MANM-CS: Data Generation for Benchmarking Causal Structure Learning from Mixed Discrete-Continuous and Nonlinear Data

Johannes Huegle, Christopher Hagedorn, Lukas Böhme*, Mats Pörschke*, Jonas Umland*, Rainer Schlosser
Hasso Plattner Institute, University of Potsdam, Germany
{firstname.lastname}@hpi.de, *{firstname.lastname}@student.hpi.de

Abstract

In recent years, the growing interest in methods of causal structure learning (CSL) has been confronted with a lack of access to a well-defined ground truth within real-world scenarios to evaluate these methods. Commonly used synthetic benchmarks are limited in their scope as they are either restricted to a "static" low-dimensional data set or do not allow examining mixed discrete-continuous or nonlinear data. This work introduces the mixed additive noise model that provides a ground truth framework for generating observational data following various distribution models. Moreover, we present our reference implementation MANM-CS that provides easy access and demonstrate how our framework supports researchers and practitioners. Further, we propose future research directions and possible extensions.

1 Introduction and Background

Methods of causal structure learning (CSL) have received widespread attention in the scientific field as the knowledge of underlying causal structures is the basis for decision support within many real-world scenarios [40]. In recent years, the corresponding research addressing challenges of CSL in practice has led to a broad spectrum of different methods (see Sec. 1.1). In this context, the evaluation of CSL methods encounter requirements concerning well-defined benchmark data, e.g., for mixed or nonlinear data (Sec. 1.2). Therefore, we propose the mixed additive noise model (MANM) to establish a flexible yet well-defined ground-truth model, allowing the data generation under various evaluation perspectives (Sec. 1.3).

1.1 Causal Structure Learning and Challenges

In CSL, the following standard notation is used. The causal structures between a finite set of p random variables $\mathbf{V} = \{V_1, \dots, V_p\}$ are encoded in a causal graphical model (CGM) consisting of a directed acyclic graph (DAG) $\mathcal{G}$, where directed edges $V_j \to V_i$ depict a direct causal relationship between two respective nodes V_j and V_i, $i, j = 1, \dots, p$, and the joint distribution over the variables $\mathbf{V}$, denoted by $P^{\mathbf{V}}$, e.g., cf. [27, 40].

Within this framework, CSL aims to derive as many of the underlying causal relationships in $\mathcal{G}$ from independent and identically distributed observational data as possible. Therefore, methods of CSL either leverage probabilistic characteristics of the variables' joint probability distribution $P^{\mathbf{V}}$ or prescribe a specific functional causal model (FCM) to the relation between variables, e.g., cf. [6]. In this context, the causal Markov condition states a coincidence between the causal structure within $\mathcal{G}$ and the conditional independence (CI) characteristics of the joint distribution $P^{\mathbf{V}}$ [40]. On this basis, probabilistic methods either exploit CI tests, called constraint-based, or optimize a score function over

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

the space of equivalence classes, called score-based to recover the causal structures [40]. In contrast, algorithms based on properly defined FCMs benefit from additional but restrictive assumptions and can distinguish between different DAGs within the same equivalence class [6]. In this context, a direct edge $V_j \to V_i$ in $\mathcal{G}$ encodes that V_i is a function of its cause V_j and some noise term N_i independent to V_j, i.e., $V_i = f_i(V_j, N_i)$ where f_i is assumed to be a function from an appropriately constraint functional class $\mathcal{F}$, e.g., linearity with additive noise [14].

While all CSL methods require that several assumptions hold, observational data of real-world scenarios often violates the constraints made for CSL. For example, in practice, it may be impossible to observe all variables to ensure causal sufficiency, i.e., that there are no latent confounding variables [40]. Moreover, real-world data often does not follow a simple functional form but includes nonlinear and mixed discrete-continuous relationships [6, 23]. Therefore, a wide spectrum of scientific publications focus on different extensions to improve the accuracy under weakened constraints, e.g., given latent variables [4, 39], assuming a nonlinear function f_i within FCMs [9, 46] or considering CSL from mixed discrete continuous data [1, 10, 36, 42].

1.2 Requirements on Modeling Causal Structures for Benchmarking CSL Methods

As (novel) CSL methods are commonly evaluated against their own synthetic benchmarks and compared within a limited scope, e.g., [30, 38, 43], it is difficult to compare individual methods against each other; in particular, if they may require different assumptions. In this context, Glymour et al. [6] summarized the current state as follows: "There are multiple algorithms available, many of them are poorly tested [...] all of them have choices of parameters [...], and all of them have conditions on the data distributions and other assumptions under which they will be informative rather than misleading". Hence, methods for CSL should be validated within different scenarios, including a varying number of variables or sensitivity of parameters, aiming to understand the method's behaviors in specific edge cases, e.g., when underlying assumptions on the causal relationships are violated [17]. Therefore, a thorough evaluation of CSL methods requires the introduction of an easily customizable framework for generating observational data supplemented with precise definitions of underlying causal structures that connects and extends existing ideas, see Sec. 2. In particular, a data generating model should satisfy the following requirements:

(R1) be formalized as a FCM to ensure interpretability, e.g., concerning causal inference;
(R2) allow for continuous, discrete, and mixed discrete-continuous causal relationships;
(R3) be flexible and easily extendable, e.g., to allow for interventional data;
(R4) be implemented as an easy-to-use open access package.

1.3 Contribution

In our work, we propose the mixed additive noise model (MANM) as a flexible yet easy-to-use synthetic data generation process for benchmarking CSL methods under a wide range of conditions. Our main contributions can be summarized as follows: First, we introduce the MANM as an FCM to model causal structures within various distribution models from discrete over mixed discrete-continuous to nonlinear, cf. (R1)-(R3). Second, to provide easy access to the research community, we present our reference implementation, called `MANM-CS`, cf. (R4). Third, we demonstrate the usability of `MANM-CS` in comparison to the well-known benchmark data sets and in a simple benchmarking experiment on the accuracy of CSL from mixed discrete-continuous and nonlinear data.

The remainder of the paper is structured as follows. We consider related work on available benchmarking methods of CSL in Section 2. In Section 3, we introduce the MANM as a benchmarking framework and demonstrate its application within several common scenarios. We present our reference implementation `MANM-CS` in Section 4 and its application for benchmarking in Section 5. In Section 6, we conclude our work, point out limitations, and discuss future directions.

2 Related Work on Benchmarking of CSL Methods

Commonly, data for evaluating CSL methods is generated according to the following approaches:

(I) predefined benchmark data sets supplemented by an expected ground truth;
(II) well-established parameterized benchmark models to generate data; and
(III) flexible models based upon a probabilistic or functional formalization.

Approach	Continuous	Mixed	Discrete
(I)	[24], [25]‡, [33]	[7]	[31]
(II)	[26], [32], [35]	[22], [34], [44]	[2], [3], [5], [15], [16]
(III)	[9]‡, [13]	[1]§, [13], [18]*	[29]‡

Table 1: Implementations for modeling continuous, mixed, or discrete data based upon (I) predefined benchmark data sets, (II) well-established parameterized models, and (III) functional or probabilistic models (*undirected; ‡ two-variable case; § discretized auxiliary variables).

