Estimating Causal Effects in Gaussian Linear SCMs with Finite Data

Aurghya Maiti [* 1]   **Prateek Jain** [* 1]

Abstract

Estimating causal effects from observational data remains a fundamental challenge in causal inference, especially in the presence of latent confounders. This paper focuses on estimating causal effects in Gaussian Linear Structural Causal Models (GL-SCMs), which are widely used due to their analytical tractability. However, parameter estimation in GL-SCMs is often infeasible with finite data, primarily due to overparameterization. To address this, we introduce the class of Centralized Gaussian Linear SCMs (CGL-SCMs), a simplified yet expressive subclass where exogenous variables follow standardized distributions. We show that CGL-SCMs are equally expressive in terms of causal effect identifiability from observational distributions and present a novel EM-based estimation algorithm that can learn CGL-SCM parameters and estimate identifiable causal effects from finite observational samples. Our theoretical analysis is validated through experiments on synthetic data and benchmark causal graphs, demonstrating that the learned models accurately recover causal distributions.

1. Introduction

Identifying causal effects from observational data is a foundational challenge in causal inference, largely due to the pervasive issue of unobserved confounding. Over the past decades, significant progress has been made in addressing this challenge in the non-parametric setting, culminating in robust theoretical frameworks such as Pearl's do-calculus, and systematic algorithmic approaches for identification (Pearl, 1995; Tian & Pearl, 2002; Shpitser & Pearl, 2007; Huang & Valtorta, 2006; Bareinboim & Pearl, 2012; 2016; Correa & Bareinboim, 2017; Lee et al., 2019; Correa et al., 2021; Lee et al., 2020; Correa & Bareinboim, 2024). In particular, the problems of identifying both interventional distributions (L_2 queries) and counterfactual distributions (L_3 queries) have been extensively explored and largely resolved in the non-parametric domain.

By contrast, the fields of econometrics and statistics have traditionally approached causal inference under parametric assumptions, focusing on the identifiability of causal effects in structured models—most notably in linear systems (Brito & Pearl, 2002; Tian, 2005; Chen et al., 2017; Kumor et al., 2019; 2020; Shimizu, 2014). However, many of these works assume either access to infinite data, partial knowledge of distributional parameters, or the Markovian assumption, which excludes latent confounding and can be overly restrictive in practical settings. On the other hand, recent efforts have aimed to estimate causal effects from finite data, but these have primarily focused on non-parametric models (Jung et al., 2021a;b).

This work seeks to bridge the gap between these two perspectives by addressing the problem of estimating identifiable causal effects in Gaussian Linear Structural Causal Models (GL-SCMs) using only finite observational data. GL-SCMs are widely used for their interpretability and analytic convenience, but they often suffer from high parameter complexity—particularly when modeling both observed and latent confounding—making accurate parameter estimation infeasible from limited samples.

To address this, we introduce the class of Centralized Gaussian Linear SCMs (CGL-SCMs), a simplified yet expressive subclass of GL-SCMs in which all exogenous variables are standardized to have zero mean and unit variance. This centralization significantly reduces the number of free parameters and enables tractable learning from finite data. Our key contributions are twofold:

1. We prove that CGL-SCMs retain the same expressive power as GL-SCMs with respect to identifying causal effects from observational distributions (Th. 2.4).

2. Second, we develop a scalable Expectation-Maximization (EM) algorithm to estimate the parameters of CGL-SCMs from finite data (Alg. 1 and Alg. 2), even in the presence of latent variables, which can then be used to identify causal effects.

---

[*]Equal contribution   [1]Department of Computer Science, Columbia University NY 10025.   Correspondence to: Aurghya Maiti <aurghya@cs.columbia.edu>, Prateek Jain <prateek.jain@columbia.edu>.

2. Centralized Gaussian Linear SCMs

In this section, we first investigate the class of Gaussian Linear SCMs and then introduce a new class of Centralized Gaussian Linear SCMs which have the same expressive power for effect identification from observation. In a Gaussian Linear SCM, the exogenous variables are independently sampled from normal distributions, and endogenous variables are a linear combination of their parents variables. Next, we provide a formal definition for such a model.

