

Discrete Causal Representation Learning

Wenjin Zhang* Yixin Wang† Yuqi Gu*

*Department of Statistics, Columbia University

†Department of Statistics, University of Michigan

Abstract

Causal representation learning seeks to uncover causal relationships among high-level latent variables from low-level, entangled, and noisy observations. Existing approaches often either rely on deep neural networks, which lack interpretability and formal guarantees, or impose restrictive assumptions like linearity, continuous-only observations, and strong structural priors. These limitations particularly challenge applications with a large number of discrete latent variables and mixed-type observations. To address these challenges, we propose *discrete causal representation learning (DCRL)*, a generative framework that models a directed acyclic graph among discrete latent variables, along with a sparse bipartite graph linking latent and observed layers. This design accommodates continuous, count, and binary responses through flexible measurement models while maintaining interpretability. Under mild conditions, we prove that both the bipartite measurement graph and the latent causal graph are identifiable from the observed data distribution alone. We further propose a three-stage *estimate-resample-discovery* pipeline: penalized estimation of the generative model parameters, resampling of latent configurations from the fitted model, and score-based causal discovery on the resampled latents. We establish the consistency of this procedure, ensuring reliable recovery of the latent causal structure. Empirical studies on educational assessment and synthetic image data demonstrate that DCRL recovers sparse and interpretable latent causal structures.

Keywords: Causal Discovery; Causal Representation Learning; Directed Acyclic Graph; Identifiability; Discrete Latent Variables.

1 Introduction

Causal representation learning (CRL) seeks to recover high-level latent variables and their causal structure from low-level, entangled observations such as images, text, or time series

Correspondence to: Yuqi Gu. Email: yuqi.gu@columbia.edu. Address: 928 SSW Building, 1255 Amsterdam Avenue, New York, NY 10025.

(Schölkopf et al., 2021; Moran and Aragam, 2026). While deep generative modeling approaches to CRL have shown strong empirical performance on complex data (Yang et al., 2021; Khemakhem et al., 2020; Javaloy et al., 2023; Fan et al., 2025), their neural architectures remain black boxes with limited interpretability, impeding validation and understanding of what latent variables represent (Moran and Aragam, 2026).

Interpretability and identifiability of models are therefore central to uncovering latent causal mechanisms in complex datasets in a trustworthy manner. Informally, a causal representation is *identifiable* if the observed distribution uniquely determines the parameters of the latent variable model and the causal relations among these latents, up to a specified equivalence relation capturing the unavoidable indeterminacies. Without identifiability, representation learning is prone to practical failures such as underspecification and posterior collapse (D’Amour et al., 2022; Wang et al., 2021).

A growing body of CRL work has established that, with continuous latent variables, one can typically recover latents only up to permutations and per-coordinate reparameterizations (von Kügelgen et al., 2023; Jin and Syrgkanis, 2024), rather than to a unique canonical form. Moreover, such equivalence classes are essentially tight: without additional structure or side information, stronger identifiability is unattainable (Varici et al., 2025). Even in a fully linear model with perfect single-node interventions, identifiability is limited to scaling and permutation (Squires et al., 2022; Buchholz et al., 2023). This inherent indeterminacy prevents specific numerical values of latent variables from carrying stable semantic meaning, even in identifiable continuous-latent CRL models.

By contrast, discrete latent models with even highly nonlinear measurements can achieve identifiability up to latent-coordinate permutation alone (Lee and Gu, 2024, 2025). Consequently, while continuous variables appear more expressive, only a limited portion of the information they encode is invariant under allowable reparameterizations and therefore robustly interpretable. Discrete latent variable models thus offer a more stable form of in-

interpretability: the equivalence classes are smaller and easier to characterize, and latent coordinates can be more directly aligned with the ground truth causal factors.

From a practical perspective, this discreteness is also often the right abstraction: in many settings, the goal is to infer an unobserved state that drives observations and supports downstream decisions, rather than to estimate a calibrated real-valued quantity. For instance, in medicine, probabilistic models often represent diseases as discrete latent variables that generate observed symptoms or test results, so different latent-state configurations correspond to different clinical regimes (Shwe et al., 1991). In education, cognitive diagnosis models treat a student’s mastery of skills as discrete latent variables that explain item responses (Helm et al., 2022). In such domains, learning a continuous latent coordinate first and then imposing cutoffs yields non-canonical thresholds whose meanings can shift under scaling without additional anchoring. Discrete latent variables instead represent the abstractions directly.

Motivated by these considerations, we develop a *discrete causal representation learning* (DCRL) framework in which (i) discrete latent variables follow a latent DAG, and (ii) observations are generated through a sparse measurement graph linking the latent and observed layers with flexible mixed-type likelihoods. This framework accommodates continuous, count, and binary responses while allowing highly nonlinear latent-observation relationships. Within this new framework, our contributions are threefold.

First, despite the expressiveness of the proposed framework, we establish formal identifiability guarantees from a single observational distribution, without requiring interventions, multiple environments, or observed auxiliary variables. Our main contribution is generic identifiability: under mild conditions, outside a measure-zero set of parameter values, the latent distribution, measurement layer, and latent DAG are identifiable from the observed data distribution, uniquely up to latent label permutations, so the unavoidable equivalence class consists solely of relabelings of latent variables. Generic identifiability is directly analogous to how faithfulness excludes measure-zero violations of conditional independences in

causal graphical models (Spirtes et al., 2000; Ghassami et al., 2020). Under additional design conditions, we also obtain a stronger strict identifiability statement.

Second, we propose and analyze a novel, simple, and modular three-stage estimation pipeline. Stage I fits the discrete generative process by penalized maximum likelihood via a stochastic approximation expectation-maximization (SAEM) algorithm with spectral initialization, yielding estimates of the latent distribution and measurement graph while remaining computationally efficient. Stage II resamples latent configurations from the fitted latent law to construct a synthetic dataset in the latent space. Stage III applies Greedy Equivalence Search (GES) (Chickering, 2002) to this resampled latent dataset to recover the latent DAG. The validity of this algorithm relies on a key theoretical question: whether GES remains valid when it operates on samples from an *estimated* latent law rather than from the true one. We answer this by extending classical notions of consistency and local consistency for scoring criteria to a rate-robust setting that permits the scoring distribution to converge to the truth at a controlled rate. Our analysis provides an explicit coupling between the convergence rate in Stage I and the resampling size in Stage II that guarantees GES applied to the resampled latents still recovers the Markov equivalence class of the true latent DAG.

Third, we show that DCRL can reveal meaningful latent causal structure from data and yield interpretable discrete causal factors in practice. Through two empirical studies in educational assessment data and high-dimensional image data, we find that the learned latents align closely with domain-specific concepts and that the recovered latent DAG captures the underlying causal dependencies.

Organization. Section 2 introduces the DCRL framework. In Section 3, we establish identifiability results for DCRL. Section 4 describes the proposed three-stage estimation pipeline and establishes its theoretical consistency guarantees. Section 5 and Section 6 present simulation studies and real data applications, respectively, to demonstrate the effectiveness of our approach. Section 7 concludes the paper with a discussion of potential extensions. All

proofs are deferred to the Supplementary Material.

Notation. We write $a_N = \omega(b_N)$ (resp. $a_N = o(b_N)$) if $\frac{a_N}{b_N} \rightarrow \infty$ (resp. $\frac{a_N}{b_N} \rightarrow 0$) as $N \rightarrow \infty$. For a positive integer n , we write $[n] = \{1, \dots, n\}$. For a matrix \mathbf{A} , we use $\mathbf{A}_{i,:}$ (resp. $\mathbf{A}_{:,j}$) for its i -th row (resp. j -th column). For vectors $x, y \in \mathbb{R}^d$, write $x \succ y$ (resp. $x \preceq y$) if $x_k \geq y_k$ (resp. $x_k \leq y_k$) for all $k = 1, \dots, d$. We write $G \rtimes H$ for the semidirect product of groups G and H . For a set A in a topological space, we write A° for the interior of A .

2 Discrete Causal Representation Learning Framework

Causal Graphical Models. We begin by reviewing essential definitions to fix notation. Let $\mathbf{R} = (R_1, \dots, R_d)$ be random variables with joint distribution $p^*(\mathbf{R})$. We consider graphs $G = (V, E)$, where $V = 1, \dots, d$ corresponds to the variables R_i and $E \subseteq V \times V$ is the set of edges. A directed acyclic graph (DAG) is a directed graph with no cycles.

We now relate these graphs to conditional independencies. For a distribution p on $\mathbf{R} = (R_1, \dots, R_d)$, let $\mathcal{I}(p) := \{(A \perp\!\!\!\perp B \mid C)_p : R_A \perp\!\!\!\perp_p R_B \mid R_C\}$ denote the set of all conditional independence statements that hold under p , where $A, B, C \subseteq V$ are pairwise disjoint and $R_A = \{R_i : i \in A\}$. Let $\mathcal{I}(G) := \{(A \perp\!\!\!\perp B \mid C)_G : R_A \perp\!\!\!\perp_G R_B \mid R_C\}$ be the collection of conditional independences encoded by a DAG G via d-separation. We say p is Markov with respect to G if $\mathcal{I}(G) \subseteq \mathcal{I}(p)$. Write $\mathcal{M}(G) := \{p : \mathcal{I}(G) \subseteq \mathcal{I}(p)\}$. For DAGs, $p \in \mathcal{M}(G)$ is equivalent to the factorization of the joint density according to G : $p(\mathbf{R}) = \prod_{i=1}^d p(R_i \mid R_{\mathbf{Pa}_i^G})$, where \mathbf{Pa}_i^G is the parent set of node i in G (Lauritzen, 1996). We call p faithful to G if $\mathcal{I}(p) \subseteq \mathcal{I}(G)$ (Koller and Friedman, 2009). Combining the two inclusions, a distribution is DAG-perfect (Chickering, 2002) to G if $\mathcal{I}(G) = \mathcal{I}(p)$, so that G encodes exactly all conditional independences of p . In this case, G is a perfect map of p .

We say G_1 and G_2 are Markov equivalent if $\mathcal{I}(G_1) = \mathcal{I}(G_2)$, and write $G_1 \equiv G_2$. Markov-equivalent DAGs form a Markov equivalence class. Two DAGs are Markov equivalent if and only if they have the same skeleton and v-structures (Verma and Pearl, 1990). Each Markov equivalence class can be uniquely represented by a completed partially directed acyclic graph

(CPDAG), in which an edge $i \rightarrow k$ is directed if and only if it has the same orientation in every DAG in the class, and an edge $i - k$ is undirected if both orientations $i \rightarrow k$ and $i \leftarrow k$ occur among the DAGs in the class (Verma and Pearl, 1990; Chickering, 2002).

A *causal graphical model* consists of a DAG G and a distribution p that is Markov to G , with directed edges interpreted causally. In this paper we are primarily interested in *discrete causal graphical models*, where each variable R_i takes values in a finite state space and (G, p) is a causal graphical model.

Discrete Causal Representation Learning. We take the causal structure as primitive and place probabilistic distributions on top of it. We work with a collection of variables consisting of both observable and latent components. Let $\mathbf{X} = (X_1, \dots, X_J) \in \times_{j=1}^J \mathcal{X}_j$ denote the observed variables, where $\mathcal{X}_j \subseteq \mathbb{R}$ is allowed to be general. In particular, our formulation accommodates a wide range of data types, including continuous measurements, count-valued observations, and binary or categorical responses. We consider binary latent variables and denote them by $\mathbf{Z} = (Z_1, \dots, Z_K) \in \{0, 1\}^K$, where $K \geq 2$. The causal structure is specified by (i) a directed acyclic graph (DAG) \mathcal{G} on the latent variables Z_1, \dots, Z_K , and (ii) a directed bipartite structure $\mathbf{Q} = (q_{j,k}) \in \{0, 1\}^{J \times K}$ from the latent variables to the observed variables, where $q_{j,k} = 1$ if and only if Z_k is a direct cause of X_j for $j \in [J]$ and $k \in [K]$. The bipartite graph describes the measurement modeling structure. Together, $(\mathcal{G}, \mathbf{Q})$ determines a full acyclic causal graph on (\mathbf{Z}, \mathbf{X}) . An illustration is provided in Figure 1.

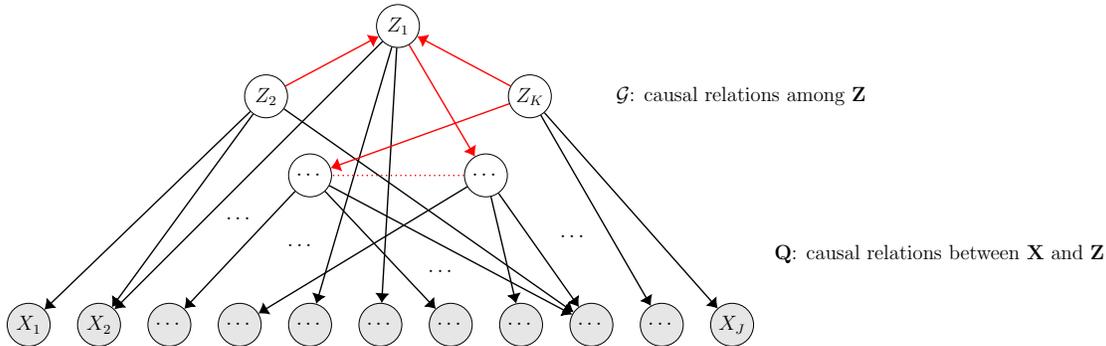


Figure 1: Latent DAG and measurement graph in discrete causal representation learning

To obtain a data-generating process from $(\mathcal{G}, \mathbf{Q})$, we view the joint distribution of the latent variables as the primitive object and assume that it satisfies the Markov factorization associated with \mathcal{G} . Let $\mathbf{p} = (\mathbb{P}(\mathbf{Z} = \mathbf{z}))_{\mathbf{z} \in \{0,1\}^K}$ denote the 2^K -dimensional probability vector of \mathbf{Z} with $\mathbf{p} \in \mathcal{M}(\mathcal{G})$. Equivalently, \mathbf{p} factorizes according to \mathcal{G} as $\mathbf{p}(\mathbf{Z}) = \prod_{k=1}^K \mathbf{p}(Z_k | Z_{\mathbf{pa}_k^{\mathcal{G}}})$, which encodes the directed dependencies prescribed by the latent DAG \mathcal{G} .

Next, we specify how the latent causes act on the observed variables according to \mathbf{Q} . For each item $j \in [J]$, let $K_j = \{k \in [K] : q_{j,k} = 1\}$ be the index set of its latent parents. We write the linear predictor $\eta_j(\mathbf{z})$ as a multilinear polynomial in the binary latent vector $\mathbf{z} = (z_1, \dots, z_K) \in \{0,1\}^K$. For any subset $S \subseteq [K]$, define the monomial feature $\phi_S(\mathbf{z}) := \prod_{k \in S} z_k$, with the convention $\phi_{\emptyset}(\mathbf{z}) \equiv 1$. Fix once and for all an ordering (S_1, \dots, S_{2^K}) of all subsets of $[K]$. Let $\phi(\mathbf{z}) \in \{0,1\}^{2^K}$ be the corresponding feature vector with entries $\phi_m(\mathbf{z}) = \phi_{S_m}(\mathbf{z})$. For item j , collect coefficients into $\beta_j \in \mathbb{R}^{2^K}$ by $(\beta_j)_m := \beta_{j,S_m}$, and impose $(\beta_j)_m = 0$ whenever $S_m \not\subseteq K_j$, so that only main effects and interactions among coordinates in K_j are allowed. Equivalently, $\eta_j(\mathbf{z}) = \beta_j^\top \phi(\mathbf{z}) = \sum_{S \subseteq K_j} \beta_{j,S} \prod_{k \in S} z_k$. Stacking the rows yields the matrix $\mathbf{B} = [\beta_1 \ \dots \ \beta_J]^\top \in \mathbb{R}^{J \times 2^K}$.

In practice, it is often sufficient to restrict attention to the main effects of the latent causes, which corresponds to setting $\beta_{j,S} = 0$ whenever $|S| > 1$. In this case, the linear predictor reduces to $\eta_j(\mathbf{z}) = \beta_{j,\emptyset} + \sum_{k \in K_j} \beta_{j,k} z_k$, which is analogous to a generalized linear specification with an intercept and main effects in the binary latent vector \mathbf{z} as covariates. Compared with the all-effect specification, this restriction reduces the number of parameters from 2^K to $|K_j| + 1$ per item, yielding a more parsimonious and interpretable specification. In our subsequent estimation procedure, we will primarily focus on this main-effect specification.

Conditionally on \mathbf{Z} , we model each X_j by an item-specific parametric family consistent with \mathbf{Q} and assume conditional independence across j given \mathbf{Z} :

$$X_j \mid \mathbf{Z} \sim \text{ParFam}_j\left(g_j(\eta_j(\mathbf{Z}), \gamma_j)\right), \quad j \in [J]. \quad (1)$$

Here $\text{ParFam}_j = \{\mathbb{P}_{j,\boldsymbol{\theta}} : \boldsymbol{\theta} \in H_j\}$ is a known family with parameter space $H_j \subseteq \mathbb{R}^{h_j}$, and $g_j : \mathbb{R} \times [0, \infty) \rightarrow H_j$ is a known link mapping the linear predictor $\eta_j(\cdot)$ and, when applicable, a dispersion parameter $\gamma_j > 0$ to the parameter of ParFam_j . Write $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_J)$.

Integrating out the latent \mathbf{Z} yields the marginal law of the observables,

$$\mathbb{P}(\mathbf{X}) = \sum_{\mathbf{z} \in \{0,1\}^K} \mathbb{P}(\mathbf{X} \mid \mathbf{Z} = \mathbf{z}) \mathbb{P}(\mathbf{Z} = \mathbf{z}), \quad (2)$$

which is determined by the triple $(\mathbf{p}, \mathbf{B}, \boldsymbol{\gamma})$ together with the causal structure $(\mathcal{G}, \mathbf{Q})$.

Definition 1. *We consider the following discrete causal representation learning data-generating process parameterized by $(\mathbf{p}, \mathcal{G}, \mathbf{B}, \mathbf{Q}, \boldsymbol{\gamma})$, where the marginal law of the observed data is given by (2). When \mathcal{G} and \mathbf{Q} are known and fixed, this induces a family of probability distributions on \mathbf{X} , indexed by $(\boldsymbol{\Theta}, \mathcal{G}, \mathbf{Q})$, which we denote by $\mathbb{P}_{\boldsymbol{\Theta}, \mathcal{G}, \mathbf{Q}}$, where $\boldsymbol{\Theta} := (\mathbf{p}, \mathbf{B}, \boldsymbol{\gamma})$.*

Remark 1. *For clarity, Section 2 presents the framework under binary latent attributes $Z_k \in \{0, 1\}$. All components extend to ordered polytomous attributes $Z_k \in [M_k] = \{0, 1, \dots, M_k - 1\}$. In the extension, η_j is still a linear combination of coefficients $\{\beta_{j,\mathbf{u}}\}$ indexed by latent “states” \mathbf{u} , and a term $\beta_{j,\mathbf{u}}$ contributes to $\eta_j(\mathbf{z})$ only when $\mathbf{u} \preceq \mathbf{z}$ coordinatewise. Moreover, $\beta_{j,\mathbf{u}}$ is nonzero only if $\text{supp}(\mathbf{u}) \subseteq K_j$. The extension and its generic identifiability result are stated in Section 3.2, with full definitions in Appendix S.4.*

Taken together, the DCRL framework yields a highly flexible and potentially strongly nonlinear measurement layer from the latent configuration to the observed responses. On the one hand, the all-effect specification for $\eta_j(\mathbf{z})$ allows arbitrary interaction patterns among the latent variables, including highest-order interactions over all attributes in K_j . On the other hand, by allowing a general parametric family ParFam_j and link mapping g_j , we impose no single fixed response family, so that a wide range of nonlinear conditional distributions $X_j \mid \mathbf{Z}$ can be accommodated.

Under DCRL, the complete causal structure comprises two components: the DAG \mathcal{G}

among latent variables Z_1, \dots, Z_K , and the directed bipartite structure \mathbf{Q} from latent variables to observed variables. Our goal is to jointly recover \mathcal{G} and \mathbf{Q} from \mathbf{X} .

Connections with Existing Studies. Most identifiable models for causal discovery with partially unobserved variables rely on linearity (Anandkumar et al., 2013; Squires et al., 2022; Huang et al., 2022; Dong et al., 2026), an assumption that often fails in practice. Beyond linearity, identifiability has also been established for certain nonlinear latent hierarchical models (Prashant et al., 2025). However, those results impose much stronger restrictions on the latent causal architecture than ours. See Supplement S.10 for more details.

3 Identifiability

We study identifiability of the DCRL framework under two notions: strict identifiability and generic identifiability. Our main new contribution is Theorem 1, which provides a generic identifiability theorem for the all-effect measurement specification under mild conditions. Since the strict identifiability results and several related extensions are adapted from existing work (Liu et al., 2025; Lee and Gu, 2025), we include only a brief discussion in the main text and defer the formal statements to Supplement Section S.1.

Strict identifiability formalizes the requirement that the model parameters are uniquely determined by the marginal law of the observable vector \mathbf{X} , up to the unavoidable label switching of the K latent variables and the Markov equivalence class ambiguity of the latent DAG. Under the equivalence relation $\sim_{\mathcal{K}}$ and the admissible parameter space introduced below, strict identifiability can be guaranteed by mild structural conditions on the measurement design together with a monotonicity-type assumption that rules out sign-flipping of latent coordinates. See Supplement Section S.1 for more details.

3.1 Generic Identifiability

Before introducing our identifiability notion, we first relate our framework to the statistical-identifiability formulation used in recent CRL work (Xi and Bloem-Reddy, 2023; Moran and

Aragam, 2026). Consider the model $X = f(Z) + \epsilon$, where $Z \in \mathcal{Z}$ is the latent variable, $f : \mathcal{Z} \rightarrow \mathcal{X}$ is the representation map, and ϵ is noise. Let \mathcal{F} be a class of admissible maps f , and let \mathcal{P} be a class of admissible latent distributions p on \mathcal{Z} . For any bijection $\xi : \mathcal{Z} \rightarrow \mathcal{Z}$, one may rewrite the model as $X = (f \circ \xi^{-1})(\xi(Z)) + \epsilon$. Therefore, without restrictions on \mathcal{F} and \mathcal{P} , the model is trivially nonidentifiable. To formalize this, let $\xi_{\#}p$ denote the push-forward of p by ξ . One calls ξ an indeterminacy transformation if $f \circ \xi^{-1} \in \mathcal{F}$ and $\xi_{\#}p \in \mathcal{P}$. The collection of all such transformations is the indeterminacy set $A(\mathcal{F}, \mathcal{P})$. Equivalently, $A(\mathcal{F}, \mathcal{P})$ indexes the “transformation-based” comparison class $\{(f \circ \xi^{-1}, \xi_{\#}p) : \xi \in A(\mathcal{F}, \mathcal{P})\}$ associated with a fixed pair (f, p) . The key question is to determine which restrictions on \mathcal{F} and \mathcal{P} make $A(\mathcal{F}, \mathcal{P})$ small while remaining flexible.

Our DCRL framework fits naturally into this framework. Here $\mathcal{Z} = \{0, 1\}^K$. For Gaussian responses and identity link, we may write $X_j = \eta_j(Z) + \varepsilon_j$ with $\varepsilon_j \sim N(0, \gamma_j^2)$, so that $f(Z) = (\eta_1(Z), \dots, \eta_J(Z)) \in \mathbb{R}^J$. For a fixed latent DAG \mathcal{G} and measurement graph \mathbf{Q} , the corresponding classes can be viewed as $\mathcal{P}_{\mathcal{G}} := \{p \text{ on } \{0, 1\}^K : p \text{ is Markov to } \mathcal{G}\}$, and $\mathcal{F}_{\mathbf{Q}} := \{f = (\eta_1, \dots, \eta_J) : \{0, 1\}^K \rightarrow \mathbb{R}^J : \eta_j(z) = \eta_j(z') \text{ whenever } z_{K_j} = z'_{K_j}, \forall j\}$. Additional assumptions used later in the paper can be understood precisely as further restrictions on $\mathcal{P}_{\mathcal{G}}$ and $\mathcal{F}_{\mathbf{Q}}$, imposed to shrink the admissible indeterminacy set.

This viewpoint is conceptually useful, but it is important to distinguish settings where the full statistical equivalence class coincides with the transformation-based class indexed by $A(\mathcal{F}, \mathcal{P})$ from those where it does not. In identifiable CRL, it is widely assumed that the generative map f is injective (Ahuja et al., 2023; Hartford et al., 2023) and that the observation model is well-posed in the sense that the distribution of $f(Z)$ is determined by that of $X = g(f(Z), \epsilon)$. A simple sufficient condition is an additive and independent observation-noise model $X = f(Z) + \epsilon$ with Gaussian noise distribution. Under these commonly used injective-generator and well-posed observation assumptions, Xi and Bloem-Reddy (2023, Lemma 2.1) shows that if two parameterizations (f_a, p_a) and (f_b, p_b) induce the same obser-

vational distribution, then they must be related by a latent-space automorphism $\xi \in \text{Aut}(\mathcal{Z})$ in the sense that $f_b = f_a \circ \xi^{-1}$ and $p_b = \xi_{\#} p_a$ up to null sets. Hence, for such models, restricting attention to the transformation-based comparison class (equivalently, studying the indeterminacy set $A(\mathcal{F}, \mathcal{P})$) entails no loss of generality. In the discrete case $\mathcal{Z} = \{0, 1\}^K$, such a ξ is simply a permutation of the finite latent state space, so $\xi_{\#} \mathbf{p}$ is merely a relabeling of \mathbf{p} and the remaining transformation-based comparison is finite.

In contrast, our DCRL framework does not directly assume a well-posed observation model in the above sense. Without this end-to-end well-posedness assumption, which would bypass much of the substantive difficulty, the reduction from general parameter-level equivalence to the transformation-based comparison class indexed by $A(\mathcal{F}, \mathcal{P})$ is no longer automatic. Accordingly, we begin by comparing arbitrary admissible parameter triples $(\Theta, \mathcal{G}, \mathbf{Q})$ and $(\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}})$ that induce the same observational distribution, and then enforce this reduction by imposing explicit, verifiable structural restrictions encoded by \mathbf{Q} , together with more delicate techniques tailored to our model class. This additional collapse step is precisely what makes our analysis more involved than approaches that assume well-posedness directly.

Before proceeding, we first specify the parameter spaces. Define

$$\begin{aligned} \Omega_K(\Theta; \mathcal{G}, \mathbf{Q}) &:= \left\{ \Theta : \mathcal{G} \text{ is a perfect map of } \mathbf{p}, \beta_{j,S} = 0 \text{ if } S \not\subseteq K_j, \beta_{j,\{k\}} \neq 0 \text{ iff } k \in K_j \right\}. \\ \Omega_K(\Theta, \mathcal{G}, \mathbf{Q}) &:= \left\{ (\Theta, \mathcal{G}, \mathbf{Q}) : \Theta \in \Omega_K(\Theta; \mathcal{G}, \mathbf{Q}) \right\}. \end{aligned}$$

Now we introduce the equivalence relation that specifies the unavoidable ambiguity, and then define generic identifiability relative to this equivalence.

Definition 2. *For the discrete CRL framework, define an equivalence relationship “ \sim_K ” by setting $(\Theta, \mathcal{G}, \mathbf{Q}) \sim_K (\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}})$ iff $\gamma = \tilde{\gamma}$ and there exists a permutation $\sigma \in S_{[K]}$ such that the following hold. First, $\mathbf{p}_{(z_{\sigma(1)}, \dots, z_{\sigma(K)})} = \tilde{\mathbf{p}}_{\mathbf{z}}$ for all $\mathbf{z} \in \{0, 1\}^K$ and $\mathcal{G} \equiv \sigma(\tilde{\mathcal{G}})$, where $\sigma(\tilde{\mathcal{G}})$ denotes the DAG obtained from $\tilde{\mathcal{G}}$ by relabeling each node k as $\sigma(k)$. Second, $q_{j,k} = \tilde{q}_{j,\sigma(k)}$ for*

all $j \in [J]$, $k \in [K]$, and for all j and $S \subseteq [K]$, $\beta_{j,S} = \tilde{\beta}_{j,\sigma(S)}$, where $\sigma(S) := \{\sigma(k) : k \in S\}$.

Definition 3. Let $(\Theta^*, \mathcal{G}^*, \mathbf{Q}^*) \in \Omega_K(\Theta, \mathcal{G}, \mathbf{Q})$ be the true parameter triple of the discrete causal representation learning framework. The framework is generically identifiable up to $\sim_{\mathcal{K}}$ if $\{\Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) : \exists (\tilde{\Theta}, \tilde{\mathbf{Q}}, \tilde{\mathcal{G}}) \not\sim_{\mathcal{K}} (\Theta, \mathbf{Q}^*, \mathcal{G}^*) \text{ such that } \mathbb{P}_{\tilde{\Theta}, \tilde{\mathbf{Q}}, \tilde{\mathcal{G}}} = \mathbb{P}_{\Theta, \mathbf{Q}^*, \mathcal{G}^*}\}$ is a measure-zero set with respect to $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$.

The equivalence relation $\sim_{\mathcal{K}}$ is the discrete analogue of the transformation-based indeterminacies encoded by $A(\mathcal{F}, \mathcal{P})$, specialized to latent-coordinate relabelings.

The measure-zero qualifier in Definition 3 parallels the faithfulness convention in causal discovery: for a fixed DAG, distributions that are Markov but unfaithful form a Lebesgue-null subset of the parameter space, so faithfulness excludes only a negligible set of degenerate configurations. Our generic-identifiability definition plays the same role here. After imposing the latent Markov-plus-faithfulness condition on \mathbf{p} , non-identifiability can still occur for exceptional values of the continuous measurement parameters $(\mathbf{B}, \boldsymbol{\gamma})$, but these exceptional configurations form a Lebesgue-null set. Hence, restricting attention to generic identifiability excludes only a measure-zero subset of “bad” $(\mathbf{B}, \boldsymbol{\gamma})$ configurations. In this sense, the loss incurred by discarding measure-zero subsets of parameters is as harmless as the loss incurred when imposing faithfulness in the first place.

We now introduce the assumptions needed for our generic identifiability result.

Assumption 1. (a) \mathcal{G} is a perfect map of \mathbf{p} and $\mathbf{p}_{\mathbf{z}} \in (0, 1)$ for all $\mathbf{z} \in \{0, 1\}^K$.

(b) For each item j , $\eta_j(\mathbf{z}) > \eta_j(\mathbf{z}')$ whenever $\mathbf{z} \succeq \mathbf{Q}_{j,:}$ and $\mathbf{z}' \not\succeq \mathbf{Q}_{j,:}$.

Assumption 1(a) is a restriction on the latent distribution class $\mathcal{P}_{\mathcal{G}}$. Assumption 1(b) is a restriction on $\mathcal{F}_{\mathbf{Q}}$: it imposes a monotonicity condition on each item response function η_j . This type of condition is also used in Liu et al. (2025); Lee and Gu (2025) to avoid the sign-flipping for each latent variable.

To ensure generic identifiability, we introduce an additional analytic assumption, which holds for regular minimal exponential families on the interior of natural parameter spaces.

Assumption 2. For each $j \in [J]$, define the canonical countable separating class $\mathcal{C}_j^{\text{can}} := \{\mathcal{X}_j \cap (a, b) : a, b \in \mathbb{Q}, a < b\} \cup \{\mathcal{X}_j\}$. Assume that $H_j^\circ \neq \emptyset$, and that the following hold.

- (i) For every $S \in \mathcal{C}_j^{\text{can}}$, the map $\boldsymbol{\theta} \mapsto \mathbb{P}_{j,\boldsymbol{\theta}}(S)$ is real-analytic on H_j° .
- (ii) $\mathbb{P}_{j,\boldsymbol{\theta}} = \mathbb{P}_{j,\boldsymbol{\theta}'}$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}'$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in H_j^\circ$.
- (iii) The link g_j maps $\mathbb{R} \times (0, \infty)$ into H_j° and is real-analytic. Moreover, exactly one of the following holds. (a) (No dispersion) The link is independent of γ and the slice map $\eta \mapsto g_j(\eta, \gamma_0)$ is injective for some (equivalently, any) fixed $\gamma_0 \in [0, \infty)$. (b) (With dispersion) The full map $(\eta, \gamma) \mapsto g_j(\eta, \gamma)$ is injective on $\mathbb{R} \times (0, \infty)$.

Our main identifiability result is stated in the following theorem.

Theorem 1. Under Assumptions 1 and 2, DCRL is generic identifiable if the following hold.

- (i) After a row permutation, we can write $\mathbf{Q}^* = [\mathbf{Q}_1^\top, \mathbf{Q}_2^\top, \mathbf{Q}_3^\top]^\top$, where $\mathbf{Q}_1, \mathbf{Q}_2 \in \{0, 1\}^{K \times K}$ have unit diagonals (off-diagonals arbitrary), and \mathbf{Q}_3 has no all-zero column.
- (ii) No column of \mathbf{Q}^* contains another: for any $p \neq q$, neither $\mathbf{Q}_{:,p}^* \succeq \mathbf{Q}_{:,q}^*$ nor $\mathbf{Q}_{:,q}^* \succeq \mathbf{Q}_{:,p}^*$.

Condition (i) is best viewed as a weak coverage requirement on the measurement design: it guarantees that every latent coordinate affects at least one observed variable, and that there exist some anchor-like items in which each latent is forced to appear. However, it is weak in the sense that such anchor-like items may still depend on many other latents. Our condition (ii) coincides with Condition 3.1 (the subset condition) in [Kivva et al. \(2021\)](#). As they observed, violating this subset condition can lead to non-identifiability. Similarly, we emphasize that condition (i) alone is not sufficient for generic identifiability. In particular, [Appendix S.2](#) constructs a model satisfying condition (i) in which \mathbf{Q} has distinct columns and each column contains at least one zero, yet the framework fails to be generically identifiable.

Our proof is a three-step reduction of the comparison class. First, a Kruskal-type tensor argument collapses the original continuous parameter comparison to a finite, transformation-generated comparison class: any remaining competitor with the same observational law must be of the form $(\xi_{\#}\mathbf{p}, \boldsymbol{\eta} \circ \xi^{-1})$, where $\xi \in S_{2^K}$ is a permutation of the 2^K latent states of \mathbf{Z} . Thus, after the tensor step, we are essentially in the setting of [Moran and Aragam \(2026\)](#), and the problem becomes how to shrink the indeterminacy set $\mathcal{A}(\mathcal{F}, \mathcal{P}) = S_{2^K}$. Before further reducing $\mathcal{A}(\mathcal{F}, \mathcal{P})$, we first identify \mathbf{Q} : using Assumption 1(b) and an inclusion–exclusion argument on the transformed $\boldsymbol{\eta}$ -array, we show that all these 2^K admissible competitors must have measurement matrices that agree with the true \mathbf{Q} up to a coordinate permutation. This observation makes the structural constraint on $\mathcal{F}_{\mathbf{Q}}$ much clearer. We then return to reducing $\mathcal{A}(\mathcal{F}, \mathcal{P})$: the subset condition, together with the structural constraint on $\mathcal{F}_{\mathbf{Q}}$, reduces the indeterminacy set from S_{2^K} to $(\mathbb{Z}_2)^K \rtimes S_K$, corresponding to coordinate permutations combined with coordinatewise bit-flips, and Assumption 1(b) further rules out bit-flips, leaving only coordinate relabelings in S_K . Finally, $\boldsymbol{\beta}$ and \mathcal{G} are recovered from the invertible linear map $\boldsymbol{\eta} \leftrightarrow \boldsymbol{\beta}$ and the perfect-map condition, yielding identifiability.

3.2 Extension to Identifiability of the Polytomous-attribute Case

We extend the binary-latent-attribute framework in Section 2 to the case where each latent attribute is polytomous with possibly different numbers of categories. Fix integers $M_k \geq 2$ for $k \in [K]$ and let $\mathbf{Z} = (Z_1, \dots, Z_K)$ take values in $\prod_{k=1}^K [M_k]$ where $[M_k] = \{0, 1, \dots, M_k - 1\}$. The latent law \mathbf{p} is generated by a categorical Bayesian network on a latent DAG \mathcal{G} . Conditionally on \mathbf{Z} , the items are independent and each $X_j \mid \mathbf{Z}$ follows the same item-specific family and link specification as in (1).

Similar to the binary case, we still use $\mathbf{Q} = (q_{j,k}) \in \{0, 1\}^{J \times K}$ to encode the bipartite measurement graph, where $q_{j,k} = 1$ means that the latent variable Z_k is a direct cause of the observed variable X_j , and set $K_j := \{k \in [K] : q_{j,k} = 1\}$. For any latent configuration vector $\mathbf{u} \in \prod_{k=1}^K [M_k]$, define $\text{supp}(\mathbf{u}) := \{k \in [K] : u_k \geq 1\}$. For each $j \in [J]$, introduce coefficients

$\{\beta_{j,\mathbf{u}}\}$ indexed by $\mathbf{u} \in \prod_{k=1}^K [M_k]$, and define the linear predictor for every latent state $\mathbf{z} \in \prod_{k=1}^K [M_k]$ by $\eta_j(\mathbf{z}) := \sum_{\mathbf{u}: \text{supp}(\mathbf{u}) \subseteq K_j, \mathbf{u} \preceq \mathbf{z}} \beta_{j,\mathbf{u}}$. Equivalently, $\eta_j(\mathbf{z})$ is a linear combination of the coefficients $\{\beta_{j,\mathbf{u}}\}$, and a term $\beta_{j,\mathbf{u}}$ contributes to $\eta_j(\mathbf{z})$ only when $\text{supp}(\mathbf{u}) \subseteq K_j$ and $\mathbf{u} \preceq \mathbf{z}$. Collect $\boldsymbol{\beta}_j = (\beta_{j,\mathbf{u}})_{\mathbf{u} \in \prod_{k=1}^K [M_k]} \in \mathbb{R}^r$ and let $\mathbf{B} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_J^\top)^\top \in \mathbb{R}^{J \times r}$.

With this parameterization in place, \mathbf{Q} still admits a direct causal interpretation in the polytomous-attribute regime: if $q_{j,k} = 0$, then varying Z_k while holding the other latents fixed does not change $\eta_j(\mathbf{z})$ and hence does not affect the conditional law of X_j ; whereas if $q_{j,k} = 1$, increasing Z_k from level s to $s+1$ activates new contributions in $\eta_j(\mathbf{z})$, including a level- $(s+1)$ main-effect contribution of Z_k and potentially additional interaction contributions with other latent causes in K_j that become available only after Z_k reaches level $s+1$. Because these newly activated effects are assigned their own parameters rather than being constrained to scale linearly across levels, the causal effect of Z_k on X_j can vary across levels and across configurations of the other latents, making the model highly flexible.

