(Horizon Europe Grant agreement ID: 101084642)

D3.2 A Prototype in Causal Discovery

Type: Demonstrator/Pilot/Prototype

CTU

January 8, 2024

Status: Live Document

Scheduled Delivery Date: 31/12/2023

Document History

- (December 31th, 2023) Version 1.0 Submitted to the EC and uploaded to CoDiet website.

1 Executive Summary

Learning causal models involves extracting meaningful relationships and dependencies between variables from observational or experimental data. This process often employs statistical and machine-learning techniques to infer causal structures and identify the directionality of causal links. Understanding causal models is crucial in various fields, from healthcare to economics, as it allows for accurate predictions, interventions, and a deeper understanding of complex systems.

We formulated learning of causal models as a non-convex operator-valued problem, without assumptions on the dimension of the hidden state. This follows recent progress in system identification based on an asymptotically convergent hierarchy of convexifications of a non-convex operator-valued problem, which is known as non-commutative polynomial optimization (NCPOP).

Here, we document our work on the prototype code in causal discovery at `https://github.com/codiet-eu/d32`, which is based on two research papers:

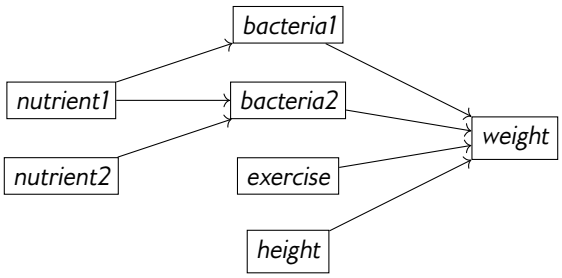
- Learning of Linear Dynamical Systems as a Non-Commutative Polynomial Optimization Problem, accepted in IEEE Transactions on Automatic Control, `https://doi.org/10.1109/TAC.2023.3313351`
- Joint Problems in Learning Multiple Dynamical Systems, submitted with a pre-print at `https://arxiv.org/abs/2311.02181`.

capturing joint work of teams at CTU (Xiaoyu He, Petr Rysavy, Jakub Marecek) and ICL (Quan Zhou, Mengjia Niu). The work has also benefitted from the insights of the teams at NKUA (Dimitrios Gunopulos, Vana Kalogeraki, Kleopatra Markou) and Technion (Shie Mannor, Mark Kozdoba). The purpose of this document is to provide further details in terms of motivation, related work, and results on benchmarks from causal learning.

2 Introduction

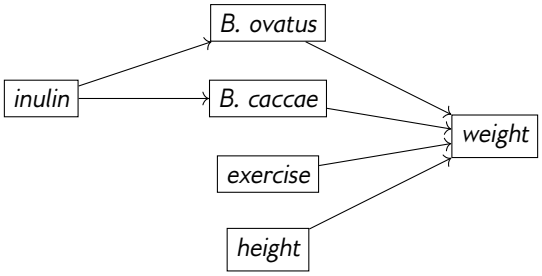
To motivate our work, let us consider an intentionally simplistic example:

Example 1. Bacteria *that live in our gut help us process food including* nutrient1 *and* nutrient1. *What metabolites become available in the blood stream (metabolome), depends on the composition of the population of bacteria (microbiome). In many settings, we could model the metabolome as a high-dimensional unobserved state. If one wishes to study the impact on an easily observable quantity such* weight, *one should like to consider confounders including* height *and the amount of* exercise. *This relationship can be represented by a directed acyclic graph (DAG) below.*

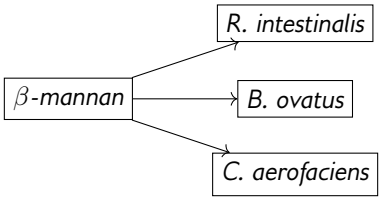


In practice, these relationships are much more complex and our goal is to infer quantitative aspects of such causal relationships from measurements of random variables, often available in the form of high-dimensional time series that are not sampled uniformly. Although some of the random variables are easily observed (e.g., weight), some others (e.g., related to the metabolome) need not be. Consider the following.

Example 2. *Metabolome depends not only on the diet and microbiome, but also the microbiome is affected by the metabolome and the diet. For example, inulin, a polysaccharide that is found in the cell walls of certain plants, promotes the growth of intestinal bacteria, which modulate the intake of energy from food. In the microbiome, bacteria such as* Bacteroides ovatus *and* Bacteroides caccae *compete for inulin and their prevalence depends on both their past prevalence, inulin levels, and concentrations of other dietary fiber, at least up to some level of inulin.*



Example 3. *A similar interaction network can be found in relation to β-mannan. The primary degrader of β-mannan is* Roseburia intestinalis, *together with others such as* Bacteroides ovatus. *For more details, see publication [15]. Therefore, a diet rich in β-mannans positively influences the growth of these two strains. In contrast, a fiber-free diet decreases their levels while promoting other bacteria, such as* Collinsella aerofaciens. *This can be crudely represented in the following causal network.*



Ideally, we would be able to make such causal models quantitative, not least to distinguish that β-mannan-rich diet promotes *Roseburia intestinalis*, while β-mannan-poor diet may promote *Collinsella aerofaciens*.

