Directed Graphical Models and Causal Discovery for Zero-Inflated Data

Author names withheld

Editor:
Under Review for CLeaR 2023

Abstract

With advances in technology, gene expression measurements from single cells can be used to gain refined insights into regulatory relationships among genes. Directed graphical models are well-suited to explore such (cause-effect) relationships. However, statistical analyses of single cell data are complicated by the fact that the data often show zero-inflated expression patterns. To address this challenge, we propose directed graphical models that are based on Hurdle conditional distributions parametrized in terms of polynomials in parent variables and their $0/1$ indicators of being zero or nonzero. While directed graphs for Gaussian models are only identifiable up to an equivalence class in general, we show that, under a natural and weak assumption, the exact directed acyclic graph of our zero-inflated models can be identified. We propose methods for graph recovery, apply our model to real single-cell gene expression data on T helper cells, and show simulated experiments that validate the identifiability and graph estimation methods in practice.

Keywords: Bayesian network, causal discovery, directed acyclic graph, identifiability

1. Introduction

Graphical models specify conditional independence relations among variables in a random vector $\boldsymbol{Y}$ indexed by the nodes $\mathcal{V}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with edge set $\mathcal{E}$ (Maathuis et al. 2019). Models based on undirected graphs may be used to explore conditional independence between any two variables Y_V and Y_U given all others $(Y_W)_{W \neq U,V}$, as represented by the absence of an edge between V and U in $\mathcal{E}$. Models based on directed acyclic graphs (DAGs), for which $\mathcal{E}$ is comprised of directed edges, capture conditional independence structure that naturally arises from cause-effect relationships between the variables.

In biology and genetics, graphical models have been applied to infer the structure of gene regulatory networks based on measurements of gene expression (Maathuis et al. 2019, Sections 20-21). Traditional technologies produce expression levels aggregated over hundreds or thousands of individual cells, and these bulk measurements are frequently modeled using the assumption of Gaussianity. In directed Gaussian graphical models, the exact structure of the underlying DAG cannot be identified from purely observational data, and the target of inference becomes an equivalence class of DAGs. For instance, one cannot differentiate between $V \to U$ and $U \to V$ when the variables are assumed bivariate normal. In the Gaussian case, directed graphical models posit linear functional relationships between the variables coupled with additive Gaussian noise. A more recent line of work emphasizes that directed graphical models that alter this assumption to nonlinear functional relationships and additive noise (Peters et al. 2014), or linear relations and non-Gaussian noise (Shimizu et al. 2006; Wang and Drton 2020), or linear relations with homoscedastic Gaussian noise (Peters and Bühlmann 2013; Chen et al. 2019) are amenable to causal discovery in the sense that different DAGs are no longer equivalent.

© 2023 .

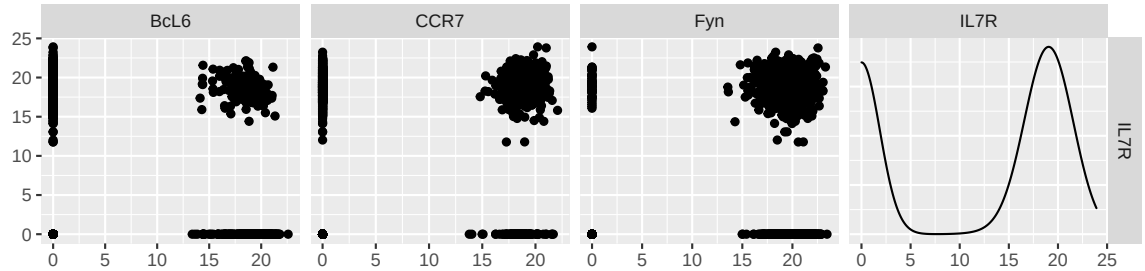

Figure 1: Kernel density of a selected gene (IL7R) and scatter plots of its relationship with three other genes (BcL6, CCR7, Fyn) from the T helper cell data analyzed in Section 6.

More recent technology obtains sequencing measurements of mRNA present in single cells. This new technology, as well as the larger sample sizes it provides, promise to give more information than bulk measurements, but at the same time bring in a unique new challenge. At the single cell level, genes appear as "on" with positive single cell gene expression levels, or as "off" with the recorded measurements zero or negligible (McDavid et al. 2019). This pattern is shown in Figure 1, which is based on a single-cell dataset with 1951 measurements from eight healthy donors, which we analyze in Section 6. The figure clearly shows the large number of zero values in each gene as well as the nonlinear relationships between genes.

In this paper, we propose two versions of directed graphical models for zero-inflated data, and prove that under a weak assumption the exact DAG can be recovered from the joint distribution. Our new graphical models build on the recently-proposed Hurdle graphical model of McDavid et al. (2019), but facilitate estimation of DAGs from observational single-cell sequencing data. In contrast to McDavid et al. (2019), our models are also not limited to zero-inflated Gaussian distributions, as we allow variables that are "on" to be non-linear polynomial functions of other variables and stochastic noise. After presenting two directed graphical models for zero-inflated data in Section 2, in Section 3, we show that under our models, the distributions that can be represented by two different DAGs must be distributions of *two-Gaussian type* (Definition 7). We then prove that such distributions do not exist for dimension $m = 2$ and $m = 3$; we also conjecture they do not exist for $m > 3$. Moreover, we are able to prove that under a natural and practical assumption, we have full identifiability in the sense of being able to identify the exact DAG underlying the model. In Section 4, we introduce different methods for estimation of the DAG. Simulation studies supporting the use of these methods are given in Section 5, and they are then applied to the T-follicular helper cell dataset (Section 6). Throughout the paper, we use subscripts to refer to entries in vectors and columns in matrices. When used as a subscript of a vector, a set of nodes/indices selects the corresponding entries from the vector, e.g., $\boldsymbol{y}_{\mathcal{V}} = (y_V)_{V \in \mathcal{V}}$.