Latent relabeling can only permute coordinates with the same number of categories. In particular, coordinates with different M_k cannot be swapped. All formal definitions, including the parameter spaces and the generic identifiability notion, are given in Appendix S.4.

Conditions in Theorem 2 are parallel to those in Theorem 1, and Assumption 3 is the ordered-level analogue of the monotonicity in Assumption 1(b). With these in place, we obtain generic identifiability for the polytomous-attribute extension of DCRL. It is worth pointing out that, if $M_k = 2$ for all k (i.e., the binary-attribute case), then all assumptions and conditions in Theorem 2 reduce to Theorem 1 exactly.

Assumption 3. For each item j , define the top-set $\mathcal{M}_j := \{\mathbf{Z} \in \mathcal{Z} : Z_k = M_k - 1 \text{ for all } k \in K_j\}$. Assume $\max_{\mathbf{Z} \in \mathcal{Z} \setminus \mathcal{M}_j} \eta_j(\mathbf{Z}) < \min_{\mathbf{Z} \in \mathcal{M}_j} \eta_j(\mathbf{Z})$ for $j \in [J]$. Moreover, for each item j , each $k \in K_j$, and each threshold $u \in [M_k - 1]$, define $\mathcal{T}_{1,j}^{(k,u)} := \{\mathbf{Z} \in \mathcal{Z} : Z_k \geq u, Z_h = M_h - 1 \forall h \in K_j \setminus \{k\}\}$ and $\mathcal{T}_{0,j}^{(k,u)} := \{\mathbf{Z} \in \mathcal{Z} : Z_k \leq u - 1, Z_h = M_h - 1 \forall h \in K_j \setminus \{k\}\}$, and assume $\max_{\mathbf{Z} \in \mathcal{T}_{0,j}^{(k,u)}} \eta_j(\mathbf{Z}) < \min_{\mathbf{Z} \in \mathcal{T}_{1,j}^{(k,u)}} \eta_j(\mathbf{Z})$.

Theorem 2. *Under Assumption 1(a), Assumption 2, and Assumption 3, the polytomous-attribute DCRL is generically identifiable if the following conditions hold.*

(i) *For each $k \in [K]$, let $m_k = \lceil \log_2 M_k \rceil$ and $\tilde{d} = \sum_{k=1}^K m_k$. After a row permutation, $\mathbf{Q}^* = [\mathbf{Q}_1^\top, \mathbf{Q}_2^\top, \mathbf{Q}_3^\top]^\top$ with $\mathbf{Q}_1, \mathbf{Q}_2 \in \{0, 1\}^{\tilde{d} \times K}$. For each $a \in \{1, 2\}$, $\mathbf{Q}_a = [\mathbf{Q}_{a,1}^\top, \dots, \mathbf{Q}_{a,K}^\top]^\top$, where $\mathbf{Q}_{a,k} \in \{0, 1\}^{m_k \times K}$ and every row of $\mathbf{Q}_{a,k}$ has a 1 in column k . Other entries are arbitrary. Moreover, \mathbf{Q}_3 has no all-zero column.*

(ii) *For any $p \neq q$, neither $\mathbf{Q}_{:,p} \succeq \mathbf{Q}_{:,q}$ nor $\mathbf{Q}_{:,q} \succeq \mathbf{Q}_{:,p}$.*

Condition (i) implies that at least $2 \sum_{k=1}^K (\lceil \log_2 M_k \rceil) + 1$ items are required to achieve generic identifiability. For the binary-attribute case, condition (i) reduces to $J \geq 2K + 1$, matching the binary generic-identifiability requirement in Theorem 1. The logarithmic order $m_k = \lceil \log_2 M_k \rceil$ here is optimal up to constants and the ceiling since this theorem is required to hold uniformly over the general response families. It suffices to consider the binary-response submodel contained in our framework. Fix a coordinate k , and consider any collection of l items with $q_{j,k} = 1$. Write their joint response as $\mathbf{X} \in \{0, 1\}^l$. Choose a latent DAG \mathcal{G} in which Z_k is a root so that its marginal $\mathbf{p}^{(k)}$ can vary freely on an open set, and fix all remaining latent parameters as well as parameters $(\mathbf{B}, \boldsymbol{\gamma})$. Then for each $s \in \{0, \dots, M_k - 1\}$, the conditional law of \mathbf{X} given $Z_k = s$ is a vector $\mathbf{v}_s \in \mathbb{R}^{2^l}$, and the marginal law of Y is the mixture $\mathbb{P}(\mathbf{X} = \cdot) = \sum_{s=0}^{M_k-1} p_s^{(k)} \mathbf{v}_s$, so the map $\mathbf{p}^{(k)} \mapsto \mathbb{P}(\mathbf{X} = \cdot)$ is linear. If $2^l < M_k$, this linear map cannot be injective on any open set: there exists $0 \neq \mathbf{h} \in \mathbb{R}^{M_k}$ with $\sum_s h_s = 0$ such that $\sum_s h_s \mathbf{v}_s = 0$, and hence for every interior $\mathbf{p}^{(k)}$ and all sufficiently small ε , the two distinct marginals $\mathbf{p}^{(k)}$ and $\mathbf{p}^{(k)} + \varepsilon \mathbf{h}$ induce the same law of \mathbf{X} . Thus the set of parameters yielding non-identifiability contains an open subset, so generic identifiability fails in this binary submodel unless $2^l \geq M_k$. Consequently, the logarithmic scaling in condition (i) is sharp in magnitude for the general-response setting, reflecting that polytomous attributes with more categories require more items to achieve identifiability.

Our proof of Theorem 2 follows the same spirit as Theorem 1 but contains delicate arguments to address the extra challenge in the polytomous-attribute case. First, a Kruskal-type tensor argument again collapses the original continuous parameter comparison to a transformation-generated comparison class: any remaining competitor with the same observational law must be of the form $(\xi_{\#}\mathbf{p}, \boldsymbol{\eta} \circ \xi^{-1})$, where now $\xi \in S_r$ and $r = \prod_{k=1}^K M_k$. Then the remaining task is again to shrink $\mathcal{A}(\mathcal{F}, \mathcal{P}) = S_r$. But the next step is more subtle than in the binary case. The inclusion–exclusion reconstruction of \mathbf{Q} does not extend directly to the polytomous setting, so we instead use the subset condition to identify \mathbf{Q} and then reduce the indeterminacy set from S_r to $(\prod_{k=1}^K S_{M_k}) \times S_{K, \mathbf{M}}$, where $S_{K, \mathbf{M}} = \{\sigma \in S_K : M_{\sigma(k)} = M_k \forall k \in [K]\}$. We then impose the monotonicity condition to eliminate the within-coordinate permutations, reducing the indeterminacy set further to $S_{K, \mathbf{M}}$. This difference causes the group-reduction step to exclude a Lebesgue-null exceptional set, whereas the corresponding step in the binary case does not, reflecting the genuine difficulty of the polytomous case. Extracting the second step specifically yields the following corollary.

Corollary 1. *Let $\mathcal{Z} = \prod_{k=1}^K [M_k]$, where $M_k \geq 2$, and consider the representation model $\mathbf{X} = f(\mathbf{Z}) + \epsilon$, where $f = (f_1, \dots, f_J) : \mathcal{Z} \rightarrow \mathbb{R}^J$. For a binary matrix $\mathbf{Q} = (q_{j,k}) \in \{0, 1\}^{J \times K}$, write $K_j := \{k \in [K] : q_{j,k} = 1\}$. Define $\mathcal{P} := \{\mathbf{p} \text{ on } \mathcal{Z} : \sum_{\mathbf{z} \in \mathcal{Z}} p_{\mathbf{z}} = 1\}$ and*

$$\mathcal{F} := \left\{ f : \mathcal{Z} \rightarrow \mathbb{R}^J \mid \begin{array}{l} \text{neither } \mathbf{Q}_{:,a} \succeq \mathbf{Q}_{:,b} \text{ nor } \mathbf{Q}_{:,b} \succeq \mathbf{Q}_{:,a}, \quad (\forall a \neq b), \\ \exists \mathbf{Q} \in \{0, 1\}^{J \times K} \text{ s.t., } f_j(\mathbf{z}) = f_j(\mathbf{z}') \text{ whenever } \mathbf{z}_{K_j} = \mathbf{z}'_{K_j} \quad (j \in [J]), \\ f_j(\mathbf{z}) \neq f_j(\mathbf{z}') \text{ whenever } \mathbf{z}_{K_j} \neq \mathbf{z}'_{K_j} \quad (j \in [J]) \end{array} \right\}$$

Then every admissible indeterminacy transformation $\xi \in \mathcal{A}(\mathcal{F}, \mathcal{P})$ can only permute coordinates (with the same number of categories) and relabel the levels within each coordinate.

Corollary 1 establishes disentanglement in the sense of Moran and Aragam (2026): among all admissible latent bijections $\xi : \mathcal{Z} \rightarrow \mathcal{Z}$ that preserve membership in $(\mathcal{F}, \mathcal{P})$, the only ones that remain are element-wise transformations and permutations. Notably, this con-

clusion is obtained with f treated nonparametrically, subject only to the restrictions in \mathcal{F} , rather than under a specific parameterization like all-effect forms. Moreover, under injective-generator and well-posed observation assumptions that are common in identifiable CRL (Ahuja et al., 2023; Khemakhem et al., 2020), the full statistical equivalence class coincides with the transformation-based indeterminacy class (Xi and Bloem-Reddy, 2023), so the corollary yields a direct disentanglement conclusion from observational data alone. This contrasts with much of the existing literature, where even after adopting the same baseline, additional sources of variation—such as auxiliary variables, multiple environments, or interventions—are typically invoked to further shrink the indeterminacy set down to permutations and component-wise transformations (Khemakhem et al., 2020; Ahuja et al., 2023). In our framework, mild structural constraints on f encoded by \mathbf{Q} already enforce this shrinkage at the level of $A(\mathcal{F}, \mathcal{P})$, which can be viewed as a natural restriction when \mathbf{Q} is used to describe which latents affect which measurements. If one further imposes a monotonicity condition, the within-coordinate relabelings can also be removed.

4 Estimation Procedure and Theoretical Guarantees

With a slight abuse of notation, we let \mathbf{X} denote the $N \times J$ data matrix with rows $\mathbf{X}_1, \dots, \mathbf{X}_N$. Similarly, let \mathbf{Z} denote the corresponding $N \times K$ latent variable matrix with rows $\mathbf{Z}_1, \dots, \mathbf{Z}_N$. Given the bottom-layer data $\mathbf{X} \in \mathbb{R}^{N \times J}$ with unknown latent causal structures and parameters $(\mathbf{p}, \mathcal{G}, \mathbf{B}, \mathbf{Q}, \gamma)$, our objective is to recover \mathbf{Q} and \mathcal{G} . In this section, we provide the complete pipeline for recovering the measurement graph and latent DAG (Algorithm 1).

Next, we describe the pipeline in detail. In Stage I, we apply a penalized maximum likelihood estimator to obtain $(\hat{\mathbf{p}}, \hat{\mathbf{B}}, \hat{\gamma})$ and hence $\hat{\mathbf{Q}}$, as detailed in Section 4.1. In Stage II, we fix a strictly increasing sampling rule $f : \mathbb{Z}_+ \rightarrow \mathbb{Z}_+$ and draw $f(N)$ i.i.d. latent samples from $\hat{\mathbf{p}} = (\hat{p}_{\mathbf{z}})_{\mathbf{z} \in \{0,1\}^K}$ to form the resampled matrix $\hat{\mathbf{Z}} \in \{0,1\}^{f(N) \times K}$. The requirement that f be strictly increasing is imposed purely for notational convenience, and our analysis depends only on the growth rate of $f(N)$ relative to N . In Stage III, we run Greedy Equivalence

Algorithm 1: Discrete Causal Representation Learning: Estimate-Resample-Discovery

Data: \mathbf{X} , K , strictly increasing sampling rule $f : \mathbb{Z}_+ \rightarrow \mathbb{Z}_+$.

Stage I: Parameter Estimation;

Obtain the estimated $(\hat{\boldsymbol{\rho}}, \hat{\mathbf{B}}, \hat{\boldsymbol{\gamma}})$;

Set $\hat{\mathbf{Q}}$ by the support of $\hat{\mathbf{B}}$: $\hat{q}_{j,k} \leftarrow \mathbf{1}(\hat{\beta}_{j,k} \neq 0)$;

Stage II: Latent Resampling from $\hat{\boldsymbol{\rho}}$;

Draw $f(N)$ i.i.d. samples $\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(f(N))} \stackrel{\text{i.i.d.}}{\sim} \hat{\boldsymbol{\rho}}$ and stack them into $\hat{\mathbf{Z}} \in \{0, 1\}^{f(N) \times K}$;

Stage III: Causal Discovery on Latent Space;

Perform a causal discovery method on $\hat{\mathbf{Z}}$ to obtain the estimated $\hat{\mathcal{G}}$;

Output: $(\hat{\boldsymbol{\rho}}, \hat{\mathbf{B}}, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{Q}}, \hat{\mathcal{G}})$.

Search (GES; [Chickering, 2002](#)) on $\hat{\mathbf{Z}}$ to obtain an estimate $\hat{\mathcal{G}}$; details on GES are given in Section 4.2. In Section 4.3, we specify suitable ranges for the resampling size $f(N)$ and conditions to ensure that this three-stage pipeline enjoys rigorous consistency guarantees.

While our identifiability results hold for the all-effect specification, in practice we focus on the main-effect specification for parsimony, interpretability, and computational tractability.

4.1 Penalized likelihood estimation via Gibbs–SAEM

Let $\ell(\boldsymbol{\Theta} \mid \mathbf{X}) = \sum_{i=1}^N \log \mathbb{P}(\mathbf{X}_i \mid \boldsymbol{\Theta})$ be the marginal log-likelihood from (2), and define

$$\hat{\boldsymbol{\Theta}} \in \arg \max_{\boldsymbol{\Theta}} \left\{ \ell(\boldsymbol{\Theta} \mid \mathbf{X}) - p_{\lambda_N, \tau_N}(\mathbf{B}) \right\}, \quad (3)$$

where p_{λ_N, τ_N} is an entrywise penalty extended additively over $\mathbf{B} = (\beta_{j,k})$: $p_{\lambda_N, \tau_N}(\mathbf{B}) = \sum_{k=1}^K \sum_{j=1}^J p_{\lambda_N, \tau_N}(\beta_{j,k})$. The bipartite matrix is then estimated by thresholding:

$$\hat{q}_{j,k} = \mathbf{1}(\hat{\beta}_{j,k} \neq 0), \quad j \in [J], k \in [K]. \quad (4)$$

Throughout, we assume that the penalty p_{λ_N, τ_N} satisfies the regularity conditions in [Lee and Gu \(2025, Supplement S.1.4\)](#), including standard truncated sparsity-inducing penalties such as the truncated Lasso penalty (TLP) ([Shen et al., 2012](#)) and the smoothly clipped absolute deviation (SCAD) penalty ([Fan and Li, 2001](#)), and we do not reproduce them here

The following result establishes consistency of the parameters and the bipartite graph.

Theorem 3 (Theorem 3 in Lee and Gu (2025)). Let $(\Theta^*, \mathbf{Q}^*, \mathcal{G}^*)$ denote the true parameters in the discrete CRL framework. Suppose the parameter space is compact and all entries of \mathbf{B} are bounded, the data-generating process at Θ^* is identifiable and has nonsingular Fisher information, and the tuning parameters satisfy $1/\sqrt{N} \ll \tau_N \ll \lambda_N/\sqrt{N} \ll 1$. If $\hat{\Theta}$ solves (3), then there exists a relabeling $\tilde{\Theta} \sim_{\mathcal{K}} \hat{\Theta}$ such that $\|\tilde{\Theta} - \Theta^*\| = O_p(1/\sqrt{N})$. Moreover, with $\tilde{\mathbf{Q}}$ computed from $\tilde{\Theta}$ via (4), one has $\mathbb{P}(\tilde{\mathbf{Q}} \neq \mathbf{Q}^*) \rightarrow 0$.

We compute (3) via a penalized Gibbs–SAEM algorithm (Delyon et al., 1999; Kuhn and Lavielle, 2004; Lee and Gu, 2025). A direct EM implementation would require evaluating the conditional expectation of the complete-data log-likelihood by summing over all 2^K latent configurations. This leads to an $O(NJ2^K)$ cost per outer iteration, which becomes prohibitive once K is moderate. Gibbs–SAEM sidesteps this by replacing the exact E–step with low-cost Gibbs updates. We summarize the resulting algorithm in Algorithm 2.

In the E–step we run an alternating coordinate Gibbs sampler (Gu and Xu, 2023) targeting the posterior $\mathbb{P}(\mathbf{Z} \mid \mathbf{X}; \Theta^{[t]})$ under the parameter iterate $\Theta^{[t]} = (\boldsymbol{\beta}^{[t]}, \boldsymbol{\gamma}^{[t]}, \mathbf{p}^{[t]})$. In our implementation we take a single Gibbs sweep per outer iteration ($C = 1$) following Delyon et al. (1999). Within this sweep we visit each coordinate $Z_{i,k}$ in turn. Conditional on the current value of all other coordinates $\mathbf{Z}_{i,-k}$, the conditional distribution of $Z_{i,k}$ is Bernoulli with success probability $\mathbb{P}(Z_{i,k} = 1 \mid \mathbf{Z}_{i,-k}, \mathbf{Y}; \Theta^{[t]}) \propto \mathbb{P}(Z_{i,k} = 1, \mathbf{Z}_{i,-k}; \mathbf{p}^{[t]}) \prod_{j=1}^J \mathbb{P}(X_{ij} \mid \mathbf{Z}_i; \boldsymbol{\beta}_j^{[t]}, \gamma_j^{[t]})$, and similarly for $Z_{i,k} = 0$. Equivalently, the log-odds of $Z_{i,k} = 1$ versus 0 is given by the difference in log joint densities when flipping that bit. This conditional has a closed form and is straightforward to sample from. To keep each flip inexpensive, we maintain for every sample–item pair the linear score $\psi_{ij} = \beta_{j,0}^{[t]} + \sum_{k=1}^K \beta_{j,k}^{[t]} Z_{ik}$, so that the likelihood contribution of \mathbf{Z}_i to X_{ij} can be updated by rank-one changes in ψ_{ij} . Flipping a single bit $Z_{i,k}$ modifies all ψ_{ij} by adding or subtracting $\beta_{j,k}^{[t]}$, which costs $O(J)$ operations. A full Gibbs sweep visits all K coordinates for each of the N individuals, so the E–step costs $O(NJK)$ per outer iteration, in contrast to the $O(NJ2^K)$ cost of exact EM when K grows.

The Gibbs sweep produces an updated latent matrix $\mathbf{Z}^{[t+1]}$, inducing an empirical distribution on $\{0, 1\}^K$ that we use to update \mathbf{p} via Robbins–Monro stochastic approximation. Writing $\hat{\mathbf{p}}^{[t+1]}$ for the empirical proportions of the sampled configurations at iteration $t + 1$, we set $\mathbf{p}^{[t+1]} = (1 - \theta_{t+1})\mathbf{p}^{[t]} + \theta_{t+1}\hat{\mathbf{p}}^{[t+1]}$, with stepsizes $\{\theta_t\}_{t \geq 1}$ satisfying $\sum_t \theta_t = \infty$ and $\sum_t \theta_t^2 < \infty$.

In the M–step, for each row j of \mathbf{B} we set

$$F_j^{[t+1]}(\boldsymbol{\beta}_j, \gamma_j) = (1 - \theta_{t+1})F_j^{[t]}(\boldsymbol{\beta}_j, \gamma_j) + \frac{\theta_{t+1}}{C} \sum_{r=1}^C \sum_{i=1}^N \log \mathbb{P}(X_{ij} \mid \mathbf{Z}_i = \mathbf{Z}_i^{[t+1],r}; \boldsymbol{\beta}_j, \gamma_j),$$

and then solve the penalized maximization $(\boldsymbol{\beta}_j^{[t+1]}, \gamma_j^{[t+1]}) = \arg \max_{\boldsymbol{\beta}_j, \gamma_j} \left\{ F_j^{[t+1]}(\boldsymbol{\beta}_j, \gamma_j) - p_{\lambda_N, \tau_N}(\boldsymbol{\beta}_j) \right\}$. Together, the Gibbs E–step and penalized SAEM M–step provide an efficient stochastic optimization scheme for the penalized objective (3).

In complex latent variable models, the penalized SAEM algorithm may converge to local maxima if poorly initialized. To mitigate this, we adopt a spectral initialization that first applies the universal singular value thresholding (Chatterjee, 2015) procedure for low-rank generalized linear factor models (Zhang et al., 2020), followed by an SVD-based Varimax rotation to obtain sparse factor loadings (Rohe and Zeng, 2023). See Section S.9.2 for details.

Algorithm 2 yields estimates of \mathbf{p} and \mathbf{Q} , which suffices to infer the latent causal structure. It remains to sample latent variables from $\hat{\mathbf{p}}$ and apply a causal discovery method. In this work, we employ GES and therefore first review its relevant details.

4.2 Greedy Equivalence Search

Greedy Equivalence Search (GES) is a score-based causal discovery method that searches over Markov equivalence classes (MECs) and is guaranteed to identify the MEC of the true DAG under suitable conditions on the scoring criterion. This subsection briefly reviews these conditions, as they will guide our choice of score in the presence of estimation error from $\hat{\mathbf{p}}$.

Let \mathcal{D} denote N i.i.d. samples from a distribution p^* that is faithful to DAG G^* . In a

Algorithm 2: Penalized Gibbs–SAEM Algorithm

Data: \mathbf{X}, K , tuning parameters λ_N, τ_N , stepsizes $\{\theta_t\}_{t \geq 1}$, number of Gibbs sweeps $C \geq 1$.

Initialize $\mathbf{Z}^{[0]}, \Theta^{[0]} = (\mathbf{p}^{[0]}, \boldsymbol{\beta}^{[0]}, \boldsymbol{\gamma}^{[0]})$, $F_j^{[0]} \equiv 0$ for $j \in [J]$; set $t \leftarrow 0$.

while $\|\Theta^{[t+1]} - \Theta^{[t]}\|$ is larger than a threshold **do**

$\mathbf{Z}_{\text{cur}} \leftarrow \mathbf{Z}^{[t]}$;

for $r \in [C]$ **do**

for $i \in [N]$ **do**

for k in a random permutation of $\{1, \dots, K\}$ **do**

$\mathbf{z}^{(1)} \leftarrow (1, \mathbf{Z}_{\text{cur}, i, -k}), \quad \mathbf{z}^{(0)} \leftarrow (0, \mathbf{Z}_{\text{cur}, i, -k})$;

$\Delta \ell_{ik} \leftarrow \log \mathbf{p}^{[t]}(\mathbf{z}^{(1)}) - \log \mathbf{p}^{[t]}(\mathbf{z}^{(0)}) + \sum_{j=1}^J \log \frac{\mathbb{P}(X_{ij} \mid \mathbf{z}^{(1)}; \boldsymbol{\beta}_j^{[t]}, \gamma_j^{[t]})}{\mathbb{P}(X_{ij} \mid \mathbf{z}^{(0)}; \boldsymbol{\beta}_j^{[t]}, \gamma_j^{[t]})}$;

$\mathbf{Z}_{\text{cur}, i, k} \sim \text{Bernoulli}(\text{expit}(\Delta \ell_{ik}))$;

end

end

$\mathbf{Z}^{[t+1], r} \leftarrow \mathbf{Z}_{\text{cur}}$;

end

$\mathbf{Z}^{[t+1]} \leftarrow \mathbf{Z}^{[t+1], C}$;

$\hat{\mathbf{p}}^{[t+1]}(\mathbf{z}) \leftarrow \frac{1}{CN} \sum_{r=1}^C \sum_{i=1}^N \mathbf{1}\{\mathbf{Z}_{i, \cdot}^{[t+1], r} = \mathbf{z}\}$ for $\mathbf{z} \in \{0, 1\}^K$;

$\mathbf{p}^{[t+1]}(\mathbf{z}) \leftarrow (1 - \theta_{t+1}) \mathbf{p}^{[t]}(\mathbf{z}) + \theta_{t+1} \hat{\mathbf{p}}^{[t+1]}(\mathbf{z})$;

for $j \in [J]$ **do**

$\hat{F}_j^{[t+1]}(\boldsymbol{\beta}_j, \gamma_j) \leftarrow \frac{1}{C} \sum_{r=1}^C \sum_{i=1}^N \log \mathbb{P}(X_{ij} \mid \mathbf{Z}_i = \mathbf{Z}_i^{[t+1], r}; \boldsymbol{\beta}_j, \gamma_j)$;

$F_j^{[t+1]}(\boldsymbol{\beta}_j, \gamma_j) \leftarrow (1 - \theta_{t+1}) F_j^{[t]}(\boldsymbol{\beta}_j, \gamma_j) + \theta_{t+1} \hat{F}_j^{[t+1]}(\boldsymbol{\beta}_j, \gamma_j)$;

$(\boldsymbol{\beta}_j^{[t+1]}, \gamma_j^{[t+1]}) \leftarrow \arg \max_{\boldsymbol{\beta}_j, \gamma_j} \{F_j^{[t+1]}(\boldsymbol{\beta}_j, \gamma_j) - p_{\lambda_N, \tau_N}(\boldsymbol{\beta}_j)\}$;

end

$t \leftarrow t + 1$;

end

Let $\hat{\mathbf{p}} \leftarrow \mathbf{p}^{[T]}$ and $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \leftarrow (\boldsymbol{\beta}^{[T]}, \boldsymbol{\gamma}^{[T]})$ at convergence;

Estimate the measurement graph \mathbf{Q} from the sparsity pattern of (4);

Output: $\hat{\Theta} = (\hat{\mathbf{p}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ and \mathbf{Q} .

score-based framework, each DAG G is assigned a score $S(G; \mathcal{D})$, and the estimation problem is formulated as $G^* \in \arg \max_G S(G; \mathcal{D})$. Based on a scoring criterion, GES performs a two-phase greedy search over equivalence classes represented by CPDAGs: a forward phase that repeatedly applies a valid insertion whenever it increases the score, followed by a backward

phase that greedily applies valid deletions until no further score increase is possible.

We recall several key properties of scoring criteria. The theoretical guarantees of GES rely on the following standard notions (Chickering, 2002): score equivalence, decomposability, consistency and local consistency.

First, a score is score equivalent if it gives identical scores to all DAGs in the same MEC.

Second, decomposability requires that the score can be expressed as $S(G, \mathcal{D}) = \sum_{i=1}^n s(R_i, \mathbf{Pa}_i^G)$ where each local term is a function only of R_i and its parent set \mathbf{Pa}_i^G .

Third, the score is said to be consistent if, in the limit as $N \rightarrow \infty$, we have $S(H, \mathcal{D}) > S(G, \mathcal{D})$ whenever $p^* \in \mathcal{M}(H)$ and $p^* \notin \mathcal{M}(G)$, and $S(H, \mathcal{D}) < S(G, \mathcal{D})$ whenever $p^* \in \mathcal{M}(H) \cap \mathcal{M}(G)$ and G contains fewer parameters than H .

Finally, the score is locally consistent if, when G' is obtained from G by adding an edge $u \rightarrow v$, we have in the limit as $N \rightarrow \infty$ that $S(G', \mathcal{D}) > S(G, \mathcal{D})$ if $R_v \not\perp_{p^*} R_u \mid \mathbf{Pa}_v^G$, and $S(G', \mathcal{D}) < S(G, \mathcal{D})$ if $R_v \perp_{p^*} R_u \mid \mathbf{Pa}_v^G$.

A score commonly used in score-based methods is BIC. The BIC score for graphical models (Schwarz, 1978; Koller and Friedman, 2009) is defined as $S(G; \mathcal{D}) = \log p_{\hat{\theta}}(\mathcal{D}) - \frac{1}{2}d_G \log N$, where $\hat{\theta}$ is the maximum likelihood estimate over $\mathcal{M}(G)$ and $d_G = \dim(\Theta(G))$ denotes the number of free parameters in the model associated with G . Since we focus on discrete Bayesian networks, we also consider the BDeu (Bayesian Dirichlet equivalent uniform) score, defined as the log marginal likelihood under a uniform Dirichlet prior on each conditional probability table (Heckerman et al., 1995), which is score equivalent. For a fixed equivalent sample size, the BDeu score and BIC differ only by an $O(1)$ term (Koller and Friedman, 2009), so they are asymptotically equivalent. For the remainder of the paper, we adopt the BDeu score to recover \mathcal{G} in the proposed causal discrete measurement models.

If a score-equivalent criterion is locally consistent, GES returns the MEC of G^* as $N \rightarrow \infty$ (Chickering, 2002, Theorem 4). Moreover, BDeu is locally consistent for discrete Bayesian networks (Chickering, 2002), guaranteeing that GES recovers the MEC of G^* as $N \rightarrow \infty$.

4.3 Theoretical Guarantees

Our main result is stated as follows.

Theorem 4. *Consider the discrete causal representation learning framework with true parameters $(\mathbf{p}^*, \mathcal{G}^*, \mathbf{B}^*, \mathbf{Q}^*, \gamma^*)$. Suppose Assumption 1 holds, the framework is identifiable up to $\sim_{\mathcal{K}}$, the Fisher information at Θ^* is nonsingular, and the entries of \mathbf{B} are uniformly bounded. Assume that Stage I of Algorithm 1 employs the penalized optimization problem (3) with tuning parameters (τ_N, λ_N) satisfying $\frac{1}{\sqrt{N}} \ll \tau_N \ll \frac{\lambda_N}{\sqrt{N}} \ll 1$, and that Stage III applies Greedy Equivalence Search with the BDeu score to the resampled latent data $\widehat{\mathbf{Z}}$. Further assume that the sampling rule $f : \mathbb{Z}_+ \rightarrow \mathbb{Z}_+$ is strictly increasing and satisfies $f(N) = o(N \log N)$. Then, as $N \rightarrow \infty$:*

- (i) *There exists $\tilde{\Theta} \sim_{\mathcal{K}} (\widehat{\mathbf{p}}, \widehat{\mathbf{B}}, \widehat{\gamma})$ such that $\|\tilde{\Theta} - \Theta^*\| = O_p(N^{-1/2})$ and $\mathbb{P}(\widehat{\mathbf{Q}} = \mathbf{Q}^*) \rightarrow 1$.*
- (ii) *$\widehat{\mathcal{G}}$ recovers the MEC of \mathcal{G}^* with probability tending to 1.*

In practice, we specifically employ Penalized Gibbs–SAEM (Algorithm 2) to solve (3) and thereby complete Stage I of Algorithm 1. Below, we provide a detailed account of how the theorem is established and highlight the main technical challenges involved in the argument.

Because Stage III in Algorithm 1 applies GES to resampled latents drawn from an estimated law rather than the unobserved p^* , our setting is more complex than Chickering (2002), where the data are drawn directly from a fixed p^* . Classical consistency and local consistency must therefore be strengthened to tolerate sampling from a sequence $\{p_N\}$ approaching p^* at a specified rate. This refinement is crucial for applying GES to $\widehat{\mathbf{Z}}$ rather than the unobserved \mathbf{Z} . We now introduce the rate-robust notions needed for this purpose.

Definition 4. *Let $p^* = P_{\theta^*}$ and $\mathcal{D}_N = \{\mathbf{R}_{N,1}, \dots, \mathbf{R}_{N,N}\}$ be i.i.d. from some $p_N = P_{\theta_N}$. We say a score S is $\{c_N\}$ -consistent if $\|\theta_N - \theta^*\| = O_p(1/c_N)$ with $c_N \rightarrow \infty$, and as $N \rightarrow \infty$ the following hold: (i) if $p^* \in \mathcal{M}(H)$ and $p^* \notin \mathcal{M}(G)$, then $S(H, \mathcal{D}_N) > S(G, \mathcal{D}_N)$ with probability $\rightarrow 1$; (ii) if $p^* \in \mathcal{M}(H) \cap \mathcal{M}(G)$ and G contains fewer parameters than H , then $S(G, \mathcal{D}_N) > S(H, \mathcal{D}_N)$ with probability $\rightarrow 1$.*

By Lemma 7 in [Chickering \(2002\)](#), a decomposable consistent score is locally consistent; hence $\{c_N\}$ -consistency and decomposability yield $\{c_N\}$ -local consistency.

Definition 5. Let $p^\star = P_{\boldsymbol{\theta}^\star}$ and $\mathcal{D}_N = \{\mathbf{R}_{N,1}, \dots, \mathbf{R}_{N,N}\}$ be i.i.d. from some $p_N = P_{\boldsymbol{\theta}_N}$. Let G' be obtained from G by adding the edge $i \rightarrow j$. We say a score S is $\{c_N\}$ -locally consistent if $\|\boldsymbol{\theta}_N - \boldsymbol{\theta}^\star\| = O_p(1/c_N)$ with $c_N \rightarrow \infty$, and as $N \rightarrow \infty$ the following hold: (i) if $R_j \not\perp_{p^\star} R_i \mid \mathbf{Pa}_j^G$, then $S(G', \mathcal{D}_N) > S(G, \mathcal{D}_N)$ with probability $\rightarrow 1$; (ii) if $R_j \perp_{p^\star} R_i \mid \mathbf{Pa}_j^G$, then $S(G', \mathcal{D}_N) < S(G, \mathcal{D}_N)$ with probability $\rightarrow 1$.

If a score-equivalent score criterion S is $\{c_N\}$ -locally consistent and the sampling law \mathbf{p}_N used for resampling remains within the prescribed $\{c_N\}$ -local tolerance of \mathbf{p}^\star , then it can be shown that applying GES with S to samples \mathcal{D}_N drawn from \mathbf{p}_N still returns the MEC of G^\star as N increases. Our first task is therefore to identify a $\{c_N\}$ -locally consistent criterion; a standard choice already does so. The next theorem places BDeu in our $\{c_N\}$ framework.

Theorem 5. The BDeu score is $\{c_N\}$ -consistent for discrete causal graphical models, where $c_N = \omega(\sqrt{\frac{N}{\log N}})$.

Since BDeu is decomposable, we immediately have:

Corollary 2. Using the notation in [Theorem 5](#), the BDeu score is $\{c_N\}$ -locally consistent for discrete causal graphical models, where $c_N = \omega(\sqrt{\frac{N}{\log N}})$.

Leveraging this key property, we apply GES with the BDeu score to the resampled latent data. Concretely, in our pipeline we first use the N observed samples to construct an estimator $\tilde{\boldsymbol{\theta}}_N$ of the true $\boldsymbol{\theta}^\star$, then draw $f(N)$ samples from $P_{\tilde{\boldsymbol{\theta}}_N}$ and run GES scored by BDeu on these resamples. For ease of further discussion, assume the estimator obeys the rate $\|\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^\star\| = O_p(1/g(N))$, where $g : \mathbb{Z}_+ \rightarrow \mathbb{R}$ is strictly increasing with $g(N) \rightarrow \infty$.

Under the indexing convention of [Definitions 4](#) and [5](#), the distribution generating \mathcal{D}_N should be labeled by the sample size that produced it; in particular, it is an estimate from $f^{-1}(N)$ samples rather than from N . Equivalently, $\tilde{\boldsymbol{\theta}}_{f^{-1}(N)} = \boldsymbol{\theta}_N$, so $\|\boldsymbol{\theta}_N - \boldsymbol{\theta}^\star\| =$

$O_p(1/g(f^{-1}(N)))$. Invoking Corollary 2, it thus suffices that $g(f^{-1}(N)) = \omega(\sqrt{\frac{N}{\log N}})$ to guarantee that GES on the new data recovers the correct MEC. Recall that Theorem 3 gives $\|\widehat{\mathbf{p}} - \mathbf{p}^*\| = O_p(N^{-1/2})$. Combining this $N^{-1/2}$ rate with $g(f^{-1}(N)) = \omega(\sqrt{\frac{N}{\log N}})$ yields $f^{-1}(N) = \omega(N/\log N)$, which holds whenever $f(N) = o(N \log N)$. This determines an appropriate sampling rule $f(N)$ guaranteeing consistency for the causal graph.

5 Simulation Studies

We conduct simulation studies across diverse settings to evaluate the proposed pipeline (Algorithm 1), using Algorithm 2 in Stage I. We begin with the comparison experiments and the generative setup shared across all simulations. These settings are chosen to satisfy the strict identifiability conditions in Theorem S.1 and create scenarios of varying difficulty for structure recovery. In all simulations, we use three measurement families—Gaussian, Poisson, and Bernoulli—to represent continuous, count, and binary observations, respectively. The exact constructions of two possible measurement graphs \mathbf{Q}_1 , \mathbf{Q}_2 , and parameters \mathbf{p} , (\mathbf{B}, γ) are given in Supplement S.9.3. For the resampling step, we fix $f(N) = N$ in all main-text experiments, while Supplement S.9.4 reports a sensitivity analysis over $f(N) \in \{2N, 3N\}$.

Simulation Study I: Comparison experiments. We benchmark our method against the influential work of Kivva et al. (2021), which proposed a theoretically well-founded framework for learning DAGs with discrete latent variables. Its publicly available implementation currently supports latent dimension only up to $K \leq 5$, so in this comparison we focus on Gaussian models with $K = 5$. For these comparison runs we set $\mathbf{Q} = \mathbf{Q}_1$. We adopt three standard benchmark latent DAGs with $K = 5$ shown in Figure 2.

We fix $f(N) = N$ with $N \in \{1000, 5000, 10000\}$. Structural Hamming Distance (SHD) is computed on the full composite graph obtained by combining the latent DAG \mathcal{G} with the bipartite measurement graph \mathbf{Q} . Since the composite graph contains both edges between the latent variables and edges in the bipartite graph, the absolute SHD values can be numerically large, yet the comparison remains fair because both methods are evaluated on the same

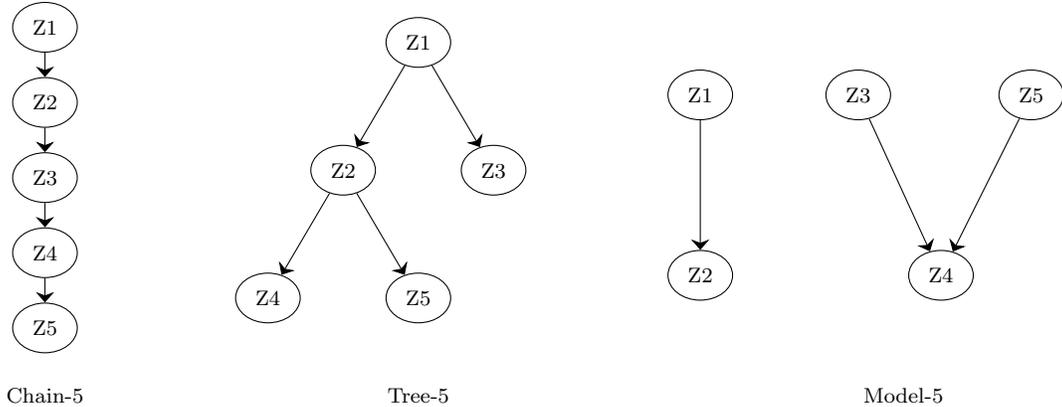


Figure 2: Simulation benchmarks for the comparison experiments

composite graph. Table 1 summarizes results from 100 independent replicates and shows that our method substantially outperforms [Kivva et al. \(2021\)](#).