3 Related Work

First, we set our work in the context of related work, including a brief overview of *traditional* causal models pioneered by Pearl [20], which yield directed acyclic graphs such as in Examples 1–3, but without the quantitative aspects, and non-commutative polynomial optimization, pioneered by [23] and nicely surveyed by [4], which is our key technical tool.

Traditional Causal Models Learning causal models is traditionally phrased in terms of learning structural causal models (SCMs) or structural equations models (SEMs) including observational and intervention distributions, and a causal graph with the aim of reasoning about counterfactual scenarios. These elements serve as a powerful tool for understanding and formalising causal relationships in *some* complex systems. Current methodologies in SCMs employ a diverse range of techniques based on constraint-, function-, gradient- and score-based methods [20, 21, 11, 43]. With respect to methods, causal models are classified as independence-based, additive-noise, and invariant-prediction methods:

- Independence-based

  Bayesian networks [20], provide a graphical representation of dependencies among variables, which aids causal inference.

  - Constraint-based

    Pearl's Causal (PC) [20] algorithm, with its stable and parallel variants, utilises conditional independence tests to identify causal relationships within a network.

  - Function-based

    Direct linear non-Gaussian acyclic model (DirectLiNGAM) [31] aims to discover causal structures in the presence of non-Gaussian and linearly mixed variables. As an improvement independent Component Analysis-based LiNGAM (ICALiNGAM) [30] employs Independent Component Analysis to separate independent components and infer causality.

    Nonlinear optimization of causal effects with methods such as NOTEARS and NOTEARS LOW RANK algorithms [44] address causal discovery in the presence of latent variables, emphasizing the importance of sparsity in causal graphs. GraNDAG Mindspore [43] leverages MindSpore, an open-source deep learning framework, to integrate neural networks for causal discovery.

- Instrumental-variable (IV) methods [26, 2, 1] leverage statistical modeling to represent causal dependencies through latent variables.

- Score-based

  Unlike relying on independent components, Greedy Equivalence Search (GES) [5] algorithm focus on causal graph by iteratively adding and removing edges based on statistical tests.

- Temporal causality

  Additionally, causal inference from time-series data has been addressed through Granger causality [24] and dynamic causal modeling [10]. The challenges of confounding and selection bias are actively tackled through propensity score matching [25].

These algorithms collectively contribute to advancing our understanding of causality in a variety of domains, offering valuable tools for unraveling complex relationships in real-world data. As we point out in Section 4, these algorithms also leave much room for improvement, and there are recent attempts [7, 19, 3] to redefine causality to improve upon the traditional causal models.

Additive Noise Models As suggested above, various algorithms have been developed to infer causal structures from observational data, each with its unique strengths and assumptions. Additive noise models (ANM, [21]) assume that the observed variables are affected by independent noise, and aims to capture the true causal relationships by testing the independence between variables and additive noise. Given a set of random variables $\boldsymbol{X} = (X_1, ..., X_N)$ with index set $\boldsymbol{V} := \{1, ..., N\}$, there exists a $\mathcal{G} = (\boldsymbol{V}, \mathcal{E})$, where $\boldsymbol{V}$ denotes the nodes and $\mathcal{E} \subseteq \boldsymbol{V}^2$ with $(v, v) \notin \mathcal{E}$

denotes the edges of the graph [21]. For each V_j, the set of its parents is represented as $\mathbf{PA}_j$. If the structural assignments (f) of variables are of the form

$$X_j := f_j(\mathbf{PA}_j) + N_j, \qquad j = 1, 2, ..., N, \tag{1}$$

that is, if the noise N_j is additive, the structural causal model (SCM) is called an additive noise model (ANM). ANM thus model the true relationship between the input and the output by accounting for a nonlinear function and an additive noise term. The independence between input and noise improves the causal interpretability of the model. This allows for a clearer understanding of how changes in the input variable causally influence the output variable without the interference of correlated noise.

Example 4. *In our running Example 1, the weight grows with cube of height, decreases with the square root of the amount of exercise and is linearly dependent on the amount of bacteria 1 and bacteria 2 in our guts. Therefore, for some non-negative coefficients β_i, it holds that*

$$X_{weight} = \beta_{height} \cdot X^3_{height} + \beta_{exercise} \cdot \sqrt{X_{exercise}} - \beta_{b1} X_{b1} + \beta_{b2} X_{b2} + N_{weight}.$$

The noise term includes random differences between humans, accounts for unknown hidden factors, and also includes uncertainty between the measurements.