Definition 2.1 (Gaussian Linear SCM). A `GL-SCM` is defined by the tuple $\langle (\mathbf{U'}, \boldsymbol{\varepsilon'}), \mathbf{X}, \mathbb{P}, \mathbf{F_X} \rangle$ where (i) $\mathbf{U'} \sim \mathcal{N}(\boldsymbol{\mu_{U'}}, \boldsymbol{\Sigma^2})$ is the set of exogenous confounders ($\boldsymbol{\Sigma^2}$ is a diagonal matrix) (ii) $\boldsymbol{\varepsilon'} \sim \mathcal{N}(\boldsymbol{\mu_{\varepsilon'}}, \boldsymbol{\Psi^2})$ is the set of exogenous non-confounding variables (iii) The endogenous variables X_i are then given by

$$ X_i = \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki} U'_k + \mu'_i + \varepsilon'_i $$

where $Pa^o(X_i)$ are the observed parents of X_i, and $Pa^u(X_i)$ are the unobserved confounding parents of X_i. μ'_i are the bias for X_i.

Estimation of the exact parameters in such a model can be hard, and often impossible, since the number of parameters can be too large compared to the information available from the observed data. In the following part of the section, we introduce Centralized Linear SCMs, in which the exogenous variables are only sampled from standard distribution $\mathcal{N}(0, 1)$ and the errors are sampled from normal distributions with zero-mean. Such models can have much lesser number of parameters compared to GL-SCMs and these parameters can in many cases be estimated from data alone.

Definition 2.2 (Centralized Gaussian Linear SCM). A `CGL-SCM` is defined by the tuple $\langle (\mathbf{U}, \boldsymbol{\varepsilon}), \mathbf{X}, \mathbb{P}, \mathbf{F_X} \rangle$ where (i) $\mathbf{U}$: set of exogenous confounders and $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (ii) $\boldsymbol{\varepsilon}$: set of exogenous non-confounding variables, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi^2})$ (iii) The endogenous variables X_i are then given by

$$ X_i = \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U_k \in Pa^u(X_i)} \alpha_{ki} U_k + \mu_i + \varepsilon_i $$

where $Pa^o(X_i)$ are the observed parents of X_i, and $Pa^u(X_i)$ are the unobserved confounding parents of X_i. μ_i are the bias for X_i.

A natural question now is how much do we lose from GL-SCM to CGL-SCM is there any model which can be represented in GL-SCM, that cannot be modeled in CGL-SCM? Is there an effect that can be identified in GL-SCM with graphical assumptions and observational data, but not in CGL-SCM. The following theorem shows that as far as observational distributions are concerned, both GL-SCM and CGL-SCM has the same expressive power.

Theorem 2.3 (Expressivity of CGL-SCM). *For any model M' in* `GL-SCM`*, there exists a model M in* `CGL-SCM` *with same causal graph such that $P^{M'}(\mathbf{X}) = P^M(\mathbf{X})$.*

An advantage of CGL-SCM is that it can be estimated from data when GL-SCM cannot be estimated because of the higher number of parameters. Next, we state the most important observation that follows from the above discussion. Suppose a query Q is identifiable in the causal graph from observational data in the GL-SCM M'. Now, by the definition of identifiabiity, it implies that any SCM with the same causal graph and $P(\mathbf{X})$ will result in the same result for Q. Hence, estimating Q in the CGL-SCM with the same $P(\mathbf{X})$ gives us an estimate of Q in the GL-SCM, even though the exact parameters of the later are impossible to know from the available data. We formally state the theorem as follows:

Theorem 2.4 (Identification in GL-SCM). *Let M' and M be a GL-SCM and CGL-SCM respectively, with the same causal graph G and $P^{M'}(\mathbf{X}) = P^M(\mathbf{X})$. If a query Q is identifiable in causal graph G, then*

$$ P^{M'}(Q) = P^M(Q) \tag{1} $$

All proofs and derivations are in the Appendix.

3. Estimation of `CGL-SCM` using Finite Data

For our problem, we assume that the causal graph is known and that we are given finite samples from an observational (L_1) distribution. In this section, we present a procedure to estimate the parameters of a CGL-SCM using only finite observational data. It is important to note that the parameter estimation problem in this setting may not have a unique solution. However, our goal is not necessarily to recover the true data-generating model, but rather to find a CGL-SCM that induces the same L_1 distribution. Once such a model is estimated, any identifiable counterfactual or interventional query can be computed using this estimated SCM. By Theorem 2.4, the results of these queries will match those from the original model that generated the data.

To estimate the parameters of a CGL-SCM, we adopt the Expectation-Maximization (EM) algorithm, a well-established method for maximum likelihood estimation in the presence of latent variables. However, to apply EM in our setting, we must first reformulate the CGL-SCM in a vectorized form that aligns with the classical linear-Gaussian framework.

