The Cartesian Gaussian Additive Noise Model for Causal Inference with Dependent Samples

Bailey Andrew
Department of Computer Science
University of Leeds
sceba@leeds.ac.uk

David R.M Westhead
School of Molecular and Cellular Biology
University of Leeds

Luisa Cutillo
School of Mathematics
University of Leeds

Abstract

We study the task of causal structure discovery from observational data when both the features and samples of our observation have causal structure. An example application is single-cell RNA-sequencing data, in which both the genes (features) and the cells (samples) interact in meaningful ways. We introduce the Cartesian Linear Gaussian Additive Noise Model to account for sample interactions, and generalize it to tensor-variate datasets with arbitrary interactions along each axis. We prove identifiability conditions analogous to those for the standard Linear Gaussian Additive Model, and produce a fast algorithm to learn the causal structure. Our method performs well on real data.

1 Introduction

Causal networks are of keen interest in various domains and come in many forms, such as gene regulatory networks (GRNs), signaling networks, and behavioral networks. In particular, directed acyclic graphs (DAGs) are often imbued with a causal interpretation; if the edge $i \to j$ exists in our DAG, we say i caused j.

The gold standard for estimating causality is an interventional experiment. However, not all problems are amenable to such an approach - it may be infeasible or even unethical. In such cases, we are forced to rely on estimating causality from observational data. Even when an interventional study is possible, if the space of possible causal relationships is large, estimates of causality from observational data can still be useful to narrow down the space of plausible hypotheses.

Observational datasets often come in the form of matrices $\mathbf{X}_{\text{samples}\times\text{features}}$, in which we assume our features have causal structure and our samples are independent. This assumption, although convenient, is rarely satisfied in practice! For single-cell RNA-sequencing data (scRNA-seq), we wish to estimate GRNs even under the knowledge that our samples (individual cells) come from a highly interlinked community. In general, we would like methods to work when both the rows and columns of our matrix have some kind of dependency structure. This can naturally be generalized to tensor-variate datasets as well.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: New Perspectives in Advancing Graph Machine Learning.

In this workshop paper, we are interested in estimating the parameters $\{\mathbf{C}_i\}$ generated by the following process (see Section 3 for a tensor-variate version):

$$\begin{aligned} \mathbf{X} &= \mathbf{C}_1\mathbf{X} + \mathbf{X}\mathbf{C}_2^T + \mathbf{E} \\ \mathbf{E}_{ij} &\sim_{i.i.d.} \mathcal{N}(0,1) \end{aligned} \tag{1}$$

This is a generalization of the common *linear Gaussian additive noise model* (LGAM), for which $\mathbf{X} = \mathbf{C}\mathbf{X} + \mathbf{E}$. If our dataset is ordered causally, i.e. x_i happens upstream of x_{i+k}, then we can write the LGAM as:

$$\begin{aligned} x_0 &= \epsilon_0 \\ x_1 &= c_{10}x_0 + \epsilon_1 \\ &\vdots \\ x_N &= c_{N0}x_0 + \cdots + c_{N,N-1}x_1 + \epsilon_N \end{aligned} \tag{LGAM}$$

Defining $\mathfrak{C}$ as the causal graph, we have for $i \neq j$ that $c_{ij} \neq 0 \iff (j \to i) \in \mathfrak{C}$.

Our more complicated generative process (Equation 1) can also be framed causally: $\mathbf{C}_1$ is the weighted adjacency matrix of the causal graph of the rows $\mathfrak{C}_1$ and $\mathbf{C}_2$ is likewise the adjacency matrix of the causal graph of the columns $\mathfrak{C}_2$. As we will see in a moment, our model is related to the Cartesian product of graphs, so we will call it the Cartesian LGAM, or $\boxplus$-LGAM for short.

It can be counter-intuitive to think about two co-existing notions of causality (that on rows and that on columns). An alternative way to interpret this model is in terms of the causality of matrix elements. For single-cell RNA sequencing data (scRNA-seq), "did cell i_1's gene j_1 cause cell i_2's gene j_2's expression levels" can be answered by looking at whether the edge $(i_1, j_1) \to (i_2, j_2)$ exists in the Cartesian product graph $\mathcal{C} = \mathcal{C}_1 \boxplus \mathcal{C}_2$. It is this property that motivates the structure of our $\boxplus$-LGAM model.

