

Hierarchical Causal Models

Eli N. Weinstein

*Department of Chemistry
Technical University of Denmark
Kgs. Lyngby, DK 2800, Denmark*

ENAWE@DTU.DK

David M. Blei

*Department of Computer Science and Department of Statistics
Columbia University
New York, NY 10027, USA*

DAVID.BLEI@COLUMBIA.EDU

Editor: Kun Zhang

Abstract

Causal questions often arise in settings where data are hierarchical: subunits are nested within units. Consider students in schools, cells in patients, or cities in states. In these settings, unit-level variables (e.g., a school’s budget) may affect subunit-level outcomes (e.g., student test scores), and subunit-level characteristics may aggregate to influence unit-level outcomes. In this paper, we show how to analyze hierarchical data for causal inference. We introduce hierarchical causal models, which extend structural causal models and graphical models by incorporating inner plates to represent nested data structures. We develop a graphical identification technique for these models that generalizes do-calculus. We show that hierarchical data can enable causal identification even when it would be impossible with non-hierarchical data—for example, when only unit-level summaries are available. We develop estimation strategies, including using hierarchical Bayesian models. We illustrate our results in simulation and through a reanalysis of the classic “eight schools” study.

Keywords: causality, graphical models, multiple environments, Bayesian models

1. Introduction

We propose *hierarchical causal models* (HCMs), a graphical modeling framework for causal inference from nested or hierarchical data. Hierarchical graphical models are a mainstay of modern Bayesian statistics, but causal graphical models have largely remained “flat”. In contrast, HCMs describe causal relationships among nested variables. We develop the semantics and theory behind HCMs, and we derive identification and estimation tools for using them to answer causal questions from hierarchical data. We show how HCMs can help us exploit the nested structure of the data to correct for confounders and other sources of causal bias.

Why study hierarchical causal models? Many phenomena across the natural and social sciences can be framed in terms of nested data. Consider the following domains where hierarchical causal modeling could be useful.

1. *Political science.* We observe citizens (subunits) within states (units). How do individual citizens’ political preferences (a subunit variable) determine which political party governs (a unit

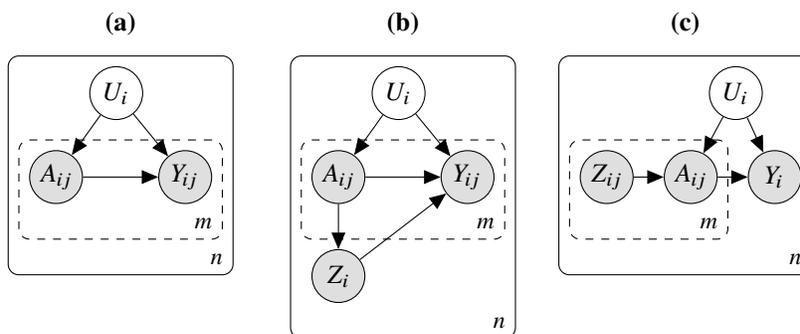


Figure 1: Example hierarchical causal models. Figure 1a: the UNIT CONFOUNDER graph. Figure 1b: the UNIT CONFOUNDER & UNIT INTERFERENCE graph. Figure 1c: the SUBUNIT INSTRUMENT graph.

variable)? How do states’ economic policies (a unit variable) affect citizens’ incomes (a subunit variable)?

2. *Biology.* We observe cells (subunits) within patients (units). How do individual cells’ genetic mutations (a subunit variable) determine whether the patient develops cancer (a unit variable)? How do patients’ chemotherapy treatments (a unit variable) determine cells’ survival (a subunit variable)?
3. *Physical chemistry.* We observe molecules (subunits) within a gas (unit). How does altering molecules’ motion (a subunit variable) alter the pressure exerted by the gas (a unit variable)? How does increasing the temperature of the gas’s container (a unit variable) alter individual molecules’ motion (a subunit variable)?

In all these settings, we may encounter complex causal graphs, with many different unit and subunit-level variables affecting one another. This paper shows how to estimate causal effects from data following given graph.

As a simplest first example, consider Figure 1a, where we observe a treatment a and an outcome y , and our goal is estimate the causal effect of a on y . For example, consider data from students in different schools, where a is a test preparation course and y is test score. The challenge is that this graph also contains an unobserved confounder u , which affects both a and y . In a typical flat causal graph, u would prevent us from learning the causal effect. However, in Figure 1a, the data is nested: we observe a_{ij} and y_{ij} for different subunits (j) within each unit (i), e.g. different students (j) within each school (i). The confounder u_i , meanwhile, is at the unit level, so it is fixed across subunits. For example, a school’s budget is a per-school quantity that affects individual students’ test preparation and test scores. Using our HCM methodology, we can show that the causal effect of a on y is identifiable, thanks to the nested structure of the data. Indeed, identification recovers classical results on fixed-effect models (Wooldridge, 2005; Angrist and Pischke, 2009; Witty et al., 2020).

Besides Figure 1a, there are many other causal graphs a scientist might consider, and causal questions they might ask (Figure 1; Definition 1). Besides confounding, there may be interference, as tutoring one student influences the test scores of others. Besides student outcomes, there may be outcomes at the school level, such as whether the school is ranked highly on a statewide list. For

many graphs, and many causal questions, fixed-effect estimators do not provide a reliable estimate, or cannot be applied to the data at all. Our identification and estimation tools let us reason about different hierarchical causal graphs, and derive appropriate methods for each.

We begin by formally defining hierarchical causal models (HCMs), a causal generalization of hierarchical probabilistic models. Following Pearl’s hierarchy, we propose a definition of HCMs in terms of structural equations with deterministic causal mechanisms (Pearl, 2009; Bareinboim et al., 2021). Next we analyze causal identification in HCMs, to determine when and how we can estimate causal effects from data. We consider arbitrary graphs and nonparametric causal mechanisms, and we develop techniques for causal identification in HCMs that build on the do calculus. Finally, we craft new causal estimation techniques, by extending methods for inference in hierarchical probabilistic models.

Overall, this paper presents a broad toolkit for accomplishing causal inference with nested data. The approach systematically extends hierarchical probabilistic and Bayesian models, to encompass interventions and causality. This framework encompasses many previous causal methods as special cases, such as fixed-effect models and clustered interference models (Wooldridge, 2005; Angrist and Pischke, 2009; Goetgeluk and Vansteelandt, 2008; Neuhaus and Kalbfleisch, 1998; Greenland, 2000; Jensen et al., 2020) (Appendix M). More importantly, however, it enables causal inference across novel problem settings and datasets, with different causal graphs.

Related work. There has been substantial research into causal inference from hierarchical data. Many of these proposed models can be understood as instances of HCMs with particular graphs and particular parametric assumptions (Appendix M).

Fixed-effects models are widely used for correcting for unit-level confounding (Wooldridge, 2005, 2010, Chapters 10-11). One way to interpret these methods is as HCMs that follow the `UNIT CONFOUNDER` graph, with a particular linear parameterization for the mechanism generating Y and sometimes with additional observed subunit-level and unit-level confounders; see Appendix M.1. Closely related are difference-in-difference and synthetic control methods, which can be understood as following the same graph, but use alternate model parameterizations (Chapter 5 Angrist and Lavy, 1999; Abadie et al., 2010). Hierarchical Bayesian methods are often used for inference in fixed-effects models (Gelman and Hill, 2006; Hill, 2013; Feller and Gelman, 2015). More recently, researchers have considered extensions of fixed-effects models that allow for nonparametric causal mechanisms. Specifically, Witty et al. (2020) propose an HCM following the `UNIT CONFOUNDER` graph with a nonparametric parameterization for the mechanism generating Y , along with a Gaussian process-based inference method. We build on this work, studying arbitrary graphs with nonparametric mechanisms.

This paper also relates to clustered interference, spillover effects, and peer effects (Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012; Wooldridge, 2005, 2010, Chapter 11). Many of these models can be understood as HCMs with an unobserved unit-level variable between the subunit treatment and outcome, and a linear parameterization of the causal mechanisms; see Appendix M.2. The unobserved unit-level variable gives rise to interference among the subunits within each unit. Other models, however, go beyond by allowing for general interaction between subunits, not just those mediated by a unit-level variable (Hudgens and Halloran, 2008; Ogburn and VanderWeele, 2014; Sävje et al., 2021).

Another line of related work considers instrumental variable models in multi-site trials, that is, repeated across multiple units (Raudenbush et al., 2012; Reardon et al., 2014). These models

correspond to HCMs with an IV graph entirely at the subunit level, together with an unobserved unit confounder, and using linear causal mechanisms; see Appendix M.3.

At a high level, our contribution to this long literature is to formalize and study HCMs under broad assumptions and conditions. Specifically we study (1) arbitrary causal graphs and (2) arbitrary (nonparametric) causal mechanisms. Our key finding is that there are many novel scenarios where causal inference is possible.

Finally, an important thread of research has studied grouped data from multiple “environments”, focusing especially on problems where the causal graph is not fully known. In some work the principal aim is to discover the causal graph (Chapters 2,4 Peters et al., 2017; Tian and Pearl, 2001; Peters et al., 2016; Perry et al., 2022; Guo et al., 2022). In other work, the aim is to develop prediction or estimation methods that are robust to the unknown graph (Arjovsky et al., 2019; Yin et al., 2021; Shi et al., 2021; Krueger et al., 2021). In either case, we can interpret multi-environment models as HCMs in which each unit corresponds to an environment, and the graph of the observed subunit variables is not entirely known. In this paper, however, we focus on problems where the causal graph is known. One consequence is that we do not require invariance assumptions or independent causal mechanism assumptions, as is standard in the multi-environment literature (Chapters 2,4 Peters et al., 2017, 2016; Guo et al., 2025); see Appendix M.4.

2. A first look at hierarchical causal models

Three examples of hierarchical causal models. Causal models describe the world in terms of variables and their impact on one another, mapped out in a graph. A hierarchical causal model (HCM) contains a plate, which denotes a systematic replication of variables within each unit. In an HCM, variables that fall inside the plate are called *subunit-level variables*; variables that fall outside the plate are called *unit-level variables*.

Figure 1a shows a hierarchical causal model we call `UNIT CONFOUNDER`. The confounder variable U_i is a hidden unit-level variable, while the treatment A_{ij} and outcome Y_{ij} are observed subunit-level variables. For example, consider a scenario where the units are schools and the subunits are students within the schools. The treatment A_{ij} is the amount of tutoring each student receives, and the outcome Y_{ij} is their test score. Both tutoring and test scores can be impacted by unobserved school-level variables such as resources, budget, or teaching philosophy. These quantities are captured by the hidden unit-level confounder U_i .

The `UNIT CONFOUNDER` model in Figure 1a captures the setup of many existing methods for causal inference from nested data. For example, we will see that fixed-effect models can be re-derived from Figure 1a, through our identification and estimation procedure. Figures 1b and 1c show two other motivating HCM graphs that represent novel scenarios. Here existing methods for causal inference from nested data do not apply. Our approach will enable nonparametric identification and inference in these new graphs, as well as many others.

Figure 1b is called `UNIT CONFOUNDER & UNIT INTERFERENCE`. It includes unobserved unit-level confounding. But it further captures unit-level interference: the subunit-level treatment variables (A_{ij}) might affect an observable unit-level variable (Z_i) that, in turn, affects the outcome of all subunits (Y_{ij}). For example, if more students are tutored then there may be more school-wide discussion about academic subjects. This discussion might then lead to better academic performance for everyone in the school.

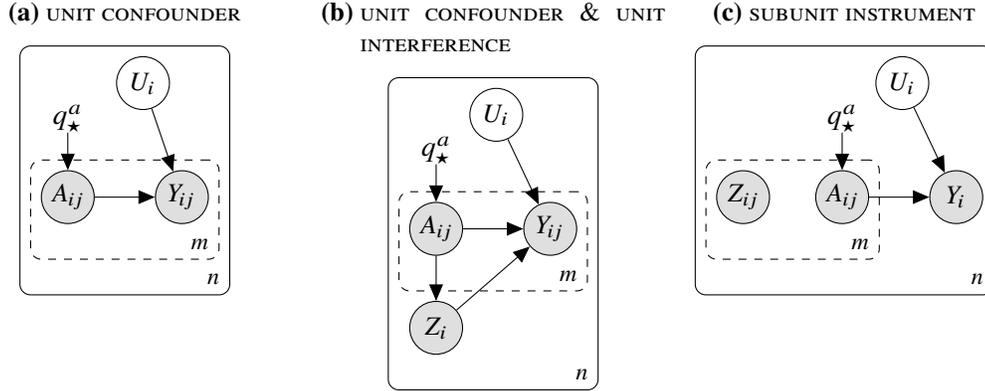


Figure 2: Examples of unconditional soft interventions in HCMs. In these examples, each of the HCMs of Figure 1 is intervened on by drawing $A_{ij} \sim q_{\star}^a(a)$. So, arrows into A_{ij} from other causal variables are removed.

Figure 1c is called **SUBUNIT INSTRUMENT**. Unlike the other two graphs, the outcome variable Y_i is now at the unit level. Further, there is a subunit-level instrument Z_{ij} , which affects the treatment A_{ij} . For example, Y_i might be a school-level outcome variable, such as whether school i is published in a list of best schools. The instrument Z_{ij} might be a randomly administered incentive for students to enroll in extra tutoring.

We will use these three examples as illustrations in our discussion of HCMs, but they are just examples; we are interested in HCMs with any graph. See Figure A3 and Figure A4 for many more examples.

Interventions in hierarchical causal models.. The goal of causal modeling is to study the effect of an intervention. What types of interventions can we consider in an HCM? One possibility is to ask about the effect of assigning every student to receive a_{\star} hours of tutoring, setting $a_{ij} = a_{\star}$ for all i and j . This is a deterministic or *hard* intervention. Another possibility is to ask about the effect of drawing a_{ij} stochastically from a distribution $q_{\star}^a(a)$, of randomly distributing tutoring hours to the students. This is an example of a stochastic policy or *soft* intervention (Chap. 4 Pearl, 2009; Dawid, 2002; Correa and Bareinboim, 2020).

Graphically, an intervention can disconnect the treatment from its parents. Figure 2 illustrates stochastic interventions on the three motifs. Note that these stochastic interventions generalize the deterministic intervention; we write a deterministic intervention as a point-mass $q_{\star}^a(a) = \delta_{a_{\star}}(a)$.

Finally, we consider more targeted interventions. Suppose the HCM contains an additional subunit variable X_{ij} , such as the student’s previous grades. The past performance might naturally affect a student’s seeking out tutoring (A_{ij}) and their test performance (Y_{ij}); see Figure 3a. With this model, we can ask about the effect of providing more tutoring to students with low grades, where the tutoring hours are drawn from a fixed conditional distribution $q_{\star}^{a|x}(a | x_{ij})$; see Figure 3b. This is a *conditional soft* intervention (Chap. 4 Pearl, 2009; Dawid, 2002; Didelez et al., 2006; Correa and Bareinboim, 2020).

A roadmap for formalizing hierarchical causal models.. We study hierarchical causal models with arbitrary graphs. Pearl (2009)’s theory of causality involves models at two levels of descriptive detail: structural causal models and causal graphical models (or causal Bayesian networks). We will

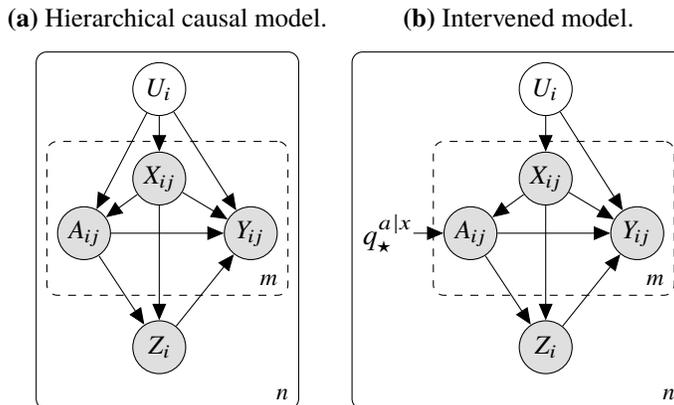


Figure 3: Example of a conditional soft intervention. The HCM (Figure 3a) is intervened on (Figure 3b) by drawing $A_{ij} \sim q_{\star}^{a|x}(a | x_{ij})$. In this intervention, the arrow from the parent X_{ij} to A_{ij} remains, but the unobserved confounder U_i no longer affects A_{ij} .

use and build on these ideas. We first define hierarchical structural causal models, which describe how each variable is generated according to a deterministic causal mechanism (Section 3). We then derive hierarchical causal graphical models, which involve stochastic causal mechanisms (Section 4).

With these formalisms in place, we will turn to identification and estimation in HCMs (Section 5). In the HCM identification problem, we consider infinite data from both units and subunits. In this setting, we will develop a systematic procedure to identify the effects of interventions. We will find, for example, that for each of the HCMs in Figure 1 and interventions in Figure 2, we can identify the effect of the treatment A on the outcome Y . We provide a detailed theory in Section 6. We define our notation where it is first used, but Table A1 also provides a central reference. Code reproducing our empirical results is at <https://github.com/EWeinstein/HCM>.

3. Hierarchical Structural Causal Models

We begin by defining the structural equations of hierarchical causal models, extending classical “flat” structural causal models by introducing subunit-level variables that lie inside an inner plate. In flat structural causal models, we have unit-level variables, affected by unit-level causes. In hierarchical structural causal models, we have both unit-level and subunit-level variables. Each type of variable can affect and be affected by unit-level and subunit-level causes.

3.1 Structural causal models

We first review flat structural causal models (Pearl, 2009; Peters et al., 2017). In a *structural causal model* (SCM), each causal variable is generated through a deterministic function of other causal variables and independent noise.

Figure 4a shows a causal graph. The corresponding structural causal model is

$$\begin{aligned}
 \gamma_i^x &\sim p(\gamma^x) & x_i &= f^x(\gamma_i^x) \\
 \gamma_i^{\bar{a}} &\sim p(\gamma^{\bar{a}}) & \bar{a}_i &= f^{\bar{a}}(\gamma_i^{\bar{a}}) \\
 \gamma_i^{\bar{y}} &\sim p(\gamma^{\bar{y}}) & \bar{y}_i &= f^{\bar{y}}(\bar{a}_i, x_i, \gamma_i^{\bar{y}}),
 \end{aligned} \tag{1}$$

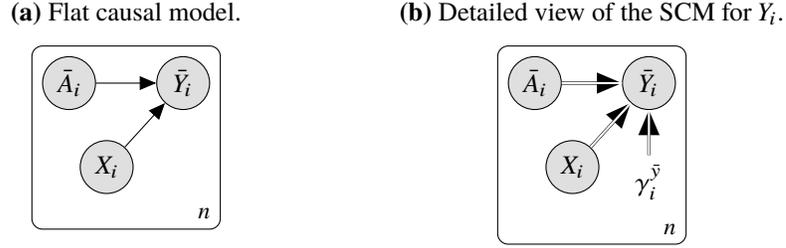


Figure 4: Generating variables in structural causal models. Figure 4a shows a flat causal model. In the corresponding SCM, \bar{Y}_i is generated from its parents \bar{A}_i, X_i and an exogenous noise variable $\gamma_i^{\bar{y}}$ via a deterministic mechanism $f^{\bar{y}}$ (Figure 4b). We depict deterministic mechanisms with double arrows.

for all $i \in \{1, \dots, n\}$. The random variables X_i, \bar{A}_i and \bar{Y}_i are called *endogenous* causal variables; the variables $\gamma_i^x, \gamma_i^a, \gamma_i^{\bar{y}}$ are called *exogenous* noise variables; the deterministic functions $f^x, f^a, f^{\bar{y}}$ are called *mechanisms*.

Each endogenous variable depends on a set of direct causes through its mechanism. Some direct causes are other causal variables in the model—the parents in the causal graph. The other direct cause is the noise variable, which accounts for the remaining unobserved factors that influence the endogenous variable. In Figure 4a, suppose \bar{Y}_i represents average test scores at a school, \bar{A}_i represents average studying hours, and X_i represents teacher quality. The value of the score $\bar{Y}_i = \bar{y}_i$ is determined through the mechanism $f^{\bar{y}}(\bar{a}_i, x_i, \gamma_i^{\bar{y}})$. It is a function of teacher quality, tutoring hours, and its noise variable $\gamma_i^{\bar{y}}$. Here the “noise” might account for textbook choice, school budget, cafeteria menu, and other unobserved factors which affect the test score.

A requirement of an SCM is that the noise variables are i.i.d. across units and independent of one another, and that each appears as an argument in a single mechanism. If an unobserved cause affects more than one causal variable, i.e., it is a *confounder*, then we must include it in the model as an unobserved causal variable.

Given an SCM, we can describe the effects of a hypothetical intervention by modifying its structural equations. For example, consider an unconditional soft intervention on \bar{A} , where it is drawn from $q_\star^{\bar{a}}$. This intervention corresponds to the following modified SCM:

$$\begin{aligned} \gamma_i^x &\sim p(\gamma^x) & x_i &= f^x(\gamma_i^x) \\ & & \bar{A}_i &\sim q_\star^{\bar{a}}(\bar{a}) \\ \gamma_i^{\bar{y}} &\sim p(\gamma^{\bar{y}}) & \bar{y}_i &= f^{\bar{y}}(\bar{a}_i, x_i, \gamma_i^{\bar{y}}). \end{aligned} \tag{2}$$

We can form a hard intervention on \bar{a} with a point-mass $q_\star^{\bar{a}} = \delta_{\bar{a}_\star}$.

3.2 Hierarchical structural causal models

While Figure 4a shows a flat causal graph, Figure 5a shows a hierarchical causal graph, a model of n units where each one contains m subunits. In this graph, the variable X_i is an endogenous unit-level variable; the variables A_{ij} and Y_{ij} are endogenous subunit-level variables. For example, suppose \bar{A}_i and \bar{Y}_i are per-group averages in the flat model of Figure 4a. Then A_{ij} and Y_{ij} from Figure 5a might be disaggregated variables, the individual values that formed the averages.

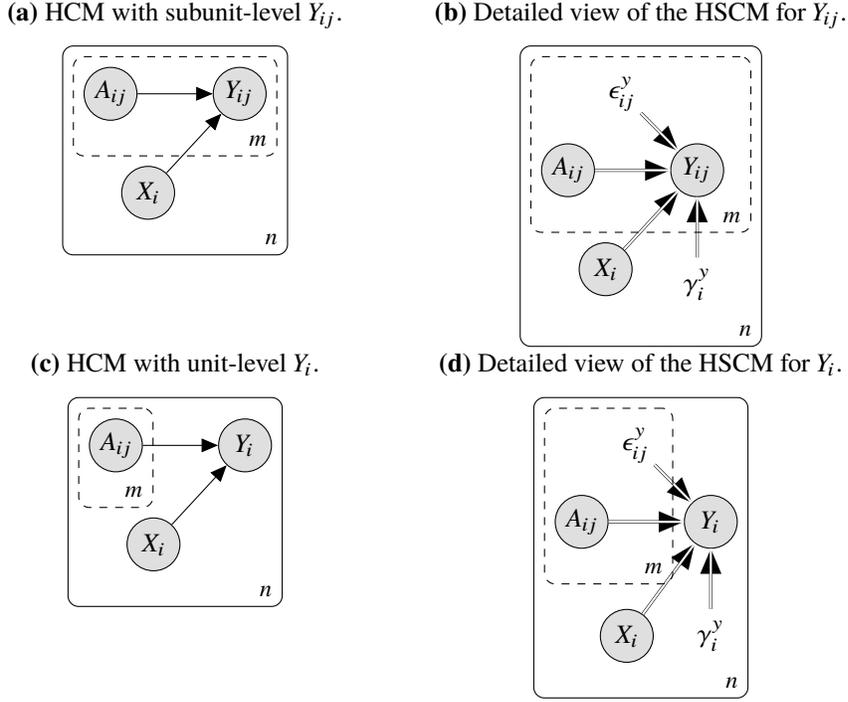


Figure 5: Generating subunit and unit variables in hierarchical structural causal models. Figure 5a and Figure 5c show HCMs. In the corresponding HSCMs, the subunit variable Y_{ij} (Figure 5b) and unit variable Y_i (Figure 5d) are generated from their parents A_{ij}, X_i and from exogenous unit and subunit noise variables $\gamma_i^y, \epsilon_{ij}^y$ via a deterministic mechanism f^y .

Given a hierarchical causal graph, how do we write its *hierarchical structural causal model* (HSCM)? Flat structural causal models only contain unit variables. Now we must account for unit variables and subunit variables.

Subunit-level variables. We first describe how an HSCM generates endogenous subunit variables. In a flat structural causal model, endogenous unit variables are affected by other unit-level variables, including their endogenous parents and exogenous noise. In an HSCM, endogenous subunit variables can be affected by both unit-level and subunit-level variables. These can include both endogenous parents and exogenous noise.

Consider the graph in Figure 5a. The HSCM posits that Y_{ij} is generated as (Figure 5b),

$$\gamma_i^y \sim p(\gamma^y) \quad \epsilon_{ij}^y \sim p(\epsilon^y) \quad y_{ij} = f^y(x_i, \gamma_i^y, a_{ij}, \epsilon_{ij}^y). \quad (3)$$

In this equation, the subunit-level variable Y_{ij} depends on its endogenous unit-level parent X_i and exogenous unit-level noise γ_i^y . It also depends on its endogenous subunit-level parent A_{ij} and exogenous subunit-level noise ϵ_{ij}^y . Since Y_{ij} is specific to subunit j , its subunit-level parents must be in the same subunit, that is, $A_{ij'}$ and $\epsilon_{ij'}^y$ for $j' \neq j$ do not enter into the mechanism for Y_{ij} .

Continuing the running example, suppose Y_{ij} is the test performance of student j in school i . Then, X_i might describe the budget of their school (an endogenous unit-level cause) and A_{ij} might describe the number of hours of tutoring they received (an endogenous subunit-level cause). The unit

level noise variable γ_i^y accounts for unobserved school-level causes, such as the school’s learning environment, the textbooks on their syllabi, or the quality of their teachers. The subunit level noise variable ϵ_{ij}^y accounts for unobserved student-level causes, such as the student’s academic interests or what they had for breakfast the morning of the test.

Notice that for subunit variables, unlike in a flat model, there are two sources of noise. The unit noise γ_i^y is shared across all values of Y_{ij} in unit i . The subunit noise ϵ_{ij} is involved only in determining Y_{ij} . More generally, all the unit- i subunit variables Y_{i1}, \dots, Y_{im} are affected by the same unit level variables, the observed X_i and the unobserved unit-level noise γ_i^y . This pattern of dependence helps capture the intuition that subunits within a unit are more similar to each other than they are to subunits in other units.

Also note the subunit noise ϵ_{ij}^y is independent across subunits. This ensures that the subunits within each unit are exchangeable, just like in hierarchical probabilistic models with inner plates.

Unit-level variables. We now describe how the HSCM generates unit-level endogenous variables. In a flat SCM, endogenous variables are only affected by other unit-level variables. In an HSCM, unit variables can also be affected by subunit variables.

Figure 5c shows a unit-level variable Y_i that depends on subunit variables A_{ij} . Its HSCM posits that Y_i is generated as (Figure 5d),

$$\gamma_i^y \sim p(\gamma^y) \quad \epsilon_{ij}^y \sim p(\epsilon^y) \quad y_i = f^y(x_i, \gamma_i^y, \{(a_{ij}, \epsilon_{ij}^y)\}_{j=1}^m). \quad (4)$$

In this equation, the unit-level variable Y_i depends on its unit-level parent X_i and exogenous unit-level noise γ_i^y . It also depends on its subunit-level parents A_{ij} and exogenous subunit-level noise ϵ_{ij}^y , for $j = 1, \dots, m$. Note Y_i can depend on all the subunits within the unit.

For example, suppose Y_i indicates whether school i is published in a list of best schools. Then X_i might describe school budget and A_{ij} might describe student tutoring hours. The noise γ_i^y accounts for unobserved unit-level causes of inclusion in the list (e.g., teacher quality, school location) while ϵ_{ij}^y accounts for unobserved subunit-level causes (e.g., student extracurricular activities, student achievements).

In an HSCM, we require that the dependence of unit-level variables Y_i on subunit variables A_{ij} and ϵ_{ij}^y is expressed as a dependence on a set of m items $\{(a_{ij}, \epsilon_{ij}^y)\}_{j=1}^m$, so that the value of f^y does not depend on the order of the elements of this set.¹ This assumption reflects the idea that subunits are exchangeable within each unit, as in hierarchical probabilistic models with inner plates. Intuitively, the requirement ensures there are no *a priori* privileged subunits. For example, the school’s appearance on the “best schools” list Y_i cannot just depend on how much the first student in the data is tutored, A_{i1} .

3.3 Theory

In this section, we propose a formal definition of hierarchical structural causal models. We clarify what is required to specify them, and discuss further their relationship to flat structural causal models. To see concrete examples of HSCMs, skip to Section 3.4.

An HSCM has endogenous variables X^1, \dots, X^V , each generated according to its mechanism f^1, \dots, f^V . The variables are ordered causally, such that X^v can only depend on X^1, \dots, X^{v-1} . Let

1. More precisely, we use $\{x_j\}_{j=1}^m$ to denote a *multiset* or *bag*, as identical elements are allowed to appear more than once, e.g. if $x_j = x_{j'}$ for $j' \neq j$.

$\mathcal{V} = \{1, \dots, V\}$. Some variables X_{ij}^v are subunit-level, and they fall inside an inner plate. We denote their coordinate $\mathcal{S} \subseteq \mathcal{V}$ and they are indexed by both the unit i and subunit j . The rest of the variables X_i^v are unit-level, and fall outside the inner plate. We denote their coordinates \mathcal{U} , where $\mathcal{S} \cup \mathcal{U} = \mathcal{V}$; they are indexed only by their unit i . Let $\text{pa}(v) \subseteq \{1, \dots, v-1\}$ denote the indices of the parents of X^v in the graph. The parents that are subunit-level are denoted $\text{pa}_{\mathcal{S}}(v)$; the parents that are unit-level are denoted $\text{pa}_{\mathcal{U}}(v)$.

Each endogenous variable X^v depends on parents, unit-level noise, and subunit-level noise. For the subunit-level variables X_{ij}^v , they depend on unit-level variables $X_i^{\text{pa}_{\mathcal{U}}(v)}$, unit-level noise γ_i^v , and the subunit variables from subunit j , specifically, subunit-level noise ϵ_{ij}^v and subunit-level parents $X_{ij}^{\text{pa}_{\mathcal{S}}(v)}$. For the unit-level variables X_i^v , they depend on other unit-level variables $X_i^{\text{pa}_{\mathcal{U}}(v)}$ and γ_i^v , and *all* the subunit variables ϵ_{ij}^v and parents $X_{ij}^{\text{pa}_{\mathcal{S}}(v)}$ for $j \in \{1, \dots, m\}$.

Thus we define an HSCM as follows.

Definition 1 (Hierarchical structural causal model) *A hierarchical structural causal model (HSCM) \mathcal{M}^{scm} is defined by (1) a directed acyclic graph \mathcal{G} , (2) a set of endogenous variables $X^{\mathcal{V}}$, of which $X^{\mathcal{S}}$ are subunit-level, (3) probability distributions $\text{p}(\gamma^v)$ and $\text{p}(\epsilon^v)$ over unit-level and subunit-level noise variables for all $v \in \mathcal{V}$, and (4) a set of mechanisms $f^{\mathcal{V}}$. Each endogenous variable X^v for $v \in \mathcal{V}$ is generated as,*

$$\begin{aligned} \gamma_i^v &\sim \text{p}(\gamma^v) \\ \epsilon_{ij}^v &\sim \text{p}(\epsilon^v) \\ x_{ij}^v &= f^v(x_i^{\text{pa}_{\mathcal{U}}(v)}, \gamma_i^v, x_{ij}^{\text{pa}_{\mathcal{S}}(v)}, \epsilon_{ij}^v) && \text{if } v \in \mathcal{S} \\ x_i^v &= f^v(x_i^{\text{pa}_{\mathcal{U}}(v)}, \gamma_i^v, \{(x_{ij}^{\text{pa}_{\mathcal{S}}(v)}, \epsilon_{ij}^v)\}_{j=1}^m) && \text{if } v \in \mathcal{U}, \end{aligned} \tag{5}$$

for $j \in \{1, \dots, m\}$ and $i \in \{1, \dots, n\}$.

This gives an axiomatic definition for hierarchical causal models (Pearl, 2009, Chap. 7).

With this class of models in place, we now describe the interventions we will consider. Our analysis focuses on hard interventions on unit variables, and soft interventions on subunit variables. In particular, we study soft interventions that may condition on parent subunit variables, though not unit variables. It is possible to define other types of interventions in HSCMs, such as soft interventions on subunit variables. But these are outside the scope of the identification and estimation methods we develop in this paper, and are left for future work.

Definition 2 (HSCM after intervention) *An intervention Δ on a hierarchical structural causal model is defined by (1) a set of variables $\mathcal{I} \subseteq \mathcal{V}$ that are intervened on, (2) the values $\{x_{\star}^v : v \in \mathcal{I} \cap \mathcal{U}\}$ that the unit-level variables are set to, and (3) the distributions $\{q_{\star}^{v|\text{pa}_{\mathcal{S}}(v)}(x^v | x^{\text{pa}_{\mathcal{S}}(v)}) : v \in \mathcal{I} \cap \mathcal{S}\}$ that the subunit-level variables are drawn from. Post-intervention, the mechanisms generating the variables $X^v : v \in \mathcal{I}$ are,*

$$\begin{aligned} X_{ij}^v &\sim q_{\star}^{v|\text{pa}_{\mathcal{S}}(v)}(x^v | x_{ij}^{\text{pa}_{\mathcal{S}}(v)}) && \text{if } v \in \mathcal{I} \cap \mathcal{S} \\ x_i^v &= x_{\star}^v && \text{if } v \in \mathcal{I} \cap \mathcal{U}, \end{aligned} \tag{6}$$

for $j \in \{1, \dots, m\}$ and $i \in \{1, \dots, n\}$.