In Table 1, we recap a selection of the above approaches that has been used for evaluation within work on CSL, e.g., [1, 30, 36, 42, 43]. We do not claim completeness but focus on the most well-known and representative data sets or models. In this context, currently used models and data sets of the three approaches come with limitations that restrict the evaluation opportunities.

Predefined benchmark data sets (I) allow a direct comparison given a common and enclosed ground truth model. However, they do not allow for performance comparison concerning a varying complexity or data set size. For example, the "`DREAM5 SYSGEN A` - In-silico network challenge" [24] is based on simulated gene expression data from [20] restricted to 1000 variables and sample sizes of $n = 100$, $n = 300$, or $n = 999$. To allow for performance evaluation of large sample properties, models from (II) sample observational data from well-established "static" parameterized models with fixed model complexity. While this allows to evaluate and compare CSL within the provided distribution and model assumption, they do not allow for an examination given a varying model complexity. For example, within the mixed case, the well-known conditional Gaussian distributed `MEHRA` model from [44] is restricted to 24 nodes that incorporate a mixture of 8 discrete and 16 continuous variables. Further, within the discrete case, the `ALARM` model from [2] fixes the number of possible discrete values each variable can take to the model's assumptions. Therefore, following approach (III), most CSL methods are evaluated within their respective scenarios, e.g., linear relationships with i.i.d. Gaussian noise, cf. [13], the mixed graphical model (MGM) of [18], or the two-variable case, cf., [9, 28]. As a basis for a comprehensive evaluation, these models are quite restrictive or vary strongly on their assumptions, e.g., the solely undirected edges of the mixed MGM model [18], and require manual implementation overhead as they are often not open accessible.

In summary, there exist numerous different data sets and models that allow for examining the performance of CSL methods within their specific scenarios. Apart from that, they do not allow the generation of observational data with varying complexity needed within a comprehensive accuracy examination of CSL. Especially, as this requires to vary the model complexity, e.g., concerning the number of considered variables, the ratio of discrete nodes, or the number of possible discrete values, cf. (R2). Moreover, currently, no common framework provides a functional formalization that ensures interpretability cf. (R1). In contrast, we aim to establish a flexible yet well-defined and unified ground-truth model, allowing the generation of observational data under various evaluation perspectives.

3 The Mixed Additive Noise Model (MANM)

In the following, we introduce the MANM as a framework for generating causal structures with mixed discrete-continuous and nonlinear relationships. In Sec. 3.1, we define the MANM in its functional form, cf. (R1), and provide some exemplary distribution models from continuous (Sec. 3.2), discrete (Sec. 3.3), and mixed discrete-continuous space (Sec. 3.4).

3.1 MANM for Modeling Mixed Discrete-Continuous and Nonlinear Data

In general, we say that variables $V_i, i = 1, \dots, p$, (cf. Sec. 1.1) of a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ are discrete if they have a (finite) discrete domain, i.e., where $V_i : \Omega \to Z_i \subseteq \mathbb{R}$ with countable subset Z_i, or continuous if they have a continuous domain, i.e., where $V_i : \Omega \to \mathbb{R}$, such that all V_i have Lebesgue measurable domains in $\mathcal{A}$. Modeling causal relationships requires to define a FCM that generates a variable $V_i \in \mathbf{V}$ according to the sets of possible discrete or continuous parents of V_i in $\mathcal{G}$, denoted by $\mathcal{P}^{dis}(V_i)$ and $\mathcal{P}^{con}(V_i)$, respectively. Therefore, we introduce the mixed additive

noise model (MANM) with mutually independent noise N_i where $N_i \perp\!\!\!\perp V_j$ for all $V_j \in \mathbf{V}$ as

$$V_i = \sum_{V_j \in \mathcal{P}^{dis}(V_i)} f_{j,i}(V_j) + \sum_{V_k \in \mathcal{P}^{con}(V_i)} f_{k,i}(V_k) + N_i, \text{ for all } V_i \in \mathbf{V}, \tag{1}$$

with functions $f_{j,i} : Z_j \to Z_i$ and $f_{k,i} : \mathbb{R} \to Z_i$ if V_i has a discrete domain Z_i, or $f_{j,i} : Z_j \to \mathbb{R}$ and $f_{k,i} : \mathbb{R} \to \mathbb{R}$ if V_i has a continuous domain. Moreover, we require that the independent noise variable N_i either is a continuous distributed random variable, e.g., $N_i \sim \mathcal{N}(0, 1)$, or discrete distributed over Z_i with $\mathbf{P}(N_i = 0) \geq \mathbf{P}(N_i = k)$ for all $k \in Z_i$ with $k \neq 0$ if V_i is continuous or discrete, respectively. Therefore, the proposed MANM of (1) extends the well-known "simple" data generating causal models and incorporates recent work on causal discovery in the two-variable model, which either are defined within a fully continuous space, e.g., cf. [9, 37, 46], or a fully discrete space, e.g., cf. [28, 29]. While we focus on CSL from purely observational data, the modeling of causal structures via the MANM allows for the generation of interventional data as described by [27] as well, e.g., see [8].

3.2 Scenario 1: MANM within the Continuous Space

In the following scenario, we provide intuitive and relatively well-established examples of the MANM within the continuous domain, i.e., $\mathcal{P}^{dis}(V_i) = \emptyset$ for all $V_i \in \mathbf{V}$.

Linear Additive Noise Models Given that $f_{k,i} : \mathbb{R} \to \mathbb{R}$ in (1) is linear, i.e., $f_{k,i}(x) = \beta_{k,i}x$, the MANM reduces to the most common form of FCMs [6]. In particular, when the additive noise N_i is i.i.d. standard Gaussian distributed such that (1) reduces to $V_i = \sum_{V_k \in \mathcal{P}^{con}(V_i)} \beta_{k,i} V_k + N_i$ with i.i.d. $N_i \sim \mathcal{N}(0, 1)$ for all $V_i \in \mathbf{V}$. Then, $\mathbf{V} = \{V_1, \ldots, V_p\}$ is multivariate Gaussian distributed with mean zero and covariance matrix $\Sigma = (I_p - \mathcal{B})^{-1}(I_p - \mathcal{B})^{-1}$, where I_p is the $p \times p$ identity matrix and $\mathcal{B}$ the $p \times p$ weighted adjacency matrix with non-zero entries $\mathcal{B}_{i,j} = \beta_{j,i}$ if there is an edge $V_j \to V_i$. Multivariate Gaussianity allows constraint-based methods to infer conditional independencies within $\mathbf{V}$ by testing for zero partial correlation, which makes CSL feasible for sparse graphs with up to thousands of variables, e.g., cf. [12]. Due to the applicability to high-dimensional settings, the linear Gaussian model has found wide use in systems biology, e.g., to infer gene regulatory networks from observational gene expression data [40].