	Proposed DCRL			Mixture-Oracle		
	1000	5000	10000	1000	5000	10000
Chain-5	0.38	0.02	0	23.66	22.68	21.7
Tree-5	0.47	0.07	0	23.16	22.3	21.94
Model-5	1.42	0.06	0	22.6	22.38	22.37

Table 1: SHD on the composite graph $\mathcal{G} \cup \mathbf{Q}$ under two methods with $f(N) = N$; penalized Gibbs-SAEM attains far smaller SHD across all settings and improves to near-perfect recovery as N grows.

To interpret these results, recall that [Kivva et al. \(2021\)](#) formulate their identifiability theory at a high level of generality: they do not assume a specific likelihood, nor fix K or the number of categories in advance. Instead, such information is, in principle, recoverable from the geometry of the mixture distribution over \mathbf{X} . In practice, this is implemented via a mixture oracle that is approximated by clustering on the observed space, effectively treating each latent configuration as a separate mixture component and thus requiring the recovery of up to 2^K clusters on the full \mathbf{X} . For moderate or large K , especially when the mixture components are only weakly separated, this reliance on clustering becomes statistically fragile. In addition to these statistical challenges, Theorem 4.8 of [Kivva et al. \(2021\)](#) shows that even the subroutine for recovering the bipartite graph Γ has complexity

$O(N^4)$, further limiting the applicability of their implementation to relatively small sample sizes and latent dimensions, consistent with the constraint $K \leq 5$ in the released code.

By contrast, we work within a more structured but still flexible framework: we posit a sparse measurement graph and an explicit item-wise likelihood for $\mathbf{X} \mid \mathbf{Z}$. This additional structure allows us to exploit the likelihood via a penalized SAEM algorithm, which leads to substantially more accurate estimation in the weakly separated regimes considered here.

Simulation Study II: Larger K and more challenging settings. We next consider larger latent dimensions and a broader collection of latent DAGs, illustrated in Figure 3. The five benchmark latent DAGs in Figure 3 (Chain-10, Tree-10, Model-7, Model-8, and Model-13) have latent dimensions $K = 10, 10, 7, 8,$ and 13 , respectively. The latent dimension K here is substantially larger than in Kivva et al. (2021); Huang et al. (2022); Prashant et al. (2025), adding to the difficulty of accurate recovery. All three distributional types (Gaussian, Poisson, Bernoulli) and both \mathbf{Q}_1 and \mathbf{Q}_2 are considered.

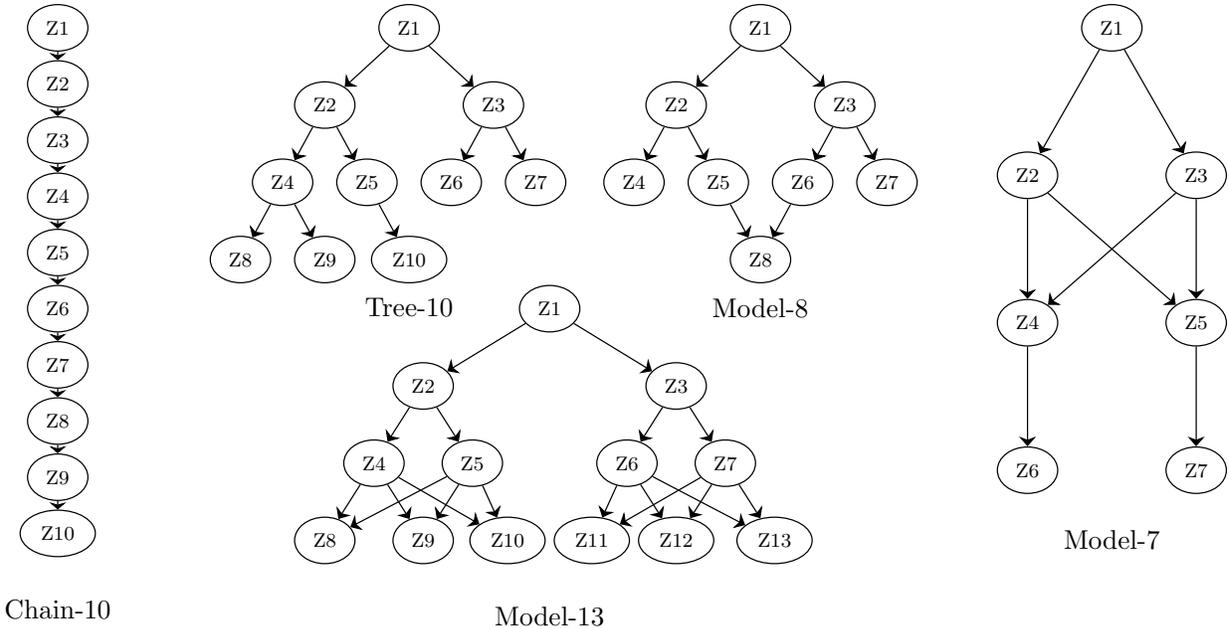


Figure 3: True latent DAGs for Simulation Study II.

We conduct 500 independent replicates in each simulation setting, and report in Table 2 the average SHD computed on the full composite graph $\mathcal{G} \cup \mathbf{Q}$. Table 2 reveals a clear pattern:

Model	\mathbf{Q}	#edges	Bernoulli			Poisson			Gaussian		
			N			N			N		
			3000	5000	7000	3000	5000	7000	3000	5000	7000
Chain-10	\mathbf{Q}_1	57	5.55	3.294	2.938	2.248	1.362	0.638	0.412	0.254	0.122
	\mathbf{Q}_2	73	5.664	3.258	3.042	2.174	0.704	0.488	0.406	0.208	0.146
Tree-10	\mathbf{Q}_1	57	4.372	2.308	1.308	1.85	1.692	1.36	0.94	0.51	0.308
	\mathbf{Q}_2	73	4.3	1.936	1.26	2.676	1.556	1.146	1.14	0.618	0.346
Model-7	\mathbf{Q}_1	41	7.692	6.334	5.798	5.838	5.422	4.892	0.196	0.042	0
	\mathbf{Q}_2	51	7.554	6.304	5.68	5.848	5.526	5.082	0.262	0.054	0.004
Model-8	\mathbf{Q}_1	46	4.336	2.682	2.19	2.106	1.916	1.878	0.132	0.048	0
	\mathbf{Q}_2	58	4.342	2.72	2.374	2.264	1.9	1.782	0.202	0.052	0.002
Model-13	\mathbf{Q}_1	81	22.37	16.454	14.062	24.65	16.472	14.162	3.206	1.646	1.008
	\mathbf{Q}_2	103	22.252	16.29	14.606	25.134	15.626	12.934	3.032	1.872	0.994

Table 2: Average SHD on the composite graph $\mathcal{G} \cup \mathbf{Q}$ with $f(N) = N$; SHD decreases with N across all designs and is smallest for Gaussian, intermediate for Poisson, and largest for Bernoulli. The column #edges reports the number of edges in the composite graph.

Bernoulli data are the most challenging, followed by Poisson, and Gaussian is the easiest. This is consistent with the intuition that discrete observations carry less information, which is why most existing simulation studies focus on the continuous Gaussian case. Although some of the reported SHD values may look large in absolute terms, they should be interpreted relative to the size of the underlying graph: the composite object contains not only the latent DAG \mathcal{G} but also the bipartite layer induced by \mathbf{Q} , which contributes a substantial number of edges. For instance, in the Model-13 design, the target structure consists of a DAG on 13 latent variables together with a bipartite graph between 13 latent and 39 observed variables; when $\mathbf{Q} = \mathbf{Q}_1$ the resulting composite graph already contains 81 edges, and when $\mathbf{Q} = \mathbf{Q}_2$ it contains 103 edges. Consequently, even a small relative error can translate into a seemingly large SHD. Viewed in this light, the results in Table 2 already indicate accurate recovery across all three measurement families. Moreover, within each simulation configuration, the average SHD decreases systematically as the sample size N increases, which provides empirical support for our identifiability theory.

6 Applications to Educational Data and Image Data

We evaluate our method on an educational assessment dataset and a synthetic ball-image dataset. In the educational dataset, we examine whether the recovered causal structure among latent cognitive skills is interpretable. The image dataset is a high-dimensional benchmark with a known generative process and ground-truth latent DAG, inspired by “balls” image setups in causal representation learning (Ahuja et al., 2023, 2024). The image data allows us to assess whether our method simultaneously recovers the true causal relationships and learns interpretable latent representations from high-dimensional observations.

6.1 TIMSS 2019 Response Time Data

We apply DCRL to response time data from the TIMSS 2019 eighth-grade mathematics assessment for U.S. students (Fishbein et al., 2021), recording each student’s time (in seconds) on each item screen. The assessment evaluates seven skills: four content skills (“Number”, “Algebra”, “Geometry”, “Data and Probability”) and three cognitive skills (“Knowing”, “Applying”, “Reasoning”). We follow the preprocessing steps of Lee and Gu (2024), but pursue the more demanding goal of recovering causal relationships among the latent skills.

We consider students who received booklet 14. Following Lee and Gu (2024), we log-transform and truncate response times to mitigate outlier influence, yielding a dataset of $N = 620$ students and $J = 29$ items. We fit this dataset under a lognormal observation specification within the DCRL framework.

Since the TIMSS 2019 database already specifies which skills are assessed by each item, the skill-item association matrix \mathbf{Q} is known and can be directly constructed as a 29×7 binary matrix (see Table S.4). Notably, this matrix satisfies the identifiability conditions in Corollary S.4, ensuring that the latent causal structure can be uniquely recovered up to label permutations and Markov equivalence. Therefore, we accordingly make minor modifications to our previous algorithm: we eliminate the sparsity-inducing penalty term and restrict

each latent variable to be connected only by those items that are designed to measure it, as determined by the known matrix \mathbf{Q} . Moreover, the screen-level response time matrix contains some missing entries. Similar to Lee and Gu (2025), we treat all such entries as missing at random. Algorithmically, objective functions are computed by summing only over person-item cells with observed response times.

With these modifications, the algorithm outputs the causal structure as a CPDAG (Figure 4), since the DAG is identifiable only up to its Markov equivalence class.

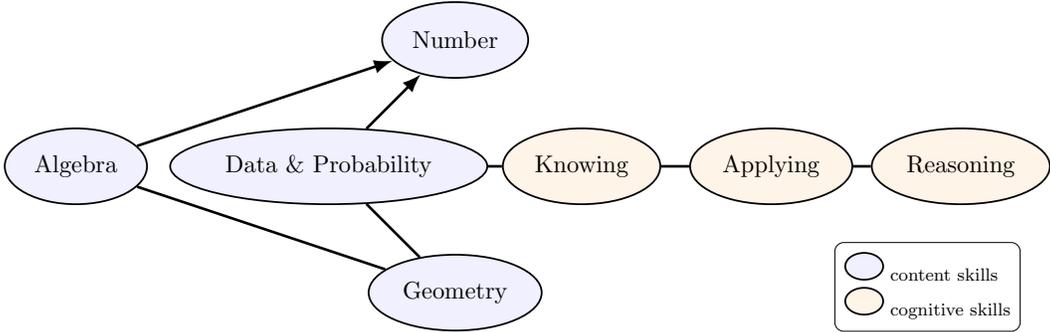


Figure 4: Learned Causal Relationships Among Seven Latent Skills

The recovered causal graph aligns with cognitive expectations and curriculum structure. The three cognitive skills—“Knowing”, “Applying”, and “Reasoning”—form a directed path, consistent with the progressive nature of cognitive processing. Among the content skills, “Number” is foundational and shows strong links to both “Algebra” and “Data and Probability”, reflecting their shared reliance on numerical reasoning. Although “Geometry” typically requires less numerical computation, it remains connected to both “Algebra” and “Data and Probability”, suggesting overlapping problem-solving strategies. Among these four content skills, “Data and Probability” is likely the most comprehensive, as it requires a broad range of skills, making it directly linked to the three cognitive skills. This supports the interpretation that tasks involving data interpretation demand a combination of factual knowledge, application, and reasoning, positioning it as an integrative skill in mathematical cognition.

6.2 Ball Image Data

We build a “seesaw + occlusion” experiment in which each latent variable represents a visibility, presence, or configuration event in the observed image. Two binary variables indicate the presence of balls on two well-separated slots of a forked tray, which acts as a load on the right side of a seesaw. Let $Z_1, Z_2 \in \{0, 1\}$ denote the presence indicators of these two tray balls. The seesaw’s mechanical response determines whether the left-side ball rises to an “up” configuration, denoted $Z_3 \in \{0, 1\}$. To reflect natural heterogeneity in physical conditions (e.g., slight variations in ball weights or instrument failures), we model this mechanism stochastically rather than deterministically by setting $\mathbb{P}(Z_3 = 1 \mid Z_1 = 1, Z_2 = 1) = 0.8$ and $\mathbb{P}(Z_3 = 1 \mid (Z_1, Z_2) \neq (1, 1)) = 0.2$, so that the tray balls increase the probability of the “rise” event without forcing it.

Finally, we introduce a fourth ball that is physically present but typically occluded by the left ball when the seesaw is in the down configuration. When the seesaw rises ($Z_3 = 1$), the fourth ball may become visible; we define $Z_4 \in \{0, 1\}$ as its visibility indicator, with $\mathbb{P}(Z_4 = 1 \mid Z_3 = 1) = 0.99$ and $\mathbb{P}(Z_4 = 1 \mid Z_3 = 0) = 0$. Overall, this construction yields a physically motivated latent causal structure $Z_1, Z_2 \rightarrow Z_3 \rightarrow Z_4$ while keeping the latent variables binary for a principled reason: each Z_k corresponds to a discrete, image-level event (presence, configuration, or visibility).

Figure 5 shows representative images; full rendering and preprocessing details are in Supplement S.9.6. Each rendered grayscale image is converted to a balls-only binary mask, resized to 96×96 , and then pooled to a 16×16 binary image. We fit the model using this pooled representation, so each sample is a 256-dimensional binary vector, where 1 denotes a bright pixel and 0 denotes a dark pixel. We generate 10000 images.

We fit DCRL with $K = 4$ binary latent variables and Bernoulli responses, a high-dimensional setting with $J = 256$ observed dimensions substantially larger than in our simulation studies. The observed variables are modeled as: $X_j \mid \mathbf{Z} \sim \text{Ber}(g_{\text{logistic}}(\beta_{j,0} +$

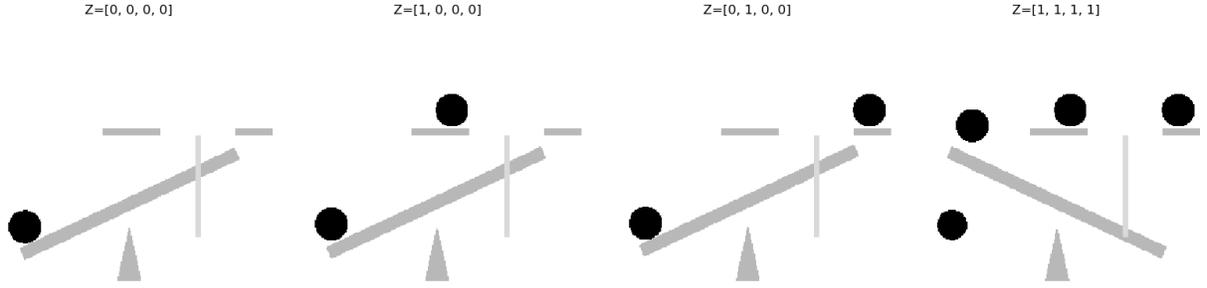


Figure 5: Representative samples from the “seesaw + occlusion” image generator. The tray balls correspond to (Z_1, Z_2) , the up/down state of the left seesaw-side ball corresponds to Z_3 , and the occluded ball corresponds to Z_4 .

$\sum_{k=1}^K \beta_{j,k} Z_k$), where g_{logistic} is the sigmoid function. Our goal is to recover both the latent DAG G and the bipartite structure \mathbf{Q} from the pooled binary data. Although \mathbf{Q} has 256 rows, the estimated $\hat{\mathbf{Q}}$ remains highly sparse. Most rows are either zero vectors (blocks where no ball appears) or nearly one-hot vectors (blocks primarily associated with a single latent variable), which matches the generative design in which most spatial cells contain at most one object. The overall sparse support pattern also satisfies the identifiability conditions in Corollary S.4. The recovered DAG over the latents matches the data-generating mechanism.

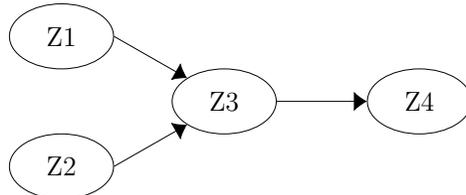


Figure 6: Estimated causal relationships among four latent variables by DCRL. This latent DAG matches the ground-truth causal relations exactly.

Since the learned causal structure matches the data-generating mechanism, we expect each recovered latent coordinate to correspond to a spatially localized event (a tray ball being present, the up/down state of the seesaw-side ball, or the occluded ball becoming visible). To interpret the factors, we visualize the effect of activating one latent coordinate at a time on the pixelwise Bernoulli probabilities in the pooled representation. Recall that $\mathbf{B} \in \mathbb{R}^{J \times (K+1)}$ stacks the intercept and main-effect parameters across the $J = 256$ pixels, so that \mathbf{Bz} gives the logits of the Bernoulli success probabilities for any latent feature vector

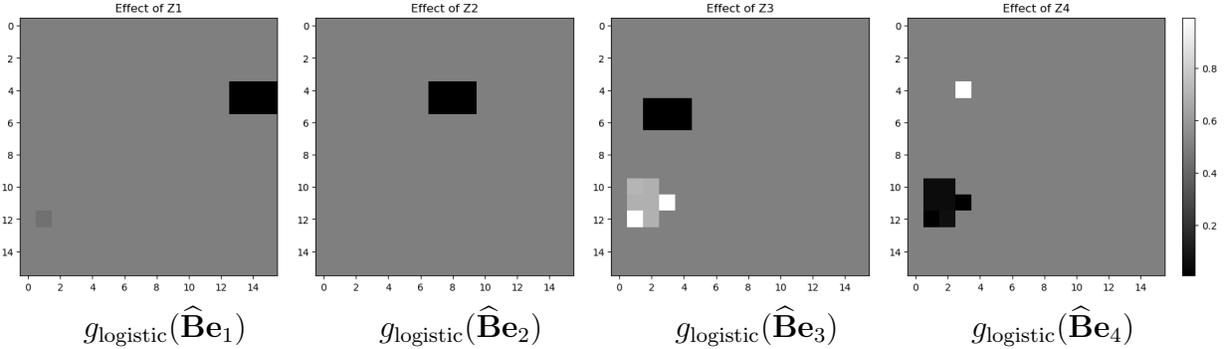


Figure 7: Effect maps obtained by activating one latent coordinate at a time. Mid-gray corresponds to probability 0.5 (no effect). Since the pooled representation is coded as background = 1 and ball = 0, white indicates an increased probability of background (ball less likely), while black indicates a decreased probability of background (ball more likely). The dominant localized regions align with the tray balls, the configuration-dependent movement of the seesaw-side ball, and the occluded ball’s visibility.

$\mathbf{z} \in \mathbb{R}^{K+1}$. For $k \in \{1, 2, 3, 4\}$, we set the intercept to 0 and activate only the k th coordinate, compute $g_{\text{logistic}}(\widehat{\mathbf{B}}\mathbf{e}_k) \in (0, 1)^{256}$, and reshape it into a 16×16 image, where \mathbf{e}_k is the k th standard basis vector. Since $g_{\text{logistic}}(0) = 0.5$, mid-gray indicates no effect, and white indicates an increased probability of background (equivalently, a decreased probability that a ball occupies that cell), while black indicates a decreased probability of background. Figure 7 reports the resulting effect maps. The tray-ball factors appear as localized dark patches at the corresponding tray locations, indicating that activating those coordinates increases the probability of a ball in those cells. The factor for the seesaw-side ball shows a signed bright/dark pattern, reflecting the ball’s up/down state: one location becomes more likely to contain a ball while another becomes less likely. The visibility factor is concentrated near the occluded ball region, with a localized effect consistent with the intended semantics. The maps are not perfectly clean, as expected given the mild positional jitter and simple preprocessing pipeline. Despite these nuisances and the coarse pooled representation, the recovered causal graph and latent factors remain interpretable and align with the data-generating mechanism.

7 Discussion

This paper develops a computationally efficient and provable estimate-resample-discovery pipeline for causal representation learning with discrete latent variables. Our procedure has a clean structure: (i) we estimate the measurement layer and the joint distribution of latent variables via a penalized Gibbs–SAEM algorithm, (ii) we resample pseudo-latent datasets from the fitted latent distribution, and (iii) we perform score-based causal discovery on the resampled latents using GES. Theoretically, we establish strict and generic identifiability for the proposed discrete causal representation learning framework. We prove $\{c_N\}$ -consistency and $\{c_N\}$ -local consistency of BDeu scores in discrete Bayesian networks, and show that, under mild conditions, this estimate-resample-discovery pipeline consistently recovers both the measurement structure and the Markov equivalence class of the latent DAG.

Although our exposition focuses on binary latent variables, the Gibbs–SAEM updates can also be modified for polytomous latent variables. For simplicity and clarity, we focus on the binary case in this paper. Although we state our results for the BDeu score, the same analysis and guarantees apply to BIC, which appears as the leading term in the BDeu expansion in our proofs. Since most score-based methods in practice adopt either BIC or BDeu (Kitson et al., 2023), this BIC-type class already covers the dominant use cases.

Several avenues remain for future work. Incorporating procedures for estimating K , such as information criteria tailored to the latent layer, would make the pipeline automatic and reduce sensitivity to model size. Additionally, since our approach provides a general framework, it is natural to explore replacing the current estimation or causal discovery components with other alternatives (Spirtes et al., 2000; Ramsey et al., 2016). These methods’ empirical performance and theoretical guarantees remain open for future investigation.

References

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. (2023). Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.

- Ahuja, K., Mansouri, A., and Wang, Y. (2024). Multi-domain causal representation learning via weak distributional invariances. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 865–873. PMLR.
- Allman, E. S., Matias, C., and Rhodes, J. A. (2008). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37:3099–3132.
- Anandkumar, A., Hsu, D., Javanmard, A., and Kakade, S. (2013). Learning linear bayesian networks with latent variables. *30th International Conference on Machine Learning, ICML 2013*.
- Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Schölkopf, B., and Ravikumar, P. (2023). Learning linear causal representations from interventions under general nonlinear mixing. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA. Curran Associates Inc.
- Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, pages 177–214.
- Chickering, D. M. (1995). A transformational characterization of equivalent bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, page 87–98.
- Chickering, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554.
- D’Amour, A., Heller, K., Moldovan, D., et al. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61.
- de la Torre, J. (2011). The generalized dina model framework. *Psychometrika*, 76:179–199.
- Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, pages 94–128.
- Dong, X., Ng, I., Dai, H., Sun, J., Song, X., Spirtes, P., and Zhang, K. (2026). Score-based greedy search for structure identification of partially observed linear causal models. In *The Fourteenth International Conference on Learning Representations*.
- Evans, R. J. (2025). Graphical models. University of Oxford.
- Fan, D., Kou, Y., and Gao, C. (2025). Causal flow-based variational auto-encoder for disentangled causal representation learning. *ACM Trans. Intell. Syst. Technol.*, 16(5).
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
- Fishbein, B., Foy, P., and Yin, L. (2021). *TIMSS 2019 User Guide for the International Database*.

- Ghassami, A., Yang, A., Kiyavash, N., and Zhang, K. (2020). Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*.
- Gu, Y. and Xu, G. (2023). A joint MLE approach to large-scale structured latent attribute analysis. *Journal of the American Statistical Association*, 118(541):746–760.
- Hartford, J., Ahuja, K., Bengio, Y., and Sridhar, D. (2023). Beyond the injective assumption in causal representation learning.
- He, S., Culpepper, S. A., and Douglas, J. (2023). *A Sparse Latent Class Model for Polytomous Attributes in Cognitive Diagnostic Assessments*, pages 413–442. Springer International Publishing, Cham.
- Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.
- Helm, C., Warwas, J., and Schirmer, H. (2022). Cognitive diagnosis models of students’ skill profiles as a basis for adaptive teaching: an example from introductory accounting classes. *Empirical Research in Vocational Education and Training*, 14:1–30.
- Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554.
- Huang, B., Low, C. J. H., Xie, F., Glymour, C., and Zhang, K. (2022). Latent hierarchical causal structure discovery with rank constraints. In *Advances in Neural Information Processing Systems*, volume 35, pages 5549–5561.
- Javaloy, A., Martin, P. S., and Valera, I. (2023). Causal normalizing flows: from theory to practice. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Jin, J. and Syrgkanis, V. (2024). Learning linear causal representations from general environments: identifiability and intrinsic ambiguity. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS ’24, Red Hook, NY, USA. Curran Associates Inc.
- Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR, Proceedings of Machine Learning Research, pages 2207–2216. ADDISON-WESLEY PUBL CO.
- Kitson, N., Constantinou, A., Zhigao, G., Liu, Y., and Chobtham, K. (2023). A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56:1–94.
- Kivva, B., Rajendran, G., Ravikumar, P. K., and Aragam, B. (2021). Learning latent causal graphs via mixture oracles. In *Advances in Neural Information Processing Systems*.
- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

- Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131.
- Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.
- Lee, S. and Gu, Y. (2024). New paradigm of identifiable general-response cognitive diagnostic models: Beyond categorical data. *Psychometrika*, 89(4):1304–1336.
- Lee, S. and Gu, Y. (2025). Deep discrete encoders: Identifiable deep generative models for rich data with discrete latent layers. *Journal of the American Statistical Association*, (just-accepted):1–25.
- Liu, J., Lee, S., and Gu, Y. (2025). Exploratory general-response cognitive diagnostic models with higher-order structures. *Psychometrika*, page 1–42.
- Minchen, N. D., de la Torre, J., and Liu, Y. (2017). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics*, 42(6):651–677.
- Mityagin, B. (2015). The zero set of a real analytic function. *Mathematical Notes*, 107:529–530.
- Moran, G. and Aragam, B. (2026). Towards interpretable deep generative models via causal representation learning. *Journal of the American Statistical Association Review*, pages 1–32.
- Nazaret, A. and Blei, D. (2024). Extremely greedy equivalence search. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2716–2745. PMLR.
- Prashant, P., Ng, I., Zhang, K., and Huang, B. (2025). Differentiable causal discovery for latent hierarchical causal models. In *13th International Conference on Learning Representations, ICLR 2025*, pages 23212–23237.
- Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2016). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3:121 – 129.
- Rohe, K. and Zeng, M. (2023). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060.
- Salakhutdinov, R. (2015). Learning deep generative models. *Annual Review of Statistics and Its Application*, 2(1):361–385.
- Salakhutdinov, R. and Hinton, G. (2009). Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455. PMLR.

- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232.
- Shwe, M., Middleton, B., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., and Cooper, G. (1991). Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. i. the probabilistic model and inference algorithms. *Methods of information in medicine*, 30(4):241–255.
- Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.
- Squires, C., Yun, A., Nichani, E., Agrawal, R., and Uhler, C. (2022). Causal structure discovery between clusters of nodes induced by latent factors. In Schölkopf, B., Uhler, C., and Zhang, K., editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 669–687. PMLR.
- Teicher, H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302.
- Varici, B., Acartürk, E., Shanmugam, K., Kumar, A., and Tajer, A. (2025). Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90.
- Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI '90*, page 255–270, USA. Elsevier Science Inc.
- von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. (2023). Nonparametric identifiability of causal representations from unknown interventions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*. Curran Associates Inc.
- Wang, Y., Blei, D., and Cunningham, J. P. (2021). Posterior collapse and latent variable non-identifiability. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5443–5455. Curran Associates, Inc.
- Xi, Q. and Bloem-Reddy, B. (2023). Indeterminacy in generative models: Characterization and strong identifiability. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6912–6939. PMLR.

- Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214.
- Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2021). CausalVAE: Disentangled representation learning via neural structural causal models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9588–9597.
- Zhang, H., Chen, Y., and Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85:358–372.

Supplementary Material

This Supplementary Material collects technical results, implementation details, and additional empirical summaries. Section S.2 presents a non-generic identifiability example. Sections S.3-S.8 contain the main proofs for our identifiability and consistency results. Section S.9 records implementation details. Section S.10 discusses additional related works.

Notation. For $d \geq 2$, let $\Delta_{d-1} = \{x \in \mathbb{R}^d : x_k \geq 0, \sum_{k=1}^d x_k = 1\}$ denote the $(d-1)$ -dimensional probability simplex, and let $\Delta_{d-1}^\circ = \{x \in \Delta_{d-1} : x_k > 0 \text{ for all } k\}$ denote its interior.

S.1 More identifiability results

We present additional identifiability results mentioned in Section 3. These results are adapted from existing work (Liu et al., 2025; Lee and Gu, 2025). Throughout, identifiability is understood up to the equivalence relation $\sim_{\mathcal{K}}$ defined in Section 3.1.

Definition 6. Let $(\Theta^*, \mathcal{G}^*, \mathbf{Q}^*) \in \Omega_K(\Theta, \mathcal{G}, \mathbf{Q})$ be the true parameter triple of the discrete causal representation learning framework. The framework is strictly identifiable up to the equivalence relation $\sim_{\mathcal{K}}$ if, for every alternative admissible triple $(\Theta', \mathcal{G}', \mathbf{Q}') \in \Omega_K(\Theta, \mathcal{G}, \mathbf{Q})$ satisfying the equality of marginal laws $\mathbb{P}_{\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}}} = \mathbb{P}_{\Theta^*, \mathcal{G}^*, \mathbf{Q}^*}$, it necessarily holds that $(\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}}) \sim_{\mathcal{K}} (\Theta^*, \mathcal{G}^*, \mathbf{Q}^*)$. Here, $\mathbb{P}_{\Theta, \mathcal{G}, \mathbf{Q}}$ denotes the marginal distribution of the observable vector \mathbf{X} , defined through (1) and (2).

Now we show our main results for the identifiability, which is from Proposition 1 in (Liu et al., 2025) and the definition of faithfulness.

Theorem S.1. Under Assumption 1, the framework is strictly identifiable if the following hold.

- (i) \mathbf{Q}^* contains two identity matrices after permuting the rows. Without the loss of generality, suppose that the first $2K$ rows of \mathbf{Q}^* are $[\mathbf{I}_K, \mathbf{I}_K]^\top$.

(ii) For any $\mathbf{z} \neq \mathbf{z}' \in \{0, 1\}^K$, there exists $j > 2K$ such that $\eta_j^*(\mathbf{z}) \neq \eta_j^*(\mathbf{z}')$.

When each latent cause affects the observables only through its main effects, without any interaction terms, Assumption 1(b) could be replaced by a weaker requirement.

Assumption 1'. (a) \mathcal{G} is a perfect map of \mathbf{p} and $\mathbf{p}_{\mathbf{z}} \in (0, 1)$ for all $\mathbf{z} \in \{0, 1\}^K$.

(b) $\sum_{j=1}^J \beta_{j,k} > 0$ for $k = 1, \dots, K$.

Under a main-effect measurement specification, the condition can be further weakened, which are from Proposition 1, 2 in Lee and Gu (2025) and the definition of faithfulness.

Corollary S.3. Suppose the measurement is main-effect only (no interaction terms). Under Assumption 1', the framework is strictly identifiable if the following conditions hold.

(i) \mathbf{Q}^* contains two identity matrices after permuting the rows. Without the loss of generality, suppose that the first $2K$ rows of \mathbf{Q}^* are $[\mathbf{I}_K, \mathbf{I}_K]^\top$.

(ii) For any $\mathbf{z} \neq \mathbf{z}' \in \{0, 1\}^K$, there exists $j > 2K$ such that $\sum_{k=1}^K \beta_{j,k}^* (z_k - z'_k) \neq 0$.

Remark 2. In many applications of the proposed framework, the number of observed items J is quite large, as is common in modern machine-learning settings. In such regimes, the strict identifiability requirement that \mathbf{Q}^* contain two identity blocks is less restrictive than it may initially appear.

Corollary S.4. Suppose the measurement is main-effect only (no interaction terms). Under Assumption 1' and Assumption 2, the framework is generic identifiable if the following hold.

(i) After a row permutation, we can write $\mathbf{Q}^* = [\mathbf{Q}_1^\top, \mathbf{Q}_2^\top, \mathbf{Q}_3^\top]^\top$, where $\mathbf{Q}_1, \mathbf{Q}_2 \in \{0, 1\}^{K \times K}$ have unit diagonals (off-diagonals arbitrary), and \mathbf{Q}_3 has no all-zero column.

S.2 Non-generic identifiability if Condition (ii) in Theorem 1 is violated

In this subsection we construct a concrete counterexample showing that Condition (ii) in Theorem 1 is indispensable. The example is chosen so that Assumption 1 and Assumption 2 hold, and Condition (i) of Theorem 1 is satisfied. The only assumption we deliberately violate is Condition (ii). Nevertheless, we exhibit a positive-measure subset of the parameter space on which the framework is not identifiable, so the framework is not generically identifiable.

We consider a one-layer saturated all-effect Bernoulli–logistic model with $K = 4$ latent variables and $J = 12$ items. In particular, we take ParFam_j to be the Bernoulli family and g_j to be the logistic link in (1).

Let

$$\mathbf{Q}^* = \left[\mathbf{Q}_1^\top, \mathbf{Q}_1^\top, \mathbf{Q}_1^\top \right]^\top, \quad \mathbf{Q}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is straightforward to verify that this measurement design satisfies Condition (i) in Theorem 1. However, Condition (ii) in Theorem 1 fails, since $\mathbf{Q}_{:,1}^* \succeq \mathbf{Q}_{:,3}^*$.

Because we work with Bernoulli–logistic responses, Assumption 2 is automatically satisfied. We also fix a strictly positive latent distribution $(p_{\mathbf{z}})_{\mathbf{z} \in \{0,1\}^4}$ so that Assumption 1(a) holds.

We now construct an explicit positive-measure subset of the parameter space on which the framework is not identifiable. Define

$$\tilde{B} = \{ \boldsymbol{\beta} \in \mathbb{R}^{16} : 2\beta_3 + \beta_{13} = \beta_0 \} \cup \{ \boldsymbol{\beta} \in \mathbb{R}^{16} : 2\beta_3 + \beta_{13} = \beta_1 \} \cup \{ \boldsymbol{\beta} \in \mathbb{R}^{16} : 2\beta_3 + \beta_{13} = 0 \}.$$

The set \tilde{B} is a finite union of proper algebraic varieties and therefore has Lebesgue measure

zero in \mathbb{R}^{16} .

Index the items as $j = 4m + r$ with $m \in \{0, 1, 2\}$ and $r \in \{1, 2, 3, 4\}$. If $r \in \{1, 2, 4\}$, select β_j from

$$\{\beta \in \mathbb{R}^{16} : \beta_S \neq 0 \text{ if and only if } S \subseteq \{r\}, \beta_r > 0\},$$

and if $r = 3$, select β_j from

$$\{\beta \in \mathbb{R}^{16} : \beta_S \neq 0 \text{ if and only if } S \subseteq \{1, 3\}, \beta_1 + \beta_3 + \beta_{13} > 0, \beta_1 + \beta_{13} > 0, \beta_3 + \beta_{13} > 0\} \setminus \tilde{B},$$

which has positive Lebesgue measure. Indeed, each inequality describes an open set in \mathbb{R}^{16} and hence has positive relative measure. Removing \tilde{B} , a measure-zero set, preserves positive relative measure. By construction, all such choices of β_j satisfy the monotonicity requirement in Assumption 1(b).

Next, we define a transformed parameterization $(\mathbf{B}', \mathbf{p}')$ that induces the same marginal distribution of \mathbf{X} but cannot be obtained from (\mathbf{B}, \mathbf{p}) through any latent-coordinate permutation.

For $j = 4m + r$ with $m \in \{0, 1, 2\}$ and $r \in \{1, 2, 4\}$, set

$$\beta'_{j,0} = \beta_{j,0}, \quad \beta'_{j,r} = \beta_{j,r},$$

and set all other entries of β'_j to zero. For items with indices $j = 4m + 3$ ($m = 0, 1, 2$), define

$$\beta'_{j,0} = \beta_{j,0} + \beta_{j,3}, \quad \beta'_{j,1} = \beta_{j,1} - \beta_{j,3}, \quad \beta'_{j,3} = -\beta_{j,3}, \quad \beta'_{j,13} = 2\beta_{j,3} + \beta_{j,13},$$

and again set all remaining entries of β'_j to zero.