This workshop paper concerns the estimation of $\mathcal{C}_1 \boxplus \mathcal{C}_2$, and more generally for K-axis tensors the estimation of $\boxplus_i^K \mathfrak{C}_i$, under the $\boxplus$-LGAM model. We do this using maximum likelihood estimation equipped with an L1 and L2 penalty (analogous to ElasticNet). We focus on the case where the causal order is known *a priori*, but also consider when it is unknown.

1.1 Our contributions

In this workshop paper, we introduce the $\boxplus$-LGAM model (Section 3), a method to learn directed graphs describing observational data. If a dataset comes from the $\boxplus$-LGAM distribution and the causal ordering is known, then this graph has a causal interpretation. We prove that our optimization problem is well-founded (Theorem 1) and that the solution has good statistical recovery rates (Theorem 2).

1.2 Notation

$\mathbf{X}$ will always refer to a generic dataset of size $d_1 \times d_2$ with covariance $\mathbf{S} = \operatorname{vec}[\mathbf{X}]\operatorname{vec}[\mathbf{X}]^T$. In general, scalars are always standard script x, vectors are lowercase bold $\mathbf{x}$, matrices are uppercase bold $\mathbf{X}$, tensors are uppercase caligraphic $\mathcal{X}$, spaces are blackboard bold $\mathbb{X}$, and graphs are always gothic uppercase $\mathfrak{X}$. We'll denote the space of Kronecker-sum-structured lower triangular matrices with positive diagonals by $\mathbb{L}_{++}^{\oplus}$. The tangent space of $\mathbb{L}_{++}^{\oplus}$ (the space of directions locally preserving membership in $\mathbb{L}_{++}^{\oplus}$) is $\mathbb{L}^{\oplus}$, the space of all lower-triangular Kronecker-sum-structured matrices. In general, superscripts denote which axis a term refers to; for example, $\mathfrak{C}^{(1)}$ refers to the causal graph of features, $\mathfrak{C}^{(2)}$ refers to the causal graph of observations, and $\mathfrak{C}$ refers to the Cartesian product of the two.

Our methodology is extensible to the tensor-variate case. For a K-axis $d_1 \times \cdots \times d_K$ tensor dataset $\mathcal{X}$, we let $\mathbf{L} = \bigoplus_{\ell=K}^{1} \mathbf{L}^{(k)} = \sum_{\ell}^{K} \mathbf{I} \otimes \mathbf{L}^{(k)} \otimes \mathbf{I}$ for appropriately-sized identity matrices. hen it is clear from context, we will use subscripts $\mathbf{L}_i$ instead of superscripts $\mathbf{L}^{(i)}$ - especially in the presence of other superscript-dwelling notations such as powers and transposes. $\mathcal{T} \times_i \mathbf{M}$ is the i-mode matrix product, which is notationally cumbersome to define (see the review by Kolda and

Bader) but is equivalent to batch matrix multiplication of $\mathcal{T}$ by $\mathbf{M}$ along its ith axis. For matrices, $\mathbf{M} \times_1 \mathbf{N} = \mathbf{N}^T\mathbf{M}$ and $\mathbf{M} \times_2 \mathbf{N} = \mathbf{M}\mathbf{N}$. Let $d_\forall = \prod_\ell d_\ell$, $d_{\setminus\ell} = \frac{d_\forall}{d_\ell}$, and $d_{\setminus\ell\setminus\ell'} = \frac{d_\forall}{d_\ell d_{\ell'}}$.

We will use hats $\hat{\mathbf{L}}$ to represent estimated parameters. If our data follows some (potentially non-$\boxplus$-LGAM) distribution $\mathcal{D}^*$, then we let $\mathbf{L}^* = \underset{\mathbf{L}}{\operatorname{argmin}}\ \mathbb{E}_{\mathcal{X}\sim\mathcal{D}^*}\left[\mathrm{NLL}\left(\mathbf{L}, \mathcal{X}\right)\right]$, where NLL is the negative-log-likelihood of our $\boxplus$-LGAM model. In other words, $\mathbf{L}^*$ is the unknown 'oracle' parameter that one would recover if one had knowledge of the true data distribution.