3.1. CGL-SCM in Vectorized Form

Given a CGL-SCM M, we first define a matrix representation of its causal structure. Let $\mathbf{X}$ denote the vector of endogenous variables, and let G be the known causal graph over $\mathbf{X}$. Construct a $|\mathbf{X}| \times |\mathbf{X}|$ matrix T where each entry

(i, j) is defined as:

$$t_{ij} = \begin{cases} \alpha_{ij} & \text{if causal edge } X_i \rightarrow X_j \text{ exists in } G \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Next, define the matrix B to capture the total influence of one variable on another through all directed paths in the graph. Let d be the length of the longest path in G. The matrix B is computed as:

$$B = I + \sum_{i=1}^{d} T^i \quad (3)$$

This sum includes the identity matrix to account for self-influence, which we define to have unit weight. The (i, j)-th entry of B thus represents the total (additive) influence of X_i on X_j across all directed paths in the graph.

In addition to B, define a $|\mathbf{U}| \times |\mathbf{X}|$ matrix C, where each entry (i, j) gives the direct effect of exogenous variable U_i on endogenous variable X_j.

Following this, the endogenous variables $\mathbf{X}$ in a CGL-SCM can be written in terms of only $\mathbf{U}$ and ε in the vectorized form as follows:

$$\mathbf{X} = B^T \boldsymbol{\mu} + B^T C^T \mathbf{U} + B^T \varepsilon \quad (4)$$

3.2. Algorithm

In the E-step, we compute the distribution of $\mathbf{U}^i \mid \mathbf{x}^i; \boldsymbol{\mu}, B, C \sim \mathcal{N}\big(\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i}, \Sigma_{\mathbf{U}^i|\mathbf{x}^i}\big)$ where,

$$\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i} = CB((CB)^T CB + B^T B)^{-1}(\mathbf{x}^i - B^T \boldsymbol{\mu})$$

$$\Sigma_{\mathbf{U}^i|\mathbf{x}^i} = \mathbf{I} - CB((CB)^T CB + B^T B)^{-1} B^T C^T$$

In the M-step, we want to maximize

$$max_{B,C,\mu} - n \log |B^T B|$$
$$- \sum_{i=1}^{N} \mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i} \Big[\big(\mathbf{x}^i - B^T \boldsymbol{\mu} - B^T C^T \mathbf{U}^i\big)^T$$
$$\big(B^T B\big)^{-1} \big(\mathbf{x}^i - B^T \boldsymbol{\mu} - B^T C^T \mathbf{U}^i\big) \Big]$$

We apply EM Algorithm for estimating B, C and $\boldsymbol{\mu}$. Note that naively applying EM Algorithm may result in a B, C which does not satisfy the graphical constraints. In order to do that we define masks B_m and C_m. Let T_m be a matrix where the element (i, j) is 1 is there is an edge $X_i \rightarrow X_j$. Then B_m is defined as

$$A = \sum_{i=1}^{d} T_m^i, \qquad b_{ij}^m = \begin{cases} 1 & \text{if } a_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 1 CGL-Go

1: **Initialization:** Initialize $B, C, \boldsymbol{\mu}$ and , B_m, C_m as defined in 3.2
2: $B = I + B \cdot B_m, \quad C = C \cdot C_m$
3: **repeat**
4:     **E-step (Expectation):**

$$\mu_{U^i|X^i} = CB((CB)^T CB + B^T B)^{-1}(\mathbf{X}^i - \mathbf{B}^T \boldsymbol{\mu})$$

$$\Sigma_{U^i|X^i} = I - CB((CB)^T CB + B^T B)^{-1}(CB)^T$$

5:     **M-step (Maximization):**
6:     Maximize $l(\mathbf{B}, \mathbf{C}, \boldsymbol{\mu})$ given by

$$- n \log |\mathbf{B}^T \mathbf{B}|$$
$$- \sum_{i=1}^{N} \mathbb{E}_{U^i|x^i} [(\mathbf{X}^i - (B^T \boldsymbol{\mu} + (CB)^T \mathbf{U}^i))^T$$
$$(B^T B)^{-1}(\mathbf{X}^i - (B^T \boldsymbol{\mu} + (CB)^T \mathbf{U}^i))]$$