2. Directed Graphical Models for Zero-Inflated Data

2.1. Hurdle Joint Distributions for Zero-Inflated Continuous Observations

We start by reviewing the undirected Hurdle graphical model. McDavid et al. (2019) proposed a *Hurdle joint distribution* with density

$$f(\boldsymbol{y}; \mathbf{A}, \mathbf{B}, \mathbf{K}) \propto \exp\left(\mathbb{1}_{\boldsymbol{y}}^{\top} \mathbf{A} \mathbb{1}_{\boldsymbol{y}} + \mathbb{1}_{\boldsymbol{y}}^{\top} \mathbf{B} \boldsymbol{y} - \frac{1}{2} \boldsymbol{y}^{\top} \mathbf{K} \boldsymbol{y}\right), \qquad \boldsymbol{y} \in \mathbb{R}^m, \tag{1}$$

with respect to λ^m, where λ is the sum of a point mass at 0 and the Lebesgue measure on $\mathbb{R}$, and $\mathbf{A} = (\alpha_{ij})_{i,j}, \mathbf{B} = (\beta_{ij})_{i,j}, \mathbf{K} = (k_{ij})_{i,j} \in \mathbb{R}^{m\times m}$ are matrices of interaction parameters with $\mathbf{K}$ positive definite. The indicator vector $\mathbb{1}_{\boldsymbol{y}} \equiv (\mathbb{1}_{\{y_1\neq 0\}}, \ldots, \mathbb{1}_{\{y_m\neq 0\}}) \in \{0,1\}^m$ captures which components of $\boldsymbol{y}$ are non-zero.

Suppose $\boldsymbol{Y} \in \mathbb{R}^m$ follows the Hurdle joint distribution. Intuitively, the density in (1) is obtained by combining an Ising model for the indicator vector $\mathbb{1}_{\boldsymbol{Y}}$ and a conditional normal distribution for $\boldsymbol{Y}$ given its nonzero pattern $\mathbb{1}_{\boldsymbol{Y}}$. The Ising model postulates a probability mass function proportional to $\exp\left(\mathbb{1}_{\boldsymbol{y}}^\top \mathbf{A}\mathbb{1}_{\boldsymbol{y}}\right)$. The conditional normal distribution has density $p\left(\boldsymbol{Y} = \boldsymbol{y} | \mathbb{1}_{\boldsymbol{Y}} = \mathbb{1}_{\boldsymbol{y}}; \mathbf{B}, \mathbf{K}\right) \propto \exp\left(\mathbb{1}_{\boldsymbol{y}}^\top \mathbf{B}\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^\top \mathbf{K}\boldsymbol{y}\right)$ with respect to the Lebesgue measure restricted to the subspace of $\mathbb{R}^m$ compatible with $\mathbb{1}_{\boldsymbol{y}}$. The exponential specification in (1) entails that conditional independence between two variables is equivalent to the corresponding entries in all interaction matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{K}$ being 0. In other words, $\alpha_{ij} = \alpha_{ji} = \beta_{ij} = \beta_{ji} = k_{ij} = k_{ji} = 0$ if and only if Y_i and Y_j are conditionally independent given all other variables. Indeed, it is easy to see that the induced conditional distribution of Y_i given all other variables $\boldsymbol{Y}_{-i}$ in $\boldsymbol{Y}$, has density

$$p(Y_i = y_i | \boldsymbol{Y}_{-i} = \boldsymbol{y}_{-i}) = f(y_i; \alpha_{ii} + \boldsymbol{\alpha}_{i,-i}^\top \mathbb{1}_{\boldsymbol{y}_{-i}} + \boldsymbol{\beta}_{i,-i}^\top \boldsymbol{y}_{-i}, \beta_{ii} + \boldsymbol{\beta}_{-i,i}^\top \mathbb{1}_{\boldsymbol{y}_{-i}} - \boldsymbol{k}_{i,-i}^\top \boldsymbol{y}_{-i}, k_{ii}), \quad (2)$$

that is, the distribution is a Hurdle distribution in $m = 1$ dimension with parameters α, β, and k being linear functions in $\boldsymbol{Y}_{-i}$ and $\mathbb{1}_{\boldsymbol{Y}_{-i}}$; here f is the univariate version of (1).

2.2. Hurdle Conditionals

The observation in (2) above gives rise to the following definition. Recall that λ is the sum of a point mass at 0 and the Lebesgue measure on $\mathbb{R}$.

Definition 1 ((α, β, k)-Hurdle conditionals) *Given a scalar random variable X and an m-dimensional random vector $\boldsymbol{Z}$, we say that the conditional distribution of X given $\boldsymbol{Z}$ is of* (α, β, k)-Hurdle type *if it admits conditional densities with respect to λ of the form*

$$p(X = x | \boldsymbol{Z} = \boldsymbol{z}) = f^{(m)}_{\alpha,\beta,k}(X|\boldsymbol{Z}) \equiv \frac{\exp\left(\alpha(\boldsymbol{z})\mathbb{1}_x + \beta(\boldsymbol{z})x - kx^2/2\right)}{\sqrt{2\pi/k}\exp\left(\alpha(\boldsymbol{z}) + \beta^2(\boldsymbol{z})/(2k)\right) + 1}. \quad (3)$$

Here, α and β are functions of $\boldsymbol{Z}$ (and its indicator vector).

Reparametrizing, we give another intuitive formulation of Hurdle conditionals that clearly exhibits their nature of a mixture between a point mass at 0 and a conditional Gaussian distribution.

Definition 2 ((p, μ, σ^2)-Hurdle conditionals) *Given a scalar random variable X and an m-dimensional random vector $\boldsymbol{Z}$, we say that the conditional distribution of X given $\boldsymbol{Z}$ is of* (p, μ, σ^2)-Hurdle type *if it admits conditional densities with respect to λ of the form*

$$p(X = x | \boldsymbol{Z} = \boldsymbol{z}) = f^{(m)}_{p,\mu,\sigma^2}(X|\boldsymbol{Z}) \equiv (1 - p(\boldsymbol{z}))(1 - \mathbb{1}_x) + p(\boldsymbol{z})\mathbb{1}_x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu(\boldsymbol{z}))^2}{2\sigma^2}\right). \quad (4)$$

Here, p and μ are functions of $\boldsymbol{Z}$ (and its indicator vector).

It is easy to show that the two parametrizations (3) and (4) are connected through

$$\log\frac{p}{1-p} = \alpha + \frac{\beta^2}{2k} - \frac{1}{2}\log\left(\frac{k}{2\pi}\right), \qquad \mu = \frac{\beta}{k}, \qquad \sigma^2 = \frac{1}{k}. \quad (5)$$

That is, the conditional log odds of being nonzero is linear in α and quadratic in β, and the conditional Gaussian mean is proportional to β. While the (α, β, k)-parametrization takes canonical parameters $\alpha(\boldsymbol{Z})$, $\beta(\boldsymbol{Z})$ and k using a representation as exponential family, the moment parametrization directly models the conditional mixing probability $p(\boldsymbol{Z})$, and the mean $\mu(\boldsymbol{Z})$ and variance σ^2 parameters of the conditional Gaussian distribution. We thus refer to (3) as the *canonical parametrization*, and (4) as the *moment parametrization*.