2 Background

2.1 Kronecker-structured covariance estimation

Broadly, Kronecker-structured covariance estimation aims to estimate covariance (and related quantities) when the covariance matrix can be factorized according to some function involving Kronecker products. The most basic such model is the matrix normal distribution $\mathcal{MN}$:

$$\mathbf{X} \sim \mathcal{MN}\left(\mathbf{M}, \mathbf{\Sigma}_{\text{rows}}, \mathbf{\Sigma}_{\text{columns}}\right) \qquad \iff \mathrm{vec}\left[\mathbf{X}\right] \sim \mathcal{N}\left(\mathrm{vec}\left[\mathbf{M}\right], \mathbf{\Sigma}_{\text{columns}} \otimes \mathbf{\Sigma}_{\text{rows}}\right)$$

The factor matrices $\mathbf{\Sigma}_{\text{rows}}, \mathbf{\Sigma}_{\text{columns}}$ represent the covariances along their respective axes, and their Kronecker product represents the covariance structure of matrix elements.

Rather than encoding Kronecker structure on the covariance matrix, it is often advantageous to encode it on the inverse covariance matrix (the 'precision' matrix). For normally-distributed data, the sparsity structure of the precision matrix encodes the graph of conditional dependencies. Two elements i, j are conditionally dependent if they are independent after conditioning out the rest of the variables $\mathbf{k}$ in the dataset:

$$\mathbb{P}(i|j, \mathbf{k}) = \mathbb{P}(i|\mathbf{k}) \qquad \iff i \text{ is conditionally independent of } j$$
$$\text{(definition of conditional independence in dataset } \{i, j, \mathbf{k}\})$$

Choice of the exact Kronecker structure for the precision matrix (Kronecker product, Kronecker sum, etc) typically coincides with either a graph product on the conditional dependencies or a generative process. Only the Kronecker product structure is interpretable as both. In our case, our method is interpretable as both a Cartesian product on the causal graph and as a generative process ($\boxplus$-LGAM).

Various structures have been considered on precision matrices (Kronecker product, Kronecker sum [5], squared Kronecker sum [14, 13], and the strong product [1]). Such structures have also been imposed on the Laplacian of the graph of conditional dependencies [11]. For a more complete picture, we refer the reader to the excellent review by Wang et al.. Our model is similar to the squared Kronecker sum approach, although we optimize over DAGs and they optimize over symmetric matrices.

2.2 Causal structure discovery

The LGAM model is arguably the simplest causal model for continuous variables. It's a type of *additive noise model* [4], which (for arbitrary functions f_i, causal graph $\mathcal{C}$, and distribution $\mathcal{D}$) takes the general form:

$$\begin{aligned} x_i &= f_i(\mathbf{x}_{j|(j\to i)\in\mathfrak{C}}) + \epsilon_i \\ \epsilon_i &\sim_{independently} \mathcal{D} \end{aligned} \qquad \text{(Additive Noise Model)}$$

In our case, f_i are linear functions and $\mathcal{D}$ is the Gaussian distribution. This aids in computational tractability, but comes at the cost of identifiability when the causal ordering is unknown.

Typical additive noise models assume independence of samples. Peters et al. study the case where the samples form a time series, in which autocorrelation becomes a concern. Causal inference for time series has been extensively studied [9, 2, 10], but not every scenario with non-independent variables forms a time series. In this workshop paper, we combine the causal LGAM approach with the work already done on Kronecker-structured covariance models to allow inference of causal structure of both the samples and the features simultaneously.

3 The proposed model

In this section, we will introduce our model. We defer discussion of the theory (such as existence and uniqueness of solutions) to Section 4. For tensor-variate data $\mathcal{X}$, our model is as follows:

$$\mathcal{X} = \left(\sum_{i}^{K} \mathcal{X} \times_i \mathbf{C}_i^T \right) + \mathcal{E}$$
$$\mathcal{E}_{\mathbf{i}} \sim_{i.i.d.} \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

Note that Equation 2 is a straightforward generalization of Equation 1 from the introduction. It is shown in the supplementary material that, letting $\mathbf{L}_\ell = \sqrt{d_{\setminus \ell}} \left(\frac{1}{K}\mathbf{I} - \mathbf{C}_\ell \right)$, any $\mathcal{X}$ generated from Equation 2 satisfies:

$$\text{vec}\left[\mathcal{X}\right] \sim \mathcal{N}\left(\mathbf{0}, \left(\bigoplus_{\ell} \frac{1}{\sqrt{d_{\setminus \ell}}} \mathbf{L}_\ell \right)^{-T} \left(\bigoplus_{\ell} \frac{1}{\sqrt{d_{\setminus \ell}}} \mathbf{L}_\ell \right)^{-1} \right)$$

