Supplementary Material for Directed Graphical Models and Causal Discovery for Zero-Inflated Data

Shiqing Yu SYU.PHD@GMAIL.COM
Department of Statistics, University of Washington, Seattle, Washington, 98195, U.S.A.

Mathias Drton MATHIAS.DRTON@TUM.DE
Department of Mathematics, Technical University of Munich, 85748 Garching bei München, Germany

Ali Shojaie ASHOJAIE@UW.EDU
Department of Biostatistics, University of Washington, Seattle, Washington, 98195, U.S.A.

Editors:
Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

A. Proofs

In this appendix we present proofs for the theorems and corollaries in the paper.

We first prove the following lemma that states that if two sums of distinct (ignoring the multiplicative constant) exponentials of polynomials in $\boldsymbol{y} \in \mathbb{R}^m$ agree almost everywhere in $\mathbb{R}^m$, then they must have the same number of terms and there must be a 1-1 correspondence between the terms.

Lemma 1 *Let the number of variable be $m \geq 1$ and the degree be $p \geq 1$. Let $\mathcal{D} \equiv \{\boldsymbol{d} \in \mathbb{Z}_{\geq 0}^m : 1 \leq \sum_{j=1}^m d_j \leq p\}$ be the set of nonnegative integer-valued m-vectors with ℓ_1 norm $\in [1, p]$. Given a vector $\boldsymbol{a} \in \mathbb{R}^{|\mathcal{D}|}$ indexed by $\boldsymbol{d} \in \mathcal{D}$ (i.e. $a_{\boldsymbol{d}} \in \mathbb{R}$ for all $\boldsymbol{d} \in \mathcal{D}$), define*

$$f^{(m)}(\boldsymbol{y}; \boldsymbol{a}) \equiv \exp\left(\sum_{\boldsymbol{d}\in\mathcal{D}} a_{\boldsymbol{d}} \prod_{j=1} y_i^{d_j}\right),$$

the exponential of the corresponding polynomial of degree $\leq p$ in $\boldsymbol{y} \in \mathbb{R}^m$. Note that $f^{(m)}$ does not have a constant term, and has degrees $\boldsymbol{d} \in \mathcal{D}$ and coefficients $\boldsymbol{a}$.

Suppose we have

$$\sum_{i=1}^{N_a} a_0^i f^{(m)}(\boldsymbol{y}; \boldsymbol{a}^i) = \sum_{i=1}^{N_b} b_0^i f^{(m)}(\boldsymbol{y}; \boldsymbol{b}^i) \tag{1}$$

for almost every $\boldsymbol{y} \equiv (y_1, \ldots, y_m) \in \mathbb{R}^m$ with respect to the Lebesgue measure, where $N_a \geq 0$, $N_b \geq 0$, $\{\boldsymbol{a}^i\}_{i=1}^{N_a}$ are N_a distinct vectors in $\mathbb{R}^{|\mathcal{D}|}$, $\{\boldsymbol{b}^i\}_{i=1}^{N_b}$ are N_b distinct vectors in $\mathbb{R}^{|\mathcal{D}|}$ (otherwise just combine the coefficients), and $a_0^i, b_0^i \in \mathbb{R}\backslash\{0\}$ for all i. In other words, both sides of (1) are a sum of distinct exponentials of polynomials.

Then we must have $N_a = N_b$ and there is a permutation π of $\{1, \ldots, N_a\}$ such that $\boldsymbol{a}^i = \boldsymbol{b}^{\pi(i)}$ and $a_0^i = b_0^{\pi(i)}$, i.e. there is a 1-1 correspondence between the summands on both sides of (1).

Proof [Proof of Lemma 1] First note that both sides of (1) are continuous functions, and so is their difference, which is 0 almost everywhere by assumption. Thus, the inverse image of the open set $\mathbb{R}\backslash\{0\}$ under the difference is also open, and must be the empty set since it has measure 0. (1) thus holds for all $\boldsymbol{y} \in \mathbb{R}^m$.

We prove by induction on m, and first show the result for $m = 1$. In this case, $f^{(1)}(y_1; \boldsymbol{a}) \equiv \exp(a_1 y_1 + \cdots + a_p y_1^p)$, and $\boldsymbol{a}$ is just a p-vector.

First suppose $N_a \neq 0$ and $N_b \neq 0$. Observe that as $x \nearrow +\infty$, if $a_0 \neq 0$, the function $a_0 \exp(a_1 x + \cdots + a_p x^p)$ goes to

(i) $a_0 \neq 0$ if $a_1 = \cdots = a_p = 0$, or

(ii) 0 if $a_{d_{\max\neq 0}(\boldsymbol{a})} < 0$ where $d_{\max\neq 0}(\boldsymbol{a})$ is the largest $d \in \{1, \ldots, p\}$ such that $a_d \neq 0$, or

(iii) $+\infty$ if $a_{d_{\max\neq 0}(\boldsymbol{a})} > 0$.

Rearrange the terms on the left of (1) so that for each $1 \leq i < j \leq N_a$ we have $(\boldsymbol{a}^i - \boldsymbol{a}^j)_{d_{\max\neq 0}(\boldsymbol{a}^i - \boldsymbol{a}^j)} > 0$, and denote this total order as $\boldsymbol{a}^i > \boldsymbol{a}^j$. Rearrange the right-hand side similarly. By the assumption that $\{\boldsymbol{a}^i\}_{i=1}^{N_a}$ are distinct, $\boldsymbol{a}^i - \boldsymbol{a}^j \neq 0$, so $d_{\max\neq 0}(\boldsymbol{a}^i - \boldsymbol{a}^j)$ exists and this rearrangement is possible. Now dividing both sides of (1) by $f^{(1)}(y_1; \boldsymbol{a}^1) = \exp(a_1^1 y_1 + \cdots + a_p^1 y_1^p)$ we have

$$a_0^1 + \sum_{i=2}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^1) = \sum_{i=1}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}^i - \boldsymbol{a}^1). \tag{2}$$

Since $a_0^1 \neq 0$, and by the unique maximality of $\boldsymbol{a}^1$, as $y_1 \nearrow +\infty$, all terms in the summation on the left go to 0 (case (ii)). Thus, the right-hand side necessarily also goes to $a_0^1 \neq 0$, landing us in case (i) for at least one (and only one because $\boldsymbol{b}^i$ are unique) term on the right, i.e. $\boldsymbol{b}^i - \boldsymbol{a}^1 = \mathbf{0}$. (A nonzero finite limit cannot come from a sum of terms that go to $+\infty$ with positive and negative weights, since they must grow at different rates by uniqueness of $\boldsymbol{b}^i - \boldsymbol{a}^1$.) Since summands on both sides are sorted, we must have $\boldsymbol{b}^1 = \boldsymbol{a}^1$.

Then (2) becomes $a_0^1 - b_0^1 + \sum_{i=2}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^1) = \sum_{i=2}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}^i - \boldsymbol{a}^1)$. If $a_0^1 \neq b_0^1$, by the same reasoning there exists another $i \in \{2, \ldots, N_b\}$ such that $\boldsymbol{b}^i - \boldsymbol{a}^1 = \mathbf{0}$, violating uniqueness of $\{\boldsymbol{b}^i\}_{i=1}^{N_b}$. Thus, $a_0^1 = b_0^1$ and $\boldsymbol{a}^1 = \boldsymbol{b}^1$, and we have reduced the number of summands on both sides of (2) by 1 to

$$\sum_{i=2}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^1) = \sum_{i=2}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}^i - \boldsymbol{a}^1).$$

Continuing this process by each time dividing both sides by $f^{(1)}(y_1; \boldsymbol{a}^j - \boldsymbol{a}^{j-1})$, we would have matched $\min\{N_a, N_b\}$ pairs of coefficients between the a and the b groups. If $N_a \neq N_b$, assume $N_a > N_b$ without loss of generality, then

$$\sum_{i=N_b+1}^{N_1} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^{N_b}) = \text{const}.$$

Here the right-hand side is a constant that could be nonzero, because the argument for $a_0^1 = b_0^1$ in our first elimination step does not apply here. Dividing both sides by $f^{(1)}(y_1; \boldsymbol{a}^{N_b+1} - \boldsymbol{a}^{N_b})$, we have $a_0^{N_b+1} + \sum_{i=N_b+2}^{N_1} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^{N_b+1}) = f^{(1)}(y_1; \boldsymbol{a}^{N_b} - \boldsymbol{a}^{N_b+1})$. By maximality of $\boldsymbol{a}^{N_b+1}$ among $\boldsymbol{a}^{N_b+1}, \ldots, \boldsymbol{a}^{N_a}$, the left-hand side goes to $a_0^{N_b+1} \neq 0$ as $y_1 \nearrow +\infty$, while since $\boldsymbol{a}^{N_b} > \boldsymbol{a}^{N_b+1}$, the right-hand side goes to $+\infty$, a contradiction. Thus, $N_a = N_b$, $a_0^i = b_0^i$ and $\boldsymbol{a}^i = \boldsymbol{b}^i$ for $i = 1, \ldots, N_a$, proving the $m = 1$ case when $N_a \neq 0$ and $N_b \neq 0$.

Now consider the case where one of N_a and N_b is 0; assume without loss of generality that $N_b = 0$, then by division by $f^{(1)}(y_1; \boldsymbol{a}^1)$, the right-hand side is constant 0, while the left-hand side goes to $a_0^1 \neq 0$ unless $N_a = 0$, so $N_a = N_b = 0$.

Now suppose the result holds for some $m - 1 \geq 1$, and suppose either $N_a \neq 0$ or $N_b \neq 0$, otherwise there is nothing to prove. We denote $\boldsymbol{a}_1$ as the subvector of $\boldsymbol{a}$ corresponding to $\boldsymbol{d}$ with $d_1 \geq 1$, i.e. $\{a_{\boldsymbol{d}}\}_{\boldsymbol{d}\in\mathcal{D},\, d_1 \geq 1}$, and $\boldsymbol{a}_{-1}$ as that of $\boldsymbol{a}$ with $d_1 = 0$. Separating out the terms involving y_1,

$$\begin{aligned} f^{(m)}(\boldsymbol{y}; \boldsymbol{a}^i) &= \exp\left\{ \sum_{d=1}^{p} \left(\sum_{\boldsymbol{d}\in\mathcal{D},\, d_1=d} a_{\boldsymbol{d}}^i \prod_{j=2}^{m} y_j^{d_j} \right) y_1^d \right\} \exp\left(\sum_{\boldsymbol{d}\in\mathcal{D},\, d_1=0} a_{\boldsymbol{d}}^i \prod_{j=2}^{m} y_j^{d_j} \right) \\ &= f^{(1)}\left(y_1; \boldsymbol{a}_{1*}^i(\boldsymbol{y}_{-1})\right) f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^i), \end{aligned}$$

where $\boldsymbol{a}_{1*}^i(\boldsymbol{y}_{-1}) : \mathbb{R}^{m-1} \to \mathbb{R}^p$ is a vector-valued function in $\boldsymbol{y}_{-1}$, with d-th coordinate a polynomial $\sum_{\boldsymbol{d}\in\mathcal{D},\, d_1=d} a_{\boldsymbol{d}}^i \prod_{j=2}^{m} y_j^{d_j}$, and coefficients corresponding to $\boldsymbol{a}_1^i$. Note that there is a one-to-one correspondence between such a function $\boldsymbol{a}_{1*}^i$ and vector $\boldsymbol{a}_1^i$. So we can rewrite (1) as

$$\sum_{i=1}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}_{1*}^i(\boldsymbol{y}_{-1})) f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^i) = \sum_{i=1}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}_{1*}^i(\boldsymbol{y}_{-1})) f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{b}_{-1}^i)$$

for all $\boldsymbol{y} \in \mathbb{R}^m$. Then collecting terms with the same $f^{(1)}$ (same $\boldsymbol{a}_1^i$ ($\boldsymbol{a}_{1*}^i$) or $\boldsymbol{b}_1^i$ ($\boldsymbol{b}_{1*}^i$)),

$$\sum_{\ell=1}^{C} f^{(1)}(y_1; \boldsymbol{c}_{1*}^\ell(\boldsymbol{y}_{-1})) \left\{ \sum_{j=1}^{n_\ell^a} a_0^{k_{\ell j}^a} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^{k_{\ell j}^a}) + \sum_{j=1}^{n_\ell^b} b_0^{k_{\ell j}^b} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{b}_{-1}^{k_{\ell j}^b}) \right\} = 0, \tag{3}$$

where $C > 0$, each $\boldsymbol{c}_1^\ell$ (coefficients for $\boldsymbol{c}_{1*}^\ell$) is some $\boldsymbol{a}_1^i$ or $\boldsymbol{b}_1^i$, and $\{\boldsymbol{c}_1^\ell\}_{\ell=1}^C$ are distinct. Here, let $\{k_{11}^a, \ldots, k_{1,n_1^a}^a, \ldots, k_{C1}^a, \ldots, k_{Cn_C^a}^a\}$ be a permutation of $\{1, \ldots, N_a\}$, and $\{k_{11}^b, \ldots, k_{1,n_1^b}^b, \ldots, k_{C1}^b, \ldots, k_{Cn_C^b}^b\}$ a permutation of $\{1, \ldots, N_b\}$.