Note this class of interventions includes hard interventions on subunit variables, which correspond to the special case of using a fixed delta mass as a soft intervention.

Which variables should be included in an HSCM? In a flat SCM, any variable that has more than one child – i.e. a confounder – must be included as an endogenous variable. Variables with one or no children, by contrast, can be marginalized out of the model, soaked into their noise variables and mechanisms (see e.g. Spirtes, 2010; Richardson and Spirtes, 2002; Janzing and Meija, 2022, Def. 5,6). In an HSCM, unlike an SCM, we cannot marginalize out any variable with only one child. Rather, we must include in an HSCM those single-child unit-level variables for which both their child and a parent are subunit-level; see Appendix A. The variable Z in the “interference” graph of Figure 1b is an example of such an “interferer” variable. Just as we must include confounders, whether observed or not, in a flat SCM, we must include interferers (as well as confounders) in an HSCM.

One might ask why the distribution of subunit noise $p(\epsilon^v)$ does not vary across units. In fact, Definition 1 can cover such situations without loss of generality; variation in $p(\epsilon^v)$ can be absorbed into the unit noise γ_i^v and mechanism f^v . One may also ask why subunit noise is included in the equation for generating unit variables. It is true the subunit noise does not increase the model’s expressivity in describing observational and interventional distributions (see Section 4.2 below). But it does increase the model’s expressivity in describing counterfactual distributions (Appendix B).

Relationship to classical SCMs How does the HSCM formalism relate to classical SCMs?

First, HSCMs add subunit-level variables to SCMs. We recover a classical SCM if there are no subunit-level variables, i.e. if $\mathcal{S} = \emptyset$. (We can absorb the subunit noise $\{\epsilon_{ij}^v\}_{j=1:m}$ into the unit noise γ_i^v in Equation (5).) Second, we recover a classical SCM if there is no variability among subunits. That is, if ϵ^v is constant for all i and j . (Absorb ϵ^v into f^v in Equation (5).) The resulting model has the same graph as the HSCM, but with the inner plate erased. Third, we recover a classical SCM if there is just one subunit. That is, if $m = 1$. (Absorb ϵ_{i1}^v into γ_{i1}^v in Equation (5).) Fourth, we recover a flat SCM if there are no unit-level variables and no unit-level variability. That is, if $\mathcal{U} = \emptyset$ and $p(\gamma^{\mathcal{Z}}) = \delta_{\tilde{\gamma}}$. (Absorb $\gamma_i^v = \tilde{\gamma}$ into f^v in Equation (5).) From this perspective, HSCMs allow for variation in causal mechanisms across different units.

Finally, one might ask also how an HSCM model relates to an SCM with the same graph, but where the inner plate is expanded (Figure A1). The difference is that the HSCM places the restriction that each subunit’s variables are generated by the same mechanism, i.e. the mechanism generating a_{i1} from z_{i1} is the same as that generating a_{i2} from z_{i2} . Further, the mechanism for each unit variable must be invariant to the ordering of its parent subunit variables and subunit noise.

3.4 Examples of hierarchical structural causal models

We now return to the three examples of hierarchical causal models in Figure 1. For each, we detail the corresponding hierarchical structural causal model.

The UNIT CONFOUNDER graph in Figure 1a has a unit-level variable U_i and subunit-level variables A_{ij} and Y_{ij} . It corresponds to the following HSCM:

$$\begin{aligned}
 \gamma_i^u &\sim p(\gamma^u) & \epsilon_{ij}^u &\sim p(\epsilon^u) & u_i &= f^u(\gamma_i^u, \{\epsilon_{ij}^u\}_{j=1}^m) \\
 \gamma_i^a &\sim p(\gamma^a) & \epsilon_{ij}^a &\sim p(\epsilon^a) & a_{ij} &= f^a(u_i, \gamma_i^a, \epsilon_{ij}^a) \\
 \gamma_i^y &\sim p(\gamma^y) & \epsilon_{ij}^y &\sim p(\epsilon^y) & y_{ij} &= f^y(u_i, \gamma_i^y, a_{ij}, \epsilon_{ij}^y),
 \end{aligned} \tag{7}$$

for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. Note in some scenarios, subunit noise is unlikely to contribute to U_i ; for example, if U_i consists only of factors that are determined before the start of the school year, they are unlikely to depend on student-level variables. In this special case, f^u will not depend on any subunit noise ϵ_{ij}^u , and so $u_i = f^u(\gamma_i^u)$.

The UNIT CONFOUNDER & UNIT INTERFERENCE graph in Figure 1b includes a unit-level variable, Z_i , with an observed subunit-level parent, A_{ij} . It corresponds to the following HSCM:

$$\begin{aligned}
 \gamma_i^u &\sim p(\gamma^u) & \epsilon_{ij}^u &\sim p(\epsilon^u) & u_i &= f^u(\gamma_i^u, \{\epsilon_{ij}^u\}_{j=1}^m) \\
 \gamma_i^a &\sim p(\gamma^a) & \epsilon_{ij}^a &\sim p(\epsilon^a) & a_{ij} &= f^a(u_i, \gamma_i^a, \epsilon_{ij}^a) \\
 \gamma_i^z &\sim p(\gamma^z) & \epsilon_{ij}^z &\sim p(\epsilon^z) & z_i &= f^z(\gamma_i^z, \{(a_{ij}, \epsilon_{ij}^z)\}_{j=1}^m) \\
 \gamma_i^y &\sim p(\gamma^y) & \epsilon_{ij}^y &\sim p(\epsilon^y) & y_{ij} &= f^y(z_i, u_i, \gamma_i^y, a_{ij}, \epsilon_{ij}^y),
 \end{aligned} \tag{8}$$

for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. In this model, different subunits can indirectly impact one another: for $j' \neq j$, A_{ij} does not directly impact $Y_{ij'}$ via f^y , but A_{ij} does impact Z_i , which in turn impacts $Y_{ij'}$. Hierarchical structural causal models can thus describe a special case of interference or spillover between subunits, namely interference that goes through a unit-level variable (Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012). (See Appendix M.2 for more detailed connections to notions of ‘‘clustered’’ interference.)

Finally, the SUBUNIT INSTRUMENT graph in Figure 1c corresponds to the following HSCM:

$$\begin{aligned}
 \gamma_i^u &\sim p(\gamma^u) & \epsilon_{ij}^u &\sim p(\epsilon^u) & u_i &= f^u(\gamma_i^u, \{\epsilon_{ij}^u\}_{j=1}^m) \\
 \gamma_i^z &\sim p(\gamma^z) & \epsilon_{ij}^z &\sim p(\epsilon^z) & z_{ij} &= f^z(\gamma_i^z, \epsilon_{ij}^z) \\
 \gamma_i^a &\sim p(\gamma^a) & \epsilon_{ij}^a &\sim p(\epsilon^a) & a_{ij} &= f^a(u_i, \gamma_i^a, z_{ij}, \epsilon_{ij}^a) \\
 \gamma_i^y &\sim p(\gamma^y) & \epsilon_{ij}^y &\sim p(\epsilon^y) & y_i &= f^y(u_i, \gamma_i^y, \{(a_{ij}, \epsilon_{ij}^y)\}_{j=1}^m),
 \end{aligned} \tag{9}$$

for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$.

4. Hierarchical Causal Graphical Models

While structural causal models (SCMs) describe causal processes using deterministic mechanisms and exogenous noise, *causal graphical models* (CGMs), also known as *causal Bayesian networks* or *agnostic causal DAGs*, describe causal processes with stochastic mechanisms (Shalizi, 2013; Lauritzen, 2001; Bareinboim et al., 2021; Richardson and Robins, 2013). Any classical flat SCM can be written as a CGM by integrating out the noise. Here we apply the same idea to develop *hierarchical causal graphical models* (HCGMs), deriving them from hierarchical structural causal models (Definition 1).

4.1 Causal graphical models

We first review how to derive a classical causal graphical model from a classical SCM. Consider again the flat causal model in Figure 4a and recall the structural equation for \bar{Y} in Equation (1). We can integrate out the noise γ^y to form a stochastic mechanism describing how \bar{Y} is generated given X and \bar{A} ,

$$\bar{Y}_i \sim p(\bar{y} \mid x_i, \bar{a}_i) \quad \text{where } P(\bar{Y} \in \Xi \mid X = x, \bar{A} = \bar{a}) = \int \mathbb{1}(f^y(x, \bar{a}, \gamma^y) \in \Xi) p(\gamma^y) d\gamma^y.$$

Note that this derivation is made possible by the assumption in SCMs that the generative process is *stable* across units, in the sense that the mechanism $f^{\bar{y}}$ is fixed and the noise $\gamma_i^{\bar{y}}$ is i.i.d. across units.

If we repeat this derivation for all the endogenous variables in an SCM, we find an alternative characterization of the model as a cascade of random variables. The distribution of each, conditional on its parents, is formed by integrating out the random noise from its deterministic mechanism. For example, the model of Figure 4a and Equation (1) corresponds to the following CGM,

$$\begin{aligned} X_i &\sim p(x) \\ \bar{A}_i &\sim p(\bar{a}) \\ \bar{Y}_i &\sim p(\bar{y} \mid \bar{a}_i, x_i). \end{aligned} \tag{10}$$

A probabilistic graphical model represents the family of joint distributions that respect the factorization implied by the graph. A causal graphical model additionally implies a distribution under intervention on the underlying SCM. Interventions in a CGM work the same way as in an SCM, where we replace the mechanism of a variable with an intervened mechanism. Consider a soft unconditional intervention on \bar{A}_i , where it is drawn from an intervention distribution $q_{\star}^{\bar{a}}$. This intervention results in the following CGM,

$$\begin{aligned} X_i &\sim p(x) \\ \bar{A}_i &\sim q_{\star}^{\bar{a}}(\bar{a}) \\ \bar{Y}_i &\sim p(\bar{y} \mid \bar{a}_i, x_i). \end{aligned} \tag{11}$$

The intervened CGM derives directly from the intervened SCM, again by integrating out the noise. Note while CGMs can describe interventions, they differ from SCMs in that they cannot describe counterfactuals (Pearl, 2009, Chap. 1). This paper focuses on interventions.

4.2 Hierarchical causal graphical models

We described how flat CGMs are derived from flat SCMs; now we apply the same reasoning to derive hierarchical causal graphical models from hierarchical structural causal models.

Subunit variables. We first consider the subunit variables. Recall the example graph in Figure 5a and the structural equation for subunit variable Y_{ij} in Equation (3). We will form the distribution of Y_{ij} in two stages. First, we consider it as a random variable within unit i . Then we consider the variation in its conditional distribution across units.

What is the distribution of Y_{ij} *within* its unit i ? To derive this distribution, we hold the unit variables γ_i^y and X_i fixed, and marginalize out the subunit noise ϵ_{ij}^y . The result is the conditional distribution of $Y_{ij} \mid A_{ij}$ for subunits of unit i , which we denote $Q_i^{y|a}$. In the running example, $Q_i^{y|a}(Y = y \mid A = a)$ describes the probability that a student at school i will receive a test score y after studying for a hours.

Now consider the variation of $Q_i^{y|a}$ *across* units. This conditional distribution depends deterministically on the unit noise γ_i^y as well as the unit-level parents x_i . If we marginalize out γ_i^y , we produce a two-stage generative process for Y_{ij} :

$$\begin{aligned} Q_i^{y|a} &\sim p(q^{y|a} \mid x_i) \\ Y_{ij} &\sim q_i^{y|a}(y \mid a_{ij}). \end{aligned} \tag{12}$$

Notice that $p(q^{y|a} | x_i)$ is a *distribution over distributions*. It describes how the conditional distribution of Y given A varies across units, due to unit-level noise, i.e., unobserved unit-level causes of Y . Continuing the running example, $p(q^{y|a} | x_i)$ tells us how the effectiveness of the tutoring program at producing good test scores changes across schools.

More formally, we can derive the conditional distribution of $Q^{y|a}$ in two stages. First we define the distribution that marginalizes out the subunit-level noise from a mechanism, holding the unit-level noise fixed. Then we marginalize out the unit-level noise. The result is,

$$\begin{aligned} [g_q^{y|a}(x, \gamma^y)](Y \in \Xi | A = a) &\triangleq \int \mathbb{1}(f^y(x, \gamma^y, a, \epsilon^y) \in \Xi) p(\epsilon^y) d\epsilon^y \\ P(Q^{y|a} \in \Pi | X = x) &\triangleq \int \mathbb{1}([g_q^{y|a}(x, \gamma^y)] \in \Pi) p(\gamma^y) d\gamma^y. \end{aligned}$$

Here $g_q^{y|a}(x, \gamma^y)$ is a function which takes as input x and γ^y and returns a distribution $q^{y|a}$. Thus $q_i^{y|a}(y | a) = [g_q^{y|a}(x_i, \gamma_i^y)](y | a)$. The second line defines $p(q^{y|a} | x_i)$. Note this derivation is made possible by the HSCM assumption that the generative process is stable across subunits (as well as units), in the sense that the mechanism f^y is fixed and the noise ϵ_{ij}^y is i.i.d. across subunits.

Setting aside for a moment the possibility of interventions, Equation (12) takes the form of a hierarchical *probabilistic* model. First, for each unit i , we draw a conditional distribution over subunit variables $Q_i^{y|a}$. Second, for each subunit j within each unit i , we draw the subunit variable Y_{ij} from $Q_i^{y|a}(\cdot | a_{ij})$. The subunit variables Y_{i1}, \dots, Y_{im} within each unit i are thus similar, as they are drawn from the same distribution $Q_i^{y|a}$. Such hierarchical models of grouped data are a mainstay of Bayesian statistics (e.g. Gelman and Hill, 2006), and the idea of building hierarchical models by drawing random distributions is at the foundations of Bayesian nonparametric statistics (e.g. Ghosh and Ramamoorthi, 2003).

Unit variables. We now turn to the unit-level variables. Consider the variable Y in the example graph in Figure 5c. Recall from Equation (4) that Y is generated from X and A as,

$$\gamma_i^y \sim p(\gamma^y) \quad \epsilon_{ij}^y \sim p(\epsilon^y) \quad y_i = f^y(x_i, \gamma_i^y, \{(a_{ij}, \epsilon_{ij}^y)\}_{j=1}^m).$$

Since Y_i is unit-level, we can form the random variable by simultaneously marginalizing out both unit noise γ_i^y and subunit noise $\epsilon_{i1}^y, \dots, \epsilon_{im}^y$. The result is,

$$Y_i \sim p(y | x_i, \{a_{ij}\}_{j=1}^m). \tag{13}$$

This form of the conditional stems from the fact that the deterministic mechanism f^y in the HSCM is invariant to permutations of $(a_{i1}, \epsilon_{i1}^y), \dots, (a_{im}, \epsilon_{im}^y)$ and because $\epsilon_{i1}^y, \dots, \epsilon_{im}^y$ are i.i.d.. Thus the stochastic mechanism $p(y | x_i, \{a_{ij}\}_{j=1}^m)$ must also be permutation invariant, depending only on the set of values $\{a_{ij}\}_{j=1}^m$. Equation (13) defines how a unit endogenous variable is generated in a hierarchical causal graphical model.

General case. We now define hierarchical causal graphical models in general.

Definition 3 (Hierarchical causal graphical model) Consider a hierarchical structural causal model \mathcal{M}^{scm} (Definition 1). The corresponding hierarchical causal graphical model \mathcal{M}^{cgm} has the same

graph and endogenous variables, with stochastic mechanisms,

$$\begin{aligned}
 Q_i^{v|\text{pa}_{\mathcal{S}}(v)} &\sim \text{p}(q^{v|\text{pa}_{\mathcal{S}}(v)} | x_i^{\text{pa}_{\mathcal{Z}}(v)}) && \text{for } v \in \mathcal{S} \\
 X_{ij}^v &\sim q_i^{v|\text{pa}_{\mathcal{S}}(v)}(x^v | x_{ij}^{\text{pa}_{\mathcal{S}}(v)}) && \text{for } v \in \mathcal{S} \\
 X_i^v &\sim \text{p}(x^v | x_i^{\text{pa}_{\mathcal{Z}}(v)}, \{x_{ij}^{\text{pa}_{\mathcal{S}}(v)}\}_{j=1}^m) && \text{for } v \in \mathcal{U},
 \end{aligned} \tag{14}$$

for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$.

A derivation of hierarchical causal graphical models from hierarchical structural causal models is given in Appendix C. Note the distribution $Q^{v|\text{pa}_{\mathcal{S}}(v)}$ over a subunit level variable X^v describes the subunit-level variable's dependence on its subunit-level ancestors, while the mechanism $\text{p}(q^{v|\text{pa}_{\mathcal{S}}(v)} | x_i^{\text{pa}_{\mathcal{Z}}(v)})$ generating $Q^{v|\text{pa}_{\mathcal{S}}(v)}$ only depends on the unit-level ancestors, $X_i^{\text{pa}_{\mathcal{Z}}(v)}$. Interventions in hierarchical causal graphical models work just as in hierarchical structural causal models, with the mechanism for the intervened variable replaced by its intervened value (Definition 2).

4.3 Examples

We illustrate hierarchical causal graphical models through the three examples in Figure 1. Consider the UNIT CONFOUNDER graph in Figure 1a. It corresponds to the HSCM in Equation (7), which becomes the following HCGM:

$$\begin{aligned}
 U_i &\sim \text{p}(u) \\
 Q_i^a &\sim \text{p}(q^a | u_i) && A_{ij} \sim q_i^a(a) \\
 Q_i^{y|a} &\sim \text{p}(q^{y|a} | u_i) && Y_{ij} \sim q_i^{y|a}(y | a_{ij}),
 \end{aligned} \tag{15}$$

for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. Figure 6b depicts this model with the Q variables shown explicitly.²

Now turn to the UNIT CONFOUNDER & UNIT INTERFERENCE graph in Figure 1b. Its HSCM is in Equation (8). It yields the following HCGM:

$$\begin{aligned}
 U_i &\sim \text{p}(u) \\
 Q_i^a &\sim \text{p}(q^a | u_i) && A_{ij} \sim q_i^a(a) \\
 Z_i &\sim \text{p}(z | \{a_{ij}\}_{j=1}^m) \\
 Q_i^{y|a} &\sim \text{p}(q^{y|a} | z_i, u_i) && Y_{ij} \sim q_i^{y|a}(y | a_{ij}),
 \end{aligned} \tag{16}$$

for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$ (Figure 6f).

Finally, consider the SUBUNIT INSTRUMENT graph of Figure 1c. Its HSCM is in Equation (9). It yields the following HCGM:

$$\begin{aligned}
 U_i &\sim \text{p}(u) \\
 Q_i^z &\sim \text{p}(q^z) && Z_{ij} \sim q_i^z(z) \\
 Q_i^{a|z} &\sim \text{p}(q^{a|z} | u_i) && A_{ij} \sim q_i^{a|z}(a | z_{ij}) \\
 Y_i &\sim \text{p}(y | \{a_{ij}\}_{j=1}^m),
 \end{aligned} \tag{17}$$

2. Notationally, in HCM graphs such as Figure 1a, we use a dashed rather than a solid line for the inner plate because subunit variables are not conditionally independent given their parent endogenous variables. They are, however, conditionally independent given their Q variable and parent subunit variables. We therefore use a solid line for the inner plate in Figure 6b.

for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$ (Figure 6j).

Hierarchical causal graphical models produce the same post-intervention distribution as the hierarchical structural causal model from which they are derived, since the process generating each variable from its parents is unchanged. A hard intervention on A in the SUBUNIT INSTRUMENT model corresponds to replacing the third line of Equation (17) with $A_{ij} = a_\star$. A conditional soft intervention given Z uses instead $A_{ij} \sim q_\star^{a|z}(a | z_{ij})$. Equivalently, however, we can model either intervention on A as a hard intervention on $Q^{a|z}$, which leaves unchanged the expression $A_{ij} \sim q^{a|z}(a | z_{ij})$. For a hard intervention on A , we replace $Q_i^{a|z} \sim p(q^{a|z} | u_i)$ with $q_i^{a|z} = \delta_{a_\star}$. For a conditional soft intervention, we set $q_i^{a|z} = q_\star^{a|z}$. In short, we can describe a soft intervention on a subunit variable as a hard intervention on their underlying Q distribution.

5. Identification and Estimation

The purpose of causal modeling is to understand the effect of a hypothetical intervention on the system. A crucial step in this process is to solve the problem of *causal identification* (causal ID).

In causal ID we posit a causal model and consider a hypothetical intervention together with a chosen causal quantity. We assume that some variables are observed while others are not. We ask: if given an infinite number of data points from the pre-intervention distribution, can we calculate the post-intervention distribution over the causal quantity? If we can then the quantity is identified. If we cannot, for example because of which variables are unobserved, then the quantity is not identified.

In a hierarchical causal ID problem, we again consider a hypothetical intervention and assume some variables are observed and some are hidden. But we now consider infinite data at both the subunit and unit level. With the data from infinite subunits, we effectively observe the subunit joint $q_i(x^{\delta_{\text{obs}}})$, which is the joint distribution of the observable subunit variables within each unit i . Notice that it is a *random distribution*, its randomness governed by the population distribution over units. With the data from infinite units, we effectively observe the distribution $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\delta_{\text{obs}}}))$, which is the joint distribution of the observable unit variables and the observable subunit distributions. (There are technical details to these claims; see Appendix G.)

Increasing the number of subunits can change how causal variables are generated in the model. We will assume the model converges to a stable limit as the number of subunits increases (Section 6.1). Note this assumption restricts what models can be identified.

We study hierarchical causal ID for the class of interventions in Definition 2. In particular, we focus on (a) hard interventions on unit variables and (b) soft interventions on subunit variables, which may condition on other subunit variables. Note, however, these are in some sense both hard interventions on unit-level variables: Section 4.3 showed that a soft intervention on a subunit variable is equivalent to a hard intervention on the Q distribution for that variable, which is at the unit level.

The idea behind our strategy for hierarchical causal ID is to form a flat causal model from the hierarchical causal model such that causal identification in the flat model is equivalent to identification in the hierarchical model. To transform the hierarchical model, we develop three types of graphical steps: *collapsing*, *augmenting*, and *marginalizing* (Figure 6). When we collapse, we promote the Q variables to endogenous variables, and remove the subunit endogenous variables, forming a flat causal model; when we augment, we add new unit-level endogenous variables; and when we marginalize, we remove some of the unit-level endogenous variables. Finally, with the flat model in hand, we apply the do-calculus to determine whether and how identification is possible. The next sections demonstrate these steps with the three motifs of Figure 1, and then show how

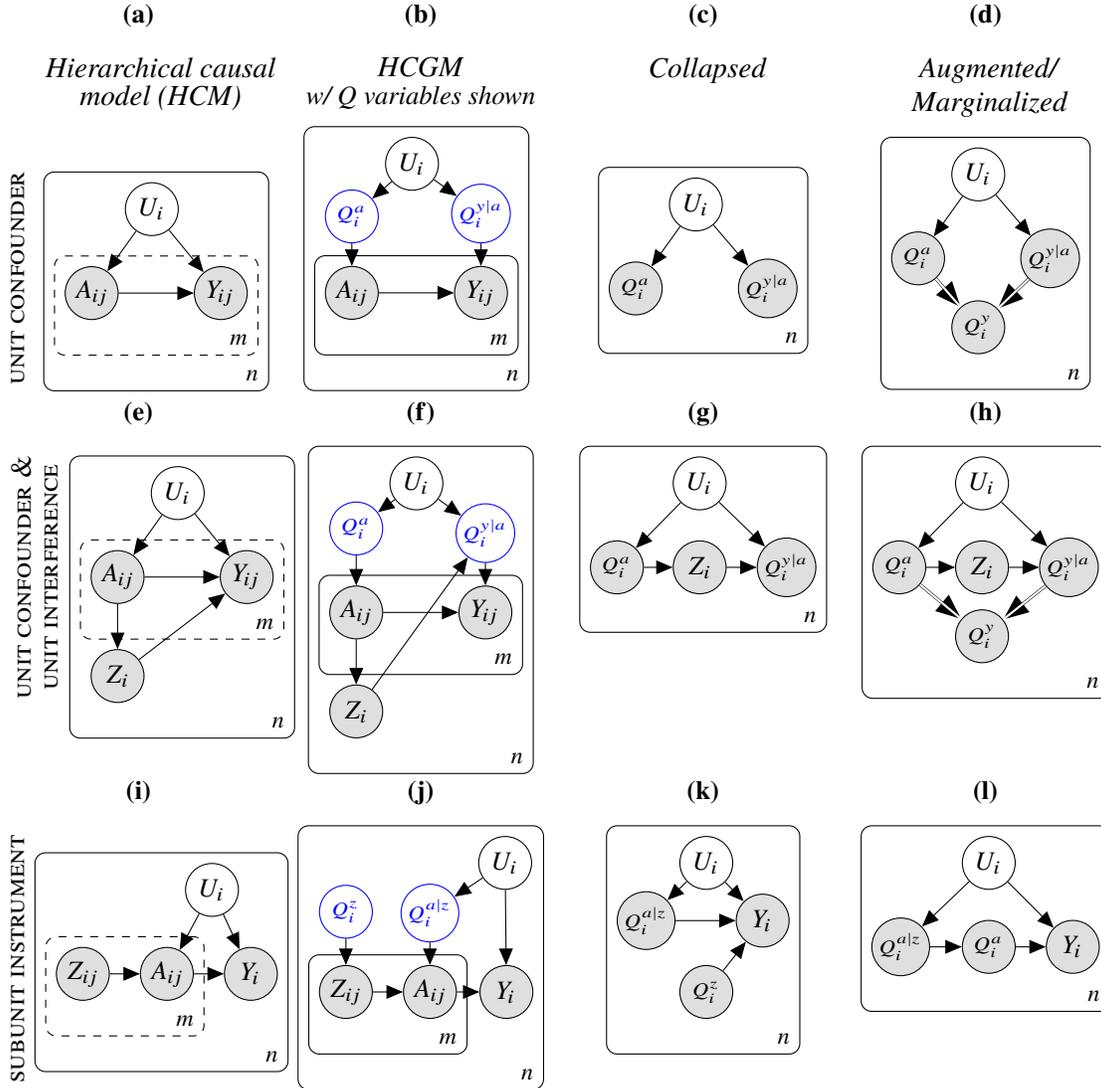


Figure 6: Transformations of example hierarchical causal models. The first column shows three example HCMs, introduced in Section 2 (repeated from Figure 1). The second column shows explicitly each HCM’s latent Q variables, as described in Section 4. The third column shows each HCM’s matching collapsed model, a flat causal model (Section 5). The final column shows the augmented and/or marginalized collapsed model (Section 5). We apply do-calculus to this final model to perform identification in the original HCM.

the resulting ID formulae can be translated into practical estimators in concrete settings. Section 6 presents the theory that justifies these steps, and algorithms for performing them on arbitrary graphs.

5.1 The unit confounder graph

We first study identification for the UNIT CONFOUNDER graph. The hierarchical causal model is in Figure 6a (repeated from Figure 1a); the Q variables are shown explicitly in Figure 6b. In the

example application, the units are schools and the subunits are students within them. The variable A is the number of tutoring hours for a student, Y is their score on a standardized test, and there are school-level confounders U that affect both how tutoring is dispersed and the performance of the students. The target intervention is one where we provide a_\star tutoring hours to all students in all schools, $\text{do}(a = a_\star)$. At the unit level, this intervention is equivalent to setting the subunit distribution of tutoring hours equal to a point mass, $q_\star^a(a) = \delta_{a_\star}(a)$.

The outcome of interest is the average test score Y . Since Y is a subunit variable, we write this estimand as an iterated expectation over units and subunits. The subunit variables are drawn from the subunit distribution $Q(a, y)$. We have:

$$\mathbb{E}[Y; \text{do}(q^a = q_\star^a)] = \mathbb{E}[\mathbb{E}[Y | Q]; \text{do}(q^a = q_\star^a)] = \mathbb{E}_p [\mathbb{E}_Q [Y]; \text{do}(q^a = q_\star^a)]. \quad (18)$$

The inner expectation is

$$\mathbb{E}_Q [Y] = \int \int y Q(a, y) da dy = \int y Q(y) dy.$$

The outer expectation is over the post-intervention distribution of unit-level quantities, including unit variables and subunit distributions. Here,

$$Q(a, y) \sim p(q(a, y); \text{do}(q^a = q_\star^a)).$$

The double expectation $\mathbb{E}_p [\mathbb{E}_Q [Y]; \text{do}(q^a = q_\star^a)]$ expresses the same quantity as $\mathbb{E}[Y; \text{do}(a = a_\star)]$, but expands the expectation to decompose the unit-level and subunit-level randomness.

Step 1: Collapse. The first step of hierarchical causal ID is to derive a *collapsed* model from the HCGM in Figure 6b. The collapsed model is a flat causal model that only contains unit-level variables. It serves as a mathematical tool to prove identification in the original HCGM. The collapsed model includes both the original unit variables and the Q variables, which are now treated as endogenous causal variables in their own right. To obtain the collapsed model, we take $m \rightarrow \infty$, effectively observing the Q_i variables, and then erase the subunit variables from the graph. For the UNIT CONFOUNDER, the collapsed graph is in Figure 6c. It corresponds to the following generative process,

$$\begin{aligned} U_i &\sim p(u) \\ Q_i^a &\sim p(q^a | u_i) \\ Q_i^{y|a} &\sim p(q^{y|a} | u_i). \end{aligned} \quad (19)$$

(In other examples, the collapsing step will be more involved.)

What is important about the collapsed model is that its distribution of unit-level variables is the same as in the hierarchical causal model, both pre- and post-intervention. (This theory is developed in general in Section 6.1.)

Step 2: Augment. In the second step we *augment* the collapsed model, adding new variables that represent quantities which depend on the subunit distribution $Q(a, y)$. In the estimand of Equation (18), the target outcome $\mathbb{E}_Q [Y]$ is an expectation relative to $Q(y)$, the marginal distribution over Y within a unit. $Q(y)$ can be written in terms of unit-level Q variables,

$$q(y) = \int q^a(a) q^{y|a}(y | a) da \triangleq m(q^a, q^{y|a}). \quad (20)$$

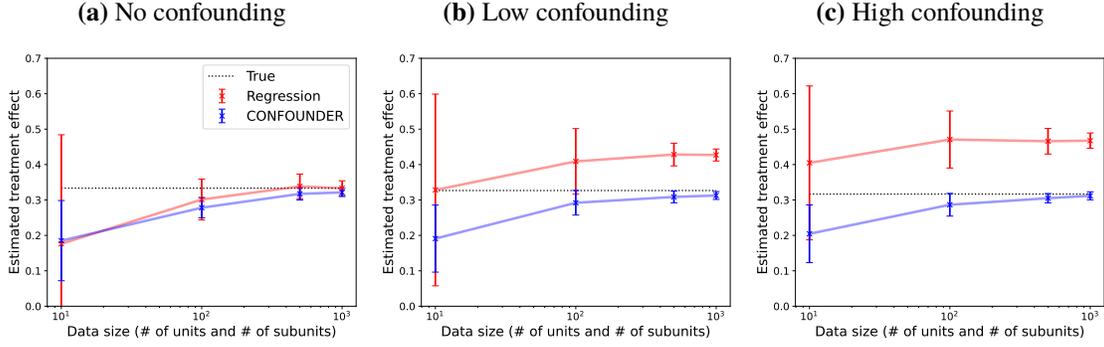


Figure 7: Estimated effects in a UNIT CONFOUNDER simulation. The UNIT CONFOUNDER estimate is based on the identification formula for the true HCM. The “Regression” estimate is based on aggregated data, and so fails to account for confounding. Error bars show standard deviation across 20 independent simulations.

We augment the collapsed model to include $q(y)$ as an additional variable, denoted q^y ; see Figure 6d. This new variable is generated according to a deterministic mechanism $q_i^y = m(q_i^a, q_i^{y|a})$, indicated by double arrows in the graph. In the augmented model, the causal estimand can be written,

$$\mathbb{E}_p \left[\mathbb{E}_{Q^y} [Y] ; \text{do}(q^a = q_\star^a) \right]. \quad (21)$$

Here, recall, the outer expectation $\mathbb{E}_p [\cdot ; \text{do}(q^a = q_\star^a)]$ is over the post-intervention distribution of all unit-level quantities, including unit variables and subunit distributions, observed and hidden.

Step 3: Identify. We reason with the augmented model graph to identify the causal estimand. The causal estimand concerns the effect that an intervention on q^a has on q^y . We apply do-calculus to Figure 6d to identify the intervention distribution via a backdoor correction,

$$\begin{aligned} p(q^y ; \text{do}(q^a = q_\star^a)) &= \int p(q^{y|a}) p(q^y | q_\star^a, q^{y|a}) dq^{y|a} \\ &= \int p(q^{y|a}) m(q_\star^a, q^{y|a}) dq^{y|a}, \end{aligned} \quad (22)$$

where the second line follows since q^y is a deterministic function of its parents. The identified causal estimand uses this distribution in the double expectation of Equation (21).

Step 4: Estimate. We observe data from n units, and m_i subunits within each unit, $\{\{a_{ij}, y_{ij}\}_{j=1}^{m_i}\}_{i=1}^n$. We use this data to approximate the terms of Equation (22) and take the expectation in Equation (21).