This is well-founded, as $\mathbf{L}_\ell$ is invertible whenever $\mathbf{C}_\ell$ represents a DAG [12]. Importantly, $\mathbf{L}_\ell$ has the same sparsity pattern as $\mathbf{C}_\ell$, and thus is also an adjacency matrix for $\mathfrak{C}_\ell$. If we assume $\mathbf{C}_\ell$ is lower triangular (which we can ensure, given the causal ordering), $\mathbf{L}_\ell$ is also lower triangular, and hence $\mathbf{L} = \bigoplus_\ell \frac{1}{\sqrt{d_{\setminus \ell}}} \mathbf{L}_\ell$ represents the Cholesky factor of the inverse covariance matrix: $\text{vec}\left[\mathcal{X}\right] \sim \mathcal{N}\left(\mathbf{0}, \left(\mathbf{L}\mathbf{L}^T\right)^{-1}\right)$. We can frame this as a convex maximum likelihood estimation problem:

$$\hat{\mathbf{L}} = \underset{\mathbf{L} \in \mathbb{L}_{++}^{\oplus}}{\text{argmin}} \ \text{tr}\left[\mathbf{L}\mathbf{L}^T\mathbf{S}\right] - \log\det\left[\mathbf{L}\mathbf{L}^T\right] \tag{3}$$

Most real-world networks are sparse; we also place an L1 penalty on the off-diagonals ($\lambda_\ell \|\mathbf{L}^{(\ell)}\|_{od}$) to ensure sparse solutions. Additionally, we will see soon that, if our data is 'too lopsided' (one axis is much larger than the other), a solution is not guaranteed to exist. To counteract this, we add an optional Frobenius-norm penalty as well. Our final optimization problem, below, is thus analogous to ElasticNet.

$$\hat{\mathbf{L}} = \underset{\mathbf{L} \in \mathbb{L}_{++}^{\oplus}}{\text{argmin}} \ \text{tr}\left[\mathbf{L}\mathbf{L}^T\mathbf{S}\right] - \log\det\left[\mathbf{L}\mathbf{L}^T\right] + \sum_\ell \left(\rho_\ell \|\mathbf{L}^{(\ell)}\|_F^2 + \lambda_\ell \|\mathbf{L}^{(\ell)}\|_{od} \right) \tag{4}$$

As long as the dataset is not 'too lopsided', Equation 4 is both smooth and strongly convex, even without Frobenius regularization. This guarantees existence, uniqueness, and efficient estimation of solutions, with accelerated proximal gradient descent requiring only $O(\log \frac{1}{\epsilon})$ iterations to be within ϵ of the optimal value. We make this lopsidedness criteria more formal below:

Theorem 1 (Lopsidedness Theorem) *Consider a dataset $\mathcal{X}_{d_1 \times \cdots \times d_K} \neq 0$. As long as the probability distribution has positive definite covariance, if $\prod_\ell d_\ell + K \geq 1 + \sum_\ell \frac{d_\ell^2 - d_\ell}{2}$, then Equation 3 is almost surely strongly convex.*

Corollary 1 *Consider a matrix-variate dataset $\mathbf{X}_{d_1 \times d_2}$, for $d_2 = d_1 + \delta \geq d_1$. Then, a solution exists almost surely if $\delta \leq \frac{1 + \sqrt{9 + 8d_1}}{2}$.*

Corollary 2 *A sufficient condition for almost-sure solution existence is that, for the largest axis $d_{\max}$, we have $d_{\setminus \max} \geq \frac{K}{2} d_{\max}$ (where $d_{\setminus \max} = \frac{d_\forall}{d_{\max}}$).*

4 Theory

In this section, we aim to prove that our optimization problem is well-founded: there always exists a unique solution $\hat{\mathbf{L}}$, and that this solution is close to the oracle $\mathbf{L}^*$. For very heavy-tailed distributions, this cannot be guaranteed: our dataset $\mathcal{X} \sim \mathcal{D}^*$ would likely be too much of an outlier to be informative for our model; thus, we need to make an assumption about tail decay rates.