2.3. Directed Graphical Models for Zero-Inflation Data

Consider an m-dimensional random vector $\boldsymbol{Y}$ whose components are indexed by the vertices of a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and whose distribution is dominated by a product measure on $\mathbb{R}^m$. A graphical model based on $\mathcal{G}$ requires that the density of the joint distribution admits a factorization as

$$f(\boldsymbol{y}) = \prod_{V \in \mathcal{V}} f_V \left(y_V | \boldsymbol{y}_{\mathrm{pa}(V)} \right), \tag{6}$$

where each factor $f_V \left(y_V | \boldsymbol{y}_{\mathrm{pa}(V)} \right)$ is a conditional density for y_V given its parent variables $\boldsymbol{y}_{\mathrm{pa}(V)}$. The set of parents is defined to be $\mathrm{pa}(V) \equiv \{U : U \to V \in \mathcal{E}\}$.

In Section 2.1, we observed that, for the Hurdle joint distributions (1), the conditional distribution of any Y_i given the others is an (α, β, k)-Hurdle with k constant, and α and β linear functions of those variables (and their indicators) that are conditionally dependent on Y_i; see (2). Motivated by this fact, we specify directed graphical models for zero-inflated data by assuming the conditional densities in the factorization in (6) to be (α, β, k)- or (p, μ, σ^2)-Hurdle conditionals. We then assume the parameters in these conditionals to be *Hurdle polynomials* in its parents, as defined now.

Definition 3 (Hurdle polynomials) *Let $\boldsymbol{Y} = (Y_V)_{V \in \mathcal{V}} \in \mathbb{R}^m$ be a random vector indexed by a set $\mathcal{V}$, and suppose $\mathcal{S} \subseteq \mathcal{V}$. If $\mathcal{S} \neq \varnothing$, define the space of* Hurdle polynomials *in $\boldsymbol{y}_{\mathcal{S}}$ as*

$$\mathcal{H}(\boldsymbol{Y}; \mathcal{S}) \equiv \Bigg\{ c_0 + \sum_{j=1}^{T} c_j \prod_{U \in \mathcal{U}_j} Y_U^{d_{j,U}} \prod_{V \in \mathcal{V}_j} \mathbb{1}_{Y_V}, \quad c_0 \in \mathbb{R},\, T \in \mathbb{N}, \\ c_j \neq 0,\, \mathcal{U}_j \subseteq \mathcal{S},\, \mathcal{V}_j \subseteq \mathcal{S} \backslash \mathcal{U}_j,\, d_{j,U} \in \mathbb{N} \quad \forall U \in \mathcal{U}_j \quad \forall j = 1, \ldots, T \Bigg\}, \tag{7}$$

where $\mathbb{N} = \{1, 2, \ldots\}$. This is the set of polynomials in values and indicators of nodes in $\mathcal{S}$. If $\mathcal{S} = \varnothing$, define $\mathcal{H}(\boldsymbol{Y}; \mathcal{S}) \equiv \mathbb{R}$. The degree *of a Hurdle polynomial as specified in (7) is $\max_{j=1,\ldots,T} \sum_{U \in \mathcal{U}_j} d_{j,U} + |\mathcal{V}_j|$. Here $|\cdot|$ denotes the set cardinality.*

In the definition (7), for the j-th term in the polynomial, c_j is its polynomial coefficient, $\mathcal{V}_j$ is the set of nodes that define the term only through their indicators, while $\mathcal{U}_j$ is the set of those whose values directly define the term, with $\{d_{j,U}\}_{U \in \mathcal{U}_j}$ the corresponding exponents.

We are now ready to formally define our models.

Definition 4 (DAG models for zero-inflated data) *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG with $|\mathcal{V}| = m$ nodes. A zero-inflated conditional Gaussian DAG model associated with $\mathcal{G}$ is a set of joint distributions on $\mathbb{R}^m$ that admit a density (with respect to λ^m) that factors as in (6) with each conditional density $f_V \left(y_V | \boldsymbol{y}_{\mathrm{pa}(V)} \right)$ being a Hurdle conditional*

(1) in the (α, β, k)-parametrization with parameters α_V, β_V and k_V, where k_V is constant, α_V and β_V are Hurdle polynomials in $\boldsymbol{y}_{\mathrm{pa}(V)}$; or

(2) in the (p, μ, σ^2)-parametrization with parameters p_V, μ_V and σ_V^2, where σ_V^2 is constant, $\log(p_V/(1-p_V))$ and μ_V are Hurdle polynomials in $\boldsymbol{y}_{\mathrm{pa}(V)}$.

It is clear from (5) that if we let the relevant parameters to be Hurdle polynomials of *any* degree, the two parametrizations are equivalent, meaning that given an underlying DAG, they share the same space of all possible joint distributions. However for computational convenience it is useful to bound the degree. In later applications, we will only consider degrees up to three.

3. Identifiability

3.1. Strong Identifiability

As we show next, the directed graphical models from Definition 4 are amenable to causal discovery in the sense that the DAG underlying the model is uniquely identifiable from a given joint distribution. More precisely, we prove identifiability under a mild assumption on the Hurdle conditionals. Let $\pi(\boldsymbol{y}_{\mathcal{S}}) \in \mathcal{H}(\boldsymbol{Y}; \mathcal{S})$ be a Hurdle polynomial for a subset $\mathcal{S} \subseteq \mathcal{V}$. For $U \in \mathcal{S}$, let $\pi_U(y_U) \equiv \pi(y_U, \mathbf{0})$ be the restriction of $\pi(\boldsymbol{y}_{\mathcal{S}})$ obtained by setting all entries other than y_U to zero. Then $\pi_U(y_U) \in \mathcal{H}(\boldsymbol{Y}; \{U\})$ is a univariate Hurdle polynomial.