Nonlinear Additive Noise Models While linear models are well understood and easy to work with, causal structures within many real-world scenarios are not necessarily linear [6]. Therefore, several CSL methods consider nonlinear FCMs of the form $V_i = \sum_{V_k \in \mathcal{P}^{con}(V_i)} f_{k,i}(V_k) + N_i$ for all $V_i \in \mathbf{V}$, where $f_{k,i} : \mathbb{R} \to \mathbb{R}$ is not required to be linear, cf., e.g., [9, 41, 45].

3.3 Scenario 2: MANM within the Discrete Space

In the discrete case, a functional relationship $f_{j,i} : Z_j \to Z_i$ provides an "interpretable" formalization of causal structures and enables generating observational and interventional data. Therefore, we consider that all variables $\mathbf{V}$ have a discrete domain, i.e., $\mathcal{P}^{con}(V_i) = \emptyset$ for all $V_i \in \mathbf{V}$. Then, causal relationships between discrete variables can be modeled in two different ways, cf. [28, 29]: First, V_i has the domain $Z_i = \mathbb{Z}$ with support $supp(V_i)$, such that the MANM can be defined analogously to the continuous case. Second, V_i has the domain $Z_i \subset \mathbb{Z}$, which allows to define $+$ as addition within the respective modulo ring $\mathbb{Z}/m_i\mathbb{Z}$, where $m_i = |supp(V_i)|$.

Integer Additive Noise Models Let $V_i : \Omega \to \mathbb{Z}$, $V_i \in \mathbb{V}$, be a discrete random variable with (maybe finite) support $supp(V_i)$. In this scenario, the MANM of (1) reduces to $V_i = \sum_{V_j \in \mathcal{P}^{dis}(V_i)} f_{j,i}(V_j) + N_i$, for all $V_i \in \mathbf{V}$, with a function $f_{j,i} : \mathbb{Z} \to \mathbb{Z}$ and mutually independent noise N_i such that $\mathbf{P}(N_i = 0) \geq \mathbf{P}(N_i = k)$ for all $k \in Z_i$ with $k \neq 0$. Note that $f_{j,i}$ can be any probabilistic or deterministic assignment from $\mathbb{Z}$ to $\mathbb{Z}$.

For illustration, consider the two variable case $V_2 = f_{1,2}(V_1) + N_2$ with the following simplified example adapted from [28]. Let V_1 be uniformly distributed over $\{-2, -1, 0, 1, 2\}$ and let N_2 be characterized by $\mathbf{P}(N_2 = -2) = \mathbf{P}(N_2 = 2) = 0.05$, and $\mathbf{P}(N_2 = -1) = \mathbf{P}(N_2 = 0) = \mathbf{P}(N_2 = 1) = 0.3$. Then, $f_{1,2}(x)$ can either be deterministic, e.g., $f_{1,2}(x) = \lceil 0.5\, x^2 \rceil$ or probabilistic, e.g.,

$$f_{1,2}(x) = \begin{cases} Binomial(0.8, 2), & \text{if } x \in \{-2, 2\} \\ Binomial(0.5, 2), & \text{if } x \in \{-1, 1\} \\ Binomial(0.2, 2), & \text{if } x \in \{0\}. \end{cases} \tag{2}$$

Cyclic Additive Noise Models Following the idea of [28], we consider the concept of m-cyclic random variables. Therefore, let $V_i : \Omega \to Z_i = \mathbb{Z}/m_i\mathbb{Z}$, i.e., taking values in $\{0, \dots, m_i - 1\}$, such that the MANM incorporates functions $f_{j,i} : \mathbb{Z}/m_j\mathbb{Z} \to \mathbb{Z}/m_i\mathbb{Z}$. Contrary to the integer additive noise model (ANM), this scenario bounds the values each variable V_i can take to be from $\{0, \dots, m_i - 1\}$, i.e., to targeted domain $\mathbb{Z}/m_i\mathbb{Z}$.

For illustration, again consider the two-variable case $V_1 \to V_2$ with V_1 taking values $\{0, 1\}$, i.e., $V_1 : \Omega \to \mathbb{Z}/2\mathbb{Z}$, and $V_2 : \Omega \to \mathbb{Z}/3\mathbb{Z}$. Let $V_1 \sim Bernoulli(0.75)$ and N_2 be characterized by $\mathbf{P}(N_2 = 0) = 0.5$, $\mathbf{P}(N_2 = 1) = 0.3$ and $\mathbf{P}(N_2 = 2) = 0.2$. We then can define $V_1 \to V_2$ as $V_2 = f_{1,2}(V_1) + N_2$ with $f_{1,2} : \mathbb{Z}/2\mathbb{Z} \to \mathbb{Z}/3\mathbb{Z}$ as mapping $0 \mapsto 1$ and $1 \mapsto 2$. Moreover, the cyclic ANM enables categorical variables with discrete values that do not have any order, cf. [29].

3.4 Scenario 3: MANM within the Mixed Space (Continuous & Discrete)

When considering CSL from mixed discrete-continuous variables, mostly, a conditional linear Gaussian (CLG) is the assumed FCM [6]. While the CLG restricts discrete variables to have discrete parents only [1, 30], the MANM allows for both directions of causal relationships between discrete and continuous variables, e.g., following the augmented conditional linear Gaussian (ACLG).

Conditional Linear Gaussian Models First, we examine how the MANM enables to generate $\mathbf{V}$ being CLG distributed. Then, the MANM of (1) is given by

$$V_i = \begin{cases} \sum_{V_j \in \mathcal{P}^{dis}(V_i)} f_{j,i}(V_j) + \sum_{V_k \in \mathcal{P}^{con}(V_i)} \beta_{k,i} V_k + \mathcal{N}(\mu_i, \sigma_i), & \text{for continuous } V_i \in \mathbf{V} \\ \sum_{V_j \in \mathcal{P}^{dis}(V_i)} f_{j,i}(V_j) + N_i, & \text{for discrete } V_i \in \mathbf{V}, \end{cases} \tag{3}$$

where $f_{j,i}$ can be any functional assignment $f_{j,i} : \mathbb{Z} \to Z_i \subseteq \mathbb{Z} \subset \mathbb{R}$, continuous Gaussian noise $\mathcal{N}(\mu_i, \sigma_i)$, and the discrete noise term N_i as defined in Sec. 3.3. As all continuous variables are multivariate Gaussian by definition, all continuous V_i are CLG distributed given the vectors of realisations $\boldsymbol{v}^{dis}$ and $\boldsymbol{v}^{con}$ of the respective discrete and continuous parents $\mathcal{P}^{dis}(V_i)$ and $\mathcal{P}^{con}(V_i)$, i.e., we have $\mathbf{P}\big(V_i \mid \boldsymbol{v}^{dis}, \boldsymbol{v}^{con}\big) \sim \mathcal{N}\big(\sum_{v_j \in \boldsymbol{v}^{dis}} f_{j,i}(v_j) + \sum_{v_k \in \boldsymbol{v}^{con}} \beta_{k,i} v_k + \mu_i, \sigma_i\big)$, cf., e.g., [19].

