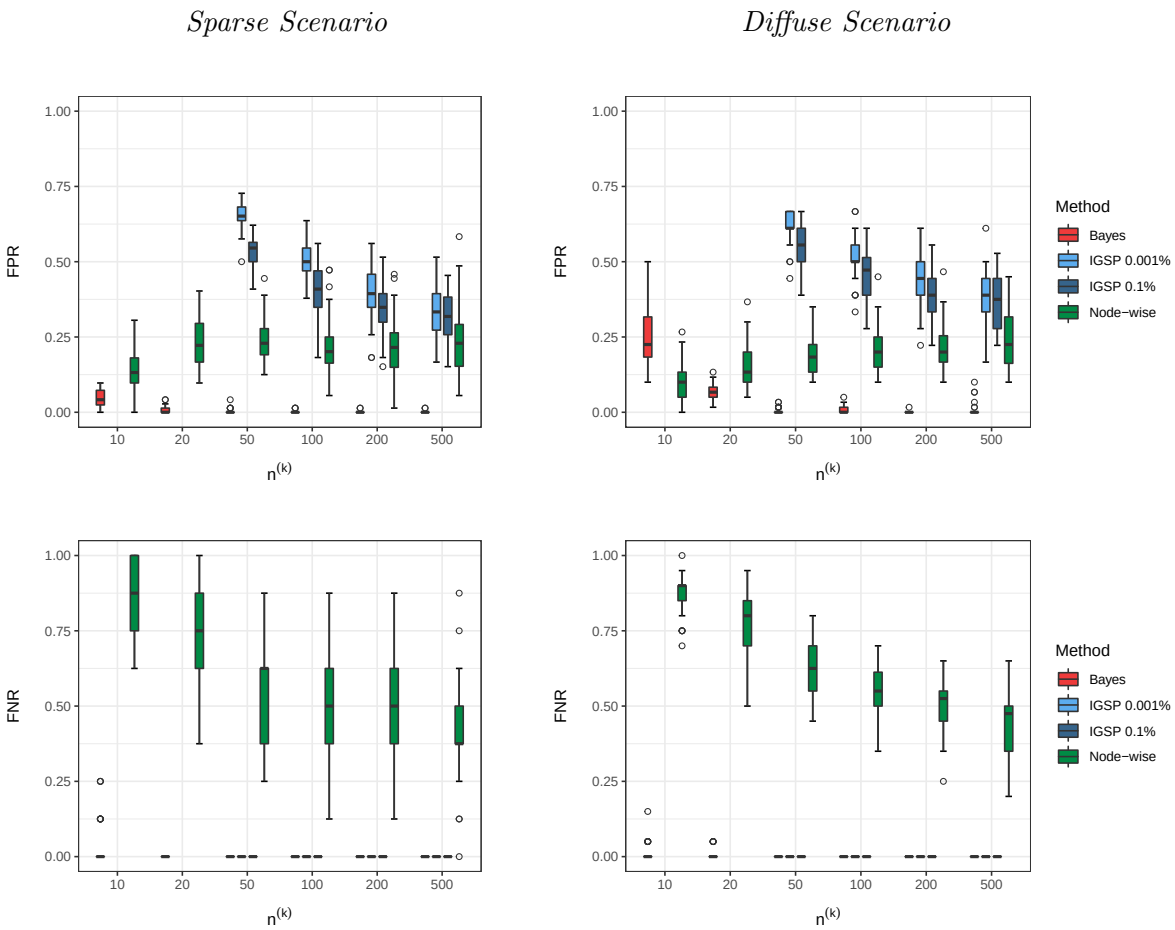

Figure 3: Simulation 1, interventional data. Distribution of the False Positive Rate (FPR, first row) and False Negative Rate (FNR, second row) across $N = 40$ simulated datasets under *Sparse* (first column) and *Diffuse* (second column) Scenarios, for number of nodes $q = 20$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1% and IGSP 0.001%) and node-wise regression (Node-wise).