Since $\{\boldsymbol{c}_1^\ell\}_{\ell=1}^C$ are distinct, $\{\boldsymbol{c}_{1*}^\ell\}_{\ell=1}^C$ are distinct finite polynomials in $\boldsymbol{y}_{-1} \in \mathbb{R}^{m-1}$. For each pair of such distinct polynomials, the lemma of Okamoto (1973) implies that they only agree at a Lebesgue-null subset of $\mathbb{R}^{n-1}$, so all polynomials are distinct except on a null set. Thus, for almost every fixed $\boldsymbol{y}_{-1} \in \mathbb{R}^{m-1}$, the left-hand side of (3) is a sum of $C > 0$ distinct $f^{(1)}$'s in y_1 multiplied by constant weights depending on $\boldsymbol{y}_{-1}$. But the right-hand side is a sum of 0 terms, so by the result for $m = 1$ we necessarily have

$$\sum_{j=1}^{n_\ell^a} a_0^{k_{\ell j}^a} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^{k_{\ell j}^a}) = \sum_{j=1}^{n_\ell^b} -b_0^{k_{\ell j}^b} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{b}_{-1}^{k_{\ell j}^b}) \tag{4}$$

for all $\ell = 1, \ldots, C$ for almost every $\boldsymbol{y}_{-1}$. Fixing $\ell \in \{1, \ldots, C\}$, for any $1 \leq j_1 < j_2 \leq n_\ell^a$, $\boldsymbol{a}^{k_{\ell j_1}^a} \neq \boldsymbol{a}^{k_{\ell j_2}^a}$ and $\boldsymbol{a}_1^{k_{\ell j_1}^a} = \boldsymbol{a}_1^{k_{\ell j_2}^a}$ implies $\boldsymbol{a}_{-1}^{k_{\ell j_1}^a} \neq \boldsymbol{a}_{-1}^{k_{\ell j_2}^a}$, and similarly for $\boldsymbol{b}$. Thus, each term on the left-hand side of (4) has its unique coefficients, and similarly for the right-hand side. Since (4) holds for almost every $\boldsymbol{y}_{-1}$, by the result for $m-1$ variables, we must have $n_\ell^a = n_\ell^b$ and each $a_0^{k_{\ell j}^a} = b_0^{k_{\ell\pi(j)}^b}$ and $\boldsymbol{a}_{-1}^{k_{\ell j}^a} = \boldsymbol{b}_{-1}^{k_{\ell\pi(j)}^b}$ for some permutation π of $\{1,\dots,n_\ell^a\}$, which in turn implies $\boldsymbol{a}^{k_{\ell j}^a} = \boldsymbol{b}^{k_{\ell\pi(j)}^b}$ for all $j=1,\dots,n_\ell^a$ by construction of the groups $\ell = 1,\dots,C$. Since this holds for all ℓ, $N_a = \sum_{\ell=1}^C n_\ell^a = \sum_{\ell=1}^C n_\ell^b = N_b$, and we have thus again matched each $\boldsymbol{a}^\ell$ with a $\boldsymbol{b}^\ell$ as well as the corresponding a_0's with b_0's. This ends the proof for m, and the entire proof. ■

Proof [Proof of Theorem 6] Suppose $\mathcal{G}$ and $\mathcal{G}'$ have the same node set $\mathcal{V}$ and are Markov equivalent, otherwise the distributions represented by them are trivially not identical.

Now suppose $p(\boldsymbol{Y})$ is Markov and faithful with respect to $\mathcal{G}$ and $\mathcal{G}'$, and factorize w.r.t. both graphs with *strong Hurdle polynomial* parameters. Then by Proposition 8, there exist V_1 and V_2 such that $V_1 \to V_2$ in $\mathcal{G}$, $V_2 \to V_1$ in $\mathcal{G}'$ and $\mathcal{P} \equiv \mathrm{pa}_{\mathcal{G}}(V_2)\backslash\{V_1\} = \mathrm{pa}_{\mathcal{G}'}(V_1)\backslash\{V_2\}$. Following the arguments in the proof of Proposition 8 in Peters et al. (2014), recursively marginalizing out nodes without children but having the same parents in both graphs, we eventually obtain structures as follows, where $\mathcal{A}$ and $\mathcal{B}$ are some unknown node sets and V_2 does not have any children in Graph $\mathcal{G}$:

$$
\begin{array}{ccccccc}
\mathbf{V}_{\mathcal{A}} & & \mathbf{V}_{\mathcal{P}} & \qquad & \mathbf{V}_{\mathcal{A}} & & \mathbf{V}_{\mathcal{P}} \\
\searrow & \swarrow & \searrow & & \nwarrow & \swarrow & \diagdown \\
& V_1 \to V_2 & & & & V_1 \leftarrow V_2 & \\
& \downarrow & & & & \downarrow & \\
& \mathbf{V}_{\mathcal{B}} & & & & \mathbf{V}_{\mathcal{B}} & \\
& \text{Graph } \mathcal{G} & & & & \text{Graph } \mathcal{G}' &
\end{array}
$$

We consider the (α,β,k)-parametrization only, since the result for the (p,μ,σ^2) naturally follows from their relationship (5). For notational simplicity write V_1 and V_2 as nodes 1 and 2. Suppose after marginalization above we are left with nodes $\mathcal{V}_0 \subseteq \mathcal{V}$ which include 1, 2, $\mathbf{V}_{\mathcal{A}}$, $\mathbf{V}_{\mathcal{B}}$ and $\mathbf{V}_{\mathcal{P}}$ illustrated above. Now let $Y_U = 0$ for all $U \in \mathcal{V}_0\backslash\{2\}$, and let $Y_2 \neq 0$. Then the joint distribution $p(Y_2 = y_2 \neq 0, \boldsymbol{y}_{\mathcal{V}_0} = \mathbf{0})$ using $\mathcal{G}$ is proportional to

$$
\begin{aligned}
&\prod_{V\in\mathcal{V}_0} \left.\frac{\exp\{\alpha_V(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}}(V)})\mathbb{1}_{y_v} + \beta_v(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}}(V)})y_V - k_V y_V^2/2\}}{\sqrt{2\pi/k_V}\exp\{\alpha_V(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}}(V)}) + \beta_V(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}}(V)})^2/(2k_V)\}+1}\right|_{y_2\neq 0, \boldsymbol{y}_{\mathcal{V}_0\backslash\{2\}}=\mathbf{0}} \\
&\propto \exp\{\beta_2(\mathbf{0})y_2 - k_2 y_2^2/2\}
\end{aligned}
$$

since 2 does not have any child in $\mathcal{G}$. But using $\mathcal{G}'$, the same joint distribution is proportional to

$$
\begin{aligned}
&\prod_{V\in\mathcal{V}_0} \left.\frac{\exp\{\alpha'_V(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}'}(V)})\mathbb{1}_{y_V} + \beta'_V(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}'}(V)})y_V - k'_V y_V^2/2\}}{\sqrt{2\pi/k'_V}\exp\{\alpha'_V(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}'}(V)}) + \beta'_V(\boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}'}(V)})^2/(2k'_V)\}+1}\right|_{y_2\neq 0, \boldsymbol{y}_{\mathcal{V}_0\backslash\{2\}}=\mathbf{0}} \\
&\qquad\propto \exp\{\beta'_2(\mathbf{0})y_2 - k'_2 y_2^2/2\} \\
&\qquad\qquad \times \prod_{U\in\mathcal{P}\cup\{1\},\, 2\in\mathrm{pa}_{\mathcal{G}'}(U)} \frac{1}{\sqrt{2\pi/k'_U}\exp\{\alpha'_U(y_2,\mathbf{0}) + \beta'_U(y_2,\mathbf{0})^2/(2k'_U)\}+1}
\end{aligned}
$$

where in the case where $\mathrm{pa}_{\mathcal{G}'}(2) = \varnothing$ replace $\alpha_2'(\mathbf{0})$ and $\beta_2'(\mathbf{0})$ by constants α_2' and β_2', and $\alpha_U'(y_2, \mathbf{0})$ and $\beta_U'(y_2, \mathbf{0})$ denote setting all parents other than 2 in the Hurdle polynomials α_U' and β_U' to $\mathbf{0}$. Since the two joint distributions derived from both graphs must be proportional to each other, we get for $y_2 \neq 0$

$$\begin{aligned} &\exp\left[y_2\{\beta_2'(\mathbf{0}) - \beta_2(\mathbf{0})\} - (k_2' - k_2)y_2^2/2\right] \\ &\qquad \propto \prod_{U \in \mathcal{P} \cup \{1\},\, 2 \in \mathrm{pa}_{\mathcal{G}'}(U)} \left[\sqrt{2\pi/k_U'} \exp\left\{\alpha_U'(y_2, \mathbf{0}) + \beta_U'(y_2, \mathbf{0})^2/(2k_U')\right\} + 1\right]. \end{aligned} \tag{5}$$

Note that $2 \in \mathrm{pa}_{\mathcal{G}'}(1)$ and thus the product on the right of (5) has at least one term. Thus, supposing that for at least one of $U \in \mathcal{P} \cup \{1\}$ such that $2 \in \mathrm{pa}_{\mathcal{G}'}(U)$, $\alpha_U'(Y_2, \mathbf{0}) + \beta_U'(Y_2, \mathbf{0})^2/(2k_U')$ is nonconstant in $Y_2 \neq 0$, then the right-hand side of (5) can be expanded into a sum of at least two exponentials of polynomials in y_2 (including the constant 1 as a degenerated exponential polynomial), while the left-hand side is a single polynomial in y_2. This is a contradiction according to Lemma 1, and thus the assumption of having *strong Hurdle polynomials* as the parameters in the Hurdle conditionals implies that $p(\boldsymbol{Y})$ cannot be represented by both $\mathcal{G}$ and $\mathcal{G}'$, which ends the proof. ■

Proof [Proof of Theorem 10] As in the proof of Theorem 6 using Proposition 8, under the assumptions there exist V_1 and V_2 such that $\mathcal{P} \equiv \mathrm{pa}_{\mathcal{G}}(V_2)\backslash\{V_1\} = \mathrm{pa}_{\mathcal{G}'}(V_1)\backslash\{V_2\}$ with $V_1 \to V_2$ in $\mathcal{G}$ and $V_2 \to V_1$ in $\mathcal{G}'$. Following the arguments in the proof of Proposition 8 in Peters et al. (2014), recursively marginalizing out nodes without children but having the same parents in both graphs, we again obtain structures as follows:

$$\begin{array}{ccc} \mathbf{V}_{\mathcal{A}} \quad \mathbf{V}_{\mathcal{P}} & \qquad & \mathbf{V}_{\mathcal{A}} \quad \mathbf{V}_{\mathcal{P}} \\ \searrow \swarrow \searrow & & \nwarrow \swarrow \diagdown \\ V_1 \to V_2 & & V_1 \leftarrow V_2 \\ \downarrow & & \downarrow \\ \mathbf{V}_{\mathcal{B}} & & \mathbf{V}_{\mathcal{B}} \\ \text{Graph } \mathcal{G} & & \text{Graph } \mathcal{G}' \end{array}$$

To ease the notation assume we again write $V_1 = 1$ and $V_2 = 2$.Note that the distribution of each node conditional on some other nodes is the sum of a point mass at 0 and a continuous distribution over $\mathbb{R}$, which follows by induction and the fact that the indefinite integral of a continuous density is continuous and that the sum of continuous densities is continuous. We focus on the continuous components, and wish to reach the conclusion using the factorization

$$\begin{aligned} P(y_1, y_2 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) &= P(y_1 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) P(y_2 | y_1, \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) \\ &= P(y_2 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) P(y_1 | y_2, \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}), \end{aligned}$$

where the second terms in both decompositions are a regular Hurdle conditional w.r.t. $\mathcal{G}$ and $\mathcal{G}'$, respectively, and we write the first terms as

$$P(y_1 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) \propto \exp\{\mathbb{1}_{y_1}\delta_1 + f_1(y_1)\}$$

and

$$P(y_2|\boldsymbol{Y}_{\mathcal{P}}=\boldsymbol{y}_{\mathcal{P}}) \propto \exp\{\mathbb{1}_{y_2}\delta_2'+f_2'(y_1)\}$$

in terms of the conditional densities w.r.t. λ. Here f_1 and f_2' are continuous functions in $\mathbb{R}$ with no additive constant term, and δ_1 and δ_2' are constants.

We prove the results in the (α,β,k)-parameterization only, since results for the (p,μ,σ^2)-parameterization would follow from their relationship (5). In our model, we assumed the α and β parameters for each node to be polynomial in the parents and their indicators. We also assumed that for each node, either the β function is nonconstant in any of the parents, or α depends on the value of all of its parents.

Consider a generic β function associated with some generic parent set $\mathcal{P}\equiv\mathcal{P}_1\sqcup\{p_0\}$ with $p_0\notin\mathcal{P}_1\neq\varnothing$ and suppose that β is nonconstant in any of $\mathcal{P}$, and write $\beta(\boldsymbol{y}_{\mathcal{P}})$ equivalently as $\beta(y_{p_0},\boldsymbol{y}_{\mathcal{P}_1})$. Then $\beta(\boldsymbol{y}_{\mathcal{P}})$ has the form $\beta_{-1}(\boldsymbol{y}_{\mathcal{P}_1})+\beta_0(\boldsymbol{y}_{\mathcal{P}_1})\mathbb{1}_{y_1}+\sum_{i=1}^k\beta_i(\boldsymbol{y}_{\mathcal{P}_1})y_1^i$, where by construction β_{-1} through β_k are (potentially constant or even zero) Hurdle polynomials in $\boldsymbol{y}_{\mathcal{P}_1}$, but there must exist some $j=0,\ldots,k$ such that β_j is nonzero. By the lemma of Okamoto (1973), $\beta_j(\boldsymbol{y}_{\mathcal{P}_1})\neq 0$ for (Lebesgue) almost every $\boldsymbol{y}_{\mathcal{P}_1}\in\mathbb{R}^{|\mathcal{P}_1|}$. Thus, $\beta(y_{p_0},\boldsymbol{y}_{\mathcal{P}_1})$ is nonconstant in y_{p_0} for almost every $\boldsymbol{y}_{\mathcal{P}_1}\in\mathbb{R}^{|P_1|}$. Formally, define

$$\mathcal{Y}_{\beta,p_0,\mathcal{P}_1}\equiv\left\{\boldsymbol{y}_{\mathcal{P}_1}\in\mathbb{R}^{|\mathcal{P}_1|}:\beta\left(y_{p_0},\boldsymbol{y}_{\mathcal{P}_1}\right)\text{ nonconstant function in }y_{p_0}\right\}.$$

Thus $\mathbb{R}^{|\mathcal{P}_1|}\backslash\mathcal{Y}_{\beta,p_0,\mathcal{P}_1}$ has zero Lebesgue measure assuming β is nonconstant in its any of $\mathcal{P}$. Hence, by a similar argument, under the assumptions of the theorem, letting

$$\begin{aligned}\mathcal{Y}_{\alpha,\beta,p_0,\mathcal{P}_1}\equiv\Big\{\boldsymbol{y}_{\mathcal{P}_1}\in\mathbb{R}^{|\mathcal{P}_1|}:\beta\left(y_{p_0},\boldsymbol{y}_{\mathcal{P}_1}\right)&\text{ nonconstant function in }y_{p_0}\text{ or}\\ &\alpha\left(y_{p_0},\boldsymbol{y}_{\mathcal{P}_1}\right)\text{ depends on the value of }y_{p_0}\Big\},\end{aligned}$$

the set $\mathbb{R}^{|\mathcal{P}_1|}\backslash\mathcal{Y}_{\alpha,\beta,p_0,\mathcal{P}_1}$ has zero Lebesgue measure.