Definition 5 (Strong Hurdle polynomials) *Let $\pi(\boldsymbol{y}_{\mathcal{S}}) \in \mathcal{H}(\boldsymbol{Y}; \mathcal{S})$. We say $\pi(\boldsymbol{y}_{\mathcal{S}})$ is a* strong Hurdle polynomial *if all of its restrictions $\pi_U(y_U)$ take at least three different values. In other words, for each $U \in \mathcal{S}$, the Hurdle polynomial $\pi(\boldsymbol{y}_{\mathcal{S}})$ contains at least one term of the form $c_j y_U^d$ with $c_j \neq 0$ and $d \geq 1$.*

Theorem 6 (DAG identifiability with strong Hurdle polynomials)
Let $f(\boldsymbol{y})$ be a joint density with respect to λ^m that factors according to a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, as in (6). Suppose for each $V \in \mathcal{V}$, the conditional $f_V\left(y_V | \boldsymbol{y}_{\mathrm{pa}(V)}\right)$ is of Hurdle type with parameters (α_V, β_V, k_V) or (p_V, μ_V, σ_V^2). If for each V, $\alpha_V + \beta_V^2/(2k_V)$, or equivalently $\log(p_V/(1-p_V))$, is a strong Hurdle polynomial, then $f(\boldsymbol{y})$ does not factor with respect to any other DAG $\mathcal{G}' \neq \mathcal{G}$.

In the proof in Appendix A, we show that the restriction on the parameters of the Hurdle conditionals is actually stronger than what we need for identifiability. However, the assumption of *strong* Hurdle polynomials is very natural in that it specifies a weak form of hierarchy among interactions by requiring that the conditional distributions are parametrized to include at least one univariate power term in every parent variable and not just indicators or interaction terms with other parents.

3.2. Weak Identifiability

Without assuming the Hurdle polynomials for the conditional distributions to be *strong*, we can still offer a weaker identifiability result that shows that the distributions in the intersection between the models obtained from two Markov equivalent DAGs with Hurdle polynomial parameters always have to be of what we call *two-Gaussian type*. In our definition of this concept, we write $\phi(\,\cdot\,; \mu, \nu)$ for the univariate normal density function with mean μ and inverse variance ν.

Definition 7 *Let $\boldsymbol{Y} = (Y_V)_{V\in\mathcal{V}}$ be a random vector, and let $W, U \in \mathcal{V}$ be the indices for two of its components. Further, let $\mathcal{P} \subseteq \mathcal{V}\backslash\{W, U\}$ be a set of additional indices. Then the joint distribution of $\boldsymbol{Y}$ is of* two-Gaussian type w.r.t. $(W, U, \mathcal{P})$ *if the following holds for both $V = W$ and $V = U$: There exist a constant ν_1^V, polynomials $\mu_1^V(\boldsymbol{y}_\mathcal{P})$, $\mu_2^V(\boldsymbol{y}_\mathcal{P})$, $\nu_2^V(\boldsymbol{y}_\mathcal{P})$, and functions $c_1^V(\boldsymbol{y}_\mathcal{P})$ and $c_2^V(\boldsymbol{y}_\mathcal{P})$ such that for almost every $\boldsymbol{y}_\mathcal{P} \in \mathbb{R}^{|\mathcal{P}|}$, $c_1^V(\boldsymbol{y}_\mathcal{P}) > 0$, $c_2^V(\boldsymbol{y}_\mathcal{P}) > 0$, either (a) $\mu_1^V(\boldsymbol{y}_\mathcal{P}) \neq \mu_2^V(\boldsymbol{y}_\mathcal{P})$; or (b) $\nu_1^V \neq \nu_2^V(\boldsymbol{y}_\mathcal{P})$ and the conditional density*

$$\mathbb{P}\left(Y_V = y \,|Y_V \neq 0, \boldsymbol{Y}_\mathcal{P} = \boldsymbol{y}_\mathcal{P}\right) = c_1^V(\boldsymbol{y}_\mathcal{P})\phi\left(y; \mu_1^V(\boldsymbol{y}_\mathcal{P}), \nu_1^V\right) + c_2^V(\boldsymbol{y}_\mathcal{P})\phi\left(y; \mu_2^V(\boldsymbol{y}_\mathcal{P}), \nu_2^V(\boldsymbol{y}_\mathcal{P})\right)$$

is a mixture of exactly two distinct Gaussian distributions with means polynomial in $\boldsymbol{y}_\mathcal{P}$; the inverse variance parameter is an absolute constant for one of these distributions and polynomial in $\boldsymbol{y}_\mathcal{P}$ for the other. If $\mathcal{P} = \varnothing$, then two-Gaussian type w.r.t. $(W, U, \varnothing)$ requires that both $\mathbb{P}(Y_W|Y_W \neq 0)$ and $\mathbb{P}(Y_U|Y_U \neq 0)$ are mixtures of exactly two distinct univariate Gaussian distributions with constant parameters, respectively.

We next recall an observation from Proposition 29(ii) in Peters et al. (2014); see Sections 1.8 and 15.3.2 of Maathuis et al. (2019) for definitions of the Markov property and faithfulness.

Proposition 8 *Suppose the distribution of $\boldsymbol{Y}$ is Markov and faithful with respect to two distinct Markov equivalent graphs $\mathcal{G}$ and $\mathcal{G}'$. Then, there must exist nodes W and U such that $W \to U$ in $\mathcal{G}$ and $U \to W$ in $\mathcal{G}'$, while $\mathcal{P} \equiv \mathrm{pa}_\mathcal{G}(U)\backslash\{W\} = \mathrm{pa}_{\mathcal{G}'}(W)\backslash\{U\}$.*

Remark 9 *Proposition 8 is at the heart of many proofs of DAG identifiability, which combine it with suitable probabilistic conditioning to reduce the comparison of two DAG models to bivariate problems involving the two graphs $W \to U$ and $W \leftarrow U$. However, in our setting, a key new challenge arises because the form of the Hurdle conditionals precludes us from applying conditioning to form sets of bivariate distributions that are of the considered Hurdle type. Indeed, conditioning on descendants of the considered variables (i.e., other variables that in the graph can be reached along directed paths) generally gives conditional distributions that are no longer of the Hurdle type used in the definition of our model class.*

We claim that the intersection of sets of joint distributions represented by two distinct Markov equivalent $\mathcal{G}$ and $\mathcal{G}'$ must be a subset of 2-Gaussian type distributions with respect to a triplet $(W, U, \mathcal{P})$ obtained from Proposition 8.