7:     Update $\mathbf{B}$:
8:     **repeat**
9:         **for** each training example $\mathbf{X}^i$ **do**
10:         Compute the gradient: $\nabla_B l(B, C, \boldsymbol{\mu})$
11:         Update parameters: $B \leftarrow B + \eta \nabla_B l(B, C, \boldsymbol{\mu}) \cdot B_m$
12:         **end for**
13:     **until** Convergence
14:     Update $\mathbf{C}$:
15:     **repeat**
16:         **for** each training example $\mathbf{X}^i$ **do**
17:         Compute the gradient: $\nabla_C l(B, C, \boldsymbol{\mu})$
18:         Update parameters: $C \leftarrow C + \eta \nabla_C l(B, C, \boldsymbol{\mu}) \cdot C_m$
19:         **end for**
20:     **until** Convergence
21:     Update $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \Big((B^T)^{-1}\mathbf{X}^i - C^T \boldsymbol{\mu}_{U^i|X^i}\Big)$$

22: **until** Convergence

---

where a_{ij} and b_{ij}^m are the element (i, j) of A and B_m respectively and d is the length of the longest path in the graph. In C_m, the element (i, j) is 1 if there is an edge $U_i \rightarrow U_j$ otherwise 0. We did not find any closed form solution for B, C during maximization so we used gradient ascent to find the matrices during this step. While optimizing using gradient ascent we mask the gradients so that graphical constraints are maintained. The final algorithm is presented in Algorithm 1. The derivation is shown in Appendix.

Matrix C directly provides edge weight from each exoge-

Algorithm 2 `CGL-Edge`

1: **for** each node X_i in $\mathbf{X}$ **do**
2:     Identify a topological order τ_i of endogenous variables $\bar{X}_i$ connected by direct outward edge from X_i using causal graph G
3:     **for** each X_k in $\bar{X}_i$ in topological order τ_i **do**
4:         Calculate edge weight t_{ik}

$$t_{ik} = b_{ik} - \sum_{\tau_i(X_j) < \tau_i(X_k)} t_{ij} b_{jk}$$

5:     **end for**
6: **end for**

---

nous confounder in $\mathbf{U}$ to endogenous variables $\mathbf{X}$. But matrix B only provides aggregated weight between each pair of endogenous variables. In order to identify matrix T we use Algorithm 2.

4. Experiments

We evaluate our algorithm on synthetic datasets generated from two synthetic Gaussian Linear SCMs. The causal graphs, known as Frontdoor Graph (Fig. 1) and Napkin Graph (Fig. 2) are widely studied in the causal inference literature. For each graph, we generated 10,000 samples of endogenous variables using its associated `CGL-SCM` and applied Alg. 1 and 2 to estimate the SCM parameters from observational data. We then compared the resulting causal distributions with those of the original models. Full parameter matrices are provided in the Appendix; representative comparisons are shown below in Tab. 1 and 2.

4.1. Frontdoor Graph

We used following `CGL-SCM` $M = \langle (\boldsymbol{U}, \boldsymbol{\varepsilon}), \mathbf{X}, \mathbb{P}, \mathbf{F_X} \rangle$ inducing a frontdoor graph as shown in Figure 1 with $\mu = \begin{bmatrix} 0.3 & 0.1 & 0.2 \end{bmatrix}^T$.

Table 1. Comparing Original and Estimated Causal Distributions for Frontdoor Graph

	ORIGINAL	ESTIMATED
$P(X_3\|do(X_2 = 1))$	$\mathcal{N}(1.1, 1.09)$	$\mathcal{N}(1.1018, 1.069)$
$P(X_3\|do(X_1 = 1))$	$\mathcal{N}(0.74, 1.9)$	$\mathcal{N}(0.7391, 1.881)$

4.2. Napkin Graph

We used following `CGL-SCM` $M = \langle (\boldsymbol{U}, \boldsymbol{\varepsilon}), \mathbf{X}, \mathbb{P}, \mathbf{F_X} \rangle$ inducing a napkin graph as shown in Figure 2 with $\mu = \begin{bmatrix} 0.8 & -0.9 & 0.01 & -0.5 \end{bmatrix}^T$.

As we can see for both examples, estimated causal distribu-

Table 2. Comparing Original and Estimated Causal Distributions for Napkin Graph

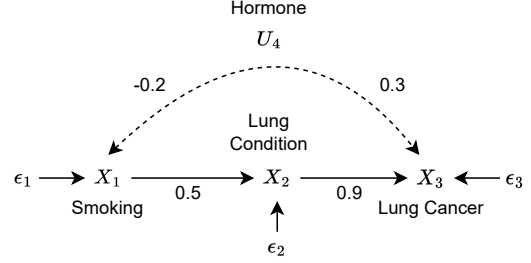|	ORIGINAL	ESTIMATED
$P(X_4\|do(X_3 = 1))$	$\mathcal{N}(0.3, 1.16)$	$\mathcal{N}(0.3051, 1.1692)$
$P(X_4\|do(X_1 = 1))$	$\mathcal{N}(-1.068, 2.3248)$	$\mathcal{N}(-0.9721, 2.3274)$

