Anchored Causal Inference in the Presence of Measurement Error

Basil Saeed
MIT
bsaeed@mit.edu

Anastasiya Belyaeva
MIT
belyaeva@mit.edu

Yuhao Wang
MIT
yuhaow@mit.edu

Caroline Uhler
MIT
cuhler@mit.edu

Abstract

We consider the problem of learning a causal graph in the presence of measurement error. This setting is for example common in genomics, where gene expression is corrupted through the measurement process. We develop a provably consistent procedure for estimating the causal structure in a linear Gaussian structural equation model from corrupted observations on its nodes, under a variety of measurement error models. Namely, we provide an estimator based on the method-of-moments and an associated test which can be used in conjunction with constraint-based causal structure discovery algorithms. We prove asymptotic consistency of the procedure and also discuss finite-sample considerations. We demonstrate our method's performance through simulations and on real data, where we recover the underlying gene regulatory network from zero-inflated single-cell RNA-seq data.

1 INTRODUCTION

Determining causal relationships between a set of variables is a central task in causal inference with applications in many scientific fields including economics, biology and social sciences (Friedman et al., 2000; Pearl, 2003; Robins et al., 2000). Directed acyclic graph (DAG) models are commonly used to represent the causal structure among variables. Learning a DAG from observations on the nodes is intrinsically hard (Chickering et al., 2004), and in general a DAG is only identifiable up to its Markov equivalence class (Verma and Pearl, 1990). In addition, in many applications there may be latent variables. While various algorithms have been developed to learn DAGs with latent variables (Spirtes et al., 2000; Colombo et al., 2012), without restrictions on the latent variables there may be many DAGs that explain the data (Spirtes and Richardson, 2002); e.g. the relationship among the latent variables cannot be determined.

Restrictions on the latent variables can improve model identifiability and allow for a causal analysis. In this paper, we consider the problem of causal discovery with measurement error, where each latent variable (representing the true quantity without measurement error) has exactly one corresponding observed variable (representing the corrupted observation of the latent variable), and the goal is to infer the causal relationships among the latent variables; see Figure 1. For instance in social sciences, the beliefs of people cannot be directly measured, but surveys can provide a noisy version of the latent variables, and we may be interested in inferring the causal structure among the latent beliefs. Similarly, in biological applications measurement error needs to be taken into account, e.g. when measuring brain signals using functional magnetic resonance (fMRI) or gene expression using RNA sequencing. The observed variable for each latent variable serves as its *anchor*. Hence, we use the term *anchored causal inference* to parallel its usage for the discrete setting considered in Halpern et al. (2015). This does not directly relate to its usage in "anchored regression" (Rothenhäusler et al., 2018).

While the method developed in this paper can be applied generally to causal inference in the presence of noise, we will showcase its use on learning the underlying gene regulatory network from single-cell RNA-seq data (Klein et al., 2015). Such data is known to suffer from *dropout* (Ziegenhain et al., 2017), which manifests itself as false zeros due to too little starting RNA or technical noise. We model this dropout as a type of measurement noise. In single-cell RNA-seq experiments, it is estimated that such false zeros occur with a probability of 24-76% across current state-of-the-art technologies (Ziegenhain et al., 2017).

Some of the most prominent causal structure discovery algorithms are based on conditional independece test-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

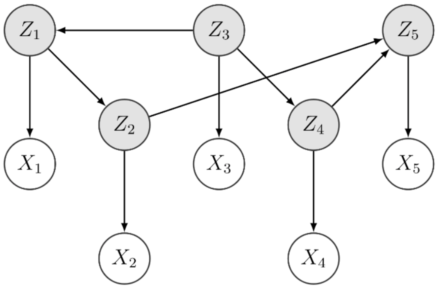
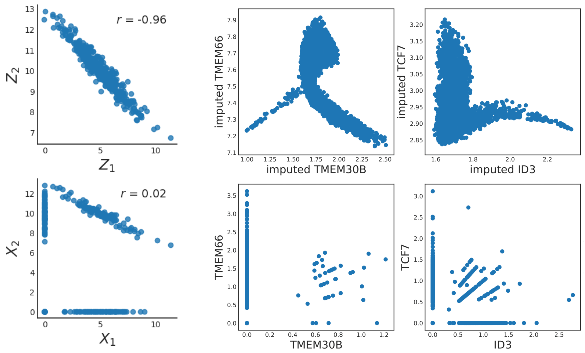


Figure 1: *Anchored Causal Model* with latent variables Z_i and corrupted observed counterparts X_i (*anchors*).

ing (Spirtes et al., 2000; Solus et al., 2017). Estimating these directly from corrupted data may lead to biased estimates since the corruption may lead to changes in the values of the correlations; see Figure 2a. Moreover, dependencies may be introduced in the observed variables that are not present in the latent variables. For instance, consider the model shown in Figure 1: While the latent variables Z_2 and Z_3 are *independent* given Z_1, the observed variables X_2 and X_3 are *dependent* given any conditioning set among the observed variables X.

Currently, the typical approach for dealing with dropout is to first impute gene expression data (Van Dijk et al., 2018). However, imputation may also introduce spurious dependencies (see Figure 2b). Furthermore, since current imputation methods were designed to recover (*unconditional*) correlations/dependence relations among genes but not *conditional* dependence relations, which are critical for causal structure discovery algorithms, the causal graph learned based on imputed data is usually inaccurate. It is therefore of great interest to develop algorithms that directly learn the causal structure among the latent variables from the corrupted data.

Zhang et al. (2017) considered the problem of learning a causal DAG model under measurement error as in Figure 1, but restricted the measurement error to be independent from the latent variables. For many applications, including modeling dropout in single-cell RNA-seq data, this assumption is too restrictive. Halpern et al. (2015) considered more general anchored causal models than in Figure 1, but only in the binary setting. Silva et al. (2006) considered a similar model for continuous distributions, but under the assumption that the dependence between latent and observed variables is linear, an assumption that is too restrictive for many applications. Inspired by topic modeling, Anandkumar et al. (2013) proposed a causal discovery method for DAGs with various levels of latent variables, but under the assumption that the latent variables are non-Gaussian and have sufficient outgoing edges for identifiability of the model.



(a) Dropout (b) Imputation

Figure 2: (a) Simulated Gaussian random variables before (top) and after dropout with rate 0.5 (bottom). (b) Imputed RNA-seq data (top) and raw data with dropout (bottom).