Define a permutation π of the 2^4 latent states by

$$\pi(0000) = 0010, \quad \pi(1000) = 1000, \quad \pi(0100) = 0110, \quad \pi(0010) = 0000,$$

$$\begin{aligned}
\pi(0001) &= 0011, & \pi(1100) &= 1100, & \pi(1010) &= 1010, & \pi(1001) &= 1001, \\
\pi(0110) &= 0100, & \pi(0101) &= 0111, & \pi(0011) &= 0001, & \pi(1110) &= 1110, \\
\pi(1101) &= 1101, & \pi(1011) &= 1011, & \pi(0111) &= 0101, & \pi(1111) &= 1111.
\end{aligned}$$

Set the transformed mixing weights by π :

$$\begin{aligned}
p'_{0000} &= p_{0010}, & p'_{1000} &= p_{1000}, & p'_{0100} &= p_{0110}, & p'_{0010} &= p_{0000}, \\
p'_{0001} &= p_{0011}, & p'_{1100} &= p_{1100}, & p'_{1010} &= p_{1010}, & p'_{1001} &= p_{1001}, \\
p'_{0110} &= p_{0100}, & p'_{0101} &= p_{0111}, & p'_{0011} &= p_{0001}, & p'_{1110} &= p_{1110}, \\
p'_{1101} &= p_{1101}, & p'_{1011} &= p_{1011}, & p'_{0111} &= p_{0101}, & p'_{1111} &= p_{1111}.
\end{aligned}$$

For all \mathbf{z} and j ,

$$\eta'_j(\mathbf{z}) = \eta_j(\pi(\mathbf{z})), \quad q'_{j,\mathbf{z}} = \sigma(\eta'_j(\mathbf{z})) = q_{j,\pi(\mathbf{z})}.$$

Consequently, for every $\mathbf{x} \in \{0, 1\}^J$,

$$\mathbb{P}'(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{z}} p'_{\mathbf{z}} \prod_{j=1}^J (q'_{j,\mathbf{z}})^{\mathbf{x}_j} (1 - q'_{j,\mathbf{z}})^{1 - \mathbf{x}_j} = \sum_{\mathbf{z}} p_{\pi(\mathbf{z})} \prod_{j=1}^J q_{j,\pi(\mathbf{z})}^{\mathbf{x}_j} (1 - q_{j,\pi(\mathbf{z})})^{1 - \mathbf{x}_j} = \mathbb{P}(\mathbf{X} = \mathbf{x}).$$

It is straightforward to verify that $\eta'_j(\mathbf{z}) > \eta'_j(\mathbf{z}')$ whenever $\mathbf{z} \succeq q_j$ and $\mathbf{z}' \not\succeq q_j$, for $1 \leq j \leq 12$, so the monotonicity condition in Assumption 1(b) continues to hold under the transformed parameterization.

By construction of β_j , it further follows that β'_j cannot be obtained from β_j through any latent-coordinate permutation if $j \equiv 3 \pmod{4}$, because the value of $\beta'_{j,13}$ differs from all four entries of the original vector β_j . Therefore, (\mathbf{B}, \mathbf{p}) and $(\mathbf{B}', \mathbf{p}')$ are not related by any latent-coordinate permutation but induce the same observable law. Since the set of admissible

$(\beta_j)_{j=1}^{12}$ has positive Lebesgue measure, the framework is not generically identifiable. In particular, this shows that even when Assumption 1, Assumption 2, and Condition (i) of Theorem 1 all hold, generic identifiability can fail once Condition (ii) is violated.

S.3 Proof of Theorem 1

Before presenting the proof of Theorem 1, we introduce some additional notations.

For each j , fix an enumeration of $\mathcal{C}_j^{\text{can}} \setminus \{\mathcal{X}_j\}$ as

$$\mathcal{C}_j^{\text{can}} \setminus \{\mathcal{X}_j\} = \{T_{1,j}, T_{2,j}, \dots\}, \quad (j \in [J]).$$

For each $t \geq 1$, define the parameter-independent finite discretization

$$\overline{\mathcal{D}}_j^{(t)} := \{T_{1,j}, \dots, T_{t,j}\} \cup \{\mathcal{X}_j\} \subseteq \mathcal{C}_j^{\text{can}}, \quad \overline{\mathcal{D}}^{(t)} := (\overline{\mathcal{D}}_j^{(t)})_{j \in [J]}.$$

Then $\kappa_j^{(t)} := |\overline{\mathcal{D}}_j^{(t)}| = t + 1$, and we index $\overline{\mathcal{D}}_j^{(t)} = (S_{1,j}^{(t)}, \dots, S_{\kappa_j^{(t)},j}^{(t)})$ with $S_{\kappa_j^{(t)},j}^{(t)} = \mathcal{X}_j$.

For each $t \geq 1$, define $\mathbf{N}_1^{(t)}$ to be a $\kappa_1^{(t)} \cdots \kappa_K^{(t)} \times 2^K$ matrix with entries

$$\mathbf{N}_1^{(t)}((l_1, \dots, l_K), \mathbf{z}) := \mathbb{P}(X_1 \in S_{l_1,1}^{(t)}, \dots, X_K \in S_{l_K,K}^{(t)} \mid \mathbf{z}).$$

Columns are indexed by $\mathbf{z} \in \{0, 1\}^K$ and rows by $\xi_1 = (l_1, \dots, l_K)$ with $l_j \in [\kappa_j^{(t)}]$. Similarly, let $\mathbf{N}_2^{(t)}$ be the $\kappa_{K+1}^{(t)} \cdots \kappa_{2K}^{(t)} \times 2^K$ matrix whose $((l_{K+1}, \dots, l_{2K}), \mathbf{z})$ -entry is

$$\mathbb{P}(X_{K+1} \in S_{l_{K+1},K+1}^{(t)}, \dots, X_{2K} \in S_{l_{2K},2K}^{(t)} \mid \mathbf{z}),$$

and let $\mathbf{N}_3^{(t)}$ be the $\kappa_{2K+1}^{(t)} \cdots \kappa_J^{(t)} \times 2^K$ matrix whose $((l_{2K+1}, \dots, l_J), \mathbf{z})$ -entry is

$$\mathbb{P}(X_{2K+1} \in S_{l_{2K+1},2K+1}^{(t)}, \dots, X_J \in S_{l_J,J}^{(t)} \mid \mathbf{z}).$$

For brevity, set

$$v_1^{(t)} = \prod_{k=1}^K \kappa_k^{(t)}, \quad v_2^{(t)} = \prod_{k=K+1}^{2K} \kappa_k^{(t)}, \quad v_3^{(t)} = \prod_{k=2K+1}^J \kappa_k^{(t)}.$$

Since $S_{\kappa_j^{(t)}, j}^{(t)} = \mathcal{X}_j$, the last row of each $\mathbf{N}_a^{(t)}$ equals $\mathbf{1}_{2K}^\top$.

Define the 3-way marginal probability tensor $\mathbf{P}_0^{(t)}$ of size $v_1^{(t)} \times v_2^{(t)} \times v_3^{(t)}$ by

$$\begin{aligned} \mathbf{P}_0^{(t)}(\xi_1, \xi_2, \xi_3) &= \mathbb{P}(X_1 \in S_{l_1,1}^{(t)}, \dots, X_J \in S_{l_J,J}^{(t)}) \\ &= \sum_{\mathbf{z}} \pi_{\mathbf{z}} \mathbf{N}_1^{(t)}((l_1, \dots, l_K), \mathbf{z}) \mathbf{N}_2^{(t)}((l_{K+1}, \dots, l_{2K}), \mathbf{z}) \mathbf{N}_3^{(t)}((l_{2K+1}, \dots, l_J), \mathbf{z}). \end{aligned}$$

Equivalently,

$$\mathbf{P}_0^{(t)} = [\mathbf{N}_1^{(t)} \text{Diag}(\mathbf{p}), \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}]. \quad (\text{S.1})$$

We record two lemmas whose proofs are deferred to the end of the subsection. The first establishes uniqueness of the tensor decomposition of $\mathbf{P}_0^{(t)}$ up to a common column permutation.

Lemma S.1. *Consider a discrete causal representation learning framework with parameters $(\mathbf{p}^*, \mathcal{G}^*, \mathbf{B}^*, \mathbf{Q}^*, \boldsymbol{\gamma}^*)$ satisfying the conditions of Theorem 1. For each $t \geq 1$, let $\mathbf{P}_0^{(t)}$ be the tensor induced by the parameter-independent discretization $\overline{\mathcal{D}}^{(t)}$ defined above, with factor matrices $\mathbf{N}_1^{(t)}, \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}$, so that $\mathbf{P}_0^{(t)} = [\mathbf{N}_1^{(t)} \text{Diag}(\mathbf{p}), \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}]$. Then there exists a Lebesgue-null set $\mathcal{N}_\infty \subset \Omega_K(\boldsymbol{\Theta}; \mathcal{G}^*, \mathbf{Q}^*)$ which constrains only $(\mathbf{B}, \boldsymbol{\gamma})$ such that the following holds.*

For every $\boldsymbol{\Theta} \in \Omega_K(\boldsymbol{\Theta}; \mathcal{G}^, \mathbf{Q}^*) \setminus \mathcal{N}_\infty$, there exists an integer $t_0 = t_0(\boldsymbol{\Theta}) < \infty$ such that for all $t \geq t_0$ the rank- 2^K CP decomposition of $\mathbf{P}_0^{(t)}$ is unique up to a common column permutation. Moreover, since $\mathcal{X}_j \in \overline{\mathcal{D}}_j^{(t)}$, each $\mathbf{N}_a^{(t)}$ contains a row equal to $\mathbf{1}_{2K}^\top$, hence the uniqueness contains no nontrivial scaling ambiguity.*

The next lemma constrains how the 2^K columns can be permuted.

Lemma S.2. *Let $(\mathbf{p}, \mathbf{B}, \mathcal{G}, \mathbf{Q})$ and $(\mathbf{p}', \mathbf{B}', \mathcal{G}', \mathbf{Q}')$ satisfy Assumption 1, and suppose \mathbf{Q} also meets the conditions of Theorem 1. Assume there exists a permutation $\mathfrak{S} \in S_{\{0,1\}^K}$ such that*

$$\eta_j(\mathbf{z}) = \eta'_j(\mathfrak{S}(\mathbf{z})) \quad \text{for all } j, \mathbf{z}.$$

Then $(\mathbf{p}, \mathbf{B}, \mathcal{G}, \mathbf{Q}) \sim_{\mathcal{K}} (\mathbf{p}', \mathbf{B}', \mathcal{G}', \mathbf{Q}')$ for $\mathbf{B} \in \Omega(\mathbf{B}; \mathbf{Q})$, where

$$\Omega(\mathbf{B}; \mathbf{Q}) = \{\mathbf{B} : \beta_{j,S} = 0 \text{ whenever } S \not\subseteq K_j, \beta_{j,\{k\}} \neq 0 \text{ if and only if } k \in K_j\}.$$

Assume \mathbf{Q}^* satisfies the conditions of Theorem 1 and that $\Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$. Suppose there exist alternative parameters $(\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}})$ such that $\mathbb{P}_{\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}}} = \mathbb{P}_{\Theta, \mathbf{Q}^*, \mathcal{G}^*}$. We will show that $(\Theta, \mathcal{G}^*, \mathbf{Q}^*) \sim_K (\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}})$ for Θ outside a Lebesgue-null set.

By Lemma S.1, there exists a Lebesgue-null set \mathcal{N}_∞ such that for every $\Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) \setminus \mathcal{N}_\infty$, we could find an integer $t_0 = t_0(\Theta) < \infty$ such that for every $t \geq t_0$ the tensor decomposition of $\mathbf{P}_0^{(t)}$ is unique up to a common column permutation. Fix any $t \geq t_0$. Then

$$\mathbf{P}_0^{(t)} = [\mathbf{N}_1^{(t)} \text{Diag}(\mathbf{p}), \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}] = [\tilde{\mathbf{N}}_1^{(t)} \text{Diag}(\tilde{\mathbf{p}}), \tilde{\mathbf{N}}_2^{(t)}, \tilde{\mathbf{N}}_3^{(t)}],$$

where the equality holds up to a common permutation of the 2^K columns. Hence there exists a permutation $\mathfrak{S}^{(t)} \in S_{\{0,1\}^K}$ such that

$$\mathbf{N}_a^{(t)}(\cdot, \mathbf{z}) = \tilde{\mathbf{N}}_a^{(t)}(\cdot, \mathfrak{S}^{(t)}(\mathbf{z})), \quad a = 1, 2, 3, \quad \mathbf{p}_z = \tilde{\mathbf{p}}_{\mathfrak{S}^{(t)}(\mathbf{z})}.$$

In particular, for every $j \in [J]$, every $l \in [\kappa_j^{(t)}]$, and every $\mathbf{z} \in \{0, 1\}^K$,

$$\mathbb{P}_{j, g_j(\eta_j(\mathbf{z}), \gamma_j)}(S_{l,j}^{(t)}) = \mathbb{P}_{j, g_j(\tilde{\eta}_j(\mathfrak{S}^{(t)}(\mathbf{z})), \tilde{\gamma}_j)}(S_{l,j}^{(t)}). \quad (\text{S.2})$$

We now justify the passage from the setwise equalities in (S.2) to equality of the full condi-

tional laws as probability measures, and simultaneously show that the aligning permutation stabilizes as t increases.

Lemma S.3. *Fix a countable separating class \mathcal{C}_j for each $j \in [J]$. Let $\mathcal{D}_j \subseteq \mathcal{D}_j^+ \subseteq \mathcal{C}_j$ be two finite collections for each j . Construct the corresponding factor matrices $(\mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3)$ and $(\mathbf{N}_1^+, \mathbf{N}_2^+, \mathbf{N}_3^+)$, and similarly $(\tilde{\mathbf{N}}_1, \tilde{\mathbf{N}}_2, \tilde{\mathbf{N}}_3)$ and $(\tilde{\mathbf{N}}_1^+, \tilde{\mathbf{N}}_2^+, \tilde{\mathbf{N}}_3^+)$ under an alternative parameterization. Assume that both tensors admit unique rank- 2^K CP decompositions up to a common column permutation, so that there exist $\mathfrak{S}, \mathfrak{S}^+ \in S_{\{0,1\}^K}$ satisfying*

$$\mathbf{N}_a(\cdot, \mathbf{z}) = \tilde{\mathbf{N}}_a(\cdot, \mathfrak{S}(\mathbf{z})), \quad \mathbf{N}_a^+(\cdot, \mathbf{z}) = \tilde{\mathbf{N}}_a^+(\cdot, \mathfrak{S}^+(\mathbf{z})), \quad a = 1, 2, 3.$$

If \mathbf{N}_1 has full column rank 2^K , then $\mathfrak{S}^+ = \mathfrak{S}$.

Proof. Since \mathbf{N}_1 has full column rank 2^K , its 2^K columns are pairwise distinct. Because $\mathcal{D}_j \subseteq \mathcal{D}_j^+$ for all j , each row event used to define \mathbf{N}_1 (i.e., each product event determined by choosing one set from each \mathcal{D}_j) also appears among the row events defining \mathbf{N}_1^+ . Thus, for each $\mathbf{z} \in \{0, 1\}^K$, the column $\mathbf{N}_1(\cdot, \mathbf{z})$ is obtained from $\mathbf{N}_1^+(\cdot, \mathbf{z})$ by restricting to those rows corresponding to product events formed from \mathcal{D} .

Fix \mathbf{z} . From the two Kruskal conclusions, we have $\mathbf{N}_1(\cdot, \mathbf{z}) = \tilde{\mathbf{N}}_1(\cdot, \mathfrak{S}(\mathbf{z}))$ and $\mathbf{N}_1^+(\cdot, \mathbf{z}) = \tilde{\mathbf{N}}_1^+(\cdot, \mathfrak{S}^+(\mathbf{z}))$. Restricting the second equality to the rows corresponding to \mathcal{D} gives $\mathbf{N}_1(\cdot, \mathbf{z}) = \tilde{\mathbf{N}}_1(\cdot, \mathfrak{S}^+(\mathbf{z}))$. Therefore, $\tilde{\mathbf{N}}_1(\cdot, \mathfrak{S}(\mathbf{z})) = \tilde{\mathbf{N}}_1(\cdot, \mathfrak{S}^+(\mathbf{z}))$. Since $\tilde{\mathbf{N}}_1$ is a column permutation of \mathbf{N}_1 , its columns are also pairwise distinct, so the above equality forces $\mathfrak{S}(\mathbf{z}) = \mathfrak{S}^+(\mathbf{z})$. As \mathbf{z} was arbitrary, $\mathfrak{S} = \mathfrak{S}^+$. \square

Since $\Theta \notin \mathcal{N}_\infty$, Lemma S.1 implies that for every $t \geq t_0(\Theta)$ the Kruskal conclusion holds for $\mathbf{P}_0^{(t)}$, and hence the associated aligning permutation $\mathfrak{S}^{(t)}$ is well-defined. Moreover, for every $t \geq t_0(\Theta)$, the Kruskal argument in the proof of Lemma S.1 yields $\text{rk}_k(\mathbf{N}_1^{(t)}) = 2^K$ and hence $\mathbf{N}_1^{(t)}$ has full column rank 2^K . Applying Lemma S.3 to the nested discretizations $\overline{\mathcal{D}}^{(t)} \subseteq \overline{\mathcal{D}}^{(t+1)}$ for $t \geq t_0(\Theta)$ yields that $\mathfrak{S}^{(t)}$ is constant in t . Denote the common permutation

by \mathfrak{S} .

Now fix $j \in [J]$ and let $S \in \mathcal{C}_j^{\text{can}}$ be arbitrary. Since $\bigcup_{t \geq t_0} \overline{\mathcal{D}}_j^{(t)} = \mathcal{C}_j^{\text{can}}$, there exists $t \geq t_0$ such that $S \in \overline{\mathcal{D}}_j^{(t)}$. Therefore (S.2) implies

$$\mathbb{P}_{j, g_j(\eta_j(\mathbf{z}), \gamma_j)}(S) = \mathbb{P}_{j, g_j(\tilde{\eta}_j(\mathfrak{S}(\mathbf{z})), \tilde{\gamma}_j)}(S) \quad \text{for all } \mathbf{z} \in \{0, 1\}^K.$$

Because $\mathcal{C}_j^{\text{can}}$ is separating, we conclude that

$$\mathbb{P}_{j, g_j(\eta_j(\mathbf{z}), \gamma_j)} = \mathbb{P}_{j, g_j(\tilde{\eta}_j(\mathfrak{S}(\mathbf{z})), \tilde{\gamma}_j)} \quad \text{as probability measures on } \mathcal{X}_j, \quad \text{for all } \mathbf{z}.$$

By Assumption 2(ii) and injectivity of g_j in Assumption 2(iii), this further implies

$$\eta_j(\mathbf{z}) = \tilde{\eta}_j(\mathfrak{S}(\mathbf{z})) \quad \text{and} \quad \gamma_j = \tilde{\gamma}_j, \quad \text{for all } j, \mathbf{z}.$$

Finally, Lemma S.2 yields $(\mathbf{p}, \mathbf{B}, \mathcal{G}, \mathbf{Q}) \sim_{\mathcal{K}} (\tilde{\mathbf{p}}, \tilde{\mathbf{B}}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}})$, completing the proof.

S.3.1 Proof of Lemma S.1

Fix $t \geq 1$ and consider the tensor $\mathbf{P}_0^{(t)}$ induced by the parameter-independent discretization $\overline{\mathcal{D}}^{(t)}$ constructed at the beginning of this subsection, with factor matrices $\mathbf{N}_1^{(t)}, \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}$ satisfying (S.1). Write $\text{rk}_k(M)$ for the Kruskal column rank of a matrix M .

By Kruskal's theorem, it suffices to prove that there exist a Lebesgue-null set $\mathcal{N}_\infty \subset \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$ which constrains only $(\mathbf{B}, \boldsymbol{\gamma})$ and, for every $\Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) \setminus \mathcal{N}_\infty$, an integer $t_0 = t_0(\Theta) < \infty$ such that

$$\text{rk}_k(\mathbf{N}_1^{(t)}) = 2^K, \quad \text{rk}_k(\mathbf{N}_2^{(t)}) = 2^K, \quad \text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2 \quad (\text{S.3})$$

for all $t \geq t_0(\Theta)$.

We first establish that $\text{rk}_k(\mathbf{N}_1^{(t)}) = 2^K$ generically for all sufficiently large t . Let $\mathcal{J}_{\text{disp}}^{(K)} \subseteq$

$[K]$ denote the indices of items among $\{1, \dots, K\}$ whose response family genuinely includes an unknown dispersion parameter γ_j . For one-parameter families we do not treat γ_j as a coordinate.

Define the local parameter block

$$\Theta_1 := \left(\{ \beta_{j,S} : j \in [K], S \subseteq K_j \}, \{ \gamma_j : j \in \mathcal{J}_{\text{disp}}^{(K)} \} \right),$$

which we identify with a vector in a Euclidean space \mathbb{R}^{D_1} of dimension $D_1 := \sum_{j=1}^K 2^{|K_j|} + |\mathcal{J}_{\text{disp}}^{(K)}|$. Throughout we restrict attention to the open connected domain

$$U_1 := \left\{ \Theta_1 : \gamma_j > 0 \text{ for all } j \in \mathcal{J}_{\text{disp}}^{(K)} \right\} \subset \mathbb{R}^{D_1}.$$

Because the discretization $\overline{\mathcal{D}}^{(t)}$ is parameter-independent, every entry of $\mathbf{N}_1^{(t)}$ is a finite product of terms of the form

$$\mathbb{P}_{j, g_j(\eta_j(\mathbf{z}), \gamma_j)}(S), \quad S \in \overline{\mathcal{D}}_j^{(t)},$$

with S fixed. For $j \in \mathcal{J}_{\text{disp}}^{(K)}$, Assumption 2(i) and Assumption 2(iii) imply that $(\eta, \gamma) \mapsto \mathbb{P}_{j, g_j(\eta, \gamma)}(S)$ is real-analytic on $\mathbb{R} \times (0, \infty)$. For $j \notin \mathcal{J}_{\text{disp}}^{(K)}$, g_j is independent of γ and $\mathbb{P}_{j, g_j(\eta, \gamma)}(S) = \mathbb{P}_{j, g_j(\eta, \gamma_0)}(S)$ for any fixed $\gamma_0 \in [0, \infty)$; in particular, $\eta \mapsto \mathbb{P}_{j, g_j(\eta, \gamma_0)}(S)$ is real-analytic on \mathbb{R} . Since each $\eta_j(\mathbf{z})$ is a polynomial in the coefficients $\{\beta_{j,S}\}$, it follows that every entry of $\mathbf{N}_1^{(t)}$ is real-analytic on the domain U_1 . Consequently, $f_{1,t}(\Theta_1) := \det((\mathbf{N}_1^{(t)})^\top \mathbf{N}_1^{(t)})$ is a real-analytic function on the domain $U_1 \subset \mathbb{R}^{D_1}$.

Next we describe the projection of $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$ onto $\Theta_1 \subseteq U_1$. Let $E_1 = \{(j, k) : j \in [K], k \in K_j\}$ index the main-effect coefficients $\beta_{j, \{k\}}$ for $j \leq K$. For each sign pattern $\sigma \in \{-1, +1\}^{E_1}$ define the orthant

$$\mathcal{E}_1^{(\sigma)} := \left\{ \Theta_1 \subseteq U_1 : \beta_{j, \{k\}} \sigma_{j,k} > 0 \text{ for all } (j, k) \in E_1 \right\},$$

where all remaining coordinates in Θ_1 are unrestricted. Each $\mathcal{E}_1^{(\sigma)}$ is an open, connected domain of \mathbb{R}^{D_1} , and

$$\Pi_1(\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)) = \bigcup_{\sigma} \mathcal{E}_1^{(\sigma)},$$

where Π_1 denotes the projection onto the coordinates Θ_1 .

Fix any sign pattern $\sigma \in \{-1, +1\}^{E_1}$. We will show that there exist an explicit parameter point $\bar{\Theta}_1^{(\sigma)} \in \mathcal{E}_1^{(\sigma)}$ and an integer $t_\sigma < \infty$ such that $f_{1,t}(\bar{\Theta}_1^{(\sigma)}) > 0$ for all $t \geq t_\sigma$. In particular, for every $t \geq t_\sigma$ the restriction of $f_{1,t}$ to $\mathcal{E}_1^{(\sigma)}$ is a nontrivial real-analytic function on the open, connected domain $\mathcal{E}_1^{(\sigma)}$. By [Mityagin \(2015\)](#), the zero set $\mathcal{V}_{1,\sigma,t} := \{\Theta_1 \in \mathcal{E}_1^{(\sigma)} : f_{1,t}(\Theta_1) = 0\}$ has Lebesgue measure zero in $\mathcal{E}_1^{(\sigma)}$.

To construct the point, we now explicitly use condition (i) of [Theorem 1](#). There exists a permutation ϱ_1 of $[K]$ such that the permuted $K \times K$ block $\bar{\mathbf{Q}}_1 := \mathbf{Q}_{\varrho_1(1:K),:}$ has unit diagonal, equivalently $q_{\varrho_1(r),r} = 1$ for all $r \in \{1, \dots, K\}$. Define the induced bijection $\rho : [K] \rightarrow [K]$ by $\rho(j) := \varrho_1^{-1}(j)$ for $j \in \{1, \dots, K\}$. Then, for every $j \in [K]$, we have $q_{j,\rho(j)} = 1$, hence $\rho(j) \in K_j$.

Fix a sign pattern $\sigma \in \{-1, +1\}^{E_1}$. Define a boundary point $\Theta_{1,0}^{(\sigma)}$ by setting all interaction terms to zero and keeping only the single admissible main effect $\beta_{j,\{\rho(j)\}}$ for each $j \in [K]$

$$\beta_{j,S} = 0 \text{ for all } j \in [K] \text{ and } |S| \geq 2, \quad \beta_{j,\{\rho(j)\}} = \sigma_{j,\rho(j)}, \quad \beta_{j,\{k\}} = 0 \text{ for all } (j,k) \in E_1, k \neq \rho(j),$$

with all remaining coordinates in Θ_1 arbitrary, and $\gamma_j > 0$ for $j \in \mathcal{J}_{\text{disp}}^{(K)}$. Because $\rho(j) \in K_j$, the coordinate $\beta_{j,\{\rho(j)\}}$ is indeed part of Θ_1 , so this assignment is admissible. Under $\Theta_{1,0}^{(\sigma)}$, for each $j \in [K]$ the conditional law of X_j depends on \mathbf{z} only through $\mathbf{z}_{\rho(j)}$, since $\eta_j(\mathbf{z}) = \beta_{j,\emptyset} + \beta_{j,\{\rho(j)\}} \mathbf{z}_{\rho(j)}$.

Fix $j \in [K]$. Let $\mu_{j,0}$ and $\mu_{j,1}$ denote the two conditional laws of X_j under $\Theta_{1,0}^{(\sigma)}$ corresponding to $\mathbf{z}_{\rho(j)} = 0$ and $\mathbf{z}_{\rho(j)} = 1$, respectively. Because $\beta_{j,\{\rho(j)\}} \neq 0$, we have $\eta_j(0) \neq \eta_j(1)$ in the $\mathbf{z}_{\rho(j)}$ coordinate. By [Assumption 2\(iii\)](#) the map $\eta \mapsto g_j(\eta, \gamma_j)$ is injective for fixed γ_j ,

hence the induced parameters are distinct. By identifiability in Assumption 2(ii), it follows that $\mu_{j,0} \neq \mu_{j,1}$ as probability measures. Since $\mathcal{C}_j^{\text{can}}$ is separating, there exists a set $B_j \in \mathcal{C}_j^{\text{can}}$ such that $\mu_{j,0}(B_j) \neq \mu_{j,1}(B_j)$.

For each $r \in [K]$, define $j_r := \varrho_1(r)$, so that $\rho(j_r) = r$ and $q_{j_r,r} = 1$. For j_r , choose a distinguishing set $B_{j_r} \in \mathcal{C}_{j_r}$ as above and write $B_{j_r} = T_{i_r, j_r}$ for some $i_r \geq 1$. Define $t_\sigma := \max_{r \in [K]} i_r < \infty$. For every $t \geq t_\sigma$ we have $B_{j_r} \in \overline{\mathcal{D}}_{j_r}^{(t)}$ for all $r \in [K]$, and moreover $S_{\kappa_{j_r, j_r}^{(t)}}^{(t)} = \mathcal{X}_{j_r} \in \overline{\mathcal{D}}_{j_r}^{(t)}$.

Fix any $t \geq t_\sigma$. Consider the $2^K \times 2^K$ submatrix of $\mathbf{N}_1^{(t)}$ obtained by restricting to the 2^K rows indexed by $l_{j_r} \in \{i_r, \kappa_{j_r}^{(t)}\}$ ($r = 1, \dots, K$), that is, for each r we use either the event B_{j_r} or the event \mathcal{X}_{j_r} . Under $\Theta_{1,0}^{(\sigma)}$ and conditional independence of $(X_{j_1}, \dots, X_{j_K})$ given \mathbf{z} , this submatrix factorizes as a Kronecker product:

$$\mathbf{N}_{1,\text{sub}}^{(t)}(\Theta_{1,0}^{(\sigma)}) = \bigotimes_{r=1}^K \begin{pmatrix} \mu_{j_r,0}(B_{j_r}) & \mu_{j_r,1}(B_{j_r}) \\ 1 & 1 \end{pmatrix}.$$

Each 2×2 factor has nonzero determinant $\mu_{j_r,0}(B_{j_r}) - \mu_{j_r,1}(B_{j_r}) \neq 0$, hence the Kronecker product is invertible. Therefore $\mathbf{N}_1^{(t)}(\Theta_{1,0}^{(\sigma)})$ has full column rank 2^K .

Next, we perturb $\Theta_{1,0}^{(\sigma)}$ into the open orthant $\mathcal{E}_1^{(\sigma)}$ while preserving full column rank of $\mathbf{N}_1^{(t)}$. For $\varepsilon > 0$, define $\Theta_{1,\varepsilon}^{(\sigma)}$ by keeping all coordinates of $\Theta_{1,0}^{(\sigma)}$ unchanged except setting, for every $(j,k) \in E_1$ with $k \neq \rho(j)$, $\beta_{j,\{k\}} = \sigma_{j,k}\varepsilon$, and leaving $\beta_{j,\{\rho(j)\}} = \sigma_{j,\rho(j)}$. Then $\Theta_{1,\varepsilon}^{(\sigma)} \in \mathcal{E}_1^{(\sigma)}$ for every $\varepsilon > 0$. Since the determinant of the fixed $2^K \times 2^K$ submatrix $\mathbf{N}_{1,\text{sub}}^{(t)}(\Theta_1)$ is continuous in Θ_1 on U_1 and is nonzero at $\Theta_{1,0}^{(\sigma)}$, there exists $\varepsilon_\sigma(t) > 0$ such that for all $\varepsilon \in (0, \varepsilon_\sigma(t))$ the submatrix remains invertible, and hence $\mathbf{N}_1^{(t)}(\Theta_{1,\varepsilon}^{(\sigma)})$ has full column rank 2^K . Choose any such ε and set $\overline{\Theta}_1^{(\sigma)} := \Theta_{1,\varepsilon}^{(\sigma)}$. Then $f_{1,t}(\overline{\Theta}_1^{(\sigma)}) > 0$.

Since there are only finitely many sign patterns, define $t_1 := \max_{\sigma \in \{-1,+1\}^{E_1}} t_\sigma < \infty$. Then for every $t \geq t_1$ and every sign pattern σ , the restriction of $f_{1,t}$ to $\mathcal{E}_1^{(\sigma)}$ is a nontrivial real-analytic function, and hence the set $\mathcal{V}_{1,\sigma,t} := \{\Theta_1 \in \mathcal{E}_1^{(\sigma)} : f_{1,t}(\Theta_1) = 0\}$ has Lebesgue

measure zero in $\mathcal{E}_1^{(\sigma)}$.

For each fixed $t \geq t_1$, define

$$\mathcal{V}_{1,t} := \Pi_1(\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)) \cap \bigcup_{\sigma} \mathcal{V}_{1,\sigma,t}.$$

Then $\mathcal{V}_{1,t}$ has Lebesgue measure zero in $\Pi_1(\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*))$. Define

$$\mathcal{N}_{1,t}^{(1)} := \{(\mathbf{p}, \mathbf{B}, \gamma) \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) : \Theta_1 \in \mathcal{V}_{1,t}\}.$$

Then $\mathcal{N}_{1,t}^{(1)}$ has Lebesgue measure zero in $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$ and depends only on (\mathbf{B}, γ) . For every $t \geq t_1$ and every $\Theta \notin \mathcal{N}_{1,t}^{(1)}$ we have $f_{1,t}(\Theta_1) \neq 0$, hence $\mathbf{N}_1^{(t)}$ has full column rank 2^K , which implies $\text{rk}_k(\mathbf{N}_1^{(t)}) = 2^K$.

An entirely analogous argument yields an integer $t_2 < \infty$ and, for each $t \geq t_2$, a Lebesgue-null set $\mathcal{N}_{1,t}^{(2)} \subset \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$ depending only on (\mathbf{B}, γ) such that

$$\text{rk}_k(\mathbf{N}_2^{(t)}) = 2^K \quad \text{for all } t \geq t_2, \Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) \setminus \mathcal{N}_{1,t}^{(2)}.$$

We now prove that $\text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2$ generically. It suffices to verify that Condition B of [Lee and Gu \(2024\)](#) holds for generic parameters in $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$.

Fix any pair $\mathbf{z} \neq \mathbf{z}'$ and choose an index $\ell = \ell(\mathbf{z}, \mathbf{z}') \in [K]$ such that $z_\ell \neq z'_\ell$. Since \mathbf{Q}_3 has no all-zero column, there exists an item $j = j(\mathbf{z}, \mathbf{z}') \in \{2K + 1, \dots, J\}$ such that $\mathbf{Q}_{j,\ell} = 1$, hence $\ell \in K_j$.

Similarly, let $\mathcal{J}_{\text{disp}} \subseteq [J]$ denote the indices of items whose response family genuinely includes an unknown dispersion parameter γ_j . Collect the local measurement parameters for

item j into the free block

$$\Theta_j := \begin{cases} (\{\beta_{j,S} : S \subseteq K_j\}, \gamma_j) \in \mathbb{R}^{2^{|K_j|}} \times (0, \infty), & j \in \mathcal{J}_{\text{disp}}, \\ (\{\beta_{j,S} : S \subseteq K_j\}) \in \mathbb{R}^{2^{|K_j|}}, & j \notin \mathcal{J}_{\text{disp}}. \end{cases}$$

By the definition of $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$, we have $\beta_{j,\{k\}} = 0$ for all $k \notin K_j$ and $\beta_{j,\{\ell\}} \neq 0$.

For this fixed pair $(\mathbf{z}, \mathbf{z}')$, consider the difference

$$h_{j,\mathbf{z},\mathbf{z}'}(\Theta_j) := \eta_j(\mathbf{z}) - \eta_j(\mathbf{z}') = \sum_{S \subseteq K_j} \beta_{j,S} \left(\prod_{k \in S} z_k - \prod_{k \in S} z'_k \right).$$

This is a linear function of the coefficients $\{\beta_{j,S}\}$ (and is independent of γ_j when $j \in \mathcal{J}_{\text{disp}}$). Moreover, the coefficient of $\beta_{j,\{\ell\}}$ in $h_{j,\mathbf{z},\mathbf{z}'}$ equals $z_\ell - z'_\ell \neq 0$, hence $h_{j,\mathbf{z},\mathbf{z}'}$ is a nontrivial linear functional.

Therefore the zero set

$$E_{j,\mathbf{z},\mathbf{z}'} := \{\Theta_j : h_{j,\mathbf{z},\mathbf{z}'}(\Theta_j) = 0\}$$

has Lebesgue measure zero in the free-coordinate space of Θ_j : it is an affine hyperplane in $\mathbb{R}^{2^{|K_j|}}$ when $j \notin \mathcal{J}_{\text{disp}}$, and it is an affine hyperplane in the β -coordinates times $(0, \infty)$ when $j \in \mathcal{J}_{\text{disp}}$.

Define the corresponding exceptional set in the full parameter space by

$$\mathcal{N}_{\mathbf{z},\mathbf{z}'}^{(3)} := \left\{ \Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) : h_{j(\mathbf{z},\mathbf{z}'),\mathbf{z},\mathbf{z}'}(\Theta_{j(\mathbf{z},\mathbf{z}')}) = 0 \right\}.$$

Since $h_{j(\mathbf{z},\mathbf{z}'),\mathbf{z},\mathbf{z}'}$ is a nontrivial linear functional of the local block $\Theta_{j(\mathbf{z},\mathbf{z}')}$ (through its β -coordinates), the set $\mathcal{N}_{\mathbf{z},\mathbf{z}'}^{(3)}$ has Lebesgue measure zero in $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$.

Now fix $\Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) \setminus \mathcal{N}_{\mathbf{z},\mathbf{z}'}^{(3)}$. Then $\eta_j(\mathbf{z}) \neq \eta_j(\mathbf{z}')$. Writing

$$\mathbb{P}_{j,\mathbf{z}} = \mathbb{P}_{j,\theta_{j,\mathbf{z}}}, \quad \theta_{j,\mathbf{z}} := g_j(\eta_j(\mathbf{z}), \gamma_j) \in H_j^\circ,$$

injectivity of $\eta \mapsto g_j(\eta, \gamma_j)$ in Assumption 2(iii) implies $\theta_{j,\mathbf{z}} \neq \theta_{j,\mathbf{z}'}$. By identifiability of ParFam_j in Assumption 2(ii), we have $\mathbb{P}_{j,\mathbf{z}} \neq \mathbb{P}_{j,\mathbf{z}'}$ as probability measures. Since $\mathcal{C}_j^{\text{can}}$ is separating, there exists a set $S_{\mathbf{z},\mathbf{z}'} \in \mathcal{C}_j^{\text{can}}$ such that $\mathbb{P}_{j,\mathbf{z}}(S_{\mathbf{z},\mathbf{z}'}) \neq \mathbb{P}_{j,\mathbf{z}'}(S_{\mathbf{z},\mathbf{z}'})$. By the enumeration $\mathcal{C}_j^{\text{can}} = \{T_{1,j}, T_{2,j}, \dots\}$, we may write $S_{\mathbf{z},\mathbf{z}'} = T_{m(\mathbf{z},\mathbf{z}'),j}$ for some index $m(\mathbf{z},\mathbf{z}') \geq 1$. Consequently, for every $t \geq m(\mathbf{z},\mathbf{z}')$, we have $S_{\mathbf{z},\mathbf{z}'} \in \overline{\mathcal{D}}_j^{(t)}$, and therefore Condition B holds for this pair $(\mathbf{z}, \mathbf{z}')$.

Since there are finitely many pairs $(\mathbf{z}, \mathbf{z}')$ with $\mathbf{z} \neq \mathbf{z}'$, define $t_3 := \max_{\mathbf{z} \neq \mathbf{z}'} m(\mathbf{z}, \mathbf{z}') < \infty$ and $\mathcal{N}_1^{(3)} := \bigcup_{\mathbf{z} \neq \mathbf{z}'} \mathcal{N}_{\mathbf{z},\mathbf{z}'}^{(3)}$. Then $\mathcal{N}_1^{(3)}$ has Lebesgue measure zero in $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$ and constrains only the measurement parameters. Moreover, for every $t \geq t_3$ and every $\Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) \setminus \mathcal{N}_1^{(3)}$, Condition B holds for the discretization $\overline{\mathcal{D}}^{(t)}$.