Augmented Conditional Linear Gaussian Models To overcome the restrictions of the CLG models, [19] introduced the so-called augmented CLG model, in which discrete variables with continuous parents are generated by using the softmax function. Then, $f_{k,i} : \mathbb{R} \to \mathbb{Z}$ assigns a probability to each realization v_i within the support $supp(V_i)$, i.e., $v_i \in Z_i$ with $\mathbf{P}(V_i = v_i) > 0$, given the realization v_k of V_k. Therefore, let $f_{k,i}$ be given by the probabilistic mapping

$$f_{k,i} := \mathbf{P}(V_i = v_i | V_k = v_k) = \frac{exp(\alpha_{k,i} + \beta_{k,i} v_k)}{\sum_{s \in \{1,\dots,m_i - 1\} :\, v_s \in supp(V_i)} exp(\alpha_{k,s} + \beta_{k,s} v_k)} \tag{4}$$

with soft-max parameters $\alpha_{k,s}$ and $\beta_{k,s}$ defined for all $v_s \in supp(V_i)$ given the realization $V_k = v_k$. Then, the MANM of (1) allows for generation of data according to the ACLG model via

$$V_i = \begin{cases} \sum_{V_j \in \mathcal{P}^{dis}(V_i)} f_{j,i}(V_j) + \sum_{V_k \in \mathcal{P}^{con}(V_i)} \beta_{k,i} V_k + \mathcal{N}(\mu_i, \sigma_i), & \text{for continuous } V_i \in \mathbf{V} \\ \sum_{V_j \in \mathcal{P}^{dis}(V_i)} f_{j,i}(V_j) + \sum_{V_k \in \mathcal{P}^{con}(V_i)} f_{k,i}(V_k) + N_i, & \text{for discrete } V_i \in \mathbf{V}, \end{cases} \tag{5}$$

with $f_{j,i}$ as defined in the context of (3) and the discrete noise term N_i as defined in Sec. 3.3. In this context, the MANM is flexible enough to model causal structures within discrete-continuous mixtures by incorporating various deterministic and probabilistic mappings $f_{j,i}$ from $\mathbb{Z}$ to $\mathbb{R}$, and vice versa via $f_{k,i}$. For example, consider simple step functions for $f_{k,i} : \mathbb{R} \to \mathbb{Z}$.

In summary, the MANM provides a flexible functional framework to model causal structures with various characteristics of an edge $V_j \to V_i$, cf. (R1) - (R2).

4 `MANM-CS`: A CSL Benchmarking Framework

To provide the research community easy access to the MANM for benchmarking CSL methods, cf. (R4), we present our reference implementation `MANM-CS`[1] It generates mixed and nonlinear observational data (Sec. 4.1), allows for high-dimensional scenarios (Sec. 4.2), and covers the various distribution models included in the MANM (Sec. 4.3).

[1] https://github.com/hpi-epic/manm-cs

4.1 Implementation of Data Sampling Process

Data generation of `MANM-CS` follows the common two-step approach, where first, a DAG $\mathcal{G}$ with respective parameterized FCM is generated, and second, each observation is sampled by iterating over the nodes considering the functional relationships regarding their parents. Therefore, several parameters (see Table 2 in Appendix) provide easy specification of the underlying MANM introduced in Sec. 3 as basis for generation of mixed and nonlinear data. In Algorithm 1 (in Appendix), `MANM-CS` generates a DAG that incorporates `num_nodes` number of ordered nodes with edge density `edge_density`. Moreover, nodes are chosen to be discrete with a number of classes between $\texttt{discrete\_class}_{\mathbf{min}}$ and $\texttt{discrete\_class}_{\mathbf{max}}$ or continuous distributed according to `discrete_ratio`. If the joint distribution is conditional Gaussian the first `discrete_ratio` $\times$ `num_nodes` are discrete, otherwise for augmented conditional Gaussian each variable is chosen to be discrete with probability `discrete_ratio`. According to the MANM noise terms for discrete and continuous variables are chosen to be discrete with corresponding `discrete_noise_ratio` or normal distributed with standard deviation `std_continuous_noise`, respectively.

Moreover, functional relationships for each edge in $\mathcal{G}$ are either sampled from self-chosen `functions` within the continuous space (see Sec. 3.2), follow the cyclic additive noise model within discrete (see Sec. 3.3), or defined as a soft-max and CLG model within the mixed space, respectively (see Sec. 3.4). Hence, Algorithm 1 returns a fully parameterized CGM following the specifications of the MANM as basis for Algorithm 2 (in Appendix). In particular, Algorithm 2 implements the sampling of `num_samples` observations by iterating over the nodes considering the noise terms and functional relationships regarding their parents.

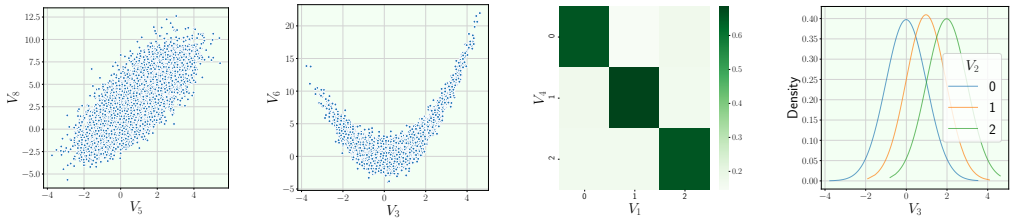
4.2 Runtime Performance of Data Generation Process

To speed-up data generation within high-dimensional scenarios, `MANM-CS`' data sampling can be executed in parallel by specifying the number of processes `num_processes`. Although ideal speed-up is not achieved, parallelization reduces data generation significantly - in particular for many nodes. For example, the execution time of one million samples generated according to a DAG with 1 000 nodes and edge density 0.4 decreases from 4 322 seconds (16 cores), over 2 214 seconds (32 cores) to 1 367 seconds (64 cores) .

4.3 Exemplary Characteristics of Data Generated by `MANM-CS`

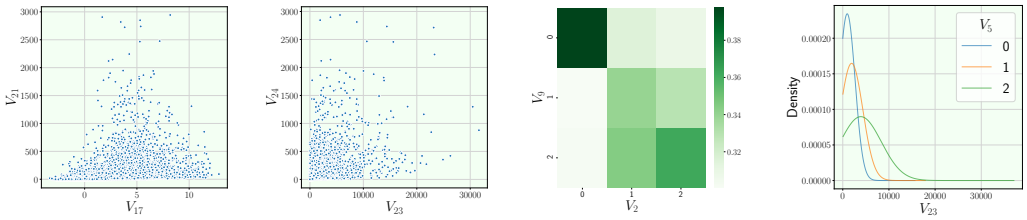
The following examples illustrate the range of causal relationships and respective data characteristics.