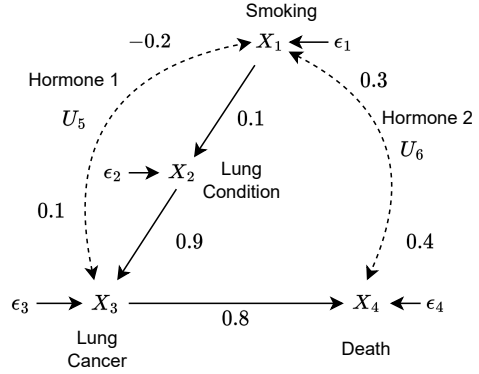*Figure 1.* Frontdoor Causal Graph with edge weights

Figure 2. Napkin Causal Graph with edge weights

tions are very close to original distributions.

5. Conclusion

In this work, we introduced Centralized Gaussian Linear SCMs as a tractable subclass of Gaussian SCMs that preserve the identifiability of causal effects from observational data. We demonstrated their theoretical equivalence to GL-SCMs in terms of expressivity and identifiability and proposed an EM-based learning algorithm for parameter estimation, and then estimating causal effects once we know the parameters. Through experiments on canonical graphs like the frontdoor and napkin structures, we showed that our method accurately recovers interventional distributions from finite samples. Future works can explore more general distributions or finding bounds for non-identifiable queries from finite data.

6. Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bareinboim, E. and Pearl, J. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 113–120, 2012.

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. In Shiffrin, R. M. (ed.), *Proceedings of the National Academy of Sciences*, volume 113, pp. 7345–7352. National Academy of Sciences, 2016.

Brito, C. and Pearl, J. A graphical criterion for the identification of causal effects in linear models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 2, pp. 533–540, 2002. URL https://cdn.aaai.org/AAAI/2002/AAAI02-080.pdf.

Chen, B., Kumor, D., and Bareinboim, E. Identification and model testing in linear structural equation models using auxiliary variables. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1156–1164, International Convention Centre, Sydney, Australia, Aug 2017. PMLR.

Correa, J. and Bareinboim, E. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 3740–3746, San Francisco, CA, Feb 2017. AAAI Press.

Correa, J. and Bareinboim, E. Counterfactual graphical models: Constraints and inference. Technical Report R-115, Causal Artificial Intelligence Lab, Columbia University, August 2024.

Correa, J., Lee, S., and Bareinboim, E. Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34: 6856–6867, 2021.

Huang, Y. and Valtorta, M. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, pp. 217–224, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922.

Jung, Y., Tian, J., and Bareinboim, E. Estimating identifiable causal effects through double machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12113–12122, 2021a.

Jung, Y., Tian, J., and Bareinboim, E. Estimating identifiable causal effects on markov equivalence class through double machine learning. In *International Conference on Machine Learning*, pp. 5168–5179. PMLR, 2021b.

Kumor, D., Chen, B., and Bareinboim, E. Efficient identification in linear structural causal models with instrumental cutsets. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/2a845d4d23b883acb632fefd814e175f-Paper.pdf.

Kumor, D., Cinelli, C., and Bareinboim, E. Efficient identification in linear structural causal models with auxiliary cutsets. In *International Conference on Machine Learning*, pp. 5501–5510. PMLR, 2020.

Lee, S., Correa, J. D., and Bareinboim, E. General identifiability with arbitrary surrogate experiments. In *In Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, 2019.

Lee, S., Correa, J. D., and Bareinboim, E. General identifiability with arbitrary surrogate experiments. In *Uncertainty in artificial intelligence*, pp. 389–398. PMLR, 2020.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Shimizu, S. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shpitser, I. and Pearl, J. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, pp. 352–359, Arlington, Virginia, USA, 2007. AUAI Press. ISBN 0974903930.

Tian, J. Identifying direct causal effects in linear models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, pp. 346–353, 2005. URL https://cdn.aaai.org/AAAI/2005/AAAI05-055.pdf.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pp. 567–573, USA, 2002. American Association for Artificial Intelligence. ISBN 0262511290.

A. Background

It is pivotal to introduce and elucidate the concepts of intervention and identifiability—crucial aspects in the realm of causality, integral to this paper.