When assignments f are non-linear, let the joint probability distribution $\mathbf{P_X}$ be induced by an ANM with (1), where noise variables are normally distributed as $N_j \backsim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable functions f_j are nonlinear (See Theorem 7.7 in [21]). Specifically, the parents $\mathbf{PA}_j$ of X_j are denoted as $X_{k_1}, ..., X_{k_l}$. The function $f_j(x_{k_1}, ..., x_{k_{a-1}}, \cdot, x_{k_{a+1}}, ..., x_{k_l})$ is assumed to be non-linear for all a and some $x_{k_1}, ..., x_{k_{a-1}}, x_{k_{a+1}}, ..., x_{k_l} \in \mathbb{R}^{l-1}$. In this case, we can identify the corresponding graph $\mathcal{G}$ from the joint distribution $\mathbf{P_X}$ (See proof of Corollary 31 in [22]).

Linear Additive Noise Models In some cases, it may be preferable to assume that the functions f_j in additive noise models (1) are linear.

Example 5. *If we restrict Example 1 to people of common height, the relationship between the variables is locally linear. The same holds for the amount of exercise, assuming that we exclude outliers as professional athletes. In such a case, a linear approximation of the formula in 4 might be useful. Then, for some non-negative coefficients β_i, it holds that*

$$X_{weight} = \beta_{height} \cdot X_{height} + \beta_{exercise} \cdot X_{exercise} - \beta_{b1} X_{b1} + \beta_{b2} X_{b2} + N_{weight}.$$

Under the assumption that the structural assignments (f) are linear, noises $N_j, j = 1, \ldots, N$ are $i.i.d.$ and follow the same Gaussian distribution, or alternatively, noises $N_j, j = 1, \ldots, N$ are jointly independent, non-Gaussian with strictly positive density, the ANM structure may be identifiable (cf. Proposition 7.5 & Theorem 7.6 in [21]). In particular, the identifiability of linear ANM is reduced to identifiability of linear dynamical systems, where there is a recent understanding of the sample complexity (cf. Table 1 for fully-observed systems and Table 2 more broadly).

System Identification and Linear Dynamic Systems (LDS) Let us formalize the connection in more detail. Let m be the hidden state dimension and n be the observational dimension. A linear dynamic system (LDS) $\mathbf{L}$ is defined as a quadruple $(\mathbf{F}, \mathbf{G}, \mathbf{\Sigma}, \mathbf{V})$, where $\mathbf{F}$ and $\mathbf{G}$ are *system matrices* of dimension $m \times m$ and $n \times m$, respectively. $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$ are covariance matrices [42]. Hence, a single realization of the LDS of length T, denoted $\mathbf{X} = \{x_1, x_2, \ldots, x_T\} \in \mathbb{R}^{n \times s \times T}$, is nonlinear, and is defined by *initial conditions* ϕ_0, and realization of noises υ_t and ω_t as

$$\phi_t = \mathbf{F}\phi_{t-1} + \omega_t, \tag{2}$$

$$x_t = \mathbf{G}'\phi_t + \upsilon_t, \tag{3}$$

where $\phi_t \in \mathbb{R}^{m \times s}$ is the vector autoregressive processes with hidden components and $\{\omega_t, \upsilon_t\}_{t \in \{1,2,\ldots,T\}}$ are normally distributed process and observation noises with zero mean and covariance of $\mathbf{\Sigma}$ and $\mathbf{V}$ respectively, i.e., $\omega_t \sim N(0, \mathbf{\Sigma}) \in \mathbb{R}^{m \times s}$ and $\upsilon_t \sim N(0, \mathbf{V}) \in \mathbb{R}^{n \times s}$. The transpose of $\mathbf{G}$ is denoted as $\mathbf{G}'$. Vector $x_t \in \mathbb{R}^{n \times s}$ serves

Table 1: Sample complexities of fully-observed system identification according to [35]. Dimension $d = d_{\mathsf{x}} + d_{\mathsf{u}}$, i.e., the dimension of the state and the dimension of the control. The total number of non-zero elements is denoted by d_{s}. $snr^\star$ denote the snr under the best possible active exploration policy. For [27], we only show the result for $\rho(A) \leq 1$. The sample complexities are given in terms of $N_{\mathrm{tot}} = N_{\mathrm{traj}} T$, i.e. the total number of samples, where T is the horizon and N_{traj} is the number of trajectories. For single trajectory data, we have $N_{\mathrm{tot}} = T$. All bounds are non-asymptotic and we only use the big-O notation to simplify the presentation of the bounds.