Theorem 10 (General Identifiability) *Let $\boldsymbol{Y}$, $\mathcal{G}$, $\mathcal{G}'$, W, U, $\mathcal{P}$ be as in Proposition 8. Let $\boldsymbol{Y}$ have a λ^m-density that factors w.r.t. both graphs $\mathcal{G}$ and $\mathcal{G}'$. For each $\mathcal{H} = \mathcal{G}, \mathcal{G}'$, let the node conditionals in the factorization be Hurdle conditionals with the parameters $(\alpha_V^\mathcal{H})_{V\in\mathcal{V}}$ and $(\beta_V^\mathcal{H})_{V\in\mathcal{V}}$ from (3), or equivalently $(p_V^\mathcal{H})_{V\in\mathcal{V}}$ and $(\mu_V^\mathcal{H})_{V\in\mathcal{V}}$ from (4), that are Hurdle polynomials of the form (7), where for $(V, T, \mathcal{H}) = (U, W, \mathcal{G})$ and $(V, T, \mathcal{H}) = (W, U, \mathcal{G}')$ it holds that*

(i) $\beta_V^\mathcal{H}(y_T, \boldsymbol{y}_\mathcal{P})$ (or $\mu_V^\mathcal{H}(y_T, \boldsymbol{y}_\mathcal{P})$) depends on at least one of $\mathbb{1}_{y_T}$ and y_T, or

(ii) $\alpha_V^\mathcal{H}(y_T, \boldsymbol{y}_\mathcal{P})$ (or $p_V^\mathcal{H}(y_T, \boldsymbol{y}_\mathcal{P})$) depends on the value of y_T (and maybe additionally on $\mathbb{1}_{y_T}$).

Then the distribution of $\boldsymbol{Y}$ must be of two-Gaussian type w.r.t. $(W, U, \mathcal{P})$. In this case we also say the distribution is of two-Gaussian type w.r.t. $\mathcal{G}$ and $\mathcal{G}'$.

Note that the assumption of faithfulness in Proposition 8 implies that we have (i) or (ii) or a condition (iii) that states that $\alpha_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$ (or $p_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$) depends on $\mathbb{1}_{y_T}$ only and $\beta_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$ (or $\mu_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$) is constant in y_T. It is case (iii) that we rule out in our assumption of Theorem 10.

The result is proved in Appendix A. It is easy to show that the result also holds if we make modifications such as restricting the maximum degree of the polynomial or excluding interactions between the discrete and continuous components. In the two- and three-dimensional cases (i.e., $m = 2, 3$) we show in Appendix A that there does not exist a joint distribution for $\boldsymbol{Y}$ that is of two-Gaussian type with respect to two distinct Markov equivalent graphs. We thus have the following result on full identifiability for graphs with two or three nodes.

Corollary 11 (Identifiability in two and three dimensions) *If $|\mathcal{V}| \leq 3$, i.e., in a binary/triary setting, there does not exist a joint distribution that is of two-Gaussian type w.r.t. two distinct Markov equivalent DAGs $\mathcal{G}$ and $\mathcal{G}'$. Thus, strong identifiability is guaranteed as in Theorem 6, meaning that the sets of Markov and faithful distributions associated to $\mathcal{G}$ and $\mathcal{G}'$ must be disjoint.*

Theorem 6 and Corollary 11 state that the DAGs are perfectly identifiable if $m = 2, 3$ or if we assume the Hurdle polynomials to be *strong*; Theorem 10 claims that without assuming *strong* Hurdle polynomials, the distributions for $m > 3$ from which the graph is not identifiable must be a subset of the *two-Gaussian type* distributions. We conjecture that in general, with $m > 3$, the set of two-Gaussian type distributions with respect to any two graphs is an empty set. In Appendix B we show scatter plots of simulated data that give some indication of how Markov equivalent graphs may be differentiated under our models.

4. Estimation of DAGs from Zero-Inflated Data

Suppose now that we are given an i.i.d. sample $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}$ comprised of m-variate observations. The log-likelihood function ℓ of any DAG model can be decomposed into the sum of conditional (or nodewise) log-likelihood functions ℓ^V for the V-th variable conditional on its parent variables. Let $y_V^{(1)}, \ldots, y_V^{(n)}$ be the n observations of the V-th variable. For the canonical (α, β, k)-parametrization from (3), the nodewise log-likelihood function is

$$\begin{aligned}\ell^V\left(\alpha_V, \beta_V, k_V \,\middle|\, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}\right) = \sum_{i=1}^{n} \Bigg(& \alpha_V\left(\boldsymbol{y}_{\mathrm{pa}(V)}^{(i)}\right) \mathbb{1}_{y_V^{(i)}} + \beta_V\left(\boldsymbol{y}_{\mathrm{pa}(V)}^{(i)}\right) y_V^{(i)} - k_V {y_V^{(i)}}^2/2 \\ & - \log\left[\sqrt{2\pi/k_V}\exp\left\{\alpha_v\left(\boldsymbol{y}_{\mathrm{pa}(V)}^{(i)}\right) + \beta_V^2\left(\boldsymbol{y}_{\mathrm{pa}(V)}^{(i)}\right)/(2k_V)\right\} + 1\right]\Bigg);\end{aligned}$$

for the moment (p, μ, σ^2)-parametrization from (4) it is

$$\begin{aligned}\ell^V\left(p_V, \mu_V, \sigma_V^2 \,\middle|\, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}\right) = & \sum_{i: y_V^{(i)} = 0} \log\left\{1 - p_V\left(\boldsymbol{y}_{\mathrm{pa}(V)}^{(i)}\right)\right\} \\ & + \sum_{i: y_V^{(i)} \neq 0} \left[\log p_V\left(\boldsymbol{y}_{\mathrm{pa}(V)}^{(i)}\right) - \frac{1}{2}\log(2\pi\sigma_V^2) - \left\{y_V^{(i)} - \mu_V\left(\boldsymbol{y}_{\mathrm{pa}(V)}^{(i)}\right)\right\}^2/(2\sigma_V^2)\right].\end{aligned}$$

In the latter case, we see the sum of the log-likelihood functions from the logistic regression model for p_V and the linear regression for μ_V restricted to the observations with $y_V^{(i)} \neq 0$. Here we recall

that the parameters $\alpha_V, \beta_V, p_V, \mu_V$ are themselves polynomials in $\boldsymbol{y}_{\mathrm{pa}(V)}$ and their indicators, and we are using them as a shorthand notation on the left-hand sides where we really mean ℓ^V as a function of the parameters (i.e., coefficients) in those polynomials.

