Identification of causal directionality using conditional DAG models: supplementary information

CHRIS. J. OATES

School of Mathematical and Physical Sciences,
University of Technology Sydney,
NSW 2007 Australia
christopher.oates@uts.edu.au

JIM Q. SMITH

Department of Statistics,
University of Warwick,
Coventry, CV4 7AL UK
j.q.smith@warwick.ac.uk

SACH MUKHERJEE

German Center for Neurodegenerative Diseases (DZNE),
53175 Bonn, Germany
sach.mukherjee@dzne.de

Simulated Data

Data $(X_i^l, Y_i^l)_{i=1,\dots,p}^{l=1,\dots,n}$ were generated as described below. In summary

1. Generate a DAG $G \in \mathcal{G}$
2. Simulate secondary variables $(X_i)_{i=1,\dots,p}^{l=1,\dots,n}$
3. Simulate primary variables $(Y_i)_{i=1,\dots,p}^{l=1,\dots,n}$ conditional upon both G and $(X_i)_{i=1,\dots,p}^{l=1,\dots,n}$

Below we describe each of these three steps in more detail:

Step 1: Without loss of generality we assume an ordering of the variables $Y_1, \dots, Y_p$. Then, for each node i in turn, the parents of node i in G are selected from $\{\pi \subseteq \{1, \dots, i-1\} : |\pi| \leq 2\}$ with probability $p(\pi) \propto \binom{p}{|\pi|}^{-1}$. This defines a DAG $G \in \mathcal{G}$.

Step 2: Pick a permutation τ uniformly at random from the set $\{1, \dots, p\}$ and generate a second DAG H on the variables $X_{\tau(1)}, \dots, X_{\tau(p)}$, similarly to Step 1, using the ordering τ. Then generate $(X_i)_{i=1,\dots,p}^{l=1,\dots,n}$ as follows:

$$X_i^l = \theta \frac{1}{|\mathrm{pa}_H(i)|} \sum_{j \in \mathrm{pa}_H(i)} \gamma_{i,j} \frac{X_j^l}{\overline{X_j}} + (1-\theta)\epsilon_i^l \quad (1)$$

Here $\overline{X}_j = \left(\frac{1}{n}\sum_{l=1}^{n}(X_j^l)^2\right)^{1/2}$ is the standard deviation of X_j over all l samples, $\gamma_{i,j}$ are regression coefficients that are particular to parent j and child i, θ is a tuning parameter that controls the amount of correlation between the secondary variables, and $\epsilon_i^l \sim N(0,1)$ are independent noise terms. We generated the regression coefficients independently as $\gamma_{i,j} \sim \frac{1}{2}N(-1,\frac{1}{4}) + \frac{1}{2}N(1,\frac{1}{4})$.

Step 3: Generate $(Y_i)_{i=1,\ldots,p}^{l=1,\ldots,n}$ conditional upon both G and $(X_i)_{i=1,\ldots,p}^{l=1,\ldots,n}$ as follows:

$$Y_i^l \quad = \quad \beta_{i,0}X_i + \frac{1}{|\mathrm{pa}_G(i)|}\sum_{j\in \mathrm{pa}_G(i)} \beta_{i,j}\frac{Y_j^l}{\overline{Y}_j} + \epsilon_i^l \tag{2}$$

Here $\overline{Y}_j = \left(\frac{1}{n}\sum_{l=1}^{n}(Y_j^l)^2\right)^{1/2}$ is the standard deviation of Y_j over all l samples, $\beta_{i,j}$ are regression coefficients that are particular to parent j and child i, and $\epsilon_i^l \sim N(0,1)$ are independent noise terms (not related to the ϵ_i^l in Step 2). We generated the regression coefficients independently as $\beta_{i,0}$, $\beta_{i,j} \sim \frac{1}{2}N(-1,\frac{1}{4}) + \frac{1}{2}N(1,\frac{1}{4})$.

Additional Results

Here, in SFigs. 1 and 2, we report two additional simulation studies (described in the main text) that consider violations of the partial faithfulness assumption. In SFigs. 3 and 4 we report additional results from the biological application.