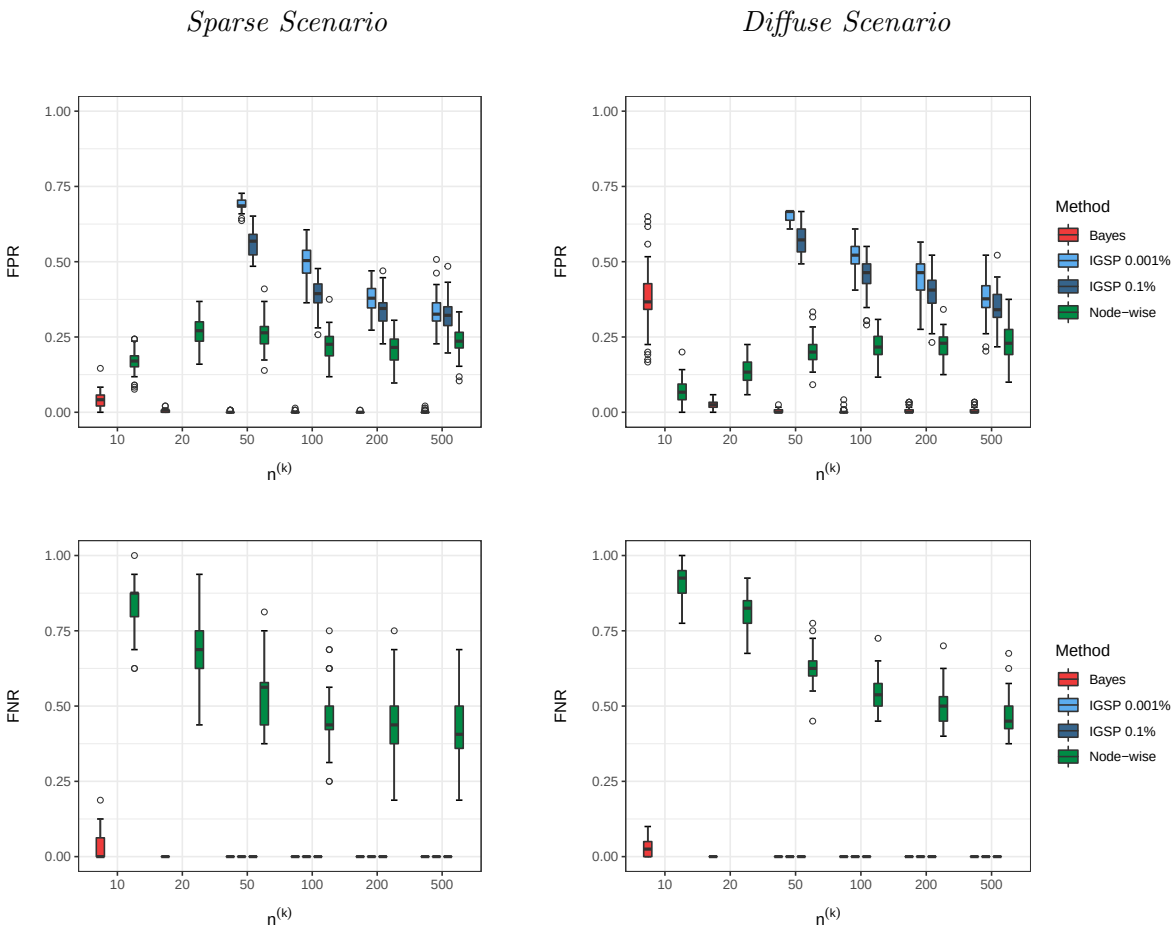

Figure 4: Simulation 1, interventional data. Distribution of the False Positive Rate (FPR, first row) and False Negative Rate (FNR, second row) across $N = 40$ simulated datasets under *Sparse* (first column) and *Diffuse* (second column) Scenarios, for number of nodes $q = 40$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1%, IGSP 0.001%) and node-wise regression (Node-wise).

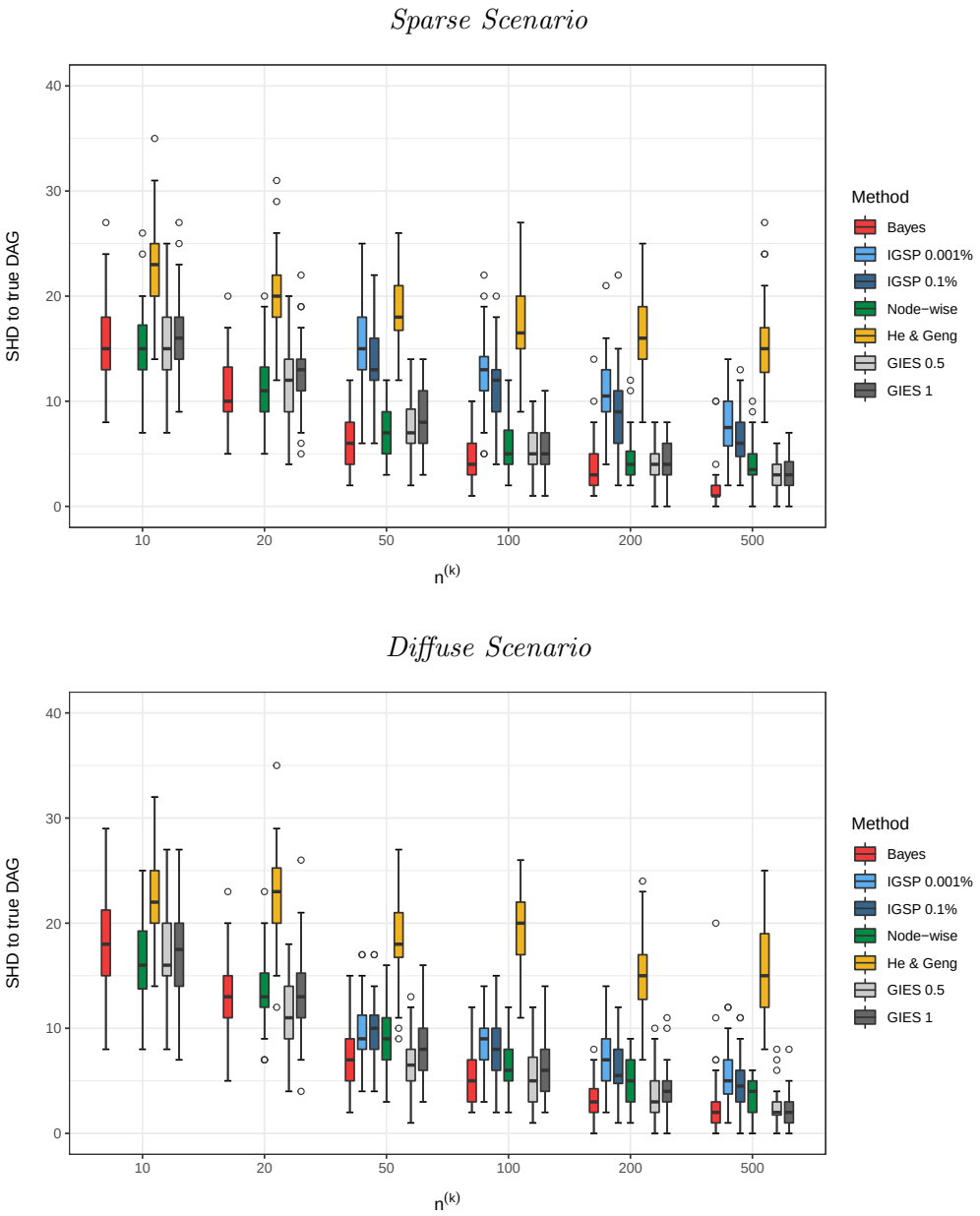

Figure 5: Simulation 1, interventional data. Distribution across $N = 40$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true DAG under *Sparse* (top) and *Diffuse* (bottom) Scenarios, for number of nodes $q = 20$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1%, IGSP 0.001%), node-wise regression (Node-wise), Algorithm 1 of He and Geng (2016) (He & Geng) and the *Greedy Interventional Equivalence Search* method with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5, GIES 1).