Paper	Trajectory	Stability	Actuation	Upper Bound	Burn-in time	Lower Bound
[6]	multiple	any	white-noise	$\tilde{O}(T\frac{d\log 1/\delta}{\varepsilon^2 snr_T})$	$T\tilde{O}(d+\log 1/\delta)$	-
[32]	single	$\rho(A)\leq 1$	white-noise	$\tilde{O}(\frac{d\log d/\delta}{\varepsilon^2 snr_T})$	$\tilde{O}(\tau d\log d/\delta)$	$\Omega(\frac{d+\log 1/\delta}{\varepsilon^2 snr_T})$
[27]	single	any	white-noise	$\tilde{O}(\frac{d\log d/\delta}{\varepsilon^2 snr_1})$	$\tilde{O}(d\log d/\delta)$	-
[12]	single	any	active	-	-	$\Omega(\frac{\log 1/\delta}{\varepsilon^2 snr_T^\star})$
[13]	single	$\rho(A)<1$	white-noise	$\tilde{O}(\frac{d+\log 1/d}{\varepsilon^2 snr_T})$	$\tilde{O}(\frac{d+\log 1/d}{(1-\rho(A))^2})$	-
[37]	single	$\rho(A)<1$	active	$\tilde{O}(\frac{d+\log 1/\delta}{\varepsilon^2 snr_\tau^\star})$	$\mathrm{poly}(\frac{1}{1-\rho(A)})\tilde{O}(d+\log 1/\delta)$	$\Omega(\frac{\log 1/\delta}{\varepsilon^2 snr_\infty^\star})$
[9]	single	$\rho(A)<1$	white-noise	$\tilde{O}\big(\frac{d_{\mathsf{s}}\log d/\delta}{\varepsilon^2 snr_\infty(1-\rho(A))}\big)$	$\tilde{O}\big(\frac{d_{\mathsf{s}}^2\log d/\delta}{(1-\rho(A))^4}\big)$	-
[34]	single	$\rho(A)\leq 1$	any	$\tilde{O}(\exp(d)\frac{\log 1/\delta}{\varepsilon^2})$	$\tilde{O}(d\log d/\delta)$	$\Omega(\exp(d)\frac{\log 1/\delta}{\varepsilon^2})$

as an observed output of the system. Recently, Zhou and Marecek [46] proposed to find the global optimum of the objective function subject to the feasibility constraints arising from (2) and (3):

$$\min_{f_t,\phi_t,\mathbf{G},\mathbf{F},\omega_t,\upsilon_t} \sum_{t\in\{1,2,\ldots,T\}} \|X_t - f_t\|_2^2 + \|\omega_t\|_2^2 + \|\upsilon_t\|_2^2, \tag{4}$$

for a L_2-norm $\|\cdot\|_2$. In the causality problem we are given N variables $\mathbf{X_j} \in \mathbb{R}^{s\times T}$. A natural problem is to find the estimation values f_t of the LDS that generated the observation data. In other words, we are interested in finding the optimal objective values and the residual vectors υ_t, ω_t that belong to each LDS.

Operator-Valued Trace Optimization Problems The formulation of trace optimization on DAGs typically revolves around the unconstrained trace minimization problem, represented as:

$$\text{tr min}(p) := \inf\{\text{tr}\, p(X) \,|\, X \in \mathcal{S}_n\}$$

where $\mathcal{S}_n$ is the space of n-tuple real symmetric matrices. Additionally, there exists a constrained version of trace minimization, denoted as:

$$\text{tr min}(p, Q) := \inf\{\text{tr}\, p(X) \,|\, X \in \mathcal{S}_n,\, q_i(X) \succeq 0,\, i = 1, \ldots, m\}$$

where $q_i(X) \succeq 0$ enforces positive semidefiniteness constraints on polynomials in Q. The optimization problem becomes more intricate when considering the interaction between the trace minimization objective and these constraints.

Recent advancements in non-commutative polynomial optimization, as introduced by [46], present a comprehensive approach to solving trace optimization problems. This work explores a sequence of natural linear matrix inequalities, showcasing the efficacy of non-commutative polynomial optimization in recovering causal structures. Furthermore, [41, 40, 39] contributes to this domain by investigating the application of the non-commutative variant of the Term-Sparsity Exploiting Moment/Sum-of-Squares (TSSOS) hierarchy. The proposed methodology not only demonstrates convergence but also scalability, providing a powerful tool for handling complex causal structures inherent in trace optimization problems on DAGs.

4 Formulating Causal Learning as a NCPOP

As suggested in the Introduction, we would like to learn causal models that make it possible to capture:

Table 2: Sample complexity and error bounds on the estimated Markov parameters for selected recent methods, according to [35]. The parameters $R \leq n$ and r are respectively the order of the system and the length of the FIR impulse response; see [33] and [36] for more information. The error bounds are measured with respect to the Frobenius norm.