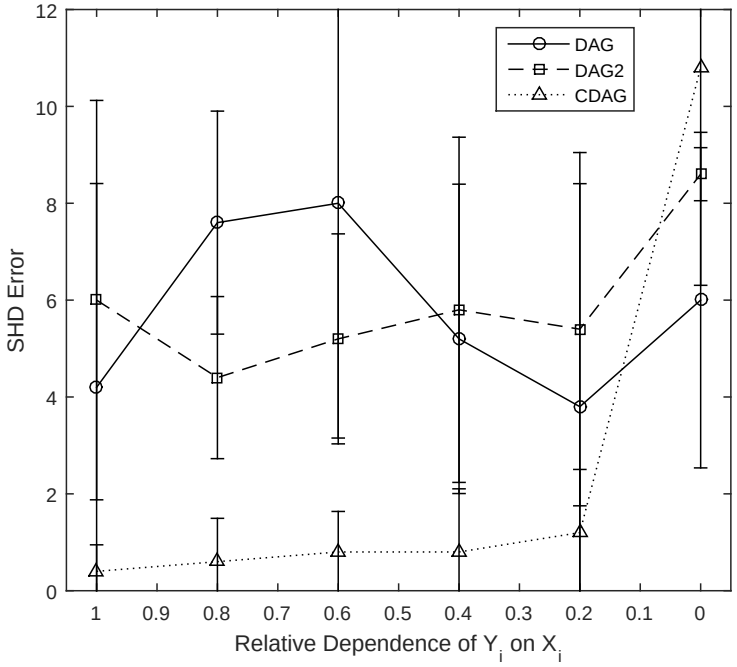

Figure 1: Simulated data results; violating partial faithfulness by weakening the instrumental effect. [Data were generated using linear-Gaussian structural equations. Here we fixed $\theta = 0$, $p = 15$, $n = 1000$ and considered decreasing the influence of the X_i on the Y_i, as described in the main text. On the x-axis we display the strength of the association $X_i \to Y_i$ as a proportion of its value in previous simulations, so that when $x = 1$ the results correspond to those in Table 1 in the main text. "DAG" = estimation based only on primary variables $(Y_i)_{i\in V}$, "DAG2" = estimation based on the full data $(X_i)_{i\in W} \cup (Y_i)_{i\in V}$, "CDAG" = CDAG estimation based on the full data $(X_i)_{i\in W} \cup (Y_i)_{i\in V}$.]

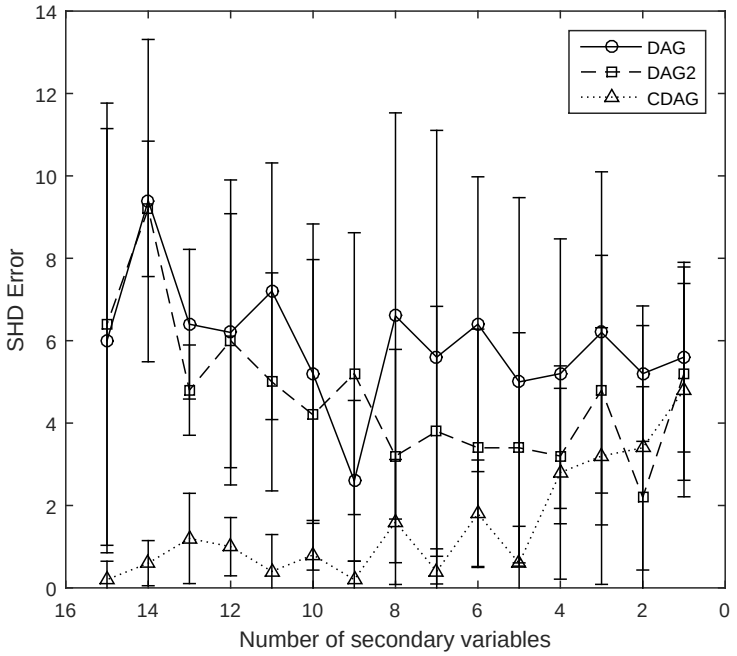

Figure 2: Simulated data results; violating partial faithfulness by removing secondary variables. [Data were generated using linear-Gaussian structural equations. Here we fixed $\theta = 0$, $p = 15$, $n = 1000$ and considered decreasing the influence of the X_i on the Y_i, as described in the main text. On the x-axis we display the number of secondary variables that were *not* removed, so that when $x = 15$ the results correspond to those in Table 1 in the main text. "DAG" = estimation based only on primary variables $(Y_i)_{i \in V}$, "DAG2" = estimation based on the full data $(X_i)_{i \in W} \cup (Y_i)_{i \in V}$, "CDAG" = CDAG estimation based on the full data $(X_i)_{i \in W} \cup (Y_i)_{i \in V}$.]