Definition A.1 (Intervention). Given two disjoint sets of variables in an SCM M, $\mathbf{Z}, \mathbf{Y} \subseteq \mathbf{X}$, the causal effect of $\mathbf{Z}$ on $\mathbf{Y}$, denoted as $P(\mathbf{y}|do(\mathbf{z}))$, is a function from $\mathbf{Z}$ to the space of probability distributions of $\mathbf{Y}$. For each realization $\mathbf{z}$ of $\mathbf{Z}$, $P(\mathbf{y}|do(\mathbf{z}))$ gives the probability $\mathbf{Y} = \mathbf{y}$ induced by deleting from the model all equations corresponding to variables in $\mathbf{Z}$ and substituting $\mathbf{Z} = \mathbf{z}$ in the remaining equations.

Definition A.2 (Identifiabilty). Given the set of observed variables $\mathbf{X}$ such that $\mathbf{Z}, \mathbf{Y} \subseteq \mathbf{X}$ the causal effect of $\mathbf{Z}$ on $\mathbf{Y}$ is said to be identifiable from a causal diagram G if the quantity $P(\mathbf{y} \mid do(\mathbf{z}))$ can be computed uniquely from a positive probability of the observed variables. That is, if for every pair of models M_1 and M_2 inducing G,

$$P^{M_1}(\mathbf{y} \mid do(\mathbf{z})) = P^{M_2}(\mathbf{y} \mid do(\mathbf{z}))$$

whenever $P^{M_1}(\mathbf{x}) = P^{M_2}(\mathbf{x}) > 0$.

B. Proofs

B.1. Proof of Theorem 2.3

Let us take a `GL-SCM` $M' = \langle (\boldsymbol{U'}, \boldsymbol{\varepsilon'}), \mathbf{X}, \mathbb{P}, \mathbf{F_X} \rangle$. The endogenous variables X_i are then given by

$$X_i = \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki} U'_k + \mu'_i + \varepsilon'_i$$

where $Pa^o(X_i)$ are the observed parents of X_i, and $Pa^u(X_i)$ are the unobserved confounding parents of X_i. μ'_i are the bias for X_i. X_i can be re-written as:

$$X_i = \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki}(\mu_{U'_k} + \sigma^2_{U'_k} U_k) + \mu'_i + \mu_{\varepsilon'_i} + \varepsilon_i$$

$$= \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U_k \in Pa^u(X_i)} \alpha_{ki} U_k + \mu_i + \varepsilon_i$$

where,

$$\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi^2}), \quad \alpha_{ki} = \alpha'_{ki} \sigma^2_{U'_k}, \quad \mu_i = \left(\sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki} \mu_{U'_k} \right) + \mu'_i + \mu_{\varepsilon'_i}$$

This is exactly the definition of a `CGL-SCM`. Thus for every `GL-SCM` M' there exists a `CGL-SCM` M such that $P^{M'}(\mathbf{X}) = P^M(\mathbf{X})$.

B.2. Proof of Theorem 2.4

The proof follows directly from the definition of identifiablity, which says that if a query Q is identifiable from observational data $P(V)$ and causal graph, then all models which match on $P(V)$, must match on Q. Now, CGL-SCM is a subclass of GL-SCM, and if a Q is identifiable from $P(V)$ and the causal graph, then it must also match in Q.

C. Derivation of EM Steps

The distributions for $\mathbf{U}$ and $\mathbf{X}$ can be written as [1]

$$\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{X} = B^T \boldsymbol{\mu} + B^T C^T \boldsymbol{U} + B^T \boldsymbol{\varepsilon}$$

---

[1] Assume that ε has variance $\mathbf{I}$ for this part. Derivation for general $\boldsymbol{\Psi}$ is similar.

From these, we can derive the join distribution of $(\mathbf{U}, \mathbf{X})$, after computing mean and variance

$$\mathbb{E}[\mathbf{U}] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{X}] = B^T \boldsymbol{\mu}$$

Now, the variances $\Sigma_{UU} = I$ (from definition). Σ_{UX} and Σ_{XX} can be computed as follows:

$$\Sigma_{UX} = \mathbb{E}[(\mathbf{U} - \mathbb{E}[\mathbf{U}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] = \mathbb{E}[\mathbf{U}(B^T \boldsymbol{\mu} + B^T C^T \mathbf{U} + B^T \varepsilon - B^T \boldsymbol{\mu})^T]$$
$$= \mathbb{E}[\mathbf{U}\mathbf{U}^T CB + \varepsilon^T B] = \mathbb{E}[\mathbf{U}\mathbf{U}^T]CB + \mathbb{E}[\varepsilon^T]B = CB$$