Distribution	AUC-PR		MCC	
	LGAM	$\boxplus$-LGAM	LGAM	$\boxplus$-LGAM
$100_{\text{ER}} \times 50_{\text{AR}(1)}$ Gaussian	0.778	**0.823** (↑6%)	0.726	**0.781** (↑8%)
$40_{\text{ER}} \times 40_{\text{ER}} \times 40_{\text{ER}}$ Gaussian	0.791	**0.801** (↑1%)	0.726	**0.756** (↑4%)
$60_{\text{ER}} \times 60_{\text{ER}} \times 60_{\text{ER}}$ Gaussian	0.758	**0.775** (↑2%)	0.684	**0.707** (↑3%)
$50_{\text{ER}} \times 50_{\text{AR}(1)}$ Gaussian	0.680	**0.762** (↑12%)	0.670	**0.715** (↑7%)
$50_{\text{ER}} \times 100_{\text{AR}(1)}$ Gaussian	0.468	**0.598** (↑28%)	0.600	**0.676** (↑13%)

Table 1: Results on synthetic data, ordered by problem difficulty. We report both the area under the precision-recall curve (AUC-PR) and the best Matthew's correlation coefficient (MCC) achieved by the methods. Results are reported on the second axis of the dataset for simplicity. All ground truth graphs were 10% dense. The subscript of an axis indicates the type of graph used (ER is Erdos-Renyi, AR(1) is autoregressive, and $\mathcal{I}$ is independent/an empty graph).

A common assumption is that our distribution be 'sub-Gaussian', i.e. its tails decay no slower than a Gaussian. For example, this assumption is made by Greenewald et al. when proving recovery rates for their Kronecker-sum-structured precision matrix model. However, scRNA-seq data is often modeled with a Poisson or negative binomial distribution, neither of which is sub-Gaussian. To ensure our recovery rates are adequate in this scenario, we will make the more general assumption of α-sub-exponentiality.

Definition 1 (α-sub-exponentiality) *A distribution $\mathcal{D}$ is α-sub-exponential if for $\alpha > 0$, some constant c and all $t > 0$, $\mathbb{P}_{x\sim\mathcal{D}}\left(|x - \mathbb{E}[x]| \geq t\right) \leq 2e^{\frac{-t^\alpha}{c}}$.*

Example 1 *The Gaussian distribution and all bounded distributions are $(\alpha \geq 2)$-sub-exponential. The Poisson and negative binomial distributions are $(\alpha = 1)$-sub-exponential. The Cauchy, log-normal, and t distributions are **not** α-sub-exponential.*

If the true distribution $\mathcal{D}^*$ is α-sub-exponential, then the tails decay fast enough that we can obtain favorable statistical recovery rates. The larger α is, the faster the tails decay, and thus the more favorable the recovery rate will be. Under this assumption, we can prove the following result:

Theorem 2 *Let $\lambda_1 = \lambda_2 = \cdots = \lambda$ and $\rho_1 = \rho_2 = \cdots = \rho$. Assume our data is α-sub-exponential with $\alpha \in (0, 2]$. Suppose the largest Frobenius norm of any factor of the regularized oracle solution $\mathbf{L}^*_\rho$ is R_F, the total amount of nonzero elements of the factors of the regularized oracle solution is s, and μ is the strong convexity constant. With probability at least $1 - O\left(2^{-\lambda^{\frac{2}{\alpha}}}\right)$, we have:*

$$\|\hat{\mathbf{L}} - \mathbf{L}^*\|_{F,\oplus} \leq \frac{2K\rho}{\mu} R_F + \frac{3\lambda\sqrt{s}}{\mu + \rho}$$

5 Results

It is often hard to validate graphical models, given that we rarely have access to a ground truth graph. We'll analyze our performance on two types of synthetic data: that which we generate ourselves from simple distributions to ensure a given causal structure (Section 5.1) and a simulated scRNA-seq dataset from the literature (Section 5.2). We'll call the former 'synthetic' and the latter 'simulated', to reflect the fact that the latter dataset is an expert-curated simulation of real-world data.

5.1 Synthetic Data

Each of the synthetic datasets will experiment with are generated from our generative model in Equation 2. We do not explore the effect of ρ here, but will in Section 5.2. We let $\lambda_1 = \cdots = \lambda_K = \lambda$.