Method	Sample Complexity	Error Bound ($\|\cdot\|_F$)	Additional Notes
Salar Fattahi [8]	$\mathcal{O}(\log^2(Tp))$	$\mathcal{O}\left(\sqrt{m}\left(\frac{\log(Tnp)}{N}\right)^{1/4}\right)$	Single trajectory
Oymak and Ozay [18]	$\tilde{\mathcal{O}}(Tq)$	$\tilde{\mathcal{O}}\left(\sqrt{m}\left(\frac{Tq}{N}\right)^{1/2}\right)$	Single trajectory
Sarkar *et. al.* [28]	$\tilde{\mathcal{O}}(n^2)$	$\tilde{\mathcal{O}}\left(\sqrt{m}\left(\frac{pn^2}{N}\right)^{1/2}\right)$	Single trajectory, Suitable for systems with unknown order
Zheng and Li [45]	$\tilde{O}(mT+q)$	$\tilde{\mathcal{O}}\left(\sqrt{m}\left(\frac{T^3q}{N}\right)^{1/2}\right)$	Multiple trajectories, Stable and unstable systems
Sun *et. al.* [33]	$\tilde{\mathcal{O}}(pR)$	$\tilde{\mathcal{O}}\left(\left(\frac{Rnp}{N}\right)^{1/2}\right)$	Multiple trajectories, MISO ($m=1$)
Tu *et. al.* [36]	$\tilde{\mathcal{O}}(r)$	$\tilde{\mathcal{O}}\left(\left(\frac{r}{T}\right)^{1/2}\right)$	Multiple trajectories, SISO ($p=m=1$)

- quantitative aspects of causality: for example, distinguishing between bacterial strains, whose growth is promoted by nutrients in fiber-rich diet, and strains, whose growth is inhibited by the same diet, as in Example 3
- non-linear aspects of causality: for example, up to some level of a metabolite, multiple strains of bacteria have to compete for the metabolite, while from some level onwards, the needs can be saturated and the growth can be constrained by other factors, as in Example 2 or Example 4
- hidden states (latent variables) of an *a priori* unknown dimension: for example, the role of hormones is widely acknowledged, but their concentrations may not be available with sufficient time resolution, and it may not be clear what hormones' concentrations to consider *a priori*. At the same time, one would like to preserve as much explainability as possible, perhaps through targeted reduction [14].
- cycles in causal relationships: for example, the metabolome depends on the diet and microbiome, but the microbiome is affected by the metabolome and the diet, as in Example 2. Related cyclic relationships are expected to involve gastrointestinal hormones [29].
- time-series aspects, such as nonanticipativity and delays: clearly, causal relationships should be established between the cause in the past and the effect in the future, with some delay between the two.
- mixture-model aspects: clearly, there are variations between the metabolism in various individuals, perhaps due to genomic differences. One should explore joint problems [17], where multiple causal models are learned without the assignment of individuals to subgroups represented by the causal models given *a priori*.

We aim to address most of these aspects of causal learning (not necessarily in linear additive noise models) using non-commutative polynomial optimization (NCPOP) techniques in:

1. **Modeling:** Express the objective function in a form suitable for NCPOP, considering the non-commutative nature of the variables and the presence of non-convex operators.
2. **NCPOP Conversion:** Use NCPOP techniques to convert the non-commutative polynomial optimization problem into a format that can be addressed numerically.
3. **Solution:** Apply numerical algorithms to find optimal solutions, taking into account the non-convexities inherent in the problem.

Non-Convex operator-valued problems, such as trace optimization, can be addressed using the powerful tools developed originally for commutative polynomial optimization. The use of NCPOP offers a versatile approach for handling non-commutative variables and non-convexities in optimization problems, making it applicable to a wide range of mathematical and computational challenges.

As a simple example, we consider the iterated additive-noise model estimation via non-commutative polynomial optimization (IANN), as presented in Algorithm 1, which learns a causal graph in the tradition of Pearl [20, 21]. Our formulation uses equations (4) subject to (2 and 3), although one could vary those easily to allow for the non-linear aspects.

5 Numerical Illustrations

5.1 Data Description

Synthetic Data Our synthetic data-generating procedure is implemented in Python. The observation data is generated according to the supplied node count, edge count, and noise types. For every data set, we perform $5, 10, 15, 20, 25, 30$ experiments, in which different random seeds are employed to construct SCM in each iteration.

A Well-Known Dataset Three real-world datasets are provided in gCastle API [43] and each contains observational records collected from the real-world. To be more precise, real-dataset-processed and true-graph tables are included in each decompress package. The real-dataset-processed table includes each row counts the occurrences of the alarms $(A_i, i = 0, 1, ..., 56)$ in 10 minutes, and the rows are arranged in the time order, i.e., first 10 mins., second 10 mins., etc. The true-graph table is the underlying causal relationships, according to expert experience.

5.2 Performance

We evaluated the estimated graphs using four metrics: F1 score, False Discovery Rate (FDR), True Positive Rate (TPR), and Structural Hamming Distance (SHD) which is the smallest number of edge additions, deletions, and reversals to convert the estimated graph into the true DAG. The SHD takes into account both false positives and false negatives and a lower SHD indicates a better estimate of the causal graph.

Baselines For a fair comparison, Table (3) lists run time and F1 score performance of multiple methods listed in the Related Work section, which illustrates the various numbers of parameters of all the baselines and our method are similar. Additionally, we experimented the same length of time window range and also applied two different random seeds to each SEM.