Example 1 *First, we consider a small and sparse CGM $\mathcal{G}$ such that the distributional characteristics of direct causal relationships are primarily induced through the corresponding functional mappings of the underlying MANM. In this sense, we consider a mixed CGM (*`conditional_gaussian`=1, `discrete_ratio`=0.5*) of* `num_nodes`=10 *with* `edge_density`=0.4 *that includes discrete variables (*$\texttt{discrete\_class}_{\mathbf{min}}$=3, $\texttt{discrete\_class}_{\mathbf{max}}$=4*) and continuous variables with nonlinear causal relationships (*`functions`={[0.4, `linear`], [0.3, `quadratic`], [0.3, `cosine`]}*) and corresponding noise terms (*`discrete_ratio`=0.5, `std_continuous_noise`=1.0*). On this basis, the data characteristics of* `num_samples`=10 000 *depicted in Fig. 1 follow the expected evidently linear, quadratic, discrete, and mixed causal relationships of direct edges within the sparse CGM $\mathcal{G}$.*



(a) Linear $V_5 \to V_8$ (b) Quadratic $V_3 \to V_6$ (c) Discrete $V_1 \to V_4$ (d) Mixed $V_2 \to V_3$

Figure 1: Unconfounded data distributions: (a) scatter plots for the linear edge $V_5 \to V_8$, (b) the nonlinear edge $V_3 \to V_6$, (c) a heatmap of conditional probabilities for the discrete edge $V_1 \to V_4$, and (d) a density plot of V_3 given the realization of a discrete parent V_2 for a mixed edge $V_2 \to V_3$.



(a) Linear $V_{17} \to V_{21}$ (b) Cosine $V_{23} \to V_{24}$ (c) Discrete $V_2 \to V_9$ (d) Mixed $V_5 \to V_{23}$

Figure 2: Confounded data distributions: (a) scatter plots for the linear edge $V_{17} \to V_{21}$, (b) the nonlinear edge $V_{23} \to V_{24}$, (c) a heatmap of conditional probabilities for the discrete edge $V_2 \to V_9$, and (d) a density plot of V_{23} given the realization of a discrete parent V_5 for a mixed edge $V_5 \to V_{23}$.

Example 2 *Next, we consider a larger and denser CGM $\mathcal{G}$ such that the distributional characteristics of direct causal relationships may be distorted through common confounders. Therefore, we change Example 1 by increasing* `num_nodes` =25 *and* `edge_density` =0.8 *while retaining all other parameters. On this basis, the data distributions of direct causal relationships depicted in Fig. 2 are now characterized by interferences of respective direct linear, cosine, discrete or mixed causal relationships with indirect causal relationships induced through confounders within a denser CGM.*

Therefore, these examples not only illustrate the achieved interpretability of a causal relationship based upon a well-defined FCM, cf. (R1), but also demonstrate the achievable complexity of data characteristics provided by `MANM-CS`. Note, the contrary interference on the data distribution between edges with no direct edge through common confounders may induce a visible but not existing direct functional relationship. Moreover, note, that variations of the noise parameters `discrete_noise_ratio` and `std_continuous_noise` yield further statistical dispersion.

5 Benchmarking Scenarios and Experimental Evaluation

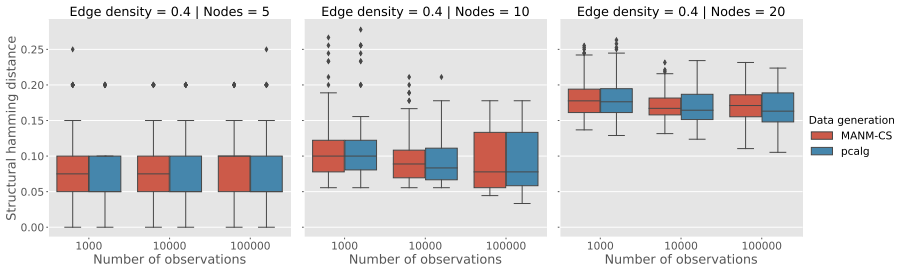
In this section, we demonstrate that `MANM-CS` not only covers common benchmarking approaches (Sec. 5.1), but allows for a more comprehensive examination of CSL methods, too (Sec.5.2).

5.1 Experiment 1: CSL in Comparison to Well-known Benchmark Approaches

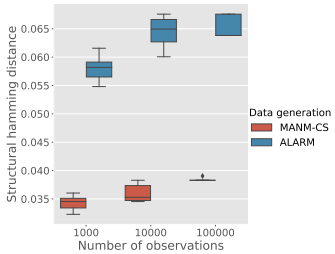
In this experiment, we compare large sample properties of the well-known PC algorithm [40] with appropriate CI tests on observational data generated by `MANM-CS` that aims to mimic data sampled from common type (II) and (III) approaches, cf. Sec. 2. In particular, we show a coincidence in improvements of the structural Hamming distance (SHD) regarding the learned CGMs from data with increasing sample size in the context of different distribution models, cf. Sec. 3.2 - 3.4.

Within continuous space, the FCM of the linear ANM (see Sec. 3.2) allows for easy data generation following approach (III), e.g., within `pcalg` [13]. Given that the MANM is a more general model class, the SHD of CGMs learned by the PC algorithm with Fisher's z-test shows a direct coincidence on multivariate Gaussian data sampled by `MANM-CS` and `pcalg` for variations in the number of variables as well as in the number of observations, see Fig. 3 (a).

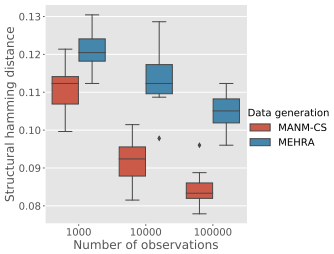
Within the mixed and discrete space, benchmarking of CSL methods are often restricted to data generated by "static" type (II) approaches, e.g., parameterized models found in the `bnlearn` repository [33]. For comparison, we generate observational data using the `MANM-CS`' capabilities for parameter adjustment to mimic the characteristics of the well-known `ALARM` [2] and `MEHRA` [44] networks. In this context, the descriptive functional restriction on modeling a causal relationship between two discrete variables within `MANM-CS` becomes recognizable. For example, `ALARM` incorporates probabilistic mappings $f_{k,i}$ between two discrete nodes V_i, V_j while `MANM-CS`'s implementation is currently restricted to the mapping of the cyclic ANM, cf. Sec. 3.3. Hence, the induced independence characteristics of the variables' joint distribution are empirically stronger within data sampled by `MANM-CS`, which yield lower SHDs in comparison to the parameterized discrete `ALARM` and CLG `MEHRA` networks, respectively, see Fig. 3 (b) and (c). Nevertheless, the coincidence in a decreasing SHD of the CGMs learned by the PC algorithm with appropriate Pearson's X^2 (discrete, cf. [33, 13]) and asymptotic mutual information χ^2 test (CLG, cf. [33]) for an increasing number of samples is visible for all approaches and distribution models.