1. Estimate the per-unit conditional distribution $\hat{q}_i^{y|a}$ from $\{a_{ij}, y_{ij}\}_{j=1}^{m_i}$. This step amounts to estimating a separate conditional model for each unit, e.g., a set of regression models.
2. Calculate the per-unit marginal distribution

$$\hat{q}_i^y(y) = m(q_\star^a, \hat{q}_i^{y|a}) = \int q_\star^a(a) \hat{q}_i^{y|a}(y | a) da.$$

If q_\star^a is a point mass at a value a_\star then $\hat{q}_i^y(y) = \hat{q}_i^{y|a}(y | a_\star)$. Further calculate the expectation with respect to each marginal, $\hat{\mu}_i^y \triangleq \mathbb{E}_{\hat{q}_i^y} [Y]$.

3. Estimate the population distribution $\hat{p}(q^{y|a})$ with the empirical distribution of $\{\hat{q}_i^{y|a}\}_{i=1}^n$.

With these ingredients, the final estimate of Equation (21) is simply the average of the per-unit expectations,

$$\mathbb{E}_p \left[\mathbb{E}_{Q^y} [Y] ; \text{do}(q^a = q_{\star}^a) \right] \approx \frac{1}{n} \sum_{i=1}^n \hat{\mu}_i^y. \quad (23)$$

As a demonstration, we consider binary A, Y, U , and draw simulated variables from a true UNIT CONFOUNDER model (details in Appendix D.1). Our goal is to estimate the average treatment effect,

$$\mathbb{E}_p \left[\mathbb{E}_{Q^y} [Y] ; \text{do}(q^a = \delta_1) \right] - \mathbb{E}_p \left[\mathbb{E}_{Q^y} [Y] ; \text{do}(q^a = \delta_0) \right]. \quad (24)$$

We observe $\{\{a_{ij}, y_{ij}\}_{j=1}^m\}_{i=1}^n$ for increasing numbers of units and subunits, and where the number of units equals the number of subunits in each (i.e., $n = m$). Figure 7 compares the UNIT CONFOUNDER HCM estimator in Equation (23) with a regression estimator that simply models the per-unit average outcome as a linear function of the per-unit average treatment. We consider three different simulations, with different levels of confounding in each. In each case, as the number of units and subunits increases, the HCM-based estimator converges to the true effect. The regression model, by contrast, only approaches the true effect if there is no confounding.

In summary, we have derived from the UNIT CONFOUNDER graph the methodology of fixed-effect regression models, including nonparametric regression (Wooldridge, 2005; Angrist and Pischke, 2009; Witty et al., 2020) (see Appendix M.1 for further details). In the following sections, we demonstrate how to derive new causal inference methods for novel hierarchical data settings.

5.2 The unit confounder and unit interference graph

We next study the UNIT CONFOUNDER & UNIT INTERFERENCE graph. The hierarchical causal model is in Figure 6e (repeated from Figure 1b); the Q variables are shown explicitly in Figure 6f. The causal estimand is in Equation (18).

Step 1: Collapse. First we collapse the HCGM, taking $m \rightarrow \infty$. In this graph, we must consider the edge between subunit variable a_{ij} and unit variable z_i . Earlier, we required that the mechanism for z_i was invariant to the ordering of the subunit variables a_{ij} . We now make a further requirement: the mechanism for z_i converges as $m \rightarrow \infty$. Specifically, $p(z | \{a_{ij}\}_{j=1}^m)$ must converge to a mechanism $p(z | q_i(a))$, which depends only on the subunit distribution over A .

For example, suppose z_i depends on an empirical average of a function of a_{ij} ,

$$Z_i \sim p \left(z | \{a_{ij}\}_{j=1}^m \right) = p \left(z | \frac{1}{m} \sum_{j=1}^m h(a_{ij}) \right).$$

If $h(a_{ij}) = a_{ij}$ then this average is the mean; if $h(a_{ij}) = \mathbb{1}(a_{ij} > 2)$ then this average is the fraction of subunits for which $a_{ij} > 2$. As $m \rightarrow \infty$ the average converges, $\frac{1}{m} \sum_{j=1}^m h(a_{ij}) \rightarrow \mathbb{E}_{q_i(a)} [h(A)]$. Thus the mechanism of z converges to one that depends only on the subunit distribution $q_i(a)$. (Convergent and divergent mechanisms are discussed in depth in Appendix E.)

With this requirement in place, we can write the collapsed model. It is in Figure 6g, and now draws an arrow directly from Q_i^a to Z_i . The generative process is,

$$\begin{aligned}
 U_i &\sim p(u) \\
 Q_i^a &\sim p(q^a \mid u_i) \\
 Z_i &\sim p(z \mid q_i^a) \\
 Q_i^{y|a} &\sim p(q^{y|a} \mid z_i, u_i).
 \end{aligned} \tag{25}$$

Notice the treatment distribution Q_i^a is connected to the conditional distribution of the outcome $Q_i^{y|a}$ through the interferer Z_i .

Step 2: Augment. Next we augment the graph. The causal estimand involves the within-unit marginal distribution of Y , so again we augment the graph to include Q_i^y . Again it is a deterministic function of Q_i^a and $Q_i^{y|a}$, namely $q_i^y = m(q_i^a, q_i^{y|a})$. The augmented graph is in Figure 6h.

Step 3: Identify. We now apply do-calculus to the augmented graph to identify the causal estimand. We can write the post-intervention distribution of Q^y as,

$$p(q^y; \text{do}(q^a = q_\star^a)) = \int p(q^y \mid q_\star^a, q^{y|a}) p(q^{y|a}; \text{do}(q_\star^a)) dq^{y|a}. \tag{26}$$

The first term is a point mass at the marginal, $q^y(y) = \int q_\star^a(a) q^{y|a}(y \mid a) da$. We identify the second term with a *front-door adjustment*,

$$p(q^{y|a}; \text{do}(q^a = q_\star^a)) = \int p(z \mid q_\star^a) \underbrace{\left(\int p(q^a) p(q^{y|a} \mid q^a, z) dq^a \right)}_{p(q^{y|a}; \text{do}(z))} dz. \tag{27}$$

We identify the estimand with the expectation of Y under the distribution in Equation (26).

Step 4: Estimate. We observe n units, each with m_i subunits: $\{z_i, \{a_{ij}, y_{ij}\}_{j=1}^{m_i}\}_{i=1}^n$. We use the data to form estimates of the elements of Equations (26) and (27).

1. Estimate the i th term \hat{q}_i^a from $\{a_{ij}\}_{j=1}^{m_i}$. For example, if a_{ij} is binary then we can estimate a Bernoulli parameter.
2. Estimate the i th conditional $\hat{q}_i^{y|a}$ from $\{y_{ij}, a_{ij}\}_{j=1}^{m_i}$. For example, if y_{ij} is binary then this estimate can be a pair of Bernoulli parameters $\hat{\pi}_{i,0}$ and $\hat{\pi}_{i,1}$, to parameterize the conditional distribution of Y given $A = 0$ and $A = 1$.
3. Estimate the population conditional $\hat{p}(z \mid q^a)$ from $\{z_i, \hat{q}_i^a\}_{i=1}^n$. For example, if the interference variable z_i is binary then this estimate can be a logistic regression, conditional on the parameters determining \hat{q}_i^a .
4. Estimate the population conditional $\hat{p}(q^{y|a} \mid q^a, z)$ from $\{z_i, \hat{q}_i^a, \hat{q}_i^{y|a}\}_{i=1}^n$. If z_i is binary then we can use two different regression models, each conditional on the parameters determining $\hat{q}_i^{y|a}$.
5. Estimate the population distribution $\hat{p}(q^a)$ with the empirical distribution of $\{\hat{q}_i^a\}_{i=1}^n$.

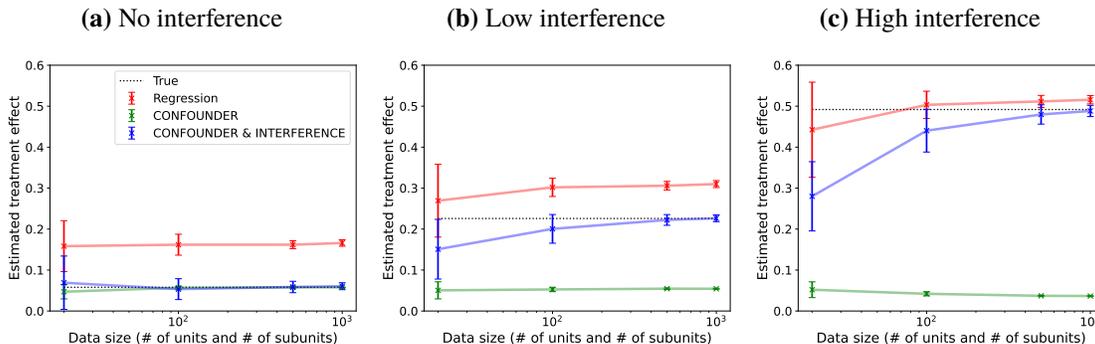


Figure 8: Estimated effects in a CONFOUNDER & INTERFERENCE simulation. The UNIT CONFOUNDER & UNIT INTERFERENCE estimate is based on the identification formula for the true HCM. The UNIT CONFOUNDER estimate is based on the UNIT CONFOUNDER model, which is incorrect in this simulation, as it ignores confounding. The “Regression” estimate is based on aggregated data, and fails to account for confounding or interference. Error bars show standard deviation across 20 independent simulations.

We plug these estimates into Equations (26) and (27) to estimate the intervention distribution.

As a demonstration, we consider binary A , Y , and Z , and continuous U , and draw simulated variables from a true UNIT CONFOUNDER & UNIT INTERFERENCE model (details are in Appendix D.2). We consider three different simulations, each with different levels of interference, i.e., an increasing effect of Z on Y . Here our goal is to estimate a difference between two soft interventions,

$$\mathbb{E}_p \left[\mathbb{E}_{Q^y} [Y] ; \text{do}(q^a = \text{Bern}(0.75)) \right] - \mathbb{E}_p \left[\mathbb{E}_{Q^y} [Y] ; \text{do}(q^a = \text{Bern}(0.25)) \right]. \quad (28)$$

$\text{Bern}(\mu)$ is the Bernoulli distribution with mean μ .

We observe $\{z_i, \{a_{ij}, y_{ij}\}_{j=1}^m\}_{i=1}^n$ for increasing numbers of units and subunits, and where the number of units equals the number of subunits in each. Figure 8 compares the UNIT CONFOUNDER & UNIT INTERFERENCE estimator with the UNIT CONFOUNDER estimator from Section 5.1, as well as the aggregated linear regression estimator discussed in Section 5.1. The linear regression estimator does not account for confounding or interference. The UNIT CONFOUNDER estimator does not account for interference, and we can see its error grow as interference increases. Said another way, fixed-effect models fail with interference. But regardless of the level of interference, the UNIT CONFOUNDER & UNIT INTERFERENCE estimator converges to the true effect with increasing data. As we increase the number of subunits per unit, the bias at low N decreases (Figure A6) (Rainforth et al., 2018, Appendix B).

5.3 The subunit instrument graph

Last we study the SUBUNIT INSTRUMENT graph. The HCM is in Figure 6i (repeated from Figure 1b); the HCGM is in Figure 6j. Here the causal estimand involves a unit-level outcome,

$$\mathbb{E}_p \left[Y ; \text{do}(q^a = q_\star^a) \right]. \quad (29)$$

Step 1: Collapse. We take $m \rightarrow \infty$, remove the subunit variables, and correctly connect the Q variables to the unit-level variables. The collapsed graph is in Figure 6k. The generative process is,

$$\begin{aligned}
 U_i &\sim p(u) \\
 Q_i^z &\sim p(q^z) \\
 Q_i^{a|z} &\sim p(q^{a|z} | u_i) \\
 Y_i &\sim p(y | q_i^{a|z}, q_i^z) = p\left(y \mid q_i(a) = \int q_i^{a|z}(a | z) q_i^z(z) dz\right).
 \end{aligned} \tag{30}$$

Notice the distribution of Y depends on the marginal distribution of subunit variable $Q(a)$, which is formed by the two Q variables, Q^z and $Q^{a|z}$.

Step 2: Augment and Marginalize. Here we augment the graph to form the distribution of the treatment variable q^a , the variable on which we intervene (Figure A2). This variable is a deterministic function of its parents, $q_i^a(a) = m(q_i^z, q_i^{a|z})$ and the outcome variable Y now depends only on the augmentation variable, $Y_i \sim p(y | q_i(a))$. Again we require that the mechanism for y_i converges as $m \rightarrow \infty$.

To analyze the SUBUNIT INSTRUMENT graph, we also need a new idea: *marginalization*. We marginalize out Q_i^z so that Q_i^a no longer depends deterministically on its parents; see Figure 6l, and note the double arrows denoting a deterministic mechanism have been replaced by a single arrow from $Q^{a|z}$ to Q^a . In the marginalized graph, Q_i^a depends stochastically on its remaining parent, $Q_i^{z|a}$. This step ensures positivity. In the marginalized graph, we can have $p(q^a = q_\star^a | q^{y|a}) > 0$ for all $q^{y|a}$ (we discuss this assumption further in Appendix I).

Step 3: Identify. With positivity in place, we can use a backdoor adjustment to identify the intervention distribution,

$$p(y; \text{do}(q^a = q_\star^a)) = \int p(q^{a|z}) p(y | q_\star^a, q^{a|z}) dq^{a|z}. \tag{31}$$

We identify the estimand in Equation (29) from the expectation of Y under the distribution in Equation (31).

Step 4: Estimate. We observe n units, each with m_i subunits: $\{\{z_{ij}, a_{ij}\}_{j=1}^{m_i}, y_i\}_{i=1}^n$. We use the data to form estimates of the elements of Equation (31).

1. Estimate the i th conditional treatment distribution $\hat{q}_i^{a|z}$ from $\{z_{ij}, a_{ij}\}_{j=1}^{m_i}$.
2. Estimate the i th marginal treatment distribution \hat{q}_i^a from $\{a_{ij}\}_{j=1}^{m_i}$.
3. Estimate the population outcome distribution $\hat{p}(y | q^a, q^{a|z})$ from $\{\hat{q}_i^{a|z}, \hat{q}_i^a, y_i\}_{i=1}^n$.
4. Estimate the population distribution of $\hat{p}(q^{a|z})$ with the empirical distribution of $\{\hat{q}_i^{a|z}\}_{i=1}^n$.

We plug these estimates into Equation (31), and calculate the expectation.

As a final demonstration, we consider binary A, Y, U, Z , and draw simulated variables from a true SUBUNIT INSTRUMENT model (details are in Appendix D.3). We consider three different simulations,

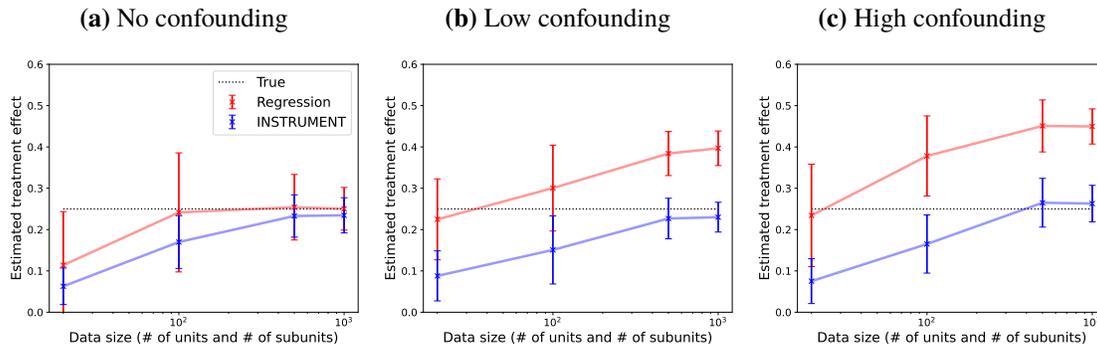


Figure 9: Estimated effects in an SUBUNIT INSTRUMENT simulation. The SUBUNIT INSTRUMENT estimate is based on the identification formula for the true HCM. The “Regression” estimate ignores the instrument and hence ignores confounding. Error bars show standard deviation across 20 independent simulations.

each with different levels of confounding. Our goal is to estimate a difference between two soft interventions,

$$\mathbb{E}_p [Y ; \text{do}(q^a = \text{Bern}(0.75))] - \mathbb{E}_p [Y ; \text{do}(q^a = \text{Bern}(0.25))] . \quad (32)$$

We observe $\{\{z_{ij}, a_{ij}, y_{ij}\}_{j=1}^m\}_{i=1}^n$ for increasing numbers of units and subunits, and where the number of units equals the number of subunits in each. Figure 9 compares the SUBUNIT INSTRUMENT estimator with a straightforward regression of Y on the average of A , which does not account for the hidden confounder. The error of the regression estimator is larger when there is more confounding. As the number of units and subunits increases, the SUBUNIT INSTRUMENT estimator converges to the true effect. Note that here both estimators use a nonparametric model, a Gaussian process classifier, rather than a well-specified parametric model as in the previous simulations; partially as a result, we see more substantial bias and variance in the estimator than in the previous simulations. As the number of subunits per unit increases, the bias at low N decreases (Figure A7) (Rainforth et al., 2018, Appendix B).

6. Theory

We demonstrated with the three graphs of Figure 1 how collapsing, augmenting and marginalizing can be used for hierarchical causal ID. In this section we elaborate the key assumptions behind these steps, and explain how they can be used on hierarchical causal models with arbitrary graphs. Note the notation introduced in this section is also summarized in Table A1.

6.1 Collapsed models

Our central tool for identification in HCMs is the collapsed model. The collapsed model is a flat CGM that matches a given HCGM, and in which the Q variables are endogenous variables. Here we derive the collapsed model for arbitrary HCGMs, and explain the assumptions that justify it.

We first define two sets of variables. Let $\text{da}_s(w)$ denote the indices of the *direct subunit ancestors* of a unit variable X^w , that is, those ancestral subunit variables which are either parents of X^w or

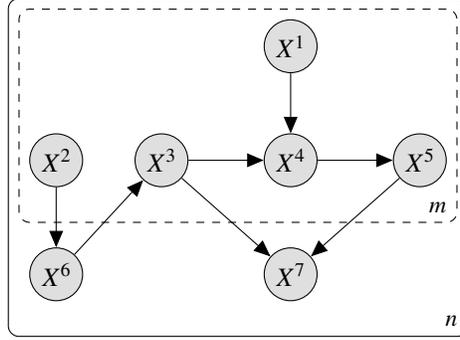


Figure 10: Direct subunit ancestors and descendants. In this example, the direct subunit ancestors of X^7 are X^1, X^3, X^4 and X^5 . Conversely, X^7 is the direct unit descendant of X^1, X^3, X^4 and X^5 . The only direct subunit ancestor of X^6 is X^2 , and X^6 is the direct unit descendant of X^2 .

are connected via a directed path containing only other subunit-level variables; see Figure 10. Let $\text{dd}_{\mathcal{U}}(v)$ denote the indices of the *direct unit descendants* of a subunit variable X^v , that is, those unit variables who have X^v as a direct subunit ancestor.

Definition 4 (Collapsed model) Consider a hierarchical causal graphical model \mathcal{M}^{cgm} , as in Definition 3. The corresponding collapsed model \mathcal{M}^{col} is a flat causal graphical model. It is

$$\begin{aligned} Q_i^{v|\text{pa}_{\mathcal{S}}(v)} &\sim \text{p}(q^v|\text{pa}_{\mathcal{S}}(v) \mid x_i^{\text{pa}_{\mathcal{U}}(v)}) && \text{for subunit variables } v \in \mathcal{S} \\ X_i^w &\sim \text{p}(x^w \mid x_i^{\text{pa}_{\mathcal{U}}(w)}, q_i(x^{\text{pa}_{\mathcal{S}}(w)})) && \text{for unit variables } w \in \mathcal{U}, \end{aligned} \quad (33)$$

for $i \in \{1, \dots, n\}$. Note the unit variables X^w are drawn conditional on the subunit marginal of their subunit ancestors $q_i(x^{\text{pa}_{\mathcal{S}}(w)})$,

$$q_i(x^{\text{pa}_{\mathcal{S}}(w)}) = \int \dots \int \prod_{w' \in \text{da}_{\mathcal{S}}(w)} q_i^{w'|\text{pa}_{\mathcal{S}}(w')}(x^{w'} \mid x^{\text{pa}_{\mathcal{S}}(w')}) \prod_{w' \in \text{da}_{\mathcal{S}}(w) \setminus \text{pa}_{\mathcal{S}}(w)} dx^{w'}. \quad (34)$$

It is through this dependence that X^w connects to $Q^{v|\text{pa}(v)}$ variables.

In the collapsed model, each endogenous Q -variable $Q_i^{v|\text{pa}_{\mathcal{S}}(v)}$ for $v \in \mathcal{V}$ is observed if and only if X^v and all its subunit parents are observed in the original HCGM, i.e. $v \in \mathcal{S}_{\text{obs}}$ and $\text{pa}_{\mathcal{S}}(v) \subset \mathcal{S}_{\text{obs}}$. Each endogenous non- Q -variable X_i^w for $w \in \mathcal{U}$ is observed if and only if it is observed in the original HCGM, i.e. $w \in \mathcal{U}_{\text{obs}}$.

Algorithm 1 is a graphical algorithm for deriving the collapsed causal graphical model from the hierarchical causal graphical model. Figure 1 gives three examples of collapsed models, which we have discussed. Other examples are in Figure A3 and Figure A4.

We next show that the effects of interventions in a collapsed model match the effects of corresponding interventions in the original HCGM. A hard intervention on a Q variable of a collapsed model, i.e., $\text{do}(q^{v|\text{pa}_{\mathcal{S}}(v)} = q_{\star}^{v|\text{pa}_{\mathcal{S}}(v)})$, corresponds to a soft intervention $X^v \sim q_{\star}^{v|\text{pa}_{\mathcal{S}}(v)}(x^v \mid x^{\text{pa}_{\mathcal{S}}(v)})$ on the subunit variable X^v in the original HCGM. A hard intervention on a unit variable X^w in the collapsed model corresponds to the same intervention in the original HCGM.

Algorithm 1 Graphical algorithm for collapsing an HCGM. This algorithm transforms the graph of an HCM into the graph of its collapsed model, following Definition 4.

Input: a hierarchical causal graphical model \mathcal{M}^{cgm} (Definition 3)
Output: the collapsed flat causal graphical model \mathcal{M}^{col} (Definition 4)
for each subunit variable $v \in \mathcal{S}$ **do**
 create a unit endogenous variable $Q^{v|\text{pa}_{\mathcal{S}}(v)}$.
 if X^v and its parent subunit variables are observed ($v \in \mathcal{S}_{\text{obs}}, \text{pa}_{\mathcal{S}}(v) \subset \mathcal{S}_{\text{obs}}$) **then**
 mark $Q^{v|\text{pa}_{\mathcal{S}}(v)}$ as observed (filled circle)
 else
 mark $Q^{v|\text{pa}_{\mathcal{S}}(v)}$ as hidden (empty circle)
 end if
 disconnect the unit parents $X^{\text{pa}_{\mathcal{Z}}(v)}$ from X^v ; connect them to $Q^{v|\text{pa}_{\mathcal{S}}(v)}$
 for each direct unit descendant $w \in \text{dd}_v(\mathcal{S})$ **do**
 connect $Q^{v|\text{pa}_{\mathcal{S}}(v)}$ to X^w (see Equation (34))
 end for
 erase the subunit variable X^v
end for
erase the inner plate

HCGMs converge to collapsed models. We now equate the HCGM to its collapsed model, in the infinite subunit limit. The key assumption is that mechanisms converge. Here, KL denotes the Kullback-Leibler divergence.

Definition 5 (Mechanism convergence) Consider a unit variable $w \in \mathcal{U}$ in an HCGM, and its mechanism $p(x^w | x^{\text{pa}_{\mathcal{Z}}(w)}, \{x^{\text{pa}_{\mathcal{S}}(w)}\}_{j=1}^m)$. We say the mechanism converges with infinite subunits if there exists a limiting conditional distribution $p(x^w | x^{\text{pa}_{\mathcal{Z}}(w)}, q(x^{\text{pa}_{\mathcal{S}}(w)}))$ such that,

$$\mathbb{E}_{X_{1:m}^{\text{pa}_{\mathcal{S}}(w)} \sim q(x^{\text{pa}_{\mathcal{S}}(w)})} \left[\text{KL} \left(p(x^w | x^{\text{pa}_{\mathcal{Z}}(w)}, q(x^{\text{pa}_{\mathcal{S}}(w)})) \parallel p(x^w | x^{\text{pa}_{\mathcal{Z}}(w)}, \{X_j^{\text{pa}_{\mathcal{S}}(w)}\}_{j=1}^m) \right) \right] \xrightarrow[m \rightarrow \infty]{a.s.} 0. \quad (35)$$

We define the left hand side as $D_m^w(x^{\text{pa}_{\mathcal{Z}}(w)}, q(x^{\text{pa}_{\mathcal{S}}(w)}))$.

Heuristically, if a mechanism depends smoothly on the empirical distribution of subunit variables $\hat{q}_m(x^{\text{pa}_{\mathcal{S}}(w)}) = \frac{1}{m} \sum_{j=1}^m \delta_{x_j^{\text{pa}_{\mathcal{S}}(w)}}$ and does not depend on the total number of subunits m , we can expect it to converge (Appendix E). An HCGM converges to its collapsed model so long as its unit variable mechanisms converge.

Theorem 6 (Collapsing a hierarchical causal model) Let $p_{\Delta, m}(x^{\mathcal{U}}, q(x^{\mathcal{S}}))$ be the joint distribution over $x^{\mathcal{U}}$ and $q(x^{\mathcal{S}})$ given by an HCGM with m subunits, under an intervention Δ (Definition 2). Let $p_{\Delta}^{\text{col}}(x^{\mathcal{U}}, q(x^{\mathcal{S}}))$ be the distribution given by the corresponding collapsed model, under the corresponding intervention. Assume each unit variable mechanism converges, such that for all $w \in \mathcal{U}$ we have $\mathbb{E}_{p_{\Delta}^{\text{col}}} [D_m^w(X^{\text{pa}_{\mathcal{Z}}(w)}, Q(x^{\text{pa}_{\mathcal{S}}(w)}))] \rightarrow 0$ as $m \rightarrow \infty$ a.s.. Then the hierarchical causal graphical model converges to the collapsed model,

$$\text{KL}(p_{\Delta}^{\text{col}} \parallel p_{\Delta, m}) \xrightarrow[m \rightarrow \infty]{a.s.} 0. \quad (36)$$

The proof is in Appendix F. This is a key result, as it allows us to analyze identification in hierarchical causal models by analyzing identification in a corresponding flat model.

Do-calculus in the collapsed model. To study identification, we apply do-calculus to the collapsed model. Do-calculus rests on the assumption that we know the joint distribution over observed variables in the model. This is justified by two key technical assumptions on the HCGM. First, with infinite data, we can observe the joint distribution over observable unit variables and the observable subunit distribution.

Assumption 1 (Known observable joint) *The distribution $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$ of the HCGM is known.*

In Appendix G we prove that under the conditions of Theorem 6, $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$ can be learned from data in the limit of infinite units and infinite subunits per unit, sampled from the HCGM. In particular, we show that the empirical distribution over units, of the empirical distribution over subunits, will converge in Wasserstein distance to $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$. We also characterize the convergence rate in N and M .

Knowledge of the joint distribution of subunits, $q(x^{\mathcal{S}_{\text{obs}}})$, does not always imply full knowledge of the conditional distributions. For example, in the UNIT CONFOUNDER graph, even if we know $q(a, y)$ we do not necessarily know $q^{y|a}(y | a')$ for all values of a' . The reason is that $q^a(a)$ may put zero probability on some values of A , in which case $q^{y|a}(y | a')$ will be unobservable for these values. So, we make the following positivity assumption. Here \mathcal{X}^v is the domain of a variable X^v , and $p(q(x^{\mathcal{S}}))$ is the distribution over subunit distributions in the HCGM.

Assumption 2 (Subunit-level positivity) *Consider $Q(x^{\mathcal{S}}) \sim p(q(x^{\mathcal{S}}))$. With probability one, for all $v \in \mathcal{S}$ and $x^{\text{pa}_{\mathcal{S}}(v)} \in \mathcal{X}^{\text{pa}_{\mathcal{S}}(v)}$, we have $Q(x^{\text{pa}_{\mathcal{S}}(v)}) > 0$.*

In short, to use the collapsed model for identification, we need variability among subunits. (If a subunit variable has the same value for all subunits then it is effectively a unit variable.)

Assumption 1 and Assumption 2 together imply that the distribution over observable variables in the collapsed models is known. Let $q^{\mathcal{Q}} \triangleq \{q^{v|\text{pa}_{\mathcal{S}}(v)} : v \in \mathcal{S}\}$ denote the set of Q variables in the collapsed model, and let $q^{\mathcal{Q}_{\text{obs}}} \triangleq \{q^{v|\text{pa}_{\mathcal{S}}(v)} : v \in \mathcal{S}_{\text{obs}}, \text{pa}_{\mathcal{S}}(v) \subset \mathcal{S}_{\text{obs}}\}$ denote the subset of Q variables that are observed.

Proposition 7 (Observed collapsed model) *Given the conditions of Theorem 6, Assumption 1 and Assumption 2, the joint distribution over the observed endogenous variables of the collapsed model is known, $p^{\text{col}}(x^{\mathcal{U}_{\text{obs}}}, q^{\mathcal{Q}_{\text{obs}}})$.*

Do-calculus proceeds from the assumption that the joint distribution over the observed endogenous variables in a flat causal graphical model is known. Proposition 7 says that this distribution is known for the collapsed model. So, Proposition 7 implies we can apply do-calculus to identify the effects of interventions in the collapsed model. Then, by Theorem 6, we can equate effects in the collapsed model to effects in the original HCGM.

Do-calculus also rests on assumptions about unit-level positivity (Shpitser and Pearl, 2006). Most salient is that the intervention we are studying always has non-zero probability.

Assumption 3 (Unit-level positivity) *Let \tilde{Z} be a variable in a collapsed model (it may be a unit variable X^w or a Q variable). Let $Z^{\text{pa}(\tilde{z})}$ denote the parents of \tilde{Z} in the collapsed model graph. For any hard intervention $\text{do}(\tilde{z} = \tilde{z}_{\star})$, we require $p^{\text{col}}(\tilde{z}_{\star} | z^{\text{pa}(\tilde{z})}) > 0$ a.s. for $Z^{\text{pa}(\tilde{z})} \sim p^{\text{col}}(z^{\text{pa}(\tilde{z})})$.*

When applying do-calculus, one sometimes finds that identifying one effect in the model depends on first identifying the effects of other interventions. In such cases, positivity most hold for those other interventions as well (Assumption 4, Appendix H).

6.2 Augmentation and marginalization

Augmentation and marginalization are graphical proof techniques that help establish identification in HCMs. We use them when do-calculus on the collapsed model does not immediately yield identification. We saw examples of augmented and marginalized models in Section 5. Here we outline the approach; more details are in Appendix J.

When we augment a collapsed model, we add an additional endogenous variable, which describes some quantity of interest. For the augmentation to be valid, we must recover the original model when we marginalize out the augmentation variable (details on marginalization are in Appendix A).

Definition 8 (Valid augmented model) Consider a collapsed model with distribution $p^{\text{col}}(x^{\mathcal{U}}, q^{\mathcal{Q}})$ over endogenous variables $X^{\mathcal{U}}, Q^{\mathcal{Q}}$. An augmented model \mathcal{M}^{aug} includes an additional endogenous variable \tilde{Q} generated from parents $Q^{\text{pa}(\tilde{Q})}$ according to a chosen deterministic mechanism $\tilde{q}_i = \tilde{f}_i(q_i^{\text{pa}(\tilde{Q})})$.

This augmentation variable is observed so long as it can always be computed from observed variables in the original collapsed model. The augmentation is valid so long as if \tilde{Q} is marginalized out of the model, we recover the original collapsed model.

As we have seen, we focus on augmentation variables that describe marginal or conditional distributions over subunit variables, i.e. new Q variables. Algorithm 2 provides a graphical algorithm for augmenting a collapsed model. Figure 6d, Figure 6h and Figure A2 give three examples of augmented models, which we have discussed. Other examples are in Figure A3 and Figure A4.

We finally turn to marginalization. We marginalize an augmented model to identify the effects of interventions on its augmentation variable. Note we may be interested in these effects for their own sake, or as part of our effort to identify the effects of other interventions in the model. Once we drop a parent of the augmentation variable from the model, it depends stochastically rather than deterministically on its remaining parents. This allows interventions on the augmentation variable to satisfy positivity (Assumption 3). For example, we used this approach to achieve identification in the SUBUNIT INSTRUMENT graph in Section 5.3.

In detail, we can marginalize out any variable with one or zero children (Appendix A). Graphically, we erase the variable and connect its parents directly to its child. We modify the mechanism for the child accordingly, absorbing the mechanism for the marginalized variable. Algorithm 3 provides a graphical algorithm for marginalizing an augmented model. In addition to the SUBUNIT INSTRUMENT model (Figure 6l), Figure A3 and Figure A4 give further examples of marginalized models.

Augmented and marginalized models match the original model. The following propositions justify the use of augmented and marginalized models. They show that we can equate effects in these models to effects in the original collapsed model, and hence to effects in the original HCM. For simplicity, we state these results under the assumption that the collapsed model has been augmented with just one augmentation variable.

Algorithm 2 Graphical algorithm for augmenting a collapsed model. This algorithm adds an augmentation variable to a collapsed HCGM, following Definition 8.