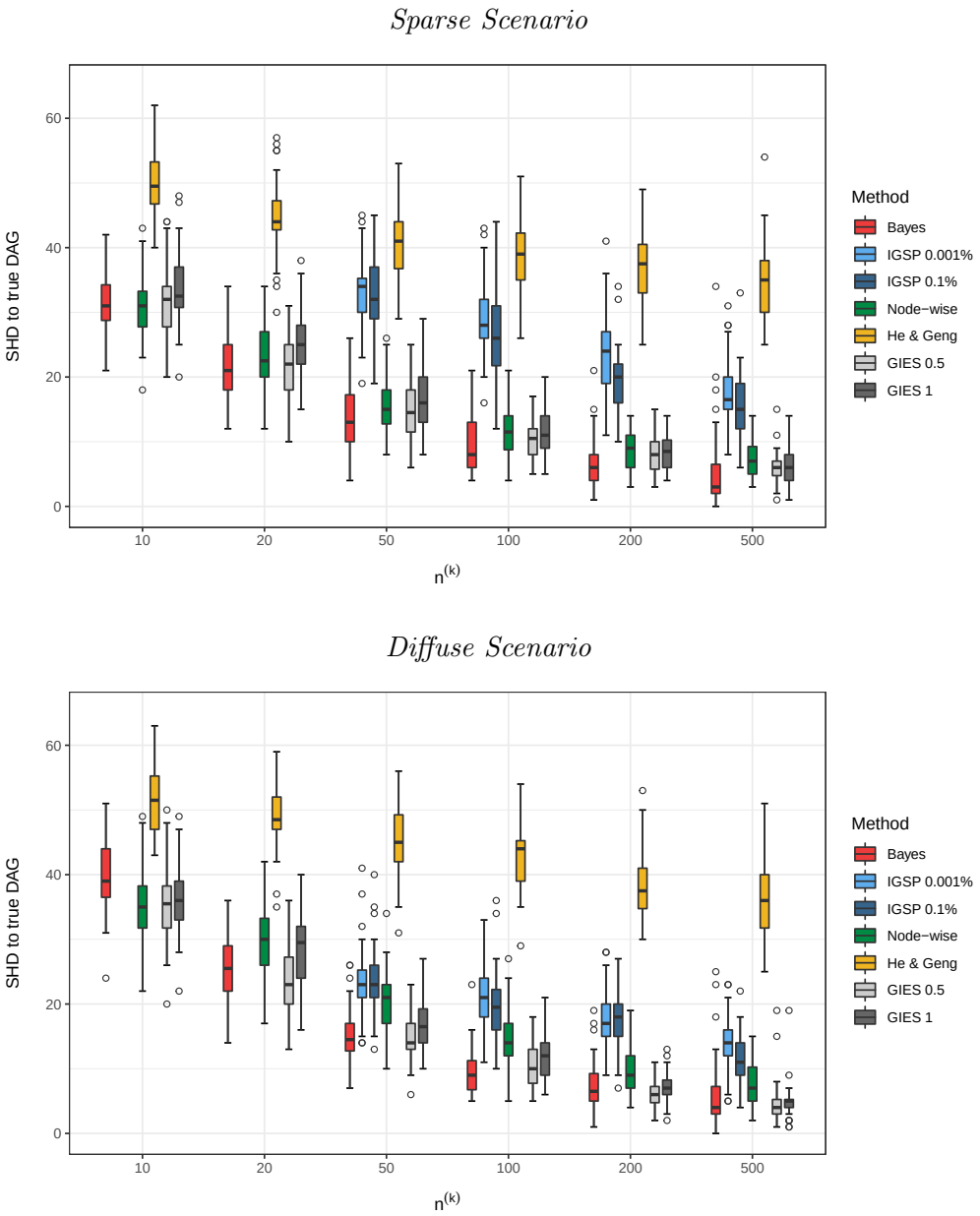

Figure 6: Simulation 1, interventional data. Distribution across $N = 40$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true DAG under *Sparse* (top) and *Diffuse* (bottom) Scenarios, for number of nodes $q = 40$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1% IGSP 0.001%), node-wise regression (Node-wise), Algorithm 1 of He and Geng (2016) (He & Geng) and the *Greedy Interventional Equivalence Search* method with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5, GIES 1).