Comparison with ANCPOP We evaluated F1 score of both simulator data and real-world dataset, and learned causal structure from the data. For the artificial data, linear and nonlinear SEM samples including Gauss, Exp, Gumbel, Uniform, Logistic (for linear); and mlp, mim, Gaussian process, additive Gaussian process (gp-add), Quadratic (for nonlinear) noise types simulation were estimated. To make sure the variety of the application of the ANCPOP algorithm, we extended our result based on testing 6, 9 and 12 variables, as well as 10, 15 and 20 edges. The F1 score of applying ANCPOP to simulate samples from linear SEM with Gauss noise and nonlinear SEM with gp-add noise are performed in Figures (1) and (2) , respectively.

While the F1 scores leave a lot of room for improvement, we stress that the dimensions of the hidden state, and thus the system matrices are not assumed. The corresponding NCPOP is mathematically challenging, and runtime of current methods increases exponentially. In the NPA hierarchy [16, 23], this is exponential and high, butif sparity is exploited [41, 40, 38], the run-time stays relatively modest. We hope to improve upon the computational aspects further.

Algorithm 1: The procedure for causal learning.

function Independence Test (Input two-variables $\overrightarrow{X}, \overrightarrow{Y}$ and residuals $\widetilde{N}_X$, $\widetilde{N}_Y$)

if $\widetilde{N}_X \perp\!\!\!\perp X \quad and \quad \widetilde{N}_Y \not\perp\!\!\!\perp Y$ **then return** X causes Y;
if $\widetilde{N}_X \not\perp\!\!\!\perp X \quad and \quad \widetilde{N}_Y \perp\!\!\!\perp Y$ **then return** Y causes X;
if $\widetilde{N}_X \not\perp\!\!\!\perp X \quad and \quad \widetilde{N}_Y \not\perp\!\!\!\perp Y$ **then return** bad model;
if $\widetilde{N}_X \perp\!\!\!\perp X \quad and \quad \widetilde{N}_Y \perp\!\!\!\perp Y$ **then return** both directions possible;

end function

function NCPOP Residuals MIMIC (Input a variable X or Y)

For X (or Y, respectively), find error-free estimates f_t^X (or f_t^Y) using minimization subject to (2) and (3), to identify system:

$$\min_{f_t,\phi_t,\mathbf{G},\mathbf{F},\omega_t,\upsilon_t} \sum_{t\in\{1,2,\ldots,T\}} \|X_t - f_t\|_2^2 + \|\omega_t\|_2^2 + \|\upsilon_t\|_2^2.$$

$\widetilde{N}_X = \overrightarrow{Y} - f_t^X$ (or $\widetilde{N}_Y = \overrightarrow{X} - f_t^Y$);
return $\widetilde{N}_X$ or $\widetilde{N}_Y$;
end function

function Casual Model Construction

▷ Construct adjacency matrix $\mathbf{C}$ by fitting and testing variables X, Y;
while $d \neq n$ **do**
Initialize the adjacency matrix: $\mathbf{C} \leftarrow []$;
for X *in measured data* **do**
$\widetilde{N}_X \leftarrow$ NCPOP Residuals MIMIC(X);
for Y *in measured data* **do**
$\widetilde{N}_Y \leftarrow$ NCPOP Residuals MIMIC(Y);
if Independence Test(X, Y, $\widetilde{N}_X$, $\widetilde{N}_Y$) = *'a causes b'* **then**
$c_{ab} \leftarrow 1$;
else
$c_{ab} \leftarrow 0$;
end
end
end
return Adjacency Matrix $\mathbf{C}$;
▷ Check for Matrix Exponential Constraint, so that G is a DAG;
Calculate $d = \mathrm{tr}(e^{\mathbf{C}})$;
end
Result: Structure causal model
end function