Input: a collapsed model \mathcal{M}^{col} and an augmentation variable \tilde{Q}
Output: an augmented model \mathcal{M}^{aug}
 add the augmentation variable \tilde{Q}_i to the graph, with a mechanism $\tilde{q}_i = \tilde{f}(q_i^{\text{pa}(\tilde{q})})$
 if $\tilde{f}(q_i^{\text{pa}(\tilde{q})})$ can be computed from $q(x^{\text{obs}})$, mark \tilde{Q}_i as observed (filled circle)
for each parent of \tilde{Q}_i **do**
 connect the parent to \tilde{Q}_i via a double arrow
end for
for each variable X^w for $w \in \mathcal{U}$ in the collapsed model **do**
 if $\tilde{f}(q_i^{\text{pa}(\tilde{q})})$ appears in the mechanism for X_i^w **then**
 connect \tilde{Q} to X^w
 in X^w 's mechanism, replace $\tilde{f}(q_i^{\text{pa}(\tilde{q})})$ with \tilde{q}_i
 if one or more parents of \tilde{Q} no longer appear in X^w 's mechanism **then**
 erase the arrow from the parent(s) to X^w
 end if
 end if
end for

Algorithm 3 Graphical algorithm for marginalizing an augmented model. This algorithm marginalizes out parent(s) of an augmentation variable (Section 6.2).

Input: an augmented collapsed model \mathcal{M}^{aug} with augmentation variable \tilde{Q}
Input: a set $Q^{\mathcal{E}} \subseteq Q^{\text{pa}(\tilde{q})}$ of one or more parents of \tilde{Q} , for whom \tilde{Q} is their only child
Output: a marginalized model \mathcal{M}^{mar}
for each variable Q^c in the set of parents $Q^{\mathcal{E}}$ **do**
 connect any parents of Q^c to \tilde{Q}
 erase Q^c
end for
 replace all double arrows into \tilde{Q} with single arrows

The first result equates causal effects in an augmented model to causal effects in the original collapsed model. It applies to interventions on any variable besides the augmentation variable itself. Let $p(x^{\mathcal{U}}, q^{\tilde{Q}})$ denote the distribution of the original collapsed model, and let $p^{\text{aug}}(x^{\mathcal{U}}, q^{\tilde{Q}}, \tilde{q})$ denote the distribution of the augmented model.

Proposition 9 (Augmented model matches original model) *For any intervention Δ (soft or hard) on one or more endogenous variable in the original collapsed model, we have, for a valid augmented model, $p_{\Delta}^{\text{aug}}(x^{\mathcal{U}}, q^{\tilde{Q}}, \tilde{q}) = p_{\Delta}^{\text{col}}(x^{\mathcal{U}}, q^{\tilde{Q}}, \tilde{f}(q^{\text{pa}(\tilde{q})}))$ a.e..*

The result follows immediately from Definition 8. It says the post-intervention distribution over the augmentation variable in the augmented model matches the post-intervention distribution over the quantity the augmentation variable describes in the original model. For instance, in Figure 6d the augmentation variable is q^y , and we can conclude $p^{\text{aug}}(q^y; \text{do}(q^a = q_{\star}^a)) = p^{\text{col}}(\int q^y |^a(y | a)q^a(a)da; \text{do}(q^a = q_{\star}^a))$.

In marginalized models, we are interested in interventions on augmentation variables themselves. We now show that we can equate the effects of interventions on these augmentation variables to the effects of the corresponding intervention in the original HCGM. We will focus on augmentation variables that describe the conditional distribution of one subunit variable given some (or none) of its parents, i.e. augmentation variables $Q^{v|\mathcal{R}}$ where $\mathcal{R} \subset \text{pa}_S(v)$ (see Appendix J.1 for a full definition).

The following result equates the effects of the intervention $\text{do}(q^{v|\mathcal{R}} = q_{\star}^{v|\mathcal{R}})$ in a marginalized model to the effects of the intervention $\text{do}(X^v \sim q_{\star}^{v|\mathcal{R}}(x^v|x_{ij}^{\mathcal{R}}))$ in the original HCGM. Let p^{mar} denote the marginalized model distribution.

Proposition 10 (Augmentation interventions match original model) *Consider an augmentation variable $Q^{v|\mathcal{R}}$ where $\mathcal{R} \subset \text{pa}_S(v)$ (with mechanism given by Equation (71) in the appendix). Recall $Q^{v|\text{pa}_S(v)}$ is the original Q variable describing X^v . Let Y denote one or more outcome variables. Assume either (a) $Q^{v|\text{pa}_S(v)}$ does not appear in the marginalized model, or (b) all directed paths from $Q^{v|\text{pa}_S(v)}$ to Y go through $Q^{v|\mathcal{R}}$. Then, $p^{\text{mar}}(y; \text{do}(q^{v|\mathcal{R}} = q_{\star}^{v|\mathcal{R}})) = p^{\text{col}}(y; \text{do}(q^{v|\text{pa}_S(v)} = q_{\star}^{v|\mathcal{R}}))$ a.e..*

A proof is in Appendix J. As an example, in Figure 6l, the augmentation variable is Q^a , while the original Q variable describing X^a is $Q^{a|z}$. The only directed path from $Q^{a|z}$ to the outcome Y goes through Q^a . So, we can equate the effect $p^{\text{mar}}(y; \text{do}(q^a = q_{\star}^a))$ in the marginalized model to $p^{\text{col}}(y; \text{do}(q^{a|z} = q_{\star}^a))$ in the collapsed model, which in turn corresponds to $p(y; \text{do}(A \sim q_{\star}^a(a)))$ in the original HCGM.

6.3 When does hierarchy enable identification?

We have described a procedure for proving identification in hierarchical causal graphical models. Figure A3 and Figure A4 give examples of graphs where it does and does not lead to identification. In this section, we investigate general features of HCGM graphs that enable identification, and compare them to flat causal models. Our results reveal when and where hierarchy enables identification.

We build on the bi-directed path criterion for flat causal models (Tian and Pearl, 2002; Pearl, 2009, Chap. 3). A path is *bi-directed* if it follows the pattern $Z^1 \leftarrow U^1 \rightarrow Z^2 \leftarrow U^2 \rightarrow \dots \leftarrow U^k \rightarrow Z^{k+1}$ where the U^v variables are hidden and the Z^v are observed.³ For example, in Figure A4i, there is a bi-directed path $A_i \leftarrow U_i \rightarrow Q_i^{y|w} \leftarrow U'_i \rightarrow Q_i^w$ from A_i to Q_i^w . Let X^{obs} denote the set of observed endogenous variables in the flat model.

Theorem 11 (Bi-directed path criterion (Tian and Pearl, 2002), Thm. 3) *The effect $p(x^{\text{obs}}; \text{do}(a = a_{\star}))$ is identified if and only if there is no bi-directed path between A and any of its children.*

We now develop similar criteria for hierarchical causal models, by applying the bi-directed path criterion to collapsed/augmented/marginalized models. Call a node $X^{v'}$ a *subunit instrument* of X^v if both X^v and $X^{v'}$ are subunit-level, X^v is the only child of $X^{v'}$, and X^v has no parents.

Theorem 12 (Sufficient conditions for identification in hierarchical models) *Consider an HCGM with no hidden subunit-level confounders, and assume the treatment variable A is subunit-level. We*

3. The term *bi-directed* comes from the graphical notation in which one draws a dashed, bidirectional arc between every pair of observed variables affected by the same confounder. In this case, a bi-directed path is one in which every edge is bidirectional.

are interested in the effect $p(y; \text{do}(A \sim q_\star^a))$ if Y is unit-level or $p(q(y); \text{do}(A \sim q_\star^a))$ if Y is subunit-level. Delete from the graph any variable that is not either an ancestor of Y or Y itself. The effect is identifiable if (1) there is no bi-directed path from A to a direct unit descendant of A , or (2) A has a subunit-level instrument.

The proof is in Appendix K.1. Part 1 says that in hierarchical causal models, we can ignore unit-level confounding between subunit variables, except insofar as it leads (via a bi-directed path) to confounding with a direct descendant outside the inner plate (examples: Figure 1a, Figure 1b, Figure A3a, Figure A3c and Figure A3e). Part 2 says that subunit-level instruments are a license to ignore unit-level confounding entirely (examples: Figure 1c, Figure A3j and Figure A3m)

Theorem 12 tells us broadly about the advantages of hierarchy. When we disaggregate some quantity and make fine-grained measurements instead, we change it from a unit-level variable to a subunit-level variable (e.g. instead of a school’s average test score, we have the per-student test score). This disaggregation allows us to ignore confounding (bi-directed paths) between the treatment variable and some or all of its children. In each of the examples in Figure A3, the effect is identified in the HCGM, but would not be if all the subunit variables were unit-level.

There are also general situations where disaggregation is not helpful, that is, making subunit-level rather than unit-level measurements does not aid identification. Below, we compare an HCGM directly to a flat causal model where all the subunit variables are unit-level, but the graph and the observed variables are the same. We refer to this as the *erased inner plate* model, with distribution $p^{\text{ep}}(x^{\text{obs}})$ over the observed endogenous variables.

Theorem 13 (No benefits of hierarchy for unit treatments) *Consider an HCGM with no hidden subunit-level confounders, and assume the treatment variable A is unit-level. If the effect $p^{\text{ep}}(x^{\text{obs}}; \text{do}(a = a_\star))$ is not identified in the erased inner plate model, then the effect $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}); \text{do}(a = a_\star))$ is not identified in the HCGM.*

The proof is in Appendix K.2. Theorem 13 tells us, for example, that the effect of A on W and Y in Figure A4h is not identified. A caveat, however, is that the result only deals with effects on all the observed endogenous variables, not a specific outcome variable Y (Tian and Pearl, 2002).

Naively, we might expect that by measuring in finer detail the mechanisms by which a unit-level treatment affects an outcome, we might better be able to infer the treatment’s effects. However, Theorem 13 suggests the benefits of hierarchy for causal identification only accrue when we can measure the treatment itself in finer detail. Intuitively, subunit-level data is useful for causal identification because it provides information about a natural experiment in which subunit treatments are randomized within each unit. There is no such natural experiment for unit-level treatments, regardless of whether or not other subunit-level variables are observed.

7. Application: Eight Schools

We now illustrate the use of hierarchical causal models on a real-world problem. We analyze data from a well-known study describing a set of randomized experiments conducted at eight secondary schools in the United States in 1977 (Alderman and Powers, 1979). At each school, students were randomly assigned to attend special test preparation programs or to not attend; the scores of each student on the SAT verbal component were measured at the end of the program. The goal was to understand the effects of test preparation programs on test scores.

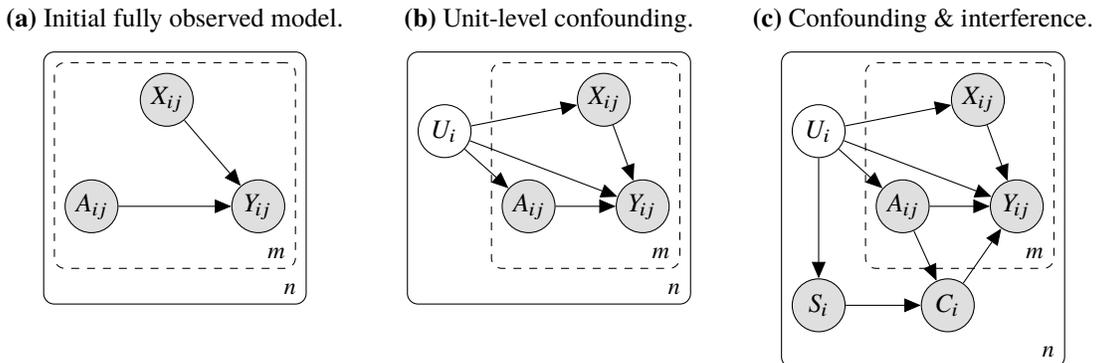


Figure 11: Models for the eight schools data. (a) The initial model, in which all variables are observed (Section 7.1). (b) An extended model that includes an unobserved unit-level confounder (Section 7.2). (c) An extended model that also includes an observed interferer (Section 7.3).

This “eight schools” study is used for textbook illustrations of the principles of hierarchical Bayesian modeling and inference (Rubin, 1981; Gelman et al., 2013). Here we reanalyze the data in the framework of hierarchical causal models. We first show how the standard textbook analysis can be derived as estimation under a hierarchical causal model. We then show how we can account for a plausible source of interference and refine the inferences of the standard analysis.

In this data, each student is randomly assigned to the treatment, the test preparation program. Let $a_{ij} = 1$ if student j in school i is treated, and $a_{ij} = 0$ otherwise. At the end of the preparation program, each student takes the SAT verbal component; let y_{ij} indicate the score of student j in school i . Besides the test preparation program, a student’s pre-treatment academic ability likely contributes to their outcome y_{ij} . For this reason, researchers also recorded each student’s scores on several tests taken before the program began: PSAT verbal, PSAT mathematics and Test of Standard Written English. Let x_{ij} denote the scores of student j in school i on these earlier tests.

We are interested in the average treatment effect on test scores if all students were enrolled in the special preparation program versus if the special preparation program were discontinued,

$$\text{ATE} = \mathbb{E}_p[Y ; \text{do}(a = 1)] - \mathbb{E}_p[Y ; \text{do}(a = 0)]. \quad (37)$$

We will consider several hierarchical causal models. Details of this study are in Appendix L.

7.1 A fully observed model

We first study a fully-observed hierarchical causal model, Figure 11a. Since treatment is randomized, there are no confounders inside the inner plate, and nor is there an arrow from X_{ij} to A_{ij} . Figure A9 in the appendix shows the collapsed, augmented and marginalized models. Applying do-calculus, we can identify the effect as

$$\text{ATE} = \mathbb{E}_p[\mathbb{E}_{Q^{y|a}}[Y | A = 1]] - \mathbb{E}_p[\mathbb{E}_{Q^{y|a}}[Y | A = 0]]. \quad (38)$$

We estimate this effect with hierarchical Bayesian methods. We parameterize $q_i^{y|a}$ with a linear regression, and parameterize $p(q^{y|a})$ as a normal distribution over its coefficients, with unknown mean and variance. We perform Bayesian inference on all the unknown parameters. (There are some

subtleties, since the eight-schools study does not make public its per-student data; see the details in Appendix L.1.) The resulting Bayesian model matches the textbook eight schools model (Gelman et al., 2013). We compute the posterior over the ATE using MCMC, specifically the No-U-turn Hamiltonian Monte Carlo sampler (NUTS) in NumPyro (Hoffman and Gelman, 2014; Phan et al., 2019; Bingham et al., 2019). Figure 12 shows the results (blue distribution). This analysis suggests the treatment is likely to increase test scores a modest amount: each question on the SAT verbal is worth an average of 7 points, and the posterior mean and standard deviation of the ATE are 4.4 points and 3.4 points respectively.

7.2 Unit-level confounding

Whether each student attends the test preparation program is randomized within each school, so there are no subunit-level confounders between treatment and outcome. But there may still be unit-level confounders. For example, each school’s financial and administrative resources may affect both student test scores and student enrollment in the program. Indeed, Alderman and Powers (1979) report that “where student interest far exceeded the program’s capacity [...], a larger number of students went into the control group than into the treatment group.” In the data, the fraction of students treated varies between 25% and 55% across different schools (Figure A12d). It seems plausible that schools with greater financial resources could have larger and more effective programs.

Figure 11b considers the possibility of unit-level confounders. Note the graph also allows confounders to impact X , the pre-treatment test scores. To identify the causal effect, we first collapse, augment, and marginalize (Figure A10). Then, applying do-calculus recovers the same identification formula for the ATE as above in Equation (38). Intuitively, the impact of the confounder is through the conditional distribution of Y given A within each school, so to correct for confounding it suffices to control for this conditional. In short, the estimation method of the previous section is robust to unobserved unit-level confounding.

7.3 Confounding & interference

There is evidence that increased class size can negatively impact students’ academic performance (e.g. Angrist and Lavy, 1999). Since the level of enrollment in each school’s tutoring programs presumably influences their class size, this leads to the possibility of interference. In the data, Alderman and Powers (1979) report the class sizes for each test preparation program at each school, and we can visually examine its relationship to the per-school estimated treatment effect; see Figure A12a. It seems plausible that class size impacts test scores.

To address this formally, we consider the hierarchical causal model in Figure 11c. Here C_i is the class size and S_i is the total number of students who expressed interest in the tutoring program. This model allows for the possibility of interference: enrolling more students in the test preparation program may drive up class size, which in turn may drive down students’ test scores.

The collapsed, augmented, and marginalized models are shown in Figure A11 in the appendix. Applying do-calculus, the interventional expectation $\mathbb{E}_p [Y ; \text{do}(q^a = \delta_{a_\star})]$ can be identified as

$$\int \int p(s) p(c | q^a = \delta_{a_\star}, s) ds \int p(q^a, s') \mathbb{E}_p [\mathbb{E}_{Q^{y|a}} [Y | A = a_\star] | q^a, s', c] dq^a ds' dc. \quad (39)$$

We identify the ATE by identifying the expectation for $a_\star = 1$ and $a_\star = 0$.

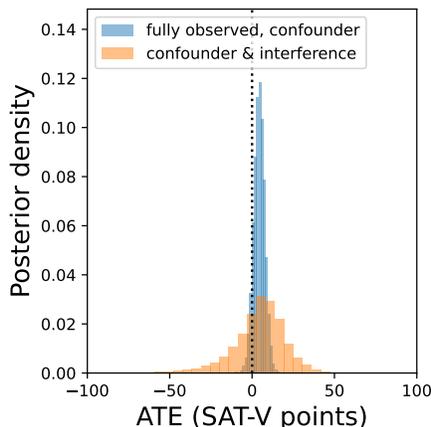


Figure 12: Posterior over average treatment effect for eight schools study. In blue is the posterior over the ATE for the models discussed in Section 7.1 and Section 7.2. In orange is the posterior for the model in Section 7.3. We find, when we take into account possible interference, more uncertainty in the effect of the preparation program on test scores.

We develop a hierarchical Bayesian estimation strategy. Again we parameterize $q_i^{y|a}$ with a linear model. We parameterize q_i^a with a Bernoulli distribution with unknown mean. We parameterize $p(c | q^a, s)$ with a linear regression that predicts c from s and the mean of q^a . We parameterize $p(q^{y|a} | q^a, s, c)$ with a linear regression that predicts the coefficient of the linear model of $q^{y|a}$ based on s, c and the mean of q^a . We place priors on all parameters and perform Bayesian inference, again using MCMC. We use each sample from the MCMC procedure to form a Monte Carlo approximation of Equation (39). All the details of this procedure are in Appendix L.2.

With this estimation strategy, Figure 12 plots the posterior ATE (orange). The posterior mean of the ATE (4.2 points) is similar to that for the classical analysis of the initial fully observed/confounder model (4.4 points). But there is substantially more uncertainty when we account for interference. The 5th percentile of the ATE posterior is -26 points and the 95th percentile is 28 points, whereas for the classic analysis they are -1.2 points and 9.8 points respectively. In summary, the standard eight schools analysis suggests that the test preparation program is likely to be modestly effective, if rolled out to all students. But when we allow for the possibility of school-level confounders and interference through class size, the program’s effectiveness is more uncertain.

8. Discussion

We proposed hierarchical causal models and studied identification and estimation. HCMs are a general tool for addressing causal questions using hierarchical data. We developed proof techniques for identifying causal effects in arbitrary HCMs, without parametric assumptions on causal mechanisms. We developed estimation methods based on hierarchical probabilistic models, and found deep connections to hierarchical Bayesian methods. These methods enable causal inference in novel hierarchical data settings with novel causal graphs.

A popular rule-of-thumb for handling hierarchical data in causal inference is to include the index of the environment or group as a dummy variable in the regression. This approach is well justified

when the unit just contributes hidden confounding (the `UNIT CONFOUNDER` graph). But in many real world scenarios, there is more complexity at the unit level, such as sources of interference or outcomes of interest. To model what is happening at the unit level, we cannot simply replace all these unit variables with a dummy variable in a regression. This is reflected in the distinct identification formula of our three HCM examples.

8.1 Assumptions and limitations

Our methods have several important assumptions and limitations. A defining aspect of the HCM framework is that units are exchangeable and subunits are exchangeable within units. This assumption is what connects HCMs to hierarchical probabilistic graphical models with inner plates, ensuring that any HCM reduces to such a model in the absence of intervention. However, exchangeability is not always appropriate, for instance when different units or subunits correspond to nearby points in time or space (Bertrand et al., 2004). An important area of investigation is how to relax the exchangeability assumption. As one example in this direction, Christiansen et al. (2022) present a nonparametric causal model similar to the `UNIT CONFOUNDER` model that accounts for spatiotemporal correlation.

A key restriction of our identification results is that they hold only in the limit of infinite subunits, and only for HCMs whose mechanisms converge in this limit. An advantage of the infinite subunit limit is that it allows us to avoid placing parametric assumptions on the distribution over subunit variables. In reality, however, there may not be very many subunits within each unit. To help address such situations, it will be important to explore reasonable assumptions that enable identification from small numbers of subunits.

Our identification theory addresses unobserved unit-level confounders and interferers, but there are many other threats to valid causal inference in HCMs. Going forward, it will be important to investigate issues such as unobserved subunit-level confounders, selection bias among units or subunits, and more general forms of interference among units and subunits.

8.2 Outlook

Broadly speaking, HCMs help formalize the question of when reductionism—in the sense of analyzing an aggregate phenomena in terms of its individual component parts—enables understanding of cause and effect. On one hand, we see that reductionism can be an enabler for causal inference. By looking at individual subunits, instead of aggregate unit-level variables, we can effectively hold unit-level confounders fixed while randomizing subunit treatments. On the other hand, HCMs also show that reductionism offers no advantage when we are interested in unit-level treatments. Without randomness at the subunit level, we are left with the usual requirements for causal inference.

HCMs provide data analysis methods that leverage technological progress in measurement and intervention methods. Across many scientific domains, technological advances lead to unit-level data being supplemented or supplanted by subunit-level data. Consider, for example, a political scientist interested in the impact of news consumption on political behavior. In the past, they may have had to rely data such as the subscription levels of different newspapers in different cities (Gentzkow et al., 2014). Modern media apps, however, enable measurement of the exact news articles read by individuals (González-Bailón et al., 2023). So, unit-level data about groups of citizens can be replaced by subunit-level data about individual citizens. These apps further enable interventions on the articles recommended to individuals, a targeted intervention on subunits (Guess et al., 2023).

HCMs thus offer one potential tool for leveraging this novel technology to better understand the effects of news media. Moreover, because our HCM identification results are nonparametric, they do not only apply to binary or continuous variables. Instead, they can be applied to complex structured data, such as that recorded by apps. For example, one can treat the entire text of news articles as a treatment or an outcome variable (Feder et al., 2022; Egami et al., 2022).

Analogous technological advances are also being made in fields outside social science. Consider a biologist interested in the effects of gene expression on disease progression (Tejada-Lapuerta et al., 2023). In the past, they would have had to rely on bulk gene expression measurements, taking the average expression levels within a tissue. With the development of single cell RNA sequencing, expression levels can be measured in individual cells (Klein et al., 2015). So, unit-level data about tissues can be replaced by subunit-level data about their constituent cells. Meanwhile, advances in synthetic biology enable targeted modification of gene expression levels in specific cell types, i.e. conditional soft subunit interventions (Hrvatín et al., 2019). HCMs thus offer a possible tool for leveraging single cell data to inform emerging therapeutic strategies.

Theoretically, a central open problem is finding an identification method for HCMs that is *complete*, in the sense that if an effect cannot be identified via the method then it is not identified. The do-calculus is complete (Shpitser and Pearl, 2006; Huang and Valtorta, 2006), and our identification method rests on application of do-calculus to the collapsed model. But many HCMs have collapsed models that are not fully nonparametric, even though the HCM itself is nonparametric. For example, in the collapsed SUBUNIT INSTRUMENT graph, the outcome variable Y_i depends on its parents $Q^{a|z}$ and Q^z only through the marginal that they induce. Consequently, there can be effects that are identified even when do-calculus says they are not.

Another open direction is to advance estimation. We gave some estimation methods, including both hierarchical point estimators (Section 5) and hierarchical Bayesian models, which account for statistical uncertainty (Section 7). To improve sample efficiency and robustness, it may be valuable to develop influence function-based corrections, along the lines of doubly robust estimators or targeted learning (Kennedy, 2022). It is an open question how to extend these and other causal estimation approaches to estimands from hierarchical causal models.

References

- Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *J. Am. Stat. Assoc.*, 105(490):493–505, 2010.
- Donald L Alderman and Donald E Powers. The effects of special preparation on SAT-Verbal scores. *ETS Res. Rep. Ser.*, 1979(1):i–37, 1979.
- Joshua D Angrist. The perils of peer effects. *Labour Econ.*, 30:98–108, 2014.
- Joshua D Angrist and Victor Lavy. Using Maimonides’ rule to estimate the effect of class size on scholastic achievement. *Q. J. Econ.*, 114(2):533–575, 1999.
- Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, 2009.

- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. 2019.
- Elias Bareinboim, Juan Correa, Duligur Ibeling, and Thomas Icard. On Pearl’s hierarchy and the foundations of causal inference. 2021.
- Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *Q. J. Econ.*, 119(1):249–275, 2004.
- Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20(28):1–6, 2019.
- Vladimir I Bogachev and Aleksandr V Kolesnikov. The Monge-Kantorovich problem: Achievements, connections, and perspectives. *Russian Math. Surveys*, 67(5):785, 2012.
- Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In James J Heckman and Edward E Leamer, editors, *Handbook of Econometrics*, volume 6, pages 5633–5751. Elsevier, 2007.
- Rune Christiansen, Matthias Baumann, Tobias Kuemmerle, Miguel D Mahecha, and Jonas Peters. Toward causal inference for spatio-temporal data: Conflict and forest loss in Colombia. *J. Am. Stat. Assoc.*, 117(538):591–601, 2022.
- Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *AAAI Conference on Artificial Intelligence*, 2020.
- A P Dawid. Influence diagrams for causal modelling and inference. *Int. Stat. Rev.*, 70(2):161–189, 2002.
- Vanessa Didelez, A Philip Dawid, and Sara Geneletti. Direct and indirect effects of sequential treatments. In *Conference on Uncertainty in Artificial Intelligence*, 2006.
- Richard M Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.
- Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *Am. Econ. Rev.*, 101(5):1739–1774, 2011.
- Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *Science Advances*, 8(42), 2022.
- Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, Brandon M Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.
- Avi Feller and Andrew Gelman. Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences*, pages 1–16, May 2015.

- Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *J. Appl. Stat.*, 31(7):799–815, 2004.
- Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.
- Matthew Gentzkow, Jesse M Shapiro, and Michael Sinkinson. Competition and ideological diversity: Historical evidence from US newspapers. *Am. Econ. Rev.*, 104(10):3073–3114, 2014.
- J K Ghosh and R V Ramamoorthi. *Bayesian Nonparametrics*. Series in Statistics. Springer, 2003.
- Sylvie Goetgeluk and Stijn Vansteelandt. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780, September 2008.
- Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A Tucker. Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398, 2023.
- S Greenland. When should epidemiologic regressions use random coefficients? *Biometrics*, 56(3): 915–921, September 2000.
- Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A Tucker. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381(6656):404–408, 2023.
- Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de Finetti: On the identification of invariant causal structure in exchangeable data. *arXiv*, 2022.
- Siyuan Guo, Chi Zhang, Karthika Mohan, Ferenc Huszár, and Bernhard Scholkopf. Do finetti: on causal effects for exchangeable data. In *Neural Information Processing Systems (NeurIPS)*, 2025.
- Jennifer Hill. Multilevel models and causal inference. In Marc A Scott, Jeffrey S Simonoff, and Brian D Marx, editors, *The SAGE Handbook of Multilevel Modeling*, pages 201–219. SAGE Publications Ltd, 2013.
- Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 2014.

- Sinisa Hrvatin, Christopher P Tzeng, M Aurel Nagy, Hume Stroud, Charalampia Koutsoumpa, Oren F Wilcox, Elena G Assad, Jonathan Green, Christopher D Harvey, Eric C Griffith, and Michael E Greenberg. A scalable platform for the development of cell-type-specific viral drivers. *Elife*, 8, 2019.
- Yimin Huang and Marco Valtorta. Pearl’s calculus of intervention is complete. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.
- Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *J. Am. Stat. Assoc.*, 103(482):832–842, 2008.
- Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. Technical report, National Bureau of Economic Research, 2002.
- Dominik Janzing and Sergio Mejia. Phenomenological causality. *arXiv*, 2022.
- David Jensen, Javier Burrone, and Matthew Rattigan. Object conditioning for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, 2020.
- Edward Kennedy. Semiparametric doubly robust targeted double machine learning: A review. 2022.
- Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161:1187–1201, 2015.
- David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning*, 2021.
- S Lauritzen. Causal inference from graphical models. *Monographs on Statistics and Applied Probability*, 87:63–108, 2001.
- Chanhwa Lee, Donglin Zeng, and Michael G Hudgens. Efficient nonparametric estimation of stochastic policy effects with clustered interference. *arXiv*, 2022.
- Jeffrey W Miller and Matthew T Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.*, 15(1):3333–3370, 2014.
- J M Neuhaus and J D Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645, June 1998.
- Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.
- Xuanlong Nguyen. Borrowing strength in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016.
- Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv*, 2009.

- Elizabeth L Ogburn and Tyler J VanderWeele. Causal diagrams for interference. *Statistical Science*, 29(4):559–578, 2014.
- Judea Pearl. *Causality*. Cambridge University Press, 2009.
- Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*, 2022.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Series B Stat. Methodol.*, 78(5):947–1012, 2016.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.
- Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv*, 2019.
- Tom Rainforth, Robert Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning (ICML)*, 2018.
- Stephen W Raudenbush, Sean F Reardon, and Takako Nomi. Statistical analysis for multisite trials using instrumental variables with random coefficients. *J. Res. Educ. Eff.*, 5(3):303–332, 2012.
- Sean F Reardon, Fatih Unlu, Pei Zhu, and Howard S Bloom. Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *J. Educ. Behav. Stat.*, 39(1): 53–86, 2014.
- T Richardson and James M Robins. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. 2013.
- Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Ann. Stat.*, 30(4):962–1030, 2002.
- Donald B Rubin. Estimation in parallel randomized experiments. *J. Educ. Behav. Stat.*, 6(4): 377–401, 1981.
- Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister, and Jonas Peters. Exploiting independent instruments: Identification and distribution generalization. In *International Conference on Machine Learning*, 2022.
- Fredrik Sävje, Peter Aronow, and Michael Hudgens. Average treatment effects in the presence of unknown interference. *Ann. Stat.*, 49(2):673–701, 2021.
- Cosma Shalizi. *Advanced data analysis from an elementary point of view*. 2013.
- Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Association for the Advancement of Artificial Intelligence*, 2006.

Peter Spirtes. Introduction to causal inference. *J. Mach. Learn. Res.*, 11:1643–1662, 2010.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Stat. Methods Med. Res.*, 21(1):55–75, 2012.

Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *arXiv*, 2023.

Jin Tian and Judea Pearl. Causal discovery from changes. In *Conference on Uncertainty in Artificial Intelligence*, 2001.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Conference on Artificial Intelligence*, 2002.

Cedric Villani. *Optimal transport: Old and new*, volume 338. Springer Science & Business Media, 2008.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Sam Witty, Kenta Takatsu, David Jensen, and Vikash Mansinghka. Causal inference using Gaussian processes with structured latent confounders. In *International Conference on Machine Learning*, 2020.

Jeffrey M Wooldridge. Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Rev. Econ. Stat.*, 87(2):385–390, 2005.

Jeffrey M Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.

Mingzhang Yin, Yixin Wang, and David M Blei. Optimization-based causal estimation from heterogeneous environments. *arXiv*, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Symbol	Description	Definition
p	Distribution of an HCM.	Definitions 1 and 3
f	Deterministic mechanism in an HSCM.	Definition 1
\mathcal{G}	Graph of an HCM.	Definition 1
$\mathcal{V} = \{1, \dots, V\}$	Indices of the endogenous variables.	Section 3.3
X^v, X^w	The v th, w th endogenous variables.	Section 3.3
$\mathcal{S} \subseteq \mathcal{V}$	Indices of subunit endogenous variables.	Section 3.3
$\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$	Indices of unit endogenous variables.	Section 3.3
$\text{pa}(v) \subseteq \mathcal{V}$	Indices of the parents of X^v in the graph \mathcal{G} .	Section 3.3
$\text{pa}_{\mathcal{S}}(v), \text{pa}_{\mathcal{U}}(v)$	Indices of the subunit, unit parents of X^v .	Section 3.3
m	Number of subunits.	Definition 1
n	Number of units.	Definition 1
γ_i^v	Unit noise affecting X^v .	Definition 1
ϵ_{ij}^v	Subunit noise affecting X^v .	Definition 1
$\{x_j\}_{j=1}^m$	Shorthand for the multiset $\{x_1, \dots, x_m\}$.	Section 3.2
\mathcal{M}^{scm}	Hierarchical structural causal model (HSCM).	Definition 1
Δ	Intervention on an HCM.	Definition 2
p_{Δ}	Post-intervention distribution of an HCM.	Theorem 6.
$\mathcal{I} \subseteq \mathcal{V}$	Indices of intervened variables.	Definition 2
x_{\star}^v	Value of X^v in an intervention.	Definition 2
q_{\star}^v	Distribution of X^v in an intervention.	Definition 2.
$\text{do}(x^v = x_{\star})$	A hard intervention that sets X^v to x_{\star} .	Definition 2
$\text{do}(X^v \sim q_{\star}^v)$	A soft intervention that draws X^v from q_{\star}^v .	Definition 2
$Q(x^{\mathcal{S}})$	Within-unit distribution over subunit variables.	Section 5
$Q^v, Q^{v w}$	Q variables in an HCGM/collapsed model.	Definitions 3 and 4, Equation (71)
\mathcal{M}^{cgm}	Hierarchical causal graphical model (HCGM).	Definition 3
$\mathcal{S}_{\text{obs}} \subseteq \mathcal{S}$	Indices of observed subunit variables.	Section 5
$\mathcal{U}_{\text{obs}} \subseteq \mathcal{U}$	Indices of observed unit variables.	Section 5
$\text{da}_{\mathcal{S}}(v)$	Indices of direct subunit ancestors of X^v .	Section 6.1
$\text{dd}_{\mathcal{U}}(v)$	Indices of direct unit descendants of X^v .	Section 6.1
\mathcal{M}^{col}	Collapsed model.	Definition 4
p^{col}	Distribution of collapsed model.	Definition 4, Theorem 6
\mathcal{M}^{aug}	Augmented model.	Definition 8
p^{aug}	Distribution of augmented model.	Definition 8, Section 6.2
\mathcal{M}^{mar}	Marginalized model.	Section 6.2
p^{mar}	Distribution of marginalized model.	Section 6.2
$q^{\mathcal{Q}}$	Set of Q variables in a collapsed model.	Section 6.1
$q^{\mathcal{Q}_{\text{obs}}}$	Set of observed Q variables in a collapsed model.	Section 6.1
\mathcal{M}^{ep}	Erased inner plate model.	Section 6.3
p^{ep}	Distribution of erased inner plate model.	Section 6.3

Table A1: Notation.