The main contributions of this paper are as follows:

- We introduce *anchored causal inference* to model causal relationships among latent variables in the Gaussian setting with noisy observations .
- We develop a consistent estimator for partial correlations based on the method of moments and an associated conditional independence test that can be used with consistent structure learning algorithms to find the structure among the latent variables, resulting in a consistent procedure for structure learning.
- We present experimental results on both simulated data and real single-cell RNA-seq data, showing that our estimator, which takes into account measurement error, outperforms other methods for causal inference when applied to the corrupted data.

2 PRELIMINARIES

Let $\mathcal{G} = ([p], E)$ be a directed acyclic graph (DAG) with nodes $[p] := \{1, \ldots, p\}$ and directed edges E. We associate a random variable Z_i to each node $i \in [p]$. We denote the joint distribution of $Z = (Z_1, \ldots, Z_p)^T$ by $\mathbb{P}$ and assume that Z is generated by a *linear Gaussian structural equation model*:

$$Z = B^T Z + \epsilon, \tag{1}$$

where B is the weighted adjacency matrix of $\mathcal{G}$ and $\epsilon \sim \mathcal{N}_p(\mu, \Omega)$ with $\Omega = \text{diag}(\omega_1^2, \cdots, \omega_p^2)$. We consider the problem where only a noise-corrupted version of Z is observed. We define $X_1, \ldots, X_p$ to be the *observed* variables (*anchors*) generated from the *latent* variables

$Z_1, \dots, Z_p$ by a noise process $X_i = F_i(Z_i)$, where F_i is a possibly non-deterministic function such that X_i has non-zero variance.. We aim to learn the DAG $\mathcal{G}$ associated with the latent variables Z.

A standard approach for structure discovery when no latent variables are present is to first infer the conditional independence (CI) relations among the observed variables and then use the CI relations to learn the DAG structure (Spirtes et al., 2000). However, since multiple DAGs can encode the same CI relations, $\mathcal{G}$ can only be identified up to its *Markov equivalence class* (MEC). An MEC can be represented by a *CPDAG*, a partially directed graph whose skeleton (underlying undirected graph) is the skeleton of $\mathcal{G}$ and an edge is directed if it has the same direction for all DAGs in the MEC (Verma and Pearl, 1990). Various algorithms have been developed for learning a CPDAG when no latent variables are present (Chickering, 2002; Solus et al., 2017; Spirtes et al., 2000), most prominently the *PC algorithm* (Spirtes et al., 2000), which treats causal inference as a constraint satisfaction problem with the CI relations as constraints. The PC algorithm is provably consistent, meaning that it outputs the correct MEC when the sample size $n \to \infty$ under the so-called *faithfulness* assumption, which asserts that the CI relations entailed by $\mathbb{P}$ are the relations implied by separation in the DAG $\mathcal{G}$ (Spirtes et al., 2000).

Harris and Drton (2013) proved high-dimensional consistency of the PC-algorithm when applied directly to the observed data X for Gaussian copula or non-paranormal models, i.e., when the Gaussian random vector Z is a *deterministic* function of X. In this case, the conditional independence statements among the variables X are equivalent to the conditional independence statements among Z (Harris and Drton, 2013), which simplifies the problem greatly. When X is a random function of Z, the setting considered in this paper, this equivalence generally does not hold. Figure 1 shows an example where $Z_2 \perp\!\!\!\perp Z_3 | Z_1$, but without any further assumptions, $X_2 \not\perp\!\!\!\perp X_3 | X_1$. Furthermore, Yoon et al. (2020) provided a consistent rank-based estimator for the correlation matrix of Z when the X_i are from a *truncated* Guassian copula, meaning that there exist constants $c_1, \dots, c_p$ such that X_i is set to zero if $X_i > c_i$, and applied this for estimating correlations in gene expression data. While our model assumptions are for a different class of noise functions and our gene expression model differs from that of Yoon et al. (2020), we provide a comparison of the CPDAG learned by the PC-algorithm using this correlation matrix with the CPDAG learned using our estimation method in Section 5.1.

As compared to the fully observational setting, when latent variables are present, identifiability is further weakened. Various algorithms have been developed for learning in this setting (Spirtes et al., 2000; Colombo et al., 2012). However, these algorithms cannot estimate causal relations among the latent variables, which is our problem of interest. Moreover, Leung et al. (2016) study identifiability of directed Gaussian graphical models in the presence of a single latent variable, Blom et al. (2018) provide an upper bound for the measurement error when the noise is Gaussian and independent of the latent variables, and Zhang et al. (2017), Silva and Scheines (2005), Silva et al. (2006), Halpern et al. (2015) and Anandkumar et al. (2013) all consider the problem of learning causal edges among latent variables from the observed variables, i.e. models as in Figure 1 or generalizations thereof, but under assumptions that may not hold for our applications of interest, namely that the measurement error is independent of the latent variables (Zhang et al., 2017), that the observed variables are a linear function of the latent variables (Silva and Scheines, 2005; Silva et al., 2006), that the observed variables are binary (Halpern et al., 2015), or that each latent variable is non-Gaussian with sufficient outgoing edges to guarantee identifiability (Anandkumar et al., 2013). We note, however, that our work requires assumptions (described below) that some of these methods do not.

3 ANCHORED CAUSAL INFERENCE

In the following, we first describe the assumptions of our *Anchored Causal Model*, then motivate the model by the application to learning the underlying gene regulatory network from zero-inflated single-cell RNA-seq data, and finally provide an algorithm for anchored causal inference and prove its consistency under the model assumptions.

Model Assumptions (Anchored Causal Model)**.**

(A1). *Given a DAG $\mathcal{G} = ([p], E)$, the latent variables $Z = (Z_1, \dots Z_p)$ are generated by a linear Gaussian structural equation model (see (1)) that is faithful to $\mathcal{G}$.*

(A2). *The observed random vector $X = (X_1, \dots, X_p)$ satisfies for all $i \in [p]$*

$$X_i \perp\!\!\!\perp \{X_1, \dots, X_p, Z_1, \dots, Z_p\} \setminus \{X_i, Z_i\} \mid Z_i.$$

Furthermore, for all $i, j \in [p]$ there exists a finite-dimensional vector η_i of monomials in X_i and a finite-dimensional vector η_{ij} of monomials in X_i and X_j such that their means can be mapped to the moments of the latent variables by known continuously differentiable functions g_i and g_{ij}, i.e., $\mathbb{E}[Z_i] = g_i(\mathbb{E}[\eta_i])$ and $\mathbb{E}[Z_i Z_j] = g_{ij}(\mathbb{E}[\eta_{ij}])$, and their covariance satisfies $Cov(\eta_i, \eta_{ij}) < \infty$.