Now we go back to $\mathcal{G}$ and $\mathcal{G}'$. Suppose $\mathcal{P}\neq\varnothing$ and that the Hurdle density of node 2 conditional on $\{1\}\sqcup\mathcal{P}$ in $\mathcal{G}$ have α and β parameters $\alpha_2(y_1,\boldsymbol{y}_{\mathcal{P}})$ and $\beta_2(y_1,\boldsymbol{y}_{\mathcal{P}})$, and let those for 1 conditional on $\{2\}\sqcup\mathcal{P}$ in $\mathcal{G}'$ be $\alpha_1'(y_2,\boldsymbol{y}_{\mathcal{P}})$ and $\beta_1'(y_2,\boldsymbol{y}_{\mathcal{P}})$. We also denote $\mathcal{Y}_*\equiv\mathcal{Y}_{\alpha_2,\beta_2,1,\mathcal{P}}\cap\mathcal{Y}_{\alpha_1',\beta_1',2,\mathcal{P}}$, which by discussion above contains almost every $\boldsymbol{y}_{\mathcal{P}}\subset\mathbb{R}^{|\mathcal{P}|}$.

From now on we thus fix $\boldsymbol{y}_{\mathcal{P}}\in\mathcal{Y}_*$ and condition on $\boldsymbol{Y}_{\mathcal{P}}=\boldsymbol{y}_{\mathcal{P}}$, and omit the dependency of the α and β functions on $\mathcal{P}$, and write them as scalar functions instead notation-wise. By discussion above, β_2 becomes a nonconstant function in y_1 and β_1' becomes a nonconstant function in y_2. Note that for $\mathcal{P}=\varnothing$, we do not fix or condition on any parent variables and α_1', α_2, β_1' and β_2 are automatically univariate functions, with β_1' and β_2 nonconstant by assumption.

The joint density of $P(y_1,y_2|\boldsymbol{Y}_{\mathcal{P}}=\boldsymbol{y}_{\mathcal{P}})$ w.r.t. λ thus has two characterizations (up to normalizing constants)

$$\begin{aligned}&\frac{\exp\{\mathbb{1}_{y_1}\delta_1+f_1(y_1)+\mathbb{1}_{y_2}\alpha_2(y_1)+y_2\beta_2(y_1)-y_2^2k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}+1}\\ &\qquad\propto\frac{\exp\{\mathbb{1}_{y_2}\delta_2'+f_2'(y_2)+\mathbb{1}_{y_1}\alpha_1'(y_2)+y_1\beta_1'(y_2)-y_1^2k_1'/2\}}{\sqrt{2\pi/k_1'}\exp[\alpha_1'(y_2)+\{\beta_1'(y_2)\}^2/(2k_1')]+1},\end{aligned}\tag{6}$$

where $\alpha_2(y_1)$ has the form $c_{\alpha_2,-1}+c_{\alpha_2,0}\mathbb{1}_{y_1}+c_{\alpha_2,1}y_1+\cdots+c_{\alpha_2,k}y_1^k$ with coefficients being polynomials in $\boldsymbol{y}_{\mathcal{P}}$ and their indicators (or constants if $\mathcal{P}=\varnothing$), and similarly for $\beta_2(y_1)$, $\alpha_1'(y_2)$

and $\beta_1'(y_2)$. Note that if the values of $\mathbb{1}_{y_1}$ and $\mathbb{1}_{y_2}$ are given, these four functions are just polynomials in y_1 and y_2, respectively.

First condition on the event $\mathbb{1}_{y_1} = \mathbb{1}_{y_2} = 1$ that has a positive probability. Then (6) becomes

$$\begin{aligned}
&\frac{\exp\{f_1(y_1) + \alpha_2(y_1) + y_2\beta_2(y_1) - y_2^2 k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1}\mathbb{1}_{y_1}\mathbb{1}_{y_2}, \\
&\qquad\propto \frac{\exp\{f_2'(y_2) + \alpha_1'(y_2) + y_1\beta_1'(y_2) - y_1^2 k_1'/2\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + (\beta_1'(y_2))^2/(2k_1')\} + 1}\mathbb{1}_{y_1}\mathbb{1}_{y_2}, \qquad (7)
\end{aligned}$$

for all $(y_1, y_2) \in (\mathbb{R}\backslash\{0\})^2$. (7) has the form

$$\frac{\exp\{f_1(y_1) + P_1(y_1, y_2)\}}{\exp\{P_2(y_1)\} + 1} = \frac{\exp\{f_2'(y_2) + P_3(y_1, y_2)\}}{\exp\{P_4(y_2)\} + 1},$$

where P_1 and P_3 are polynomials in y_1 and y_2 simultaneously, possibly with interactions from the $y_2\beta_2(y_1)$ and $y_1\beta_1'(y_2)$ terms, and P_2 and P_4 are univariate polynomials in y_1, y_2, respectively. By cross-multiplication,

$$\begin{aligned}
&\exp\{f_1(y_1) + P_1(y_1, y_2) + P_4(y_2)\} + \exp\{f_1(y_1) + P_1(y_1, y_2)\} \\
&\qquad = \exp\{f_2'(y_2) + P_3(y_1, y_2) + P_2(y_1)\} + \exp\{f_2'(y_2) + P_3(y_1, y_2)\}. \qquad (8)
\end{aligned}$$

Differentiating both sides of (8) with respect to y_1,

$$\begin{aligned}
&\left[\frac{\partial}{\partial y_1}\{f_1(y_1) + P_1(y_1, y_2)\}\right]\exp\{f_1(y_1) + P_1(y_1, y_2) + P_4(y_2)\} \\
&\qquad + \exp\{f_1(y_1) + P_1(y_1, y_2))\} \\
&= \left[\frac{\partial}{\partial y_1}\{P_3(y_1, y_2) + P_2(y_1)\}\right]\exp\{f_2'(y_2) + P_3(y_1, y_2) + P_2(y_1)\} \\
&\qquad + \left\{\frac{\partial}{\partial y_1}P_3(y_1, y_2)\right\}\exp\{f_2'(y_2) + P_3(y_1, y_2)\}. \qquad (9)
\end{aligned}$$

Plugging (8) into the left-hand side of (9),

$$\begin{aligned}
&\left[\frac{\partial}{\partial y_1}\{f_1(y_1) + P_1(y_1, y_2)\}\right]\left[\exp\{f_2'(y_2) + P_3(y_1, y_2) + P_2(y_1)\}\right. \\
&\qquad \left. + \exp\{f_2'(y_2) + P_3(y_1, y_2)\}\right] \\
&= \left[\frac{\partial}{\partial y_1}\{P_3(y_1, y_2) + P_2(y_1)\}\right]\exp\{f_2'(y_2) + P_3(y_1, y_2) + P_2(y_1)\} \\
&\qquad + \left\{\frac{\partial}{\partial y_1}P_3(y_1, y_2)\right\}\exp\{f_2'(y_2) + P_3(y_1, y_2)\},
\end{aligned}$$

which simplifies to

$$\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)-P_2(y_1)\right\}\right]$$
$$\times\exp\left\{f_2'(y_2)+P_3(y_1,y_2)+P_2(y_1)\right\}$$
$$+\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)\right\}\right]$$
$$\times\exp\left\{f_2'(y_2)+P_3(y_1,y_2)\right\}=0.$$

Since $\exp\left\{f_2'(y_2)+P_3(y_1,y_2)\right\}\neq 0$, this becomes

$$\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)-P_2(y_1)\right\}\right]\exp\left\{P_2(y_1)\right\}$$
$$+\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)\right\}\right]=0. \quad (10)$$

Focusing on the components that involve y_2, we see that

$$\left[\frac{\partial}{\partial y_1}\left\{P_1(y_1,y_2)-P_3(y_1,y_2)\right\}\right]\left[\exp\left\{P_2(y_1)\right\}+1\right]$$

does not depend on y_2. Since $(\exp(P_2(y_1))+1)>0$, we have

$$\frac{\partial^2}{\partial y_1\partial y_2}\left\{P_1(y_1,y_2)-P_3(y_1,y_2)\right\}=0.$$

Recall that

$$P_1(y_1,y_2)-P_3(y_1,y_2)=\alpha_2(y_1)+y_2\beta_2(y_1)-y_2^2k_2/2-\alpha_1'(y_2)-y_1\beta_1'(y_2)+y_1^2k_1'/2. \quad (11)$$

So $0=\frac{\partial^2}{\partial y_1\partial y_2}\{P_1(y_1,y_2)-P_3(y_1,y_2)\}=\frac{\mathrm{d}\beta_2(y_1)}{\mathrm{d}y_1}-\frac{\mathrm{d}\beta_1'(y_2)}{\mathrm{d}y_2}$ implies that β_2 and β_1' are both linear with the same coefficient on the linear term. Now that β_2 has the form $\beta_2(y_1)=c_{\beta_2,-1}+c_{\beta_2,0}\mathbb{1}_{y_1}+c_{\beta_2,1}y_1$, write $\beta_{2;-1,0}\equiv c_{\beta_2,-1}+c_{\beta_2,0}=\beta_2(0)+c_{\beta_2,0}$ as a shorthand notation for β_2 with indicator set to 1 while y_1 set to 0. Similarly define $\beta_{1;-1,0}'\equiv c_{\beta_1',-1}+c_{\beta_1',0}=\beta_1'(0)+c_{\beta_1',0}$. Then for $y_1,y_2\neq 0$ since $c_{\beta_2,1}=c_{\beta_1',1}$, we necessarily have

$$\begin{aligned}y_2\beta_2(y_1)-y_1\beta_1'(y_2)&=y_2(c_{\beta_2,-1}+c_{\beta_2,0}+c_{\beta_2,1}y_1)-y_1(c_{\beta_1',-1}+c_{\beta_1',0}+c_{\beta_1',1}y_2)\\&=y_2\beta_{2;-1,0}-y_1\beta_{1;-1,0}',\end{aligned}$$

and so by (11)

$$\begin{aligned}&P_1(y_1,y_2)-P_3(y_1,y_2)\\&=\left(\alpha_2(y_1)-y_1\beta_{1;-1,0}'+y_1^2k_1'/2\right)-\left(\alpha_1'(y_2)-y_2\beta_{2;-1,0}+y_2^2k_2/2\right)\\&\equiv P_{1,3}(y_1)-(\text{function in } y_2 \text{ only}).\end{aligned}$$

Plugging this into (10), we get

$$\left[\frac{\mathrm{d}}{\mathrm{d}y_1}\left\{f_1(y_1)+P_{1,3}(y_1)-P_2(y_1)\right\}\right]\exp\left\{P_2(y_1)\right\}+\left[\frac{\mathrm{d}}{\mathrm{d}y_1}\left\{f_1(y_1)+P_{1,3}(y_1)\right\}\right]$$

equals 0, or equivalently

$$\left[\frac{\mathrm{d}}{\mathrm{d}y_1}\{f_1(y_1)+P_{1,3}(y_1)\}\right][\exp\{P_2(y_1)\}+1]=\left\{\frac{\mathrm{d}}{\mathrm{d}y_1}P_2(y_1)\right\}\exp\{P_2(y_1)\}.$$

Then

$$\begin{aligned}f_1(y_1)&=\int\frac{\exp\{P_2(y_1)\}\{\mathrm{d}P_2(y_1)/\,\mathrm{d}y_1\}}{\exp\{P_2(y_1)\}+1}\,\mathrm{d}y_1-P_{1,3}(y_1)\\&=\log\left[1+\exp\{P_2(y_1)\}\right]-P_{1,3}(y_1)+\text{const}.\end{aligned}$$

So for $y_1\neq 0$,

$$\begin{aligned}\exp(f_1(y_1))&\propto\frac{1+\exp\{P_2(y_1)\}}{\exp\{P_{1,3}(y_1)\}}\\&=\frac{1+\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}}{\exp\{\alpha_2(y_1)-\beta'_{1;-1,0}y_1+y_1^2k'_1/2\}}\\&=\exp\{-\alpha_2(y_1)+y_1\beta'_{1;-1,0}-y_1^2k'_1/2\}\\&\qquad+\sqrt{2\pi/k_2}\exp\{y_1\beta'_{1;-1,0}+\beta_2(y_1)^2/(2k_2)-y_1^2k'_1/2\}.\end{aligned}\tag{12}$$

Now condition on the event $\mathbb{1}_{y_1}=1$ and $\mathbb{1}_{y_2}=0$. Then (6) becomes

$$\frac{\exp\{f_1(y_1)\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}+1}\mathbb{1}_{y_1}\propto\exp\{y_1\beta'_1(0)-y_1^2k'_1/2\}\mathbb{1}_{y_1},$$

which implies that for $y_1\neq 0$,

$$\begin{aligned}\exp\{f_1(y_1)\}&\propto\exp\{y_1\beta'_1(0)-y_1^2k'_1/2\}\\&\quad+\sqrt{2\pi/k_2}\exp\left\{y_1\beta'_1(0)-y_1^2k'_1/2+\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\right\}.\end{aligned}\tag{13}$$

Applying Lemma 1 to (12) and (13), by matching the terms we have (conditional on $y_1\neq 0$) either

$$-\alpha_2(y_1)+y_1\beta'_{1;-1,0}=y_1\beta'_1(0)+\text{const};\quad\text{or}\tag{14}$$

$$\begin{aligned}-\alpha_2(y_1)+y_1\beta'_{1;-1,0}&=y_1\beta'_1(0)+\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)+\text{const}\quad\text{and}\\y_1\beta'_{1;-1,0}+\beta_2(y_1)^2/(2k_2)&=y_1\beta'_1(0)+\text{const}.\end{aligned}\tag{15}$$