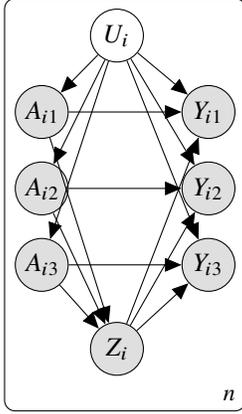


Figure A1: A flat causal model corresponding to the UNIT CONFOUNDER & UNIT INTERFERENCE hierarchical causal model (Figure 1b) with the inner plate expanded, for $m = 3$ subunits.

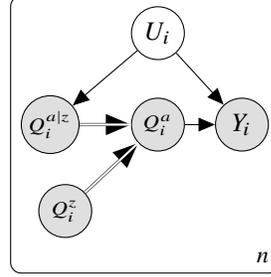


Figure A2: Augmented SUBUNIT INSTRUMENT model.

Appendix A. Marginalizing Hierarchical Causal Models

In this section we describe rules for marginalization in hierarchical causal models. In particular, we are interested in the question of when we can ignore an endogenous variable, deleting it from the graph, without affecting how the remaining endogenous variables are generated. These marginalization rules tell us what sorts of variables we can and cannot safely ignore when constructing a hierarchical causal model. In flat causal models, we can marginalize out any endogenous variable with less than two children (e.g. Richardson and Spirtes, 2002; Janzing and Mejia, 2022; Peters et al., 2017, Chap. 9). For HCMs, we have:

Proposition 14 (Marginalizing HCMs) *Consider a variable Y in an HSCM \mathcal{M}^{scm} . Assume that (1) Y does not have two or more children (Y is not a confounder), and (2) Y is not a unit level variable with both a subunit parent and a subunit child (Y is not an interferer). Then, if Y is marginalized out of \mathcal{M}^{scm} so that it is no longer an endogenous variable, we obtain a valid HSCM.*

Proof The first assumption prevents us from marginalizing out confounders. This is necessary since any HCM reduces, as a special case, to a flat causal model with the same graph (namely, the erased inner plate model, Section 6.3). Next we analyze variables with one child, on a case-by-case basis.

Subunit variable, subunit child. We start by considering the situation where the variable we are interested in marginalizing out, Y , is subunit-level, and has a single subunit child Z . We assume without loss of generality that Y has a single subunit parent X and a single unit parent W . In practice, Y may have more of each kind of parent, but this does not change the analysis, as we can group all the subunit-level parents together and all the unit-level parents together. We also assume without loss of generality that X and W are parents of Z . Now, the structural causal model generating Y and Z is,

$$\begin{aligned} \gamma_i^y &\sim p(\gamma^y) & \epsilon_{ij}^y &\sim p(\epsilon^y) & y_{ij} &= f^y(w_i, \gamma_i^y, x_{ij}, \epsilon_{ij}^y) \\ \gamma_i^z &\sim p(\gamma^z) & \epsilon_{ij}^z &\sim p(\epsilon^z) & z_{ij} &= f^z(w_i, \gamma_i^z, y_{ij}, x_{ij}, \epsilon_{ij}^z). \end{aligned} \quad (40)$$

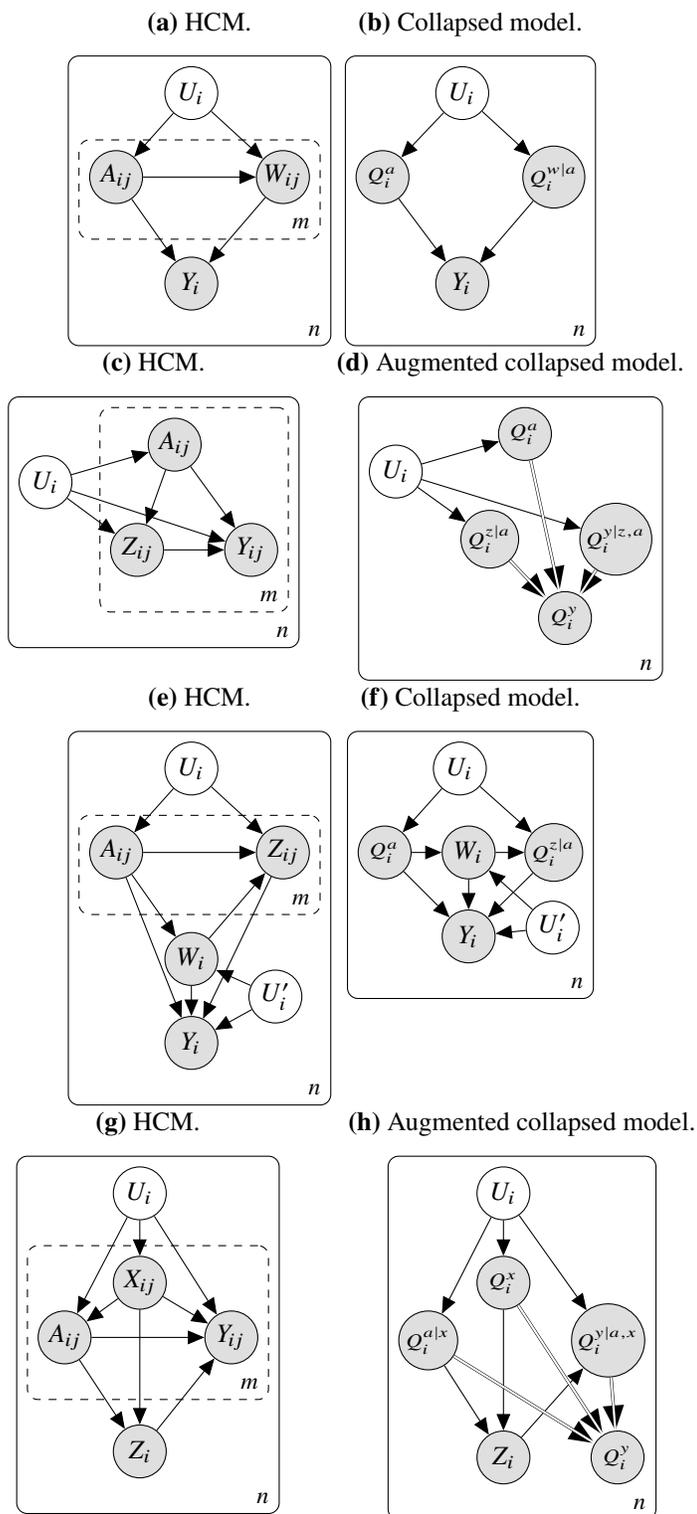


Figure A3: Examples of hierarchical causal models where the effect of A on Y is identified. Each row shows a hierarchical causal model (first plot on the left) and graphs derived from it. In each case, the effect of A on Y would not be identifiable if the inner plate were erased. *Figure continues on the next page.*

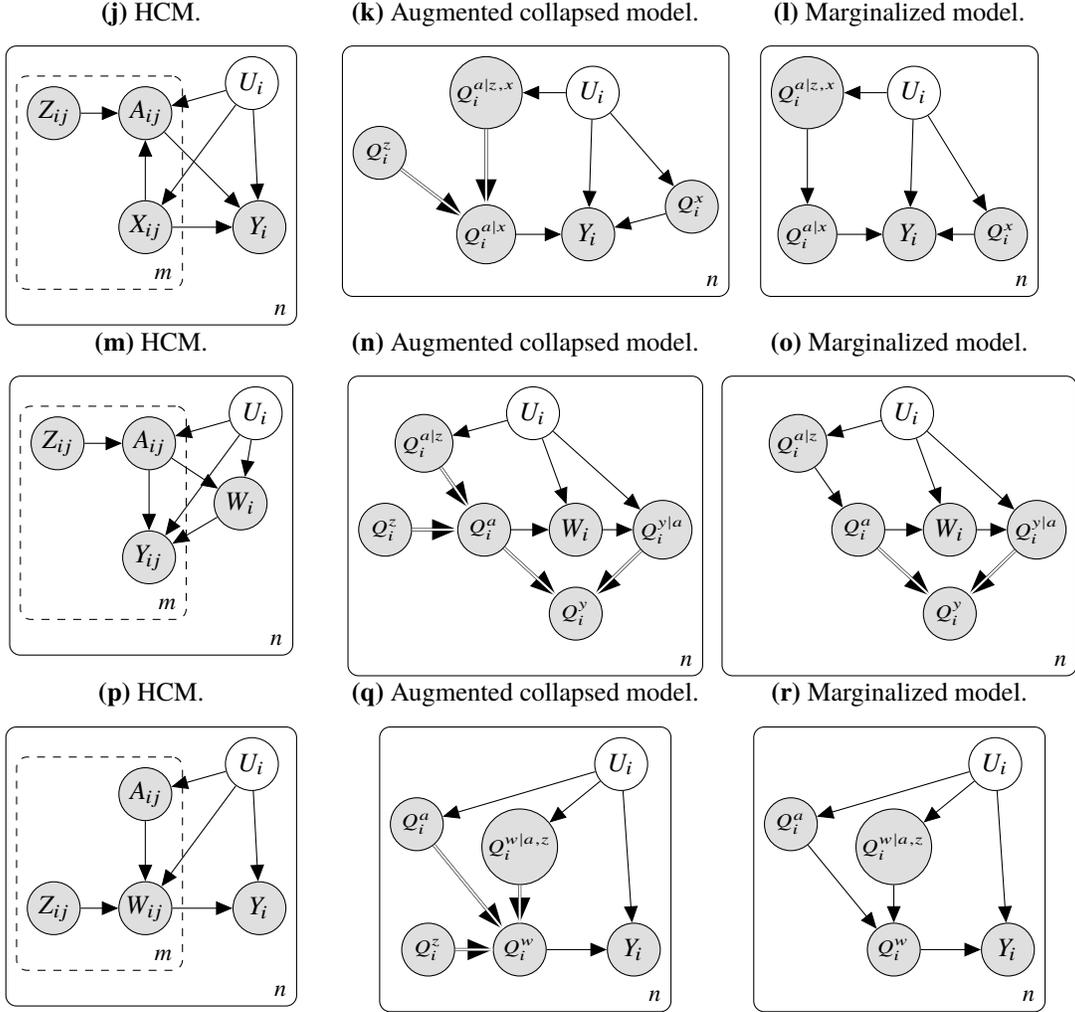


Figure A3: continued

We can marginalize out Y to obtain,

$$\begin{aligned}
 \tilde{\gamma}_i^z &\sim \tilde{p}(\tilde{\gamma}^z) & \tilde{\epsilon}_{ij}^z &\sim \tilde{p}(\tilde{\epsilon}^z) & z_{ij} &= \tilde{f}^z(w_i, \tilde{\gamma}_i^z, x_{ij}, \tilde{\epsilon}_{ij}^z) \\
 \text{where } \tilde{\gamma}_i^z &= (\gamma_i^y, \gamma_i^z), & \tilde{\epsilon}_{ij}^z &= (\epsilon_{ij}^y, \epsilon_{ij}^z), & & \\
 \text{and } \tilde{f}^z(w_i, \tilde{\gamma}_i^z, x_{ij}, \tilde{\epsilon}_{ij}^z) &= f^z(w_i, \gamma_i^z, f^y(w_i, \gamma_i^y, x_{ij}, \epsilon_{ij}^y), x_{ij}, \epsilon_{ij}^z). & & & &
 \end{aligned} \tag{41}$$

The new model for Z meets Definition 1: \tilde{f}^z depends just on x_{ij} and not on $x_{ij'}$ for $j' \neq j$, the noise $\tilde{\gamma}_i^z$ is i.i.d. across units, and the noise $\tilde{\epsilon}_{ij}^z$ is i.i.d. across subunits and units.

Subunit variable, unit child. We next consider the situation where the variable we are interested in marginalizing out, Y , is subunit-level, but its child Z is unit-level. We again have a subunit parent

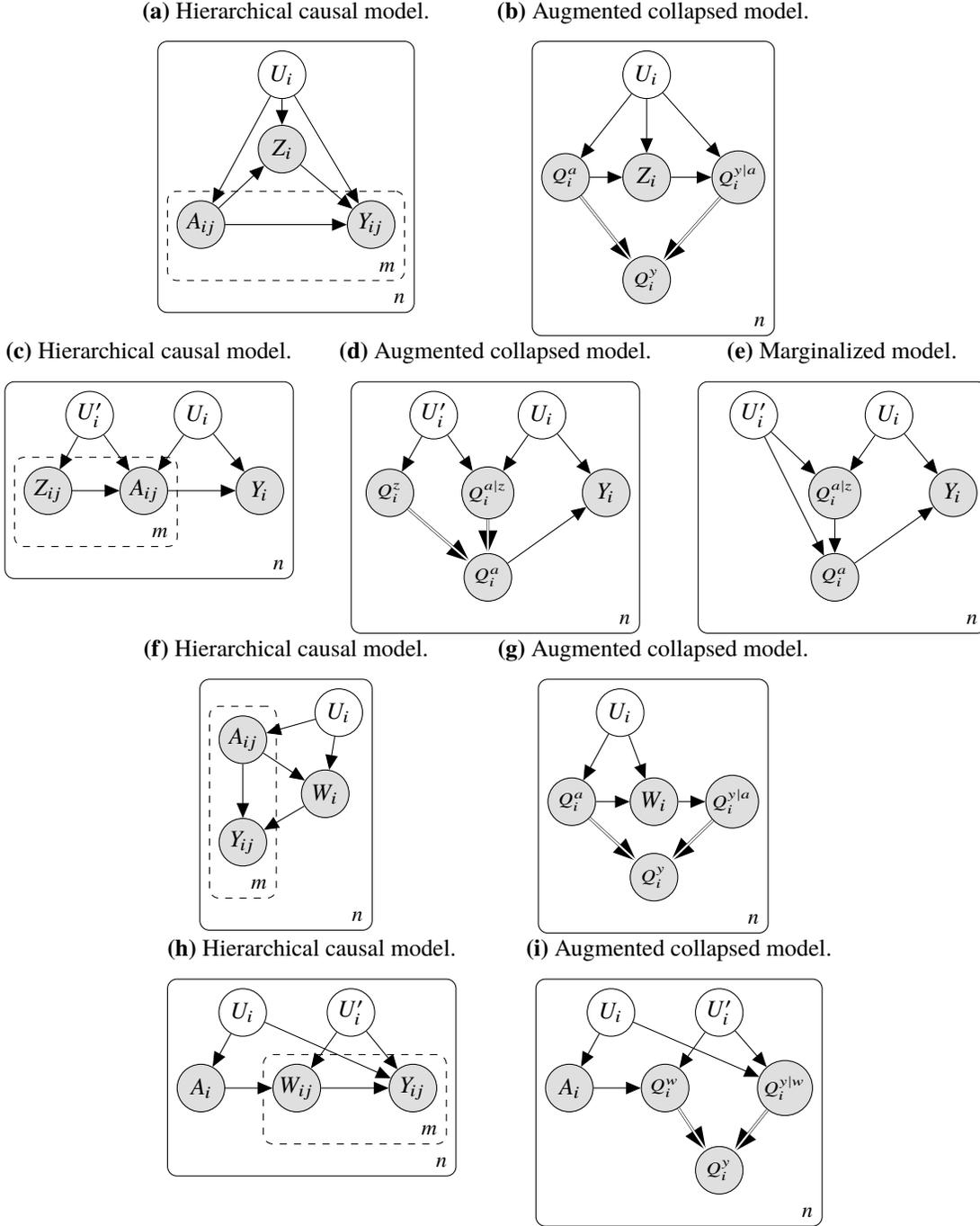


Figure A4: Examples of models where the effect of A on Y is *not* identified via our method.

X and unit parent W of Y and Z . When we marginalize out Y we obtain the model,

$$\begin{aligned} \tilde{\gamma}_i^z &\sim \tilde{p}(\tilde{\gamma}^z) & \tilde{\epsilon}_{ij}^z &\sim \tilde{p}(\tilde{\epsilon}^z) & z_i &= \tilde{f}^z(w_i, \tilde{\gamma}_i^z, \{(x_{ij}, \tilde{\epsilon}_{ij}^z)\}_{j=1}^m) \\ \text{where } \tilde{\gamma}_i^z &= (\gamma_i^y, \gamma_i^z), & \tilde{\epsilon}_{ij}^z &= (\epsilon_{ij}^y, \epsilon_{ij}^z) \end{aligned} \quad (42)$$

and $\tilde{f}^z(w_i, \tilde{\gamma}_i^z, \{(x_{ij}, \tilde{\epsilon}_{ij}^z)\}_{j=1}^m) = f^z\left(w_i, \gamma_i^z, \left\{ (x_{ij}, f^y(x_{ij}, w_i, \gamma_i^y, \epsilon_{ij}^y), \epsilon_{ij}^z) \right\}_{j=1}^m\right)$.

We can see that \tilde{f}^z is invariant to permutations of $(x_{i1}, \tilde{\epsilon}_{i1}^z), \dots, (x_{im}, \tilde{\epsilon}_{im}^z)$. So, the marginalized model meets Definition 1.

Unit variable, unit child. We next consider the case where Y is unit-level, and it has a unit child. We again have a subunit parent X and unit parent W . Marginalizing out Y gives,

$$\begin{aligned} \tilde{\gamma}_i^z &\sim \tilde{p}(\tilde{\gamma}^z) & \tilde{\epsilon}_{ij}^z &\sim \tilde{p}(\tilde{\epsilon}^z) & z_i &= \tilde{f}^z(w_i, \tilde{\gamma}_i^z, \{(x_{ij}, \tilde{\epsilon}_{ij}^z)\}_{j=1}^m) \\ \text{where } \tilde{\gamma}_i^z &= (\gamma_i^y, \gamma_i^z), & \tilde{\epsilon}_{ij}^z &= (\epsilon_{ij}^y, \epsilon_{ij}^z), & & \\ \text{and } \tilde{f}^z(w_i, \tilde{\gamma}_i^z, \{(x_{ij}, \tilde{\epsilon}_{ij}^z)\}_{j=1}^m) &= f^z\left(f^y\left(w_i, \gamma_i^y, \{(x_{ij}, \epsilon_{ij}^y)\}_{j=1}^m\right), w_i, \gamma_i^z, \{(x_{ij}, \epsilon_{ij}^z)\}_{j=1}^m\right). \end{aligned} \quad (43)$$

We see that \tilde{f}^z is invariant to permutations of $(x_{i1}, \tilde{\epsilon}_{i1}^z), \dots, (x_{im}, \tilde{\epsilon}_{im}^z)$, so Definition 1 is met.

Unit variable, subunit child. We now turn to the case where Y is unit-level, and it has a subunit child. We split this case into two sub-cases. The first is when Y has only unit parents; the second is when Y has at least one subunit parent.

In the first sub-case, when Y only has a unit parent W and no subunit parent, we obtain,

$$\begin{aligned} \tilde{\gamma}_i^z &\sim \tilde{p}(\tilde{\gamma}^z) & \tilde{\epsilon}_{ij}^z &\sim \tilde{p}(\tilde{\epsilon}^z) & z_{ij} &= \tilde{f}^z(w_i, \tilde{\gamma}_i^z, \tilde{\epsilon}_{ij}^z) \\ \text{where } \tilde{\gamma}_i^z &= (\gamma_i^y, \epsilon_{i1}^y, \dots, \epsilon_{im}^y, \gamma_i^z), & \tilde{\epsilon}_{ij}^z &= \epsilon_{ij}^z, & & \\ \text{and } \tilde{f}^z(w_i, \tilde{\gamma}_i^z, \tilde{\epsilon}_{ij}^z) &= f^z\left(f^y\left(w_i, \gamma_i^y, \{\epsilon_{ij}^y\}_{j=1}^m\right), w_i, \gamma_i^z, \epsilon_{ij}^z\right). \end{aligned} \quad (44)$$

This marginalized model meets Definition 1. Note we have absorbed the subunit-level noise ϵ^y into the unit-level noise $\tilde{\gamma}^z$ on Z , rather than the subunit-level noise $\tilde{\epsilon}^z$; this is because Z_{ij} cannot depend on $\tilde{\epsilon}_{ij}^y$, for $j' \neq j$.

Finally, we arrive at the case where marginalization is impossible: Y is unit-level and has a single subunit child, but it also has a subunit parent X . The original structural causal model for Y and Z is,

$$\begin{aligned} \gamma_i^y &\sim p(\gamma^y) & \epsilon_{ij}^y &\sim p(\epsilon^y) & y_i &= f^y(\gamma_i^y, \{(x_{ij}, \epsilon_{ij}^y)\}_{j=1}^m) \\ \gamma_i^z &\sim p(\gamma^z) & \epsilon_{ij}^z &\sim p(\epsilon^z) & z_{ij} &= f^z(y_i, \gamma_i^z, \epsilon_{ij}^z). \end{aligned} \quad (45)$$

If we attempt to marginalize out Y , we obtain,

$$z_{ij} = f^z\left(f^y\left(\gamma_i^y, \{(x_{ij}, \epsilon_{ij}^y)\}_{j=1}^m\right), \gamma_i^z, \epsilon_{ij}^z\right). \quad (46)$$

This mechanism does not meet Definition 1, because Z_{ij} can depend on $X_{ij'}$ for $j' \neq j$. Thus, when a variable is an *interferer* – a unit variable with a subunit child and subunit parent – it cannot be marginalized out of an HCM. ■

Confounding is a central problem in causal inference because it is impossible to ignore: confounders cannot be marginalized out of a causal model. Variables that have just one child, though, can be ignored in flat causal models. Intuitively, this implies we do not need to worry about the details of all the intermediate variables between a cause and its effect.

In hierarchical causal models, interference is, like confounding, a central problem: interferers cannot be marginalized out. So, while we can still ignore the intermediate variables between cause

and effect in most cases, we cannot ignore an intermediate variable if it gives rise to interference. This makes the development of techniques to correct for interference a key issue in the study of hierarchical causal models.

Appendix B. Counterfactuals in Hierarchical Causal Models

In this section we introduce counterfactuals in hierarchical structural causal models. We then explain in particular how subunit-level noise contributes to modeling unit variable counterfactuals. This motivates the inclusion of subunit-level noise in the HSCM equations for unit-level variables (Definition 1).

Given a sample of two endogenous variables A and Y from a structural causal model, counterfactuals address the question: what would Y be if A were a_\star , all else held equal? To compute a counterfactual, we (1) fix all the noise variables, (2) set a to a_\star , and (3) run the model forward to compute the resulting value of Y (Pearl, 2009, Chap. 7).

Counterfactuals in HSCMs work the same way. For example, consider an HSCM with a subunit treatment A and unit outcome Y . It has equations

$$\begin{aligned} \gamma_i^a &\sim p(\gamma^y) & \epsilon_{ij}^a &\sim p(\epsilon^y) & a_{ij} &= f^a(\gamma_i^a, \epsilon_{ij}^a) \\ \gamma_i^y &\sim p(\gamma^y) & \epsilon_{ij}^y &\sim p(\epsilon^y) & y_i &= f^y(\gamma_i^y, \{(a_{ij}, \epsilon_{ij}^y)\}_{j=1}^m). \end{aligned} \quad (47)$$

for $j \in \{1, \dots, m\}$ and $i \in \{1, \dots, n\}$. Say we sample from this model, and the value of Y in unit $i = 1$ is $Y_1 = y_1$. We can ask: what would Y_1 be if a_{11} were a_\star ? This counterfactual is given by,

$$\begin{aligned} a_{11} &= a_\star \\ a_{1j} &= f^a(\gamma_1^a, \epsilon_{1j}^a) \text{ for } j \in \{2, \dots, m\} \\ y'_1 &= f^y(\gamma_1^y, \{(a_{1j}, \epsilon_{1j}^y)\}_{j=1}^m). \end{aligned} \quad (48)$$

Here we keep the values of the noise $\gamma_1, \epsilon_{12}, \dots, \epsilon_{1m}$ the same, but plug in the counterfactual value of a_{11} .

As discussed in Section 4.2, the subunit noise in mechanisms for unit variables (e.g. ϵ^y in Equation (47)) does not increase the HSCM's expressiveness in describing interventional distributions. The subunit noise does, however, increase its expressiveness in describing counterfactuals. That is,

$$\gamma_i^y \sim p(\gamma^y) \quad \epsilon_{ij}^y \sim p(\epsilon^y) \quad y_i = f^y(\gamma_i^y, \{(a_{ij}, \epsilon_{ij}^y)\}_{j=1}^m), \quad (49)$$

is a more expressive model for describing counterfactuals than,

$$\gamma_i^y \sim p(\gamma^y) \quad y_i = f^y(\gamma_i^y, \{a_{ij}\}_{j=1}^m), \quad (50)$$

where the subunit noise ϵ^y has been dropped. To see this, consider the counterfactual scenario where we permute the values of a_{i1}, \dots, a_{im} . In the model without subunit noise (Equation (50)), the counterfactual value of Y_i cannot be different from its actual value, since $f^y(\gamma_i^y, \{a_{ij}\}_{j=1}^m) = f^y(\gamma_i^y, \{a_{i\pi(j)}\}_{j=1}^m)$ for any permutation π . With subunit noise (Equation (49)), however, we have no such restriction, and the counterfactual value of Y_i may be different.

To illustrate the importance of this expressivity, we consider a simple scenario, where Y_i represents whether school i appears on a ‘‘best schools’’ list. Say Y_i depends on whether the fraction

of students at the school who pass all their classes is 50% or more. However, we only observe a_{ij} , which indicates whether student j passes their English class. Then, the subunit noise ϵ_{ij}^y describes whether student j passes the rest of their classes, and the mechanism generating Y_i can be written,

$$y_i = f^y(\{(a_{ij}, \epsilon_{ij}^y)\}_{j=1}^m) = \mathbb{1}\left(\frac{1}{m} \sum_{j=1}^m a_{ij} \epsilon_{ij}^y \geq 0.5\right). \quad (51)$$

Consider a school $i = 1$ with just two students. One of the students passes all their classes ($a_{11} = \epsilon_{11}^y = 1$) and the other none of them ($a_{12} = \epsilon_{12}^y = 0$), so that $y_1 = 1$. Let us investigate the counterfactual scenario where the English grades of the students are switched, such that the first student fails ($a_{11} = 0$) and the second passes ($a_{12} = 1$). From Equation (51), we see the counterfactual value of y_1 is $y_1' = \frac{1}{m}(0 \cdot 1 + 1 \cdot 0) = 0$. That is, in this counterfactual scenario, the school does not appear on the “best schools” list, since none of their students passed all their classes. By contrast, a model that did not include subunit noise (Equation (50)) would demand, unreasonably, that y_1' must still be 1 in this counterfactual scenario.

Appendix C. Deriving Hierarchical Causal Graphical Models

In this section, we derive hierarchical causal graphical models (Definition 3) from hierarchical structural causal models (Definition 1). The derivation is a generalization of the special case given in Section 4.2.

For each subunit variable $v \in \mathcal{S}$, we have

$$\begin{aligned} [g_q^{v|\text{pa}_{\mathcal{S}}(v)}(x^{\text{pa}_{\mathcal{Z}}(v)}, \gamma^v)](X^v \in \Xi \mid x^{\text{pa}_{\mathcal{S}}(v)}) &\triangleq \int \mathbb{1}(f^v(x^{\text{pa}_{\mathcal{Z}}(v)}, \gamma^v, x^{\text{pa}_{\mathcal{S}}(v)}, \epsilon^v) \in \Xi) p(\epsilon^v) d\epsilon^v \\ P(Q^v|\text{pa}_{\mathcal{S}}(v)) \in \Pi \mid x^{\text{pa}_{\mathcal{Z}}(v)}) &\triangleq \int [g_q^{v|\text{pa}_{\mathcal{S}}(v)}(x^{\text{pa}_{\mathcal{Z}}(v)}, \gamma^v) \in \Pi] p(\gamma^v) d\gamma^v. \end{aligned} \quad (52)$$

In the first line, we marginalize out the subunit noise ϵ^v to produce a function $g_q^{v|\text{pa}_{\mathcal{S}}(v)}$ that takes in unit-level variables $X^{\text{pa}_{\mathcal{Z}}(v)}$ and γ^v , and returns a conditional distribution over the subunit-level variable X^v given the subunit-level variables $X^{\text{pa}_{\mathcal{S}}(v)}$. In the second line, we marginalize out the unit-level noise γ^v to produce a conditional distribution over subunit distributions.

For each unit variable $v \in \mathcal{U}$, we have

$$\begin{aligned} P(x^v \in \Xi \mid x^{\text{pa}_{\mathcal{Z}}(v)}, \{x_j^{\text{pa}_{\mathcal{S}}(v)}\}_{j=1}^m) \\ = \int \dots \int \mathbb{1}(f^v(x_i^{\text{pa}_{\mathcal{Z}}(v)}, \gamma^v, \{(x_j^{\text{pa}_{\mathcal{S}}(v)}, \epsilon_j^v)\}_{j=1}^m) \in \Xi) p(\gamma^v) d\gamma^v \prod_{j=1}^m p(\epsilon_j^v) d\epsilon_j^v. \end{aligned} \quad (53)$$

We can confirm that the stochastic mechanism $P(x^v \in \Xi \mid x^{\text{pa}_{\mathcal{Z}}(v)}, \{x_j^{\text{pa}_{\mathcal{S}}(v)}\}_{j=1}^m)$ is invariant to permutations of the subunits, as for any permutation π of $\{1, \dots, m\}$ we have,

$$= \int \dots \int \mathbb{1}(f^v(x_i^{\text{pa}_{\mathcal{Z}}(v)}, \gamma^v, \{(x_{\pi(j)}^{\text{pa}_{\mathcal{S}}(v)}, \epsilon_j^v)\}_{j=1}^m) \in \Xi) p(\gamma^v) d\gamma^v \prod_{j=1}^m p(\epsilon_j^v) d\epsilon_j^v, \quad (54)$$

since f^v is invariant to permutations of $\{(x_j^{\text{pa}_{\mathcal{S}}(v)}, \epsilon_j^v)\}_{j=1}^m$ and each ϵ_j^v is drawn i.i.d. for $j \in \{1, \dots, m\}$.

Appendix D. Details on Simulations

D.1 UNIT CONFOUNDER

In this section we describe in detail our simulation and estimation procedures for the UNIT CONFOUNDER model (Section 5.1). Each endogenous variable (U , A and Y) is binary, and the data-generating HCGM is,

$$\begin{aligned} U_i &\sim \text{Bernoulli}(\omega) \\ \mu_i^a &\sim \text{Beta}(\alpha^a(u_i), \beta^a) & q_i^a &= \text{Bernoulli}(\mu_i^a) & A_{ij} &\sim q_i^a(a) \\ \mu_i^{y|a}(a) &\sim \text{Beta}(\alpha^{y|a}(a, u_i), \beta^{y|a}) & q_i^{y|a}(\cdot | a) &= \text{Bernoulli}(\mu_i^{y|a}(a)) & Y_{ij} &\sim q_i^{y|a}(y | a_{ij}), \end{aligned}$$

for $j \in \{1, \dots, m\}$ and $i \in \{1, \dots, n\}$. We set $\alpha^a(0) = 0.5, \alpha^a(1) = 4, \alpha^{y|a}(0, 0) = 0.5, \alpha^{y|a}(1, 0) = 2, \alpha^{y|a}(0, 1) = 1, \alpha^{y|a}(1, 1) = 4, \beta^a = 1$ and $\beta^{y|a} = 2$. We simulate with different values of ω , which governs the amount of confounding. From the data generating model, we can calculate the true effect as,

$$\mathbb{E}_p[\mathbb{E}_Q[Y]; \text{do}(q^a = \delta_{a_\star})] = (1 - \omega) \frac{\alpha^{y|a}(a_\star, 0)}{\alpha^{y|a}(a_\star, 0) + \beta^{y|a}} + \omega \frac{\alpha^{y|a}(a_\star, 1)}{\alpha^{y|a}(a_\star, 1) + \beta^{y|a}},$$

for $a_\star \in \{0, 1\}$, where we have used the fact that the mean of $\text{Beta}(\alpha, \beta)$ is $\alpha/(\alpha + \beta)$.