4.1. Fitting Hurdle Conditionals

Estimation of the graphical models amounts to fitting the conditional distribution of one node given a set of others. For the canonical (α, β, k)-parametrization, the log-likelihood function is convex in α_V, β_V and k_V. Moreover, α_V and β_V are linear in the polynomial coefficients. Therefore, the log-likelihood is convex in the coefficients to estimate and can be maximized by standard methods; e.g., coordinate descent. Estimation for the moment (p, μ, σ^2)-parametrization (4), on the other hand, can be easily solved by separately fitting a logistic regression to p_V and a linear regression to μ_V. Recall again that the two parametrizations, canonical and moment, are equivalent when assuming a full polynomial model, i.e., when the degree and structure of the polynomials is unrestricted. However, when restricting, for instance, the degree the two parametrizations yield different models.

The (α, β, k)-parametrization with linear Hurdle polynomials (i.e., degree 1) naturally comes from conditional distributions of the joint distribution defined for undirected graphical models in McDavid et al. (2019). However, at least for higher degrees, the (p, μ, σ^2)-parametrization may be more intuitive and useful in practice as it leads to a decomposition into a logistic regression and a linear regression. This decomposition enables us to use optimized standard regression solvers for model fitting. The (p, μ, σ^2)-parametrization also makes it easy to apply available routines to incorporate regularization on the coefficients/parameters into our loss, which is helpful when the number of samples is small compared to the number of parameters. Such higher dimensionality of the models arises in particular when assuming a higher degree for the Hurdle polynomials. In our implementation, we use an ℓ_2 regularization and select its tuning parameter using the Bayesian information criterion (BIC). We also assume the highest degree of Hurdle polynomials and select the degree by optimizing BIC simultaneously over the degree and the ℓ_2 penalty, so the degree is separately optimized for each regression (combination of node and its candidate parent set).

4.2. Graph Search

To estimate the underlying DAG, we consider two state-of-the-art methods: (A) exhaustive score-based search and (B) greedy search. Both methods rely on a model score which we take to be the BIC defined as $\nu \log n - 2\ell$, where ν is the total number of parameters in the model, n is the sample size, and ℓ is the log-likelihood as introduced in Section 4.

Exhaustive search. Optimizing the BIC over the set of all DAGs is possible for moderately small m using the dynamic programming algorithm of Silander and Myllymäki (2006). This approach is justified by the asymptotic consistency of the BIC as well as the identifiability of our model (recall Section 3). The experiments of Silander and Myllymäki (2006) suggest that for Gaussian models the search is practical for $m < 32$. Estimation of our models is computationally more challenging but exhaustive search is feasible at least for $m < 16$.

Greedy search. Instead of optimizing BIC over all DAGs, we may apply a greedy search that iteratively improves BIC by moving to a neighboring DAG that provides the largest improvement. The neighborhood is defined using edge additions, deletions, and reversals; compare Chickering (2003). While Chickering (2003) discusses consistency of graph recovery in terms of equivalence classes,

in our case the algorithm determines individual graphs. For faster estimation in sparse settings, we consider restricting the maximum node in-degree (i.e., the maximum number of parents).

There are various approaches that may help accelerating the estimation process. As an example, one can use caching (Ramsey et al. 2017) and dynamic updating (Goudie and Mukherjee 2016) to save time on computing the likelihoods and checking acyclicity in the current estimated graph. To speed up the estimation, we cache the BICs of all the nodewise regressions that have been fit so far, which requires little memory overhead. As the greedy search may be stuck in a local minimum, the most obvious way to circumvent this is to run the greedy algorithm initialized with multiple random DAGs with the same number of nodes and different sparsity levels, and choose the output that has the lowest BIC. Alternatively, one can first estimate an undirected graph using the method of McDavid et al. (2019), and initialize the search with multiple directed graphs whose moral graph is the estimated undirected graph. Moreover, to scale to larger m, we can first use the procedure of McDavid et al. (2019) to identify the connected components of the estimated undirected graph and then estimate the directed edges in each connected component. This procedure is justified by the fact that the connected components for the underlying true undirected and directed graphs coincide.

5. Numerical Experiments

Our numerical studies in this section aim to verify identifiability and exact DAG recovery. Due to space limitation, we present the results for the exhaustive search in the main paper. Following the discussion in Section 4.2, we use our self-implemented greedy search (GDS) (Chickering 2003) with BIC score, as well as an exhaustive search with dynamic programming (Silander and Myllymäki 2006). The results for the greedy search—which facilitates estimation of much larger DAGs—show similar trends and are presented in Appendix B.

To illustrate the performance of exhaustive search, we consider three DAG structures: (i) chain graph with $m = 10$, (ii) complete graph with $m = 10$, and (iii) lattice graph with $m = 9$. For each structure, we consider true generating conditional distributions using the following parametrizations: a) (α, β, k)-(canonical) parametrization with *linear* Hurdle polynomials, b) (p, μ, σ^2)-(moment) parametrization with *linear* Hurdle polynomials, and c) (p, μ, σ^2)-(moment) parametrization with *quadratic* Hurdle polynomials. We note that the distributions represented by c) is a superset of those by a) and b). By (5), distributions represented by a) and b) are disjoint because $\log(p/(1-p))$ is a weighted sum of α and β^2.

Recall the definition of Hurdle conditionals in (3) and (4) in Section 2.2. In our experiments, whenever $\mathrm{pa}(V) = \varnothing$, we generate $y_V \sim f_0$ such that $f_0(x) = \frac{1}{2}(1 - \mathbb{1}_x) + \frac{1}{2}\phi(x; 0, 1)$, where ϕ is the standard normal density. Otherwise, for parametrization a), we use Hurdle conditionals with parameters $k_V = 1$, $\alpha_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \beta_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U \in \mathrm{pa}(V)} (\mathbb{1}_{y_U} - y_U)$; similarly for parametrization b) we take $\sigma_V^2 = 1$, $\log \frac{p_V}{1-p_V}(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U \in \mathrm{pa}(V)} (\mathbb{1}_{y_U} + y_U)$ and $\mu_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U \in \mathrm{pa}(V)} (\mathbb{1}_{y_U} - y_U)$; finally, for parametrization c) we take $\sigma_V^2 = 1$ and $\log \frac{p_V}{1-p_V}(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U \in \mathrm{pa}(V)} \left(\mathbb{1}_{y_U} + y_U + \frac{y_U^2}{10}\right) + \frac{1}{10} \sum_{\substack{U,W \in \mathrm{pa}(V) \\ U \neq W}} (\mathbb{1}_{y_U} + y_U)(\mathbb{1}_{y_V} + y_V)$, and $\mu_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U \in \mathrm{pa}(V)} \left(\mathbb{1}_{y_U} - y_U - \frac{y_U^2}{10}\right) + \frac{1}{10} \sum_{\substack{U,W \in \mathrm{pa}(V) \\ U \neq W}} (\mathbb{1}_{y_U}\mathbb{1}_{y_V} - y_U\mathbb{1}_{y_V} - y_V\mathbb{1}_{y_U} - y_V y_U)$. We then normalize the coefficients in the above expressions ($\pm 1, \pm 1/10$) such that α_V, β_V, $\log p_V/(1-p_V)$ and μ_V have means 0 and 1, respectively, across the samples. This normalization ensures that the marginal probability of being nonzero, the marginal mean, and the marginal variance for each node