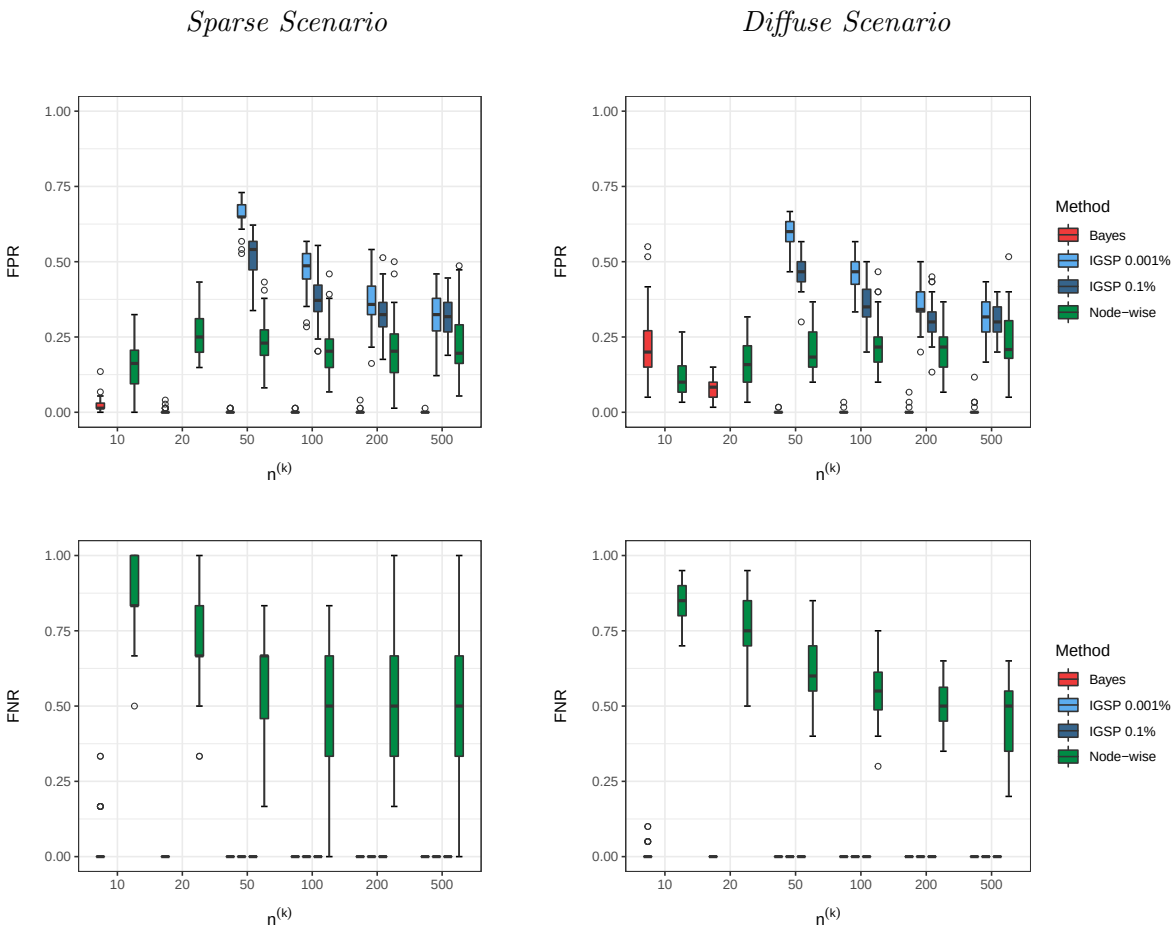

Figure 7: Simulation 2, observational and interventional data. Distribution of the False Positive Rate (FPR, first row) and False Negative Rate (FNR, second row) across $N = 40$ simulated datasets under *Sparse* (first column) and *Diffuse* (second column) Scenarios, for number of nodes $q = 20$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1%, IGSP 0.001%) and node-wise regression (Node-wise).

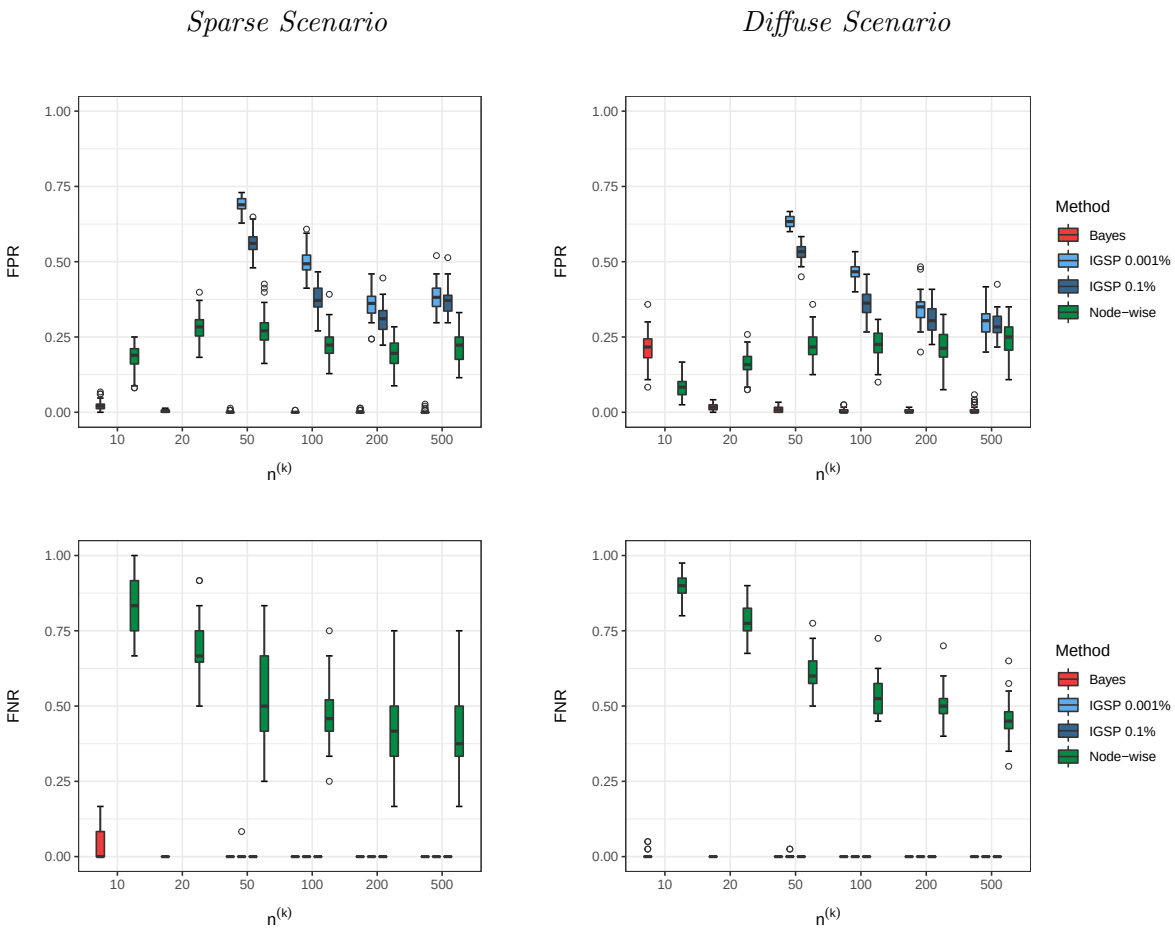

Figure 8: Simulation 2, observational and interventional data. Distribution of the False Positive Rate (FPR, first row) and False Negative Rate (FNR, second row) across $N = 40$ simulated datasets under *Sparse* (first column) and *Diffuse* (second column) Scenarios, for number of nodes $q = 40$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1%, IGSP 0.001%) and node-wise regression (Node-wise).