Conditional on $y_1\neq 0$, in the first case (14), $\alpha_2(y_1)=y_1c_{\beta'_1,0}+\text{const}$; in the second case (15), $\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)=\text{const}$ and $\beta_2(y_1)^2/(2k_2)=-y_1c_{\beta'_1,0}+\text{const}$, which implies $\beta_2(y_1)=\text{const}$ and $\alpha_2(y_1)=\text{const}$ for $y_1\neq 0$, and $c_{\beta'_1,0}=0$, which in turn implies (14). Thus, in either case, $\alpha_2(y_1)=c_{\alpha_2,0}\mathbb{1}_{y_1}+y_1c_{\beta'_1,0}+\text{const}$, i.e. α_2 is linear (or constant) in $y_1\neq 0$ with coefficient on y_1 equal to $c_{\beta'_1,0}$. By (14) for $y_1\neq 0$,

$$\begin{aligned}\exp\{f_1(y_1)\}&\propto\exp\{y_1\beta'_1(0)-y_1^2k'_1/2\}\\&\quad+\sqrt{2\pi/k_2}\exp\{y_1\beta'_{1;-1,0}+\beta_2(y_1)^2/(2k_2)-y_1^2k'_1/2\},\end{aligned}\tag{16}$$

clearly a single univariate Gaussian or a mixture of two univariate Gaussian distributions (since β_2 is at most linear in y_1). Similarly, we must have $\alpha_1'(y_2) = y_2\beta_{2;-1,0} - y_2\beta_2(0) + \text{const} = y_2 c_{\beta_2,0} + \text{const}$ for $y_2 \neq 0$, and for $y_2 \neq 0$

$$\exp\{f_2'(y_2)\} \propto \exp\{y_2\beta_2(0) - y_2^2 k_2/2\} \\ + \sqrt{2\pi/k_1'}\exp\{y_2\beta_{2;-1,0} + \beta_1'(y_2)^2/(2k_1') - y_2^2 k_2/2\}. \quad (17)$$

Now suppose by contradiction that $\exp\{f_1(y_1)\}$ given $y_1 \neq 0$ has only one Gaussian component, instead of being a sum of two Gaussian densities. Then by (16), $\beta_1'(0) = \beta_{1;-1,0}'$ and $\beta_2(y_1)$ is a constant given $\mathbb{1}_{y_1}$, i.e. $\beta_2(y_1) = c_{\beta_2,-1} + c_{\beta_2,0} = \beta_{2;-1,0}$ for $y_1 \neq 0$. Plugging this into the left-hand side of (6) and integrating w.r.t. $\lambda(y_1)$, the continuous part ($y_2 \neq 0$) of the marginal distribution of y_2 given $\boldsymbol{Y}_{\mathcal{P}} \equiv \boldsymbol{y}_{\mathcal{P}}$ is

$$\exp\{f_2'(y_2)\} \propto \frac{\exp\{f_1(0) + \alpha_2(0) + y_2\beta_2(0) - y_2^2k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(0) + \beta_2(0)^2/(2k_2)\} + 1} \\ + \exp\{y_2\beta_{2;-1,0} - y_2^2k_2/2\}\int_{\mathbb{R}} \frac{\exp\{\delta_1 + f_1(y_1) + \alpha_2(y_1)\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1}\,\mathrm{d}y_1,$$

which is a mixture between $\mathcal{N}(\beta_2(0)/k_2, 1/k_2)$ and $\mathcal{N}(\beta_{2;-1,0}/k_2, 1/k_2)$, i.e. the variance in both components are equal. Note that the integral in the second term is a Lebesgue integral. This together with (17) implies that $\beta_1'(y_2)$ cannot depend on the value of y_2 given $y_2 \neq 0$, i.e. $\beta_1'(y_2) = c_{\beta_1',-1} + c_{\beta_1',0} = \beta_{1;-1,0}'$. Since we already know that $\beta_1'(0) = \beta_{1;-1,0}'$ by discussion above, this implies that β_1' is an absolute constant in y_2 and $\mathbb{1}_{y_2}$, and also that α_2 may depend on y_1 only through $\mathbb{1}_{y_1}$, a contradiction to the assumption of the theorem.

Thus, (16) and (17) will both have to be mixtures of precisely two Gaussians, and so by definition the joint distribution $p(\boldsymbol{Y})$ of $\boldsymbol{Y}$ must be of 2-Gaussian type with respect to $\mathcal{G}$ and $\mathcal{G}'$. ■

Proof [Proof of Corollary 11] When $|\mathcal{V}| = 2$, in Proposition 8 we always have $\mathcal{P} = \varnothing$ and V_1 does not have a parent in $\mathcal{G}$, so $P(Y_{V_1} = y|Y_{V_1} \neq 0)$ by definition is just a Gaussian, not a mixture two Gaussians, and hence $p(\boldsymbol{Y})$ cannot be of 2-Gaussian type with respect to any pairs of distinct Markov equivalent graphs.

Now consider $|\mathcal{V}| = 3$, and assume the two vertices with reversible edges in Proposition 8 are V_1 and V_2, and that $V_1 \to V_2$ in $\mathcal{G}$ and $V_1 \leftarrow V_2$ in $\mathcal{G}'$. If neither V_1 or V_2 has V_3 as its parent in both graphs, then we can marginalize V_3 out and it reduces to the 2-d case. Suppose otherwise. Then we must have (1) $V_1 \to V_2 \leftarrow V_3$ in $\mathcal{G}$, or (2) $V_2 \to V_1 \leftarrow V_3$ in $\mathcal{G}'$, or (3) an additional edge between V_1 and V_3 added to (1), or (4) an additional edge between V_2 and V_3 added to (2).

For (1) and (2) both graphs are the only graph in their Markov equivalence class; for (3) the reversible edge becomes V_1—V_3 violating the assumption (and in fact one can marginalize out the common child V_2 and get back to the 2-d case), and similarly for (4). Thus, we have again ruled out the possibility of any pair of distinct Markov equivalent graphs with respect to which $p(\boldsymbol{Y})$ can be of 2-Gaussian type. ■

Remark 2 *In the proof of Theorem 10, we proved that whenever $p(\boldsymbol{Y})$ factorizes with respect to two distinct graphs $\mathcal{G}$ and $\mathcal{G}'$ (whenever identifiability does not hold), everything up to (17) in the* *proof must hold. Specifically, conditioning on almost every $\boldsymbol{y}_{\mathcal{P}}$, α_2 and β_2 in $\mathcal{G}$ as well as α_1' and β_1' in $\mathcal{G}'$ can be at most linear in y_1 and y_2, respectively, namely*

$$\begin{aligned} \beta_1'(y_2) &= c_{\beta_1',-1} + c_{\beta_1',0}\mathbb{1}_{y_2} + c_{\beta_1',1}y_2, \quad \beta_2(y_1) = c_{\beta_2,-1} + c_{\beta_2,0}\mathbb{1}_{y_1} + c_{\beta_2,1}y_1, \\ \alpha_1'(y_2) &= c_{\alpha_1',-1} + c_{\alpha_1',0}\mathbb{1}_{y_2} + c_{\alpha_1',1}y_2, \quad \alpha_2(y_1) = c_{\alpha_2,-1} + c_{\alpha_2,0}\mathbb{1}_{y_1} + c_{\alpha_2,1}y_1, \end{aligned}$$

with coefficients depending on $\boldsymbol{y}_{\mathcal{P}}$ where

$$c_{\alpha_1',1} = c_{\beta_2,0}, \quad c_{\alpha_2,1} = c_{\beta_1',0}, \quad c_{\beta_1',1} = c_{\beta_2,1}. \tag{18}$$

It is noted that, although not used in deriving our conclusion involving 2-Gaussian type distributions, we in addition also have the following results.

$$c_{\alpha_1',-1} = c_{\alpha_2,-1}, \quad c_{\alpha_1',0} = c_{\alpha_2,0}, \quad c_{\alpha_1',-1} + c_{\alpha_1',0} = c_{\alpha_2,-1} + c_{\alpha_2,0} = 0.$$

These might shed some light on how to show that distributions of 2-Gaussian type do not exist for a general $m \geq 4$.

Proof [Proof of Remark 2]

By (6), (16), (17), the joint distribution of Y_1 and Y_2 conditional on $\boldsymbol{Y}_{\mathcal{P}}$ has two characterizations (up to normalizing constants)

$$\begin{aligned} &\frac{\exp\{\mathbb{1}_{y_1}\delta_1 + y_1\beta_1'(0) - y_1^2k_1'/2 + \mathbb{1}_{y_2}\alpha_2(y_1) + y_2\beta_2(y_1) - y_2^2k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1} \\ &+ \frac{\sqrt{2\pi/k_2}\exp\{\mathbb{1}_{y_1}\delta_1 + y_1\beta_{1;-1,0}' + \beta_2(y_1)^2/(2k_2) - y_1^2k_1'/2 + \mathbb{1}_{y_2}\alpha_2(y_1) + y_2\beta_2(y_1) - y_2^2k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1} \\ &\propto \frac{\exp\{\mathbb{1}_{y_2}\delta_2' + y_2\beta_2(0) - y_2^2k_2/2 + \mathbb{1}_{y_1}\alpha_1'(y_2) + y_1\beta_1'(y_2) - y_1^2k_1'/2\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + \beta_1'(y_2)^2/(2k_1')\} + 1} \\ &+ \frac{\sqrt{2\pi/k_1'}\exp\{\mathbb{1}_{y_2}\delta_2' + y_2\beta_{2;-1,0} + \beta_1'(y_2)^2/(2k_1') - y_2^2k_2/2 + \mathbb{1}_{y_1}\alpha_1'(y_2) + y_1\beta_1'(y_2) - y_1^2k_1'/2\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + \beta_1'(y_2)^2/(2k_1')\} + 1}. \end{aligned} \tag{19}$$

Divide both sides by $\exp(y_1\beta_1'(0) + y_2\beta_2(0) - y_1^2k_1'/2 - y_2^2k_2/2)$ and expanding $\beta_1'(y_2)$ and $\beta_2(y_1)$, this becomes

$$\begin{aligned} &\frac{\exp\{\mathbb{1}_{y_1}\delta_1 + \mathbb{1}_{y_2}\alpha_2(y_1) + y_2c_{\beta_2,0}\mathbb{1}_{y_1} + y_1y_2c_{\beta_2,1}\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1} \\ &+ \frac{\sqrt{2\pi/k_2}\exp\{\mathbb{1}_{y_1}\delta_1 + y_1c_{\beta_1',0} + \beta_2(y_1)^2/(2k_2) + \mathbb{1}_{y_2}\alpha_2(y_1) + y_2c_{\beta_2,0}\mathbb{1}_{y_1} + y_1y_2c_{\beta_2,1}\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1} \\ &\propto \frac{\exp\{\mathbb{1}_{y_2}\delta_2' + \mathbb{1}_{y_1}\alpha_1'(y_2) + y_1c_{\beta_1',0}\mathbb{1}_{y_2} + y_1y_2c_{\beta_1',1}\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + \beta_1'(y_2)^2/(2k_1')\} + 1} \\ &+ \frac{\sqrt{2\pi/k_1'}\exp\{\mathbb{1}_{y_2}\delta_2' + y_2c_{\beta_2,0} + \beta_1'(y_2)^2/(2k_1') + \mathbb{1}_{y_1}\alpha_1'(y_2) + y_1c_{\beta_1',0}\mathbb{1}_{y_2} + y_1y_2c_{\beta_1',1}\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + \beta_1'(y_2)^2/(2k_1')\} + 1}. \end{aligned}$$

Now expanding $\alpha_1'(y_2)$ and $\alpha_2(y_1)$ and using the relationships in (18), we divide both sides by $\exp(y_1 c_{\alpha_2,1}\mathbb{1}_{y_2} + y_2 c_{\beta_2,0}\mathbb{1}_{y_1} + y_1 y_2 c_{\beta_2,1}) = \exp(y_1 c_{\beta_1',0}\mathbb{1}_{y_2} + y_2 c_{\alpha_1',1}\mathbb{1}_{y_1} + y_1 y_2 c_{\beta_2,1})$ and get

$$\begin{aligned}
&\frac{\exp\{\mathbb{1}_{y_1}\delta_1 + \mathbb{1}_{y_2}(c_{\alpha_2,-1} + c_{\alpha_2,0}\mathbb{1}_{y_1})\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1} \\
&+ \frac{\sqrt{2\pi/k_2}\exp\left\{\mathbb{1}_{y_1}\delta_1 + y_1 c_{\beta_1',0} + \beta_2(y_1)^2/(2k_2) + \mathbb{1}_{y_2}(c_{\alpha_2,-1} + c_{\alpha_2,0}\mathbb{1}_{y_1})\right\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1} \\
&= C_0\frac{\exp\{\mathbb{1}_{y_2}\delta_2' + \mathbb{1}_{y_1}(c_{\alpha_1',-1} + c_{\alpha_1',0}\mathbb{1}_{y_2})\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + \beta_1'(y_2)^2/(2k_1')\} + 1} \\
&+ C_0\frac{\sqrt{2\pi/k_1'}\exp\{\mathbb{1}_{y_2}\delta_2' + y_2 c_{\beta_2,0} + \beta_1'(y_2)^2/(2k_1') + \mathbb{1}_{y_1}(c_{\alpha_1',-1} + c_{\alpha_1',0}\mathbb{1}_{y_2})\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + \beta_1'(y_2)^2/(2k_1')\} + 1}
\end{aligned} \tag{20}$$

for some C_0. Setting $\mathbb{1}_{y_1} = \mathbb{1}_{y_2} = 0$ (20) becomes

$$\frac{1 + \sqrt{2\pi/k_2}\exp\{c^2_{\beta_2,-1}/(2k_2)\}}{\sqrt{2\pi/k_2}\exp\{c_{\alpha_2,-1} + c^2_{\beta_2,-1}/(2k_2)\} + 1} = C_0\frac{1 + \sqrt{2\pi/k_1'}\exp\{c^2_{\beta_1'1,-1}/(2k_1')\}}{\sqrt{2\pi/k_1'}\exp\{c_{\alpha_1',-1} + c^2_{\beta_1',-1}/(2k_1')\} + 1}, \tag{21}$$

and with $\mathbb{1}_{y_1} \neq 0$, $\mathbb{1}_{y_2} = 0$ (20) becomes

$$\begin{aligned}
&\exp(\delta_1)\frac{1 + \sqrt{2\pi/k_2}\exp\{y_1 c_{\beta_1',0} + \beta_2(y_1)^2/(2k_2)\}}{\sqrt{2\pi/k_2}\exp\{c_{\alpha_2,-1} + c_{\alpha_2,0} + c_{\alpha_2,1}y_1 + \beta_2(y_1)^2/(2k_2)\} + 1} \\
&\qquad = C_0\exp(c_{\alpha_1',-1})\frac{1 + \sqrt{2\pi/k_1'}\exp\{c^2_{\beta_1'1,-1}/(2k_1')\}}{\sqrt{2\pi/k_1'}\exp\{c_{\alpha_1',-1} + c^2_{\beta_1',-1}/(2k_1')\} + 1}.
\end{aligned} \tag{22}$$