While Assumption (A1) fixes the structural and functional relationship between the latent variables Z by a linear

Gaussian structural equation model, Assumption (A2) fixes the structural relationship between latent and observed variables by each X_i having exactly one parent Z_i for all $i \in [p]$, and ensures that the first- and second-order moments of $Z \sim \mathcal{N}_p(\mu, \Sigma)$ can be obtained from moments of X without the restriction to a specific measurement error model. This allows for more general noise models than in (Silva et al., 2006; Zhang et al., 2017). We remark that given a model specified by the noise function F_i, the functions g_i and g_{ij} are directly obtainable from F_i, as we demonstrate in the following example. Hence, Assumption (A2) could have been specified in terms of F_i. However, knowledge of F_i is, in general, a stronger condition than knowledge of g_i and g_{ij}; we therefore chose to specify our assumptions in terms of g_i and g_{ij}.

Example 3.1 (Additive Gaussian noise model)**.** *Suppose*

$$X_i = F_i(Z_i) = Z_i + U_i, \;\; \textit{where} \;\; U_i \sim \mathcal{N}(m_i, s_i^2)$$

for all $i \in [p]$. Then, it holds that $\mathbb{E}[X_i] = \mu_i + m_i$, and hence

$$g_i(y) = y - m_i, \quad \eta_i = X_i.$$

Similarly,

$$\begin{aligned} \mathbb{E}[X_i^2] &= \mathbb{E}[\mathbb{E}[X_i^2|Z_i]] \\ &= \mathbb{E}[Z_i^2] + 2m_i\mathbb{E}[Z_i] + s_i^2 + m_i^2. \end{aligned}$$

Hence,

$$g_{ii}(y) = y_2 - 2m_i g_i(y_1) - s_i^2 - m_i^2$$

with $\eta_{ii} = (X_i, X_i^2)$. Finally, for $i \neq j$, we have

$$\begin{aligned} \mathbb{E}[X_iX_j] &= \mathbb{E}[\mathbb{E}[X_i|Z_i]\mathbb{E}[X_j|Z_j]] \\ &= \mathbb{E}[Z_iZ_j] + m_i\mathbb{E}[Z_j] + m_j\mathbb{E}[Z_i] + m_im_j, \end{aligned}$$

giving

$$g_{ij}(y) = y_3 - m_i g_j(y_2) + m_j g_i(y_1) + m_i m_j$$

with $\eta_{ij} = (X_i, X_j, X_iX_j)$.

We next present examples of other models that fit into our framework. The functions g_i, g_{ij} and the vectors η_i,η_{ij} can be obtained through a calculation similar to that just presented. We therefore omit the details in what follows.

Example 3.2 (Modeling single-cell RNA-seq data)**.** *Let Z_i represent the true latent RNA values (log-transformed) and X_i the observed RNA values after dropout. Pierson and Yau (2015) considered a simple model of gene regulation represented by a linear Gaussian structural equation model among the latent variables Z and modeled dropout for single-cell RNA-seq data by*

$$X_i = F_i(Z_i) = \begin{cases} Z_i & \textit{w.p.} \;\; q_i \\ 0 & \textit{w.p.} \;\; 1 - q_i \end{cases} \qquad \textit{for all} \quad i \in [p].$$

Assuming the dropout probabilities q_i are known, this model satisfies Assumptions (A1) and (A2) with

$$\begin{array}{l} \eta_i \;\, = X_i \\ \eta_{ij} = X_iX_j \end{array}, \;\; g_i(y) = \frac{y}{q_i} \;\; \textit{and} \;\; g_{ij}(y) = \begin{cases} \frac{y}{q_iq_j} & i \neq j \\ \frac{y}{q_i} & i = j \end{cases},$$

but does not satisfy the assumptions in (Silva et al., 2006; Zhang et al., 2017). □

While this is a simple model for RNA-seq data, also more complicated models fit into our framework. For example, the Michaelis-Menten model considers dropout probabilities that depend on $\mu_i = \mathbb{E}(Z_i)$ (Andrews and Hemberg, 2018). In addition, RNA-seq data is often modeled using Poisson random variables (Pachter, 2011; Grün et al., 2014). The following more complex model with these two properties fits into our framework.

Example 3.3 (A more complex model for single-cell RNA-seq data)**.** *Consider the model defined by*

$$X_i = \begin{cases} \textit{Poisson}(Z_i) & \textit{w.p.} \;\; \frac{\mu_i}{c+\mu_i} \\ 0 & \textit{w.p.} \;\; 1 - \frac{\mu_i}{c+\mu_i} \end{cases},$$

where c is a given parameter. Using $\mathbb{E}[X_i] \geq 0$, one can solve for the moments of Z_i in terms of the moments of X_i to obtain

$$\eta_i = X_i, \;\; \eta_{ii} = (X_i, X_i^2), \; \eta_{ij} = (X_i, X_j, X_iX_j)$$

and

$$\begin{aligned} g_i(y) &= \frac{y + \sqrt{y^2 + 4c}}{2}, \\ g_{ii}(y) &= y_2 \frac{c + g_i(y_1)}{g_i(y_1)} - g_i(y_1)^2 - g_i(y_1), \\ g_{ij}(y) &= \frac{(c + g_i(y_1))(c + g_j(y_2))}{g_i(y_1)g_j(y_2)} y_3 - g_i(y_1)g_j(y_2), \end{aligned}$$

for $i \neq j$. Hence also this model satisfies the assumptions of an Anchored Causal Model. □

Having provided various examples motivating our Anchored Causal Model, we now introduce Algorithm 1, our *Anchored Causal Inference* procedure to learn the causal structure among the latent Z variables. The procedure works as follows: Given n i.i.d. samples of X denoted by $\hat{X} = (\hat{X}^{(1)}, \hat{X}^{(2)}, \ldots, \hat{X}^{(n)})$, compute the required empirical moments $\mathbb{E}[\hat{\eta}_i]$ and $\mathbb{E}[\hat{\eta}_{ij}]$. Given a particular measurement error model defined by g_i and g_{ij}, compute the first- and second-order moments of Z to obtain its covariance matrix $\hat{\Sigma}$. If we can obtain the set of CI relations involving Z, we can use causal structure discovery algorithms to learn $\mathcal{G}$ (up to its Markov equivalence class). Since Z follows a Gaussian distribution, conditional independence corresponds to zero partial correlation. Let