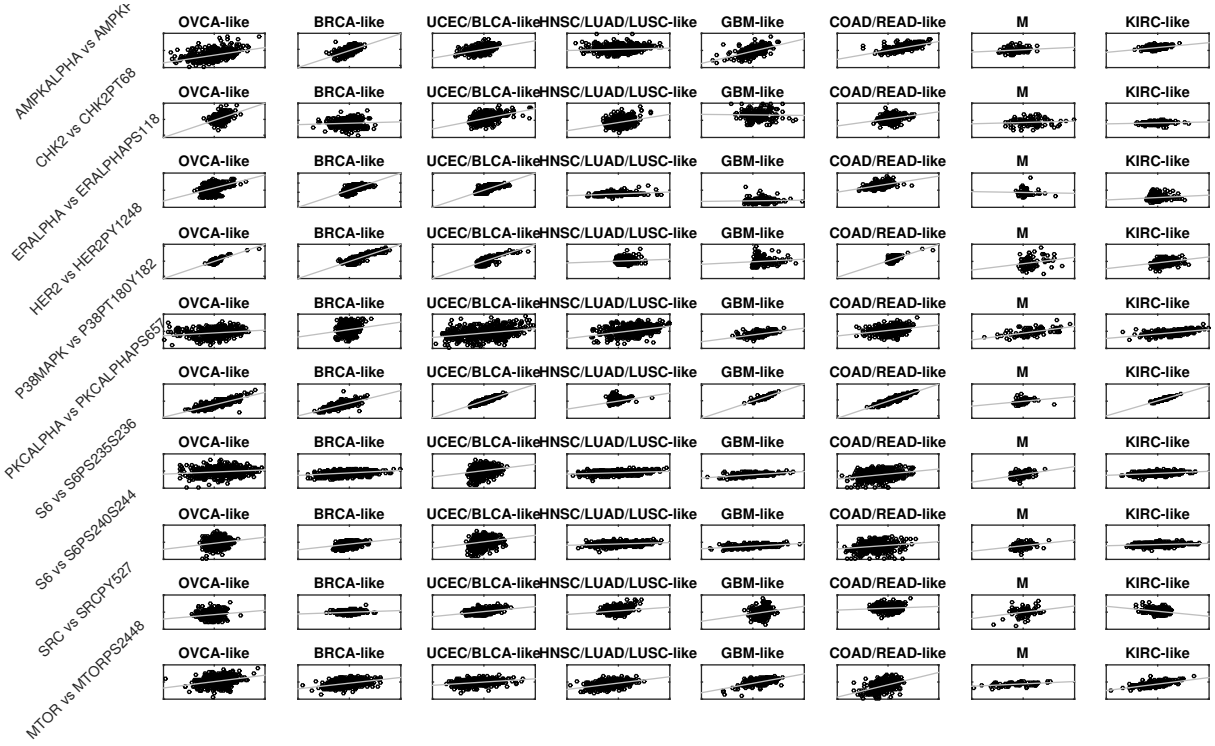

Figure 3: Correlation between total protein expression levels and phosphorylated protein expression levels, for 10 of the $p = 24$ species involved in the biological application. [The grey line in each panel is a least-squares linear regression.]

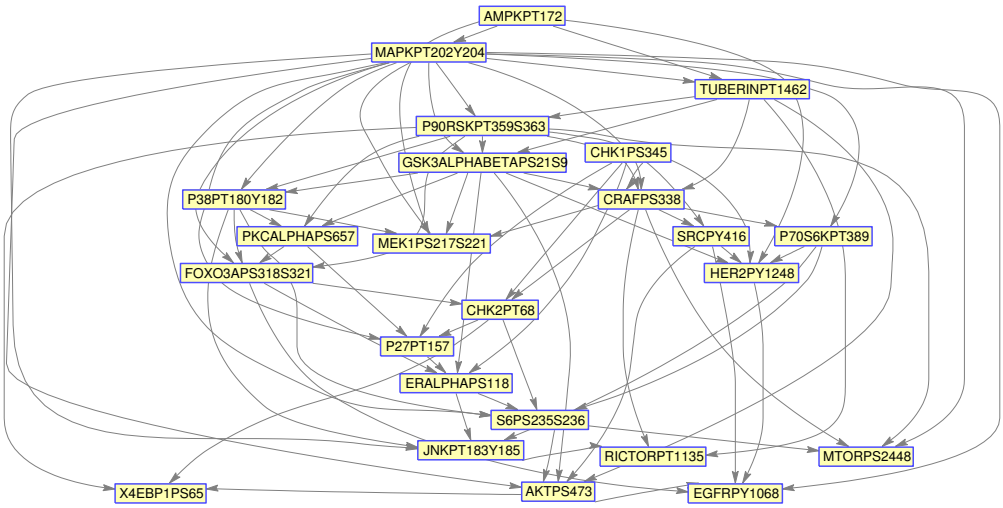

Figure 4: *Maximum a posteriori* DAG, estimated from data derived from BRCA-like cancer samples. Here vertices represent phosphorylated proteins (primary variables) and edges are reified with the interpretation that the parent protein plays a causal role in phosphorylation of the child protein. Note that this DAG estimator is based only on the primary variable data.