Since the right-hand side of (22) is a constant, by matching the numerator and the denominator of the left-hand side using Lemma 1, we must have either (i) $y_1 c_{\beta_1',0} + \beta_2(y_1)^2/(2k_2) = c_{\alpha_2,-1} + c_{\alpha_2,0} + c_{\alpha_2,1}y_1 + \beta_2(y_1)^2/(2k_2)$, or (ii) $y_1 c_{\beta_1',0} + \beta_2(y_1)^2/(2k_2) = \text{const}$ for $y_1 \neq 0$. But (ii) implies that $c_{\beta_2,1} = c_{\beta_1',0} = 0$, which by $c_{\beta_1',1} = c_{\beta_2,1}$ implies that β_1' is an absolute constant in $y_2 \in \mathbb{R}$, a violation to the assumption. Thus (i) holds, and by $c_{\beta_1',0} = c_{\alpha_2,1}$ this implies that

$$\alpha_{2;-1,0} \equiv c_{\alpha_2,-1} + c_{\alpha_2,0} = 0, \quad \text{and by symmetry} \quad \alpha'_{1;-1,0} \equiv c_{\alpha_1',-1} + c_{\alpha_1',0} = 0. \tag{23}$$

Thus the left-hand side of (22) is just $\exp(\delta_1)$. Note that the right-hand side of (22) is $\exp(c_{\alpha_1',-1})$ times the right-hand side of (21). So by equating the left-hand side of (22) with $\exp(c_{\alpha_1',-1})$ times the left-hand side of (21) we have

$$\exp(\delta_1) = \exp\left(c'_{\alpha_1,-1}\right)\frac{1 + \sqrt{2\pi/k_2}\exp\left\{c^2_{\beta_2,-1}/(2k_2)\right\}}{\sqrt{2\pi/k_2}\exp\left\{c_{\alpha_2,-1} + c^2_{\beta_2,-1}/(2k_2)\right\} + 1)} \tag{24}$$

and similarly

$$\exp(\delta_2') = \exp\left(c_{\alpha_2,-1}\right)\frac{1 + \sqrt{2\pi/k_1'}\exp\left\{c^2_{\beta_1'1,-1}/(2k_1')\right\}}{\sqrt{2\pi/k_1'}\exp\left\{c_{\alpha_1',-1} + c^2_{\beta_1',-1}/(2k_1')\right\} + 1}. \tag{25}$$

Now by (23), with $\mathbb{1}_{y_1} = \mathbb{1}_{y_2} = 1$, (20) simplifies to $\exp(\delta_1) = C_0 \cdot \exp(\delta_2')$. Thus by (21), (24) and (25), one get

$$C_0 = \frac{\exp(\delta_1)}{\exp(\delta_2')} = \frac{\exp(c'_{\alpha_1,-1})\frac{1+\sqrt{2\pi/k_2}\exp\left\{c^2_{\beta_2,-1}/(2k_2)\right\}}{\sqrt{2\pi/k_2}\exp\left\{c_{\alpha_2,-1}+c^2_{\beta_2,-1}/(2k_2)\right)+1\right\}}}{\exp(c_{\alpha_2,-1})\frac{1+\sqrt{2\pi/k'_1}\exp\left\{c^2_{\beta'_1 1,-1}/(2k'_1)\right\}}{\sqrt{2\pi/k'_1}\exp\left\{c_{\alpha'_1,-1}+c^2_{\beta'_1,-1}/(2k'_1)\right\}+1}} = \frac{\exp(c'_{\alpha_1,-1})}{\exp(c_{\alpha_2,-1})}C_0$$

and thus $c'_{\alpha_1,-1} = c_{\alpha_2,-1}$. Combining with (23), we get

$$c_{\alpha'_1,-1} = c_{\alpha_2,-1}, \quad c_{\alpha'_1,0} = c_{\alpha_2,0}, \quad c_{\alpha'_1,-1} + c_{\alpha'_1,0} = c_{\alpha_2,-1} + c_{\alpha_2,0} = 0. \tag{26}$$

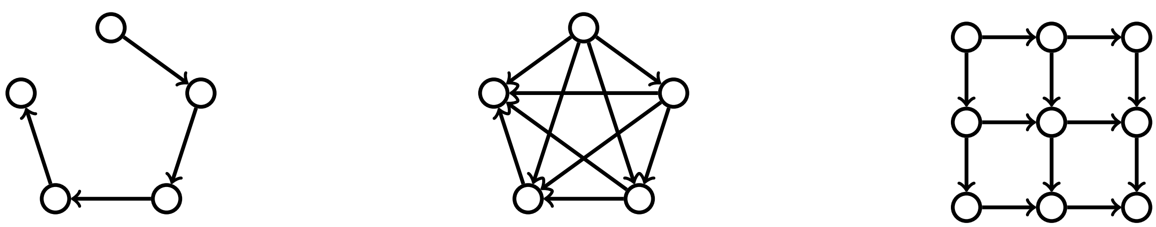
Note that this result holds as long as we assume identifiability does not hold. ■

B. Results of Additional Numerical Experiments

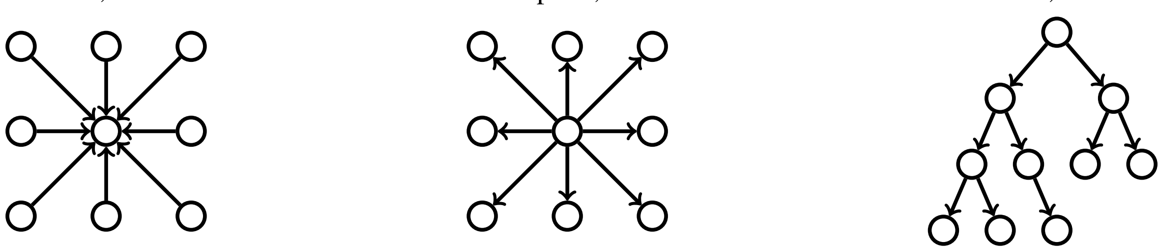
Here we provide additional details and results of numerical experiments.

Figure 1 shows the true DAG structures used in the simulation studies in Section 5.

In Figure S2, we present pairwise scatter plots of one instance of data generated with the chain graph (upper row) and the complete graph (lower row), respectively, both with (p, μ, k)-linear parametrization. Since the true topological ordering is $1 \to 2 \to 3 \to 4 \to 5$, for clarity we exclude the source and sink nodes (1 and 5) and only include nodes 2, 3 and 4. Plots on the left are



Chain, $m = 5$ Complete, $m = 5$ Lattice, $m = 9$



Star-in, $m = 9$ Star-out, $m = 9$ Tree, $m = 10$

Figure 1: Example graph structures used in our experiments.

plotted in the order $2, 3, 4$ and those on the right are reversed. In the histograms on the diagonals we only plot the continuous part.

The scatter plots indicate a slight difference in the respective marginal distributions of nodes 2 and 4 conditioned on node 3 being 0 (and vice versa). This difference intuitively explains how the orientation $2 \to 3 \to 4$ versus $4 \to 3 \to 2$ can be identified. It is worth noting that other than this difference, the marginal statistics for the three nodes are indistinguishable and there is little noticeable difference between plots on the left and on the right.

B.1. Additional Results for Exhaustive Search

Recall that we consider the following DAG structures: (i) chain graph with $m = 10$, (ii) complete graph with $m = 10$, (iii) lattice graph with $m = 9$; see Figure 1 in B for an illustration of the DAG structures.

The results for correctly specified models are shown in Figures S3–S5. Each figure has one true underlying DAG from those mentioned above. In all figures, each row indicates one choice of true data generating parametrization—(α, β, k)-linear, and (p, μ, σ^2)-linear and quadratic—and each column shows the results using each estimating parametrization. Thus, plots on the diagonal (with bold titles) correspond to correct parametrizations, where the estimating parametrization agrees with the truth. Off-diagonal plots, in contrast, correspond to cases where the model parametrization is misspecified.

The results indeed indicate that in all settings, exhaustive search with correct parametrization almost always identifies the exact DAG for large n. Surprisingly, model misspecification does not seem to negatively impact the results by a significant amount. Overall, our simulation studies confirm the identifiability theory (Theorem 6).

Figure S6–S8 show exact recovery rates. In the plots, exact success rates are measured by the percentage of times (out of $B = 100$ iterations for each setting) that the exact DAG is recovered, whereas the equivalent success rates stand for the percentage of times the equivalence class of DAG is correctly identified, and are thus no less than the exact success rates.

B.2. Results for Greedy Search (GDS)

To evaluate the performance of greedy search we consider the following graphs: (i) chain graph with $m = 100$, (ii) complete graph with $m = 10$, (iii) lattice graph with $m = 100$, (iv) star-in graph with $m = 20$ ($j \to 1$ for $j = 2, \ldots, m$), (v) star-out graph with $m = 100$ ($1 \to j$ for $= 2, \ldots, m$), (vi) tree graph with $m = 100$ ($j \to 2j$ for $j \leq \lfloor m/2 \rfloor$ and $j \to 2j + 1$ for $j \leq \lfloor (m-1)/2 \rfloor$).

Results for the greedy search algorithm are shown in Figures S9 and S10, where each row corresponds to a different true graph, and each column corresponds to one of the three aforementioned parametrizations, where for simplicity we only present the results with correctly specified parametrization.

The results indicate that GDS works reasonably well in all settings but may require larger samples for recovering the structure of complete/ very dense graphs, or graphs with high in-degrees. While exhaustive search often succeeds with high probability even with small samples, it may not be scalable for large m. In such cases, the greedy and faster GDS method, which shows promising results, provides a viable alternative. Utilizing the stability selection method of Shah and Samworth (2013) can further improve the GDS results.

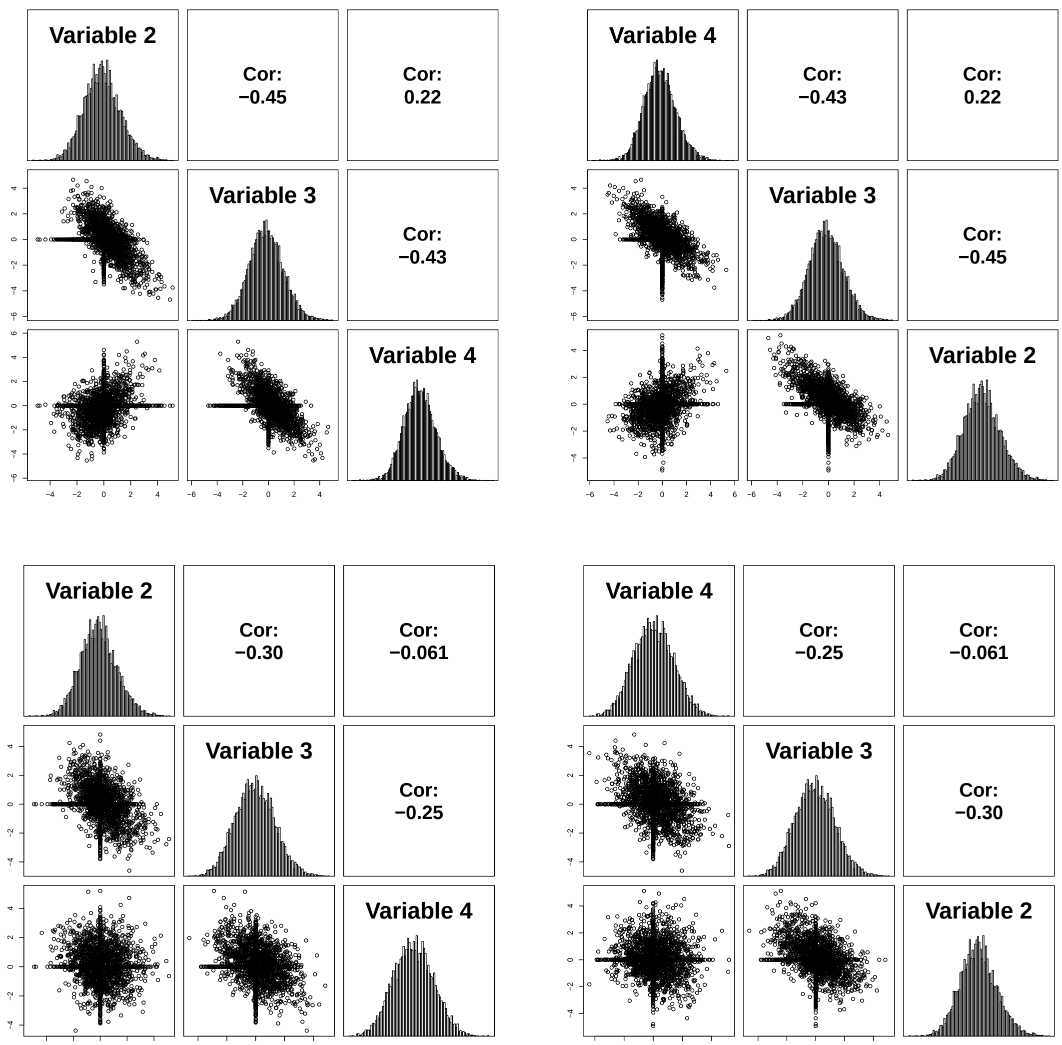


Figure S2: Pairwise scatterplots of zero-inflated data generated using chain graphs (upper row) and complete graphs (lower row), both with topological ordering $1 \to 2 \to 3 \to 4 \to 5$; only nodes 2, 3 and 4 are plotted. Plots on the left are plotted in the order $2, 3, 4$, and $4, 3, 2$ on the right. Only the continuous part is plotted in the histograms on the diagonals. There is little noticeable difference between the histograms and scatter plots when we reverse the graph order, yet our methods can still determine the correct topological ordering.