We now turn to estimation.

1. For $q_i^{y|a}$ we use the point estimate,

$$\hat{q}_i^{y|a}(\cdot | a) = \text{Bernoulli} \left(\hat{\mu}_i^{y|a}(a) = \frac{\sum_{j=1}^m y_{ij} \delta_a(a_{ij}) + 1}{\sum_{j=1}^m \delta_a(a_{ij}) + 2} \right), \quad (55)$$

for each unit $i \in \{1, \dots, n\}$ and for $a \in \{0, 1\}$. Here we have added pseudocounts for regularization, i.e. $\hat{\mu}_i^{y|a}(a)$ corresponds the posterior mean of $\mathbb{E}_{q_i^{y|a}}[Y | a]$ under a $\text{Beta}(1, 1)$ prior.

2. We can compute $\hat{\mu}_i^y(a_\star) = \hat{\mu}_i^{y|a}(a_\star)$ since the intervention distribution q_\star^a is a point mass at a_\star .

Applying Equation (23), we can estimate the treatment effect as $\frac{1}{n} \sum_{i=1}^n \hat{\mu}_i^y(1) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_i^y(0)$.

We simulated data sets of size $n = 1000$ and $m = 1000$ from the data generating model, then constructed our estimate based on subsets of increasing size (observing 10 subunits and units, 100 subunits and units, etc.). Each panel of Figure 7 shows the convergence of our estimator to the true effect with increasing data, across 20 independent simulations. We use $\omega = 0$ for the ‘‘no confounding’’ simulations (Figure 7a), $\omega = 0.2$ for ‘‘low confounding’’ (Figure 7b) and $\omega = 0.5$ for ‘‘high confounding’’ (Figure 7c).

We compare to a regression estimator that comes from naively applying linear regression to aggregated data. In particular, we estimate $\mathbb{E}_p[\bar{Y} | \bar{A} = 1] - \mathbb{E}_p[\bar{Y} | \bar{A} = 0]$, where $\bar{y} \triangleq \mathbb{E}_q[Y] = \int y q^y(y) dy$ is the within-unit average of y and $\bar{a} \triangleq \mathbb{E}_q[A] = \int a q^a(a) da$ is the within-unit average of a . We perform our estimate by running a linear regression predicting $\hat{y}_i = \frac{1}{m} \sum_{j=1}^m y_{ij}$ from $\hat{a}_i = \frac{1}{m} \sum_{j=1}^m a_{ij}$ for each unit i .

In the presence of confounding, the linear regression estimate does not converge to the true effect (Figure 7b, Figure 7c). Without confounding, the linear regression estimate does converge to the

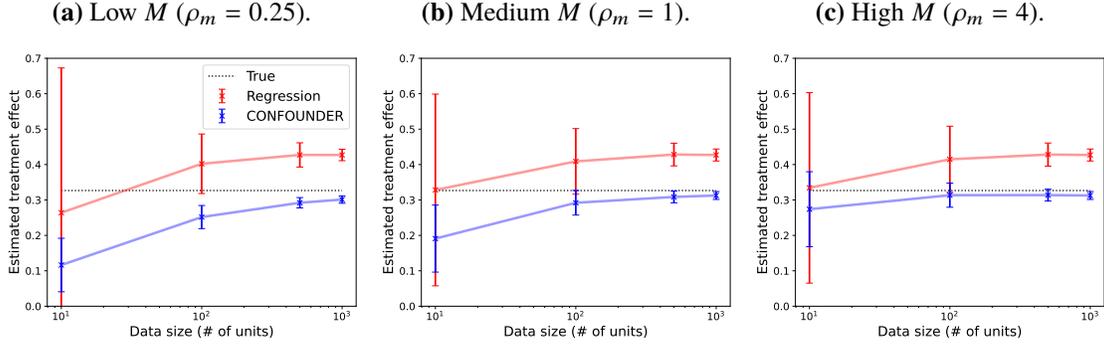


Figure A5: Estimated effects in a UNIT CONFOUNDER simulation for changing number of subunits per unit ($\rho_m = M/N$). Error bars show standard deviation across 20 independent simulations.

true effect (Figure 7a). Note that in general, a linear regression estimate will not always be accurate in the absence of confounding; rather, its success depends on the fact that A and Y are binary in our simulation.

We also evaluated estimation performance as the proportion of subunits to units changes. We set $M = \lceil \rho_m N \rceil$ for different ρ_m , and track estimator convergence as N increases (Figure A5). As ρ_m increases, the bias at low N decreases. This is in line with the theory of nested Monte Carlo estimation (Rainforth et al., 2018, Appendix B).

D.2 UNIT CONFOUNDER & UNIT INTERFERENCE

We next detail our simulations and estimation procedures for the INTERFERENCE model (Section 5.2). The data generating HCGM for our simulation is,

$$\begin{aligned}
 U_i &\sim \text{Normal}(0, 1) \\
 v_i^a &\sim \text{Normal}(0.5 u_i, 1) & q_i^a &= \text{Bernoulli}(\sigma(v_i^a)) & A_{ij} &\sim q_i^a(a) \\
 Z_i &\sim \text{Bernoulli}\left(\sigma\left(2\sigma^{-1}\left(\frac{1}{m}\sum_{j=1}^m a_{ij}\right) - 0.8\right)\right) \\
 v_i^{y|a}(a) &\sim \text{Normal}(g_\rho(a, z_i, 0.5u_i), 0.1) & q_i^{y|a}(\cdot | a) &= \text{Bernoulli}(\sigma(v_i^{y|a}(a))) & Y_{ij} &\sim q_i^{y|a}(y | a_{ij}),
 \end{aligned}$$

for $j \in \{1, \dots, m\}$ and $i \in \{1, \dots, n\}$, where $\text{Normal}(\mu, \tau)$ is a normal distribution with mean μ and variance τ^2 , $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic sigmoid function, $\sigma^{-1}(x) = \log(x/(1 - x))$ is its inverse, and $g_\rho(a, z_i, 0.5u_i) = 0.5a + \rho(2z_i - 1) + 0.5u_i$. We simulate with different values of ρ , which determines the strength of interference.

From the data generating process, we can calculate the true effect of an intervention with $q_\star^a = \text{Bernoulli}(\mu_\star)$ as

$$\begin{aligned} \mathbb{E}_p[\mathbb{E}_{Q^y}[Y]; \text{do}(q^a = q_\star^a)] &= \mathbb{E}_{q_\star^a} \left[\mathbb{E}_{p(z|q^a)} \left[\mathbb{E}_{\int p(q^{y|a}|u,z)p(u)du} \left[\mathbb{E}_{Q^{y|a}}[Y | A] | Z \right] | q_\star^a \right] \right] \\ &= \sum_{a=0}^1 (\mu_\star)^a (1 - \mu_\star)^{1-a} \sum_{z=0}^1 \sigma(2\sigma^{-1}(\mu_\star) - 0.8)^z (1 - \sigma(2\sigma^{-1}(\mu_\star) - 0.8))^{1-z} \\ &\quad \int \mathcal{N}(x | 0, 1) \sigma(g_\rho(a, z, \sqrt{0.5^2 + 0.1^2 x})) \end{aligned}$$

Here, $\mathcal{N}(x | 0, 1)$ denotes the pdf of a standard normal, and we have used the fact that the distribution $\int p(q^{y|a}|z, u)p(u)du$ can be rewritten from a sum of independent normal distributions to depend on a single standard normal x . We compute the Gaussian integral numerically.

We now turn to estimation.

1. We estimate q_i^a for each unit i as $\hat{q}_i^a = \text{Bernoulli}(\hat{\mu}_i^a = \frac{1}{m+2}(1 + \sum_{j=1}^m a_{ij}))$, again using pseudo-counts for regularization.
2. We estimate $q_i^{y|a}$ for each unit i with Equation (55).
3. We estimate $p(z | q^a)$ with logistic regression, predicting z_i from $\sigma^{-1}(\hat{\mu}_i^a)$. This gives an estimate $\hat{p}(z | \mu^a)$.
4. We estimate $p(q^{y|a} | q^a, z)$ using four separate linear regressions. For each $a \in \{0, 1\}$ and $z \in \{0, 1\}$, we predict $\sigma^{-1}(\hat{\mu}_i^{y|a}(a))$ from $\sigma^{-1}(\hat{\mu}_i^a)$ for all units i such that $z_i = z$. This gives an estimate $\hat{p}(\sigma^{-1}(\mu^{y|a}) | \mu^a, z)$.
5. We estimate $p(q^a)$ with the empirical distribution of $\hat{\mu}_i^a$ for all units $i \in \{1, \dots, n\}$.

Finally, we combine these estimates following the identification formula,

$$\begin{aligned} \mathbb{E}_p[\mathbb{E}_{Q^y}[Y]; \text{do}(q^a = q_\star^a)] &= \mathbb{E}_{A \sim q_\star^a} \left[\int p(z | q_\star^a) \int p(q^a) \mathbb{E}_p[\mathbb{E}_{Q^y}[Y | A] | q^a, z] dq^a dz \right] \\ &\approx \sum_{a=0}^1 (\mu_\star)^a (1 - \mu_\star)^{1-a} \sum_{z=0}^1 \hat{p}(z | \mu_\star) \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{p}}[\mu^{y|a}(a) | \hat{\mu}_i^a, z]. \end{aligned}$$

To compute $\mathbb{E}_{\hat{p}}[\mu^{y|a}(a) | \hat{\mu}_i^a, z]$, we use Monte Carlo integration: we draw 100 samples from $\hat{p}(\sigma^{-1}(\mu^{y|a}(a)) | \hat{\mu}_i^a, z)$, apply $\sigma(\cdot)$ to each sample, then take the average.

As before, we simulated data sets of size $n = 1000$ and $m = 1000$, and constructed our estimate based on subsets of increasing size. Note that here the ground truth has finite but large m , while our identification technique makes the approximation $m \rightarrow \infty$. Each panel of Figure 8 shows the convergence of our estimator to the true, analytically computed, $m \rightarrow \infty$ effect with increasing data set size, across 20 independent simulations. We use $\rho = 1.5$ for the ‘‘high interference’’ simulations (Figure 8c), $\rho = 0.5$ for the ‘‘low interference’’ (Figure 8b) and $\rho = 0$ for the ‘‘no interference’’ (Figure 8a).

We also evaluated estimation performance as the proportion of subunits to units changes. We set $M = \lceil \rho_m N \rceil$ for different ρ_m , and track estimator convergence as N increases (Figure A6). As ρ_m increases, the bias at low N decreases (Rainforth et al., 2018, Appendix B).

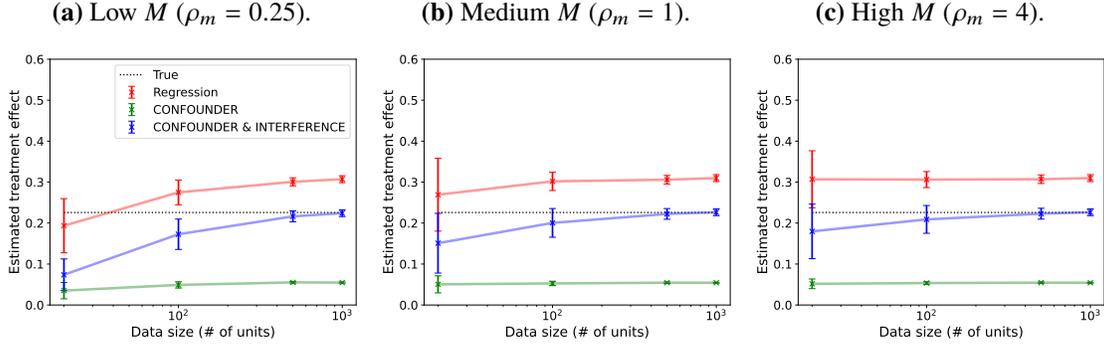


Figure A6: Estimated effects in a UNIT CONFOUNDER & UNIT INTERFERENCE simulation for changing number of subunits per unit ($\rho_m = M/N$). Error bars show standard deviation across 20 independent simulations.

D.3 SUBUNIT INSTRUMENT

Finally we detail our simulations and estimation procedures for the SUBUNIT INSTRUMENT model (Section 5.3). In this simulation, the data generating HCGM is,

$$\begin{aligned}
 U_i &\sim \text{Bernoulli}(\omega) \\
 \mu_i^z &\sim \text{Beta}(2, 2) & q_i^z &= \text{Bernoulli}(\mu_i^z) & Z_{ij} &\sim q_i^z(z) \\
 \epsilon_i^{a|z}(z) &\sim \text{Beta}(2 - 1.8u_i, 0.2 + 1.8u_i) & q_i^{a|z}(\cdot | z) &= \text{Bernoulli}(0.8z + 0.2 \epsilon_i^{a|z}(z)) & A_{ij} &\sim q_i^{a|z}(a | z_{ij}) \\
 Y_i &\sim \text{Bernoulli}\left(0.45 - 0.4 u_i + 0.5 \frac{1}{m} \sum_{j=1}^m a_{ij}\right),
 \end{aligned}$$

for $j \in \{1, \dots, m\}$ and $i \in \{1, \dots, n\}$. Here, ω governs the amount of confounding.

We are interested in the effects that soft interventions on A have on the outcome Y . In the infinite m limit, $\mathbb{E}_p[Y; \text{do}(q^a = q_\star)] = 0.45 - 0.4\omega + 0.5\mu_\star$, where the intervention distribution is $q_\star^a = \text{Bernoulli}(\mu_\star)$. Note that for the interventions we consider, namely $\mu_\star = 0.25$ and $\mu_\star = 0.75$, the effect is identified, since it satisfies the positivity condition (Assumption 6). In particular, for any $\mu_i^{a|z}$, with probability 1 there exists a value of μ^z such that $\mu_\star = (1 - \mu^z)\mu_i^{a|z}(0) + \mu^z\mu_i^{a|z}(1)$. This is because in the simulation, $\mu_i^{a|z}(1) \geq 0.8 > \mu_\star > 0.2 \geq \mu_i^{a|z}(0)$ with probability one.

We now turn to estimation. So far, in previous simulations, we have focused on parametric models. Here we explore a nonparametric outcome model (a Gaussian process classifier).

1. For each unit i , we estimate $q_i^{a|z}$ just as we estimated $q_i^{y|a}$ in Appendix D.2, obtaining $\hat{q}_i^{a|z} = \text{Bernoulli}(\hat{\mu}_i^{a|z})$.
2. For each unit i , we estimate q_i^a just as we estimated q_i^a in Appendix D.2, obtaining $\hat{q}_i^a = \text{Bernoulli}(\hat{\mu}_i^a)$.
3. We estimate $p(y | q^a, q^{a|z})$ with a Gaussian process classifier, predicting y_i from $\sigma^{-1}(\hat{\mu}_i^a)$, $\sigma^{-1}(\hat{\mu}_i^{a|z}(0))$ and $\sigma^{-1}(\hat{\mu}_i^{a|z}(1))$. This gives an estimate $\mathbb{E}_{\hat{p}}[Y | \sigma^{-1}(\hat{\mu}_i^a), \sigma^{-1}(\hat{\mu}_i^{a|z}(0)), \sigma^{-1}(\hat{\mu}_i^{a|z}(1))]$.

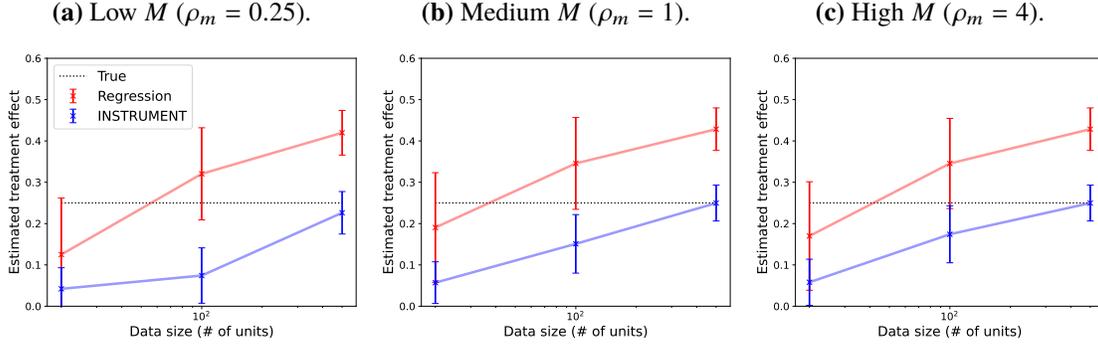


Figure A7: Estimated effects in a SUBUNIT INSTRUMENT simulation for changing number of subunits per unit ($\rho_m = M/N$). Error bars show standard deviation across 20 independent simulations.

Finally, we combine these estimates following the identification formula,

$$\mathbb{E}_p[Y; \text{do}(q^a = q_\star)] \approx \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{p}}[Y \mid \sigma^{-1}(\mu_\star), \sigma^{-1}(\hat{\mu}_i^a|z(0)), \sigma^{-1}(\hat{\mu}_i^a|z(1))].$$

We simulate data sets of size $n = 1000$ and $m = 1000$, and constructed our estimate based on subsets of increasing size. Figure 9 shows the convergence of our estimator to the true $m \rightarrow \infty$ effect with increasing data set size, across 20 independent simulations. We use $\omega = 0$ for the “no confounding” simulations (Figure 9a), $\omega = 0.2$ for the “low confounding” (Figure 9b) and $\omega = 0.5$ for the “high confounding” (Figure 9c). As a comparison, we also plot the behavior of an estimator which does not use the backdoor correction, and just predicts y_i from $\sigma^{-1}(\hat{\mu}_i^a)$ using a Gaussian process classifier. This naive approach is incorrect for this data generating model, as it ignores confounding.

We also evaluated estimation performance as the proportion of subunits to units changes. We set $M = \lceil \rho_m N \rceil$ for different ρ_m , and track estimator convergence as N increases (Figure A7). As ρ_m increases, the bias at low N decreases (Rainforth et al., 2018, Appendix B).

Appendix E. Convergent and Divergent Mechanisms

In this section we discuss the assumption that causal mechanisms converge (Definition 5), which is a key assumption on which our identification method rests. We describe examples of convergent and divergent mechanisms, and outline their general features.

E.1 Convergent mechanisms

A simple example of a mechanism that converges with infinite subunits is $p(y \mid \{a_j\}_{j=1}^m) = \text{Normal}(y \mid \frac{1}{m} \sum_{j=1}^m h(a_j), \sigma)$. Note here that the mechanism can be written in terms of the empirical distribution of samples, $\text{Normal}(y \mid \mathbb{E}_{A \sim \hat{q}_m(a)}[h(A)], \sigma)$ with $\hat{q}_m(a) = \frac{1}{m} \sum_{j=1}^m \delta_{a_j}(a)$. If $\mathbb{E}_{q(a)}[h(A)]$ is finite, then the mechanism will converge to $\text{Normal}(y \mid \mathbb{E}_{q(a)}[h(A)], \sigma)$. In particular, plugging in the formula for the KL divergence between two Gaussians, and applying the continuous

mapping theorem, we have

$$D_m^y(x^{\text{pa}_Z(y)}, q(x^{\text{pa}_S(y)})) = \frac{1}{2\sigma^2} (\mathbb{E}_{\hat{q}_m(a)}[h(A)] - \mathbb{E}_{q(a)}[h(A)])^2 \rightarrow 0, \quad (56)$$

as $m \rightarrow \infty$ a.s..

This is one example of a broad family of mechanisms based on exponential family distributions that exhibit convergence, namely mechanisms of the form,

$$p(x^v \mid x^{\text{pa}_Z(v)}, \{x_j^{\text{pa}_S(v)}\}_{j=1}^m) = r\left(x^v; g\left(\frac{1}{m} \sum_{j=1}^m h(x^{\text{pa}_Z(v)}, x_j^{\text{pa}_S(v)})\right)\right), \quad (57)$$

where h is a function, g is a continuous function, and $r(x; \eta) = \lambda(x) \exp(\eta^\top s(x) - \kappa(\eta))$, is an exponential family distribution with natural parameter $\eta \in \mathbb{R}^K$. This class of mechanisms is quite general, as for any finite m , any continuous function of $x^{\text{pa}_Z(v)}$ and $\{x_j^{\text{pa}_S(v)}\}_{j=1}^m$ can be written in the form $g\left(\frac{1}{m} \sum_{j=1}^m h(x^{\text{pa}_Z(v)}, x_j^{\text{pa}_S(v)})\right)$ for continuous g and h (Zaheer et al., 2017, Theorem 7). Under mild regularity conditions on the exponential family, the mechanism in Equation (57) will converge.

Proposition 15 *Assume $\{\eta \in \mathbb{R}^K : |\kappa(\eta)| < \infty\}$ is an open and nonempty subset of \mathbb{R}^K , and that $\eta \rightarrow r(\cdot; \eta)$ is one-to-one. Further assume $|\mathbb{E}_{q(x^{\text{pa}_S(v)})}[h(x^{\text{pa}_Z(v)}, X^{\text{pa}_S(v)})]| < \infty$. Then, the exponential family mechanism (Equation (57)) converges with infinite subunits (Definition 5).*

Proof The limiting distribution is

$$p(x^v \mid x^{\text{pa}_Z(v)}, q(x^{\text{pa}_S(v)})) = r\left(x^v; g\left(\mathbb{E}_{q(x^{\text{pa}_S(v)})}[h(x^{\text{pa}_Z(v)}, X^{\text{pa}_S(v)})]\right)\right). \quad (58)$$

Define $\eta_m = g\left(\frac{1}{m} \sum_{j=1}^m h(x^{\text{pa}_Z(v)}, x_j^{\text{pa}_S(v)})\right)$ and $\eta_0 = g\left(\mathbb{E}_{q(x^{\text{pa}_S(v)})}[h(x^{\text{pa}_Z(v)}, X^{\text{pa}_S(v)})]\right)$, so

$$\text{KL}\left(p(x^v \mid x^{\text{pa}_Z(v)}, q(x^{\text{pa}_S(v)})) \parallel p(x^v \mid x^{\text{pa}_Z(v)}, \{x_j^{\text{pa}_S(v)}\}_{j=1}^m)\right) = \text{KL}(r(x^v; \eta_0) \parallel r(x^v; \eta_m)). \quad (59)$$

The KL divergence between two members of an exponential family can be written as a Bregman divergence between their natural parameters (e.g. Nielsen and Garcia, 2009), giving

$$\text{KL}(r(x^v; \eta_0) \parallel r(x^v; \eta_m)) = B_\kappa(\eta_m \parallel \eta_0) = \kappa(\eta_m) - \kappa(\eta_0) - \nabla_\eta \kappa(\eta_0)^\top (\eta_m - \eta_0), \quad (60)$$

where B_κ is the Bregman divergence. By the strong law of large numbers and the continuous mapping theorem, $\eta_m \rightarrow \eta_0$ a.s. as $m \rightarrow \infty$. By standard properties of exponential family models (e.g. Miller and Harrison, 2014, Prop. 19), κ is C^∞ , i.e. it has continuous derivatives of all orders. Thus by the continuous mapping theorem, $B_\kappa(\eta_m \parallel \eta_0) \rightarrow 0$ a.s., and the conclusion follows. ■

E.2 Approximating divergent mechanisms

Some mechanisms diverge in the infinite subunit limit. For example, consider a situation in which Y_i depends on the total rather than an average of its parent subunit variable A_{ij} , namely $p(y \mid \{a_{ij}\}_{j=1}^m) = \text{Normal}(y \mid \sum_{j=1}^m h(a_{ij}), \sigma)$. If h is strictly positive then as $m \rightarrow \infty$ we have $\sum_{j=1}^m h(a_{ij}) \rightarrow \infty$.

For example, consider a scenario where Y_i is the school budget, A_{ij} the number of classes in which student j enrolls, and the budget depends on the total enrollment.

In such situations, however, it is often reasonable to approximate the original divergent model with an alternative model that does exhibit convergence. In particular, consider an HCM where the total number of subunits (e.g. the number of students in the school) is represented as a separate unit-level variable S_i , and we have the mechanism $p(y \mid s_i, \{a_{ij}\}_{j=1}^m) = \text{Normal}(y \mid s_i \frac{1}{m} \sum_{j=1}^m h(a_{ij}), \sigma)$. This mechanism matches the original mechanism if $s_i = m$, but if we allow s_i to be held fixed as m increases, the new mechanism converges to $\text{Normal}(y \mid s_i \mathbb{E}_{q_i(a)}[h(A)], \sigma)$. This limit corresponds to an approximation in which we ignore sampling variability among subunits, and replace the empirical mean $\frac{1}{m} \sum_{j=1}^m h(a_{ij})$ with the mean of the underlying distribution, $\mathbb{E}_{q_i(a)}[h(A)]$. This approximation can be reasonable in settings where the number of subunits is large, according to the law of large numbers.⁴

Appendix F. Proof of Theorem 6 (Collapsing a Hierarchical Causal Model)

In this section we prove Theorem 6, which says that in the limit of infinite subunits, hierarchical causal graphical models converge to collapsed models.

Before beginning the proof, we briefly review two key properties of the KL divergence. First, the KL divergence can be decomposed into a sum of conditionals: $\text{KL}(p(x, y) \parallel p'(x, y)) = \mathbb{E}_p[\log \frac{p(X)}{p'(X)}] + \mathbb{E}_p[\log \frac{p(Y|X)}{p'(Y|X)}] = \text{KL}(p(x) \parallel p'(x)) + \text{KL}(p(y \mid x) \parallel p'(y \mid x))$. Second, from Jensen's inequality, we have $\text{KL}(p(y) \parallel p'(y)) = \mathbb{E}_p[\log p(Y)] - \mathbb{E}_p[\log \int p'(Y \mid x)p'(x)dx] \leq \mathbb{E}_p[\log p(Y)] - \mathbb{E}_{Y \sim p(y)}[\mathbb{E}_{X \sim p'(x)}[\log p'(Y \mid X)]] = \mathbb{E}_{p'}[\text{KL}(p(y) \parallel p'(y \mid X))]$.

Proof The idea of the proof is to bound the KL divergence between the original HCGM and the collapsed model in terms of the sum of the expected KL divergence between each mechanism.

It will suffice to prove convergence for the observational distribution divergence, i.e. $\text{KL}(p^{\text{col}}(x^{\mathcal{U}}, q(x^{\mathcal{S}})) \parallel p_m(x^{\mathcal{U}}, q(x^{\mathcal{S}})))$. The reason is that under the interventions defined by Definition 2, the post-intervention HCGM and collapsed model take the same form as the pre-intervention HCGM and collapsed model, just with a different choice of mechanism for the intervened variables. For example, the distribution $p_{\Delta, m}(x^{\mathcal{U}}, q(x^{\mathcal{S}}))$ under the intervention $\text{do}(X^v \sim q_{\star}^{v|pa_{\mathcal{S}}(v)}(x^v \mid X_{ij}^{pa_{\mathcal{S}}(v)}))$ is identical to the distribution $\tilde{p}(x^{\mathcal{U}}, q(x^{\mathcal{S}}))$ under a modified model in which $p(q^{v|pa_{\mathcal{S}}(v)} \mid x^{pa_{\mathcal{Z}}(v)})$ is replaced by $\tilde{p}(q^{v|pa_{\mathcal{S}}(v)} \mid x^{pa_{\mathcal{Z}}(v)}) = \delta_{q_{\star}^{v|pa_{\mathcal{S}}(v)}}(q^{v|pa_{\mathcal{S}}(v)})$. Likewise for the collapsed model, $p_{\Delta}^{\text{col}}(x^{\mathcal{U}}, q(x^{\mathcal{S}}))$. Moreover, in the statement of Theorem 6, the same assumptions on the mechanisms apply for the pre- and post-intervention distributions. So in short, convergence of the post-intervention distribution will follow as a special case of convergence of the observational distribution.

Also, it is convenient to convert from distributions over subunit joint distributions $q(x^{\mathcal{S}})$ to distributions over subunit conditionals $q^{\mathcal{Q}} = \{q^{v|pa_{\mathcal{S}}(v)} : v \in \mathcal{S}\}$, where $q(x^{\mathcal{S}}) = \prod_{v \in \mathcal{S}} q^{v|pa_{\mathcal{S}}(v)}(x^v \mid$

4. Analogous approximations often occur, for example, in statistical physics, when studying how macroscopic quantities (i.e. unit-level variables) depend on microscopic quantities (i.e. subunit-level variables) by considering the limit of infinite particles (i.e. subunits). E.g. rather than use the empirical mean of the squared velocity of the molecules in a gas, one can (in the thermodynamic limit) use the mean under the Maxwell-Boltzmann distribution.

$\text{pa}_S(v)$). The mapping from conditional distributions to joint distributions is onto,⁵ which implies,

$$\text{KL}(\text{p}^{\text{col}}(x^{\mathcal{U}}, q(x^{\mathcal{S}})) \| \text{p}_m(x^{\mathcal{U}}, q(x^{\mathcal{S}}))) \leq \text{KL}(\text{p}^{\text{col}}(x^{\mathcal{U}}, q^{\mathcal{Q}}) \| \text{p}_m(x^{\mathcal{U}}, q^{\mathcal{Q}})).$$

Our task is now to show that the right hand side converges to zero as $m \rightarrow \infty$.

We take the indices of the endogenous variables to be causally ordered, so that $\text{pa}(v) \subseteq \{1, \dots, v-1\}$ for all $v \in \{1, \dots, V\}$. Let $v_k^{\mathcal{S}}$ denote the index of the k th subunit-level variable, i.e. $v_k^{\mathcal{S}} \in \mathcal{S}$ and $v_k^{\mathcal{S}} < v_{k+1}^{\mathcal{S}}$ for all $k \in \{1, \dots, |\mathcal{S}|\}$. Set $v_0^{\mathcal{S}} = 0$. Let $\mathcal{S} > v$ denote the set of subunit variables after v in the causal ordering, that is $\{v' \in \mathcal{S} : v' > v\}$, and define $\mathcal{U} > v$ analogously, as well as $\mathcal{S} \leq v$ and $\mathcal{U} \leq v$. Let $q^{\mathcal{Q} > v}$ denote the set of Q variables that describe subunit variables after v , that is $q^{\mathcal{Q} > v} = \{q^{v' | \text{pa}_S(v')} : v' > v\}$.

Our bound will decompose in terms of conditional distributions in the HCGM, over variables after $v_k^{\mathcal{S}}$ in the causal ordering given variables up to $v_k^{\mathcal{S}}$. In particular, we use,

$$\begin{aligned} & \text{p}(x^{\mathcal{U} > v_k^{\mathcal{S}}}, q^{\mathcal{Q} > v_k^{\mathcal{S}}} \mid x^{\mathcal{U} \leq v_k^{\mathcal{S}}}, \{x_j^{\mathcal{S} \leq v_k^{\mathcal{S}}}\}_{j=1}^m) \\ &= \int \cdots \int \prod_{k'=k+1}^{|\mathcal{S}|} \left[\prod_{j=1}^m q^{v_{k'}^{\mathcal{S}} | \text{pa}_S(v_{k'}^{\mathcal{S}})}(x_j^{v_{k'}^{\mathcal{S}}} \mid x_j^{\text{pa}_S(v_{k'}^{\mathcal{S}})}) \right] \text{p}(q^{v_{k'}^{\mathcal{S}} | \text{pa}_S(v_{k'}^{\mathcal{S}})} \mid x^{\text{pa}_Z(v_{k'}^{\mathcal{S}})}) \\ & \quad \times \prod_{v'=v_{k'-1}^{\mathcal{S}}+1}^{v_{k'}^{\mathcal{S}}-1} \text{p}(x^{v'} \mid x^{\text{pa}_Z(v')}, \{x_j^{\text{pa}_S(v')}\}_{j=1}^m) dx_1^{\mathcal{S} > v_k^{\mathcal{S}}} \cdots dx_m^{\mathcal{S} > v_k^{\mathcal{S}}}. \end{aligned}$$

Note the full distribution of the HCGM over all Q variables and unit X^v variables, $\text{p}_m(x^{\mathcal{U}}, q^{\mathcal{Q}})$, corresponds to the case where $k = 0$. We compare to the collapsed model, which has the conditional,

$$\text{p}^{\text{col}}(x^{\mathcal{U} > v_k^{\mathcal{S}}}, q^{\mathcal{Q} > v_k^{\mathcal{S}}} \mid x^{\mathcal{U} \leq v_k^{\mathcal{S}}}, q^{\mathcal{Q} \leq v_k^{\mathcal{S}}}) = \prod_{k'=k+1}^{|\mathcal{S}|} \text{p}(q^{v_{k'}^{\mathcal{S}} | \text{pa}_S(v_{k'}^{\mathcal{S}})} \mid x^{\text{pa}_Z(v_{k'}^{\mathcal{S}})}) \prod_{v'=v_{k'-1}^{\mathcal{S}}+1}^{v_{k'}^{\mathcal{S}}-1} \text{p}(x^{v'} \mid x^{\text{pa}_Z(v')}, q(x^{\text{pa}_S(v')})).$$

The full distribution of the collapsed model over all Q variables and unit X^v variables, $\text{p}^{\text{col}}(q^{\mathcal{Q}}, x^{\mathcal{U}})$, corresponds to the case where $k = 0$.

5. The mapping is not one-to-one because, if a distribution $q(x)$ does not have full support, multiple different values of $q(y \mid x)$ will yield the same joint $q(x, y)$.