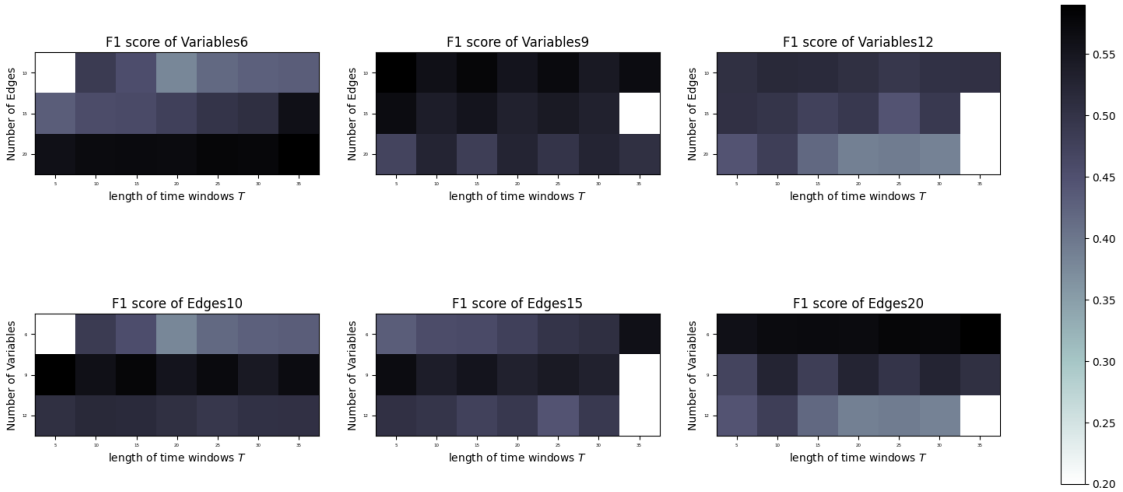


Figure 1: **Performance of Samples from Gaussian Linear Noise SEM**

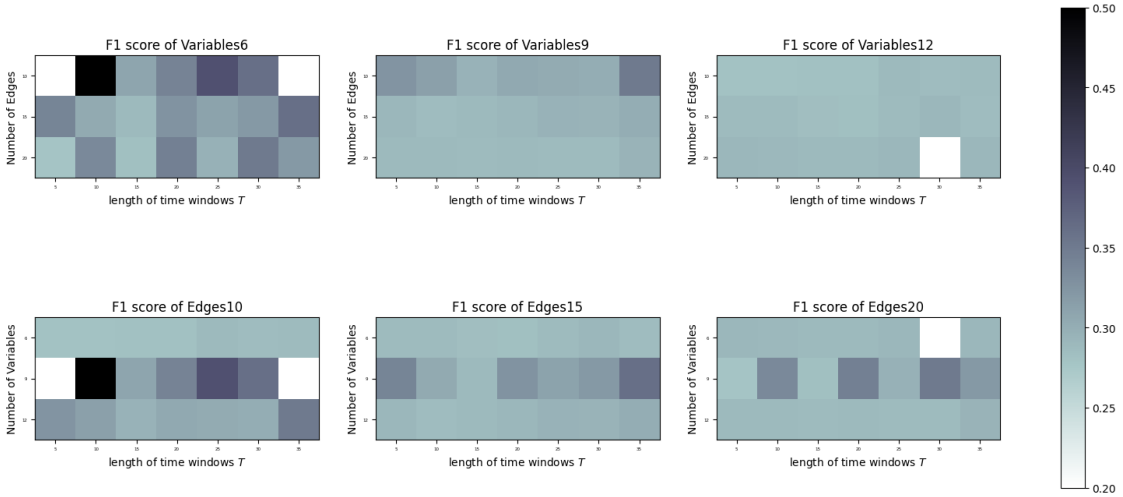


Figure 2: **Performance of Samples from gp-add Nonlinear Noise SEM**

Algorithm 2: The benchmarking procedure

function Synthetic Data Simulation (Give the number of variables(n) and edges(e), and noise distribution types of SEM)

Generate the artificial true causal graph and observation data based on the SCM;

▷ Randomly generate an adjacency matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ as a DAG;

foreach $a, b \in \{1, 2, \ldots, n\}$ *in* $\mathbf{C}$ **do**

$c_{ab} \leftarrow$ RandomInt($\{0, 1\}$);

▷ Check for Matrix Exponential Constraint;

Calculate $d = \mathrm{tr}(e^{\mathbf{C}})$;

if *d = n* **then**

return Simulated adjacency matrix $\mathbf{C}$;

end

end

end function

Table 3: **Comparison Table between different algorithms**

Algorithms	F1 Score	Duration(s)
PC(variant=stable)	0.65	636
PC(variant=parallel)	0.6486	636
ANM	0.9	647
DirectLiNGAM	0.4651	687
ICALiNGAM	0.878	57
GES	0.5926	766
NOTEARS	0.9	993
NOTEARS Low Rank	1	582
GraNDAG	0.1818	324

References

[1] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[2] Kenneth A Bollen. Structural equations with latent variables. *Wiley*, 1989.

[3] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *NeurIPS 2023, arXiv preprint arXiv:2306.02235*, 2023.

[4] Sabine Burgdorf, Igor Klep, and Janez Povh. *Optimization of polynomials in non-commuting variables*. Springer, 2016.

[5] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2003.

[6] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2017.

[7] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, Georgios Arvanitidis, and Bernhard Schölkopf. On data manifolds entailed by structural causal models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8188–8201. PMLR, 23–29 Jul 2023.

[8] Salar Fattahi. Learning partially observed linear dynamical systems from logarithmic number of samples. In *Learning for Dynamics and Control*, pages 60–72. PMLR, 2021.

[9] Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Learning sparse dynamical systems from a single sample trajectory. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2682–2689. IEEE, 2019.

[10] Karl J. Friston, Lee M. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19:1273–1302, 2003.

[11] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[12] Yassir Jedra and Alexandre Proutiere. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2676–2681. IEEE, 2019.

[13] Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 996–1001. IEEE, 2020.

[14] Armin Keki, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. *arXiv preprint arXiv:2311.18639*, 2023.