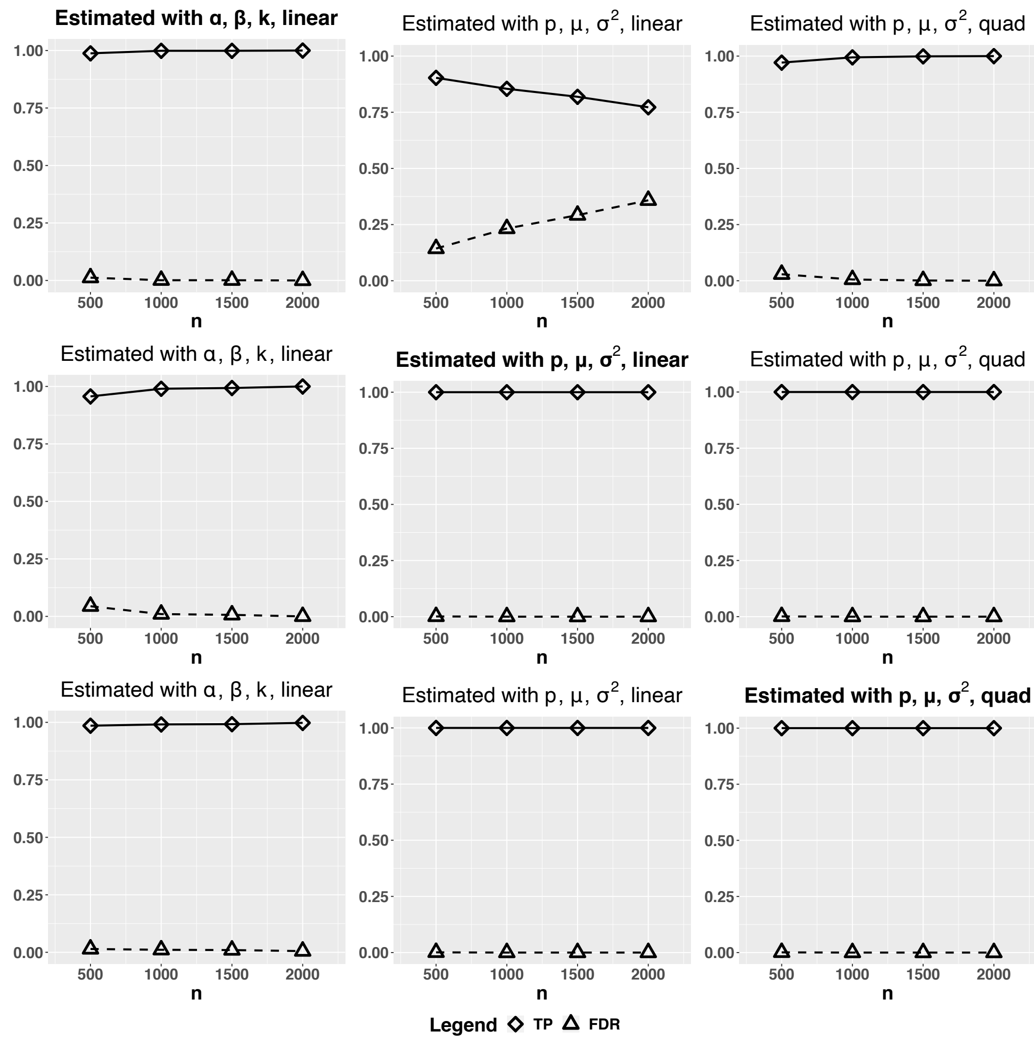


Figure S3: Chain graph, $m = 10$, exhaustive search. Each row corresponds to a different generating parametrization, and each column a different estimating parametrization. Generating and estimating parametrizations agree on the diagonal. '$\Diamond$' with solid lines: true positive rate; 'Δ' with dashed lines: false discovery rate.

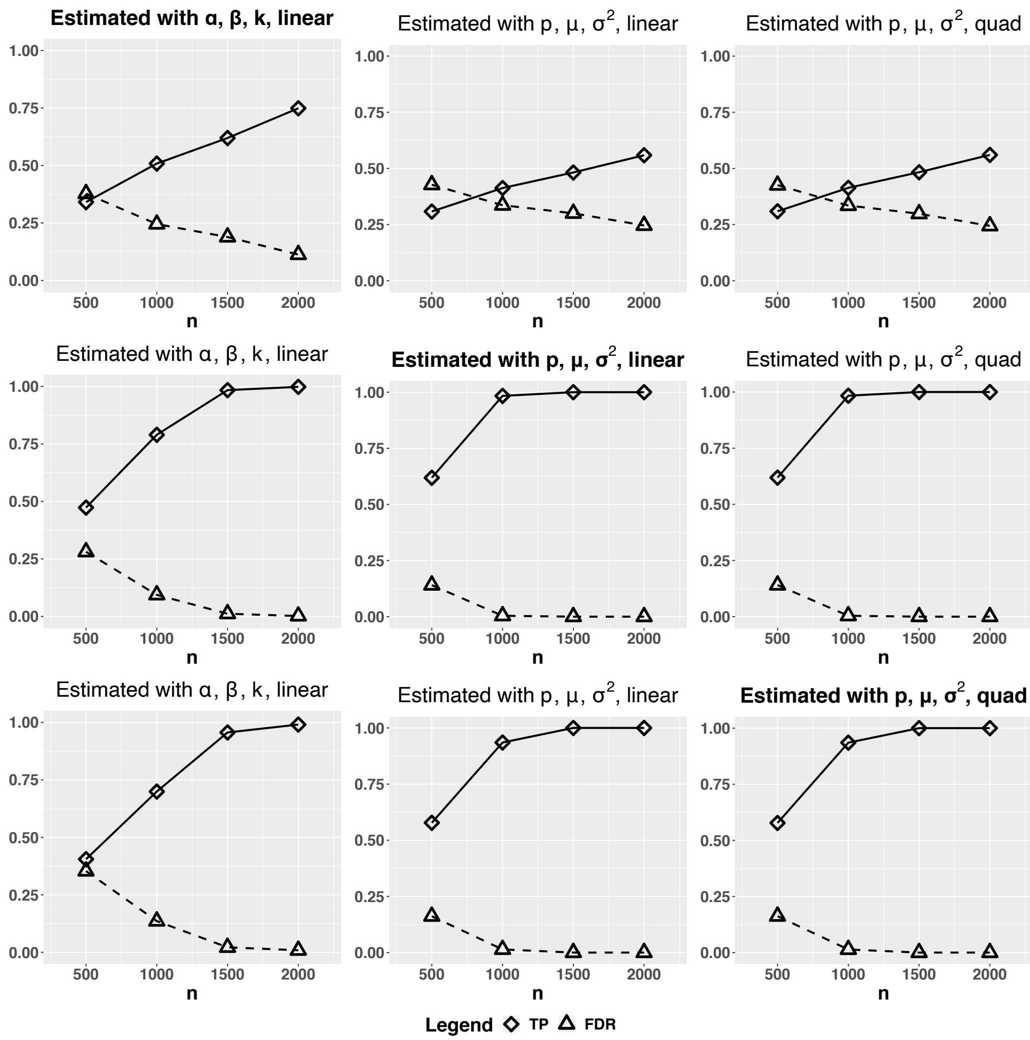


Figure S4: Complete graph, $m = 10$, exhaustive search. Each row corresponds to a different generating parametrization, and each column a different estimating parametrization. Generating and estimating parametrizations agree on the diagonal. '$\Diamond$' with solid lines: true positive rate; 'Δ' with dashed lines: false discovery rate.

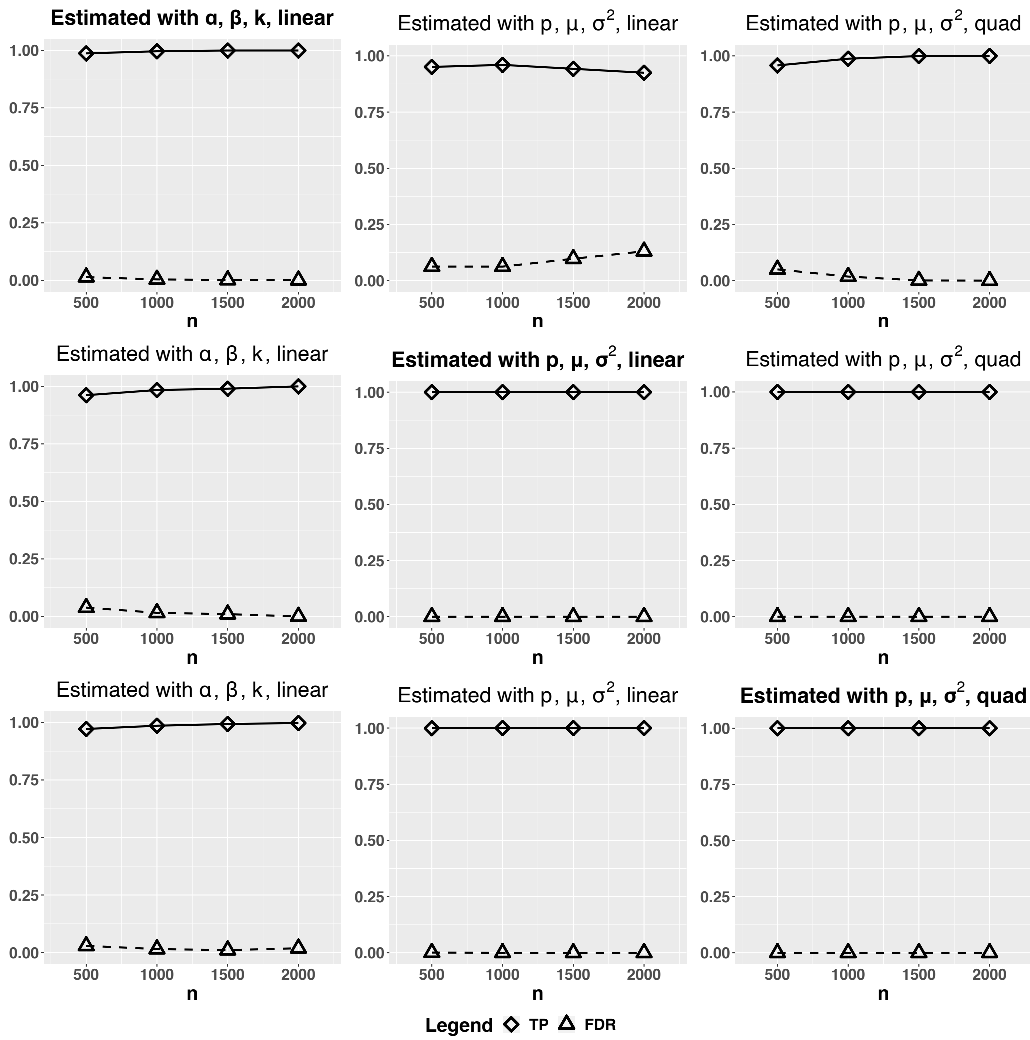


Figure S5: Lattice graph, $m = 9$, exhaustive search. Each row corresponds to a different generating parametrization, and each column a different estimating parametrization. Generating and estimating parametrizations agree on the diagonal. '$\Diamond$' with solid lines: true positive rate; 'Δ' with dashed lines: false discovery rate.

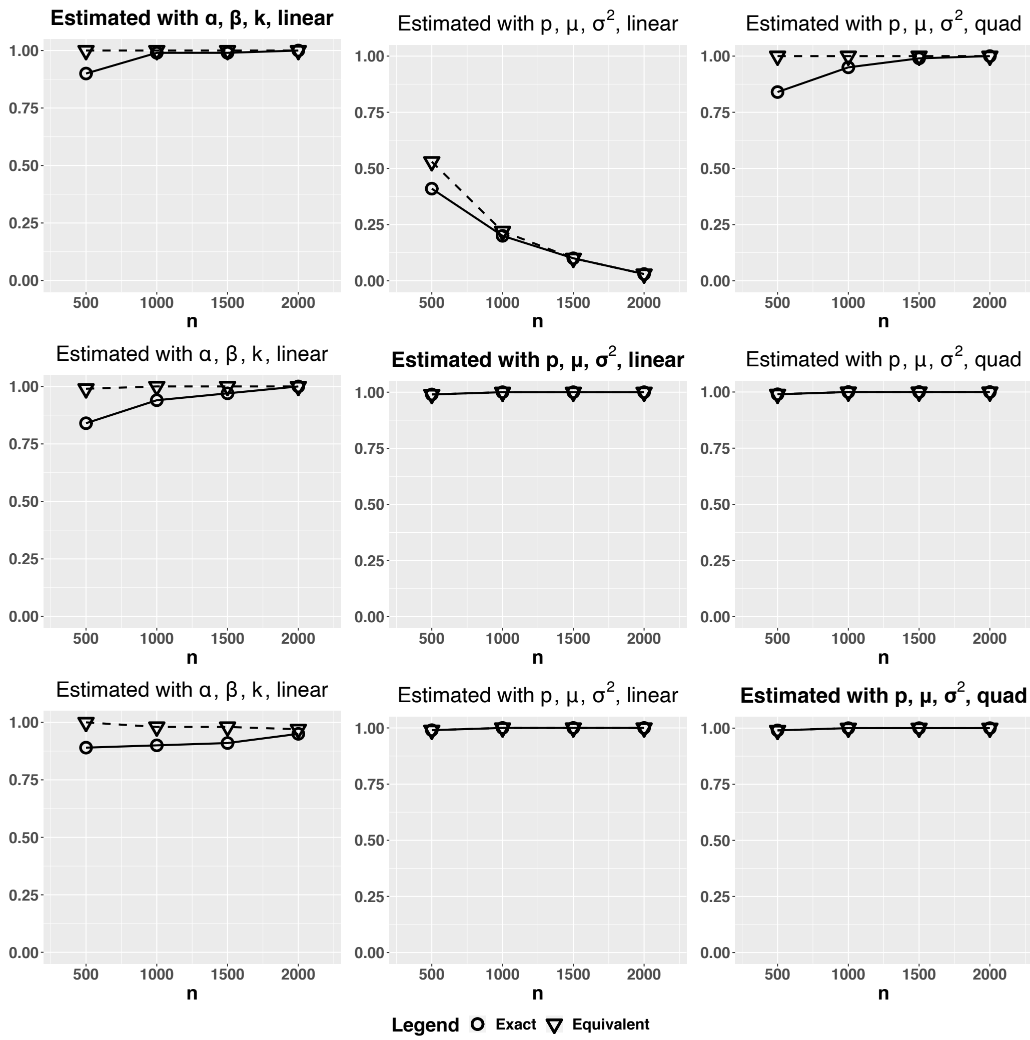


Figure S6: Chain graph, $m = 10$, exhaustive search. Each row corresponds to a different generating parametrization, and each column a different estimating parametrization. Generating and estimating parametrizations agree on the diagonal. '$\circ$' with solid lines: success rates of exact DAG recovery; '∇' with dashed lines: success rates for recovery of equivalence class.