In this section, our synthetic dataset is generated from our $\boxplus$-LGAM model, with the ground truth graphs being 10% dense. On this data, we investigate performance on out-of-distribution data in

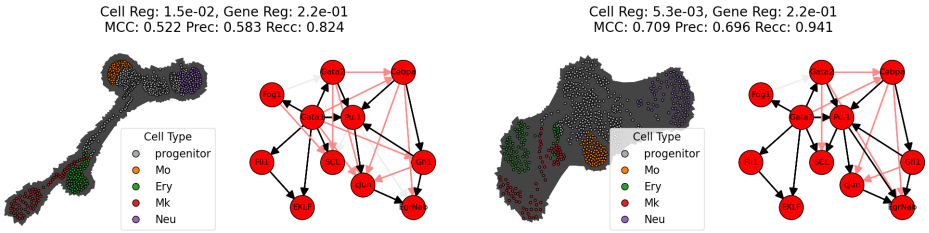

Figure 1: (Left) The third best gene graph arising from our ⊞-LGAM's grid search. Chosen for display as its cell graph qualitatively captures cell development the best. (Right) The best gene graph arising from our ⊞-LGAM model's grid search. (All) Black edges are true positives, near-white edges are false negatives, pale red edges are false positives.

Section 5.2. We summarize results in Table 1, using standard LGAM as a baseline. The method performs very well for tensor-variate and/or AR(1) data, outperforming the baseline in all scenarios given.

5.2 Simulated scRNA-seq Data

In this section, we apply our method to the simulated dataset of Krumsiek et al.. The dataset in question (the 'Krumsiek dataset') was generated from a literature-curated regulatory network of 11 genes involved in blood cell differentiation. The dataset contains 640 cells, which are labeled by cell type (erythrocyte, megakaryocyte, monocyte, neutrophil, and progenitor). Progenitor cells differentiate into the other cells; for scRNA-seq data, the term 'pseudo-time' is used to describe how far along a cell is on its development pathway.

We compare our performance against the LGAM baseline. The task for the gene axis is clear; we wish to know how well our method recovers the structure of the ground truth graph. For the cell axis, this is less clear. Given that we have access to cell labels, we can evaluate how well cells of the same type cluster in our graphs. We are also interested on how well, qualitatively, our graphs recover the pseudo-time structure of the dataset.

For our ⊞-LGAM model, we first grid-searched over the two L1 penalties λ_1, λ_2, with fixed ρ. As we don't have a quantitative metric for the cell graph's performance, we used the performance (Matthew's correlation coefficient) on the gene graph to rank results. We then picked the best-performing result, and performed an extended grid-search over ρ. For the baseline, we grid-searched over λ and ρ. See Figure 1(Right) for the gest gene graph. To investigate the cell graph qualitatively for pseudo-time alignment, we looked at the three best performing cell graphs from the original grid search, as well as the best performing cell graph from the extended grid search. Figure 1(Left) displays the one we, subjectively, felt captured pseudo-temporal structures the best.

The results are favorable to our model; our best gene graph achieved an MCC of 0.709, compared to 0.590 for the baseline. By optimizing only over gene graph performance, we automatically learned cell graphs as well, which outperformed the baseline in clustering quality measured by adjusted mutual information.

6 Discussion

This workshop paper addresses the problem of causal inference under non-independence of samples. In particular, we have generalized the classical linear Gaussian additive noise model using a Kronecker-separable framework. Our method achieves a good statistical recovery rate. By learning all graphs simultaneously, it allows metrics for one graph to be used to learn the other graphs, as done in Section 5.2. Even if we are unwilling or unable to make causal interpretations of the results $\hat{\mathbf{L}}$, they can always be interpreted as Cholesky factors.

Our ⊞-LGAM model could still be improved. For example, it is built around Gaussian noise terms; while this aids tractability, it harms identifiability and will not always be the best choice. Additionally, it would be useful to adapt the model to non-linear interactions between variables. It would be

valuable to generalize LGAM using other graph products, such as the strong product, and investigate which type of structure should be preferred in which contexts.

Additionally, there are some aspects of our model that we have not discussed in this workshop paper that will be present in the full paper version. In addition to more extensive experiments and full proofs of theorems, this also includes theory on when and how one can infer causal ordering from observational data, and a more complete analysis of what happens for lopsided data. Even when the data is too lopsided for Theorem 1, we can still establish some conditions for solution existence not given here.