Algorithm 1 Anchored Causal Inference

```
Input: $n$ samples $\hat{X} = (\hat{X}^{(1)}, \dots, \hat{X}^{(n)})$ of the random vector $X = F(Z)$; the functions $g_i$ and $g_{ij}$ as in (A2).
Output: CPDAG representing the Markov equivalence class of the DAG $\mathcal{G}$ of the latent variables $Z$.
1. For each $i, j \in [p]$ compute the sample moment vectors $\mathbb{E}[\hat{\eta}_i]$ and $\mathbb{E}[\hat{\eta}_{ij}]$ from the samples $\hat{X}$.
2. Estimate the sample moments of $Z$ via $\hat{\mu}_i \triangleq g_i(\mathbb{E}[\hat{\eta}_i])$, $\hat{\mu}_{ij} \triangleq g_{ij}(\mathbb{E}[\hat{\eta}_{ij}])$.
3. Estimate the covariance matrix $\hat{\Sigma}$ of $Z$ by $(\hat{\Sigma})_{ij} = \hat{\mu}_{ij} - \hat{\mu}_i\hat{\mu}_j$ for all $i, j \in [p]$.
4. Estimate the partial correlations of $Z$ from $\hat{\Sigma}$ using (2).
5. Calculate the test statistics defined in Corollaries 1 or 2 to infer the CI relations among the latent variables $Z$.
6. Use a consistent causal discovery algorithm (e.g. the PC algorithm) based on the inferred CI relations.
```

$i, j \in [p]$ and $K \subseteq [p] \setminus \{i, j\}$, then the sample partial correlations $\hat{\rho}_{ij \cdot K}$ can be computed recursively for increasing conditioning set sizes by

$$\hat{\rho}_{ij \cdot K} = \frac{\hat{\rho}_{ij \cdot K \setminus \{l\}} - \hat{\rho}_{il \cdot K \setminus \{l\}} \hat{\rho}_{jl \cdot K \setminus \{l\}}}{\sqrt{1 - \hat{\rho}^2_{il \cdot K \setminus \{l\}}} \sqrt{1 - \hat{\rho}^2_{jl \cdot K \setminus \{l\}}}}, \qquad (2)$$

where in the base case $\hat{\rho}_{ij \cdot \emptyset} = \hat{\rho}_{ij}$ are the correlations obtained from $\hat{\Sigma}$. The main difficulty lies in developing test statistics based on the estimated partial correlations $\hat{\rho}_{ij \cdot K}$ such that the inferred CI relations correspond as $n \to \infty$ to the set of CI relations implied by the underlying causal DAG $\mathcal{G}$. Such test statistics are developed in Corollaries 1 and 2. The inferred CI relations can then be fed into a constraint-based causal discovery algorithm such as the PC algorithm (Spirtes et al., 2000) or the hybrid GSP algorithm (Solus et al., 2017) to obtain the CPDAG of $\mathcal{G}$.

The first step in asserting consistency of Algorithm 1 is the following lemma. The proof is provided in Appendix A.

Lemma 1. *Under assumptions (A1) and (A2), the estimator $\hat{\rho}_{ij \cdot K}$ in (2) is asymptotically consistent.*

Next, we design a consistent hypothesis test for obtaining CI relations based on the estimated partial correlations of Z in (2), similar in principle to Gaussian CI tests based on Fisher's z-transform used by many causal inference algorithms (Chickering, 2002; Spirtes et al., 2000; Solus et al., 2017). When F_i is the identity function for all $i \in [p]$, it can be shown (Lehmann, 1998) that the estimated partial correlations in (2) satisfy

$$\sqrt{n}(\hat{\rho}_{ij \cdot K} - \rho_{ij \cdot K}) \xrightarrow{D} \mathcal{N}_1\Big(0, (1 - \rho^2_{ij \cdot K})^2\Big). \qquad (3)$$

Hence, applying the Delta method (van der Vaart and Wellner, 1996) to Fisher's z-transform $z_f(\rho) := \log((1 + \rho)/(1 - \rho))/2$ of the estimated partial correlations yields

$$\sqrt{n}\Big(z_f(\hat{\rho}_{ij \cdot K}) - z_f(\rho_{ij \cdot K})\Big) \xrightarrow{D} \mathcal{N}_1(0, 1). \qquad (4)$$

Hence Fisher's z-transform can be used in the test statistic $T := \sqrt{n}\, z_f(\hat{\rho}_{i,j \cdot K})$ to test conditional independence by declaring $X_i \perp\!\!\!\perp X_j | X_K$ at significance α if and only if

$$|T| \leq \Phi^{-1}(1 - \frac{\alpha}{2}), \qquad (5)$$

where Φ^{-1} denotes the inverse CDF of $\mathcal{N}(0, 1)$. The following theorem generalizes (3) to our anchored causal model class, where the partial correlations of Z are estimated from the observed moments of X.

Theorem 1. *Let η denote the vector of monomials of X required to compute the first- and second-order moments of Z. Let ν denote the vector of first- and second-order moments of η. Then under assumptions (A1) and (A2), for any $i, j \in [p]$ and $K \subseteq [p] \setminus \{i, j\}$, the estimated partial correlation $\hat{\rho}_{ij \cdot K}$ in (2) satisfies*

$$\sqrt{n}(\hat{\rho}_{i,j \cdot K} - \rho_{i,j \cdot K}) \xrightarrow{D} \mathcal{N}_1\left(0, \tau_{ij \cdot K}(\nu)\right)$$

where $\tau_{ij \cdot K}$ is a continuous function of ν.

The proof of Theorem 1 can be found in Appendix B, where we provide a procedure for computing the function $\tau_{ij \cdot K}$ for any $i, j \in [p]$ and $K \subseteq [p] \setminus \{i, j\}$. The main idea of the proof is as follows: First apply the Central Limit Theorem to the vector of sample moments $\mathbb{E}[\hat{\eta}]$. Under assumption (A2), the correlations ρ based on Σ are continuously differentiable functions of ν. Furthermore, for any $i, j \in [p]$ and $K \subseteq [p] \setminus \{i, j\}$, the partial correlation $\rho_{ij \cdot K}$ is defined recursively for increasing conditioning set sizes as a continuously differentiable function of ρ. Hence, one can iteratively apply the Delta method (Lehmann, 1998) starting from the statement of the Central Limit Theorem applied to $\mathbb{E}[\hat{\eta}]$ to obtain the asymptotic distribution of $\hat{\rho}_{ij \cdot K}$.