(a) Continuous space: `MANM-CS` and `pcalg`



(b) Discrete space: `MANM-CS` and `ALARM`



(c) Mixed space: `MANM-CS` and `MEHRA`

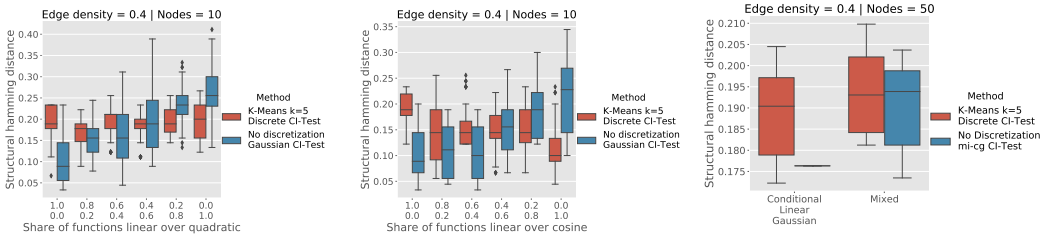
Figure 3: Median SHD (10 runs) of learned CGMs with the PC algorithm from (a) observational data sampled by `MANM-CS` or `pcalg` for the continuous space with a varying number of nodes, or from (b) the Bayesian benchmark networks `ALARM` for the discrete and (c) `MEHRA` for the mixed space.

5.2 Experiment 2: CSL within Mixed and Nonlinear Space

In this experiment, we demonstrate `MANM-CS`'s capabilities for benchmarking CSL methods in the context of various scenarios regarding mixed discrete-continuous and nonlinear data. In particular, we examine the decreasing accuracy of CLG model-based CSL when assumptions on the causal relationships, e.g., linearity, are violated. Following [30], we include a discretization-based approach as nonparametric baseline but consider distribution models beyond simple CLG or Lee Hastie data. Therfore, we compare the median SHD (10 runs) of learned CGMs with the PC algorithm, cf. [13, 33], with asymptotic mutual information χ^2 test assuming a CLG model, cf. [33], compared to Pearson's X^2 test, cf. [13], where continuous variables are discretized through the k-means algorithm, cf. [21], with $k = 5$.

Violating Linearity We consider a violation of linearity within the CLG and its implication on the accuracy of learned CGMs. In particular, Fig. 4 (a) and (b) depict the median SHD of the parametric against the discretization-based approach for an increasing ratio of quadratic and cosine functional relationships, respectively (`num_nodes` =10, `edge_density` =0.4, `discrete_ratio` =0, `num_samples`=10 000). As the asymptotic mutual information χ^2 test is based upon the partial correlation within continuous space, its good accuracy within the purely linear case decreases steadily for an increasing ratio on nonlinear functional relationships. In contrast, the discretization-based approach's accuracy behaves rather invariant in the context variations in terms of nonlinearity for both cases (a) and (b). Moreover, although not true generally, cf. [23, 30], the accuracy of the discretization-based approach even exceeds the parametric CLG-based approach in the presented edge cases of mainly quadratic or cosine functional relationships.

Violating Conditional Gaussianity We consider the implication of changing from the CLG to another mixed model (see Sec. 3.4). In this sense, Fig. 4 (c) depicts the median SHD of the parametric CLG-based against the discretization-based approach under the assumption of an CLG and an augmented CLG model (`num_nodes` =50, `edge_density` =0.4, `discrete_ratio` =0.5, `num_samples`=10 000). Although the PC algorithm with an appropriate asymptotic mutual information χ^2 shows a significantly better accuracy within the CLG model, the accuracy is slightly exceeded by the discretization-based approach if discrete nodes are allowed to have continuous parents, i.e., within the augmented CLG model.



(a) Linear to quadratic. (b) Linear to cosine. (c) CLG to mixed.

Figure 4: Median SHD (10 runs) of learned CGMs with the PC algorithm with asymptotic mutual information χ^2 test assuming a CLG model compared to Pearson's X^2 test where continuous variables are discretized (k-means with $k = 5$) given violation of the CLG model assumption for an increasing ratio of (a) quadratic and (b) cosine functional relationships, respectively, and (c) for considering the augmented CLG model.

The above examples demonstrate the importance of validating CSL methods assumptions in practice and the demand for understanding the method's accuracy within specific edge cases, e.g., when assumptions on the causal relationships are violated, cf. [17].

6 Conclusion, Limitations, and Future Work

We introduced the mixed additive noise model (MANM) to provide a framework for generating causal structures within mixed discrete-continuous and nonlinear data. Its functional formalization defined in (1) provides an interpretable characterization of causal structures as demonstrated from a theoretical and empirical perspective (see Sec. 3.2 - 3.4 and Sec. 5, respectively), cf. (R1). In particular, it connects well-established work of CSL within different distribution models, such as CLG, and allows for the generation of continuous, discrete, and mixed discrete-continuous observational data, cf. (R2). Due to the functional form, the MANM is flexible enough to support further extensions, e.g., to consider the generation of interventional data similar to [8], cf. (R3). Moreover, it allows examining methods' accuracy in case of a misspecified choice of hyperparameters or given invalidated assumptions, e.g., using an incomplete selection of $\mathbf{V}$ to model the causally insufficient case of latent variables. To provide easy access to the research community, we present our reference implementation `MANM-CS` and benchmarking scenarios, cf. (R4). In particular, `MANM-CS`'s capabilities not only provide enough opportunities to mimic common benchmarking approaches but also allows for more comprehensive evaluations with varying model complexity that exceed "static" type (II) approaches (see Sec. 5). Further, `MANM-CS` can be easily integrated in pipelines for CSL such as `MPCSL` [11], which allows researchers and practitioners to easily evaluate their methods.

While the restriction on the FCM allows for a formalization of various causal relationships, the functional constraints induce limitations that are worth to be noticed. For example, in contrast to the assumptions of the MANM within the discrete space, cf. Sec. 3.3, there may not always be a functional representation of a causal relationship between discrete variables in practice [29], cf. Sec. 5.1. Moreover, the embedding of discrete variables into the continuous space as defined in (1) restricts the functional relationship to be location-related, which may be violated within real-world scenarios. In this context, possible generalizations are weakened additivity concerning the independent noise or a post nonlinearity [9, 46]. Note that if the characteristics of the data generating mechanism do not follow the MANM, the requirements of CSL methods should be general enough to reveal the data generating processes approximately [6].

As the MANM provides a ground truth model for generating observational data following various distribution models with nonlinear and mixed discrete-continuous data, we work on a more comprehensive empirical evaluation of several popular CSL methods for future work. Moreover, we aim to provide more parameters such as concerning missing values or interventional data, different noise distributions, and functional classes to enable a more fine-grained evaluation of CSL methods. Further, we invite the research community to participate in the extension of `MANM-CS` actively.