References

[1] Bailey Andrew, David Robert Westhead, and Luisa Cutillo. The Strong Product Model for Network Inference without Independence Assumptions. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pages 5230–5238. PMLR, April 2025. URL https://proceedings.mlr.press/v258/andrew25a.html. ISSN: 2640-3498.

[2] Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Survey and Evaluation of Causal Discovery Methods for Time Series. *Journal of Artificial Intelligence Research*, 73:767–819, February 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13428. URL https://www.jair.org/index.php/jair/article/view/13428.

[3] Kristjan Greenewald, Shuheng Zhou, and Alfred Hero, III. Tensor Graphical Lasso (TeraLasso). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(5):901–931, November 2019. ISSN 1369-7412. doi: 10.1111/rssb.12339. URL https://doi.org/10.1111/rssb.12339.

[4] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://papers.nips.cc/paper_files/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

[5] Alfredo Kalaitzis, John Lafferty, Neil D. Lawrence, and Shuheng Zhou. The Bigraphical Lasso. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1229–1237. PMLR, May 2013. URL https://proceedings.mlr.press/v28/kalaitzis13.html. ISSN: 1938-7228.

[6] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Rev.*, 51(3):455–500, August 2009. ISSN 0036-1445. doi: 10.1137/07070111X. URL https://doi.org/10.1137/07070111X.

[7] Jan Krumsiek, Carsten Marr, Timm Schroeder, and Fabian J. Theis. Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. *PLOS ONE*, 6(8): e22649, August 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0022649. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0022649. Publisher: Public Library of Science.

[8] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal Inference on Time Series using Restricted Structural Equation Models. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/47d1e990583c9c67424d369f3414728e-Abstract.html.

[9] J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, July 2018. ISSN 1054-1500. doi: 10.1063/1.5025050. URL https://doi.org/10.1063/1.5025050.

[10] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, July 2023. ISSN 2662-138X. doi: 10.1038/s43017-023-00431-y. URL https://www.nature.com/articles/s43017-023-00431-y. Publisher: Nature Publishing Group.

[11] Changhao Shi and Gal Mishne. Learning Cartesian Product Graphs with Laplacian Constraints. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR, April 2024. URL https://proceedings.mlr.press/v238/shi24a.html. ISSN: 2640-3498.

[12] Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, September 2010. ISSN 0006-3444. doi: 10.1093/biomet/asq038. URL https://doi.org/10.1093/biomet/asq038.

[13] Yu Wang and Alfred Hero. SG-PALM: a Fast Physically Interpretable Tensor Graphical Model. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10783–10793. PMLR, July 2021. URL https://proceedings.mlr.press/v139/wang21k.html. ISSN: 2640-3498.

[14] Yu Wang, Byoungwook Jang, and Alfred Hero. The Sylvester Graphical Lasso (SyGlasso). In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1943–1953. PMLR, June 2020. URL https://proceedings.mlr.press/v108/wang20d.html. ISSN: 2640-3498.

[15] Yu Wang, Zeyu Sun, Dogyoon Song, and Alfred Hero. Kronecker-structured covariance models for multiway data. *Statistics Surveys*, 16(none):238–270, January 2022. ISSN 1935-7516. doi: 10.1214/22-SS139. URL https://projecteuclid.org/journals/statistics-surveys/volume-16/issue-none/Kronecker-structured-covariance-models-for-multiway-data/10.1214/22-SS139.full. Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are backed up in this workshop paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the section 'Discussion'.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: The proofs are long and technical, so are omitted. The theorems are about theoretical properties of the estimator (asymptotic bounds and solution existence in the edge case of $\rho = 0$), which while important are not the main contribution of this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our experiments as thoroughly as the space constraints allow, and are happy to go into more detail when necessary.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release our GitHub repository with all the experimental results once we finish writing the journal paper, which will be before this workshop takes place - at which time the experiments and algorithm implementation will be fully available for everyone.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: Yes

   Justification: We described our grid search

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We don't due to space reasons, but all results were run on a personal laptop; it is not a deep-learning model, and is more akin to a tool such as PCA.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Krumsiek et al. are cited for their data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This work is entirely unrelated to LLMs, nor did we use LLMs to write the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.