Finally, Condition B implies that for every $\mathbf{z} \neq \mathbf{z}'$ there exists a row of $\mathbf{N}_3^{(t)}$ (using $S_{\mathbf{z},\mathbf{z}'}$ for item j and \mathcal{X} for all other items in block 3) on which the \mathbf{z} -th and \mathbf{z}' -th columns differ, hence all columns are pairwise distinct. Together with the fact that $\mathbf{N}_3^{(t)}$ contains the all- \mathcal{X} row $\mathbf{1}_{2K}^\top$, this rules out collinearity of any two columns and yields

$$\text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2, \quad \text{for all } t \geq t_3, \Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) \setminus \mathcal{N}_1^{(3)}.$$

For each $t \geq \max\{t_1, t_2, t_3\}$ define $\mathcal{N}_1^{(t)} := \mathcal{N}_{1,t}^{(1)} \cup \mathcal{N}_{1,t}^{(2)} \cup \mathcal{N}_1^{(3)}$. By construction $\mathcal{N}_1^{(t)}$ has Lebesgue measure zero in $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$ and depends only on (\mathbf{B}, γ) , not on \mathbf{p} .

Now define the global exceptional set

$$\mathcal{N}_\infty := \left(\bigcup_{t \geq \max\{t_1, t_2, t_3\}} \mathcal{N}_{1,t}^{(1)} \right) \cup \left(\bigcup_{t \geq \max\{t_1, t_2, t_3\}} \mathcal{N}_{1,t}^{(2)} \right) \cup \mathcal{N}_1^{(3)}.$$

Since $\{\mathcal{N}_{1,t}^{(1)}\}_{t \geq \max\{t_1, t_2, t_3\}}$ and $\{\mathcal{N}_{1,t}^{(2)}\}_{t \geq \max\{t_1, t_2, t_3\}}$ are countable families of Lebesgue-null sets, the union \mathcal{N}_∞ is also Lebesgue-null, and it still constrains only (\mathbf{B}, γ) .

Fix $\Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) \setminus \mathcal{N}_\infty$. Define $t_0(\Theta) := \max\{t_1, t_2, t_3\}$. Then for all $t \geq t_0(\Theta)$ we

have simultaneously

$$\text{rk}_k(\mathbf{N}_1^{(t)}) = 2^K, \quad \text{rk}_k(\mathbf{N}_2^{(t)}) = 2^K, \quad \text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2.$$

This completes the proof.

S.3.2 Proof of Lemma S.2

To avoid ambiguity, we fix the identification

$$f_b : \{0, 1\}^K \longrightarrow \mathcal{P}([K]), \quad f_b(v) := \{k \in [K] : v_k = 1\}.$$

Conversely, for a subset $T \subseteq [K]$, we write $(f_b)^{-1}(T) \in \{0, 1\}^K$ for its indicator vector.

A permutation $\pi \in S_K$ acts on subsets by $\pi(T) := \{\pi(i) : i \in T\} \subseteq [K]$, and induces the corresponding action on vectors by permuting coordinates:

$$(\pi \cdot v)_k = v_{\pi^{-1}(k)} \quad (v \in \{0, 1\}^K, k \in [K]).$$

We first establish that $\mathbf{Q} \sim_{\mathcal{K}} \mathbf{Q}'$. Equivalently, there exists a column permutation $\pi \in S_K$ such that

$$\mathbf{Q}'_{j,:} = \pi \cdot \mathbf{Q}_{j,:} \quad \text{for all } j \in \{1, \dots, J\}.$$

In fact, for $(\boldsymbol{\eta}_j(\mathbf{z}))_{1 \leq j \leq J, \mathbf{z} \in \{0, 1\}^K} \in \mathbb{R}^{2^K \times J}$ and $(\boldsymbol{\eta}'_j(\mathbf{z}))_{1 \leq j \leq J, \mathbf{z} \in \{0, 1\}^K} \in \mathbb{R}^{2^K \times J}$ that differ only by a row permutation, the counts of coordinatewise maximal rows are preserved. In the single-index case $S = \{j\}$, a row indexed by $\mathbf{z} \in \{0, 1\}^K$ is maximal in the $2^K \times 1$ submatrix $\boldsymbol{\eta}_{:, \{j\}}$ if and only if

$$\mathbf{z} \succeq (f_b)^{-1}(K_j),$$

hence the number of maximal rows equals $2^{K-|K_j|}$. The same conclusion holds for $\boldsymbol{\eta}'$, so $2^{K-|K_j|} = 2^{K-|K'_j|}$ and therefore $|K_j| = |K'_j|$.

For any distinct j, j' , a row is coordinatewise maximal in the $2^K \times 2$ submatrix $\boldsymbol{\eta}_{\cdot, \{j, j'\}}$ if and only if

$$\mathbf{z} \succeq (f_b)^{-1}(K_j) \quad \text{and} \quad \mathbf{z} \succeq (f_b)^{-1}(K_{j'}),$$

which is equivalent to $\mathbf{z} \succeq (f_b)^{-1}(K_j \cup K_{j'})$. Consequently, the number of maximal rows is $2^{K-|K_j \cup K_{j'}|}$, and by row-permutation invariance we obtain

$$|K_j \cup K_{j'}| = |K'_j \cup K'_{j'}|.$$

More generally, for any finite $S \subseteq \{1, \dots, J\}$, a row is coordinatewise maximal in $\boldsymbol{\eta}_{\cdot, S}$ if and only if

$$\mathbf{z} \succeq (f_b)^{-1}\left(\bigcup_{j \in S} K_j\right),$$

so the number of maximal rows is $2^{K-|\cup_{j \in S} K_j|}$, which is invariant under row permutations.

Hence

$$\left|\bigcup_{j \in S} K_j\right| = \left|\bigcup_{j \in S} K'_j\right| \quad \text{for all } S \subseteq \{1, \dots, J\}.$$

For $T \subseteq [J]$, write

$$N_T := \left|\bigcap_{j \in T} K_j\right| \quad \text{and} \quad N'_T := \left|\bigcap_{j \in T} K'_j\right|.$$

It is straightforward to check

$$\left|\bigcup_{j \in S} K_j\right| = \sum_{\emptyset \neq T \subseteq S} (-1)^{|T|+1} N_T \quad \text{for all } S \subseteq [J],$$

and

$$\left|\bigcup_{j \in S} K'_j\right| = \sum_{\emptyset \neq T \subseteq S} (-1)^{|T|+1} N'_T \quad \text{for all } S \subseteq [J].$$

Since the union cardinalities match for all S , we deduce that

$$N_T = N'_T \quad \text{for all } T \subseteq [J].$$

Next, for each pattern $v \in \{0, 1\}^J$ with support $\text{supp}(v) := \{j : v_j = 1\}$, define

$$M_v := |\{k \in [K] : \mathbf{Q}_{:,k} = v\}|, \quad M'_v := |\{k \in [K] : \mathbf{Q}'_{:,k} = v\}|.$$

The intersection counts decompose as

$$N_T = \sum_{v: \text{supp}(v) \supseteq T} M_v \quad \text{for all } T \subseteq [J],$$

and likewise for N'_T in terms of M'_v . Similarly, we have

$$M_v = M'_v \quad \text{for all } v \in \{0, 1\}^J.$$

Hence the multisets of columns of \mathbf{Q} and \mathbf{Q}' coincide. Equivalently, there exists a column permutation $\pi \in S_K$ such that

$$\mathbf{Q}'_{j,:} = \pi \cdot \mathbf{Q}_{j,:} \quad \text{for all } j \in \{1, \dots, J\}.$$

Next, we have the following claim.

Claim 1. *Suppose that \mathbf{Q} satisfies the conditions of Theorem 1, and that $\mathbf{Q}'_{j,:} = \pi \cdot \mathbf{Q}_{j,:}$ for all j , then the admissible column permutation $\mathfrak{S} \in S_{2^K}$ (acting on the 2^K latent states) such that $\eta_{j,\mathbf{z}} = \eta'_{j,\mathfrak{S}(\mathbf{z})}$ ($\forall j, \mathbf{z}$) can be restricted to the right coset $(\mathbb{Z}_2)^K \pi$. In other words, every admissible \mathfrak{S} can be written in the form*

$$f_b(\mathfrak{S}((f_b)^{-1}(T))) = \pi(T) \Delta A, \quad T \subseteq [K],$$

for some subset $A \subseteq [K]$, where $A \Delta B = (A \setminus B) \cup (B \setminus A)$ for any two sets A and B

If we further suppose that both (\mathbf{B}, \mathbf{Q}) and $(\mathbf{B}', \mathbf{Q}')$ satisfy Assumption 1(b), then the only admissible permutation is π .

A detailed proof of this reduction is given in the next subsection.

Based on this claim, we directly conclude that $(\mathbf{p}, \mathbf{B}, \mathcal{G}, \mathbf{Q}) \sim_{\kappa} (\mathbf{p}', \mathbf{B}', \mathcal{G}', \mathbf{Q}')$.

S.3.3 Proof of the Claim

We now state a lemma, which is essentially an equivalent reformulation of the first part of the claim in Subsection S.3.2. However, for clarity of exposition, we still restate it in a slightly different language.

Let $\mathcal{P}([K])$ denote the family of all subsets of $[K] = \{1, \dots, K\}$. All 2^K -dimensional vectors and $2^K \times 2^K$ matrices below are indexed by $\mathcal{P}([K])$ in lexicographic order. Define

$$Y_{Q,T} = \mathbf{1}\{T \subseteq Q\}, \quad Q, T \subseteq [K].$$

Given $\boldsymbol{\beta} = (\beta_T)_{T \subseteq [K]} \in \mathbb{R}^{2^K}$, set

$$\boldsymbol{\eta} = \mathbf{Y}\boldsymbol{\beta}, \quad \text{so that } \eta_Q = \sum_{T \subseteq Q} \beta_T.$$

Let $f : \mathcal{P}([K]) \rightarrow \mathcal{P}([K])$ be a bijection, and let \mathbf{P}_f be its associated permutation matrix acting on coordinates indexed by $\mathcal{P}([K])$:

$$(\mathbf{P}_f v)_Q = v_{f^{-1}(Q)}, \quad v \in \mathbb{R}^{2^K}.$$

Hence \mathbf{P}_f simply reorders the 2^K coordinates of any vector according to f , and thus $\mathbf{P}_f \in S_{2^K}$. We then define

$$\boldsymbol{\eta}' = \mathbf{P}_f \boldsymbol{\eta}, \quad \boldsymbol{\beta}' = \mathbf{Y}^{-1} \mathbf{P}_f \mathbf{Y} \boldsymbol{\beta}.$$

For any $Q \subseteq [K]$, define

$$\Pi_Q = \text{Diag}(\mathbf{1}\{Q' \subseteq Q\})_{Q' \subseteq [K]} \in \mathbb{R}^{2^K \times 2^K},$$

$$E_Q = \text{Im}(\Pi_Q) = \{\mathbf{x} \in \mathbb{R}^{2^K} : x_{Q'} = 0 \text{ if } Q' \not\subseteq Q\}. \quad (\text{S.4})$$

Let $\mathcal{F} = \{Q_1, \dots, Q_l\} \subseteq \mathcal{P}([K])$ be a family of subsets of $[K]$, and let $\pi \in S_K$ be a permutation on $\{1, \dots, K\}$. For each subset $Q \subseteq [K]$, write

$$\pi(Q) := \{\pi(i) : i \in Q\} \subseteq [K],$$

so that π acts naturally on $\mathcal{P}([K])$. We define the set of row permutations preserving the corresponding subspaces:

$$G_\pi(Q) = \{\mathbf{P}_f : \mathbf{Y}^{-1} \mathbf{P}_f \mathbf{Y} E_Q \subseteq E_{\pi(Q)}\},$$

and their intersection over the family \mathcal{F} :

$$H_\pi(\mathcal{F}) = \bigcap_{Q \in \mathcal{F}} G_\pi(Q) = \bigcap_{i=1}^l G_\pi(Q_i).$$

To describe coordinate structure, define the signature map

$$\phi_{\mathcal{F}} : [K] \rightarrow \{0, 1\}^l, \quad \phi_{\mathcal{F}}(i) := (\mathbf{1}\{i \in Q_j\})_{1 \leq j \leq l}.$$

Lemma S.4. *Assume that for every $i \in [K]$ there exists some $Q \in \mathcal{F}$ such that $i \notin Q$, and that for any distinct $i, j \in [K]$, neither $\phi_{\mathcal{F}}(i) \preceq \phi_{\mathcal{F}}(j)$ nor $\phi_{\mathcal{F}}(j) \preceq \phi_{\mathcal{F}}(i)$ holds. Then the intersection $H_\pi(\mathcal{F})$ coincides with the coset of $(\mathbb{Z}_2)^K \rtimes S_K$ corresponding to π , namely*

$$H_\pi(\mathcal{F}) = (\mathbb{Z}_2)^K \pi,$$

where $(\mathbb{Z}_2)^K$ denotes the subgroup of coordinatewise bit-flip operations

$$T \mapsto T \Delta A, \quad A \subseteq [K].$$

Equivalently, every bijection $f \in H_\pi(\mathcal{F})$ is uniquely of the form

$$f(T) = \pi(T) \Delta A, \quad A \subseteq [K].$$

Now we are ready to give the proof of Lemma S.4. We need three propositions.

Proposition S.1. *Let $Q \subseteq [K]$ and, for each $U \in \mathcal{P}(Q)$, define the Q -blocks*

$$B^Q(U) := \{T \subseteq [K] : T \cap Q = U\} \subseteq \mathcal{P}([K]).$$

Recall that $f : \mathcal{P}([K]) \rightarrow \mathcal{P}([K])$ is a bijection and \mathbf{P}_f is its associated permutation matrix acting by $(\mathbf{P}_f v)_Q = v_{f^{-1}(Q)}$. Then the following are equivalent:

- (a) $\mathbf{P}_f \in G_\pi(Q)$.
- (b) \exists a bijection $g : \mathcal{P}(Q) \rightarrow \mathcal{P}(\pi(Q))$ such that $f(B^Q(U)) = B^{\pi(Q)}(g(U)) \quad \forall U \in \mathcal{P}(Q)$.

Proof. Recall $E_Q = \{\mathbf{x} \in \mathbb{R}^{2^K} : x_{Q'} = 0 \text{ if } Q' \not\subseteq Q\}$ (See (S.4)). For $\boldsymbol{\beta} \in E_Q$ and any $S \subseteq [K]$,

$$(\mathbf{Y}\boldsymbol{\beta})_S = \eta_S = \sum_{T \subseteq S} \beta_T = \sum_{T \subseteq S \cap Q} \beta_T,$$

so the image can be written as

$$Y_Q := \mathbf{Y}E_Q = \left\{ y \in \mathbb{R}^{2^K} : \exists h : \mathcal{P}(Q) \rightarrow \mathbb{R} \text{ such that } y_S = h(S \cap Q) \quad \forall S \subseteq [K] \right\}.$$

Similarly,

$$Y_{\pi(Q)} = \left\{ y \in \mathbb{R}^{2^K} : \exists \tilde{h} : \mathcal{P}(\pi(Q)) \rightarrow \mathbb{R} \text{ such that } y_S = \tilde{h}(S \cap \pi(Q)) \quad \forall S \subseteq [K] \right\}.$$

Since \mathbf{Y} is invertible, $\mathbf{Y}^{-1}\mathbf{P}_f\mathbf{Y}E_Q \subseteq E_{\pi(Q)}$ if and only if $\mathbf{P}_fY_Q \subseteq Y_{\pi(Q)}$.

Thus $\mathbf{P}_f \in G_\pi(Q)$ if and only if \mathbf{P}_f maps vectors that are constant on each $B^Q(U)$ to vectors that are constant on each $B^{\pi(Q)}(V)$. This holds if and only if f maps each block $B^Q(U)$, as a set, onto some block $B^{\pi(Q)}(V)$. Because $\{B^Q(U)\}_{U \in \mathcal{P}(Q)}$ is a partition and f is a bijection, these images define a unique bijection $g : \mathcal{P}(Q) \rightarrow \mathcal{P}(\pi(Q))$ with

$$f(B^Q(U)) = B^{\pi(Q)}(g(U)) \quad \forall U \in \mathcal{P}(Q).$$

This proves (a) and (b) are equivalent. □

Proposition S.2. *For any \mathcal{F} , $(\mathbb{Z}_2)^K \pi \subseteq H_\pi(\mathcal{F})$.*

Proof. Fix $A \subseteq [K]$ and define $\tau_A : \mathcal{P}([K]) \rightarrow \mathcal{P}([K])$ by

$$\tau_A(T) := \pi(T) \Delta A.$$

We claim that for every $Q \subseteq [K]$ and every $U \in \mathcal{P}(Q)$,

$$\tau_A(B^Q(U)) = B^{\pi(Q)}(\pi(U) \Delta (A \cap \pi(Q))).$$

Indeed, if $T \in B^Q(U)$ then $T \cap Q = U$, and by the identity $(X \Delta Y) \cap Z = (X \cap Z) \Delta (Y \cap Z)$ we have

$$(\pi(T) \Delta A) \cap \pi(Q) = (\pi(T) \cap \pi(Q)) \Delta (A \cap \pi(Q)) = \pi(T \cap Q) \Delta (A \cap \pi(Q)) = \pi(U) \Delta (A \cap \pi(Q)),$$

which is independent of the particular T in the block and depends only on U . Thus τ_A maps

the block $B^Q(U)$ onto the block $B^{\pi(Q)}(\pi(U)\Delta(A \cap \pi(Q)))$, so the associated permutation matrix P_{τ_A} satisfies $P_{\tau_A} \in G_\pi(Q)$ by Proposition S.1. Since $Q \subseteq [K]$ was arbitrary, we have

$$P_{\tau_A} \in \bigcap_{Q \in \mathcal{F}} G_\pi(Q) = H_\pi(\mathcal{F}).$$

Finally, the set $\{\tau_A : A \subseteq [K]\}$ is exactly the right coset $(\mathbb{Z}_2)^K \pi$, hence $(\mathbb{Z}_2)^K \pi \subseteq H_\pi(\mathcal{F})$. \square

Proposition S.3. *Assume that for any distinct $i, j \in [K]$, neither $\phi_{\mathcal{F}}(i) \preceq \phi_{\mathcal{F}}(j)$ nor $\phi_{\mathcal{F}}(j) \preceq \phi_{\mathcal{F}}(i)$ holds, then we have*

$$H_\pi(\mathcal{F}) \subseteq (\mathbb{Z}_2)^K \pi.$$

Proof. Throughout we identify a permutation matrix \mathbf{P}_f with its underlying bijection $f : \mathcal{P}([K]) \rightarrow \mathcal{P}([K])$ via $(\mathbf{P}_f v)_S = v_{f^{-1}(S)}$. Fix $i \in [K]$ and define, for $T \subseteq [K]$,

$$D_i(T) := f(T) \Delta f(T \Delta \{\pi^{-1}(i)\}).$$

We now state the following proposition, which characterizes $D_i(T)$ and will be used below.

Proposition S.4. *Under the assumptions of Proposition S.3, for every $\mathbf{P}_f \in H_\pi(\mathcal{F})$, every $i \in [K]$, and every $T \subseteq [K]$, one has*

$$D_i(T) = \{i\}.$$

With Proposition S.4, take any $j \in [K]$ and $T \subseteq [K]$. Substituting $i = \pi(j)$ in the definition of $D_i(T)$ yields

$$f(T \Delta \{j\}) = f(T) \Delta \{\pi(j)\}.$$

Let $A := f(\emptyset)$, we will prove that $f(T) = A \Delta \pi(T)$ by induction on $m := |T|$.

1. Base case $m = 0$. For $T = \emptyset$, we have $f(\emptyset) = A = A \Delta \pi(\emptyset)$.
2. Inductive step. Assume $f(S) = A \Delta \pi(S)$ holds for all $S \subseteq [K]$ with $|S| = m$. Let $T = S \cup \{j\}$ with $j \notin S$. By Proposition S.4 with $i = \pi(j)$,

$$f(S \Delta \{j\}) = f(S) \Delta \{\pi(j)\}.$$

Since $S \Delta \{j\} = S \cup \{j\} = T$ and $\pi(T) = \pi(S \cup \{j\}) = \pi(S) \Delta \{\pi(j)\}$, we obtain

$$f(T) = f(S) \Delta \{\pi(j)\} = (A \Delta \pi(S)) \Delta \{\pi(j)\} = A \Delta (\pi(S) \Delta \{\pi(j)\}) = A \Delta \pi(T).$$

This completes the induction.

Hence for every $\mathbf{P}_f \in H_\pi(\mathcal{F})$, the underlying bijection has the form $f(T) = A \Delta \pi(T)$. In other words, $H_\pi(\mathcal{F}) \subseteq (\mathbb{Z}_2)^K \pi$.

□

Combining Proposition S.3 with Proposition S.2, we complete the proof of Lemma S.4

Based on this lemma, we can obtain the sum of each column of \mathbf{B}' . Let $\mathbf{1}$ be the all-ones vector. For any $S \subseteq [K]$ and $j \in \{1, \dots, J\}$, using $\mathbf{1}^\top \mathbf{Y}^{-1} = e_{[K]}^\top$, it follows that

$$\sum_S \beta'_{j,S} = \mathbf{1}^\top \boldsymbol{\beta}' = e_{[K]}^\top \mathbf{P}_f \mathbf{Y} \boldsymbol{\beta}_j = e_{f([K])}^\top \mathbf{Y} \boldsymbol{\beta}_j = \sum_{U \subseteq f([K])} \beta_{j,U}.$$

Since $f([K]) = \pi([K]) \Delta A = [K] \setminus A$, we obtain:

$$\sum_S \beta'_{j,S} = \sum_{U \subseteq [K] \setminus A} \beta_{j,U}.$$

Now since both (\mathbf{B}, \mathbf{Q}) and $(\mathbf{B}', \mathbf{Q}')$ both satisfy Assumption 1(b), we must have

$$\sum_S \beta'_{j,S} = \sum_S \beta_{j,S} \quad j = 1, \dots, J.$$

This forces $A \cap K_j = \emptyset$ for all $j = 1, \dots, J$. Consequently, $A = \emptyset$ and $f(T) = \pi(T)$ is the only admissible permutation.

S.3.4 Proof of Proposition S.4

We first claim that, for all $f \in H_\pi(\mathcal{F})$ and all i, T ,

$$D_i(T) \subseteq \bigcap_{\substack{Q \in \mathcal{F} \\ \pi^{-1}(i) \notin Q}} (\pi(Q))^c = \{j \in [K] : \phi_{\mathcal{F}}(\pi^{-1}(j)) \preceq \phi_{\mathcal{F}}(\pi^{-1}(i))\}.$$

Note that by the assumption that for every $i \in [K]$ there must exist some $Q \in \mathcal{F}$ such that $i \notin Q$, the index set $\{Q \in \mathcal{F} : \pi^{-1}(i) \notin Q\}$ is nonempty. Hence the intersection is taken over a nonempty family

In fact, let $Q \in \mathcal{F}$ satisfy $\pi^{-1}(i) \notin Q$. Then T and $T\Delta\{\pi^{-1}(i)\}$ lie in the same Q -block:

$$T \cap Q = (T\Delta\{\pi^{-1}(i)\}) \cap Q.$$

Since $f \in H_\pi(\mathcal{F})$, Proposition S.1 implies that f maps each Q -block onto a $\pi(Q)$ -block.

Hence

$$f(T) \cap \pi(Q) = f(T\Delta\{\pi^{-1}(i)\}) \cap \pi(Q),$$

and therefore

$$D_i(T) \cap \pi(Q) = \emptyset.$$

As this holds for every $Q \in \mathcal{F}$ with $\pi^{-1}(i) \notin Q$, we obtain

$$D_i(T) \subseteq \bigcap_{\substack{Q \in \mathcal{F} \\ \pi^{-1}(i) \notin Q}} (\pi(Q))^c. \tag{S.5}$$

Next we rewrite the right-hand side using signature vectors. Fix $j \in [K]$,

$$j \in \bigcap_{\{Q \in \mathcal{F}: \pi^{-1}(i) \notin Q\}} (\pi(Q))^c$$

means precisely that for every $Q \in \mathcal{F}$ with $\pi^{-1}(i) \notin Q$ we have $j \notin \pi(Q)$. By the definition of the signature map, the condition $\pi^{-1}(i) \notin Q$ is the same as $\phi_{\mathcal{F}}(\pi^{-1}(i))_Q = 0$, and $j \notin \pi(Q)$ is the same as $\phi_{\mathcal{F}}(\pi^{-1}(j))_Q = 0$. Therefore the preceding sentence says: for all $Q \in \mathcal{F}$, $\phi_{\mathcal{F}}(\pi^{-1}(j)) \preceq \phi_{\mathcal{F}}(\pi^{-1}(i))$. Hence

$$\bigcap_{\substack{Q \in \mathcal{F} \\ \pi^{-1}(i) \notin Q}} (\pi(Q))^c = \{j \in [K] : \phi_{\mathcal{F}}(\pi^{-1}(j)) \preceq \phi_{\mathcal{F}}(\pi^{-1}(i))\},$$

and the claim follows from (S.5).

By assumption, for each i we have

$$\{j \in [K] : \phi_{\mathcal{F}}(\pi^{-1}(j)) \preceq \phi_{\mathcal{F}}(\pi^{-1}(i))\} = \{i\}.$$

Therefore, by (S.5) we already proved above,

$$D_i(T) \subseteq \{j : \phi_{\mathcal{F}}(\pi^{-1}(j)) \preceq \phi_{\mathcal{F}}(\pi^{-1}(i))\} = \{i\}.$$

Since f is a bijection and $T \neq T \Delta \{\pi^{-1}(i)\}$, we have $D_i(T) \neq \emptyset$, hence $D_i(T) = \{i\}$.

S.4 Extension to the polytomous-attribute case

Similar to He et al. (2023), to encode the intrinsic ordering of levels, we also use cumulative threshold indicators $\mathcal{I}_{k,u}(\mathbf{Z}) = \mathbf{1}\{Z_k \geq u\}$ for $u \in \{1, \dots, M_k - 1\}$ and form the Kronecker feature map

$$\Phi(\mathbf{Z}) = \otimes_{k=1}^K (1, \mathcal{I}_{k,1}(\mathbf{Z}), \dots, \mathcal{I}_{k,M_k-1}(\mathbf{Z}))^\top \in \mathbb{R}^r.$$

Then each item j has a predictor

$$\eta_j(\mathbf{Z}) = \langle \boldsymbol{\beta}_j, \Phi(\mathbf{Z}) \rangle = \sum_{\mathbf{u} \in \mathcal{Z}} \beta_{j,\mathbf{u}} \prod_{k=1}^K \mathcal{I}_{k,u_k}(\mathbf{Z}),$$

where $\boldsymbol{\beta}_j := (\beta_{j,\mathbf{u}})_{\mathbf{u} \in \mathcal{Z}} \in \mathbb{R}^r$.

The measurement structure is encoded directly by a binary matrix $\mathbf{Q} = (q_{j,k}) \in \{0, 1\}^{J \times K}$.

We impose the structural restriction that coefficients vanish outside the relevant coordinates,

$$\beta_{j,\mathbf{u}} = 0 \quad \text{whenever} \quad \text{supp}(\mathbf{u}) \not\subseteq K_j,$$

and require a full main-effect ladder whenever $q_{j,k} = 1$. Specifically, for each $k \in [K]$ and each threshold $v \in \{1, \dots, M_k - 1\}$ define the main-effect index $\mathbf{u}(k, v) = (u_h)_{h \in [K]}$ with $u_k = v$ and $u_h = 0$ for $h \neq k$. We assume

$$q_{j,k} = 1 \iff \beta_{j,\mathbf{u}(k,v)} \neq 0 \text{ for all } v \in \{1, \dots, M_k - 1\}, \quad (j \in [J], k \in [K]),$$

while interaction coordinates \mathbf{u} with $|\text{supp}(\mathbf{u})| \geq 2$ and $\text{supp}(\mathbf{u}) \subseteq K_j$ are allowed to be nonzero.

Because the latent variables may have unequal numbers of categories, admissible relabelings must preserve numbers of categories. Define the permutation group

$$S_{K,\mathbf{M}} := \left\{ \varpi \in S_K : M_{\varpi(k)} = M_k \quad \forall k \in [K] \right\}. \quad (\text{S.6})$$

Given $\varpi \in S_{K,\mathbf{M}}$, define the induced bijection $\sigma_\varpi : \mathcal{Z} \rightarrow \mathcal{Z}$ by

$$\sigma_\varpi(z_1, \dots, z_K) := (z_{\varpi^{-1}(1)}, \dots, z_{\varpi^{-1}(K)}), \quad (\text{S.7})$$

and let P_ϖ be the induced $r \times r$ Kronecker permutation matrix such that

$$\Phi(\sigma_\varpi(\mathbf{z})) = P_\varpi \Phi(\mathbf{z}), \quad \mathbf{z} \in \mathcal{Z}. \quad (\text{S.8})$$

We say that $(\mathbf{p}, \mathcal{G}, \mathbf{Q}, \{\beta_j\}_{j \in [J]}, \gamma)$ and $(\mathbf{p}', \mathcal{G}', \mathbf{Q}', \{\beta'_j\}_{j \in [J]}, \gamma')$ are equivalent, denoted $\sim_{K, \mathbf{M}}^{\text{ord}}$, if and only if $\gamma = \gamma'$ and there exists $\varpi \in S_{K, \mathbf{M}}$ such that, with $\sigma = \sigma_\varpi$,

$$p_{\mathbf{z}} = p'_{\sigma(\mathbf{z})} \quad \forall \mathbf{z} \in \mathcal{Z}, \quad \beta'_j = P_\varpi \beta_j \quad \forall j \in [J], \quad q'_{j,k} = q_{j, \varpi^{-1}(k)} \quad \forall (j, k) \in [J] \times [K], \quad (\text{S.9})$$

and the relabeled DAG $\varpi(\mathcal{G})$ is Markov equivalent to \mathcal{G}' .

Let $\Theta := (\mathbf{p}, \mathbf{B}, \gamma)$. Define the admissible parameter space

$$\begin{aligned} \Omega_K(\Theta; \mathcal{G}, \mathbf{Q}) &:= \left\{ \Theta : \mathcal{G} \text{ is a perfect map of } \mathbf{p}, \beta_{j, \mathbf{u}} = 0 \text{ if } \text{supp}(\mathbf{u}) \not\subseteq K_j, \right. \\ &\quad \left. q_{j,k} = 1 \text{ iff } \beta_{j, \mathbf{u}(k,v)} \neq 0 \text{ for all } v \in \{1, \dots, M_k - 1\} \right\}, \\ \Omega_K(\Theta, \mathcal{G}, \mathbf{Q}) &:= \left\{ (\Theta, \mathcal{G}, \mathbf{Q}) : \Theta \in \Omega_K(\Theta; \mathcal{G}, \mathbf{Q}) \right\}. \end{aligned}$$

Definition 7. Let $(\Theta^*, \mathcal{G}^*, \mathbf{Q}^*) \in \Omega_K(\Theta, \mathcal{G}, \mathbf{Q})$ be the true parameter triple. The framework is generically identifiable up to $\sim_{K, \mathbf{M}}^{\text{ord}}$ if

$$\left\{ \Theta \in \Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*) : \exists (\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}}) \not\sim_{K, \mathbf{M}}^{\text{ord}} (\Theta, \mathcal{G}^*, \mathbf{Q}^*) \text{ such that } \mathbb{P}_{\tilde{\Theta}, \tilde{\mathcal{G}}, \tilde{\mathbf{Q}}} = \mathbb{P}_{\Theta, \mathcal{G}^*, \mathbf{Q}^*} \right\} \quad (\text{S.10})$$

is a Lebesgue-null subset of $\Omega_K(\Theta; \mathcal{G}^*, \mathbf{Q}^*)$.

For condition (i) in Theorem 2, we can equivalently rewrite it as follows:

- (i) The item index set can be partitioned as $\mathcal{J}_1 = \{1, \dots, \tilde{d}\}$, $\mathcal{J}_2 = \{\tilde{d} + 1, \dots, 2\tilde{d}\}$, and $\mathcal{J}_3 = \{2\tilde{d} + 1, \dots, J\}$, and there exist bijections: $\rho_1 : \mathcal{J}_1 \rightarrow \{(k, s) : k \in [K], s \in [m_k]\}$ and $\rho_2 : \mathcal{J}_2 \rightarrow \{(k, s) : k \in [K], s \in [m_k]\}$. For each $a \in \{1, 2\}$ and each $j \in \mathcal{J}_a$, write $\rho_a(j) = (k(j), s(j))$. $q_{j, k(j)} = 1$ for all $j \in \mathcal{J}_1 \cup \mathcal{J}_2$, and for each $k \in [K]$ there exists

$j_k \in \mathcal{J}_3$ such that $q_{j_k, k} = 1$.

We will use ρ_1 and ρ_2 later to index the items in the first two blocks.

S.5 Proof of Theorem 2

S.5.1 Kruskal reduction via a parameter-independent refinement scheme

Fix an enumeration of $\mathcal{C}_j^{\text{can}} \setminus \{\mathcal{X}_j\}$ as

$$\mathcal{C}_j^{\text{can}} \setminus \{\mathcal{X}_j\} = \{T_{1,j}, T_{2,j}, \dots\}, \quad (j \in [J]).$$

For each $t \geq 1$, define the parameter-independent finite discretization

$$\overline{\mathcal{D}}_j^{(t)} := \{T_{1,j}, \dots, T_{t,j}\} \cup \{\mathcal{X}_j\} \subseteq \mathcal{C}_j^{\text{can}}, \quad \overline{\mathcal{D}}^{(t)} := (\overline{\mathcal{D}}_j^{(t)})_{j \in [J]}.$$

Let $\kappa_j^{(t)} := |\overline{\mathcal{D}}_j^{(t)}| = t + 1$ and index

$$\overline{\mathcal{D}}_j^{(t)} = (S_{1,j}^{(t)}, \dots, S_{\kappa_j^{(t)},j}^{(t)}) \quad \text{with} \quad S_{\kappa_j^{(t)},j}^{(t)} = \mathcal{X}_j.$$

For each $t \geq 1$, define factor matrices $\mathbf{N}_1^{(t)}, \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}$ whose columns are indexed by $\mathbf{z} \in \mathcal{Z}$ as follows.

For $\xi_1 = (\ell_j)_{j \in \mathcal{J}_1}$ with $\ell_j \in [\kappa_j^{(t)}]$ for all $j \in \mathcal{J}_1$, set

$$\mathbf{N}_1^{(t)}(\xi_1, \mathbf{z}) := \mathbb{P}\left(\bigcap_{j \in \mathcal{J}_1} \{X_j \in S_{\ell_j, j}^{(t)}\} \mid \mathbf{z}\right).$$

For $\xi_2 = (\ell_j)_{j \in \mathcal{J}_2}$ with $\ell_j \in [\kappa_j^{(t)}]$ for all $j \in \mathcal{J}_2$, set

$$\mathbf{N}_2^{(t)}(\xi_2, \mathbf{z}) := \mathbb{P}\left(\bigcap_{j \in \mathcal{J}_2} \{X_j \in S_{\ell_j, j}^{(t)}\} \mid \mathbf{z}\right).$$

For $\xi_3 = (\ell_j)_{j \in \mathcal{J}_3}$ with $\ell_j \in [\kappa_j^{(t)}]$ for all $j \in \mathcal{J}_3$, set

$$\mathbf{N}_3^{(t)}(\xi_3, \mathbf{z}) := \mathbb{P}\left(\bigcap_{j \in \mathcal{J}_3} \{X_j \in S_{\ell_j, j}^{(t)}\} \mid \mathbf{z}\right).$$

Define the tensor $\mathbf{P}_0^{(t)}$ by

$$\mathbf{P}_0^{(t)}(\xi_1, \xi_2, \xi_3) = \mathbb{P}\left(\bigcap_{j=1}^J \{X_j \in S_{\ell_j, j}^{(t)}\}\right), \quad \xi_a = (\ell_j)_{j \in \mathcal{J}_a},$$

where $(\ell_j)_{j=1}^J$ is the concatenation of ξ_1, ξ_2, ξ_3 in the natural item order. By conditional independence given \mathbf{z} and the law of total probability,

$$\mathbf{P}_0^{(t)}(\xi_1, \xi_2, \xi_3) = \sum_{\mathbf{z} \in \mathcal{Z}} p_{\mathbf{z}} \mathbf{N}_1^{(t)}(\xi_1, \mathbf{z}) \mathbf{N}_2^{(t)}(\xi_2, \mathbf{z}) \mathbf{N}_3^{(t)}(\xi_3, \mathbf{z}), \quad (\text{S.11})$$

$$\mathbf{P}_0^{(t)} = [\mathbf{N}_1^{(t)} \text{Diag}(\mathbf{p}), \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}].$$

Since $S_{\kappa_j^{(t)}, j}^{(t)} = \mathcal{X}_j$, each $\mathbf{N}_a^{(t)}$ contains a row equal to $\mathbf{1}_r^\top$.

Lemma S.5. *Consider a discrete causal representation learning framework with parameters $(\mathbf{p}^*, \mathcal{G}^*, \mathbf{B}^*, \mathbf{Q}^*, \boldsymbol{\gamma}^*)$ satisfying the conditions of Theorem 2. For each $t \geq 1$, let $\mathbf{P}_0^{(t)}$ be the tensor induced by the parameter-independent discretization $\overline{\mathcal{D}}^{(t)}$ defined above, with factor matrices $\mathbf{N}_1^{(t)}, \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}$, so that $\mathbf{P}_0^{(t)} = [\mathbf{N}_1^{(t)} \text{Diag}(\mathbf{p}), \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}]$. Then there exists a Lebesgue-null set $\mathcal{N}_\infty \subset \Omega_K(\boldsymbol{\Theta}; \mathcal{G}^*, \mathbf{Q}^*)$ which constrains only $(\mathbf{B}, \boldsymbol{\gamma})$ such that the following holds.*

For every $\boldsymbol{\Theta} \in \Omega_K(\boldsymbol{\Theta}; \mathcal{G}^, \mathbf{Q}^*) \setminus \mathcal{N}_\infty$, there exists an integer $t_0 = t_0(\boldsymbol{\Theta}) < \infty$ such that for all $t \geq t_0$ the rank- r CP decomposition of $\mathbf{P}_0^{(t)}$ is unique up to a common column permutation. Moreover, since $\mathcal{X}_j \in \overline{\mathcal{D}}_j^{(t)}$, each $\mathbf{N}_a^{(t)}$ contains a row equal to $\mathbf{1}_r^\top$, hence the uniqueness contains no nontrivial scaling ambiguity.*

Proof. The argument is parallel to the proof of Lemma S.1 in the binary-attribute setting.

The only places changed are the following. First, the latent state space is $\mathcal{Z} = \prod_{k=1}^K [M_k]$ of size $r = \prod_{k=1}^K M_k$, so every instance of 2^K in Lemma S.1 is replaced by r here. Second, the witness construction that certifies $\text{rank}(\mathbf{N}_1^{(t)}) = r$ uses Condition (i) in Theorem 2 to obtain a Kronecker product block built from $m_k = \lceil \log_2 M_k \rceil$ items per coordinate k , rather than the 2×2 blocks in the binary argument.