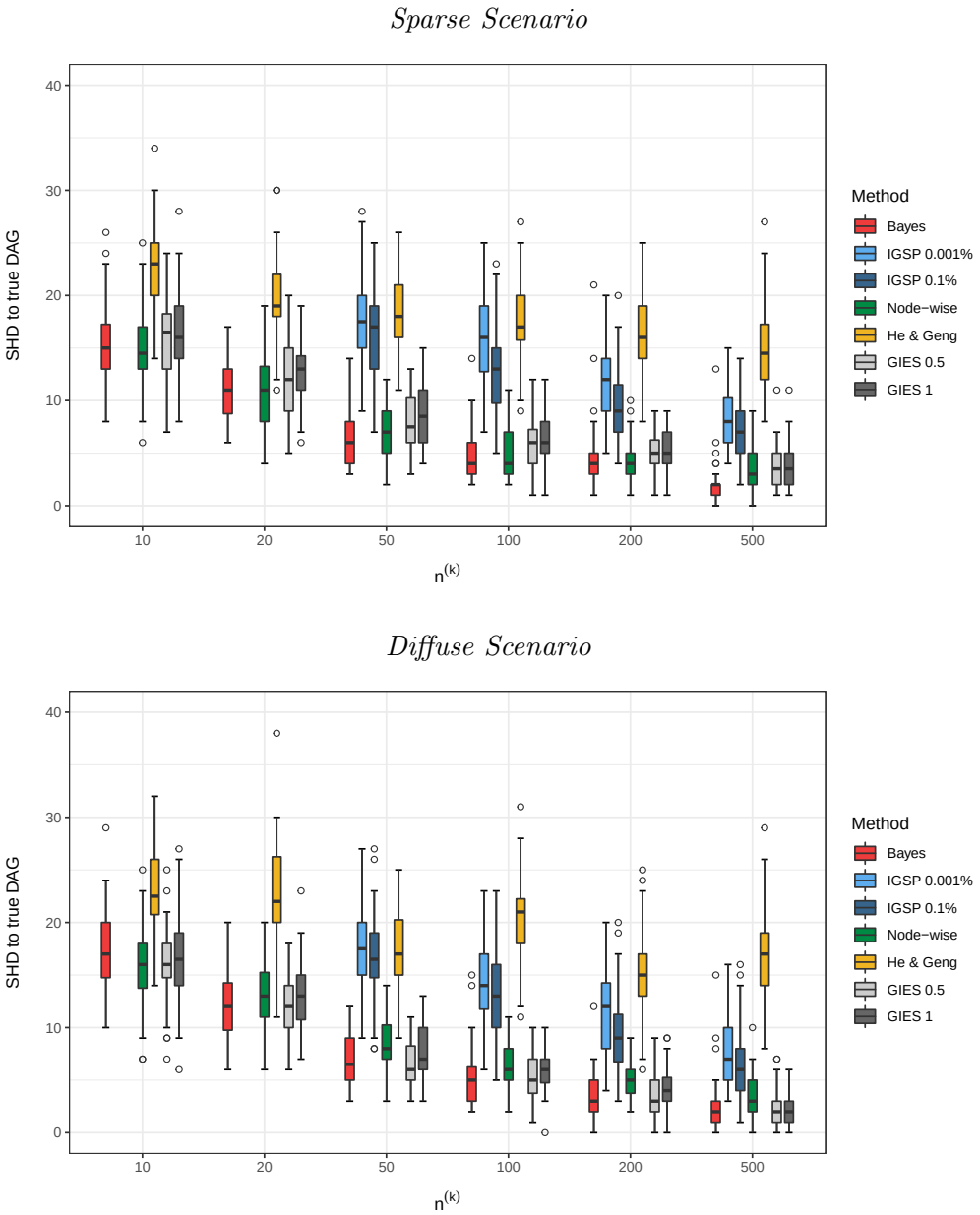

Figure 9: Simulation 2, observational and interventional data. Distribution across $N = 40$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true DAG under *Sparse* (top) and *Diffuse* (bottom) Scenarios, for number of nodes $q = 20$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1%, IGSP 0.001%), node-wise regression (Node-wise), Algorithm 1 of He and Geng (2016) (He & Geng) and the *Greedy Interventional Equivalence Search* method with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5, GIES 1).

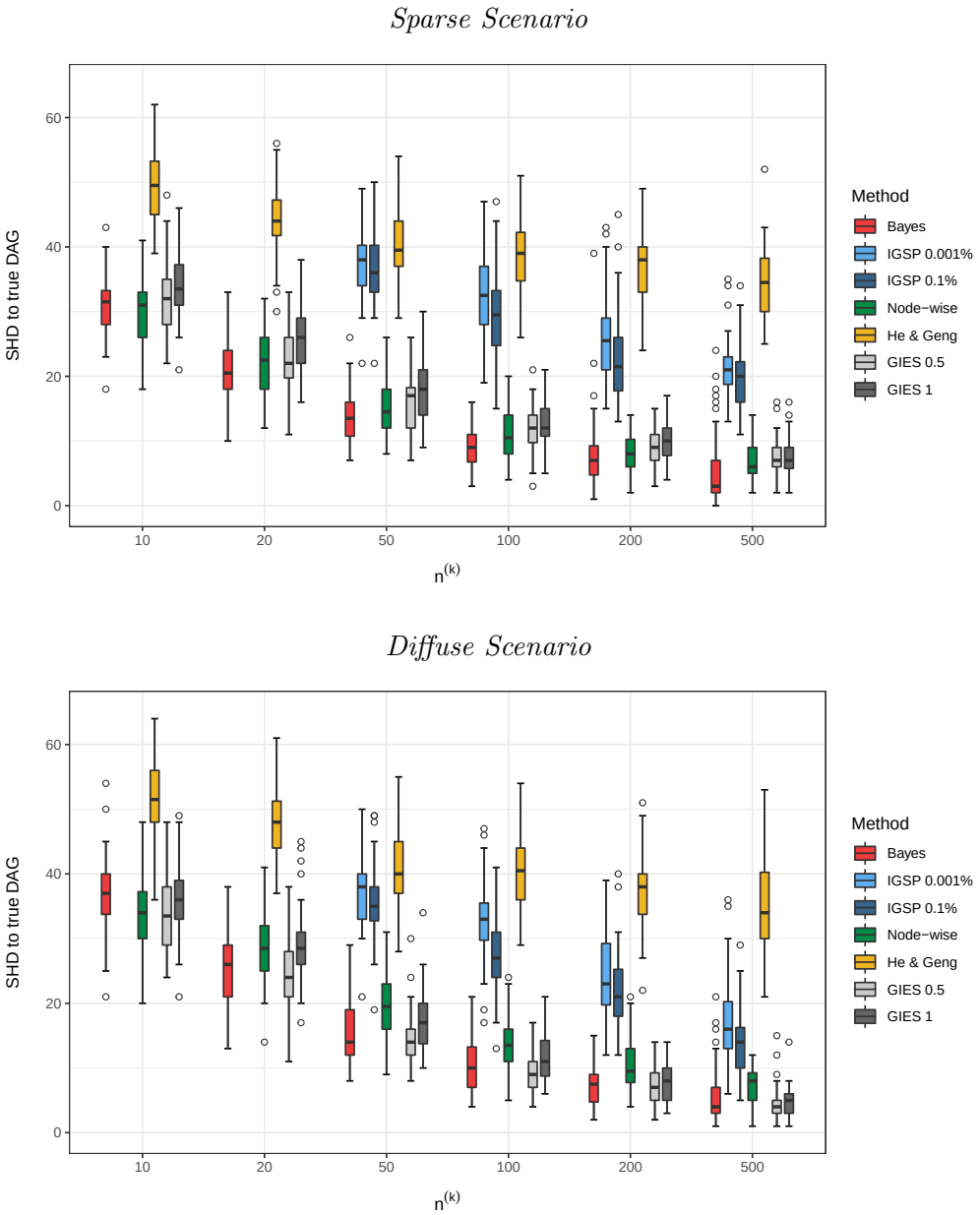

Figure 10: Simulation 2, observational and interventional data. Distribution across $N = 40$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true DAG under *Sparse* (top) and *Diffuse* (bottom) Scenarios, for number of nodes $q = 40$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1%, IGSP 0.001%) node-wise regression (Node-wise), Algorithm 1 of He and Geng (2016) (He & Geng) and the *Greedy Interventional Equivalence Search* method with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5, GIES 1).