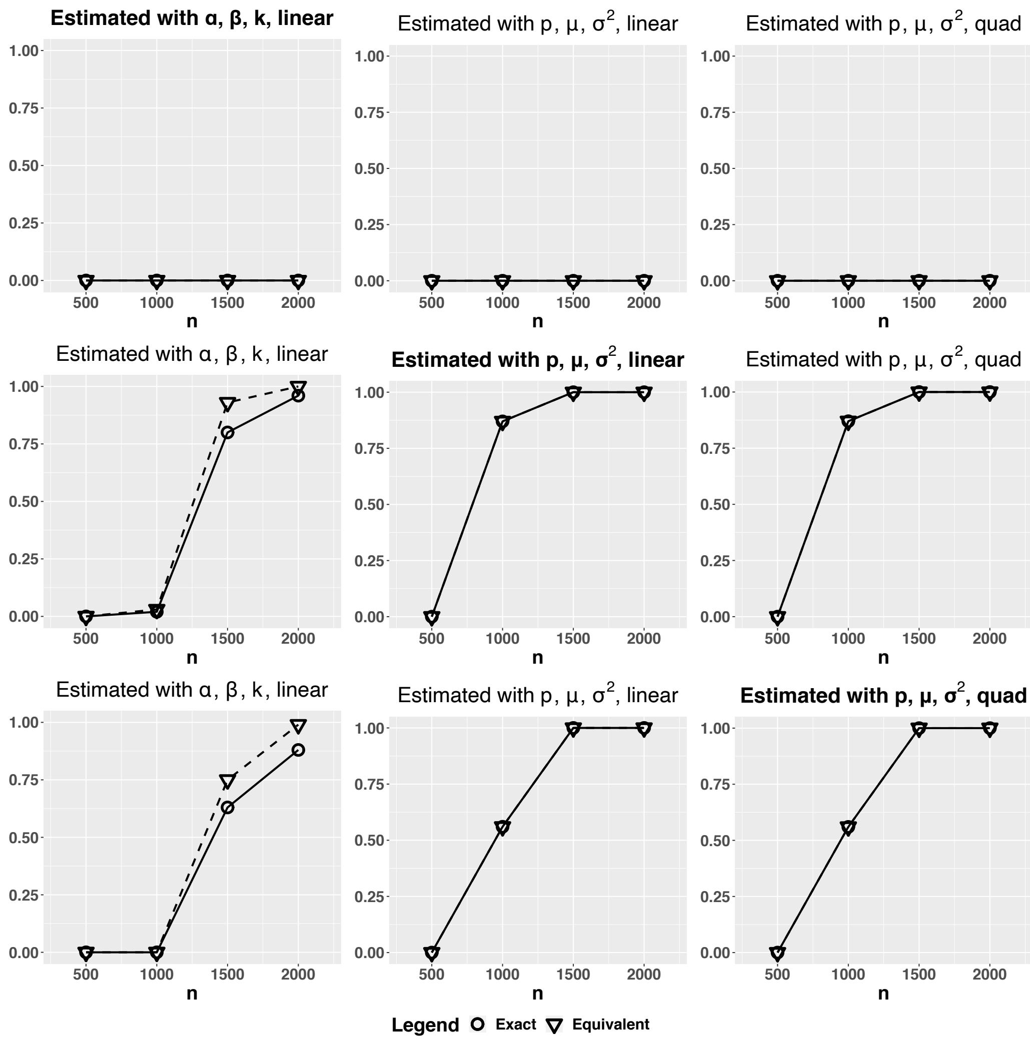


Figure S7: Complete graph, $m = 10$, exhaustive search. Each row corresponds to a different generating parametrization, and each column a different estimating parametrization. Generating and estimating parametrizations agree on the diagonal. '○' with solid lines: success rates of exact DAG recovery; '∇' with dashed lines: success rates for recovery of equivalence class.

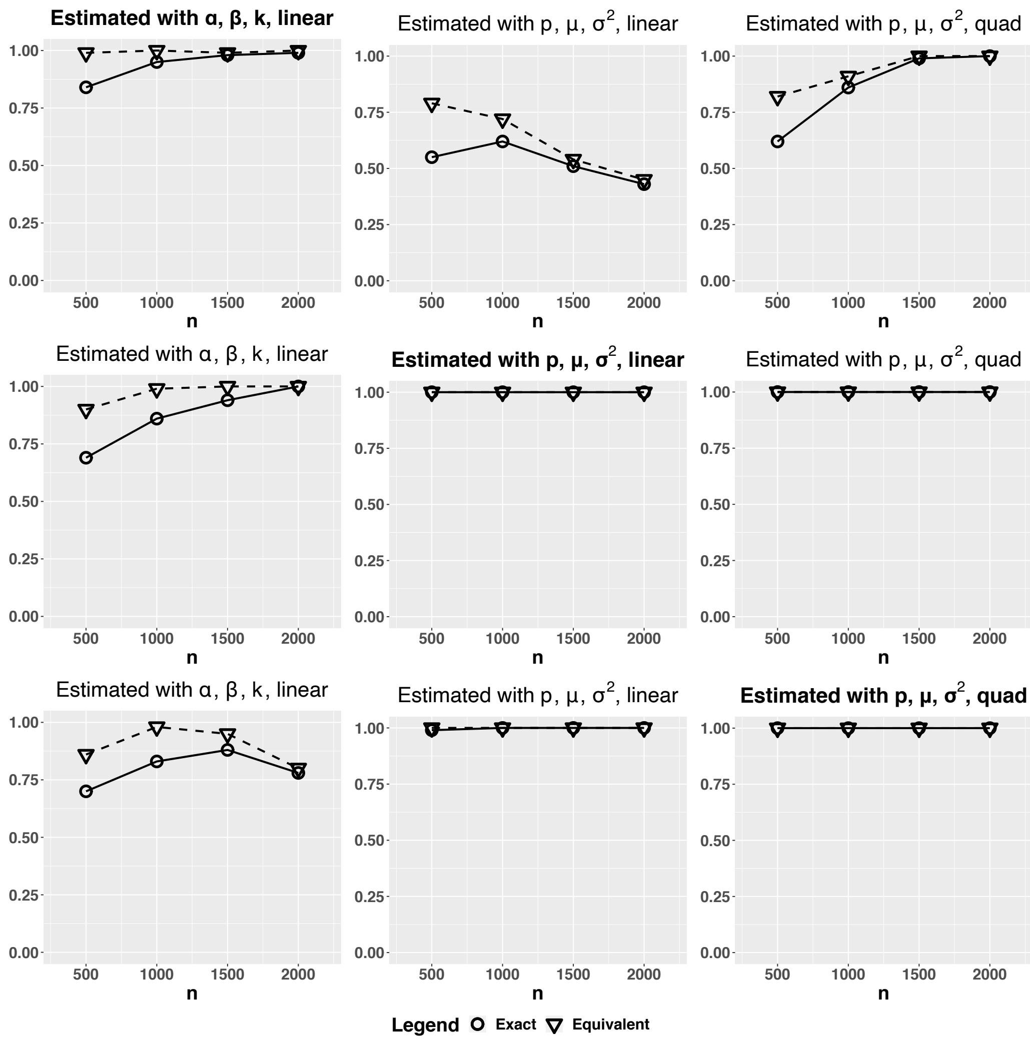


Figure S8: Lattice graph, $m = 9$, exhaustive search. Each row corresponds to a different generating parametrization, and each column a different estimating parametrization. Generating and estimating parametrizations agree on the diagonal. '$\circ$' with solid lines: success rates of exact DAG recovery; '∇' with dashed lines: success rates for recovery of equivalence class.

B.3. Details on Estimation of Connected Components

In this section we present simulation results validating our the strategy in Section 4.2, namely first applying the procedure of McDavid et al. (2019), then estimating the directed graphs inside each connected component of its estimated undirected graph. We measure the quality of the connected components (CC) estimated compared to the truth. The underlying true graph is a block-diagonal graph with $m = 100$ nodes, evenly divided into 10 connected components, where each connected component has the exact same setting as the complete graph with $m = 10$ nodes in previous sections. Specifically in each trial, we apply the method of McDavid et al. (2019), pick the estimate

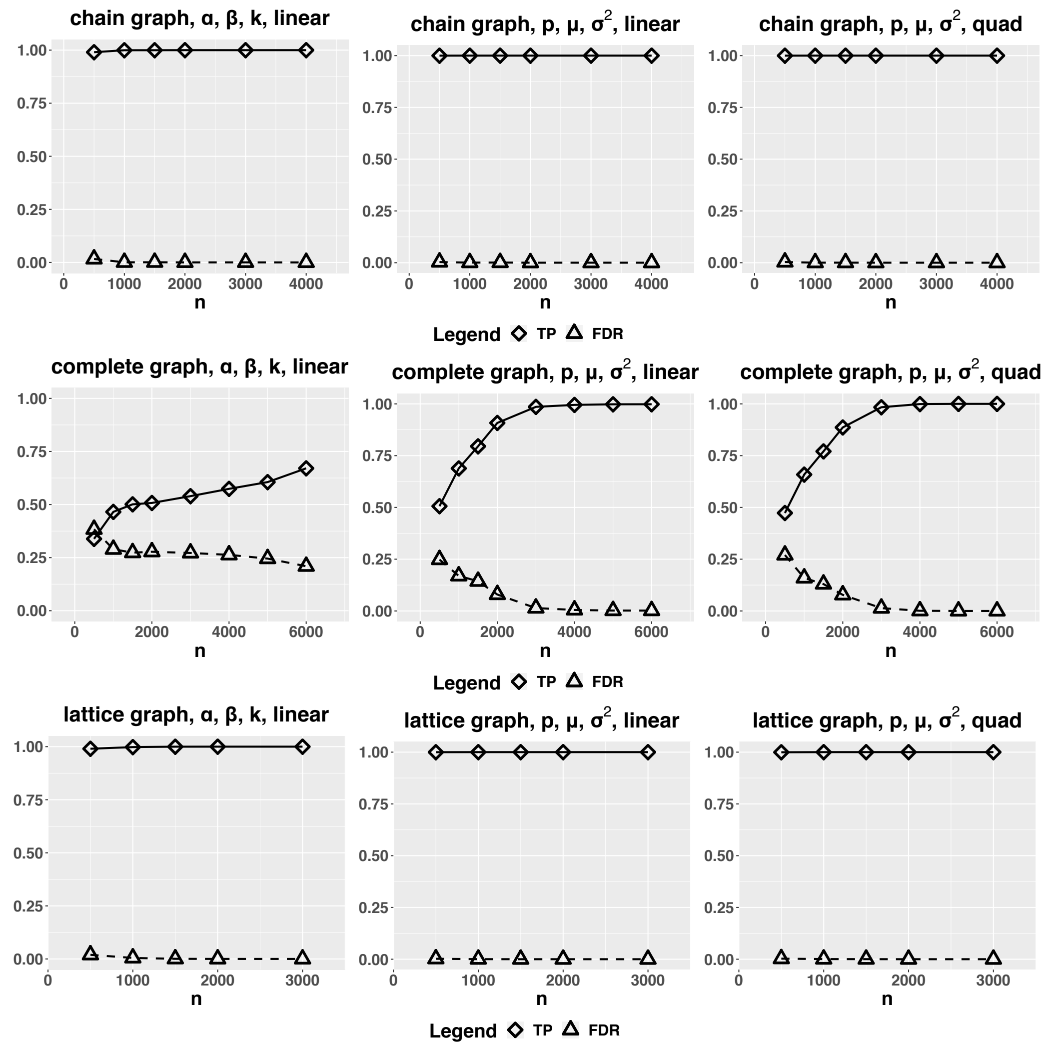


Figure S9: Results for GDS for chain graph with $m = 100$, complete graph with $m = 10$ and lattice graph with $m = 100$. Each row corresponds to a different graph structure, and each column corresponds to a different parametrization; the generating and estimating parametrizations are the same in the results. '$\diamond$' with solid lines: true positive rate; 'Δ' with dashed lines: false discovery rate.

that minimizes the BIC (with the exception of metric (b) below) and use the "and" rule to find the undirected graph (since the estimate is asymmetric due to their neighborhood selection method).

As in Section B.2, each column corresponds to a different parametrization (estimating parametrization = truth). As before all results shown are averaged over 100 trials. Each row contains two different metrics with value in $[0, 1]$, which we explain below.

(a) "Subset/eq CCs": A trial is count as successful if each true CC is a subset of some estimated CC, i.e. if any two truly connected nodes are estimated to belong to the same CC; the proportion of successful trials is reported.

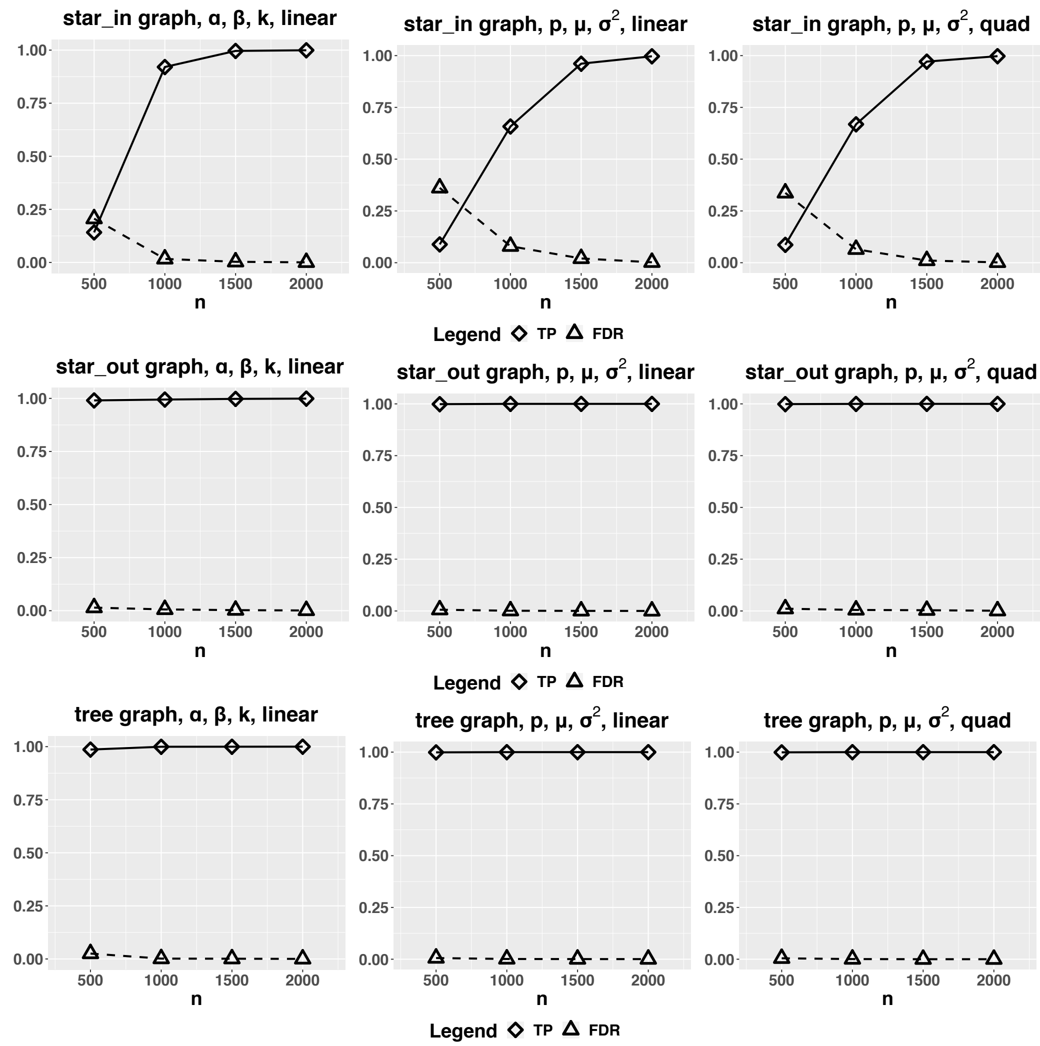


Figure S10: Results for GDS for star_in graph with $m = 20$, star_out graph with $m = 100$ and tree graph with $m = 100$. Each row corresponds to a different graph structure, and each column corresponds to a different parametrization; the generating and estimating parametrizations are the same in the results. '$\Diamond$' with solid lines: true positive rate; 'Δ' with dashed lines: false discovery rate.