References

[1] Bryan Andrews, Joseph Ramsey, and Gregory F. Cooper. Scoring Bayesian networks of mixed variables. *International Journal of Data Science and Analytics*, 6(1):3–18, Aug 2018. ISSN 2364-4168. doi: 10.1007/s41060-017-0085-7. URL https://doi.org/10.1007/s41060-017-0085-7.

[2] Ingo A. Beinlich, H. J. Suermondt, R. Martin Chavez, and Gregory F. Cooper. The ALARM Monitoring System: A Case Study with two Probabilistic Inference Techniques for Belief Networks. In Jim Hunter, John Cookson, and Jeremy Wyatt, editors, *AIME 89*, pages 247–256, Berlin, Heidelberg, 1989. Springer Berlin Heidelberg. ISBN 978-3-642-93437-7.

[3] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive Probabilistic Networks with Hidden Variables. *Machine Learning*, 29(2):213–244, Nov 1997. ISSN 1573-0565. doi: 10.1023/A:1007421730016. URL https://doi.org/10.1023/A:1007421730016.

[4] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning High-Dimensional Directed Acyclic Graphs with Latent and Selection Variables. *Annals of Statistics*, 40:294–321, 2012. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/41713636.

[5] Cristina Conati, Abigail S Gertner, Kurt VanLehn, and Marek J Druzdzel. On-Line Student Modeling for Coached Problem Solving Using Bayesian Networks. In *Proceedings of the Sixth International Conference on User Modeling*, pages 231–242. Springer, 06 1997. ISBN 978-3-211-82906-6. doi: 10.1007/978-3-7091-2670-7_24.

[6] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in genetics*, 10:524, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00524.

[7] Alex Greenfield, Aviv Madar, Harry Ostrer, and Richard Bonneau. DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. *PLOS ONE*, 5(10):1–14, 10 2010. doi: 10.1371/journal.pone.0013397. URL https://doi.org/10.1371/journal.pone.0013397.

[8] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. Causal Structure Learning. *Annual Review of Statistics and Its Application*, 5(1):371–391, 2018. doi: 10.1146/annurev-statistics-031017-100630. URL https://doi.org/10.1146/annurev-statistics-031017-100630.

[9] Patrik O. Hoyer, Dominik Janzing, Joris Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear Causal Discovery with Additive Noise Models. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, pages 689–696, Red Hook, NY, USA, 2008. Curran Associates Inc. ISBN 9781605609492.

[10] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized Score Functions for Causal Discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*, KDD '18, pages 1551–1560, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220104. URL https://doi.org/10.1145/3219819.3220104.

[11] Johannes Huegle, Christopher Hagedorn, Michael Perscheid, and Hasso Plattner. MPCSL - A Modular Pipeline for Causal Structure Learning. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 2247—-2257, New York, NY, USA, 2021. Association for Computing Machinery.

[12] Markus Kalisch and Peter Bühlmann. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8:613–636, May 2007. ISSN 1532-4435. URL http://jmlr.org/papers/v8/kalisch07a.html.

[13] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software, Articles*, 47(11):1–26, 2012. ISSN 1548-7660. doi: 10.18637/jss.v047.i11. URL https://www.jstatsoft.org/v047/i11.

[14] Yutaka Kano and Shohei Shimizu. Causal Inference Using Nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, 2003.

[15] Kevin B. Korb and Ann E. Nicholson. *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., USA, 2nd edition, 2010. ISBN 1439815917.

[16] Steffen L. Lauritzen and David J. Spiegelhalter. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224, 1988. ISSN 00359246. URL http://www.jstor.org/stable/2345762.

[17] Andrew R. Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*, 2020.

[18] Jason D. Lee and Trevor J. Hastie. Learning the Structure of Mixed Graphical Models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015. doi: 10.1080/10618600.2014.900500. URL https://doi.org/10.1080/10618600.2014.900500.

[19] Uri Lerner, Eran Segal, and Daphne Koller. Exact Inference in Networks with Discrete Children of Continuous Parents. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 319—-328, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

[20] Li Liu, Michael I. Mishchenko, and W. Patrick Arnott. A study of radiative properties of fractal soot aggregates using the superposition t-matrix method. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 109(15):2656–2663, 2008. ISSN 0022-4073. doi: https://doi.org/10.1016/j.jqsrt.2008.05.001. URL https://www.sciencedirect.com/science/article/pii/S002240730800112X.

[21] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[22] Alessandro Magrini, Stefano Di Blasi, and Federico Mattia Stefanini. A conditional linear Gaussian network to assess the impact of several agronomic settings on the quality of Tuscan Sangiovese grapes. *Biometrical Letters*, 54(1):25–42, 2017. doi: doi:10.1515/bile-2017-0002. URL https://doi.org/10.1515/bile-2017-0002.

[23] Daniel Malinsky and David Danks. Causal Discovery Algorithms: A Practical Guide. *Philosophy Compass*, 13(1):e12470, 2018. doi: https://doi.org/10.1111/phc3.12470. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12470.

[24] Daniel Marbach, James C. Costello, Robert Küffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, Manolis Kellis, James J. Collins, Andrej Aderhold, Gustavo Stolovitzky, and et al. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9 (8):796–804, 8 2012. ISSN 1548-7091.

[25] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. URL http://jmlr.org/papers/v17/14-518.html.

[26] Rainer Opgen-Rhein and Korbinian Strimmer. From Correlation to Causation Networks: a Simple Approximate Learning Algorithm and its Application to High-Dimensional Plant Gene Expression Data. *BMC Systems Biology*, 1(1):1–10, Aug 2007. ISSN 1752-0509. doi: 10.1186/1752-0509-1-37.

[27] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 1st edition, 2000. ISBN 0521773628.

[28] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Identifying Cause and Effect on Discrete Data using Additive Noise Models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 597–604, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/peters10a.html.

[29] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal Inference on Discrete Data using Additive Noise Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, December 2011. doi: 10.1109/TPAMI.2011.71.

[30] Vineet K. Raghu, Allen Poon, and Panayiotis V. Benos. Evaluation of causal structure learning methods on mixed data types. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, volume 92 of *Proceedings of Machine Learning Research*, pages 48–65, London, UK, 20 Aug 2018. PMLR. URL http://proceedings.mlr.press/v92/raghu18a.html.

[31] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308 (5721):523–529, 2005. ISSN 0036-8075. doi: 10.1126/science.1105809. URL https://science.sciencemag.org/content/308/5721/523.

[32] Juliane Schäfer and Korbinian Strimmer. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. doi: doi:10.2202/1544-6115.1175. URL https://doi.org/10.2202/1544-6115.1175.

[33] Marco Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35:1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v035.i03. URL https://www.jstatsoft.org/v035/i03.

[34] Marco Scutari. Bayesian Network Repository, 2012. URL http://www.bnlearn.com/bnrepository.

[35] Marco Scutari, Phil Howell, David J Balding, and Ian Mackay. Multiple Quantitative Trait Analysis Using Bayesian Networks. *Genetics*, 198(1):129–137, 2014. ISSN 0016-6731. doi: 10.1534/genetics.114.165704. URL https://www.genetics.org/content/198/1/129.