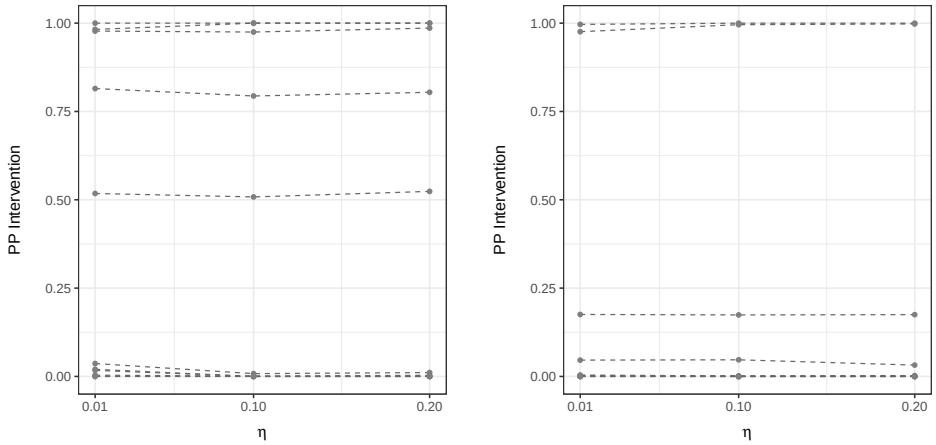

Figure 11: Sachs data. Posterior probabilities of intervention $\widehat{p}_{j\in I_k}$ as a function of $\eta \in \{0.01, 0.1, 0.2\}$ for each node $j = 1, \ldots, q$, for the two datasets with unknown targets, I_8 (left) and I_9 (right).

on the posterior probabilities of intervention $\widehat{p}_{j\in I_k}$ as defined in Equation (13) of our paper, for each node $j = 1, \ldots, q$ and target I_k, $k = 1, \ldots, K$. For both the applications, we vary $\eta \in \{0.01, 0.10, 0.20\}$.

Results for the Sachs data are reported in Figure 11 where each plot refers to one of the two datasets with unknown targets. It is clear that in this setting, characterized by large group sample sizes $n^{(k)}$, results are quite insensitive to the choice of η. Similar results were obtained for the epilepsy data: from Figure 12, which reports the behaviour of $\widehat{p}_{j\in I_k}$ as a function of η, for $j = 1, \ldots, q$ and each of the three unknown targets (drugs) I_k, $k = 2, 3, 4$, posterior probabilities of inclusion are stable around zero for most of the nodes and targets. Differently, results are more affected by the choice of η for nodes with non-zero probabilities of inclusion. However, largest variations do not exceed a range of 0.25; in addition, estimated intervention targets, obtained with a threshold for inclusion of 0.5, never change, with the only exception being one node in the last group/intervention, included in the estimated target only for $\eta = 0.10$, with a probability of inclusion just above 0.5.

7 Real data analyses: additional results and figures

In this section we include additional results and plots for the application to real data. Specifically, for the epilepsy dataset we report in Figure 13 the heat map collecting the posterior probabilities of intervention computed for each node $v \in \{1, \ldots, 100\}$ under each of the three drug therapies (interventions) as in Equation (15) of our paper. Sparsity in the map reveals that there are few genes exhibiting a high posterior probability of intervention under some of the treatments. In particular, only six genes are associated with probabilities of intervention exceeding 0.5. The same result is also clear from the posterior distribution of the number of intervened nodes (size

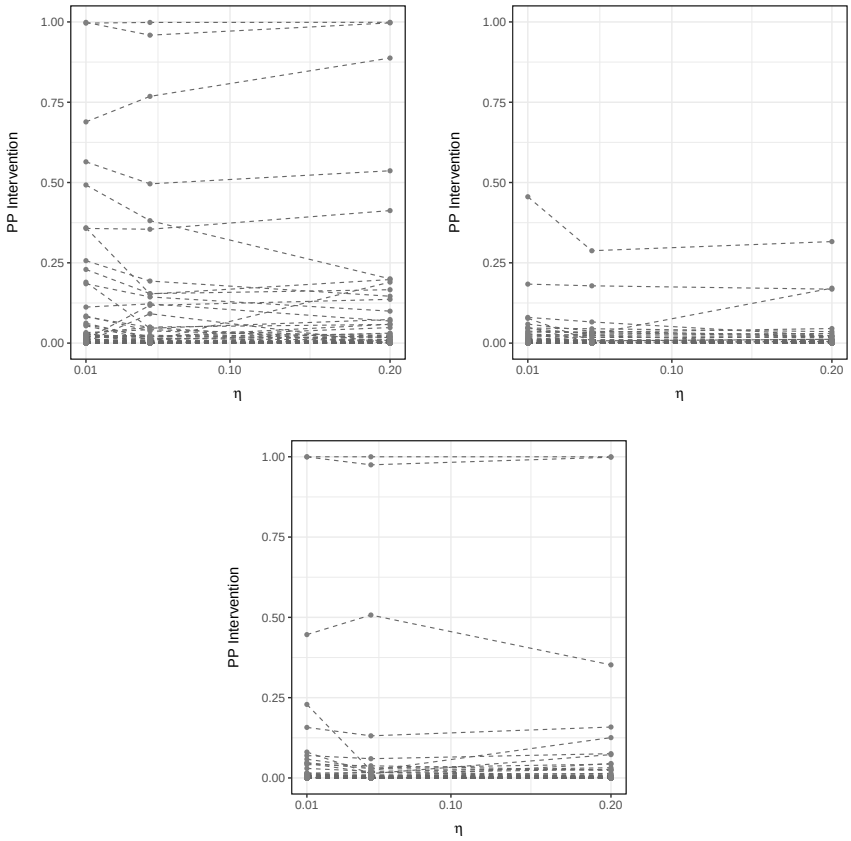

Figure 12: Epilepsy data. Posterior probabilities of intervention $\widehat{p}_{j \in I_k}$ as a function of $\eta \in \{0.01, 0.1, 0.2\}$ for each node $j = 1, \ldots, q$ and each dataset (targets I_2, I_3, I_4) corresponding to one of the three administrated drugs.