Similarly, Σ_{XX}

$$\Sigma_{XX} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] = \mathbb{E}[(B^T C^T \mathbf{U} + B^T \varepsilon)(B^T C^T \mathbf{U} + B^T \varepsilon)^T]$$
$$= \mathbb{E}[B^T C^T \mathbf{U}\mathbf{U}^T CB + B^T C^T \mathbf{U}\varepsilon^T B + B^T \varepsilon \mathbf{U}^T CB + B^T \varepsilon\varepsilon^T B]$$
$$= B^T C^T \mathbb{E}[\mathbf{U}\mathbf{U}^T]CB + B^T \mathbb{E}[\varepsilon\varepsilon^T]B = (CB)^T CB + B^T B$$

The joint distribution $(\mathbf{U}, \mathbf{X})$ can be written as

$$\begin{bmatrix} \mathbf{U} \\ \mathbf{X} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ B^T \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{I} & CB \\ (CB)^T & (CB)^T CB + B^T B \end{bmatrix} \right)$$

In the E-step, we compute the distribution of $\mathbf{U}^i \mid \mathbf{x}^i; \boldsymbol{\mu}, B, C \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{U}^i \mid \mathbf{x}^i}, \Sigma_{\mathbf{U}^i \mid \mathbf{x}^i})$ where,

$$\boldsymbol{\mu}_{\mathbf{U}^i \mid \mathbf{x}^i} = CB((CB)^T CB + B^T B)^{-1}(\mathbf{x}^i - B^T \boldsymbol{\mu}), \qquad \Sigma_{\mathbf{U}^i \mid \mathbf{x}^i} = \mathbf{I} - CB((CB)^T CB + B^T B)^{-1}B^T C^T$$

The mean and variance of $\mathbf{X}^i \mid \mathbf{u}^i; \boldsymbol{\mu}, B, C$ can be computed similarly,

$$\boldsymbol{\mu}_{\mathbf{X}^i \mid \mathbf{u}^i} = B^T \boldsymbol{\mu} + B^T C^T \mathbf{u}^i, \qquad \Sigma_{\mathbf{X}^i \mid \mathbf{u}^i} = B^T B$$

In the M-step, we want to maximize

$$\sum_{i=1}^{N} \mathbb{E}_{\mathbf{U}^i \mid \mathbf{x}^i}[\log p(\mathbf{x}^i \mid \mathbf{U}^i; \boldsymbol{\mu}, B, C)]$$
$$= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|B^T B| - \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}\left[(\mathbf{x}^i - B^T \boldsymbol{\mu} - B^T C^T \mathbf{U}^i)^T (B^T B)^{-1} (\mathbf{x}^i - B^T \boldsymbol{\mu} - B^T C^T \mathbf{U}^i) \right]$$

with respect to $B, C, \boldsymbol{\mu}$

In order to optimize this, we have to compute the third term and take expectation

$$\sum_{i=1}^{N} \mathbb{E}\left[(\mathbf{x}^i - B^T \boldsymbol{\mu} - B^T C^T \mathbf{U}^i)^T (B^T B)^{-1} (\mathbf{x}^i - B^T \boldsymbol{\mu} - B^T C^T \mathbf{U}^i) \right]$$
$$= \sum_{i=1}^{N} (\mathbf{x}^i)^T (B^T B)^{-1} \mathbf{x}^i - 2(\mathbf{x}^i)^T B^{-1}(\boldsymbol{\mu} + C^T \mathbb{E}_{\mathbf{U}^i \mid \mathbf{x}^i}[\mathbf{U}^i]) + \boldsymbol{\mu}^T \boldsymbol{\mu} + 2\boldsymbol{\mu}^T C^T \mathbb{E}_{\mathbf{U}^i \mid \mathbf{x}^i}[\mathbf{U}^i] + \mathbb{E}_{\mathbf{U}^i \mid \mathbf{x}^i}\left[(\mathbf{U}^i)^T CC^T \mathbf{U}^i \right]$$

The term $\mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}\left[(\mathbf{U}^i)^T CC^T \mathbf{U}^i\right]$ can be computed as follows:

$$
\mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}\left[(\mathbf{U}^i)^T CC^T \mathbf{U}^i\right] = \mathbb{E}\Big[\sum_j \sum_{l,k} c_{lj} c_{jk} \cdot U_l^i U_k^i \mid \mathbf{x}^i\Big] = \sum_j \sum_{l,k} c_{lj} c_{jk} \mathbb{E}[U_l^i U_k^i \mid \mathbf{x}^i] = \sum_{l,k} \sum_j c_{lj} c_{jk} \mathbb{E}[U_l^i U_k^i \mid \mathbf{x}^i]
$$

$$
= \sum_{l,k} \mathbb{E}[U_l^i U_k^i \mid \mathbf{x}^i] \sum_j c_{lj} c_{jk} = \mathbb{E}[\mathbf{U}^i \mathbf{U}^{i^T} \mid \mathbf{x}^i] \cdot CC^T = \left(\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i} \boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i}^T + \Sigma_{\mathbf{U}^i|\mathbf{x}^i}\right) \cdot CC^T
$$