Finally, the discussion of dispersion parameters γ is exactly the same as in Lemma S.1. When an item family is one-parameter, we do not treat γ_j as a free coordinate. When γ_j is genuinely present, Assumption 2 ensures that $(\eta, \gamma) \mapsto \mathbb{P}_{j, g_j(\eta, \gamma)}(S)$ is real-analytic on $\mathbb{R} \times (0, \infty)$ for each fixed $S \in \overline{\mathcal{D}}_j^{(t)}$. For simplicity, we only treat the second case here.

Fix $t \geq 1$. By Kruskal's theorem, uniqueness holds provided that

$$\text{rk}_k(\mathbf{N}_1^{(t)}) + \text{rk}_k(\mathbf{N}_2^{(t)}) + \text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2r + 2.$$

Thus it suffices to establish, outside a Lebesgue-null exceptional set,

$$\text{rk}_k(\mathbf{N}_1^{(t)}) = r, \quad \text{rk}_k(\mathbf{N}_2^{(t)}) = r, \quad \text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2, \quad (\text{S.12})$$

for all sufficiently large t .

(i) *Generic full column rank for $\mathbf{N}_1^{(t)}$ and $\mathbf{N}_2^{(t)}$.*

For each $i \in \mathcal{J}_1$ define

$$\mathcal{U}_i^{\mathbf{Q}} := \{\mathbf{u} \in \mathcal{Z} : \text{supp}(\mathbf{u}) \subseteq K_i\}. \quad (\text{S.13})$$

Let

$$\mathcal{I}_1^{\mathbf{Q}} := \{(i, \mathbf{u}) : i \in \mathcal{J}_1, \mathbf{u} \in \mathcal{U}_i^{\mathbf{Q}}\}, \quad q_1 := |\mathcal{I}_1^{\mathbf{Q}}|. \quad (\text{S.14})$$

As before, the block parameter vector Θ_1 collects exactly the free coefficients $\boldsymbol{\beta}_1 := (\beta_{i, \mathbf{u}})_{(i, \mathbf{u}) \in \mathcal{I}_1^{\mathbf{Q}}} \in \mathbb{R}^{q_1}$ and the dispersion parameters $\boldsymbol{\gamma}_1 \in (0, \infty)^{|\mathcal{J}_1^\gamma|}$, where \mathcal{J}_1^γ denote the indices of items

among the first block whose response family genuinely includes an unknown dispersion parameter γ_j . In other words, we identify

$$\Theta_1 = (\boldsymbol{\beta}_1, \boldsymbol{\gamma}_1) \in \tilde{\Omega}_1(\mathbf{Q}) := \mathbb{R}^{q_1} \times (0, \infty)^{|\mathcal{J}_1|}. \quad (\text{S.15})$$

Fix t and view $\mathbf{N}_1^{(t)}$ as a function of Θ_1 . By Assumption 2, for each fixed row index ξ_1 and column \mathbf{z} , the entry $\mathbf{N}_1^{(t)}(\xi_1, \mathbf{z})$ is a finite product of analytic maps of the form

$$(\eta, \gamma) \mapsto \mathbb{P}_{j, g_j(\eta, \gamma)}(S), \quad S \in \overline{\mathcal{D}}_j^{(t)},$$

evaluated at analytic functions of the local item parameters in Θ_1 . Hence every entry of $\mathbf{N}_1^{(t)}$ is real-analytic in the ambient Euclidean coordinates of Θ_1 , and so is $f_{1,t}(\Theta_1) := \det((\mathbf{N}_1^{(t)}(\Theta_1))^\top \mathbf{N}_1^{(t)}(\Theta_1))$.

Therefore $f_{1,t}$ is real-analytic on the open connected domain $\tilde{\Omega}_1(\mathbf{Q})$, and hence either $f_{1,t} \equiv 0$ on $\tilde{\Omega}_1(\mathbf{Q})$, or else its zero set is Lebesgue-null (Mityagin, 2015).

The feasible block- \mathcal{J}_1 set in $\Omega_K(\boldsymbol{\Theta}; \mathcal{G}^*, \mathbf{Q}^*)$ still imposes the nonzero ladder constraint for every main effect. Accordingly define

$$\Omega_1(\mathbf{Q}) := \left\{ \Theta_1 \in \tilde{\Omega}_1(\mathbf{Q}) : \beta_{i, \mathbf{u}(k,v)} \neq 0 \quad \forall i \in \mathcal{J}_1, \forall k \in K_i, \forall v \in [M_k - 1] \right\}. \quad (\text{S.16})$$

In Θ_1 -coordinates, $\Omega_1(\mathbf{Q})$ is obtained from $\tilde{\Omega}_1(\mathbf{Q})$ by removing finitely many coordinate hyperplanes $\{\beta_{i, \mathbf{u}(k,v)} = 0\}$, hence $\Omega_1(\mathbf{Q})$ is dense in $\tilde{\Omega}_1(\mathbf{Q})$ and has full Lebesgue measure in it.

Consequently, it suffices to construct a single witness point $\Theta_{1, \text{wit}} \in \tilde{\Omega}_1(\mathbf{Q})$ such that $f_{1,t}(\Theta_{1, \text{wit}}) > 0$ for all sufficiently large t . Indeed, that implies $f_{1,t} \not\equiv 0$ on $\tilde{\Omega}_1(\mathbf{Q})$, hence $\{\Theta_1 \in \tilde{\Omega}_1(\mathbf{Q}) : f_{1,t}(\Theta_1) = 0\}$ is Lebesgue-null. Therefore,

$$\{\Theta_1 \in \Omega_1(\mathbf{Q}) : f_{1,t}(\Theta_1) = 0\} = \{\Theta_1 \in \tilde{\Omega}_1(\mathbf{Q}) : f_{1,t}(\Theta_1) = 0\} \cap \Omega_1(\mathbf{Q}), \quad (\text{S.17})$$

is Lebesgue-null in $\Omega_1(\mathbf{Q})$ as well.

For each $k \in [K]$, fix an injective map

$$c_k : \{0, 1, \dots, M_k - 1\} \rightarrow \{0, 1\}^{m_k}, \quad c_k(a) = (c_{k,1}(a), \dots, c_{k,m_k}(a)).$$

For each $i \in \mathcal{J}_1$, choose numbers $b_{i,0} \neq 0$ and $b_{i,1} \neq 0$ and define a function on $[M_{k(i)}]$ by

$$f_i(a) := b_{i,0} + b_{i,1} c_{k(i),s(i)}(a), \quad a \in [M_{k(i)}].$$

Since $q_{i,k(i)} = 1$ for $i \in \mathcal{J}_1$, all main-effect coordinates $\{\mathbf{u}(k(i), v) : v \in [d_{k(i)}]\}$ are admissible.

We define the witness coefficients by

$$\beta_{i,\mathbf{0}} := f_i(0), \quad \beta_{i,\mathbf{u}(k(i),v)} := f_i(v) - f_i(v-1) \quad (v \in [d_{k(i)}]), \quad \beta_{i,\mathbf{v}} := 0 \text{ for all remaining } \mathbf{v} \in \mathcal{Z}. \quad (\text{S.18})$$

Then for all $\mathbf{z} \in \mathcal{Z}$,

$$\eta_i(\mathbf{z}) = f_i(z_{k(i)}) = b_{i,0} + b_{i,1} c_{k(i),s(i)}(z_{k(i)}). \quad (\text{S.19})$$

For each $i \in \mathcal{J}_1$, define two probability measures on \mathcal{X}_i by

$$\mu_{i,0}(\cdot) := \mathbb{P}_{i, g_i(b_{i,0}, \gamma_i)}(\cdot), \quad \mu_{i,1}(\cdot) := \mathbb{P}_{i, g_i(b_{i,0} + b_{i,1}, \gamma_i)}(\cdot).$$

Since $b_{i,1} \neq 0$, Assumption 2(iii) and Assumption 2(ii) imply $\mu_{i,0} \neq \mu_{i,1}$. Since $\mathcal{C}_i^{\text{can}}$ is separating, we can choose $B_i \in \mathcal{C}_i^{\text{can}}$ such that $\mu_{i,0}(B_i) \neq \mu_{i,1}(B_i)$.

For each $k \in [K]$ and each $s \in [m_k]$, define the unique index $i_{k,s} \in \mathcal{J}_1$ by

$$i_{k,s} := \rho_1^{-1}(k, s), \quad B_{k,s} := B_{i_{k,s}},$$

and set

$$x_{k,s,0} := \mu_{i_{k,s},0}(B_{k,s}), \quad x_{k,s,1} := \mu_{i_{k,s},1}(B_{k,s}), \quad s \in [m_k],$$

so that $x_{k,s,0} \neq x_{k,s,1}$.

Choose t_* large enough so that $B_i \in \overline{\mathcal{D}}_i^{(t_*)}$ for all $i \in \mathcal{J}_1$. Fix any $t \geq t_*$. Consider the submatrix $\mathbf{N}_{1,\text{sub}}^{(t)}$ of $\mathbf{N}_1^{(t)}$ obtained by restricting to the $2^{\tilde{d}}$ rows indexed by $\varepsilon = (\varepsilon_{k,s})_{k \in [K], s \in [m_k]} \in \{0, 1\}^{\tilde{d}}$, where for each $i_{k,s}$ we select

$$S_{\ell_{i_{k,s}}, i_{k,s}}^{(t)} = \begin{cases} B_{k,s}, & \varepsilon_{k,s} = 1, \\ \mathcal{X}_{i_{k,s}}, & \varepsilon_{k,s} = 0, \end{cases}$$

and for columns we keep all $\mathbf{z} \in \mathcal{Z}$. Under (S.19) and conditional independence across $j \in \mathcal{J}_1$ given \mathbf{z} ,

$$\mathbf{N}_{1,\text{sub}}^{(t)}(\varepsilon, \mathbf{z}) = \prod_{k=1}^K \prod_{s=1}^{m_k} (1 - \varepsilon_{k,s} + \varepsilon_{k,s} x_{k,s, c_{k,s}(z_k)}), \quad \varepsilon \in \{0, 1\}^{\tilde{d}}, \quad \mathbf{z} \in \mathcal{Z}. \quad (\text{S.20})$$

For each $k \in [K]$, define the $2^{m_k} \times 2^{m_k}$ matrix H_k^{full} indexed by $\varepsilon^{(k)} \in \{0, 1\}^{m_k}$ and $b \in \{0, 1\}^{m_k}$ via

$$H_k^{\text{full}}(\varepsilon^{(k)}, b) := \prod_{s=1}^{m_k} (1 - \varepsilon_s^{(k)} + \varepsilon_s^{(k)} x_{k,s, b_s}).$$

Then

$$H_k^{\text{full}} = \bigotimes_{s=1}^{m_k} \begin{pmatrix} 1 & 1 \\ x_{k,s,0} & x_{k,s,1} \end{pmatrix}, \quad \det(H_k^{\text{full}}) = \prod_{s=1}^{m_k} (x_{k,s,1} - x_{k,s,0})^{2^{m_k-1}} \neq 0,$$

so H_k^{full} is invertible. Let H_k be the $2^{m_k} \times M_k$ submatrix obtained by restricting the columns of H_k^{full} to the subset $\{c_k(a) : a \in [M_k]\} \subseteq \{0, 1\}^{m_k}$. Since c_k is injective, these M_k columns are linearly independent, hence $\text{rank}(H_k) = M_k$.

By (S.20), the matrix $\mathbf{N}_{1,\text{sub}}^{(t)}$ is the Kronecker product of these blocks,

$$\mathbf{N}_{1,\text{sub}}^{(t)} = \bigotimes_{k=1}^K H_k,$$

so we have

$$\text{rank}(\mathbf{N}_{1,\text{sub}}^{(t)}) = \prod_{k=1}^K \text{rank}(H_k) = \prod_{k=1}^K M_k = r.$$

Therefore, at the coefficient choice (S.18), the full matrix $\mathbf{N}_1^{(t)}$ has full column rank r for every $t \geq t_\star$.

The similar real-analytic argument applied to block \mathcal{J}_2 yields a Lebesgue-null set on which $\text{rk}_k(\mathbf{N}_2^{(t)}) = r$ for all $t \geq t_\star$.

(ii) *Eventual* $\text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2$.

Fix $\mathbf{z} \neq \mathbf{z}'$ and choose k with $z_k \neq z'_k$. Let $j_k \in \mathcal{J}_3$ be as in Condition (i) in Theorem 2, so that $q_{j_k,k} = 1$. Set $u_\star := \min\{z_k, z'_k\} + 1 \in [M_k - 1]$ so that $\mathcal{I}_{k,u_\star}(\mathbf{z}) \neq \mathcal{I}_{k,u_\star}(\mathbf{z}')$. Then $\beta_{j_k, \mathbf{u}(k,u_\star)}$ is a free coordinate in the chart for item j_k (subject only to $\beta_{j_k, \mathbf{u}(k,u_\star)} \neq 0$), so $\eta_{j_k}(\mathbf{z}) = \eta_{j_k}(\mathbf{z}')$ defines a proper affine hyperplane in that coefficient space. Hence, outside the union of these hyperplanes over all pairs $\mathbf{z} \neq \mathbf{z}'$, we have $\eta_{j_k}(\mathbf{z}) \neq \eta_{j_k}(\mathbf{z}')$ for every $\mathbf{z} \neq \mathbf{z}'$. By Assumption 2, this implies the corresponding conditional laws differ, hence for all large enough t the columns of $\mathbf{N}_3^{(t)}$ are pairwise distinct, and together with the all- \mathcal{X} row this yields $\text{rk}_k(\mathbf{N}_3^{(t)}) \geq 2$.

(iii) *Kruskal uniqueness for all large t* . Combining the three blocks yields (S.12) for all sufficiently large t outside a Lebesgue-null set, hence uniqueness up to a common column permutation. \square

Fix Θ outside a Lebesgue-null set to be specified and take t large enough so that Lemma S.5 applies. Let $\mathbf{N}_a'^{(t)}$ denote the analogous factor matrices constructed from Θ' using the same discretizations and the same item blocks $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3$. If

$$\mathbf{P}_0^{(t)} = [\mathbf{N}_1^{(t)} \text{Diag}(\mathbf{p}), \mathbf{N}_2^{(t)}, \mathbf{N}_3^{(t)}] = [\mathbf{N}_1'^{(t)} \text{Diag}(\mathbf{p}'), \mathbf{N}_2'^{(t)}, \mathbf{N}_3'^{(t)}],$$

then there exists a permutation $\mathfrak{S}^{(t)} \in S_r$ such that

$$\mathbf{N}_a^{(t)}(\cdot, \mathbf{z}) = \mathbf{N}_a'^{(t)}(\cdot, \mathfrak{S}^{(t)}(\mathbf{z})) \quad (a = 2, 3), \quad (\mathbf{N}_1^{(t)} \text{Diag}(\mathbf{p}))(\cdot, \mathbf{z}) = (\mathbf{N}_1'^{(t)} \text{Diag}(\mathbf{p}'))(\cdot, \mathfrak{S}^{(t)}(\mathbf{z})). \quad (\text{S.21})$$

Let $\xi_{1,\text{all}}$ be the row index of $\mathbf{N}_1^{(t)}$ selecting \mathcal{X}_j for every $j \in \mathcal{J}_1$. Then $\mathbf{N}_1^{(t)}(\xi_{1,\text{all}}, \mathbf{z}) = 1$ and $\mathbf{N}_1'^{(t)}(\xi_{1,\text{all}}, \mathbf{z}) = 1$ for all \mathbf{z} , hence evaluating (S.21) at $\xi_{1,\text{all}}$ yields

$$p_{\mathbf{z}} = p'_{\mathfrak{S}^{(t)}(\mathbf{z})} \quad \forall \mathbf{z} \in \mathcal{Z}. \quad (\text{S.22})$$

Dividing the last identity in (S.21) columnwise by $p_{\mathbf{z}}$ yields

$$\mathbf{N}_1^{(t)}(\cdot, \mathbf{z}) = \mathbf{N}_1'^{(t)}(\cdot, \mathfrak{S}^{(t)}(\mathbf{z})) \quad \forall \mathbf{z} \in \mathcal{Z}. \quad (\text{S.23})$$

In particular, for every $j \in [J]$, every $S \in \overline{\mathcal{D}}_j^{(t)}$, and every $\mathbf{z} \in \mathcal{Z}$,

$$\mathbb{P}_{j,g_j(\eta_j(\mathbf{z}), \gamma_j)}(S) = \mathbb{P}_{j,g_j(\eta_j'(\mathfrak{S}^{(t)}(\mathbf{z}), \gamma_j'))}(S). \quad (\text{S.24})$$

Lemma S.6 (Permutation stability under refinement). *Let $\overline{\mathcal{D}}^{(t)} \subseteq \overline{\mathcal{D}}^{(t+1)}$ be the nested discretizations above, and let $\mathfrak{S}^{(t)}, \mathfrak{S}^{(t+1)}$ be the aligning permutations obtained from Lemma S.5 at levels t and $t+1$. If $\mathbf{N}_1^{(t)}$ has full column rank r , then $\mathfrak{S}^{(t+1)} = \mathfrak{S}^{(t)}$.*

Proof. Because $\overline{\mathcal{D}}^{(t)} \subseteq \overline{\mathcal{D}}^{(t+1)}$, every row event defining $\mathbf{N}_1^{(t)}$ also appears among the rows defining $\mathbf{N}_1^{(t+1)}$. Thus the column $\mathbf{N}_1^{(t)}(\cdot, \mathbf{z})$ is obtained from $\mathbf{N}_1^{(t+1)}(\cdot, \mathbf{z})$ by restricting to a subset of rows. Using (S.23) at levels t and $t+1$ and restricting the level $t+1$ identity to the level t rows yields

$$\mathbf{N}_1'^{(t)}(\cdot, \mathfrak{S}^{(t)}(\mathbf{z})) = \mathbf{N}_1'^{(t)}(\cdot, \mathfrak{S}^{(t+1)}(\mathbf{z})).$$

Since $\mathbf{N}_1^{(t)}$ has full column rank, its columns are pairwise distinct, and the same holds for $\mathbf{N}_1'^{(t)}$ because it is a column permutation of $\mathbf{N}_1^{(t)}$. Therefore $\mathfrak{S}^{(t)}(\mathbf{z}) = \mathfrak{S}^{(t+1)}(\mathbf{z})$ for all \mathbf{z} . \square

On a generic set where $\mathbf{N}_1^{(t)}$ has full column rank for all sufficiently large t , Lemma S.6 implies that $\mathfrak{S}^{(t)}$ is constant for all large t . Denote the common permutation by $\mathfrak{S} \in S_r$.

Now fix any $j \in [J]$ and any set $S \in \mathcal{C}_j^{\text{can}}$. Since $\bigcup_{t \geq 1} \overline{\mathcal{D}}_j^{(t)} = \mathcal{C}_j^{\text{can}}$, there exists t large enough with $S \in \overline{\mathcal{D}}_j^{(t)}$. Then (S.24) implies

$$\mathbb{P}_{j, g_j(\eta_j(\mathbf{z}), \gamma_j)}(S) = \mathbb{P}_{j, g_j(\eta'_j(\mathfrak{S}(\mathbf{z})), \gamma'_j)}(S) \quad \text{for all } \mathbf{z} \in \mathcal{Z}.$$

Because $\mathcal{C}_j^{\text{can}}$ is separating, we conclude

$$\mathbb{P}_{j, g_j(\eta_j(\mathbf{z}), \gamma_j)} = \mathbb{P}_{j, g_j(\eta'_j(\mathfrak{S}(\mathbf{z})), \gamma'_j)} \quad \text{as probability measures on } \mathcal{X}_j, \quad \forall j \in [J], \mathbf{z} \in \mathcal{Z}.$$

Finally, by Assumption 2(ii) and injectivity of g_j in Assumption 2(iii), we obtain

$$\eta_j(\mathbf{z}) = \eta'_j(\mathfrak{S}(\mathbf{z})), \quad \gamma_j = \gamma'_j, \quad \forall j \in [J], \mathbf{z} \in \mathcal{Z}, \quad (\text{S.25})$$

and combining with (S.22) (stabilized to \mathfrak{S}) yields

$$p_{\mathbf{z}} = p'_{\mathfrak{S}(\mathbf{z})} \quad \forall \mathbf{z} \in \mathcal{Z}. \quad (\text{S.26})$$

S.5.2 Recovering \mathbf{Q} and restricting admissible relabelings (yielding Corollary 1)

Lemma S.7. *Fix integers $M_k \geq 2$ and $\mathcal{Z} = \prod_{k=1}^K [M_k]$. Let $\mathbf{Q} = (q_{j,k}) \in \{0, 1\}^{J \times K}$ and define $K_j = \{k : q_{j,k} = 1\}$. For each item j , let $\eta_j : \mathcal{Z} \rightarrow \mathbb{R}$ satisfy the structural restriction*

$$\eta_j(\mathbf{Z}) = \eta_j(\mathbf{Z}') \quad \text{whenever } \mathbf{Z}_{K_j} = \mathbf{Z}'_{K_j}.$$

Assume the generic injectivity condition

$$\eta_j(\mathbf{Z}) \neq \eta_j(\mathbf{Z}') \quad \text{whenever } \mathbf{Z}_{K_j} \neq \mathbf{Z}'_{K_j}. \quad (\text{S.27})$$

Let $(\mathbf{Q}', \{\eta'_j\}_{j=1}^J)$ be another design/predictor pair with the analogous property. Suppose there exists a bijection $\sigma : \mathcal{Z} \rightarrow \mathcal{Z}$ such that

$$\eta_j(\mathbf{Z}) = \eta'_j(\sigma(\mathbf{Z})) \quad \text{for all } j \in [J], \mathbf{Z} \in \mathcal{Z}. \quad (\text{S.28})$$

Assume moreover the no-containment condition for \mathbf{Q}

$$\text{for any } p \neq q, \text{ neither } \mathbf{Q}_{:,p} \succeq \mathbf{Q}_{:,q} \text{ nor } \mathbf{Q}_{:,q} \succeq \mathbf{Q}_{:,p}. \quad (\text{S.29})$$

Then the following conclusions hold.

- (a) Let $C_k := \{j \in [J] : q_{j,k} = 1\}$ denote the support set of column k of \mathbf{Q} . For any $\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}$, define the coordinate-difference set

$$S(\mathbf{Z}, \mathbf{Z}') := \{k \in [K] : Z_k \neq Z'_k\}$$

and the item-difference pattern

$$D(\mathbf{Z}, \mathbf{Z}') := \{j \in [J] : \eta_j(\mathbf{Z}) \neq \eta_j(\mathbf{Z}')\}.$$

Then

$$D(\mathbf{Z}, \mathbf{Z}') = \bigcup_{k \in S(\mathbf{Z}, \mathbf{Z}')} C_k. \quad (\text{S.30})$$

- (b) The collection of inclusion-minimal nonempty sets among $\{D(\mathbf{Z}, \mathbf{Z}') : \mathbf{Z} \neq \mathbf{Z}'\}$ equals $\{C_k : k \in [K]\}$. Consequently, the multiset of columns of \mathbf{Q} is identified from $\{\eta_j(\mathbf{Z})\}$, hence \mathbf{Q} is identified up to a column permutation.

- (c) There exists a permutation $\varpi \in S_K$ such that

$$C_k = C'_{\varpi(k)} \quad \text{for all } k \in [K],$$

equivalently $\mathbf{Q}' = \mathbf{Q}\Pi$ for the permutation matrix Π of ϖ^{-1} .

(d) The relabeling σ must lie in

$$\left(\prod_{k=1}^K S_{M_k} \right) \rtimes S_{K,\mathbf{M}}, \quad S_{K,\mathbf{M}} := \{\varpi \in S_K : M_{\varpi(k)} = M_k \ \forall k\}.$$

More explicitly, with ϖ from (c), there exist permutations $\tau_k \in S_{M_k}$ such that for all $\mathbf{Z} = (Z_1, \dots, Z_K) \in \mathcal{Z}$ and all $k \in [K]$,

$$(\sigma(\mathbf{Z}))_{\varpi(k)} = \tau_k(Z_k).$$

Equivalently, for all $(Z_1, \dots, Z_K) \in \mathcal{Z}$,

$$\sigma(Z_1, \dots, Z_K) = (\tau_{\varpi^{-1}(1)}(Z_{\varpi^{-1}(1)}), \dots, \tau_{\varpi^{-1}(K)}(Z_{\varpi^{-1}(K)})).$$

Proof. We prove (a)–(d) in order.

Proof of (a). Fix $\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}$. For any item j , by the defining restriction of K_j we have $\eta_j(\mathbf{Z}) = \eta_j(\mathbf{Z}')$ whenever $\mathbf{Z}_{K_j} = \mathbf{Z}'_{K_j}$. Conversely, by (S.27), $\eta_j(\mathbf{Z}) \neq \eta_j(\mathbf{Z}')$ whenever $\mathbf{Z}_{K_j} \neq \mathbf{Z}'_{K_j}$. Hence

$$\eta_j(\mathbf{Z}) \neq \eta_j(\mathbf{Z}') \iff \mathbf{Z}_{K_j} \neq \mathbf{Z}'_{K_j} \iff K_j \cap S(\mathbf{Z}, \mathbf{Z}') \neq \emptyset.$$

Therefore

$$D(\mathbf{Z}, \mathbf{Z}') = \{j : K_j \cap S(\mathbf{Z}, \mathbf{Z}') \neq \emptyset\} = \bigcup_{k \in S(\mathbf{Z}, \mathbf{Z}')} \{j : k \in K_j\} = \bigcup_{k \in S(\mathbf{Z}, \mathbf{Z}')} C_k,$$

which is (S.30).

Proof of (b). Let $\mathcal{D} := \{D(\mathbf{Z}, \mathbf{Z}') : \mathbf{Z} \neq \mathbf{Z}'\}$. By (a), every element of \mathcal{D} is a union of

some subcollection of $\{C_k\}_{k=1}^K$. Fix $k \in [K]$ and choose \mathbf{Z}, \mathbf{Z}' that differ only at coordinate k . Then $S(\mathbf{Z}, \mathbf{Z}') = \{k\}$ and (a) gives $D(\mathbf{Z}, \mathbf{Z}') = C_k \in \mathcal{D}$, so every C_k appears in \mathcal{D} .

Now take any nonempty $D \in \mathcal{D}$. Then $D = \bigcup_{k \in S} C_k$ for some nonempty $S \subseteq [K]$. If $|S| \geq 2$ then for any $k_0 \in S$ we have $C_{k_0} \subseteq D$. By (S.29), the inclusion is strict, hence D is not inclusion-minimal. Thus the inclusion-minimal nonempty elements of \mathcal{D} are exactly $\{C_k : k \in [K]\}$. Knowing all C_k recovers \mathbf{Q} up to a permutation of columns.

Proof of (c). From (S.28), for any $\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}$ and any $j \in [J]$,

$$\eta_j(\mathbf{Z}) \neq \eta_j(\mathbf{Z}') \iff \eta'_j(\sigma(\mathbf{Z})) \neq \eta'_j(\sigma(\mathbf{Z}')),$$

hence $D(\mathbf{Z}, \mathbf{Z}') = D'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}'))$. Because σ is a bijection on \mathcal{Z} , the map $(\mathbf{Z}, \mathbf{Z}') \mapsto (\sigma(\mathbf{Z}), \sigma(\mathbf{Z}'))$ is a bijection on $\{(\mathbf{Z}, \mathbf{Z}') : \mathbf{Z} \neq \mathbf{Z}'\}$, and therefore

$$\mathcal{D} = \{D(\mathbf{Z}, \mathbf{Z}') : \mathbf{Z} \neq \mathbf{Z}'\} = \{D'(\tilde{\mathbf{Z}}, \tilde{\mathbf{Z}}') : \tilde{\mathbf{Z}} \neq \tilde{\mathbf{Z}}'\} = \mathcal{D}'$$

as *sets of subsets* of $[J]$.

By part (b) (which uses (S.29) for \mathbf{Q}), the inclusion-minimal nonempty elements of \mathcal{D} are exactly $\{C_k : k \in [K]\}$. Since $\mathcal{D} = \mathcal{D}'$, the inclusion-minimal nonempty elements of \mathcal{D}' are also exactly $\{C_k : k \in [K]\}$.

Fix $k \in [K]$ and choose $\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}$ that differ only at coordinate k . Then $D(\mathbf{Z}, \mathbf{Z}') = C_k$, so $D'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}')) = C_k$. Applying part (a) to (\mathbf{Q}', η') gives

$$D'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}')) = \bigcup_{h \in S'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}'))} C'_h.$$

Every C'_h belongs to \mathcal{D}' (take two states that differ only at coordinate h), so each C'_h appearing in the above union is a nonempty element of \mathcal{D}' and satisfies $C'_h \subseteq \bigcup_{h \in S'} C'_h = C_k$. But C_k is inclusion-minimal among the nonempty elements of \mathcal{D}' , hence no nonempty element of \mathcal{D}'

can be a *proper* subset of C_k . Therefore every C'_h appearing in the union must equal C_k . In particular, there exists at least one index $\varpi(k) \in [K]$ such that $C'_{\varpi(k)} = C_k$.

If $\varpi(k_1) = \varpi(k_2)$ then $C_{k_1} = C'_{\varpi(k_1)} = C'_{\varpi(k_2)} = C_{k_2}$. Under (S.29), the sets $\{C_k\}_{k=1}^K$ are pairwise distinct, so $k_1 = k_2$. Thus ϖ is injective and hence a permutation of $[K]$. Consequently $C_k = C'_{\varpi(k)}$ for all $k \in [K]$, equivalently $\mathbf{Q}' = \mathbf{Q}\Pi$ for the permutation matrix Π .

Proof of (d).

Define the Hamming graph on \mathcal{Z} whose vertex set is \mathcal{Z} and where two vertices are adjacent if they differ in exactly one coordinate. For each $k \in [K]$, call an edge $\{\mathbf{Z}, \mathbf{Z}'\}$ a k -edge if \mathbf{Z}, \mathbf{Z}' differ only in coordinate k .

By (c), \mathbf{Q}' is a column permutation of \mathbf{Q} , hence (S.29) also holds for \mathbf{Q}' . Now fix a k -edge $\{\mathbf{Z}, \mathbf{Z}'\}$, i.e. \mathbf{Z}, \mathbf{Z}' differ only at coordinate k . Then $D(\mathbf{Z}, \mathbf{Z}') = C_k$, so $D'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}')) = C_k = C'_{\varpi(k)}$. If $\sigma(\mathbf{Z})$ and $\sigma(\mathbf{Z}')$ differed in at least two coordinates, then by part (a) for (\mathbf{Q}', η') the set $D'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}'))$ would be a union of at least two distinct sets among $\{C'_h\}$. Under (S.29) for \mathbf{Q}' , such a union strictly contains each constituent, hence cannot equal $C'_{\varpi(k)}$. Therefore $\sigma(\mathbf{Z})$ and $\sigma(\mathbf{Z}')$ differ in exactly one coordinate. Let that coordinate be h . Then by part (a) for (\mathbf{Q}', η') applied to the pair $\sigma(\mathbf{Z}), \sigma(\mathbf{Z}')$ we have $D'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}')) = C'_h$. Comparing with $D'(\sigma(\mathbf{Z}), \sigma(\mathbf{Z}')) = C'_{\varpi(k)}$ yields $C'_h = C'_{\varpi(k)}$, and since (S.29) implies the supports $\{C'_1, \dots, C'_K\}$ are pairwise distinct, we conclude $h = \varpi(k)$. Consequently, σ maps every k -edge to a $\varpi(k)$ -edge.

Fix $k \in [K]$ and fix a context $\mathbf{z}_{-k} \in \prod_{h \neq k} [M_h]$. For $t \in [M_k]$, write $\mathbf{z}(\mathbf{z}_{-k}, t)$ for the latent state whose $-k$ coordinates equal \mathbf{z}_{-k} and whose k th coordinate equals t . Then any two distinct vertices in the fiber

$$\mathcal{F}_k(\mathbf{z}_{-k}) := \{\mathbf{z}(\mathbf{z}_{-k}, t) : t \in [M_k]\}$$

differ in exactly one coordinate (namely k), hence are joined by a k -edge. Since σ maps

k -edges to $\varpi(k)$ -edges, it follows that for any $t \neq t'$, the images $\sigma(\mathbf{z}(\mathbf{z}_{-k}, t))$ and $\sigma(\mathbf{z}(\mathbf{z}_{-k}, t'))$ differ in exactly one coordinate (namely $\varpi(k)$). In particular, all vectors $\{\sigma(\mathbf{z}(\mathbf{z}_{-k}, t)) : t \in [M_k]\}$ agree on coordinates outside $\varpi(k)$.

Therefore there exists a map $\tau_{k, \mathbf{z}_{-k}} : [M_k] \rightarrow [M_{\varpi(k)}]$ such that

$$(\sigma(\mathbf{z}(\mathbf{z}_{-k}, t)))_{\varpi(k)} = \tau_{k, \mathbf{z}_{-k}}(t) \quad \text{for all } t \in [M_k]. \quad (\text{S.31})$$

Because σ is injective, the points $\sigma(\mathbf{z}(\mathbf{z}_{-k}, t))$ are all distinct as t varies, hence $\tau_{k, \mathbf{z}_{-k}}$ is injective. Thus $M_{\varpi(k)} \geq M_k$. Applying the same argument to σ^{-1} (which maps $\varpi(k)$ -edges back to k -edges) yields $M_k \geq M_{\varpi(k)}$. Consequently $M_{\varpi(k)} = M_k$ and $\tau_{k, \mathbf{z}_{-k}} \in S_{M_k}$ is a permutation.

Next we show that $\tau_{k, \mathbf{z}_{-k}}$ does not depend on the context \mathbf{z}_{-k} . Fix $t \in [M_k]$ and take two contexts $\mathbf{z}_{-k}, \mathbf{b}_{-k}$. In the Hamming graph on \mathcal{Z} , there is a path from $\mathbf{z}(\mathbf{z}_{-k}, t)$ to $\mathbf{z}(\mathbf{b}_{-k}, t)$ that changes only coordinates in $[K] \setminus \{k\}$. Along this path, each step is an h -edge for some $h \neq k$, hence its image under σ is a $\varpi(h)$ -edge. Since ϖ is a permutation, $\varpi(h) \neq \varpi(k)$ whenever $h \neq k$, so the $\varpi(k)$ -coordinate remains constant along the image path. Therefore $(\sigma(\mathbf{z}(\mathbf{z}_{-k}, t)))_{\varpi(k)} = (\sigma(\mathbf{z}(\mathbf{b}_{-k}, t)))_{\varpi(k)}$. By (S.31), this implies $\tau_{k, \mathbf{z}_{-k}}(t) = \tau_{k, \mathbf{b}_{-k}}(t)$ for every t , hence $\tau_{k, \mathbf{z}_{-k}}$ is the same permutation for all contexts. Denote this common permutation by $\tau_k \in S_{M_k}$.

We have shown that for every $\mathbf{Z} = (Z_1, \dots, Z_K) \in \mathcal{Z}$ and every $k \in [K]$, $(\sigma(\mathbf{Z}))_{\varpi(k)} = \tau_k(Z_k)$. Equivalently, writing $m = \varpi(k)$ and $k = \varpi^{-1}(m)$, $(\sigma(\mathbf{Z}))_m = \tau_{\varpi^{-1}(m)}(Z_{\varpi^{-1}(m)})$. Therefore, for all $(Z_1, \dots, Z_K) \in \mathcal{Z}$,

$$\sigma(Z_1, \dots, Z_K) = (\tau_{\varpi^{-1}(1)}(Z_{\varpi^{-1}(1)}), \dots, \tau_{\varpi^{-1}(K)}(Z_{\varpi^{-1}(K)})).$$

Finally, since $M_{\varpi(k)} = M_k$ for all k , we have $\varpi \in S_{K, \mathbf{M}}$. This proves that σ lies in $\left(\prod_{k=1}^K S_{M_k}\right) \rtimes S_{K, \mathbf{M}}$. \square

We now explain explicitly why Lemma S.7 implies Corollary 1. Fix any $\xi \in A(\mathcal{F}, \mathcal{P})$ and any $f \in \mathcal{F}$. By definition of the indeterminacy set, the transformed pair $(f \circ \xi^{-1}, \xi_{\#}p)$ also belongs to $(\mathcal{F}, \mathcal{P})$. Write

$$f' := f \circ \xi^{-1}.$$

Since $f \in \mathcal{F}$, there exists a binary matrix \mathbf{Q} such that $(\mathbf{Q}, \{f_j\}_{j=1}^J)$ satisfies the structural restriction

$$f_j(\mathbf{z}) = f_j(\mathbf{z}') \quad \text{whenever } \mathbf{z}_{K_j} = \mathbf{z}'_{K_j},$$

the generic injectivity condition

$$f_j(\mathbf{z}) \neq f_j(\mathbf{z}') \quad \text{whenever } \mathbf{z}_{K_j} \neq \mathbf{z}'_{K_j},$$

and the subset condition on the columns of \mathbf{Q} . Likewise, since $f' \in \mathcal{F}$, there exists another binary matrix \mathbf{Q}' such that $(\mathbf{Q}', \{f'_j\}_{j=1}^J)$ satisfies the analogous properties. Now set

$$\eta_j := f_j, \quad \eta'_j := f'_j, \quad \sigma := \xi.$$

Then for every $j \in [J]$ and every $\mathbf{z} \in \mathcal{Z}$,

$$\eta_j(\mathbf{z}) = f_j(\mathbf{z}) = f'_j(\xi(\mathbf{z})) = \eta'_j(\sigma(\mathbf{z})),$$

so all assumptions of Lemma S.7 are satisfied. Therefore Lemma S.7(d) yields

$$\xi \in \left(\prod_{k=1}^K S_{M_k} \right) \rtimes S_{K, \mathbf{M}},$$

which is exactly the conclusion of Corollary 1.

Note that the corollary is, in one respect, more restrictive in its setup than Lemma S.7. The lemma assumes the subset condition only for the true matrix \mathbf{Q} , whereas Corollary 1

is stated in terms of $\xi \in A(\mathcal{F}, \mathcal{P})$, and hence requires both f and $f \circ \xi^{-1}$ to belong to \mathcal{F} . Consequently, both associated design matrices \mathbf{Q} and \mathbf{Q}' must satisfy the defining constraints of \mathcal{F} , including the subset condition.