In the following two corollaries to Theorem 1, we provide different test statistics for CI testing based on the estimated partial correlations of the latent vector Z. We start by generalizing Fisher's transform and its asymptotic distribution given in (4).

Corollary 1. *If the asymptotic variance $\tau_{ij \cdot K}(\nu)$ can be written purely as a function of $\rho_{ij \cdot K}$, i.e., there exists $\tilde{\tau}_{ij \cdot K}$ such that $\tau_{ij \cdot K}(\nu) = \tilde{\tau}_{ij \cdot K}(\rho_{ij \cdot K})$, and there exists a variance stabilizing transformation $z_{ij \cdot K}$ such that*

$$z_{ij \cdot K}(\rho) = \int \frac{1}{\sqrt{\tilde{\tau}_{ij \cdot K}(\rho)}} d\rho + C \qquad (6)$$

with C chosen such that $z_{ij \cdot K}(0) = 0$, then under (A1) and (A2)

$$\sqrt{n}\Big(z_{ij \cdot K}(\hat{\rho}_{ij \cdot K}) - z_{ij \cdot K}(\rho_{ij \cdot K})\Big) \xrightarrow{D} \mathcal{N}_1(0, 1).$$

The proof of Corollary 1 follows by applying the Delta method to Theorem 1 (Appendix C). Whether the conditions of Corollary 1 are satisfied, depends on the measurement error model F. We show in Appendix F.1 that the conditions of Corollary 1 hold for the dropout model with $K = \emptyset$ and $\mu = 0$, and derive the corresponding variance stabilizing transformation. Note that it is sufficient if we can compute the integral in (6) numerically; a closed-form solution is not required. Corollary 1 implies that the test statistic $T = \sqrt{n}\, z(\hat{\rho}_{i,j\cdot K})$ in (5) can be used to consistently estimate the CI relations among the latent variables Z. When the assumptions of Corollary 1 are not met, then we can obtain a different test statistic that is asymptotically normal as in the following result.

Corollary 2. *Define* $\zeta_{ij\cdot K}(\hat{\rho}, \hat{\nu}) := \hat{\rho}/\sqrt{\tau_{ij\cdot K}(\hat{\nu})}$. *Then under (A1) and (A2)*

$$\sqrt{n}\Big(\zeta_{ij\cdot K}(\hat{\rho}_{ij\cdot K}, \hat{\nu}) - \zeta_{ij\cdot K}(\rho_{ij\cdot K}, \hat{\nu})\Big) \xrightarrow{D} \mathcal{N}_1(0, 1).$$

Corollary 2 follows from Theorem 1: since τ is continuous in ν, $\hat{\nu}$ converges to ν by the law of large numbers and implies that $\tau_{ij\cdot K}(\hat{\rho}_{ij\cdot K}, \hat{\nu}) \xrightarrow{a.s.} \tau_{ij\cdot K}(\hat{\rho}_{ij\cdot K}, \nu)$ as $n \to \infty$ (see Appendix D). Hence the test statistic $T = \sqrt{n}\, \zeta_{ij\cdot K}(\hat{\rho}_{ij\cdot K}, \hat{\nu})$ can be used in (5) to obtain the CI relations among the latent variables Z.

With respect to finite-sample considerations, note that $\zeta_{ij\cdot K}$ in Corollary 2 is a function of $\hat{\nu}$, and thus its convergence to its asymptotic distribution requires the convergence of $\hat{\nu}$. Hence, we expect the convergence in distribution of Corollary 1 to be faster than that of Corollary 2 and as a result the test statistic in Corollary 1 to perform better in the finite-sample regime.

We end this section with the main result of this paper.

Theorem 2. *Under assumptions (A1) and (A2), Algorithm 1 is consistent, i.e., as* $n \to \infty$ *it returns a CPDAG for the the Markov equivalence class of the true DAG* $\mathcal{G}$.

The proof is given in Appendix E, where we show that under the faithfulness assumption, the set of CI relations inferred from $\hat{\Sigma}$ in Algorithm 1 converges to those implied by the DAG $\mathcal{G}$. Hence using any consistent causal structure learning algorithm on these CI relations results in the correct equivalence class among the Z variables.

Remark 1. *The estimator in Algorithm 1, along with the results of Corollary 1 and 2 can be used with any CI-based structure learning algorithm to identify the structure among the* Z *variables. Hence, one can directly modify the assumptions of the anchored causal model to allow the* Z_i *to be generated by a Gaussian distribution that is faithful to a MAG (Spirtes and Richardson, 2002) (i.e., with latent variables) and replace the PC algorithm in the last step of Algorithm 1 for example by the FCI algorithm (Spirtes et al., 2000).*

4 IMPLEMENTATION

Next, we discuss an important aspect of implementation and show how the results in Section 3 can be applied to the dropout model in Example 3.2.

In the finite-sample setting, the estimated covariance matrix $\hat{\Sigma}$ of the latent variables is not guaranteed to be positive semidefinite. In this case, shrinkage towards a positive definite matrix can be used as a form of regularization. When $n < p$, a standard approach is to use Ledoit-Wolf shrinkage towards the identity matrix (Ledoit and Wolf, 2004). When $n > p$, the sample covariance matrix $\hat{S}$ based on the samples $\hat{X}$ is positive definite with probability 1 and hence $\hat{\Sigma}$ can also be shrunk towards $\hat{S}$ by