D. Experiments

D.1. Frontdoor Graph

Following is the vectorized form of frontdoor CGL-SCM $M = \langle (\boldsymbol{U}, \boldsymbol{\varepsilon}), \mathbf{X}, \mathbb{P}, \mathbf{F_X} \rangle$:

$$
\mathbf{U} = \begin{bmatrix} U_4 \end{bmatrix} \sim N(0,1) \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \sim \mathcal{N}(0, \mathbf{I}') \qquad \mu = \begin{bmatrix} 0.3 \\ 0.1 \\ 0.2 \end{bmatrix} \qquad T = \begin{bmatrix} 0 & 0.5 & 0 \\ 0 & 0 & 0.9 \\ 0 & 0 & 0 \end{bmatrix}
$$

$$
C = \begin{bmatrix} -0.2 & 0 & 0.3 \end{bmatrix} \qquad B = I + \sum_{i=1}^{2} T^i = \begin{bmatrix} 1 & 0.5 & 0.45 \\ 0 & 1 & 0.9 \\ 0 & 0 & 1 \end{bmatrix}
$$

$$
\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = B^T \mu + B^T C^T U + B^T \boldsymbol{\varepsilon}
$$

After running Alg. 1 we obtained following estimations for B, C and μ

$$
\hat{\mu} = \begin{bmatrix} 0.3002 \\ 0.0963 \\ 0.2007 \end{bmatrix} \qquad \hat{C} = \begin{bmatrix} -0.2149 & 0 & 0.2626 \end{bmatrix} \qquad \hat{B} = \begin{bmatrix} 1 & 0.5012 & 0.4491 \\ 0 & 1 & 0.9011 \\ 0 & 0 & 1 \end{bmatrix}
$$

After running algorithm 2 we obtained following estimation for T

$$
\hat{T} = \begin{bmatrix} 0 & 0.5012 & 0 \\ 0 & 0 & 0.9011 \\ 0 & 0 & 0 \end{bmatrix}
$$

D.2. Napkin Graph

Following is the vectorized form of napkin CGL-SCM $M = \langle (\boldsymbol{U}, \boldsymbol{\varepsilon}), \mathbf{X}, \mathbb{P}, \mathbf{F_X} \rangle$:

$$
\mathbf{U} = \begin{bmatrix} U_5 \\ U_6 \end{bmatrix} \sim N(0, \mathbf{I}) \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix} \sim \mathcal{N}(0, \mathbf{I}') \qquad \mu = \begin{bmatrix} 0.8 \\ -0.9 \\ 0.01 \\ -0.5 \end{bmatrix} \qquad T = \begin{bmatrix} 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 0 \end{bmatrix}
$$

$$
C = \begin{bmatrix} -0.2 & 0 & 0.1 & 0 \\ 0.3 & 0 & 0 & 0.4 \end{bmatrix} \qquad B = I + \sum_{i=1}^{3} T^i = \begin{bmatrix} 1 & 0.1 & 0.09 & 0.072 \\ 0 & 1 & 0.9 & 0.72 \\ 0 & 0 & 1 & 0.8 \\ 0 & 0 & 0 & 1 \end{bmatrix}
$$

$$
\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = B^T \mu + B^T C^T U + B^T \boldsymbol{\varepsilon}
$$

After running Algorithm 1 we obtained following estimations for B, C and μ

$$\hat{\mu} = \begin{bmatrix} 0.7994 \\ -0.9044 \\ 0.0113 \\ -0.4894 \end{bmatrix} \qquad \hat{C} = \begin{bmatrix} -0.1271 & 0 & 0.1665 & 0 \\ 0.3389 & 0 & 0 & 0.4113 \end{bmatrix} \qquad \hat{B} = \begin{bmatrix} 1 & 0.1024 & 0.0920 & 0.0566 \\ 0 & 1 & 0.8985 & 0.7134 \\ 0 & 0 & 1 & 0.7945 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

After running algorithm 2 we obtained following estimation for T

$$\hat{T} = \begin{bmatrix} 0 & 0.1024 & 0 & 0 \\ 0 & 0 & 0.8985 & 0 \\ 0 & 0 & 0 & 0.7945 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$