(b) "Correct CCs-Oracle": From the solution path of McDavid et al. (2019) we take the graph assuming we know the true number of CCs. Then a trial counts as successful if the estimated CCs are exactly equal to the truth, and the proportion of successful trials is reported.

(c) "Correct #CCs": A trial is treated as successful if the estimated number of CCs is equal to the truth (10), and the proportion of successful trials is reported.

(d) "Avg #CCs/10": Average of the estimated numbers of CCs over 100 trials, divided by the truth (10).

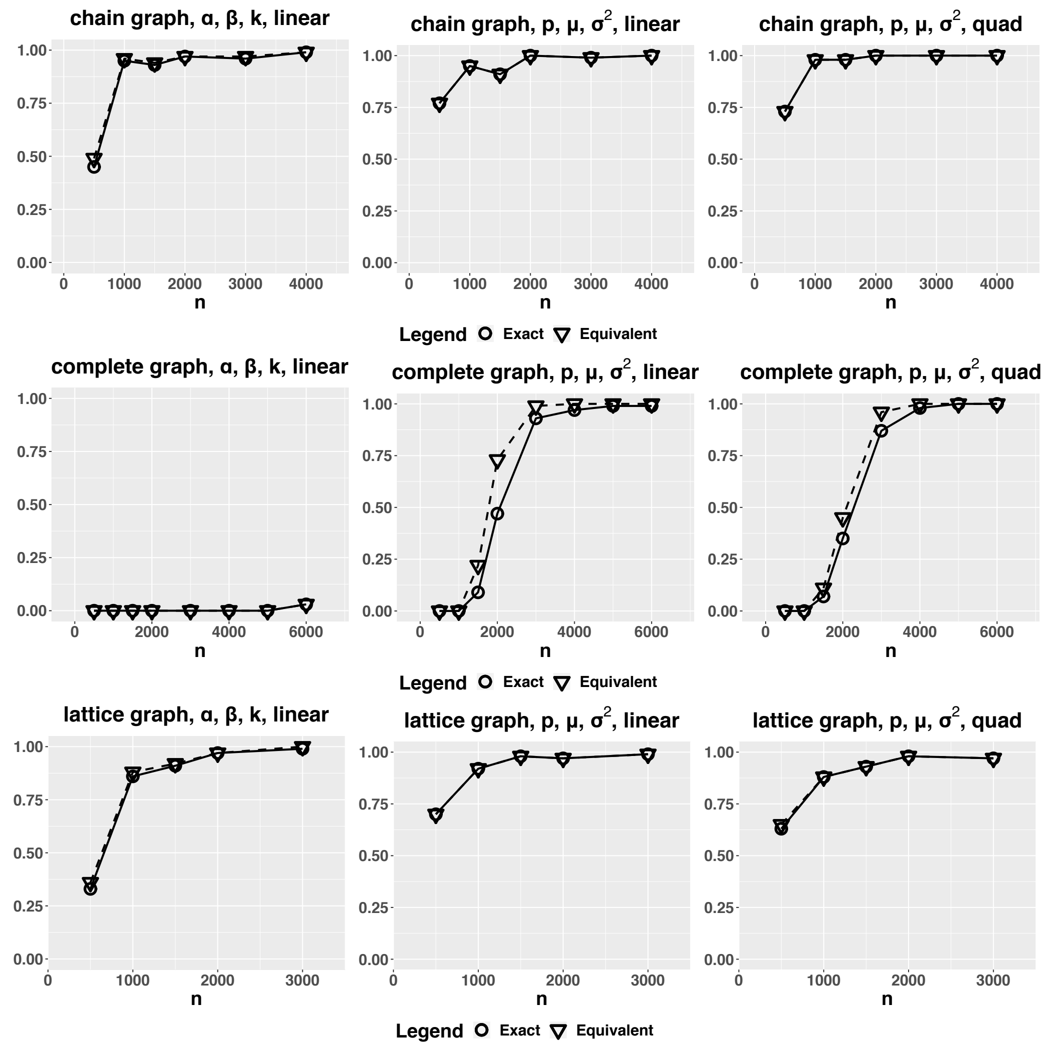


Figure S11: Results for GDS for chain graph with $m = 100$, complete graph with $m = 10$ and lattice graph with $m = 100$. Each row corresponds to a different graph structure, and each column corresponds to a different parametrization; the generating and estimating parametrizations are the same in the results. '$\circ$' with solid lines: success rates of exact DAG recovery; '∇' with dashed lines: success rates for recovery of equivalence class.

(e)&(f) "TP/FDR": The true positive rate and false discovery rate for undirected graph recovery.

A high (a) metric guards against the mistake of failing to keep two truly connected nodes in the same CC, while (c) and (d) measure how many CCs the procedure actually generates, since the trivial case where all nodes form a single CC is undesirable. Note this characterizes the statistical versus computational trade-off discussed in Section 4.2. Metric (b), on the other hand, attempts to test if the procedure can become perfect had it known the true number of CCs. Metrics (e) and (f) provide additional information from the edge recovery perspective in terms of undirected graphs.

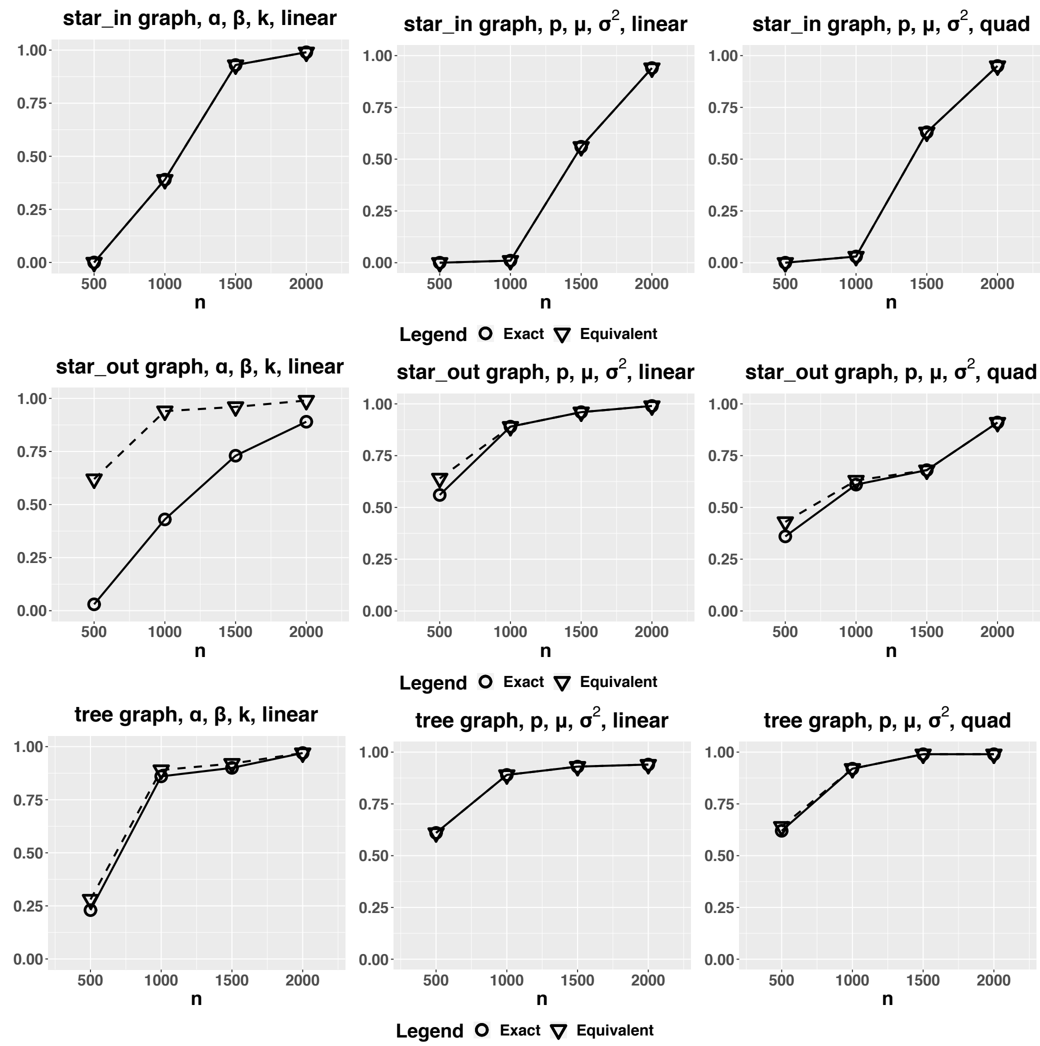


Figure S12: Results for GDS for star_in graph with $m = 20$, star_out graph with $m = 100$ and tree graph with $m = 100$. Each row corresponds to a different graph structure, and each column corresponds to a different parametrization; the generating and estimating parametrizations are the same in the results. '$\circ$' with solid lines: success rates of exact DAG recovery; '∇' with dashed lines: success rates for recovery of equivalence class.

Except for (f) which should be close to 0, one would hope for (a)–(e) to be close to 1, which indeed is the case, except for the number of CCs for the α, β, k parametrization.

References

Andrew McDavid, Raphael Gottardo, Noah Simon, and Mathias Drton. Graphical models for zero-inflated single cell gene expression. *The Annals of Applied Statistics*, 13(2):848–873, 2019.

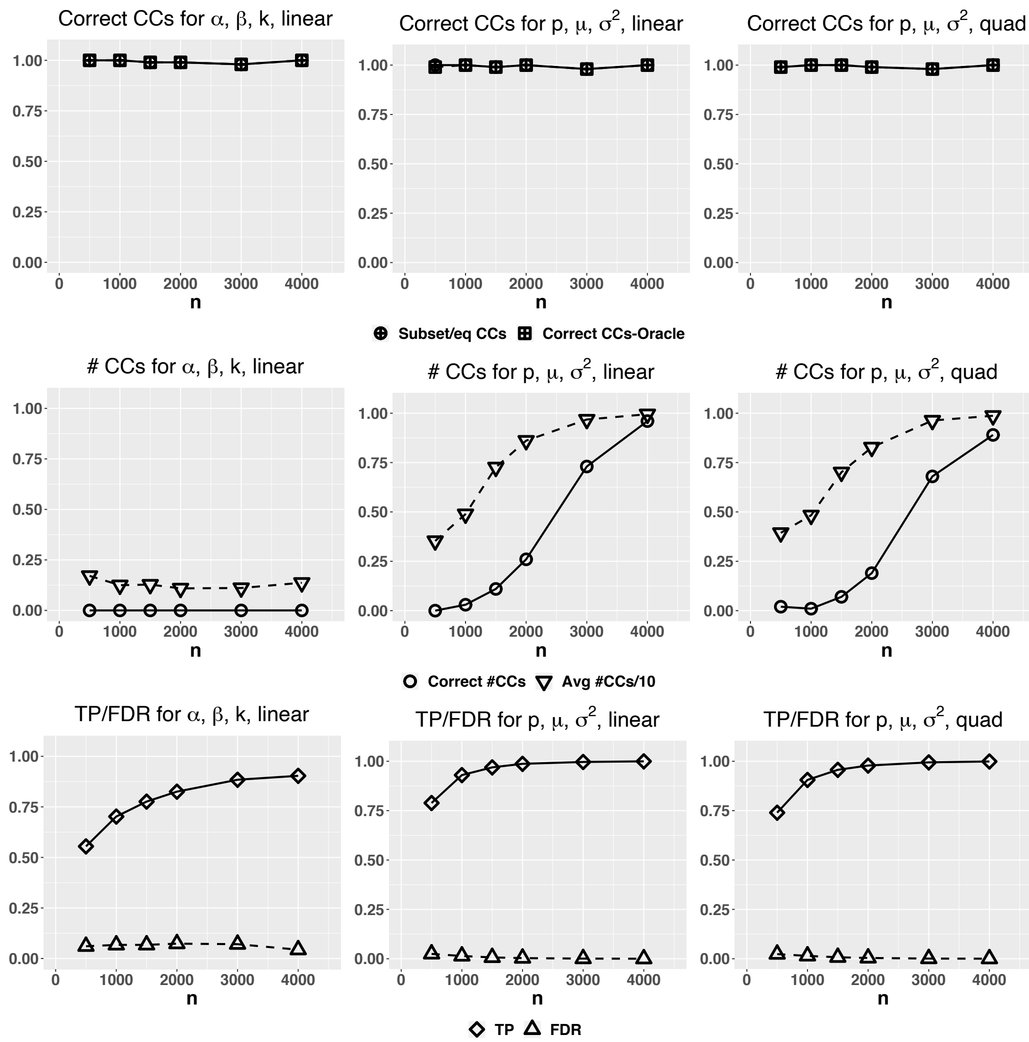


Figure S13: Results for the connected components (CC) estimated using the procedure in McDavid et al. (2019), compared against the truth (block-diagonal graph with 10 complete graphs each with $m = 10$). See discussion in Sections 4.2 and B.3.

Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1:763–765, 1973.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.