We now bound the KL by decomposing it into a sum over conditionals, and applying Jensen's inequality.

$$\begin{aligned}
 & \text{KL}(\mathbf{p}^{\text{col}}(x^{\mathcal{U}}, q^{\mathcal{Q}}) \| \mathbf{p}_m(x^{\mathcal{U}}, q^{\mathcal{Q}})) \\
 &= \mathbb{E}_{\mathbf{p}^{\text{col}}(x^{\mathcal{U} \leq v_1^{\mathcal{S}}}, q^{\mathcal{Q} \leq v_1^{\mathcal{S}}})} [\text{KL}(\mathbf{p}^{\text{col}}(x^{\mathcal{U} > v_1^{\mathcal{S}}}, q^{\mathcal{Q} > v_1^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_1^{\mathcal{S}}}, \mathcal{Q}^{v_1^{\mathcal{S}}}) \| \mathbb{E}_{\mathcal{Q}(x^{v_1^{\mathcal{S}}})} [\mathbf{p}(x^{\mathcal{U} > v_1^{\mathcal{S}}}, q^{\mathcal{Q} > v_1^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_1^{\mathcal{S}}}, \{X_j^{v_1^{\mathcal{S}}}\}_{j=1}^m)])] \\
 & \quad + \text{KL}(\mathbf{p}(q^{v_1^{\mathcal{S}}} \mid x^{\text{pa}_{\mathcal{Z}}(v_1^{\mathcal{S}})}) \| \mathbf{p}(q^{v_1^{\mathcal{S}}} \mid x^{\text{pa}_{\mathcal{Z}}(v_1^{\mathcal{S}})})) + \sum_{v'=1}^{v_1^{\mathcal{S}}-1} \text{KL}(\mathbf{p}(x^{v'} \mid x^{\text{pa}_{\mathcal{Z}}(v')}) \| \mathbf{p}(x^{v'} \mid x^{\text{pa}_{\mathcal{Z}}(v')})) \\
 & \leq \mathbb{E}_{\mathbf{p}^{\text{col}}(x^{\mathcal{U} \leq v_1^{\mathcal{S}}}, q^{\mathcal{Q} \leq v_1^{\mathcal{S}}})} [\mathbb{E}_{\mathcal{Q}(x^{v_1^{\mathcal{S}}})} [\text{KL}(\mathbf{p}^{\text{col}}(x^{\mathcal{U} > v_1^{\mathcal{S}}}, q^{\mathcal{Q} > v_1^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_1^{\mathcal{S}}}, \mathcal{Q}^{v_1^{\mathcal{S}}}) \| \mathbf{p}(x^{\mathcal{U} > v_1^{\mathcal{S}}}, q^{\mathcal{Q} > v_1^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_1^{\mathcal{S}}}, \{X_j^{v_1^{\mathcal{S}}}\}_{j=1}^m)])] \\
 &= \mathbb{E}_{\mathbf{p}^{\text{col}}(x^{\mathcal{U} \leq v_2^{\mathcal{S}}}, q^{\mathcal{Q} \leq v_2^{\mathcal{S}}})} [\mathbb{E}_{\mathcal{Q}(x^{v_1^{\mathcal{S}}})} [\text{KL}(\mathbf{p}^{\text{col}}(x^{\mathcal{U} > v_2^{\mathcal{S}}}, q^{\mathcal{Q} > v_2^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_2^{\mathcal{S}}}, \mathcal{Q}^{\mathcal{Q} \leq v_2^{\mathcal{S}}}) \| \\
 & \quad \mathbb{E}_{\mathcal{Q}^{v_2^{\mathcal{S}} | \text{pa}_{\mathcal{S}}(v_2^{\mathcal{S}})}(x^{v_2^{\mathcal{S}}} | X^{\text{pa}_{\mathcal{S}}(v_2^{\mathcal{S}})})} [\mathbf{p}(x^{\mathcal{U} > v_2^{\mathcal{S}}}, q^{\mathcal{Q} > v_2^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_2^{\mathcal{S}}}, \{X_j^{\mathcal{Q} \leq v_2^{\mathcal{S}}}\}_{j=1}^m)])] \\
 & \quad + \mathbb{E}_{\mathbf{p}^{\text{col}}(x^{\mathcal{U} \leq v_2^{\mathcal{S}}})} [\text{KL}(\mathbf{p}^{\text{col}}(q^{v_2^{\mathcal{S}} | \text{pa}_{\mathcal{S}}(v_2^{\mathcal{S}})} \mid X^{\text{pa}_{\mathcal{Z}}(v_2^{\mathcal{S}})}) \| \mathbf{p}(q^{v_2^{\mathcal{S}} | \text{pa}_{\mathcal{S}}(v_2^{\mathcal{S}})} \mid X^{\text{pa}_{\mathcal{Z}}(v_2^{\mathcal{S}})})) \\
 & \quad + \sum_{v'=v_1^{\mathcal{S}}+1}^{v_2^{\mathcal{S}}-1} \mathbb{E}_{\mathbf{p}^{\text{col}}(x^{\mathcal{U} \leq v'}, q^{\mathcal{Q} < v'})} [\mathbb{E}_{\mathcal{Q}(x^{\mathcal{S} < v'})} [\text{KL}(\mathbf{p}^{\text{col}}(x^{v'} \mid X^{\text{pa}_{\mathcal{Z}}(v')}, \mathcal{Q}(x^{\text{pa}_{\mathcal{S}}(v')})) \| \mathbf{p}(x^{v'} \mid X^{\text{pa}_{\mathcal{Z}}(v')}, \{X^{\text{pa}_{\mathcal{S}}(v')}\}_{j=1}^m))].
 \end{aligned}$$

We can recognize the terms in the final sum as $\mathbb{E}_{\mathbf{p}^{\text{col}}}[\mathbf{D}_m^{v'}(X^{\text{pa}_{\mathcal{Z}}(v')}, \mathcal{Q}(x^{\text{pa}_{\mathcal{S}}(v')}))]$. Since there are a finite number of unit variables $|\mathcal{U}|$, we can choose m large enough such that $\mathbb{E}_{\mathbf{p}^{\text{col}}}[\mathbf{D}_m^{v'}(X^{\text{pa}_{\mathcal{Z}}(v')}, \mathcal{Q}(x^{\text{pa}_{\mathcal{S}}(v')}))] < \epsilon$ for all $v' \in \mathcal{U}$. So, continuing the bound, we obtain,

$$\begin{aligned}
 & \leq \mathbb{E}_{\mathbf{p}^{\text{col}}(x^{\mathcal{U} \leq v_2^{\mathcal{S}}}, q^{\mathcal{Q} \leq v_2^{\mathcal{S}}})} [\mathbb{E}_{\mathcal{Q}(x^{\mathcal{S} \leq v_2^{\mathcal{S}}})} [\text{KL}(\mathbf{p}^{\text{col}}(x^{\mathcal{U} > v_2^{\mathcal{S}}}, q^{\mathcal{Q} > v_2^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_2^{\mathcal{S}}}, \mathcal{Q}^{\mathcal{Q} \leq v_2^{\mathcal{S}}}) \| \\
 & \quad \mathbf{p}(x^{\mathcal{U} > v_2^{\mathcal{S}}}, q^{\mathcal{Q} > v_2^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_2^{\mathcal{S}}}, \{X_j^{\mathcal{Q} \leq v_2^{\mathcal{S}}}\}_{j=1}^m)])] \\
 & \quad + \epsilon(v_2^{\mathcal{S}} - v_1^{\mathcal{S}} - 1) \\
 & \quad \dots \\
 & \leq \mathbb{E}_{\mathbf{p}^{\text{col}}(x^{\mathcal{U} \leq v_k^{\mathcal{S}}}, q^{\mathcal{Q} \leq v_k^{\mathcal{S}}})} [\mathbb{E}_{\mathcal{Q}(x^{\mathcal{S} \leq v_k^{\mathcal{S}}})} [\text{KL}(\mathbf{p}^{\text{col}}(x^{\mathcal{U} > v_k^{\mathcal{S}}}, q^{\mathcal{Q} > v_k^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_k^{\mathcal{S}}}, \mathcal{Q}^{\mathcal{Q} \leq v_k^{\mathcal{S}}}) \| \\
 & \quad \mathbf{p}(x^{\mathcal{U} > v_k^{\mathcal{S}}}, q^{\mathcal{Q} > v_k^{\mathcal{S}}} \mid X^{\mathcal{U} \leq v_k^{\mathcal{S}}}, \{X_j^{\mathcal{Q} \leq v_k^{\mathcal{S}}}\}_{j=1}^m)])] \\
 & \quad + \sum_{k'=1}^{k-1} \epsilon(v_{k'+1}^{\mathcal{S}} - v_{k'}^{\mathcal{S}} - 1) \\
 & \quad \dots \\
 & \leq |\mathcal{U}| \epsilon.
 \end{aligned}$$

Thus, the KL divergence between the collapsed model and the HCGM converges to zero as $m \rightarrow \infty$ a.s.. \blacksquare

Appendix G. Convergence of Hierarchical Empirical Distributions

In Section 5, we introduced the identification assumption that $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$ is known (Assumption 1). In this section we justify this assumption, by showing that with sufficient data we can infer $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$.

In Theorem 6 we showed that HCGMs converge to collapsed models, in the limit of infinite subunits. So in this limit, data from the HCGM can be modeled as (Definition 4),

$$\begin{aligned} X_i^{\mathcal{U}_{\text{obs}}}, Q_i &\sim p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}})) \\ X_{ij}^{\mathcal{S}_{\text{obs}}} &\sim Q_i(x^{\mathcal{S}_{\text{obs}}}). \end{aligned} \quad (61)$$

We assume now that we have a dataset $\{x_i^{\mathcal{U}_{\text{obs}}}, \{x_{ij}^{\mathcal{S}_{\text{obs}}}\}_{j=1}^M\}_{i=1}^N$, with N units and M subunits per unit, drawn from Equation (61). Here, we make a distinction here between the true underlying number of subunits in the model, m , and the number of subunits we actually observe, M . In other words, we assume that while in reality there are an effectively infinite number of subunits – which justifies the use of the collapsed model to describe the data (Theorem 6) – in practice we have a finite dataset. We similarly draw a distinction between the true and observed number of units, n versus N . We will show that as we gather more data, i.e. as $N, M \rightarrow \infty$, we can learn $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$.

In flat causal models, the data $\{x_i^{\text{obs}}\}_{i=1}^N$ is drawn as $X_i^{\text{obs}} \sim p(X^{\text{obs}})$, and identification is studied under the assumption that $p(x^{\text{obs}})$ is known. This assumption is motivated by the fact that as we gather more data, i.e. as $N \rightarrow \infty$, we can infer $p(x^{\text{obs}})$ arbitrarily well. More precisely, $p_N \rightarrow p$ a.s., where $p_N = \frac{1}{N} \delta_{x_i^{\text{obs}}}$ is the empirical distribution of the data (e.g. Dudley, 2002, Theorem 11.4.1). Here, $p_N \rightarrow p$ denotes weak convergence, though note that convergence holds in many other senses as well (for instance, the classic Glivenko–Cantelli theorem, the “fundamental theorem of statistics”, describes uniform convergence of the empirical c.d.f.).

In hierarchical causal models, the situation is somewhat different: we do not actually observe an empirical distribution of data from $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$. That is, $p_N = \frac{1}{N} \delta_{x_i^{\mathcal{U}_{\text{obs}}}, q_i}$ is unobserved, since the q_i are unobserved. Instead, we have access to the empirical distribution of empirical distributions $p_{N,M} = \frac{1}{N} \sum_{i=1}^N \delta_{x_i^{\mathcal{U}_{\text{obs}}}, q_{M,i}}$ where $q_{M,i} = \frac{1}{M} \sum_{j=1}^M \delta_{x_{ij}^{\mathcal{S}_{\text{obs}}}}$. We refer to $p_{N,M}$ as a *hierarchical* empirical distribution. In this section, we will show that $p_{N,M} \rightarrow p$ a.s., just like the non-hierarchical empirical distribution. So, despite the noise contributed by variation among subunits, we can still learn about the true distribution without making any parametric assumptions.

Both $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$ and $p_{N,M}$ are distributions over $\mathcal{X}^{\mathcal{U}_{\text{obs}}} \times \mathcal{P}(\mathcal{X}^{\mathcal{S}_{\text{obs}}})$, where $\mathcal{P}(\mathcal{X}^{\mathcal{S}_{\text{obs}}})$ is the set of distributions on $\mathcal{X}^{\mathcal{S}_{\text{obs}}}$. To establish weak convergence of the hierarchical empirical distribution, we will need a metric over this space of distributions. As in previous studies of the asymptotic behavior of hierarchical probabilistic models, we rely on the Wasserstein distance (Nguyen, 2016). In particular, we focus on the Wasserstein 1-distance, defined for two measures μ, ν on a separable and complete metric space (\mathcal{X}, d) as,

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{X, Y \sim \gamma} [d(X, Y)],$$

where $\Gamma(\mu, \nu)$ denotes the set of all couplings of μ, ν , i.e. the set of all joint distributions with marginals μ, ν .

We will use the Wasserstein distance not only to construct a metric on probability distributions but also a metric on probability distributions over probability distributions. We can do so using

Vershik's tower (Bogachev and Kolesnikov, 2012, Chap. 1.1). Let \mathcal{P}^1 be the set of all Borel probability measures on \mathcal{X} with finite first moment, i.e. $\mathbb{E}_{X \sim \mu}[d(X, x_0)] < \infty$ for all $\mu \in \mathcal{P}^1$ and an arbitrary $x_0 \in \mathcal{X}$.

Proposition 16 (Vershik's tower) *Let (\mathcal{X}, d) be complete, separable metric space. Then, $(\mathcal{P}^1(\mathcal{X}), W_1)$ is a complete and separable metric space, as is $(\mathcal{P}^1(\mathcal{P}^1(\mathcal{X})), W_{W_1})$, $(\mathcal{P}^1(\mathcal{P}^1(\mathcal{P}^1(\mathcal{X}))), W_{W_{W_1}})$, etc. Moreover, if (\mathcal{X}, d) is compact, so are all the other spaces.*

Vershik's tower is a useful tool for analyzing hierarchical probabilistic models, as it allows us to construct metrics on distributions with any level of hierarchy.

We now show that the hierarchical empirical distribution converges with increasing data.

Proposition 17 (Hierarchical empirical distributions converge) *Let $(\mathcal{X}^{\mathcal{S}_{\text{obs}}}, d^{\mathcal{S}_{\text{obs}}})$ and $(\mathcal{X}^{\mathcal{U}_{\text{obs}}}, d^{\mathcal{U}_{\text{obs}}})$ be compact, separable metric spaces. Assume $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}})) \in \mathcal{P}^1(\mathcal{X}^{\mathcal{U}_{\text{obs}}} \times \mathcal{P}^1(\mathcal{X}^{\mathcal{S}_{\text{obs}}}))$, where $\mathcal{X}^{\mathcal{U}_{\text{obs}}} \times \mathcal{P}^1(\mathcal{X}^{\mathcal{S}_{\text{obs}}})$ has metric $d^{\mathcal{U}_{\text{obs}}} + W_{1, d^{\mathcal{S}_{\text{obs}}}}$. In the limit as $M \rightarrow \infty$ and then $N \rightarrow \infty$, $p_{N, M}$ converges weakly to $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$ a.s..*

Proof Since $(\mathcal{X}^{\mathcal{S}_{\text{obs}}}, d_{\mathcal{S}_{\text{obs}}})$ is compact and separable, so is $(\mathcal{P}^1(\mathcal{X}^{\mathcal{S}_{\text{obs}}}), W_{1, d_{\mathcal{S}_{\text{obs}}}})$, and since $(\mathcal{X}^{\mathcal{U}_{\text{obs}}}, d_{\mathcal{U}_{\text{obs}}})$ is also compact and separable, so is the product space $(\mathcal{X}^{\mathcal{U}_{\text{obs}}} \times \mathcal{P}^1(\mathcal{X}^{\mathcal{S}_{\text{obs}}}), d^{\mathcal{U}_{\text{obs}}} + W_{1, d^{\mathcal{S}_{\text{obs}}}})$. This implies that the Wasserstein distance over this last space is well-defined.

Using the triangle inequality, and the explicit form of the Wasserstein distance for empirical distributions, we can bound the distance between the hierarchical empirical distribution and the true distribution as,

$$W_1(p_{N, M}, p) \leq W_1(p_{N, M}, p_N) + W_1(p_N, p) \leq \frac{1}{N} \sum_{i=1}^N d(x_i^{\mathcal{U}_{\text{obs}}}, x_i^{\mathcal{U}_{\text{obs}}}) + W_1(q_{M, i}, q_i) + W_1(p_N, p) \quad (62)$$

$$= \frac{1}{N} \sum_{i=1}^N W_1(q_{M, i}, q_i) + W_1(p_N, p). \quad (63)$$

Each term in the final expression compares a (non-hierarchical) empirical distribution to the distribution it is sampled from. For any Borel measure μ over a compact and separable metric space, the empirical distribution μ_N satisfies $W_1(\mu_N, \mu) \rightarrow 0$ a.s. (e.g. Weed and Bach, 2019). Therefore each term of Equation (63) converges and we have $W_1(p_{N, M}, p) \rightarrow 0$ a.s. when $M \rightarrow \infty$ and then $N \rightarrow \infty$.

Since the Wasserstein 1-distance metrizes weak convergence (e.g. Villani, 2008, Theorem 6.9), the conclusion follows. \blacksquare

This result tells us that with sufficient data, it is possible to learn $p(x^{\mathcal{U}_{\text{obs}}}, q(x^{\mathcal{S}_{\text{obs}}}))$, without making any parametric assumptions about this distribution.

Convergence rate We can also obtain a rate of convergence in N and M . Weed and Bach (2019) show that the Wasserstein distance between a distribution μ and its empirical distribution of samples μ_N satisfies $\mathbb{E}[W_1(\mu, \mu_N)] = O(n^{-1/d_\mu})$ where d_μ is a constant describing the effective dimension of the distribution. (If μ is a distribution on \mathbb{R}^d then $d_\mu \leq d$.) From Equation (63) we can obtain the convergence rate,

$$\mathbb{E}[W_1(p_{N, M}, p)] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[W_1(q_{M, i}, q_i)] + \mathbb{E}[W_1(p_N, p)] = O(M^{-1/d_q} + N^{-1/d_p}) \quad (64)$$

where d_q is the maximum effective dimension of q (under p), and d_p is the effective dimension of p . See Weed and Bach (2019) for details on the effective dimension.

Another way of quantifying convergence is to look at the expected value of a test function, e.g. how quickly $\frac{1}{N} \sum_{i=1}^N g(x_i)$ converges to $\mathbb{E}_p [g(X)]$. For hierarchical distributions, this can be generalized to examine the mismatch between,

$$A \triangleq \mathbb{E}_p \left[g(X^{\mathcal{U}_{\text{obs}}}, \mathbb{E}_Q [h(X^{\mathcal{S}_{\text{obs}}}, X^{\mathcal{U}_{\text{obs}}})]) \right] \quad (65)$$

$$A_{N,M} \triangleq \frac{1}{N} \sum_{i=1}^N g(x_i^{\mathcal{U}_{\text{obs}}}, \frac{1}{M} \sum_{j=1}^M h(x_i^{\mathcal{U}_{\text{obs}}}, x_{ij}^{\mathcal{S}_{\text{obs}}})) \quad (66)$$

for any test functions g and h . Under regularity conditions—when g and h are Lipschitz and output finite-dimension vectors, and $A_{N,M}$ has finite variance—Rainforth et al. (2018, Theorem 1) implies,

$$\|A - A_{N,M}\|^2 = O(1/N + 1/M). \quad (67)$$

Appendix H. Positivity for Do-Calculus

Do-calculus rests on positivity assumptions, which ensure the post-intervention distribution can be computed from the pre-intervention distribution. It is common, for the sake of simplicity, to assume that the joint distribution over all the endogenous variables is strictly positive. However, this assumption is stronger than necessary, and indeed can block identification in some HCMs, such as in the SUBUNIT INSTRUMENT model (Appendix I). We therefore employ the weaker positivity assumptions developed in Shpitser and Pearl (2006).

The first of these assumptions is that the intervention has positive probability (Assumption 3). This assumption ensures that the intervention we are considering is well-defined. The second positivity assumption stems from the fact that do-calculus often provides identification formulae with terms of the form,

$$\int p(y \mid x) \tilde{p}(x) dx, \quad (68)$$

where $\tilde{p}(x)$ denotes a non-observational distribution over x (here, x and y may each represent one or more endogenous variables). For example, if we are performing a hard intervention on x we could have $\tilde{p}(x) = \delta_{x^*}(x)$ (or see e.g. Equation (27) for another example). To compute Equation (68), we need to be able to estimate $p(y \mid x)$ for all values of x on which $\tilde{p}(x)$ has support, using observational data.

Assumption 4 (Unit-level positivity, part two (Shpitser and Pearl, 2006)) *For each term of the form Equation (68) that appears in the identification formula provided by do-calculus, we require that $p(x)$ is positive wherever $\tilde{p}(x)$ is positive, i.e. $\tilde{p}(x) \ll p(x)$, where \ll denotes absolute continuity.*

There is one additional subtlety: while do-calculus is typically studied under the simplifying assumption that the variables are discrete, in collapsed models the Q variables are necessarily non-discrete, even when all the endogenous variables of the HCGM are discrete. The above positivity assumption only applies to the discrete variable case, since in the continuous case, when $\tilde{p}(x)$ involves a delta function, we do not have $\tilde{p}(x) \ll p(x)$ in general. However, Assumption 4 can be relaxed with some technical regularity assumptions. Here we give a relaxation that is general enough to apply to the distribution-valued endogenous variables that appear in collapsed models. Intuitively,

this assumption extends Assumption 4 to consider positivity in a neighborhood of the intervention. Let \mathcal{X} and \mathcal{Y} denote the domains of x and y , and let $\mathcal{P}(\mathcal{Y})$ denote the set of distributions over \mathcal{Y} .

Assumption 5 (Positivity for general variables) *Consider each term in the identification formula of the form of Equation (68). Assume there exists a known sequence of distributions ν_1, ν_2, \dots converging weakly to $\tilde{p}(x)$ such that $\nu_k(x) \ll p(x)$ for all k . Assume for all measurable $\mathcal{A} \subseteq \mathcal{Y}$, $\Pr(y \in \mathcal{A} \mid x)$ is a continuous function of x .*

Proposition 18 *Under Assumption 5, and given that do-calculus identifies $\tilde{p}(x)$, Equation (68) is identified.*

Proof From the observational distribution $p(x, y)$, we can compute the expected value $\mathbb{E}_{X \sim \nu_k}[\Pr(y \in \mathcal{A} \mid X)]$ using the importance sampling formula, $\mathbb{E}_{X \sim \nu_k}[\Pr(y \in \mathcal{A} \mid X)] = \mathbb{E}_{X \sim p(x)}\left[\Pr(y \in \mathcal{A} \mid X) \frac{\nu_k(X)}{p(X)}\right]$, since $\nu_k(x) \ll p(x)$. Since $\Pr(y \in \mathcal{A} \mid x)$ is continuous and bounded, and ν_k converges weakly to $\tilde{p}(x)$, we have $\mathbb{E}_{X \sim \nu_k}[\Pr(y \in \mathcal{A} \mid X)] \rightarrow \mathbb{E}_{X \sim \tilde{p}(x)}[\Pr(y \in \mathcal{A} \mid X)]$. Since each term $\mathbb{E}_{X \sim \nu_k}[\Pr(y \in \mathcal{A} \mid X)]$ is identified, and since the sequence converges, we can identify $\int \Pr(y \in \mathcal{A} \mid x) \tilde{p}(x) dx$ as its limit. Since this holds for any \mathcal{A} , we can identify the entire distribution $\int p(y \mid x) \tilde{p}(x) dx$ ■

Appendix I. Instrumental Variable Assumptions

Here we discuss further the assumptions needed for identification in the SUBUNIT INSTRUMENT graph (Section 5.3), and explain how they relate to the identification assumptions used in conventional, flat instrumental variable models.

The key assumptions for applying do-calculus and achieving identification in the SUBUNIT INSTRUMENT graph are the positivity assumptions, Assumption 2 and Assumption 3. First, there must always be within-unit variation in the instrument, i.e. we must have $Q^z(z) > 0$ a.s. for $Q^z \sim p(q^z)$ (Assumption 2). Second, there must always be a non-zero probability of $q^a = q_\star^a$ given $q^{a|z}$, that is, $p(q^a = q_\star^a \mid q^{a|z}) > 0$ a.s. for $Q^{a|z} \sim p(q^{a|z})$ (Assumption 3). Said another way:

Assumption 6 (Unit-level positivity for the SUBUNIT INSTRUMENT graph) *For $Q^{a|z} \sim p(q^{a|z})$, there must exist a.s. a solution q^z to the integral equation*

$$\int q^{a|z}(a \mid z) q^z(z) dz = q_\star^a(a), \tag{69}$$

such that $p(q^z) > 0$.

In brief, for any value of $q^{a|z}$, there must exist some value of q^z that produces the marginal $q^a = q_\star^a$.

The positivity requirements in the hierarchical SUBUNIT INSTRUMENT model are related to the relevance and completeness assumptions that appear in flat instrumental variable models (Newey and Powell, 2003). Intuitively, in both the hierarchical and flat settings we need the instrument to (1) vary and (2) actually affect the treatment, i.e. it cannot be a completely unrelated quantity. To see this in

the hierarchical case, note that if $q^{a|z}(a | z)$ were constant with respect to z , then $\int q^{a|z}(a | z)q^z(z)dz$ would be constant with respect to q^z . This will in general violate Assumption 6.⁶

Despite these similarities, the positivity assumptions in the hierarchical instrumental variable model are distinct from the assumptions made in flat instrumental variable models. To see this, we compare to the completeness assumption that is widely used in flat nonparametric instrumental variable models (Newey and Powell, 2003). In a flat instrumental variable model with Q^z as the instrument and Q^a as the treatment, the completeness assumption can be stated as: There must exist a unique solution $g(\cdot)$ to the equation,

$$\int p(q^a | q^z)g(q^a)dq^a = \mathbb{E}_p[Y | q^z]. \quad (70)$$

While both Equation (69) and Equation (70) are Fredholm integral equations of the first kind, where the kernel describes a conditional distribution over A given Z , they are otherwise quite distinct (Carrasco et al., 2007). For example, Equation (69) involves a within-unit conditional distribution of A given Z , whereas Equation (70) involves a between-unit conditional distribution. Moreover, the solution to Equation (69) just needs to exist, whereas in Equation (70) the solution must be unique.

An important advantage of the hierarchical instrument variable model is that identification does not require any assumptions on the causal mechanism generating the outcome. In particular, the mechanism generating Y from Q^a does not need to have additive noise or to be monotonic with respect to the noise, as is required in the flat instrumental variable setting (Imbens and Newey, 2002; Newey and Powell, 2003; Saengkyongam et al., 2022). Moreover, in the hierarchical model we can identify the entire post-intervention distribution $p(y; \text{do}(q^a = q_*^a))$, whereas in the flat model we can only identify the mean $\mathbb{E}_p[Y; \text{do}(q^a = q_*^a)]$. This is especially relevant for problems with structured outcome variables, for example if Y is a text, graph, or molecule. In these cases, additive noise may be ill-defined, limiting the application of flat instrumental variable methods.

Appendix J. Details on Augmentation

In this section we provide further details on our augmentation approach, and prove Proposition 10.

J.1 General form of augmentation variables

We employ augmentation variables of a particular form, namely those generated as,

$$q_i^{\mathcal{L}|\mathcal{R}}(x^{\mathcal{L}}; \text{do}(x^{\mathcal{R}})) = \int \cdots \int \prod_{v \in \mathcal{L} \cup \text{da}_{\mathcal{S}}(\mathcal{L}) \setminus \mathcal{R}} q_i^{v|\text{pa}_{\mathcal{S}}(v)}(x^v | x^{\text{pa}_{\mathcal{S}}(v)}) \prod_{w \in \text{da}_{\mathcal{S}}(\mathcal{L}) \setminus \mathcal{R}} dx^w, \quad (71)$$

where $\mathcal{L} \subseteq \mathcal{S}$ is a set of subunit-level variables and $\mathcal{R} \subseteq \text{da}_{\mathcal{S}}(\mathcal{L}) = (\bigcup_{v \in \mathcal{L}} \text{da}_{\mathcal{S}}(v)) \setminus \mathcal{L}$ is a set of subunit direct ancestors of \mathcal{L} . Equation (71) describes the within-unit distribution over $X^{\mathcal{L}}$ after an intervention on $X^{\mathcal{R}}$, holding fixed the unit variables. In other words, it is the interventional effect derived from the subunit variable graph, with the unit variables and outer plate ignored. For example,

6. One implication is that, to achieve identification, we must have $p(q^{a|z} = \tilde{q}(a)) = 0$ for all distributions $\tilde{q}(a)$ that are constant with respect to z . This conflicts with a common positivity assumption made in do-calculus, that the observational distribution is positive everywhere. Hence, we employ the weaker do-calculus positivity assumptions proposed by Shpitser and Pearl (2006), and described in Appendix H.

in Figure A3k, when we erase the unit variables and outer plate we obtain a graph $Z \rightarrow A \leftarrow X$, so the augmentation variable $q_i^{a|x}$ is given by $q_i^{a|x}(a; \text{do}(x)) = \int q_i^{a|z,x}(a | z, x) q_i^z(z) dz$.

J.2 Proof of Proposition 10

Fundamentally, the purpose of Proposition 10 is to allow us to take advantage of constraints in the mechanisms of collapsed models, in order to establish identification. Even when we place no parametric restrictions on mechanisms in a hierarchical causal model, they appear in the collapsed model. For example, in the SUBUNIT INSTRUMENT HCGM, we did not constrain the mechanism generating Y , but we found that in the collapsed model Y can only depend on its parents $Q^{a|z}$ and Q^z through their marginal $\int Q^{a|z}(a | z) Q^z(z) dz$ (Equation (30), Section 5.3). Do-calculus operates under the assumption that there are no parametric constraints on the mechanisms in a causal graphical model. So, when we apply do-calculus directly to the collapsed model, we cannot take advantage of the model's constraints to prove identification. Augmentation, together with Proposition 10, allows us to use these constraints effectively.

Proof Note the augmentation variable $Q^{v|\mathcal{R}}$ follows Equation (71) with $\mathcal{L} = \{v\}$. In this proof, we use \mathcal{L} in place of v to make clear we are discussing the intervened variable rather than a generic variable.

We will show that $p^{\text{col}}(y; \text{do}(q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} = q_{\star}^{\mathcal{L}|\mathcal{R}})) = p^{\text{aug}}(y; \text{do}(q^{\mathcal{L}|\mathcal{R}} = q_{\star}^{\mathcal{L}|\mathcal{R}}))$. This implies the result $p^{\text{col}}(y; \text{do}(q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} = q_{\star}^{\mathcal{L}|\mathcal{R}})) = p^{\text{mar}}(y; \text{do}(q^{\mathcal{L}|\mathcal{R}} = q_{\star}^{\mathcal{L}|\mathcal{R}}))$, since marginalizing out a variable from the model cannot change the effect. Note that, in the augmented model, all directed paths from $Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}$ to Y must go through $Q^{\mathcal{L}|\mathcal{R}}$.

We use $\text{an}(Z)$ to denote the ancestors of a variable Z , with $\text{an}^{\text{col}}(Z)$ and $\text{an}^{\text{aug}}(Z)$ denoting the ancestors in the collapsed and augmented models respectively. We use $\text{ch}(Z)$ to denote children, with $\text{ch}^{\text{col}}(Z)$ and $\text{ch}^{\text{aug}}(Z)$ defined analogously. We use $\overline{\text{an}}(Z)$ to denote the ancestors of Z inclusive of Z , that is $\overline{\text{an}}(Z) = \{Z\} \cup \text{an}(Z)$, and likewise for $\overline{\text{ch}}(Z)$. Finally, we use Z^v to denote a generic endogenous variable in the collapsed or augmented model; it can be either a unit variable X^v or a Q variable.

We can write the effect in the collapsed model as,

$$\begin{aligned} & p^{\text{col}}(y; \text{do}(q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} = q_{\star}^{\mathcal{L}|\mathcal{R}})) \\ &= \int \cdots \int \prod_{v \in \overline{\text{an}}^{\text{col}}(Y) \setminus \overline{\text{ch}}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})} p^{\text{col}}(z^v | z^{\text{pa}^{\text{col}}(v)}) \\ & \quad \times \prod_{v \in \overline{\text{ch}}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})} p^{\text{col}}(z^v | q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} = q_{\star}^{\mathcal{L}|\mathcal{R}}, z^{\text{pa}^{\text{col}}(v)} \setminus q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}) \prod_{v \in \overline{\text{an}}^{\text{col}}(Y) \setminus \overline{\text{ch}}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})} dz^v. \end{aligned}$$

We will analyze the terms in the first and second product separately, equating them to terms in the augmented model. Consider first the terms in the first product, which describe the mechanisms generating Z^v for $v \in \overline{\text{an}}^{\text{col}}(Y) \setminus \overline{\text{ch}}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})$. We will argue that these mechanisms are the same in the augmented model, i.e. $p^{\text{col}}(z^v | z^{\text{pa}^{\text{col}}(v)}) = p^{\text{aug}}(z^v | z^{\text{pa}^{\text{aug}}(v)})$. Since the augmentation is valid (Definition 8), marginalizing out $Q^{\mathcal{L}|\mathcal{R}}$ from the augmented model must recover the collapsed model. So, the only way for a mechanism to differ in the augmented model is if Z^v is a child of $Q^{\mathcal{L}|\mathcal{R}}$ in the augmented model. However, in that case, marginalizing out $Q^{\mathcal{L}|\mathcal{R}}$ would make Z^v a child of $Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}$ in the collapsed model. This violates the condition that $v \in \overline{\text{an}}^{\text{col}}(Y) \setminus \overline{\text{ch}}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})$.