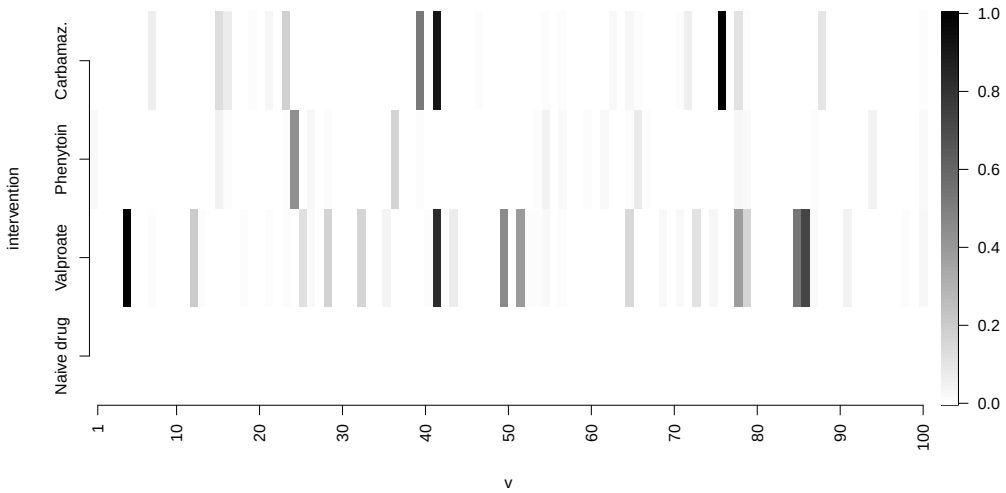

Figure 13: Epilepsy data. Heat map with estimated posterior probabilities of intervention computed for each drug-therapy (intervention) and each node/gene v ($v = 1, \ldots, 100$).

of the intervention target) obtained for each of the three drug therapies. The approximated posterior distribution resulting from our MCMC output is reported in Figure 14.

For the Sachs data, Figure 15 reports our estimated (median probability) DAG $\widehat{\mathcal{D}}$, with blue and pink circles representing nodes which are included in the estimated intervention targets $\hat{I}_8$ and $\hat{I}_9$ respectively. The same figure contains the intervention DAGs (Definition 2.1 in the main text) of $\widehat{\mathcal{D}}$ given the intervention targets $\hat{I}_8$ and $\hat{I}_9$, $\widehat{\mathcal{D}}^{\hat{I}_8}$ and $\widehat{\mathcal{D}}^{\hat{I}_9}$ respectively. The two graphs help understanding how the original DAG structure modifies after one of the reagents is applied.

We also report in Figure 16 three DAG estimates obtained from benchmark methods *Node-wise*, *He & Geng* and *IGSP 0.1%* applied to Sachs data; see also Section 5 within our paper for a brief presentation of these methods. Again, blue (pink) circles represent nodes included in the (estimated) intervention target $\hat{I}_8$ ($\hat{I}_9$); grey circles represent instead nodes which are included in both the two targets. Notice that, because *He & Geng* does not provide target estimates, this information is missing from the corresponding estimated graph. Also notice that its output is not in general a DAG but a *partially* directed graph; accordingly we name it $\widehat{\mathcal{G}}_{H\&G}$. Two of these graphs, $\widehat{\mathcal{D}}_{NW}$ and $\widehat{\mathcal{G}}_{H\&G}$, are equal to our estimate $\widehat{\mathcal{D}}$ (Figure 15) in terms of skeleton. In addition, $\widehat{\mathcal{D}}$ and $\widehat{\mathcal{D}}_{NW}$ also share the same edge orientations, with the only exception of link $p38 \leftarrow JNK$ in $\widehat{\mathcal{D}}$ which is reversed in $\widehat{\mathcal{D}}_{NW}$. By converse, orientation of most edges is not recovered in $\widehat{\mathcal{G}}_{H\&G}$ and only two oriented edges, $PLC \rightarrow PIP2$ and $PIP3 \rightarrow PIP2$, are also present in $\widehat{\mathcal{D}}$ and $\widehat{\mathcal{D}}_{NW}$. Differently, $\widehat{\mathcal{D}}_{IGSP}$ presents a different skeleton, because of the absence of two links, and also differs from $\widehat{\mathcal{D}}$ by the orientation of edges $PKA \rightarrow Erk$ and $p38 \rightarrow JNK$.

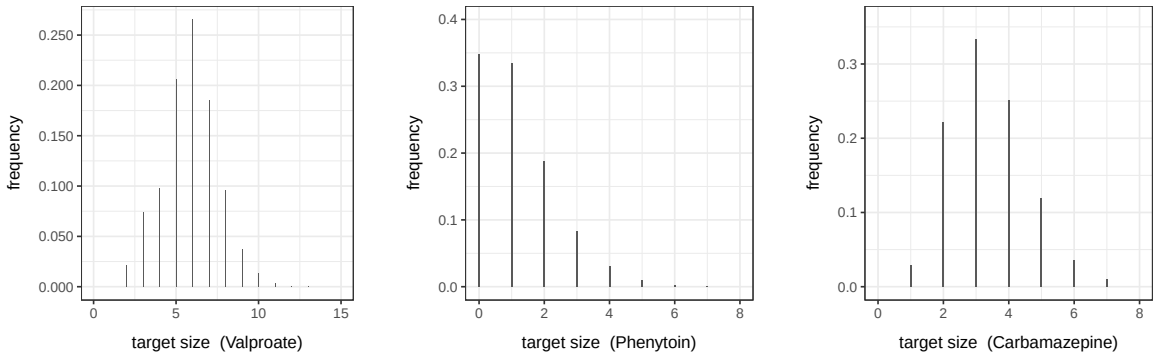

Figure 14: Epilepsy data. Posterior distribution of the number of intervened nodes (size of the intervention target) for each drug-treatment included in the study.