[15] Sabina Leanti La Rosa, Maria Louise Leth, Leszek Michalak, Morten Ejby Hansen, Nicholas A. Pudlo, Robert Glowacki, Gabriel Pereira, Christopher T. Workman, Magnus Ø. Arntzen, Phillip B. Pope, Eric C. Martens, Maher Abou Hachem, and Bjørge Westereng. The human gut firmicute roseburia intestinalis is a primary degrader of dietary β-mannans. *Nature Communications*, 10(1):905, Feb 2019.

[16] Miguel Navascués, Stefano Pironio, and Antonio Acín. SDP relaxations for non-commutative polynomial optimization. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 601–634. Springer, 2012.

[17] Mengjia Niu, Xiaoyu He, Petr Rysavy, Quan Zhou, and Jakub Marecek. Joint problems in learning multiple dynamical systems. *arXiv preprint arXiv:2311.02181*, 2023.

[18] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

[19] Junhyung Park, Simon Buchholz, Bernhard Schölkopf, and Krikamol Muandet. A measure-theoretic axiomatisation of causality. *NeurIPS 2023, arXiv preprint arXiv:2305.17139*, 2023.

[20] Judea Pearl. *Causality*. Cambridge university press, 2009. 2nd ed.

[21] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[22] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

[23] S. Pironio, M. Navascués, and A. Acín. Convergent relaxations of polynomial optimization problems with noncommuting variables. *SIAM Journal on Optimization*, 20(5):2157–2180, 2010.

[24] Alard Roebroeck, Elia Formisano, and Rainer Goebel. Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage*, 25:230–242, 2005.

[25] Paul R. Rosenbaum and Donald B. Rubin. *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, page 170–184. Cambridge University Press, 2006.

[26] John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the econometric society*, pages 393–415, 1958.

[27] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

[28] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Nonparametric finite time lti system identification. *arXiv preprint arXiv:1902.01848*, 2019.

[29] Rebecca Scott, Tricia Tan, and Stephen Bloom. Chapter seven - gut hormones and obesity: Physiology and therapies. In Gerald Litwack, editor, *Obesity*, volume 91 of *Vitamins Hormones*, pages 143–194. Academic Press, 2013.

[30] Amirhossein Shahbazinia, Saber Salehkaleybar, and Matin Hashemi. Paralingam: Parallel causal structure learning for linear non-gaussian acyclic models. *Journal of Parallel and Distributed Computing*, 176:114–127, 2023.

[31] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti J. Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

[32] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

[33] Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample system identification: Optimal rates and the role of regularization. In *Learning for Dynamics and Control*, pages 16–25. PMLR, 2020.

[34] Anastasios Tsiamis and George J Pappas. Linear systems can be hard to learn. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2903–2910. IEEE, 2021.

[35] Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

[36] Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

[37] Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

[38] Jie Wang, Martina Maggio, and Victor Magron. Sparsejsr: A fast algorithm to compute joint spectral radius via sparse sos decompositions. In *2021 American Control Conference (ACC)*, pages 2254–2259. IEEE, 2021.

[39] Jie Wang and Victor Magron. Exploiting term sparsity in noncommutative polynomial optimization. *Computational Optimization and Applications*, 80:483–521, 2021.

[40] Jie Wang, Victor Magron, and Jean-Bernard Lasserre. Chordal-tssos: a moment-sos hierarchy that exploits term sparsity with chordal extension. *SIAM Journal on Optimization*, 31(1):114–141, 2021.

[41] Jie Wang, Victor Magron, and Jean-Bernard Lasserre. Tssos: A moment-sos hierarchy that exploits term sparsity. *SIAM Journal on Optimization*, 31(1):30–58, 2021.

[42] Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

[43] Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*, 2021.

[44] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[45] Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2020.

[46] Quan Zhou and Jakub Mareek. Learning of linear dynamical systems as a non-commutative polynomial optimization problem. *IEEE Transactions on Automatic Control*, pages 1–7, 2024. 10.1109/TAC.2023.3313351.

5.3 Background: Directed Acyclic Graphs (DAG)

Directed Acyclic Graphs (DAG) are graphs $G = (V, E)$ where:

- V is the set of vertices (nodes),
- E is the set of directed edges, where each edge is an ordered pair of distinct vertices (u, v) indicating a directed connection from vertex u to vertex v,
- The graph has the acyclic property, meaning there are no directed cycles in the graph. Formally, there is no sequence of vertices $v_1, v_2, \ldots, v_k$ such that $(v_1, v_2), (v_2, v_3), \ldots, (v_{k-1}, v_k), (v_k, v_1)$ are all edges in the graph.

The adjacency matrix C is a matrix representation of the graph G, i.e.,

$$\mathbf{C} = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ c_{21} & \cdots & c_{2n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}, \tag{5}$$

where

$$c_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

According to matrix exponential constraint [44], a binary matrix $\mathbf{C} \in \{0, 1\}^{n \times n}$ is a DAG if and only if

$$\operatorname{tr}(e^{\mathbf{C}}) = n. \tag{6}$$

The proof can be found in Proposition 2 of [44].