Next, consider the terms in the second product, which describe the mechanisms generating Z^v for $v \in \text{ch}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})$. Since all paths from $Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}$ to Y go through $Q^{\mathcal{L}|\mathcal{R}}$ in the augmented model, to satisfy validity (Definition 8) each of these Z^v must be a child of $Q^{\mathcal{L}|\mathcal{R}}$ in the augmented model, and not a child of $Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}$. Moreover, validity further implies that,

$$\begin{aligned} \text{p}^{\text{col}}(z^v | q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}) &= q_{\star}^{\mathcal{L}|\mathcal{R}}, z^{\text{pa}^{\text{col}}(v)} \setminus q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} \\ &= \text{p}^{\text{aug}}(z^v | q^{\mathcal{L}|\mathcal{R}} = f^{\mathcal{L}|\mathcal{R}}(q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}) = q_{\star}^{\mathcal{L}|\mathcal{R}}, z^{\text{pa}^{\text{aug}}(Q^{\mathcal{L}|\mathcal{R}})} \setminus q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}, z^{\text{pa}^{\text{aug}}(v)} \setminus q^{\mathcal{L}|\mathcal{R}}), \end{aligned}$$

where $f^{\mathcal{L}|\mathcal{R}}$ is the mechanism generating the augmentation variable (Equation (71)). Examining Equation (71), we can see that if $q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} = q_{\star}^{\mathcal{L}|\mathcal{R}}$, then it must be the case $q^{\mathcal{L}|\mathcal{R}} = q_{\star}^{\mathcal{L}|\mathcal{R}}$, regardless of the value of the other parents of $Q^{\mathcal{L}|\mathcal{R}}$ in the augmented model. Hence,

$$\text{p}^{\text{col}}(z^v | q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} = q_{\star}^{\mathcal{L}|\mathcal{R}}, z^{\text{pa}^{\text{col}}(v)} \setminus q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}) = \text{p}^{\text{aug}}(z^v | q^{\mathcal{L}|\mathcal{R}} = q_{\star}^{\mathcal{L}|\mathcal{R}}, z^{\text{pa}^{\text{aug}}(v)} \setminus q^{\mathcal{L}|\mathcal{R}}). \quad (72)$$

Now, again using the assumption that all paths from $Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}$ to Y go through $Q^{\mathcal{L}|\mathcal{R}}$, we can rewrite the effect in the collapsed model as,

$$\begin{aligned} \text{p}^{\text{col}}(y; \text{do}(q^{\mathcal{L}|\text{pa}_S(\mathcal{L})} = q_{\star}^{\mathcal{L}|\mathcal{R}})) &= \int \cdots \int \prod_{v \in \overline{\text{an}}^{\text{col}}(Y) \setminus \overline{\text{ch}}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})} \text{p}^{\text{aug}}(z^v | z^{\text{pa}^{\text{aug}}(v)}) \\ &\quad \times \prod_{v \in \text{ch}^{\text{col}}(Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})})} \text{p}^{\text{aug}}(z^v | q^{\mathcal{L}|\mathcal{R}} = q_{\star}^{\mathcal{L}|\mathcal{R}}, z^{\text{pa}^{\text{aug}}(v)} \setminus q^{\mathcal{L}|\mathcal{R}}) \prod_{v \in \overline{\text{an}}^{\text{col}}(Y) \setminus Q^{\mathcal{L}|\text{pa}_S(\mathcal{L})}} dz^v \\ &= \int \cdots \int \prod_{v \in \overline{\text{an}}^{\text{aug}}(Y) \setminus \overline{\text{ch}}^{\text{aug}}(Q^{\mathcal{L}|\mathcal{R}})} \text{p}^{\text{aug}}(z^v | z^{\text{pa}^{\text{aug}}(v)}) \\ &\quad \times \prod_{v \in \text{ch}^{\text{aug}}(Q^{\mathcal{L}|\mathcal{R}})} \text{p}^{\text{aug}}(z^v | q^{\mathcal{L}|\mathcal{R}} = q_{\star}^{\mathcal{L}|\mathcal{R}}, z^{\text{pa}^{\text{aug}}(v)} \setminus q^{\mathcal{L}|\mathcal{R}}) \prod_{v \in \overline{\text{an}}^{\text{aug}}(Y) \setminus Q^{\mathcal{L}|\mathcal{R}}} dz^v \\ &= \text{p}^{\text{aug}}(y; \text{do}(q^{\mathcal{L}|\mathcal{R}} = q_{\star}^{\mathcal{L}|\mathcal{R}})). \end{aligned}$$

■

Appendix K. Proofs of General Identification Conditions

K.1 Proof of Theorem 12

Proof Condition 1 In the collapsed model, the only children of $Q^{a|\text{pa}_S(a)}$ are its direct unit descendants (Definition 4). By assumption, there is no bi-directed path to these variables. If Y is subunit-level, we can augment with Q^y to identify the effect of interest, but there will again be no bi-directed path from $Q^{a|\text{pa}_S(a)}$ to Q^y . So from Theorem 11, the effect of $\text{do}(q^{a|\text{pa}_S(a)} = q_{\star}^a)$ on Y is identified.

Condition 2 The general setup is shown in Figure A8a, with Z denoting the subunit instrument. In the collapsed model we have the variables $Q^{a|\text{pa}_S(a)}$ and Q^z (Figure A8a). Now, the children of $Q^{a|\text{pa}_S(a)}$ in the collapsed model are the direct unit descendants of A in the original HCGM. Since A is the only child of Z in the HCGM, the direct unit descendants of A can only depend on $Q^{a|\text{pa}_S(a)}$

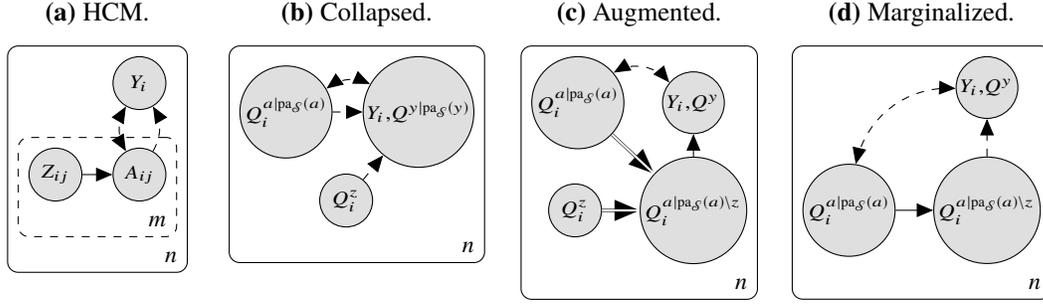


Figure A8: Steps of the proof of Theorem 12, condition 2. (a) The original HCM. Here Y_i is shown as unit-level, but it also may be subunit-level. We use a double-sided dashed arrow to denote the possibility of subunit or unit-level parents of A , Y , or both (including possibly hidden unit confounders). We use a one-sided dashed arrow to denote the possibility of directed paths from A to Y , which may be direct or run through subunit or unit-level variables. (b) Collapsed model. If Y is unit-level in the HCGM, rightmost node is also Y , but if it is subunit-level in the original HCGM, the rightmost node is $Q^{y|pa_S(y)}$. (c) Augmented model. (d) Marginalized model.

through terms of the form $\int Q^{a|pa_S(a)}(a | z, x^{pa_S(a)\setminus z})q^z(z)dz$ in the collapsed model (to see this, consider collapsing the original HCGM but with Z marginalized out). If Y is subunit level, we can augment the collapsed model with Q^y . Again, since A is the only child of Z , Q^y must depend on $Q^{a|pa_S(a)}$ only through terms of the form $\int Q^{a|pa_S(a)}(a | z, x^{pa_S(a)\setminus z})q^z(z)dz$.

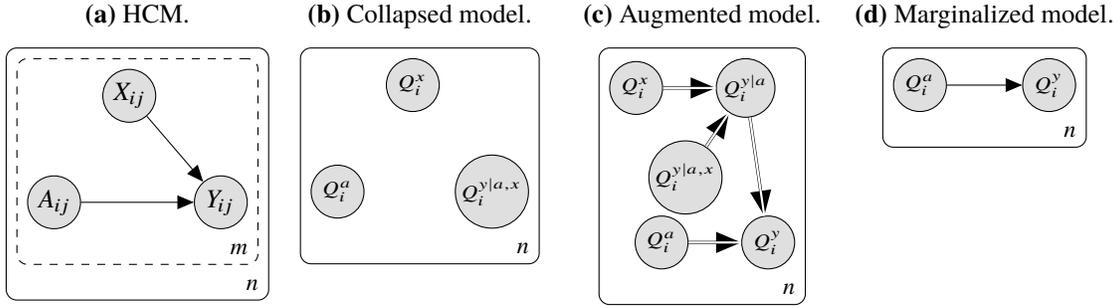
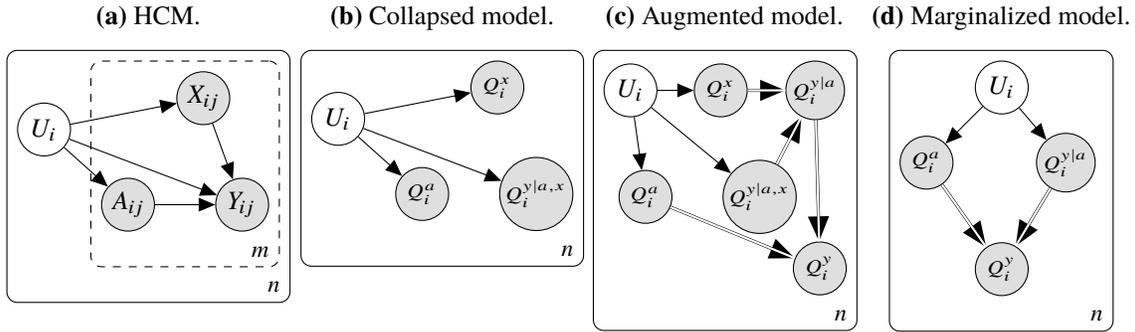
We can therefore augment the collapsed model with $Q^{a|pa_S(a)\setminus z}$ (Equation (71)) such that $Q^{a|pa_S(a)\setminus z}$ is the only child of $Q^{a|pa_S(a)}$ (Figure A8c). Moreover, the only child of Q^z will also be the augmentation variable $Q^{a|pa_S(a)\setminus z}$. Hence, we marginalize out Q^z such that the positivity assumption $p^{\text{mar}}(q^{a|pa_S(a)\setminus z} = q_\star^a | q^{a|pa_S(a)})$ can be met (Assumption 3; Figure A8d). From Proposition 10 the effect of interest is equivalent to $p^{\text{mar}}(y; \text{do}(q^{a|pa_S(a)\setminus z} = q_\star^a))$ if Y is unit-level, and to $p^{\text{mar}}(q^y; \text{do}(q^{a|pa_S(a)\setminus z} = q_\star^a))$ if Y is subunit-level.

Since there is no bi-directed path in the original HCGM between A and Z , there is no bi-directed path in the collapsed model between $Q^{a|pa_S(a)}$ and Q^z , and thus no bi-directed path in the marginalized model between $Q^{a|pa_S(a)}$ and its only child $Q^{a|pa_S(a)\setminus z}$. So from Theorem 11, the effect on Y is identified. \blacksquare

K.2 Proof of Theorem 13

Proof If the effect is not identified in the erased inner plate model, there must be a bi-directed path between A and at least one of its children (Theorem 11). Since the graph of the erased plate model is the same as that of the HCGM, there must also be a bi-directed path between A and at least one of its children in the HCGM.

Now, consider a modified HCGM in which all the outgoing arrows from subunit-level variables are erased. This is a special case of the original HCGM, so if the effect of interest is not identified in this modified model, it cannot be identified in the original model. There remains a bi-directed path between A and at least one of its children, since by assumption there are no subunit-level confounders, and hence all bi-directed paths must go through unit-level confounders.


Figure A9: Initial model for eight schools data.

Figure A10: Model for eight schools data with unobserved confounding.

We now collapse the HCGM. From Definition 4, there must remain a bi-directed path between A and a child in the collapsed model. Moreover, since there were no outgoing arrows from subunit variables, there are no constraints on the mechanisms in the collapsed model (i.e. it is fully non-parametric). Hence, from Theorem 11 the effect is not identified in the collapsed model. Since the collapsed model is equivalent to the original HCGM (Theorem 6) the result follows. ■

Appendix L. Details on Eight Schools

L.1 Fully observed and confounder models

In this section we provide further details on the estimation methods used for the initial analysis of the eight schools data, based on the fully observed (Section 7.1) and hidden unit confounder (Section 7.2) models.

The eight schools study does not make public its data at the student level. However, for each school i , the authors ran a linear regression predicting Y_{ij} from A_{ij} and X_{ij} , and reported the estimated coefficient on the treatment, $\hat{\mu}_i$. We can understand this per-school treatment effect $\hat{\mu}_i$ as a parametric estimate of $\mu_i = \mathbb{E}_{q_i^{y|a}}[Y | A = 1] - \mathbb{E}_{q_i^{y|a}}[Y | A = 0]$. Since there are only a finite number of students m_i per school, the estimate $\hat{\mu}_i$ comes with some uncertainty; (Alderman and Powers, 1979) report the standard error, σ_i . We can model $\hat{\mu}_i$ as a sample from $\text{Normal}(\mu_i, \sigma_i)$. To obtain the average treatment effect, we also need to estimate $p(q^{y|a})$, or, at minimum, its marginal $p(\mu)$. A simple parametric approach is to assume that $p(\mu)$ takes the form of a normal distribution,

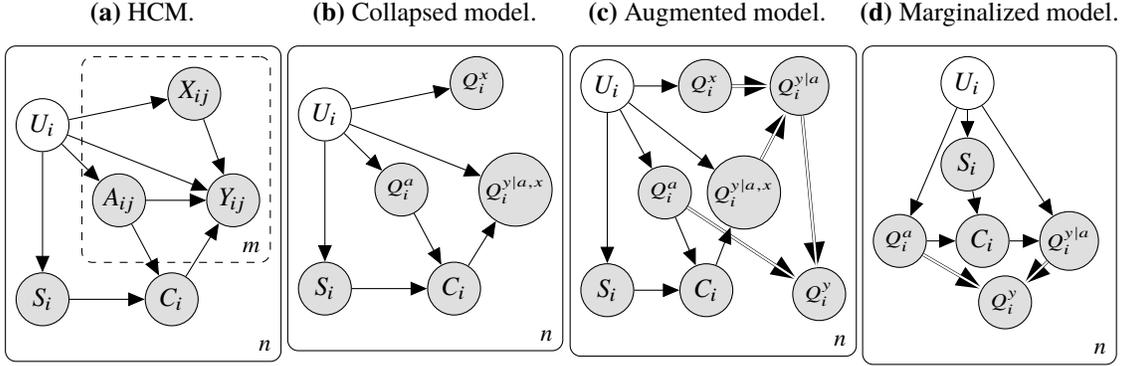


Figure A11: Model for eight schools data with unobserved confounding and interference.

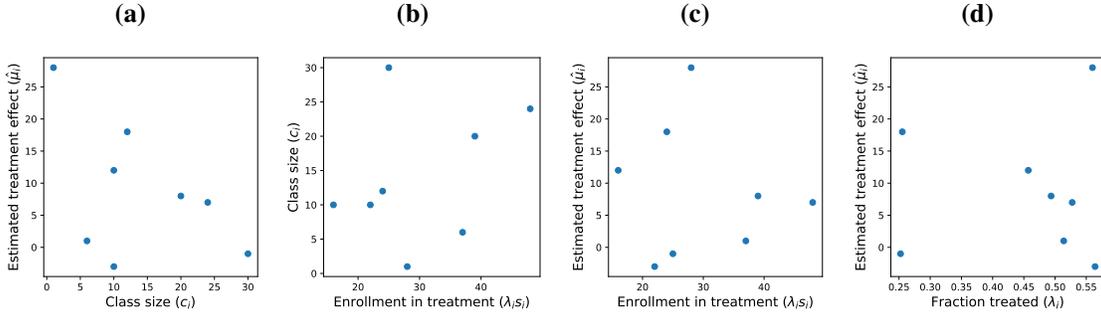


Figure A12: Scatter plots of eight schools data. Each point represents a school.

with unknown mean ν and standard deviation τ . The mean of this distribution is then the treatment effect we are interested in, $\nu = \mathbb{E}_p[\mu] = \mathbb{E}_p[\mathbb{E}_{Q^{y|a}}[Y | A = 1] - \mathbb{E}_{Q^{y|a}}[Y | A = 0]] = \text{ATE}$. Placing a diffuse prior on ν and τ , we obtain a hierarchical Bayesian model,

$$\begin{aligned} \nu &\sim \text{Normal}(0, 5) & \tau &\sim \text{HalfCauchy}(5) \\ \mu_i &\sim \text{Normal}(\nu, \tau) & \hat{\mu}_i &\sim \text{Normal}(\mu_i, \sigma_i), \end{aligned}$$

where HalfCauchy is the half Cauchy distribution with support on only positive values. We compute the posterior over ν , the estimate of the ATE, using MCMC (as described in Section 7.1). We have thus recovered, from a hierarchical causal model, the classic eight schools hierarchical Bayesian analysis.

L.2 Confounding & interference

In this section we provide further details on the model used for the analysis of the eight schools data that accounts for confounding and interference (Section 7.3).

In constructing the model, we treat the number of subunits as a separate unit variable, following the strategy described in Appendix E.2. In particular, the number of students interested in the tutoring program, S_i , corresponds to the sum of the number of students in the treatment and control groups for school i , that is m_i . (Note also that here we can have different numbers of subunits per unit.)

In our model, we assume that class size only impacts the average test scores of the treated students, such that $\mathbb{E}_p[\mathbb{E}_Q[Y | A = 0]; \text{do}(q^a = \delta_1)] = \mathbb{E}_p[\mathbb{E}_Q[Y | A = 0]; \text{do}(q^a = \delta_0)]$. Then, using Equation (39),

$$\begin{aligned} \text{ATE} &= \mathbb{E}_p[\mathbb{E}_{Q^{y|a}}[Y | A = 1] - \mathbb{E}_{Q^{y|a}}[Y | A = 0]; \text{do}(q^a = \delta_1)] = \mathbb{E}_p[\mu; \text{do}(q^a = \delta_1)] \\ &= \int \int p(c | q^a = \delta_1, s) p(s) ds \int \mathbb{E}_p[\mu | q^a, \tilde{s}, c] p(q^a, \tilde{s}) dq^a d\tilde{s} dc. \end{aligned} \quad (73)$$

So, this assumption allows us to make use of the summary results reported by Alderman and Powers (1979): we do not need information about the actual value of the test scores of the treated and untreated, only about the difference between treated and untreated.

We treat q_i^a as observed, and equal to the empirical mean of $\{a_{ij}\}_{j=1}^m$, i.e. $q_i^a = \text{Bernoulli}(\lambda_i = \frac{1}{m_i} a_{ij})$. This modeling choice reflects the fact that in this data, the real population of students is finite and fully observed, rather than a subsample of a larger population. (Note that the choice to treat q_i^a as observed rather than latent does not affect the model in the large m limit, where the identification formula applies.) We also assume C_i and μ_i depend on S_i and q_i^a only through the product $S_i \lambda_i$, the total number of students who are treated at school i .

With these assumptions in place, we consider the following hierarchical Bayesian model, which describes the joint distribution over λ, C, S and μ . On the left hand side we annotate each part of the parametric model with the term of the causal model it is describing. We use the parameterization of the Beta distribution in terms of its mean and precision (Ferrari and Cribari-Neto, 2004).

$$\begin{aligned} Q_i^a &\sim p(q^a) & \kappa_A &\sim \text{Normal}(0, 10) \text{ and } \zeta_A \sim \text{Normal}(0, 10) \\ & & \lambda_i &\sim \text{Beta}(\text{mean} = \sigma(\kappa_A), \text{precision} = \log(1 + \exp(\zeta_A))) \\ S_i &\sim p(s) & \kappa_S &\sim \text{Normal}(50, 100) \text{ and } \zeta_S \sim \text{HalfCauchy}(100) \\ C_i &\sim p(c | q^a, s) & S_i &\sim \text{Normal}(\kappa_S, \zeta_S) \\ & & \alpha_C &\sim \text{Normal}(0, 50) \text{ and } \beta_C \sim \text{Normal}(10, 100) \text{ and } \zeta_C \sim \text{HalfCauchy}(10) \\ & & C_i &\sim \text{Normal}(\alpha_C s_i \lambda_i + \beta_C, \zeta_C) \\ \mu_i &\sim p(\mu | q^a, s, c) & \alpha_Y &\sim \text{Normal}(0, 100), \beta_Y \sim \text{Normal}(0, 100), \omega_Y \sim \text{Normal}(0, 500), \tau \sim \text{HalfCauchy}(5) \\ & & \mu_i &\sim \text{Normal}(\alpha_Y c_i + \beta_Y s_i \lambda_i + \omega_Y, \tau) \\ & & \hat{\mu}_i &\sim \text{Normal}(\mu_i, \sigma_i). \end{aligned}$$

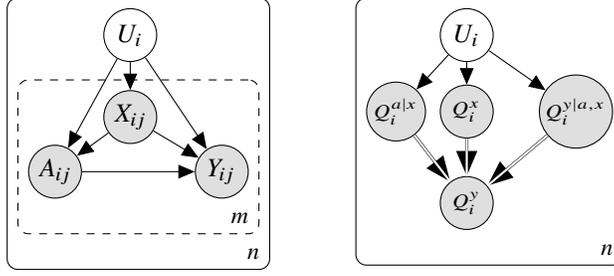
We draw samples from the posterior using the NUTS sampler in NumPyro (Hoffman and Gelman, 2014; Phan et al., 2019; Bingham et al., 2019). We use these samples to form a Monte Carlo approximation of Equation (73). Note that since we have limited data, we draw samples from the model's estimate of $p(s)$ and $p(q^a, s)$, rather than use the empirical distribution. When computing the Monte Carlo approximation of the treatment effect, we clip posterior samples of s and c that are physically impossible, setting negative or zero values of s and c to 1, and setting values of c greater than s to s .

Appendix M. Framing Previous Models as HCMs

M.1 Fixed-effects

In this section we detail how fixed-effects models can be seen as examples of HCMs. Fixed-effects models are a staple of econometrics and related fields. They are perhaps most often applied to

(a) Hierarchical causal model. (b) Collapsed and augmented.


Figure A13: Hierarchical causal model for fixed-effect models and related methods.

panel data, in which observations are made of the same set of people at different timepoints. In this context, we can think of each person as a unit, and each timepoint as a subunit.

We can understand fixed-effect models in terms of the HCM in Figure A13. The idea of the method is to correct for unobserved confounders at the unit level, U_i . In the context of econometric panel data this unobserved confounder could represent, for example, the latent ability of each individual. We observe covariates X_{ij} , treatment status A_{ij} and outcome Y_{ij} for each unit.

Standard fixed-effects models posit a parametric, linear causal mechanism. In our framework, this corresponds to a hierarchical structural causal model with,

$$y_{ij} = f^y(u_i, \gamma_i^y, a_{ij}, x_{ij}, \epsilon_{ij}^y) = \alpha a_{ij} + \beta^\top x_{ij} + \delta^\top u_i + \epsilon_{ij}^y. \quad (74)$$

(Note that although the unit noise γ_i^y does not appear in this expression, it can be absorbed into u_i without loss of generality, since u_i is latent.) The noise distribution $p(\epsilon^y)$ is assumed to have mean zero; we will take it to be Gaussian with standard deviation σ for simplicity. In the HCGM we have,

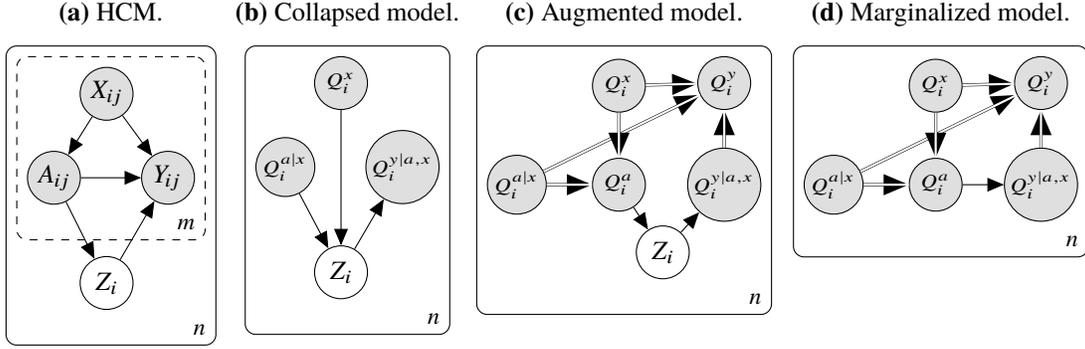
$$Y_{ij} \sim q_i^{y|a,x}(y | a_{ij}, x_{ij}) = \text{Normal}(\alpha a_{ij} + \beta^\top x_{ij} + z_i, \sigma), \quad (75)$$

where $z_i \triangleq \gamma^\top u_i$ is a latent scalar, per-unit offset. Estimation methods for fixed-effect models proceed based on Equation (75), fitting α , β and $\{z_i\}_{i=1}^n$ to the dataset and thus inferring $\{q_i^{y|a,x}\}_{i=1}^n$.

The primary goal of inference in fixed-effect models is to learn α , the coefficient on the treatment. In the HCM framework, we can understand α as the average treatment effect on Y of a hard intervention on A . In particular, we can compute, using the collapsed and augmented model (Figure A13b),

$$\begin{aligned} \mathbb{E}_p[\mathbb{E}_Q[Y]; \text{do}(q^a = \delta_{a_\star})] &= \int \int \mathbb{E}_{q^{y|a,x}}[Y | A = a_\star, x] q^x(x) dx p(q^x, q^{y|a,x}) dq^x dq^{y|a,x} \\ &= \int [\alpha a_\star + \beta^\top \mathbb{E}_{q^x}[X] + z] p(q^x, z) dq^x dz = \alpha a_\star + \beta^\top \mathbb{E}_p[\mathbb{E}_{Q^x}[X]] + \mathbb{E}_p[Z], \end{aligned} \quad (76)$$

where in the second line we have reparameterized the integral, writing it in terms of the parameter z that determines $q^{y|a,x}$ rather than $q^{y|a,x}$ itself. Now we have the average treatment effect, $\mathbb{E}_p[\mathbb{E}_{Q^y}[Y]; \text{do}(q^a = \delta_1)] - \mathbb{E}_p[\mathbb{E}_{Q^y}[Y]; \text{do}(q^a = \delta_0)] = \alpha$. In summary, we can understand fixed-effects models as examples of the HCM in Figure A13, parameterized with a linear model for Y .


Figure A14: Unobserved interference model.

There are many extensions and variations of this basic fixed-effect model that also fall within the HCM framework, by similar arguments. The coefficients α and β may be allowed to vary across units, in which case the average treatment effect in the HCM coincides with the standard target of estimation, $\mathbb{E}_p[\alpha]$ (Wooldridge, 2005). Time may be included as a covariate in X to allow for time trends in the model. Difference-in-difference estimators and synthetic control models can also be derived from the same HCM, by a similar logic (Abadie et al., 2010).

M.2 Interference

In this section we describe in more detail how existing models of interference can be understood in terms of HCMs. Consider the model in Figure A14, which has an unobserved interferer Z_i . From the collapsed, augmented and marginalized model, we can identify the effect of a soft intervention on the treatment A with a backdoor correction,

$$\begin{aligned}
 & \mathbb{E}_p[\mathbb{E}_Q[Y]; \text{do}(A \sim q_\star^a)] \\
 &= \mathbb{E}_p[\mathbb{E}_Q[Y]; \text{do}(q^{a|x} = q_\star^a)] = \mathbb{E}_p[\mathbb{E}_{Q^x}[\mathbb{E}_{Q^{a|x}}[\mathbb{E}_{Q^{y|a,x}}[Y|A, X]]]; \text{do}(q^{a|x} = q_\star^a)] \\
 &= \int \int \int \int \int y q^{y|a,x}(y | a, x) q^x(x) q_\star^a(a) da dx dy p(q^{y|a,x} | q^{a|x} = q_\star^a, q^x) p(q^x) dq^{y|a,x} dq^x \\
 &= \int \int \int \int \int y q^{y|a,x}(y | a, x) q^x(x) q_\star^a(a) da dx dy p(q^{y|a,x} | q^a = q_\star^a) p(q^x) dq^{y|a,x} dq^x.
 \end{aligned} \tag{77}$$

Note we identify the effect by equating it to a hard intervention on $Q^{a|x}$, rather than on Q^a , as Proposition 10 does not apply in Figure A14d. To obtain the final equality above, we use the fact that $Q^{y|a,x}$ must depend on Q^x and $Q^{a|x}$ only through their marginal Q^a . A crucial consequence of unobserved interference is the appearance of the term $p(q^{y|a,x} | q_\star^a)$ in the identification formula, which implies that we must predict $q^{y|a,x}$ from the distribution of the treatment, q^a .

From this HCM identification formula, we can derive, as special cases, some standard techniques for correcting for interference. Assume the treatment is binary, so that q^a is Bernoulli, and q^a is characterized entirely by its mean μ^a . Then, consider a linear HSCM mechanism for Y ,

$$\gamma_i^y \sim \text{Normal}(0, \sigma), \quad \epsilon_{ij}^y \sim \text{Normal}(0, \tau), \quad y_{ij} = \alpha a_{ij} + \beta_i^\top x_{ij} + \kappa \mu_i^a + \gamma_i^y + \epsilon_{ij}^y. \tag{78}$$

With this parameterization, the mean treatment within each unit μ^a becomes essentially just another covariate, and one can estimate the coefficients α , β and κ via a fixed-effects regression model. Now, in the HCGM, $p(q^{y|a,x} | q^a = q_\star^a)$ is given by,

$$\begin{aligned} \gamma^y &\sim \text{Normal}(0, \sigma) \\ q^{y|a,x}(\cdot | a, x) &= \text{Normal}(\alpha a + \beta^\top x + \kappa \mu_\star^a + \gamma^y, \tau). \end{aligned} \tag{79}$$

Plugging Equation (79) into Equation (77) we can derive the effect,

$$\mathbb{E}_p[\mathbb{E}_Q[Y]; \text{do}(A \sim \text{Bernoulli}(\mu_\star^a))] - \mathbb{E}_p[\mathbb{E}_Q[Y]; \text{do}(A \sim \text{Bernoulli}(\mu_{\star\star}^a))] = (\alpha + \kappa)(\mu_\star^a - \mu_{\star\star}^a). \tag{80}$$

We can understand this effect as a combination of the direct effect of A , which has coefficient α , and an interference effect, which has coefficient κ .

So, one way to account for interference is to include as a covariate, in a model for Y , an estimate of the mean μ^a of the treatment distribution within each unit. This approach is widely used in practice in studies of clustered interference, peer effects, etc. (Duflo et al., 2011; Angrist, 2014). Here we saw μ^a enter as a covariate in a linear model, but it can also enter into nonlinear models (Lee et al., 2022).

One subtlety is that our identifying equations are based on large m asymptotics, and are agnostic to how μ^a is estimated. Many interference methods recommend using the leave-one-out mean, estimated based on all the subunits except the one being predicted. For example, in Equation (78), one would replace μ^a with $\hat{\mu}_{-j}^a = \frac{1}{m-1} \sum_{j'=1}^m a_{ij'} \mathbb{1}(j' \neq j)$. This can be viewed as a sample-split estimate of μ^a .

M.3 Multi-site instrumental variables

The multi-site instrumental variable models studied by Reardon et al. (2014); Raudenbush et al. (2012) can be seen as HCMs in which the instrument, treatment, and outcome are all subunit level, and there are both unit level and subunit level confounders between the treatment and outcome. In the linear models proposed by Reardon et al. (2014); Raudenbush et al. (2012), the bias in standard IV methods introduced by the unit confounder is referred to as ‘‘compliance-effect covariance bias’’. While these models are HCMs, our nonparametric identification methods do not apply, as identification here depends on parametric assumptions.

M.4 Multi-environment learning

We next connect multi-environment causal models to HCMs. In the context of HCMs, different environments correspond to different units.

Several multi-environment methods have been developed to address problems in which the graph of subunit variables is at least partially unknown. To discover or account for the unknown graph, these methods make assumptions about the causal relationships between subunit treatments A_{ij} and subunit outcomes Y_{ij} . If these assumptions hold, there will be a detectable signature in the data that A_{ij} indeed causes Y_{ij} , and not the other way around. Broadly speaking, there are two typical classes of assumptions: the independent causal mechanism assumption, and the invariance assumption.

Independent causal mechanisms. Under the hypothesis that A causes Y , the independent causal mechanism assumption states that the mechanism generating A and the mechanism generating Y are

independent across units (Chapters 2, 4 Peters et al., 2017; Perry et al., 2022; Guo et al., 2022). In the framing of HCMs, this means that Q^a is independent of $Q^{y|a}$. So, one way to recover the independent causal mechanism assumption is to assume: There are no unit confounders nor interferers between A and Y .

Invariance. Under the hypothesis that A causes Y , the strong invariance assumption states that the stochastic mechanism generating Y is fixed across units (Peters et al., 2016; Yin et al., 2021). In the context of HCMs, this means that $q_i^{y|a}$ is constant across units, i.e. $q_1^{y|a} = q_2^{y|a} = \dots$. In an HCM, $Q_i^{y|a}$ will in general vary across units, unless there is no unit-level noise γ_i^Y or other unit-level causes. So, one way to recover the strong invariance assumption is to assume: There are no unit-level causes of Y .