Fix an item $j \in [J]$ and recall $K_j = \{k \in [K] : q_{j,k} = 1\}$ and $\text{supp}(\mathbf{u}) := \{k \in [K] : u_k \geq 1\}$. Define the admissible index set

$$\mathcal{Z}_j^{\mathbf{Q}} := \{\mathbf{u} \in \mathcal{Z} : \text{supp}(\mathbf{u}) \subseteq K_j\}, \quad (\text{S.32})$$

the corresponding feature vector

$$\Phi_j^{\mathbf{Q}}(\mathbf{Z}) := \left(\prod_{k=1}^K \mathcal{I}_{k,u_k}(\mathbf{Z}) \right)_{\mathbf{u} \in \mathcal{Z}_j^{\mathbf{Q}}} \in \{0, 1\}^{|\mathcal{Z}_j^{\mathbf{Q}}|}, \quad (\text{S.33})$$

and the free coefficient subvector

$$\boldsymbol{\beta}_j^{\mathbf{Q}} := (\beta_{j,\mathbf{u}})_{\mathbf{u} \in \mathcal{Z}_j^{\mathbf{Q}}} \in \mathbb{R}^{|\mathcal{Z}_j^{\mathbf{Q}}|}. \quad (\text{S.34})$$

Then we have the reduced representation

$$\eta_j(\mathbf{Z}) = \langle \boldsymbol{\beta}_j^{\mathbf{Q}}, \Phi_j^{\mathbf{Q}}(\mathbf{Z}) \rangle. \quad (\text{S.35})$$

Lemma S.8. *Fix item j . If $\mathbf{Z}_{K_j} = \mathbf{Z}'_{K_j}$ then $\eta_j(\mathbf{Z}) = \eta_j(\mathbf{Z}')$.*

Proof. Take any $\mathbf{u} \in \mathcal{Z}_j^{\mathbf{Q}}$. If $k \notin K_j$, then $\text{supp}(\mathbf{u}) \subseteq K_j$ forces $u_k = 0$, hence $\mathcal{I}_{k,u_k}(\cdot) = \mathcal{I}_{k,0}(\cdot) \equiv 1$. Therefore the basis product $\prod_{k=1}^K \mathcal{I}_{k,u_k}(\cdot)$ depends only on \mathbf{Z}_{K_j} . If $\mathbf{Z}_{K_j} = \mathbf{Z}'_{K_j}$ then every coordinate of $\Phi_j^{\mathbf{Q}}(\mathbf{Z})$ equals the corresponding coordinate of $\Phi_j^{\mathbf{Q}}(\mathbf{Z}')$, and (S.35) yields $\eta_j(\mathbf{Z}) = \eta_j(\mathbf{Z}')$. \square

Lemma S.9. *For any $\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}$,*

$$\mathbf{Z}_{K_j} \neq \mathbf{Z}'_{K_j} \implies \Phi_j^{\mathbf{Q}}(\mathbf{Z}) \neq \Phi_j^{\mathbf{Q}}(\mathbf{Z}').$$

Consequently, outside a Lebesgue-null set in the free coordinates $\beta_j^{\mathbf{Q}}$ (equivalently, in the non-structural-zero coordinates of β_j),

$$\mathbf{Z}_{K_j} \neq \mathbf{Z}'_{K_j} \implies \eta_j(\mathbf{Z}) \neq \eta_j(\mathbf{Z}').$$

Proof. Assume $\mathbf{Z}_{K_j} \neq \mathbf{Z}'_{K_j}$ and pick $k \in K_j$ such that $Z_k \neq Z'_k$. Without loss of generality $Z_k < Z'_k$. Set $v := Z'_k \in [M_k - 1]$. Then $\mathcal{I}_{k,v}(\mathbf{Z}) = 0$ and $\mathcal{I}_{k,v}(\mathbf{Z}') = 1$. Since $k \in K_j$, the coordinate $\mathbf{u}(k, v)$ belongs to $\mathcal{Z}_j^{\mathbf{Q}}$, and its corresponding feature in (S.33) is exactly $\prod_{h=1}^K \mathcal{I}_{h, \mathbf{u}(k,v)_h}(\cdot) = \mathcal{I}_{k,v}(\cdot)$, because $\mathbf{u}(k, v)_h = 0$ for all $h \neq k$ and $\mathcal{I}_{h,0} \equiv 1$. Hence the $\mathbf{u}(k, v)$ -coordinate of $\Phi_j^{\mathbf{Q}}(\mathbf{Z})$ differs from that of $\Phi_j^{\mathbf{Q}}(\mathbf{Z}')$, proving $\Phi_j^{\mathbf{Q}}(\mathbf{Z}) \neq \Phi_j^{\mathbf{Q}}(\mathbf{Z}')$.

For the generic injectivity, fix any distinct pair $(\mathbf{Z}, \mathbf{Z}')$ with $\mathbf{Z}_{K_j} \neq \mathbf{Z}'_{K_j}$. By the first part, $\Phi_j^{\mathbf{Q}}(\mathbf{Z}) - \Phi_j^{\mathbf{Q}}(\mathbf{Z}') \neq 0$, so the equality $\eta_j(\mathbf{Z}) = \eta_j(\mathbf{Z}')$ is the proper affine hyperplane

$$\{\beta_j^{\mathbf{Q}} : \langle \beta_j^{\mathbf{Q}}, \Phi_j^{\mathbf{Q}}(\mathbf{Z}) - \Phi_j^{\mathbf{Q}}(\mathbf{Z}') \rangle = 0\}$$

in $\mathbb{R}^{|\mathcal{Z}_j^{\mathbf{Q}}|}$ (using (S.35)). Since $|\mathcal{Z}| < \infty$, the union of these hyperplanes over all such pairs is Lebesgue-null. Intersecting with the constraint set (which only removes finitely many coordinate hyperplanes $\{\beta_{j, \mathbf{u}(k,v)} = 0\}$ and thus does not change nullness) yields the claim. \square

From Step 1 there exists a stabilized bijection $\sigma : \mathcal{Z} \rightarrow \mathcal{Z}$ such that

$$\eta_j(\mathbf{z}) = \eta'_j(\sigma(\mathbf{z})) \quad \forall j \in [J], \forall \mathbf{z} \in \mathcal{Z}. \quad (\text{S.36})$$

Apply Lemma S.8 and Lemma S.9 to every item j and intersect the corresponding generic sets over $j \in [J]$. On this intersection, the pair $(\mathbf{Q}, \{\eta_j\}_{j=1}^J)$ satisfies the hypotheses of Lemma S.7 with $K_j = \{k : q_{j,k} = 1\}$. The same holds for $(\mathbf{Q}', \{\eta'_j\}_{j=1}^J)$.

Applying Lemma S.7 to $(\mathbf{Q}, \{\eta_j\})$ and $(\mathbf{Q}', \{\eta'_j\})$ yields an permutation $\varpi \in S_{K, \mathbf{M}}$ such

that $\mathbf{Q}' = \mathbf{Q}\Pi$, where Π is the permutation matrix of ϖ^{-1} , and the stabilized relabeling σ must lie in $(\prod_{k=1}^K S_{M_k}) \times S_{K, \mathbf{M}}$. Equivalently, there exist permutations $\tau_k \in \mathfrak{S}_{M_k}$ such that

$$\sigma(z_1, \dots, z_K) = (\tau_1(z_{\varpi^{-1}(1)}), \dots, \tau_K(z_{\varpi^{-1}(K)})). \quad (\text{S.37})$$

At this stage we have identified the coordinate permutation ϖ . In the next step we remove the within-coordinate relabelings $\{\tau_k\}$, thereby concluding $\tau_k = \text{Id}$ and hence $\sigma = \sigma_\varpi$.

S.5.3 Eliminate within-coordinate relabelings

Recall that σ_ϖ for the coordinate permutation on \mathcal{Z} is defined by $\sigma_\varpi(z_1, \dots, z_K) = (z_{\varpi^{-1}(1)}, \dots, z_{\varpi^{-1}(K)})$, and write τ for the within-coordinate map $\tau(z_1, \dots, z_K) := (\tau_1(z_1), \dots, \tau_K(z_K))$, we have the factorization $\sigma = \tau \circ \sigma_\varpi$.

Define the relabeled alternative predictors

$$\tilde{\eta}'_j(\mathbf{z}) := \eta'_j(\sigma_\varpi(\mathbf{z})), \quad (j \in [J], \mathbf{z} \in \mathcal{Z}),$$

and define the relabeled stabilized permutation $\tilde{\sigma} := \sigma_\varpi^{-1} \circ \sigma$. Then $\sigma = \tau \circ \sigma_\varpi$ implies $\tilde{\sigma}$ has within-coordinate form

$$\tilde{\sigma}(z_1, \dots, z_K) = (\tilde{\tau}_1(z_1), \dots, \tilde{\tau}_K(z_K)) \quad \text{for some } \tilde{\tau}_k \in S_{M_k},$$

and the Step 2 alignment $\eta_j(\mathbf{z}) = \eta'_j(\sigma(\mathbf{z}))$ becomes

$$\eta_j(\mathbf{z}) = \tilde{\eta}'_j(\tilde{\sigma}(\mathbf{z})) \quad \forall j \in [J], \forall \mathbf{z} \in \mathcal{Z}. \quad (\text{S.38})$$

Similarly, relabel the alternative measurement matrix by permuting coordinate indices according to ϖ . Let Π denote the permutation matrix of ϖ^{-1} so that Step 2 yields $\mathbf{Q}' = \mathbf{Q}\Pi$. Define $\tilde{\mathbf{Q}}' := \mathbf{Q}'\Pi^\top$. Then $\tilde{\mathbf{Q}}' = \mathbf{Q}$ and hence $K_j^{\tilde{\mathbf{Q}}'} = K_j$ for all j . Moreover, Assumption 3

is invariant under this coordinate relabeling.

Therefore, to prove that within-coordinate relabelings are trivial, it suffices to work with the relabeled alternative model $(\tilde{\mathbf{Q}}', \{\tilde{\eta}'_j\})$ and the relabeled stabilized permutation $\tilde{\sigma}$. For notational simplicity, we may assume

$$\varpi = \text{Id}, \quad \mathbf{Q}' = \mathbf{Q}, \quad \sigma(z_1, \dots, z_K) = (\tau_1(z_1), \dots, \tau_K(z_K)),$$

with $\tau_k \in S_{M_k}$.

Thus we are reduced to the within-coordinate form

$$\sigma(z_1, \dots, z_K) = (\tau_1(z_1), \dots, \tau_K(z_K)), \quad \tau_k \in S_{M_k}, \quad (\text{S.39})$$

and we show $\tau_k = \text{Id}$ for every k .

Lemma S.10. *Assume Assumption 3 holds for both $(\mathbf{Q}, \{\eta_j\})$ and $(\mathbf{Q}', \{\eta'_j\})$. Assume also that $\mathbf{Q}' = \mathbf{Q}$ and that σ has the within-coordinate form (S.39).*

Fix $k \in [K]$ and let $j_k \in \mathcal{J}_3$ be an anchor item in the true design (Condition(i) in Theorem 2) satisfying $q_{j_k, k} = 1$. Suppose the alignment for this item holds

$$\eta_{j_k}(\mathbf{z}) = \eta'_{j_k}(\sigma(\mathbf{z})) \quad \forall \mathbf{z} \in \mathcal{Z}. \quad (\text{S.40})$$

Then $\tau_k = \text{Id}$.

Proof. Write $j := j_k$ and $R := K_j$. Since $\mathbf{Q}' = \mathbf{Q}$, we also have $K'_j = K_j = R$.

Recall the top-context block

$$\mathcal{M}_j := \{\mathbf{z} \in \mathcal{Z} : z_h = M_h - 1 \text{ for all } h \in R\}, \quad m := |\mathcal{M}_j| = \prod_{h \notin R} M_h.$$

Define \mathcal{M}'_j analogously for the alternative model. Since $K'_j = K_j$, we have $\mathcal{M}'_j = \mathcal{M}_j$.

By Assumption 3 in the true model, every $\mathbf{z} \in \mathcal{M}_j$ has $\eta_j(\mathbf{z})$ strictly larger than every $\mathbf{z} \notin \mathcal{M}_j$. Equivalently, \mathcal{M}_j is the unique subset $S \subseteq \mathcal{Z}$ with $|S| = m$ such that

$$\max_{\mathbf{z} \notin S} \eta_j(\mathbf{z}) < \min_{\mathbf{z} \in S} \eta_j(\mathbf{z}).$$

The same uniqueness statement holds for η'_j with \mathcal{M}'_j by applying Assumption 3 to the alternative model.

Now use the alignment (S.40): for any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$,

$$\eta_j(\mathbf{z}) > \eta_j(\mathbf{z}') \iff \eta'_j(\sigma(\mathbf{z})) > \eta'_j(\sigma(\mathbf{z}')).$$

Hence σ maps the set of states attaining the top m values of η_j onto the set of states attaining the top m values of η'_j . By uniqueness of the corresponding size- m separated block, this implies $\sigma(\mathcal{M}_j) = \mathcal{M}'_j = \mathcal{M}_j$.

Since σ has the coordinatewise form (S.39), for any $h \in R$ and any $\mathbf{z} \in \mathcal{M}_j$, $(\sigma(\mathbf{z}))_h = \tau_h(M_h - 1)$. But $\sigma(\mathbf{z}) \in \mathcal{M}_j$ forces $(\sigma(\mathbf{z}))_h = M_h - 1$ for all $h \in R$. Therefore

$$\tau_h(M_h - 1) = M_h - 1 \quad \forall h \in R. \tag{S.41}$$

Next, for $t \in \{0, 1, \dots, M_k - 1\}$, define a top-context state $\mathbf{z}^{(t)} \in \mathcal{Z}$ by

$$z_k = t, \quad z_h = M_h - 1 \quad (h \in R \setminus \{k\}), \quad \text{and arbitrary outside } R.$$

By Lemma S.8, the value of $\eta_j(\mathbf{z}^{(t)})$ depends only on $\mathbf{z}_R^{(t)}$, so the choice outside R is irrelevant. Define the one-dimensional profile $g(t) = \eta_j(\mathbf{z}^{(t)})$ for $t \in \{0, 1, \dots, M_k - 1\}$. We claim that g is strictly increasing. Fix any $u \in [M_k - 1]$. Since $k \in R = K_j$ (because $q_{j,k} = 1$),

Assumption 3 gives

$$\max_{\mathbf{z} \in \mathcal{T}_{0,j}^{(k,u)}} \eta_j(\mathbf{Z}) < \min_{\mathbf{z} \in \mathcal{T}_{1,j}^{(k,u)}} \eta_j(\mathbf{Z}).$$

Taking $\mathbf{Z} = \mathbf{z}^{(t)}$ with $t \leq u - 1$ in $\mathcal{T}_{0,j}^{(k,u)}$ and with $t \geq u$ in $\mathcal{T}_{1,j}^{(k,u)}$ yields $g(u - 1) < g(u)$, hence $g(0) < g(1) < \dots < g(M_k - 1)$.

The same ladder also holds in the alternative model along coordinate k . Define $\mathbf{z}'^{(t)}$ in the alternative model analogously by setting $z_k = t$, $z_h = M_h - 1$ for $h \in R \setminus \{k\}$, and arbitrary outside R , and set $g'(t) := \eta'_j(\mathbf{z}'^{(t)})$. Since $K'_j = K_j = R$, Assumption 3 applied to the alternative model yields, for each $u \in [M_k - 1]$, $g'(u - 1) < g'(u)$ and therefore $g'(0) < g'(1) < \dots < g'(M_k - 1)$.

By (S.41), for every $h \in R \setminus \{k\}$ we have $\tau_h(M_h - 1) = M_h - 1$, hence $\sigma(\mathbf{z}^{(t)}) = \mathbf{z}'^{(\tau_k(t))}$ for $t \in \{0, 1, \dots, M_k - 1\}$. Using (S.40), we obtain

$$g(t) = \eta_j(\mathbf{z}^{(t)}) = \eta'_j(\sigma(\mathbf{z}^{(t)})) = \eta'_j(\mathbf{z}'^{(\tau_k(t))}) = g'(\tau_k(t)).$$

If $s < t$, then we have $g(s) < g(t)$, so $g'(\tau_k(s)) < g'(\tau_k(t))$. Since g' is strictly increasing, it follows that $\tau_k(s) < \tau_k(t)$ for all $s < t$. Thus τ_k is a strictly increasing permutation of $\{0, 1, \dots, M_k - 1\}$, hence $\tau_k = \text{Id}$. \square

Applying Lemma S.10 to each $k \in [K]$ yields $\tau_k = \text{Id}$ for all k , so the stabilized relabeling reduces to $\sigma = \sigma_\varpi$.

S.5.4 Recover $\{\beta_j\}$ after aligning by σ_ϖ .

For each item j , define the r -vector of identified linear predictors

$$\eta_j := (\eta_j(\mathbf{z}))_{\mathbf{z} \in \mathcal{Z}} \in \mathbb{R}^r, \quad \boldsymbol{\eta}'_j := (\eta'_j(\mathbf{z}))_{\mathbf{z} \in \mathcal{Z}} \in \mathbb{R}^r.$$

After aligning by $\sigma := \sigma_\varpi$, we have

$$\boldsymbol{\eta}_j = Q_\sigma \boldsymbol{\eta}'_j, \quad (\text{S.42})$$

where Q_σ is the $r \times r$ permutation matrix acting on the state index $\mathbf{z} \in \mathcal{Z}$.

List $\mathbf{Z} \in \mathcal{Z}$ in lexicographic order and set $F = (\Phi(\mathbf{Z})^\top)_{\mathbf{z} \in \mathcal{Z}} \in \mathbb{R}^{r \times r}$. With the same one-coordinate ordering, let F_k be the $M_k \times M_k$ matrix

$$F_k(t, s) = \begin{cases} 1, & s = 0, \\ \mathbf{1}\{t \geq s\}, & s \in \{1, \dots, M_k - 1\}, \end{cases} \quad t \in \{0, \dots, M_k - 1\}.$$

Then F_k is lower triangular with ones on the diagonal, hence invertible, and the lexicographic construction gives $F = F_1 \otimes F_2 \otimes \dots \otimes F_K$, so F is invertible.

Therefore, the saturated coefficients satisfy

$$\boldsymbol{\eta}_j = F \boldsymbol{\beta}_j, \quad \boldsymbol{\eta}'_j = F \boldsymbol{\beta}'_j, \quad \text{equivalently} \quad \boldsymbol{\beta}_j = F^{-1} \boldsymbol{\eta}_j, \quad \boldsymbol{\beta}'_j = F^{-1} \boldsymbol{\eta}'_j. \quad (\text{S.43})$$

Moreover, the coordinate relabeling $\sigma = \sigma_\varpi$ acts on the design vectors by $\Phi(\sigma(\mathbf{z})) = P_\varpi \Phi(\mathbf{z})$, which implies the matrix identity

$$Q_\sigma F = F P_\varpi^\top. \quad (\text{S.44})$$

Indeed, for each $\mathbf{z} \in \mathcal{Z}$,

$$(Q_\sigma F)_{\mathbf{z},:} = F_{\sigma^{-1}(\mathbf{z}),:} = \Phi(\sigma^{-1}(\mathbf{z}))^\top = (P_\varpi^{-1} \Phi(\mathbf{z}))^\top = \Phi(\mathbf{z})^\top P_\varpi^\top = (F P_\varpi^\top)_{\mathbf{z},:}.$$

Combining (S.42)–(S.44) yields

$$\boldsymbol{\beta}_j = F^{-1} \boldsymbol{\eta}_j = F^{-1} Q_\sigma \boldsymbol{\eta}'_j = F^{-1} Q_\sigma F \boldsymbol{\beta}'_j = F^{-1} F P_\varpi^\top \boldsymbol{\beta}'_j = P_\varpi^\top \boldsymbol{\beta}'_j,$$

hence

$$\beta'_j = P_\varpi \beta_j, \quad j \in [J], \quad (\text{S.45})$$

which completes the proof.

S.6 Proof of Theorem 5

Denote the finite state space by \mathcal{Z} . By Assumption 1(a) there exists $\varepsilon \in (0, \frac{1}{2} \min_{\mathbf{z}} \mathbf{p}_{\mathbf{z}}^*)$. Define the high-probability event $\mathcal{E}_N := \{\|\mathbf{p}_N - \mathbf{p}^*\|_\infty < \varepsilon\}$, for which $\mathbb{P}(\mathcal{E}_N) \rightarrow 1$ since $\|\mathbf{p}_N - \mathbf{p}^*\| = O_p(\frac{1}{c_N})$ while $c_N \rightarrow \infty$. Since $\mathbb{P}(\mathcal{E}_N) \rightarrow 1$, it suffices to establish all subsequent asymptotic statements on \mathcal{E}_N .

On \mathcal{E}_N we have $\min_{\mathbf{z}} (\mathbf{p}_N)_{\mathbf{z}} \geq \varepsilon$. Fix any baseline \mathbf{z}_0 and enumerate the remaining $d = |\mathcal{Z}| - 1$ states as $\mathbf{z}_1, \dots, \mathbf{z}_d$. Define $\mathcal{T}_{\mathbf{z}}(\mathbf{p}) := \log\left(\frac{\mathbf{p}_{\mathbf{z}}}{\mathbf{p}_{\mathbf{z}_0}}\right)$ for $\mathbf{z}_1, \dots, \mathbf{z}_d$. Set $\boldsymbol{\theta}^* := \mathcal{T}(\mathbf{p}^*)$ and $\boldsymbol{\theta}_N := \mathcal{T}(\mathbf{p}_N)$, where \mathcal{T} is a C^∞ diffeomorphism from Δ_d° onto \mathbb{R}^d . The whole line segment $[\mathbf{p}^*, \mathbf{p}_N] := \{\mathbf{p}^* + t(\mathbf{p}_N - \mathbf{p}^*) : t \in [0, 1]\}$ is contained in the convex, compact set $K_\varepsilon := \{\mathbf{p} : \mathbf{p}_{\mathbf{z}} \geq \varepsilon \text{ for all } \mathbf{z} \text{ and } \sum_{\mathbf{z}} \mathbf{p}_{\mathbf{z}} = 1\}$, which lies strictly inside the positive simplex.

The map \mathcal{T} is a composition of an affine map and the coordinatewise logarithm on the open set $\{\mathbf{p}_{\mathbf{z}} > 0, \sum_{\mathbf{z}} \mathbf{p}_{\mathbf{z}} = 1\}$, so \mathcal{T} is continuously differentiable there and the Jacobian $\nabla \mathcal{T}(\mathbf{p})$ is a continuous function of \mathbf{p} . Since K_ε is compact, we have

$$L_\varepsilon := \sup_{\mathbf{p} \in K_\varepsilon} \|\nabla \mathcal{T}(\mathbf{p})\|_{\text{op}} < \infty.$$

Therefore \mathcal{T} is L_ε -Lipschitz on K_ε . Applying the integral form of Taylor's theorem along the segment $[\mathbf{p}^*, \mathbf{p}_N]$, we obtain

$$\|\boldsymbol{\theta}_N - \boldsymbol{\theta}^*\|_2 \leq L_\varepsilon \|\mathbf{p}_N - \mathbf{p}^*\|_2 = O_p\left(\frac{1}{c_N}\right).$$

On the event \mathcal{E}_N , every coordinate of \mathbf{p}^* and \mathbf{p}_N is strictly positive, so both $\boldsymbol{\theta}^* = \mathcal{T}(\mathbf{p}^*)$ and $\boldsymbol{\theta}_N = \mathcal{T}(\mathbf{p}_N)$ lie in the natural parameter space associated with the saturated multino-

mial family on the finite state space \mathcal{Z} . For $\mathbf{z} \in \mathcal{Z}$ define the sufficient statistic

$$\phi_j(\mathbf{z}) := \mathbf{1}\{\mathbf{z} = \mathbf{z}_j\}, \quad j = 1, \dots, d, \quad \phi(\mathbf{z}) := (\phi_1(\mathbf{z}), \dots, \phi_d(\mathbf{z}))^\top \in \mathbb{R}^d.$$

With respect to the counting measure on \mathcal{Z} , the saturated multinomial family can be written in canonical exponential-family form as

$$p_\theta(\mathbf{z}) = h(\mathbf{z}) \exp\{\langle \theta, \phi(\mathbf{z}) \rangle - A(\theta)\}, \quad \mathbf{z} \in \mathcal{Z},$$

where we take $h(\mathbf{z}) \equiv 1$ and $\theta = (\theta_1, \dots, \theta_d)^\top \in \Xi = \mathbb{R}^d$, and the log-partition function is

$$A(\theta) := \log\left(1 + \sum_{j=1}^d e^{\theta_j}\right).$$

This saturated multinomial family is regular and minimal. The mean map $\mu(\theta) := \nabla_\theta A(\theta)$ has components $\mu_j(\theta) = \mathbb{P}_\theta(\mathbf{Z} = \mathbf{z}_j)$, and the Fisher information $I(\theta) := \nabla_\theta^2 A(\theta)$ is continuous on $\Xi = \mathbb{R}^d$. In particular, at the true parameter $\theta^* = \mathcal{T}(\mathbf{p}^*)$ we have $I(\theta^*) \succ 0$.

By continuity of $I(\theta)$, we can choose a convex open neighborhood (for example, a small open ball) $U_\theta \subset \Xi$ with $\theta^* \in U_\theta$ such that

$$\lambda_+ \succeq I(\theta) \succeq \lambda_- > 0 \quad (\theta \in U_\theta).$$

Then A is λ_- -strongly convex on U_θ , so ∇A is injective and, by the inverse function theorem, the gradient map $\nabla A : U_\theta \rightarrow U_\mu := \nabla A(U_\theta)$ is a C^∞ diffeomorphism onto its image, and $\mu^* := \mu(\theta^*) \in U_\mu$. We can further pick $s > 0$ such that $B(\mu^*, s) \subseteq U_\mu$. All model comparisons are taken over feasible sets $\Xi_M = \{\theta : \mathcal{T}^{-1}(\theta) \in \mathcal{M}(M)\}$ with $M \in \{G, H\}$. Write $f(\theta; \mu) := \langle \mu, \theta \rangle - A(\theta)$ and $V_M(\mu) := \sup_{\theta \in \Xi_M} f(\theta; \mu)$.

Recall that $\mathbf{R}_{N,1}, \dots, \mathbf{R}_{N,N} \stackrel{\text{i.i.d.}}{\sim} \mathbf{p}_N$. Define $\hat{\mu}_N := N^{-1} \sum_{i=1}^N \phi(\mathbf{R}_{N,i})$. Since $\|\theta_N - \theta^*\| = O_p(\frac{1}{c_N})$ and U_θ is open, there exists $\tau > 0$ such that $\overline{B(\theta^*, \tau)} \subseteq U_\theta$ and $\mathbb{P}(\theta_N \in \overline{B(\theta^*, \tau)}) \rightarrow 1$.

Lemma S.11. V_M is continuous at $\boldsymbol{\mu}^*$.

Proof. For each $\boldsymbol{\theta} \in \Xi_M$, the map $\boldsymbol{\mu} \mapsto \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})$ is affine. As a pointwise supremum of affine functions, V_M is convex. Furthermore,

$$V_M(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Xi_M} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\} \leq \sup_{\boldsymbol{\theta} \in \text{int}(\Xi)} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\} = A^*(\boldsymbol{\mu}).$$

If $\boldsymbol{\mu} \in U_\mu$, then $\sup_{\boldsymbol{\theta} \in \text{int}(\Xi)} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\} = \langle \boldsymbol{\mu}, (\nabla A)^{-1}(\boldsymbol{\mu}) \rangle - A((\nabla A)^{-1}(\boldsymbol{\mu})) < \infty$, thus $A^*(\boldsymbol{\mu}) < \infty$ for all $\boldsymbol{\mu} \in U_\mu$. Thus V_M is finite on U_μ . As V_M is convex and finite on the open convex set $B(\boldsymbol{\mu}^*, s)$, it is continuous throughout $B(\boldsymbol{\mu}^*, s)$, in particular at $\boldsymbol{\mu}^*$. \square

Lemma S.12. $\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\| = O_p(N^{-1/2} + c_N^{-1})$.

Proof. Let $\mathcal{F}_N := \{\boldsymbol{\theta}_N \in \overline{B(\boldsymbol{\theta}^*, \tau)}\}$. Because $I(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^2 A(\boldsymbol{\theta})$ is continuous on $U_{\boldsymbol{\theta}}$ and satisfies $\lambda_- \mathbf{I}_d \preceq I(\boldsymbol{\theta}) \preceq \lambda_+ \mathbf{I}_d$ for all $\boldsymbol{\theta} \in \overline{B(\boldsymbol{\theta}^*, \tau)}$, we have $\sup_{\boldsymbol{\theta} \in \overline{B(\boldsymbol{\theta}^*, \tau)}} \text{tr } I(\boldsymbol{\theta}) \leq d\lambda_+ < \infty$.

For any $t > 0$, write the probability with the intersection as

$$\mathbb{P}(\sqrt{N}\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}(\boldsymbol{\theta}_N)\| > t, \mathcal{F}_N) = \mathbb{E}[\mathbf{1}_{\mathcal{F}_N} \mathbb{P}(\sqrt{N}\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}(\boldsymbol{\theta}_N)\| > t \mid \boldsymbol{\theta}_N)].$$

On $\{\boldsymbol{\theta}_N = \boldsymbol{\theta} \in \mathcal{F}_N\}$ we have $\text{Var}(\widehat{\boldsymbol{\mu}}_N) = I(\boldsymbol{\theta})/N$, hence $\mathbb{E}[N\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}(\boldsymbol{\theta}_N)\|^2 \mid \boldsymbol{\theta}_N] = \text{tr } I(\boldsymbol{\theta}_N)$.

Chebyshev's inequality yields

$$\mathbb{P}(\sqrt{N}\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}(\boldsymbol{\theta}_N)\| > t, \mathcal{F}_N) \leq t^{-2} \mathbb{E}[\mathbf{1}_{\mathcal{F}_N} \text{tr } I(\boldsymbol{\theta}_N)] \leq d\lambda_+/t^2,$$

equivalently $\mathbb{P}(\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}(\boldsymbol{\theta}_N)\| > t/\sqrt{N}, \mathcal{F}_N) \leq d\lambda_+/t^2$ for all $t > 0$.

By continuity of $\nabla^2 A$ on the compact $\overline{B(\boldsymbol{\theta}^*, \tau)}$, there exists $L < \infty$ with $\|\nabla^2 A(\boldsymbol{\theta})\|_{\text{op}} \leq L$ on this set, so ∇A is L -Lipschitz there. Consequently, for any $\eta > 0$ we have

$$\mathbb{P}(\|\boldsymbol{\mu}(\boldsymbol{\theta}_N) - \boldsymbol{\mu}^*\| > \eta, \mathcal{F}_N) \leq \mathbb{P}(L\|\boldsymbol{\theta}_N - \boldsymbol{\theta}^*\| > \eta) = \mathbb{P}(\|\boldsymbol{\theta}_N - \boldsymbol{\theta}^*\| > \eta/L).$$

Taking $\eta = K'/c_N$ gives $\mathbb{P}(\|\boldsymbol{\mu}(\boldsymbol{\theta}_N) - \boldsymbol{\mu}^*\| > K'/c_N, \mathcal{F}_N) = o(1)$ because $\|\boldsymbol{\theta}_N - \boldsymbol{\theta}^*\| = O_p(c_N^{-1})$.

For any $K, K' > 0$,

$$\begin{aligned} & \mathbb{P}(\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\| > \frac{K}{\sqrt{N}} + \frac{K'}{c_N}) \\ & \leq \mathbb{P}(\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}(\boldsymbol{\theta}_N)\| > \frac{K}{\sqrt{N}}, \mathcal{F}_N) + \mathbb{P}(\|\boldsymbol{\mu}(\boldsymbol{\theta}_N) - \boldsymbol{\mu}^*\| > \frac{K'}{c_N}, \mathcal{F}_N) + \mathbb{P}(\mathcal{F}_N^c). \end{aligned}$$

Using the two bounds above together with $\mathbb{P}(\mathcal{F}_N^c) \rightarrow 0$ gives $\limsup_{N \rightarrow \infty} \mathbb{P}(\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\| > K/\sqrt{N} + K'/c_N) \leq d\lambda_+/K^2$. Since the bound holds for all $K, K' > 0$, take $K' = K$ to get, for every $K > 0$, $\limsup_{N \rightarrow \infty} \mathbb{P}(\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\| > K(N^{-1/2} + c_N^{-1})) \leq \frac{d\lambda_+}{K^2}$. Given $\varepsilon > 0$, choose $K_\varepsilon \geq \sqrt{d\lambda_+/\varepsilon}$. Then there exists N_ε such that for all $N \geq N_\varepsilon$,

$$\mathbb{P}(\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\| > K_\varepsilon(N^{-1/2} + c_N^{-1})) \leq \varepsilon,$$

which is exactly $\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\| = O_p(N^{-1/2} + c_N^{-1})$. □

For discrete causal graphical models, by Thm. 18.1 of [Koller and Friedman \(2009\)](#)

$$S_{\text{BDeu}}(\mathcal{D} \mid G) = \ell(\widehat{\boldsymbol{\theta}}; \mathcal{D}) - \frac{1}{2}d_G \log N + O(1).$$

We will also utilize this fact in the subsequent proof.

Case 1: $\boldsymbol{\theta}^* \in \Xi_H \setminus \Xi_G$

Because $I \succ 0$ on $\text{int}(\Xi)$, $f(\cdot; \boldsymbol{\mu}^*)$ is strictly concave on $\text{int}(\Xi)$ and has a unique maximizer $\boldsymbol{\theta}^* = (\nabla A)^{-1}(\boldsymbol{\mu}^*)$ on $\text{int}(\Xi)$. Furthermore, since $\boldsymbol{\theta}^* \in \Xi_H$ but $\boldsymbol{\theta}^* \notin \Xi_G$, we define

$$\epsilon_0 := \sup_{\boldsymbol{\theta} \in \Xi_H} f(\boldsymbol{\theta}; \boldsymbol{\mu}^*) - \sup_{\boldsymbol{\theta} \in \Xi_G} f(\boldsymbol{\theta}; \boldsymbol{\mu}^*) = f(\boldsymbol{\theta}^*; \boldsymbol{\mu}^*) - \sup_{\boldsymbol{\theta} \in \Xi_G} f(\boldsymbol{\theta}; \boldsymbol{\mu}^*). \quad (\text{S.46})$$

Since A is λ_- -strongly convex on U_θ , $f(\boldsymbol{\theta}; \boldsymbol{\mu}^*) = \langle \boldsymbol{\mu}^*, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})$ is λ_- -strongly concave on U_θ .

Note that the independence constraints of G are given by polynomial equalities in the joint probabilities, so the set $\mathcal{M}(G) := \{\boldsymbol{p} : \text{independence constraints of } G \text{ hold}; \sum_x \boldsymbol{p}_x = 1\}$ is algebraic and hence Euclidean closed. Therefore, $\mathcal{M}(G) \cap \Delta_d^\circ$ is closed in Δ_d° under the relative topology. Since \mathcal{T} is a homeomorphism, Ξ_G is closed in $\Xi = \mathbb{R}^d$.

Because Ξ_G is relatively closed in $\text{int}(\Xi)$ and excludes $\boldsymbol{\theta}^*$, we have $\delta = \inf_{\boldsymbol{\theta} \in \Xi_G} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2 > 0$. Pick any $r \in (0, \delta)$ with $\overline{B(\boldsymbol{\theta}^*, r)} \subset U_\theta$. For any $\boldsymbol{\theta} \in \Xi_G$, set $t = \frac{r}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \in (0, 1)$ and $\boldsymbol{\theta}_r = \boldsymbol{\theta}^* + t(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. Then $\|\boldsymbol{\theta}_r - \boldsymbol{\theta}^*\| = r$ and $\boldsymbol{\theta}_r \in \overline{B(\boldsymbol{\theta}^*, r)} \subset U_\theta$. By concavity of $f(\cdot; \boldsymbol{\mu}^*)$ on the convex set $\text{int}(\Xi)$, $f(\boldsymbol{\theta}_r; \boldsymbol{\mu}^*) \geq (1-t)f(\boldsymbol{\theta}^*; \boldsymbol{\mu}^*) + tf(\boldsymbol{\theta}; \boldsymbol{\mu}^*)$, hence $f(\boldsymbol{\theta}; \boldsymbol{\mu}^*) \leq f(\boldsymbol{\theta}_r; \boldsymbol{\mu}^*)$. By λ_- -strong concavity on U_θ ,

$$f(\boldsymbol{\theta}^*; \boldsymbol{\mu}^*) - f(\boldsymbol{\theta}_r; \boldsymbol{\mu}^*) \geq \frac{\lambda_-}{2} \|\boldsymbol{\theta}_r - \boldsymbol{\theta}^*\|^2 = \frac{\lambda_-}{2} r^2.$$

Therefore, we have

$$f(\boldsymbol{\theta}; \boldsymbol{\mu}^*) \leq f(\boldsymbol{\theta}^*; \boldsymbol{\mu}^*) - \frac{\lambda_-}{2} r^2 \quad (\forall \boldsymbol{\theta} \in \Xi_G).$$

It follows that $\epsilon_0 \geq (\lambda_-/2) r^2 > 0$.

By Lemma S.11 and Lemma S.12, with probability $\rightarrow 1$,

$$\sup_{\boldsymbol{\theta} \in \Xi_H} f(\boldsymbol{\theta}; \widehat{\boldsymbol{\mu}}_N) - \sup_{\boldsymbol{\theta} \in \Xi_G} f(\boldsymbol{\theta}; \widehat{\boldsymbol{\mu}}_N) \geq \epsilon_0/2,$$

hence

$$\begin{aligned} S(H, \mathcal{D}) - S(G, \mathcal{D}) &= \frac{1}{2}(d_G - d_{\mathcal{H}}) \log N + N \left[\sup_{\boldsymbol{\theta} \in \Xi_H} f(\boldsymbol{\theta}; \widehat{\boldsymbol{\mu}}_N) - \sup_{\boldsymbol{\theta} \in \Xi_G} f(\boldsymbol{\theta}; \widehat{\boldsymbol{\mu}}_N) \right] + O_p(1) \\ &> \frac{1}{2}(d_G - d_{\mathcal{H}}) \log N + \frac{N\epsilon_0}{2} + O_p(1) \end{aligned}$$

> 0.

We make explicit why the BDeu (BD) score admits a BIC expansion with an $O_p(1)$ remainder under the triangular-array sampling $\mathbf{R}_{N,1}, \dots, \mathbf{R}_{N,N} \stackrel{\text{i.i.d.}}{\sim} p_N$, where p_N is random.