[36] Andrew J Sedgewick, Ivy Shi, Rory M Donovan, and Panayiotis V Benos. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*, 17(5):307–318, Jun 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1039-0. URL https://doi.org/10.1186/s12859-016-1039-0.

[37] 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: 2003—2030, December 2006. ISSN 1532-4435. URL http://jmlr.org/papers/v7/shimizu06a.html.

[38] Karamjit Singh, Garima Gupta, Vartika Tewari, and Gautam Shroff. Comparative Benchmarking of Causal Discovery Algorithms. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '18, pages 46–56, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450363419. doi: 10.1145/3152494.3152499. URL https://doi.org/10.1145/3152494.3152499.

[39] Peter Spirtes, Christopher Meek, and Thomas Richardson. An Algorithm for Causal Inference in the Presence of Latent Variables and Selection Bias. In *Computation, Causation, and Discovery*. AAAI Press, 05 1999. ISBN 9780262315821. doi: 10.7551/mitpress/2006.003.0009. URL https://doi.org/10.7551/mitpress/2006.003.0009.

[40] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 2000. ISBN 978-0-262-19440-2.

[41] Robert E. Tillman, Arthur Gretton, and Peter Spirtes. Nonlinear Directed Acyclic Structure Learning with Weakly Additive Noise Models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 1847–1855, Red Hook, NY, USA, 2009. Curran Associates Inc. ISBN 9781615679119.

[42] Michail Tsagris, Giorgos Borboudakis, Vincenzo Lagani, and Ioannis Tsamardinos. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics*, pages 19–30, Feb 2018. ISSN 2364-4168. doi: 10.1007/s41060-018-0097-y. URL `https://doi.org/10.1007/s41060-018-0097-y`.

[43] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian Network Structure Learning Algorithm. *Machine learning*, 65(1):31–78, Oct 2006. ISSN 1573-0565.

[44] Claudia Vitolo, Marco Scutari, Mohamed Ghalaieny, Allan Tucker, and Andrew Russell. Modeling Air Pollution, Climate, and Health Data Using Bayesian Networks: A Case Study of the English Regions. *Earth and Space Science*, 5(4):76–88, 2018. doi: https://doi.org/10.1002/2017EA000326. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017EA000326`.

[45] Kun Zhang and Aapo Hyvärinen. Causality Discovery with Additive Disturbances: An Information-Theoretical Perspective. In *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECMLPKDD'09, pages 570–585, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 3642041736.

[46] Kun Zhang and Aapo Hyvärinen. On the Identifiability of the Post-Nonlinear Causal Model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 647—655, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Appendix

Parameters for Data Generation

`MANM-CS` currently enables to generate observational data for benchmarking CSL in the context of mixed discrete-continuous and nonlinear causal relationships according to the parameters depicted in Table 2.

Parameter	Definition
`num_nodes`	Number of nodes corresponding to the number of variables p in $\mathbf{V}$
`edge_density`	Edge density of the resulting graph in range $[0, 1]$
`discrete_ratio`	Ratio of discrete nodes compared to `num_nodes`
`num_samples`	Defines the number of samples that shall be generated from the DAG.
`discrete_noise_ratio`	The probability of adding noise to a discrete node
$\texttt{discrete\_class}_{\min}$	Minimum of the range for the discrete domain $\mathbb{Z}$ of discrete nodes
$\texttt{discrete\_class}_{\max}$	Maximum of the range for the discrete domain $\mathbb{Z}$ of discrete nodes
`std_continuous_noise`	Standard deviation of continuous noise
`functions`	A list of sample probabilities and continuous functions.
`conditional_gaussian`	Flag for conditional or augmented conditional Gaussian model
`num_processes`	Number of processes used for data sampling

Table 2: Parameters and their definitions for the data generation of `MANM-CS`.

Graph Generation Algorithm

Algorithm 1 of MANM-CS generates a DAG $\mathcal{G}$ with respective parameterized FCM.

Algorithm 1 Graph Generation
Input: num_nodes, edge_density, discrete_ratio, $\texttt{discrete\_class}_{\mathbf{min}}$, $\texttt{discrete\_class}_{\mathbf{max}}$, std_continuous_noise, discrete_noise_ratio, functions, conditional_gaussian
Output: DAG $\mathcal{G}$

procedure DIS(discrete_class$_{\mathbf{min}}$, discrete_class$_{\mathbf{max}}$,discrete_noise_ratio,par)
    discrete_class$\leftarrow$ randomUniform(discrete_class$_{\mathbf{min}}$, discrete_class$_{\mathbf{max}}$)
    noise $\leftarrow$ sampleNoise(discrete_noise_ratio, discrete_class)
    return discreteNode(discrete_class, noise, par)
end procedure

procedure CON(std_continuous_noise, functions, par)
    for pcon in continuousParents(parents) **do**
        sampled_functions $\leftarrow$ choose_function(functions,pcon)
    end for
    noise $\leftarrow \mathcal{N}(0,$ std_continuous_noise$)$
    return continuousNode(sampled_functions, noise, par)
end procedure

node_list $\leftarrow [1 \ldots$ num_nodes $]$
forward_edge_list $\leftarrow [(node_u, node_v)$ in node_list $\times$ node_list $\mid node_u < node_v]$
edge_list $\leftarrow$ sampleEdges(forward_edge_list, edge_density)
num_discrete_nodes $\leftarrow$ discrete_ratio $\cdot$ num_nodes

for node in topSorted(node_list) **do**
    parents $\leftarrow$ getParents(edge_list)
    if conditional_gaussian **then** ▷ Conditional Gaussian
        if node.id $\leq$ num_discrete_nodes **then** ▷ Create node of discrete type
            node.set(DIS(discrete_class$_{\mathbf{min}}$, discrete_class$_{\mathbf{max}}$,
            discrete_noise_ratio,parents))
        else ▷ Create node of continuous type
            node.set(CON(std_continuous_noise, functions, par))
        end if
    else ▷ Augmented Conditional Gaussian
        if random(0,1) $\leq$ discrete_ratio **then** ▷ Create node of discrete type
            node.set(DIS(discrete_class$_{\mathbf{min}}$, discrete_class$_{\mathbf{max}}$,
            discrete_noise_ratio,parents))
        else ▷ Create node of continuous type
            node.set(CON(std_continuous_noise, functions, par))
        end if
    end if
end for
return constructDAG(node_list, edge_list)

Data Generation Algorithm

Algorithm 2 samples each observation by iterating over the nodes of the DAG $\mathcal{G}$ provided by Algorithm 1 considering the functional relationships regarding their parents.

Algorithm 2 Data Generation
Input: `num_samples`, DAG $\mathcal{G}$
Output: `sampled_data_matrix`

```
sampled_data_matrix ← matrix(len(G.node_list) × num_samples)
for node in topSorted(G.node_list) do
    sampled_data_matrix[node.id] ← node.sample(num_samples)
end for
return sampled_data_matrix
```