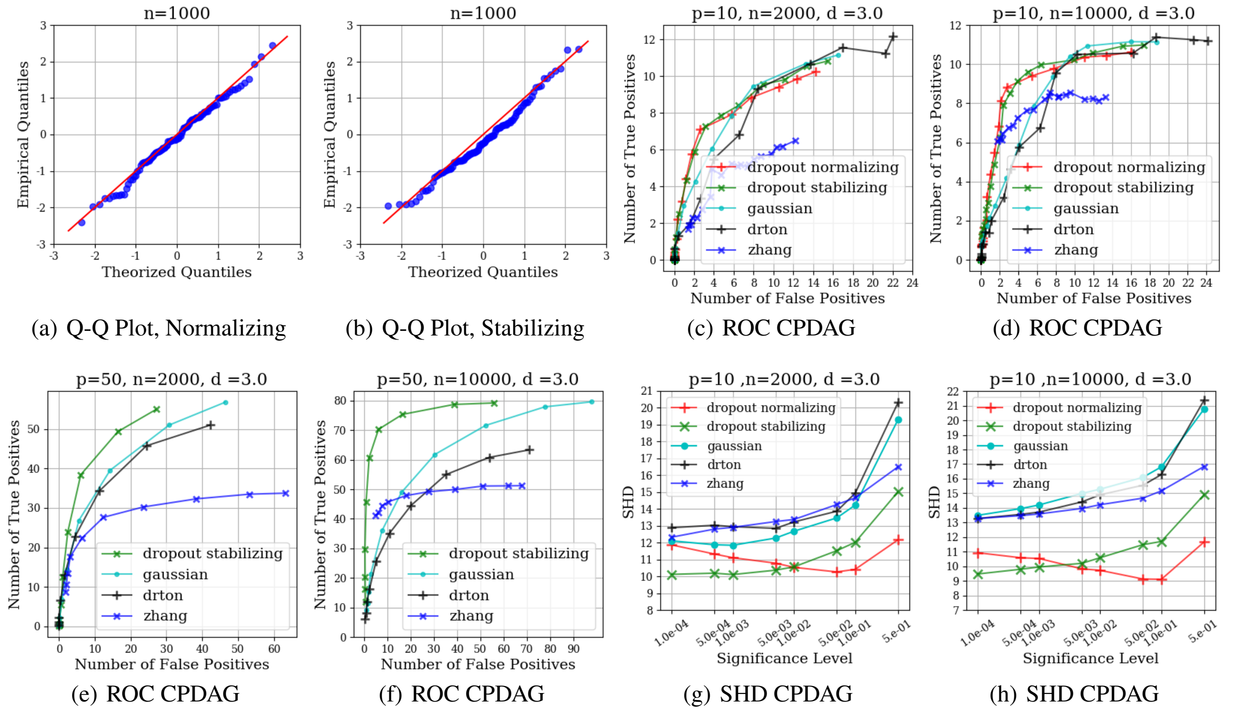
$$\hat{\Lambda} = (1-\alpha^*)\hat{\Sigma} + \alpha^*\hat{S} \quad \text{where} \quad \alpha^* = \underset{\alpha\in[0,1],\hat{\Lambda}\succeq 0}{\arg\min}\ \alpha. \tag{7}$$

The shrinkage that provides better results depends on whether $\hat{S}$ or the identity are better approximations of the true underlying covariance matrix Σ. In our experiments in Section 5, we applied shrinkage towards $\hat{S}$ as in (7). Both types of shrinkage result in consistent estimates: consistency of Ledoit-Wolf shrinkage is proven in (Ledoit and Wolf, 2004) and the consistency of shrinkage towards the sample covariance matrix in (7) follows from Theorem 1, since $\hat{\Sigma} \to \Sigma$ as $n \to \infty$ implies that $\hat{\Sigma}$ becomes positive semidefinite with large enough sample size, and therefore, $\alpha \to 0$ as $n \to \infty$, which shows that shrinkage reduces to the consistent case without shrinkage.

4.1 Application: Dropout Model

Under the dropout model in Example 3.2, the assumptions of Corollary 1 are in general not satisfied, since $\tau_{ij\cdot K}$ cannot generally be expressed as a function of $\rho_{ij\cdot K}$ only. This is shown in Appendix F.2 by plotting τ as a function of ν for fixed ρ. In the special case when $\mu = 0$, the conditions are satisfied for all $i, j \in [p]$ when $K = \emptyset$, i.e., a variance stabilizing transform $z_{ij} = z_{ij\cdot\emptyset}$ can be found for the correlations $\rho_{ij} = \rho_{ij\cdot\emptyset}$. This *dropout stabilizing transform* is provided in Appendix F.1 and can be used together with the resulting CI test as a heuristic also when $K \neq \emptyset$ or $\mu \neq 0$. Its performance is analyzed in Section 5.

In Appendix F.3, we provide a recursive formula for computing $\zeta_{ij\cdot K}$ from Corollary 1 for the dropout model, which we refer to as the *dropout normalizing transform*. From Corollary 1, $\zeta_{ij\cdot K}(\hat{\rho}_{ij\cdot K})$ should have an asymptotic variance of 1. Hence, computing $\zeta_{ij\cdot K}$ requires determining the asymptotic variance $\tau_{ij\cdot K}$ of $\hat{\rho}_{ij\cdot K}$, and then defining $\zeta_{ij\cdot K}$ appropriately to transform this variance to 1. Also note that applying shrinkage will in general change the asymptotic variance of the partial correlations.



(a) Q-Q Plot, Normalizing

(b) Q-Q Plot, Stabilizing

(c) ROC CPDAG

(d) ROC CPDAG

(e) ROC CPDAG

(f) ROC CPDAG

(g) SHD CPDAG

(h) SHD CPDAG

Figure 3: (a)-(b) Q-Q plots for the dropout normalizing and stabilizing transforms. (c)-(h) Performance of the dropout stabilizing transform, dropout normalizing transform and Gaussian CI test in simulations for structure discovery: ROC curves (c)-(f) and SHD curves (g)-(h) for evaluating the accuracy of estimating the true CPDAG.

We show how to correct $\tau_{ij\cdot K}$ as a function of the shrinkage coefficient α in Appendix F.4. This adjustment is applied in all of our experiments in Section 5.

In Section 5.2, we apply our estimation procedure based on the dropout model to single-cell RNA-seq data to infer the structure of the underlying regulatory network. For the theoretical analysis in Section 3 we assumed that the dropout probabilities q_i are known. However, in general these parameters need to be estimated, and it was proposed in (Pierson, 2015) to model q_i by $q_i = 1 - \exp^{\lambda \mu_i^2}$ for $i \in [p]$, where λ depends on the RNA-seq assay. Using this model for the dropout probabilities we can jointly estimate the parameters μ and q as follows. We can write

$$\mathbb{E}[\eta_i] = \mathbb{E}[X_i] = \mathbb{E}[(1 - \exp^{\lambda \mu_i^2}) Z_i] = (1 - \exp^{\lambda \mu_i^2})\mu_i.$$

Since μ_i corresponds to count averages, we can assume that $\hat{\mu}_i \geq 0$. Under this assumption, the equation $\mathbb{E}[\hat{\eta}_i] = (1 - \exp^{\lambda \hat{\mu}_i^2})\hat{\mu}_i$ has a unique solution for $\hat{\mu}_i$.