are stabilized, in order to show that the DAGs are truly recovered based on the conditional dependency structure instead of additional signals from these marginal quantities. In fact, in the generated samples the marginal probability is about 0.5 and the marginal mean is about 0 for all nodes, and the marginal variance for the nonzero part only is about the same for all except the source node (see Figure S2 in the Appendix for some scatter plots of the data generated). To assess the effect of misspecified parametrizations, for each combination of true DAG and true data generating parametrization—(α, β, k)-linear and (p, μ, σ^2)-linear and quadratic—we estimate the DAG using all three parametrizations for generating data.

The results are shown in Figures 2. Due to space limitation, only results for correctly specified models are presented in the main paper, and the expanded results with misspecified models are given in the Appendix. Each row of the figure corresponds to one of the three graphs (chain, complete, lattice) and each column corresponds to results using one estimating parametrization. The plots show the average true positive rate (TPR) and false discovery rate (FDR) over $B = 100$ iterations, defined as $\mathrm{TPR} = |\hat{S} \cap S_0|/|S_0|$ and $\mathrm{FDR} = |\hat{S} \backslash S_0|/\max\{|\hat{S}|, 1\}$, where $\hat{S}$ denotes the estimated set of (directed) edges, and S_0 the set of true edges.

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

6. T Helper Cell Data

In this section we present the results of applying our model to a T helper cell expression dataset. Specifically, the dataset is considered in McDavid et al. (2019) and contains both single cell and 10-cell expression measurements for T helper cells for 80 genes in eight healthy donors. We use all 1951 single cell measurements for these donors (a superset of the 465 measurements in McDavid et al. (2019)) to ensure we have a large enough sample size to produce reliable estimates. In particular, McDavid et al. (2019) consider only the T-follicular ($\mathrm{CXCR5^{+}PD1^{+}}$) cells that produce high levels of proteins CXCR5 and PD1, while we do not make this restriction. Instead, we add the indicators of $\mathrm{CXCR5^{+/-}}$ and $\mathrm{PD1^{+/-}}$ as regressors when fitting the conditional distributions. Following McDavid et al. (2019), we choose the 61 genes that have at least 5% zero and 5% nonzero values.

While the measurements are all nonnegative, the minimum, mean, and standard deviation of the nonzero values in the dataset are 7.89, 18.53, and 1.91, respectively. We thus assume zero-inflated conditional Gaussianity without considering the effect of truncation from below at 0. Following Section 4.2 we first estimate the connected componenets using the method from McDavid et al. (2019) for undirected graphs, and proceed with estimation of DAGs for each component. We use the (p, μ, σ^2)-parametrization as it is more flexible than the (α, β, k), and extra fixed covariates and controlling factors can be easily added, since fitting the conditionals only involves linear and logistic regressions. As discussed in Section 5, the (p, μ, σ^2) is also more robust than (α, β, k). We use polynomials up to degree three and data-adaptively choose the optimal degree by BIC.

To estimate the DAG, we use the greedy search (GDS) algorithm, which shows promising performance in the simulations in Appendix B. We also use the stability selection procedure of Shah and Samworth (2013) to control the FDR at 10% for each connected component. For smaller connected components, if controlling the FDR at 10% is not possible, we pick the sparsest graph that maximally maintains the connectivity. Finally, we restrict the node in-degrees to five, in order to