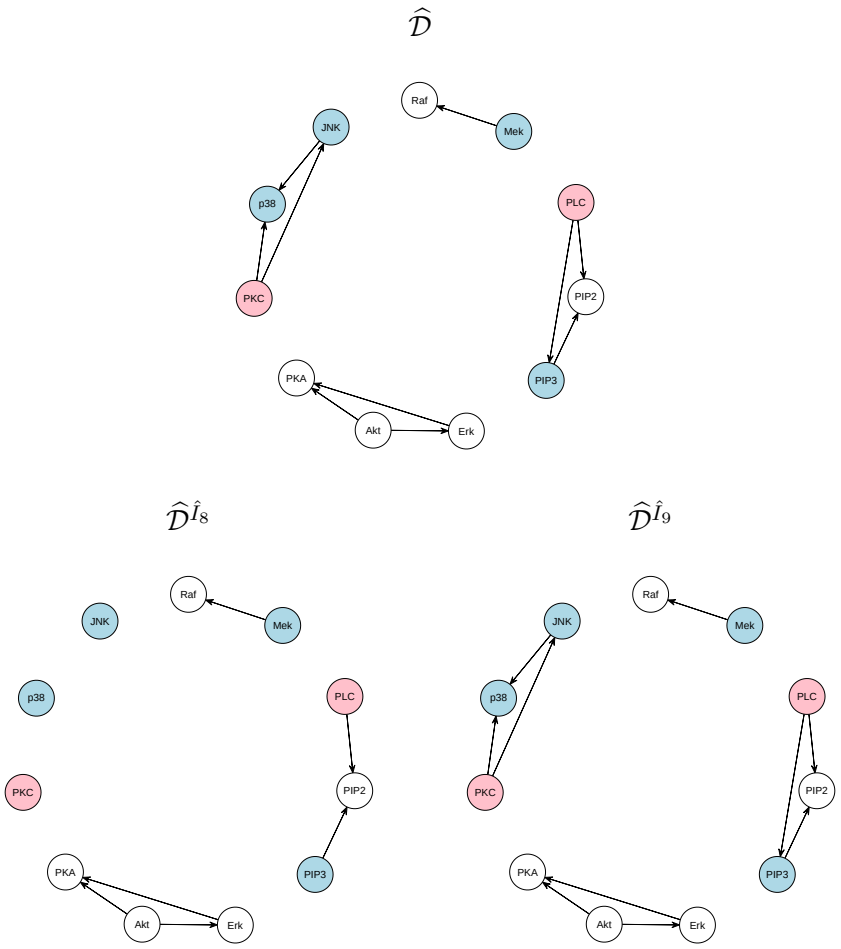

Figure 15: Sachs data. Estimated (median probability) DAG $\widehat{\mathcal{D}}$ obtained from our method with blue (pink) circles representing nodes included in targets $\hat{I}_8$ and $\hat{I}_9$; two corresponding intervention DAGs, $\widehat{\mathcal{D}}^{I_8}$ and $\widehat{\mathcal{D}}^{I_9}$.

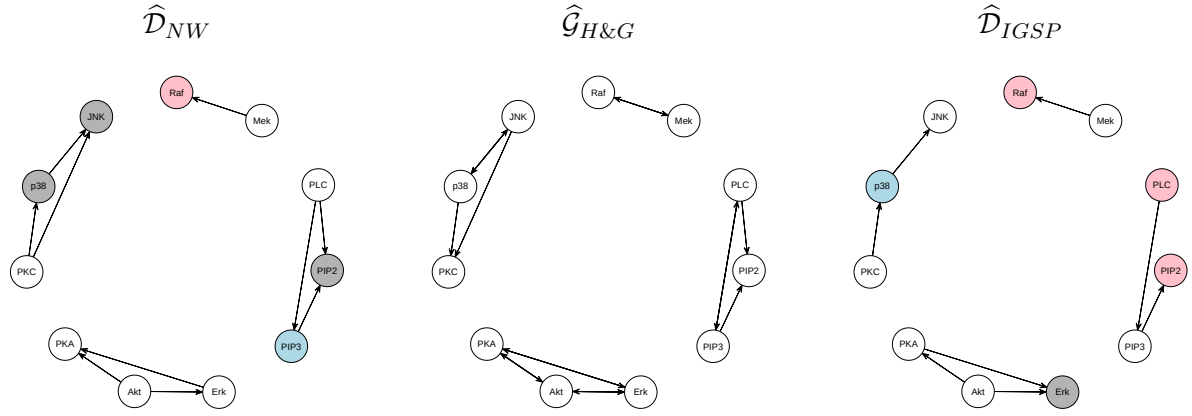

Figure 16: Sachs data. Estimated graphs obtained from *Node-wise* regression, Algorithm 1 of He and Geng (2016) and IGSP method, with blue (pink) circles representing nodes included in targets $\hat{I}_8$ and $\hat{I}_9$ (grey nodes included in both).

References

Altomare, D., G. Consonni, and L. La Rocca (2013). Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics 69*(2), 478–487.

Ben-David, E., T. Li, H. Massam, and B. Rajaratnam (2015). High dimensional Bayesian inference for Gaussian directed acyclic graph models. *arXiv pre-print*.

Cao, X., K. Khare, and M. Ghosh (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics 47*(1), 319–348.

Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 87–98. Morgan Kaufmann Publishers Inc.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research 2*(3), 445–498.

Foygel, R. and M. Drton (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems 23*, pp. 2020–2028.

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

Ghazal, G. A. and H. Neudecker (2000). On second-order and fourth-order moments of jointly distributed random matrices: a survey. *Linear Algebra and its Applications 321*(1-3), 61–93.

Han, S. W., G. Chen, M.-S. Cheon, and H. Zhong (2016). Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *Journal of the American Statistical Association 111*(515), 1004–1019.

Hauser, A. and P. Bühlmann (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research 13*(79), 2409–2464.

He, Y. and Z. Geng (2016). Causal network learning from multiple interventions of unknown manipulated targets. *arXiv preprint arXiv:1610.08611*.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of Chemical Physics 21*(6), 1087–1092.

Peluso, S. and G. Consonni (2020). Compatible priors for model selection of high-dimensional Gaussian DAGs. *Electronic Journal of Statistics 14*(2), 4110–4132.

Scott, J. G. and C. M. Carvalho (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics 17*(4), 790–808.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman (2000). *Causation, prediction, and search*. MIT press.

Squires, C., Y. Wang, and C. Uhler (2020). Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR.

Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning 65*(1), 31–78.