With respect to the parameter λ, for some single-cell RNA-seq assays it is possible to obtain an estimate for λ by including molecules with known expression as controls. However, since this estimate is often unreliable (Grün and van Oudenaarden, 2015) and not always available, we selected λ so as to minimize the amount of shrinkage required to obtain a positive semidefinite matrix.

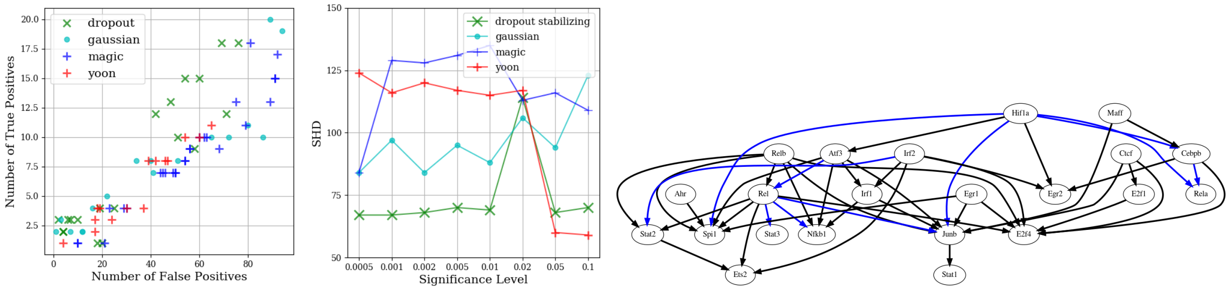
5 EXPERIMENTS

In this section, we analyze the performance of the test statistics derived earlier, along with Algorithm 1 under the dropout model from Example 3.2 both on simulated data and on single-cell RNA-seq data[1]

5.1 Simulations

Test Statistic Evaluation. First, we evaluate the performance of the dropout normalizing and dropout stabilizing transforms on the dropout model in Example 3.2 using a Q-Q plot to compare the empirical distribution of the associated statistic under the null hypothesis ($\rho_{i,j\cdot K} = 0$) with the theorized standard normal distribution. We generated data from the dropout noise model with a simple graph structure on the Z variables such that $X_i \perp\!\!\!\perp X_j | X_k, X_l$ holds for some $i, j, k, l \in [p]$. We used $p = 5$, weights in the range $[-1, -0.25] \cup [0.25, 1]$, and dropout probabilities in $[0, 0.8]$. We generated n observations of X for $n \in \{500, 1000, 5000\}$, then estimated the partial correlation $\hat{\rho}_{ij\cdot kl}$ and applied the dropout normalizing transform (Figure 3a) and the dropout stabilizing transform (Figure 3b) to obtain the associated test statistic. We repeated

[1] The code is available in the `causaldag` repository at `https://github.com/uhlerlab/causaldag`.



(a) Perturb-seq

(b) Pancreas

(c) Perturb-seq gene regulatory network

Figure 4: (a) ROC curve for predicting causal effects of interventions in Perturb-seq data. (b) SHD between CPDAGs estimated on single-cell RNA-seq data from pancreatic cells with low versus high dropout rate. (c) Gene regulatory network estimated from Perturb-seq data (blue edges indicate previously known interactions (Dixit et al., 2016) that were also detected by our method).

this generation and estimation process to obtain 100 points in the Q-Q plots. The plots for $n \in \{500, 5000\}$ are shown in Appendix G. The results show that the distribution of these statistics deviate only slightly from the expected distribution. The dropout stabilizing transform deviates slightly more, as expected, since it was derived for correlations and not for partial correlations.

Structure Learning. The data was generated from the dropout model described in Example 3.2. The structure of the matrix B in the linear Gaussian structural equation model (1) was generated by an Erdös-Renyi model with expected degree d for $d \in \{2, 3, 5\}$ and number of nodes $p \in \{10, 30, 50\}$. The weights of the matrix B were uniformly drawn from $[-1, -0.25] \cup [0.25, 1]$ to be bounded away from 0. The mean parameters μ_i were uniformly drawn from $[0, 3]$ and the probabilities q_i from $[0, 0.8]$. These ranges were chosen to match the expected ranges in the gene expression data analyzed in Section 5.2. We generated n observations of X from this generating model, for $n \in \{1000, 2000, 10000, 50000\}$.

Figure 3 (c)-(h) shows the ROC curves and the Structural Hamming Distance (SHD) evaluating the CPDAG output by Algorithm 1 for $p \in \{10, 50\}$, $n \in \{2000, 10000\}$ and $d = 3$. Each point in the plots is an average over 96 simulations. We compare 5 methods for estimating the graph. Each corresponds to a different curve in the plots. We use (1) PC with the estimator in algorithm 1 and the dropout stabilizing transform (labeled `dropout stabilizing`), and similarly (2) with the dropout normalizing transform (`dropout normalizing`); (3) PC and the Gaussian CI test applied directly to the observations (`gaussian`); (4) PC with the rank-based correlation test used by Harris and Drton (2013) (`drton`); and (5) the algorithm from Zhang et al. (2017) for the case of Gaussians (`zhang`).

In plotting the ROC curve for the CPDAG, we consider an undirected edge in the CPDAG a true positive if a directed edge exists in either direction in the true graph, and a false positive otherwise. We consider a directed edge a true positive if the same edge exists in the true graph. In Appendix G, we provide additional figures for the setting where $p = 30$, $n \in \{2000, 50000\}$ and $d \in \{2, 5\}$.

The simulation results show that the dropout stabilizing transform outperforms, or performs at least as well as the other methods in all settings we tested even though it was derived for $K = \emptyset$ and $\mu = 0$. The performance of both dropout transforms improves over the other methods with increasing sample size. A large sample size is especially important for the dropout normalizing transform because it relies on the estimation of more parameters. Since the dropout stabilizing transform is preferable to the dropout normalizing transform computationally, we concentrate on this transform in Figure 3 (e)-(f).

The simulation results show that our estimators outperform the naive Gaussian CI test applied directly to the corrupted data, and the difference in performance increases with sample size. This has important implications for the development of new single-cell RNA-seq technologies, since it indicates that an increased sample size is preferrable to minimizing dropout. Current technologies have been heading in this direction, trading off increased sample sizes (with studies containing up to a million samples) for an increased dropout rate (Zheng et al., 2017).