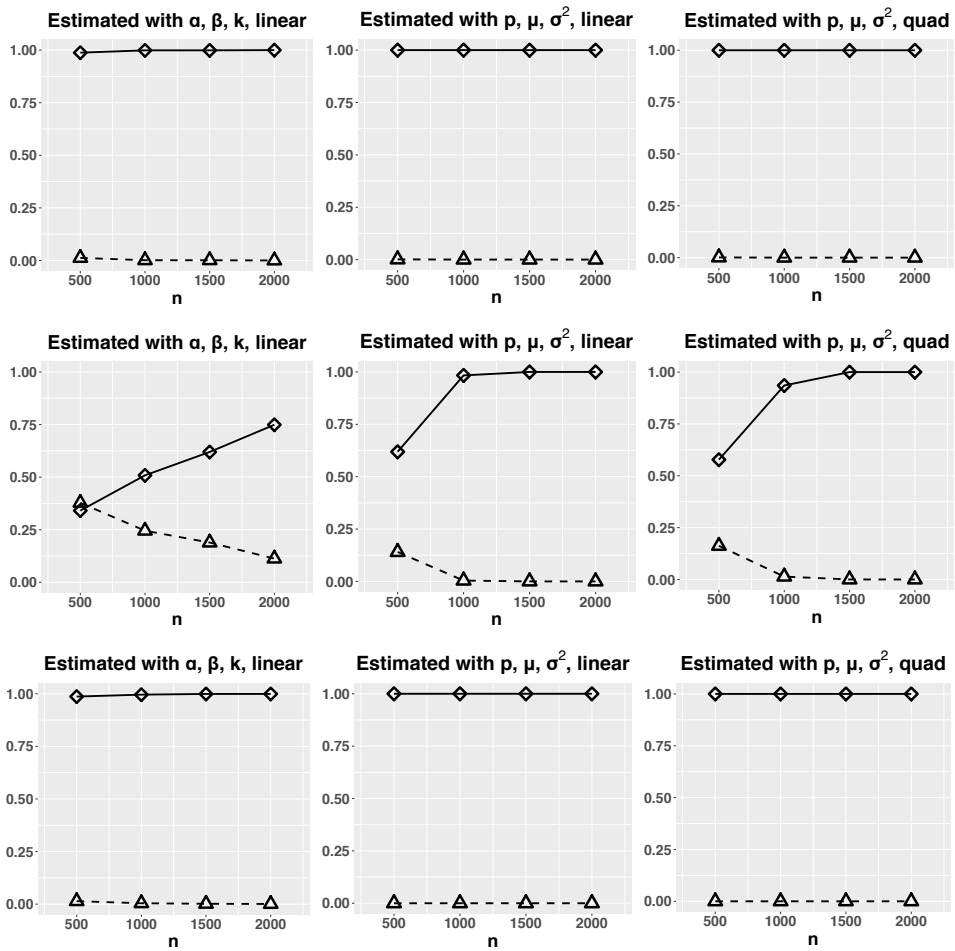

Figure 2: Simulation results for exhaustive search. Each row corresponds to a different graph (chain, complete, lattice). In all cases, estimation methods match true parametrizations. '◊' with solid lines: true positive rate; '∆' with dashed lines: false discovery rate.

both speed up estimation and to constrain the search space. This constraint is motivated by the fact that in gene regulatory networks, each gene is only expected to be regulated by a small number of other genes (Albert 2005). In contrast, since genetic networks often involve hub genes that regulate many others, we do not restrict the out-degree.

Figure 3 shows the estimated directed network along with the undirected network obtained using the method of McDavid et al. (2019). Overall, the estimated DAG structure is very similar to the undirected graph, with few differences including isolated nodes that only have a single edge with weak association in the undirected network.

7. Discussion

Motivated by the recent advent of single-cell sequencing technologies, we propose new methods for learning DAGs from zero-inflated data. Our procedures take advantage of two key features of single-cell transcrimtomics data, namely, the zero-inflation, and the large number of observations

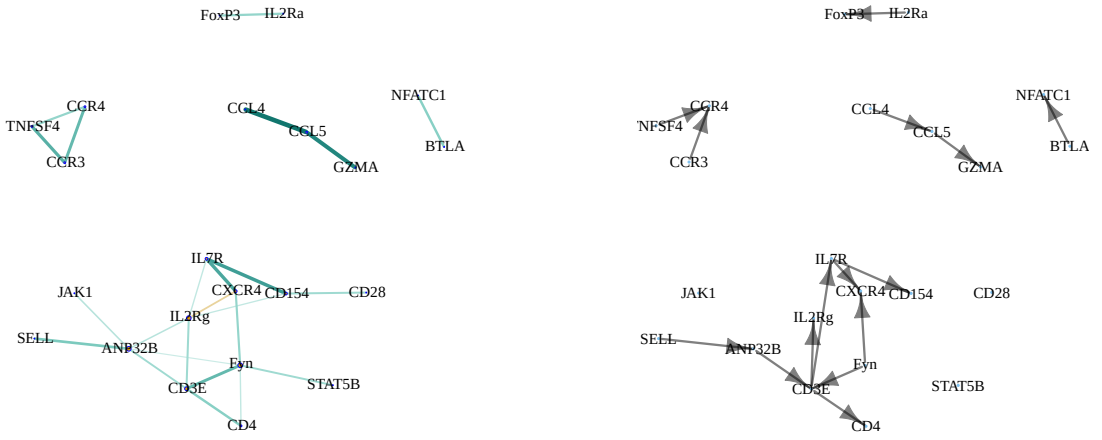

Figure 3: Estimated graph for T helper single cell data. Left: Undirected graph using the method of McDavid et al. (2019), with edge width and saturation representing the edge strength. Right: Directed graph using our method with stability selection Shah and Samworth (2013) to control FDR.

from individual samples. Our key contribution is establishing identifiability of DAGs from observational zero-inflated data. Specifically, we prove that the exact DAG can be recovered from the joint distribution under reasonable assumptions. We also show that in the most general case, the distributions from which the DAGs are not identifiable only form a small subset, which we prove to be empty in the bivariate and trivariate cases. While our proof uses a very general result on DAGs from Peters et al. (2014) as its first step, our models do not fit into the framework in that paper; we thus take a different approach that considers the zero-inflation and polynomial structures directly.

Our approach is based on factorizing the joint distribution into zero-inflated conditional Gaussian distributions with parameters polynomial in the parents and their indicators of having nonzero values. We present models in terms of two parametrizations, one called (α, β, k) that is linked to the undirected graphs studied in McDavid et al. (2019), and the other called (p, μ, σ^2) that directly models the conditional moments. Both approaches have computational appeal. In particular, the (α, β, k)-parametrization leads to convex loss functions in the parameters to be estimated, while the (p, μ, σ^2)-parametrization offers the additional benefit of allowing one to utilize standard software for logistic and linear regression. We combine these models with two state-of-the-art estimation procedures, namely greedy DAG search (GDS) and exhaustive search with dynamic programming. We also validate our identifiability theory using extensive numerical studies. These experiments indicate that the exhaustive search algorithm is effective in correctly identifying DAGs with small number of nodes. For moderate to large DAGs, the GDS algorithm offers a reasonable alternative, with performance comparable to the exhaustive search when the sample size is large enough.

Several extension of our work would be of interest. The first is to prove our conjecture that the sets of distributions from which the DAG is not identifiable are also empty for graphs with more than 3 nodes. The second is proving the consistency and investigating finite sample properties of the proposed estimation procedures. Finally, it would be interesting to extend our model to zero-inflated distributions under a truncation to the nonnegative orthant $\mathbb{R}^m_+$, which would be of interest for nonnegative *omics* data by generalizing the *score matching* loss (Hyvärinen 2005, 2007; Lyu 2009; Yu et al. 2019) to data of mixed type.

References

Reka Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957, 2005.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019. ISSN 0006-3444.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research (JMLR)*, 3(3):507–554, 2003.

Robert JB Goudie and Sach Mukherjee. A gibbs sampler for learning dags. *The Journal of Machine Learning Research*, 17(1):1032–1070, 2016.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2009.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019. ISBN 978-1-4987-8862-5.

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.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2013.

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.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129, 2017.

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.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

Tomi Silander and Petri Myllymäki. A simple approach for finding the globally optimal bayesian network structure. In *Conference on Uncertainty in Artificial Intelligence*, pages 445–452, 2006.

Y. Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-Gaussianity. *Biometrika*, 107(1):41–59, 2020. ISSN 0006-3444.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20(76):1–70, 2019.