5.2 Single-cell RNA-seq Data

Perturb-seq. We tested our method on gene expression data collected via single-cell Perturb-seq by Dixit et al. (2016) from bone marrow-derived dendritic cells (BMDCs) after log-transformation. As in most single-cell studies, the gene expression observations are affected by dropout. The data consists of 933 observational samples (after standard pre-processing), which we used for learning the gene regulatory network. The Perturb-seq data set also contains interventional samples, which we used to evaluate the estimated CPDAG and construct an ROC curve. As in Dixit et al. (2016), we focused our analysis on 24 genes, which are important transcription factors known to regulate each other as well as a variety of other genes (Garber et al., 2012). We used the dropout stabilizing transform to obtain the CI relations among the latent variables (labeled `dropout` in Figure 4) and compared the resulting CPDAG to the graph obtained using the standard Gaussian CI test applied directly to the observed corrupted data (labeled `gaussian`). In both settings we used the PC algorithm to infer the CPDAG from the CI relations. For a typical baseline used in genomics, we also imputed gene expression data using MAGIC (Van Dijk et al., 2018) and applied the PC algorithm to the imputed data (labeled `magic`). Additionally, we used the PC algorithm applied to the rank-based estimator of the correlation matrix of Z from Yoon et al. (2020) (labeled `yoon`). Figure 4(a) shows the resulting ROCs, which quantify for varying tuning parameters the accuracy of each of the learned CPDAGs in predicting the effect of each of the eight interventions. Our procedure with the dropout stabilizing transform outperforms the others. The inferred gene regulatory network is shown in Figure 4(c).

Pancreas - Type II Diabetes. We also tested our method on two gene expression datasets collected from human pancreatic cells (Baron et al., 2016; Segerstolpe et al., 2016) via different single-cell assays, one with low dropout rate and 3514 cells (Smart-seq2), and the other with high dropout rate and 8569 cells (inDrop). The expression data was log-transformed in both cases. We focused our analysis on a gene regulatory network of 20 genes, which is known to be involved in Type II Diabetes (Sharma et al., 2018). Since no interventional data is available for this application, we evaluated our estimator based on how consistent the estimated CPDAG is across the two data sets. Figure 4b shows the SHD between the CPDAGs estimated from the data set with low versus high dropout using the dropout stabilizing transform, Gaussian CI test applied directly to the data, and Gaussian CI test applied to data imputed with MAGIC. The inferred gene regulatory network is provided in Appendix G. Since the SHD is generally lower for the dropout stabilizing transform, the CPDAG estimates produced by our method are more consistent across different dropout levels, thereby suggesting that our method is more robust to dropout.

6 DISCUSSION

We proposed a procedure for learning causal structure in the presence of measurement error under the *Anchored Causal Model*, where each corrupted observed variable is generated from a latent uncorrupted variable and the aim is to learn the structure among the latent variables using the observed variables as *anchors*. We introduced an estimator and test statistics that can be used with CI-based structure discovery algorithms to learn the Markov equivalence class of the causal DAG among the latent variables. One of the main motivations for developing this algorithm was to address the problem of dropout in single-cell RNA-seq experiments. We showed how to apply our algorithm for learning the underlying gene regulatory network under a simple dropout model and analyzed its performance on synthetic data and on single-cell RNA-seq data, An interesting future application for this methodology would be to use this framework with more complex dropout models for RNA-seq data to further improve performance.

Acknowledgements

Basil Saeed was partially supported by MIT J-Clinic for Machine Learning and Health. Anastasiya Belyaeva was supported by the Abdul Latif Jameel World Water and Food Security Lab (J-WAFS) at MIT. Caroline Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), IBM, a Sloan Fellowship and a Simons Investigator Award.

References

A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade. Learning linear bayesian networks with latent variables. In *International Conference on Machine Learning*, pages 249–257, 2013.

T. S. Andrews and M. Hemberg. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics*, 2018.

M. Baron, A. Veres, S. L. Wolock, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems*, 3(4):346–360, 2016.

T. Blom, A. Klimovskaia, S. Magliacane, and J. M. Mooij. An upper bound for random measurement error in causal discovery. *arXiv preprint arXiv:1810.07973*, 2018.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5(Oct):1287–1330, 2004.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

A. Dixit, O. Parnas, B. Li, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866, 2016.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

M. Garber, N. Yosef, A. Goren, et al. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular Cell*, 47(5):810–822, 2012.

D. Grün and A. van Oudenaarden. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4): 799–810, 2015.

D. Grün, L. Kester, and A. Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637, 2014.

Y. Halpern, S. Horng, and D. Sontag. Anchored discrete factor analysis. *Preprint arXiv:1511.03299*, 2015.

N. Harris and M. Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(1):3365–3383, 2013.

A. M. Klein, L. Mazutis, I. Akartuna, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Texts in Statistics, 1998.

D. Leung, M. Drton, and H. Hara. Identifiability of directed Gaussian graphical models with one latent source. *Electronic Journal of Statistics*, 10(1):394–422, 2016.

L. Pachter. Models for transcript quantification from RNA-Seq. *Preprint arXiv:1104.3889*, 2011.

J. Pearl. Causality: Models, reasoning, and inference. *Econometric Theory*, 19(4):675–685, 2003.

E. Pierson. Statistical models for single-cell data. Master's thesis, Oxford University, UK, 2015.

E. Pierson and C. Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

Å. Segerstolpe, A. Palasantza, P. Eliasson, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabolism*, 24(4): 593–607, 2016.

A. Sharma, A. Halu, J. L. Decano, et al. Controllability in an islet specific regulatory network identifies the transcriptional factor NFATC4, which regulates type 2 diabetes associated genes. *NPJ Systems Biology and Applications*, 4(1):25, 2018.

R. Silva and R. Scheines. Generalized measurement models. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2005.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

L. Solus, Y. Wang, L. Matejovicova, and C. Uhler. Consistency guarantees for permutation-based causal inference algorithms. *Preprint arXiv:1702.03530*, 2017.

P. Spirtes and T. Richardson. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

A. W. van der Vaart and J. A. Wellner. *The Delta-Method*, pages 372–400. Springer, New York, 1996.

D. Van Dijk, R. Sharma, J. Nainys, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270, 1990.

G. Yoon, R. J. Carroll, and I. Gaynanova. Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 2020.

K. Zhang, M. Gong, J. Ramsey, K. Batmanghelich, P. Spirtes, and C. Glymour. Causal discovery in the presence of measurement error: Identifiability conditions. *Preprint arXiv:1706.03768*, 2017.

G. X. Y. Zheng, J. M. Terry, P. Belgrader, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

C. Ziegenhain, B. Vieth, S. Parekh, et al. Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, 65(4):631–643, 2017.