Fix a candidate DAG M on the discrete vector $\mathbf{Z} = (Z_v)_{v \in V}$. For each node $v \in V$, write r_v and q_v for the number of states of Z_v and the number of parent configurations of $\mathbf{Pa}_M(v)$, respectively. Index the parent configurations of $\mathbf{Pa}_M(v)$ by $u \in \{1, \dots, q_v\}$ and the states of Z_v by $k \in \{1, \dots, r_v\}$. Given the dataset $\mathcal{D} := \{\mathbf{R}_{N,1}, \dots, \mathbf{R}_{N,N}\}$ with $\mathbf{R}_{N,i} \in \mathcal{Z}$, define the empirical counts

$$N_{vuk} := \sum_{i=1}^N \mathbf{1}\left\{(\mathbf{R}_{N,i})_{\mathbf{Pa}_M(v)} = u, (\mathbf{R}_{N,i})_v = k\right\}, \quad N_{vu} := \sum_{k=1}^{r_v} N_{vuk} = \sum_{i=1}^N \mathbf{1}\left\{(\mathbf{R}_{N,i})_{\mathbf{Pa}_M(v)} = u\right\}.$$

Write $\theta_{vuk} := \mathbb{P}(Z_v = k \mid \mathbf{Pa}_M(v) = u)$ and $\theta_{vu} := (\theta_{vu1}, \dots, \theta_{vur_v}) \in \Delta_{r_v-1}$, and assume independent Dirichlet priors: for each (v, u) ,

$$\theta_{vu} \sim \text{Dir}(\alpha_{vu1}, \dots, \alpha_{vur_v}), \quad \alpha_{vuk} > 0, \quad \alpha_{vu} := \sum_{k=1}^{r_v} \alpha_{vuk}.$$

For BDeu, one uses the special choice $\alpha_{vuk} \equiv \alpha/(q_v r_v)$, hence $\alpha_{vu} \equiv \alpha/q_v$.

Now we write

$$\log P(\mathcal{D} \mid M) = \sum_{v \in V} \sum_{u=1}^{q_v} \left[\log \Gamma(\alpha_{vu}) - \log \Gamma(\alpha_{vu} + N_{vu}) + \sum_{k=1}^{r_v} \left(\log \Gamma(\alpha_{vuk} + N_{vuk}) - \log \Gamma(\alpha_{vuk}) \right) \right]. \quad (\text{S.47})$$

$$= C_M(\alpha) + \sum_{v \in V} \sum_{u=1}^{q_v} \left[\sum_{k=1}^{r_v} \log \Gamma(\alpha_{vuk} + N_{vuk}) - \log \Gamma(\alpha_{vu} + N_{vu}) \right], \quad (\text{S.48})$$

where $C_M(\alpha) = \sum_{v \in V} \sum_{u=1}^{q_v} [\log \Gamma(\alpha_{vu}) - \sum_{k=1}^{r_v} \log \Gamma(\alpha_{vuk})]$ depends only on M and the hyperparameters.

We now analyze one fixed CPT row (v, u) . To simplify notation in this row, write

$$r := r_v, \quad n_k := N_{vuk}, \quad n := N_{vu} = \sum_{k=1}^r n_k, \quad a_k := \alpha_{vuk}, \quad a := \alpha_{vu} = \sum_{k=1}^r a_k.$$

In what follows we work on the event $\min_{1 \leq k \leq r} n_k \geq 1$, so that $\log n_k$, $\log \hat{\theta}_k$, and $\log(1 + a_k/n_k)$ are well-defined. We will show this event holds with probability $\rightarrow 1$ on \mathcal{E}_N later (see (S.68)).

Define $\hat{\theta}_k := \frac{n_k}{n}$, then the row log-likelihood at the MLE is

$$\ell_{vu}(\hat{\theta}; \mathcal{D}) := \sum_{k=1}^r n_k \log \hat{\theta}_k = \sum_{k=1}^r n_k \log \left(\frac{n_k}{n} \right).$$

We apply Stirling's expansion with an explicit remainder: for all $x \geq 1$,

$$\log \Gamma(x) = \left(x - \frac{1}{2} \right) \log x - x + \frac{1}{2} \log(2\pi) + \eta(x), \quad |\eta(x)| \leq \frac{1}{12x}. \quad (\text{S.49})$$

Apply (S.49) to each $\log \Gamma(a_k + n_k)$ and to $\log \Gamma(a + n)$. We obtain

$$\begin{aligned} \sum_{k=1}^r \log \Gamma(a_k + n_k) - \log \Gamma(a + n) &= \sum_{k=1}^r \left(a_k + n_k - \frac{1}{2} \right) \log(a_k + n_k) - \left(a + n - \frac{1}{2} \right) \log(a + n) \\ &\quad + \frac{r-1}{2} \log(2\pi) + \sum_{k=1}^r \eta(a_k + n_k) - \eta(a + n). \end{aligned} \quad (\text{S.50})$$

Define

$$T_1 := \sum_{k=1}^r \left(a_k + n_k - \frac{1}{2} \right) \log n_k - \left(a + n - \frac{1}{2} \right) \log n. \quad (\text{S.51})$$

$$T_2 := \sum_{k=1}^r \left(a_k + n_k - \frac{1}{2} \right) \log \left(1 + \frac{a_k}{n_k} \right) - \left(a + n - \frac{1}{2} \right) \log \left(1 + \frac{a}{n} \right). \quad (\text{S.52})$$

Then we have:

$$T_1 = \sum_{k=1}^r \left(a_k + n_k - \frac{1}{2} \right) \left(\log n + \log \left(\frac{n_k}{n} \right) \right) - \left(a + n - \frac{1}{2} \right) \log n \quad (\text{S.53})$$

$$= \left(\sum_{k=1}^r \left(a_k + n_k - \frac{1}{2} \right) - \left(a + n - \frac{1}{2} \right) \right) \log n + \sum_{k=1}^r \left(a_k + n_k - \frac{1}{2} \right) \log \left(\frac{n_k}{n} \right). \quad (\text{S.54})$$

$$= -\frac{r-1}{2} \log n + \sum_{k=1}^r n_k \log \left(\frac{n_k}{n} \right) + \sum_{k=1}^r \left(a_k - \frac{1}{2} \right) \log \left(\frac{n_k}{n} \right) \\ = -\frac{r-1}{2} \log n + \ell_{vu}(\hat{\theta}; \mathcal{D}) + \sum_{k=1}^r \left(a_k - \frac{1}{2} \right) \log \hat{\theta}_k. \quad (\text{S.55})$$

Combining (S.50) with (S.55) and (S.52) gives

$$\sum_{k=1}^r \log \Gamma(a_k + n_k) - \log \Gamma(a + n) = \ell_{vu}(\hat{\theta}; \mathcal{D}) - \frac{r-1}{2} \log n + \sum_{k=1}^r \left(a_k - \frac{1}{2} \right) \log \hat{\theta}_k \\ + T_2 + \frac{r-1}{2} \log(2\pi) + \sum_{k=1}^r \eta(a_k + n_k) - \eta(a + n). \quad (\text{S.56})$$

We now return to the original indices. For each row (v, u) , define

$$\hat{\theta}_{vuk} := \frac{N_{vuk}}{N_{vu}}, \quad \ell_{vu}(\hat{\theta}; \mathcal{D}) := \sum_{k=1}^{r_v} N_{vuk} \log \hat{\theta}_{vuk},$$

and define

$$T_{2,vu} := \sum_{k=1}^{r_v} \left(\alpha_{vuk} + N_{vuk} - \frac{1}{2} \right) \log \left(1 + \frac{\alpha_{vuk}}{N_{vuk}} \right) - \left(\alpha_{vu} + N_{vu} - \frac{1}{2} \right) \log \left(1 + \frac{\alpha_{vu}}{N_{vu}} \right), \quad (\text{S.57})$$

$$E_{vu} := \sum_{k=1}^{r_v} \eta(\alpha_{vuk} + N_{vuk}) - \eta(\alpha_{vu} + N_{vu}). \quad (\text{S.58})$$

Then (S.56) becomes, for every (v, u) ,

$$\sum_{k=1}^{r_v} \log \Gamma(\alpha_{vuk} + N_{vuk}) - \log \Gamma(\alpha_{vu} + N_{vu}) = \ell_{vu}(\hat{\theta}; \mathcal{D}) - \frac{r_v-1}{2} \log N_{vu} + \sum_{k=1}^{r_v} \left(\alpha_{vuk} - \frac{1}{2} \right) \log \hat{\theta}_{vuk}$$

$$+ T_{2,vu} + \frac{r_v - 1}{2} \log(2\pi) + E_{vu}. \quad (\text{S.59})$$

Sum (S.59) over all v and u and substitute into (S.48). This yields

$$\log P(\mathcal{D} \mid M) = \ell(\hat{\theta}_M; \mathcal{D}) - \frac{1}{2} \sum_{v \in V} \sum_{u=1}^{q_v} (r_v - 1) \log N_{vu} + R'_{M,N}, \quad (\text{S.60})$$

$$\begin{aligned} \ell(\hat{\theta}_M; \mathcal{D}) &:= \sum_{v \in V} \sum_{u=1}^{q_v} \ell_{vu}(\hat{\theta}; \mathcal{D}) = \sum_{v \in V} \sum_{u=1}^{q_v} \sum_{k=1}^{r_v} N_{vuk} \log \left(\frac{N_{vuk}}{N_{vu}} \right), \\ R'_{M,N} &:= C_M(\alpha) + \frac{1}{2} \sum_{v \in V} \sum_{u=1}^{q_v} (r_v - 1) \log(2\pi) + \sum_{v \in V} \sum_{u=1}^{q_v} \sum_{k=1}^{r_v} \left(\alpha_{vuk} - \frac{1}{2} \right) \log \hat{\theta}_{vuk} \\ &\quad + \sum_{v \in V} \sum_{u=1}^{q_v} T_{2,vu} + \sum_{v \in V} \sum_{u=1}^{q_v} E_{vu}. \end{aligned} \quad (\text{S.61})$$

Define the model dimension $d_M = \sum_{v \in V} \sum_{u=1}^{q_v} (r_v - 1)$.

$$\begin{aligned} -\frac{1}{2} \sum_{v,u} (r_v - 1) \log N_{vu} &= -\frac{1}{2} \sum_{v,u} (r_v - 1) \log N - \frac{1}{2} \sum_{v,u} (r_v - 1) \log \left(\frac{N_{vu}}{N} \right) \\ &= -\frac{1}{2} d_M \log N - \frac{1}{2} \sum_{v,u} (r_v - 1) \log \left(\frac{N_{vu}}{N} \right). \end{aligned} \quad (\text{S.62})$$

Substitute (S.62) into (S.60) to obtain

$$\log P(\mathcal{D} \mid M) = \ell(\hat{\theta}_M; \mathcal{D}) - \frac{1}{2} d_M \log N + R_{M,N}, \quad (\text{S.63})$$

$$R_{M,N} := R'_{M,N} - \frac{1}{2} \sum_{v \in V} \sum_{u=1}^{q_v} (r_v - 1) \log \left(\frac{N_{vu}}{N} \right). \quad (\text{S.64})$$

We now prove that $R_{M,N} = O_p(1)$ under our sampling. Fix (v, u, k) and define the cylinder set $A_{vuk} := \{z \in \mathcal{Z} : z_{\mathbf{p}_{\mathbf{a}_M(v)}} = u, z_v = k\}$. On \mathcal{E}_N we then have

$$p_N(A_{vuk}) = \sum_{z \in A_{vuk}} (p_N)_z \geq \min_{z \in A_{vuk}} (p_N)_z \geq \varepsilon.$$

Conditional on p_N , the count N_{vuk} is binomial:

$$N_{vuk} = \sum_{i=1}^N \mathbf{1}\{\mathbf{R}_{N,i} \in A_{vuk}\} \mid p_N \sim \text{Bin}(N, p_N(A_{vuk})).$$

Using the Chernoff bound,

$$\mathbb{P}\left(N_{vuk} \leq \frac{1}{2}N p_N(A_{vuk}) \mid p_N\right) \leq \exp\left(-\frac{N p_N(A_{vuk})}{8}\right) \leq \exp\left(-\frac{\varepsilon N}{8}\right) \quad \text{on } \mathcal{E}_N. \quad (\text{S.65})$$

Since $p_N(A_{vuk}) \geq \varepsilon$ on \mathcal{E}_N ,

$$\mathbb{P}\left(N_{vuk} \leq \frac{\varepsilon}{2}N \mid p_N\right) \leq \exp\left(-\frac{\varepsilon N}{8}\right) \quad \text{on } \mathcal{E}_N. \quad (\text{S.66})$$

Let $m(M) := \sum_{v \in V} q_v r_v$ be the total number of triples (v, u, k) . Define the event

$$\mathcal{A}_N(M) := \left\{ \min_{v \in V} \min_{1 \leq u \leq q_v} \min_{1 \leq k \leq r_v} N_{vuk} \geq \frac{\varepsilon}{2}N \right\}.$$

By a union bound and (S.66),

$$\mathbb{P}(\mathcal{A}_N(M)^c \mid p_N) \leq \sum_{v,u,k} \mathbb{P}\left(N_{vuk} \leq \frac{\varepsilon}{2}N \mid p_N\right) \leq m(M) \exp\left(-\frac{\varepsilon N}{8}\right) \quad \text{on } \mathcal{E}_N. \quad (\text{S.67})$$

Hence

$$\begin{aligned} \mathbb{P}(\mathcal{A}_N(M)^c) &= \mathbb{P}(\mathcal{A}_N(M)^c \cap \mathcal{E}_N) + \mathbb{P}(\mathcal{E}_N^c) \\ &= \mathbb{E}[\mathbf{1}_{\mathcal{E}_N} \mathbb{P}(\mathcal{A}_N(M)^c \mid p_N)] + \mathbb{P}(\mathcal{E}_N^c) \\ &\leq m(M) \exp\left(-\frac{\varepsilon N}{8}\right) + \mathbb{P}(\mathcal{E}_N^c) \rightarrow 0. \end{aligned}$$

Therefore

$$\mathbb{P}(\mathcal{E}_N \cap \mathcal{A}_N(M)) \rightarrow 1. \quad (\text{S.68})$$

On $\mathcal{A}_N(M)$ we have, for every (v, u) , $N_{vu} = \sum_{k=1}^{r_v} N_{vuk} \geq \frac{\varepsilon}{2}N$, and for every (v, u, k) ,

$$\widehat{\theta}_{vuk} = \frac{N_{vuk}}{N_{vu}} \geq \frac{N_{vuk}}{N} \geq \frac{\varepsilon}{2}.$$

Hence, on $\mathcal{A}_N(M)$,

$$\log\left(\frac{N_{vu}}{N}\right) \in [\log(\varepsilon/2), 0], \quad \log \widehat{\theta}_{vuk} \in [\log(\varepsilon/2), 0]. \quad (\text{S.69})$$

We now bound each term in $R_{M,N}$ on $\mathcal{A}_N(M)$.

First, $C_M(\alpha)$ and $\frac{1}{2} \sum_{v,u} (r_v - 1) \log(2\pi)$ are finite and do not depend on N .

Second, by (S.69),

$$\left| \sum_{v,u,k} \left(\alpha_{vuk} - \frac{1}{2} \right) \log \widehat{\theta}_{vuk} \right| \leq \sum_{v,u,k} \left| \alpha_{vuk} - \frac{1}{2} \right| \cdot \left| \log(\varepsilon/2) \right|,$$

which is a finite constant because the sum is over finitely many (v, u, k) .

Third, we bound $T_{2,vu}$ defined in (S.57). On $\mathcal{A}_N(M)$ we have $N_{vuk} \geq (\varepsilon/2)N$ and $N_{vu} \geq (\varepsilon/2)N$. Therefore,

$$0 \leq \left(\alpha_{vuk} + N_{vuk} - \frac{1}{2} \right) \log \left(1 + \frac{\alpha_{vuk}}{N_{vuk}} \right) \leq \left(\alpha_{vuk} + N_{vuk} \right) \frac{\alpha_{vuk}}{N_{vuk}} \leq \alpha_{vuk} + \frac{2\alpha_{vuk}^2}{\varepsilon N}.$$

Similarly,

$$0 \leq \left(\alpha_{vu} + N_{vu} - \frac{1}{2} \right) \log \left(1 + \frac{\alpha_{vu}}{N_{vu}} \right) \leq \alpha_{vu} + \frac{2\alpha_{vu}^2}{\varepsilon N}.$$

Hence, on $\mathcal{A}_N(M)$,

$$|T_{2,vu}| \leq \sum_{k=1}^{r_v} \left(\alpha_{vuk} + \frac{2\alpha_{vuk}^2}{\varepsilon N} \right) + \left(\alpha_{vu} + \frac{2\alpha_{vu}^2}{\varepsilon N} \right) \leq 2\alpha_{vu} + \frac{2}{\varepsilon N} \left(\sum_{k=1}^{r_v} \alpha_{vuk}^2 + \alpha_{vu}^2 \right). \quad (\text{S.70})$$

Summing (S.70) over finitely many (v, u) shows $\sum_{v,u} T_{2,vu}$ is bounded by a constant plus $O(1/N)$ on $\mathcal{A}_N(M)$.

Fourth, we bound $\sum_{v,u} E_{vu}$ defined in (S.58). On $\mathcal{A}_N(M)$ we have $\alpha_{vuk} + N_{vuk} \geq N_{vuk} \geq (\varepsilon/2)N$ and $\alpha_{vu} + N_{vu} \geq N_{vu} \geq (\varepsilon/2)N$. Since $|\eta(x)| \leq 1/(12x)$ for $x \geq 1$, we obtain, on $\mathcal{A}_N(M)$,

$$|\eta(\alpha_{vuk} + N_{vuk})| \leq \frac{1}{12(\alpha_{vuk} + N_{vuk})} \leq \frac{1}{12N_{vuk}} \leq \frac{1}{12(\varepsilon/2)N} = \frac{1}{6\varepsilon N},$$

and similarly $|\eta(\alpha_{vu} + N_{vu})| \leq 1/(6\varepsilon N)$. Therefore, on $\mathcal{A}_N(M)$,

$$|E_{vu}| \leq \sum_{k=1}^{r_v} \frac{1}{6\varepsilon N} + \frac{1}{6\varepsilon N} = \frac{r_v + 1}{6\varepsilon N},$$

and summing over finitely many (v, u) gives $\sum_{v,u} E_{vu} = O(1/N)$ on $\mathcal{A}_N(M)$.

Finally, the extra term in $R_{M,N}$ is $-\frac{1}{2} \sum_{v,u} (r_v - 1) \log(\frac{N_{vu}}{N})$, which is bounded on $\mathcal{A}_N(M)$ because of (S.69) and finiteness of $\sum_{v,u} (r_v - 1) = d_M$.

Combining the previous bounds with (S.68), this implies $R_{M,N} = O_p(1)$ under the triangular-array sampling $\mathbf{R}_{N,i} \sim p_N$. Since we only compare finitely many models (in particular $\{G, H\}$), the same argument applies to each of them, and hence all BDeu/BIC remainders appearing in the score differences can be taken as $O_p(1)$.

Case 2: $\theta^* \in \Xi_H \cap \Xi_G$ and $d_G < d_H$

Recall $V_M(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Xi_M} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\}$ and $A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \text{int}(\Xi)} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\}$. Because $\theta^* \in \Xi_M$ and $A^*(\boldsymbol{\mu})$ maximizes over a superset $\text{int}(\Xi)$, for any $\boldsymbol{\mu}$ and any $M \in \{G, H\}$ we have

$$0 \leq V_M(\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*; \boldsymbol{\mu}) \leq A^*(\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*; \boldsymbol{\mu}). \quad (\text{S.71})$$

Recall that the gradient map $\nabla A : U_{\boldsymbol{\theta}} \rightarrow U_{\boldsymbol{\mu}}$ is a C^∞ diffeomorphism and $\boldsymbol{\theta}(\cdot) = (\nabla A)^{-1}(\cdot) : U_{\boldsymbol{\mu}} \rightarrow U_{\boldsymbol{\theta}}$ is a C^∞ inverse map. By the Fenchel–Young equality,

$$A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Xi} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\} = \langle \boldsymbol{\mu}, \boldsymbol{\theta}(\boldsymbol{\mu}) \rangle - A(\boldsymbol{\theta}(\boldsymbol{\mu})) \quad (\boldsymbol{\mu} \in U_{\boldsymbol{\mu}}),$$

whence, by the chain rule, $\nabla A^*(\boldsymbol{\mu}) = \boldsymbol{\theta}(\boldsymbol{\mu})$ and $\nabla^2 A^*(\boldsymbol{\mu}) = [\nabla^2 A(\boldsymbol{\theta}(\boldsymbol{\mu}))]^{-1}$. Evaluating at $\boldsymbol{\mu}^*$ gives $\nabla A^*(\boldsymbol{\mu}^*) = \boldsymbol{\theta}^*$ and $\nabla^2 A^*(\boldsymbol{\mu}^*) = I(\boldsymbol{\theta}^*)^{-1}$.

Therefore, by Taylor's theorem on $B(\boldsymbol{\mu}^*, s)$,

$$\begin{aligned} A^*(\boldsymbol{\mu}) &= A^*(\boldsymbol{\mu}^*) + \langle \nabla A^*(\boldsymbol{\mu}^*), \boldsymbol{\mu} - \boldsymbol{\mu}^* \rangle + \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}^*)^\top \nabla^2 A^*(\boldsymbol{\mu}^*) (\boldsymbol{\mu} - \boldsymbol{\mu}^*) + o(\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2) \\ &= (\langle \boldsymbol{\mu}^*, \boldsymbol{\theta}^* \rangle - A(\boldsymbol{\theta}^*)) + \langle \boldsymbol{\theta}^*, \boldsymbol{\mu} - \boldsymbol{\mu}^* \rangle + \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}^*)^\top I(\boldsymbol{\theta}^*)^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}^*) + o(\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2) \end{aligned} \quad (\text{S.72})$$

$$= f(\boldsymbol{\theta}^*; \boldsymbol{\mu}^*) + \langle \boldsymbol{\theta}^*, \boldsymbol{\mu} - \boldsymbol{\mu}^* \rangle + \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}^*)^\top I(\boldsymbol{\theta}^*)^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}^*) + o(\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2). \quad (\text{S.73})$$

On the other hand,

$$f(\boldsymbol{\theta}^*; \boldsymbol{\mu}) = \langle \boldsymbol{\mu}, \boldsymbol{\theta}^* \rangle - A(\boldsymbol{\theta}^*) = f(\boldsymbol{\theta}^*; \boldsymbol{\mu}^*) + \langle \boldsymbol{\theta}^*, \boldsymbol{\mu} - \boldsymbol{\mu}^* \rangle. \quad (\text{S.74})$$

Subtracting (S.74) from (S.73) yields

$$A^*(\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*; \boldsymbol{\mu}) = \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}^*)^\top I(\boldsymbol{\theta}^*)^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}^*) + o(\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2). \quad (\text{S.75})$$

Combining (S.71) and (S.75), we conclude that there exists $0 < \ell < s$ such that for $\boldsymbol{\mu} \in B(\boldsymbol{\mu}^*, \ell)$,

$$0 \leq V_M(\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*; \boldsymbol{\mu}) \leq \frac{1}{\lambda_-} \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2. \quad (\text{S.76})$$

Applying (S.76) to $M = G, H$ and subtracting, for all $\boldsymbol{\mu} \in B(\boldsymbol{\mu}^*, \ell)$ we get

$$|V_H(\boldsymbol{\mu}) - V_G(\boldsymbol{\mu})| \leq |V_H(\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*; \boldsymbol{\mu})| + |V_G(\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*; \boldsymbol{\mu})| \leq \frac{2}{\lambda_-} \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2. \quad (\text{S.77})$$

By Lemma S.12, $\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\| = O_p(N^{-1/2} + c_N^{-1})$, hence

$$V_H(\widehat{\boldsymbol{\mu}}_N) - V_G(\widehat{\boldsymbol{\mu}}_N) = O_p\left(\|\widehat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}^*\|^2\right) = O_p\left((N^{-1/2} + c_N^{-1})^2\right). \quad (\text{S.78})$$

Plugging (S.78) into the score difference, we obtain

$$\begin{aligned} S(H, \mathcal{D}) - S(G, \mathcal{D}) &= \frac{1}{2}(d_{\mathcal{G}} - d_{\mathcal{H}}) \log N + N[V_H(\widehat{\boldsymbol{\mu}}_N) - V_G(\widehat{\boldsymbol{\mu}}_N)] + O_p(1) \\ &= \frac{1}{2}(d_{\mathcal{G}} - d_{\mathcal{H}}) \log N + O_p\left(1 + \sqrt{N} c_N^{-1} + N c_N^{-2}\right) + O_p(1). \end{aligned} \quad (\text{S.79})$$

In particular, if $c_N = \omega(\sqrt{N/\log N})$, then $\sqrt{N} c_N^{-1} = o(\log N)$ and $N c_N^{-2} = o(\log N)$, so

$$S(H, \mathcal{D}) - S(G, \mathcal{D}) = \frac{1}{2}(d_{\mathcal{G}} - d_{\mathcal{H}}) \log N + o_p(\log N) + O_p(1).$$

Consequently, if $d_{\mathcal{G}} < d_{\mathcal{H}}$, then $S(H, \mathcal{D}) - S(G, \mathcal{D}) < 0$ with probability $\rightarrow 1$.

S.7 Proof of Corollary 2

By Theorem 5, it suffices to show that decomposability together with $\{c_N\}$ -consistency implies $\{c_N\}$ -local consistency of BDeu.

Fix a DAG G and an addition $i \rightarrow j$, and write $G' = G + (i \rightarrow j)$. By decomposability,

$$S(G', D_N) - S(G, D_N) = s(R_j, \mathbf{Pa}_j^{G'} \cup \{R_i\}) - s(R_j, \mathbf{Pa}_j^G),$$

so the score change depends only on the local family at X_j .

To analyze this local change, construct a convenient comparison pair (H, H') as follows. Choose a total order τ of the vertices in which every node in \mathbf{Pa}_j^G comes before i , i comes before j , and j comes before all remaining nodes. Let H' be the complete (tournament) DAG consistent with τ (i.e., orient every pair $u \prec_{\tau} v$ as $u \rightarrow v$). Then $\mathbf{Pa}_j^{H'} = \mathbf{Pa}_j^G \cup \{R_i\}$. Define H by deleting the single edge $i \rightarrow j$ from H' . Deleting an edge preserves acyclicity and yields $\mathbf{Pa}_j^H = \mathbf{Pa}_j^G$, and $H' = H + (i \rightarrow j)$. Since H and H' differ only at the family of R_j , decomposability gives

$$S(H', D_N) - S(H, D_N) = s(R_j, \mathbf{Pa}_j^{H'} \cup \{R_i\}) - s(R_j, \mathbf{Pa}_j^H) = S(G', D_N) - S(G, D_N).$$

Now apply $\{c_N\}$ -consistency to the global comparison between H and H' . In the dependence case $R_j \not\perp_{p^*} R_i \mid \mathbf{Pa}_j^G$, the complete DAG H' imposes no conditional-independence constraints and therefore contains p^* , whereas H enforces the false constraint $R_j \perp R_i \mid \mathbf{Pa}_j^G$ and thus excludes p^* . By $\{c_N\}$ -consistency, $S(H', D_N) > S(H, D_N)$ with probability $\rightarrow 1$, hence $S(G', D_N) > S(G, D_N)$ with probability $\rightarrow 1$.

In the independence case $R_j \perp_{p^*} R_i \mid \mathbf{Pa}_j^G$, both H and H' contain p^* , but H' has strictly more parameters (one extra parent for R_j). By the second clause of $\{c_N\}$ -consistency, $S(H, D_N) > S(H', D_N)$ with probability $\rightarrow 1$, hence $S(G, D_N) > S(G', D_N)$ with probability $\rightarrow 1$. This is exactly $\{c_N\}$ -local consistency.

S.8 Proof of Theorem 4

Conclusion (i) follows directly from Theorem 3. It remains to show Conclusion (ii).

Following the notation in Section 4.3, we conclude that $\tilde{\boldsymbol{\theta}}_N$ is a $\sqrt{f(N)}$ -consistent estimator of $\boldsymbol{\theta}^*$ and $g(N) = \sqrt{N}$. Combining this with $f(N) = o(N \log N)$, we immediately have

$$g(f^{-1}(N)) = \omega\left(\sqrt{\frac{N}{\log N}}\right).$$

By Corollary 2 and the analysis in Section 4.3, the following holds: let G be any DAG and G' be a different DAG obtained by adding the edge $i \rightarrow j$ to G . As $N \rightarrow \infty$, with probability $\rightarrow 1$ we have

(L1) If $Z_j \not\perp_{p^*} Z_i \mid \mathbf{Pa}_j^G$, then $S(G', \hat{\mathbf{Z}}) > S(G, \hat{\mathbf{Z}})$.

(L2) If $Z_j \perp_{p^*} Z_i \mid \mathbf{Pa}_j^G$, then $S(G', \hat{\mathbf{Z}}) < S(G, \hat{\mathbf{Z}})$.

Here $S(\cdot)$ is the BDeu score. Note that BDeu is score equivalent by Theorem 8 in Chickering (1995).

Next, we adopt the reformulation in Nazaret and Blei (2024). For a MEC M , let $\mathcal{I}(M)$ (insertions) and $\mathcal{D}(M)$ (deletions) denote, respectively, the sets of MECs reachable from M by adding or deleting a single edge in some DAG representative of M . Since our score S is

score-equivalent, we write $S(M)$ for the common value $S(G)$ over any $G \in M$. We introduce two propositions:

$P_1(M; \mathbf{p}^*)$: all conditional independencies encoded by M hold in \mathbf{p}^* ,

$P_2(M; \mathbf{p}^*)$: \mathbf{p}^* is DAG-perfect for every DAG in M ,

By Theorems 6–8 in [Nazaret and Blei \(2024\)](#), the following three statements are enough to guarantee correctness of the two-phase greedy search in MEC space:

(A) If $\max_{M' \in \mathcal{I}(M)} S(M') \leq S(M)$, then $P_1(M; \mathbf{p}^*)$ holds.

(B) If $P_1(M; \mathbf{p}^*)$ holds and $M' \in \mathcal{D}(M)$ satisfies $S(M') \geq S(M)$, then $P_1(M'; \mathbf{p}^*)$ holds as well.

(C) If $P_1(M; \mathbf{p}^*)$ holds and $\max_{M' \in \mathcal{D}(M)} S(M') \leq S(M)$, then $P_2(M; \mathbf{p}^*)$ holds.

As a result, our last task is to show these three statements hold for score equivalent scores satisfying (L1) and (L2). The following verification essentially follows the ideas of Proposition 8 and Lemmas 9–10 in [Chickering \(2002\)](#), but we reprove it here for completeness.

Verification of (A).

Assume $\max_{M' \in \mathcal{I}(M)} S(M') \leq S(M)$. We prove $P_1(M; \mathbf{p}^*)$ by contraposition. Suppose $P_1(M; \mathbf{p}^*)$ fails. Then there exist a DAG $G \in M$, a node Z_j , and a non-descendant Z_i of Z_j in G such that $Z_j \not\perp_{\mathbf{p}^*} Z_i \mid \mathbf{Pa}_j^G$. Because Z_i is a non-descendant, adding the edge $Z_i \rightarrow Z_j$ to G does not create a cycle. Moreover, since the conditional independence with respect to \mathbf{Pa}_j^G is violated, this Z_i cannot belong to \mathbf{Pa}_j^G (if $Z_i \in \mathbf{Pa}_j^G$ the statement $Z_j \perp_{\mathbf{p}^*} Z_i \mid \mathbf{Pa}_j^G$ is trivially true). Thus we can form an MEC $M' \in \mathcal{I}(M)$ by inserting $Z_i \rightarrow Z_j$ into G . By (L1) this insertion strictly increases the score, so $S(M') > S(M)$, contradicting $\max_{M' \in \mathcal{I}(M)} S(M') \leq S(M)$. Hence $P_1(M; \mathbf{p}^*)$ must hold.

Verification of (B).

Let M be an MEC such that $P_1(M; \mathbf{p}^*)$ holds, and let $M' \in \mathcal{D}(M)$ satisfy $S(M') \geq S(M)$. Pick $G \in M$ and let $G' \in M'$ be obtained from G by deleting a single edge $Z_i \rightarrow Z_j$, so that the parent set of Z_j changes from \mathbf{Pa}_j^G to $\mathbf{Pa}_j^G \setminus \{Z_i\}$.

Since $P_1(M; \mathbf{p}^*)$ holds, the law \mathbf{p}^* factorizes according to G :

$$\mathbf{p}^*(\mathbf{Z}) = \mathbf{p}^*(Z_j \mid \mathbf{Pa}_j^G) \prod_{k \neq j} \mathbf{p}^*(Z_k \mid \mathbf{Pa}_k^G). \quad (\text{S.80})$$

Now consider the reverse operation that inserts the edge $Z_i \rightarrow Z_j$ into G' . This insertion produces G . On the high-probability event where (L1)–(L2) hold for this insertion, the alternative in (L1) would imply $S(G, \widehat{\mathbf{Z}}) > S(G', \widehat{\mathbf{Z}})$, contradicting $S(G', \widehat{\mathbf{Z}}) \geq S(G, \widehat{\mathbf{Z}})$. Hence the (L2) alternative must hold, which yields

$$\mathbf{p}^*(Z_j \mid Z_i, \mathbf{Pa}_j^G \setminus \{Z_i\}) = \mathbf{p}^*(Z_j \mid \mathbf{Pa}_j^G \setminus \{Z_i\}). \quad (\text{S.81})$$

Define a set of local conditionals for G' by keeping all other conditionals the same as in G and by replacing the conditional at Z_j with

$$q^*(Z_j \mid \mathbf{Pa}_j^G \setminus \{Z_i\}) := \mathbf{p}^*(Z_j \mid \mathbf{Pa}_j^G \setminus \{Z_i\}).$$

Combining (S.80) and (S.81) we obtain,

$$\mathbf{p}^*(\mathbf{Z}) = q^*(Z_j \mid \mathbf{Pa}_j^G \setminus \{Z_i\}) \prod_{k \neq j} \mathbf{p}^*(Z_k \mid \mathbf{Pa}_k^G),$$

which is exactly the factorization of \mathbf{p}^* with respect to G' . Hence \mathbf{p}^* is also represented by G' , and therefore $P_1(M'; \mathbf{p}^*)$ holds (Theorem 6.2 in [Evans \(2025\)](#)).

Verification of (C).

Assume $P_1(M; \mathbf{p}^*)$ holds and $\max_{M' \in \mathcal{D}(M)} S(M') \leq S(M)$. We will show $P_2(M; \mathbf{p}^*)$. Suppose, toward a contradiction, that $P_2(M; \mathbf{p}^*)$ fails. Let G^* be a DAG such that \mathbf{p}^* is

DAG-perfect for G^* , and let M^* be its MEC, so that $\mathcal{I}(\mathbf{p}^*) = \mathcal{I}(M^*)$. Since $P_2(M; \mathbf{p}^*)$ fails, we have $M^* \neq M$. Since $P_1(M; \mathbf{p}^*)$ holds, we have $\mathcal{I}(M) \subseteq \mathcal{I}(\mathbf{p}^*) = \mathcal{I}(M^*)$, and hence $\mathcal{I}(M) \subsetneq \mathcal{I}(M^*)$. Pick representatives $G \in M$ and $G^* \in M^*$. By Theorem 4 in Chickering (2002), there is a finite sequence of single-edge operations (covered edge reversals and edge deletions) that transforms G into G^* , and along the entire sequence $\mathcal{I}(G^*) \supseteq \mathcal{I}(G')$ holds for every intermediate DAG G' . In particular, there exists a first deletion in the sequence, say $G \rightarrow G_1$ obtained by removing $Z_i \rightarrow Z_j$, with $\mathcal{I}(G^*) \supseteq \mathcal{I}(G_1) \supseteq \mathcal{I}(G)$. Let M_1 denote the MEC of G_1 .

Because $\mathcal{I}(G^*) \supseteq \mathcal{I}(G_1)$ and \mathbf{p}^* is DAG-perfect for G^* , every conditional independence encoded by G_1 holds in \mathbf{p}^* . Deleting $Z_i \rightarrow Z_j$ from G yields G_1 with $\mathbf{Pa}_j^{G_1} = \mathbf{Pa}_j^G \setminus \{Z_i\}$ and Z_i a non-descendant of Z_j in G_1 . Thus, by the local Markov property in G_1 , we have in particular $Z_j \perp\!\!\!\perp Z_i \mid \mathbf{Pa}_j^G \setminus \{Z_i\}$, and so this independence holds in \mathbf{p}^* , namely $Z_j \perp\!\!\!\perp_{\mathbf{p}^*} Z_i \mid \mathbf{Pa}_j^G \setminus \{Z_i\}$. By (L2) applied to the reverse insertion that adds $Z_i \rightarrow Z_j$ to G_1 (thereby recovering G), we have $S(G, \widehat{\mathbf{Z}}) < S(G_1, \widehat{\mathbf{Z}})$ for large N with probability $\rightarrow 1$. Therefore, the deletion $G \rightarrow G_1$ strictly increases the score, and by score equivalence $S(M_1) > S(M)$ with probability $\rightarrow 1$. But $M_1 \in \mathcal{D}(M)$, contradicting $\max_{M' \in \mathcal{D}(M)} S(M') \leq S(M)$.

Remark 3. In Chickering (2002, Prop. 8; Lemmas 9–10), the proof of GES correctness implicitly mixes local consistency with consistency. In this paper we follow the XGES reformulation (Nazaret and Blei, 2024) and provide a new proof using only the two local-consistency conditions (L1)–(L2), thereby avoiding any appeal to consistency.

S.9 Implementation Details

S.9.1 Assumptions regarding the penalty function

For completeness, we spell out the shape assumptions on the penalty p_{λ_N, τ_N} and tuning parameters (λ_N, τ_N) used in Theorem 3. For some $\lambda_N, \tau_N > 0$, $p_{\lambda_N, \tau_N} : \mathbb{R} \rightarrow [0, \infty)$ is a sparsity-inducing symmetric penalty that is nondecreasing on $[0, \infty)$, nondifferentiable at 0,

differentiable on $(0, \tau_N)$, and satisfies

$$p_{\lambda_N, \tau_N}(b) = 0 \text{ if } b = 0, \quad p'_{\lambda_N, \tau_N}(b) \leq C \frac{\lambda_N}{\tau_N} \text{ if } |b| \leq \tau_N, \quad p_{\lambda_N, \tau_N}(b) = \lambda_N \text{ if } |b| \geq \tau_N,$$

for some constant $C < \infty$ independent of N .

S.9.2 Detailed Initialization Algorithm

Let $\mu : H \rightarrow \mathcal{X}_j$ denotes the known mean function of the observed-layer parametric family as described in (1).

Algorithm 3: Spectral initialization

Data: \mathbf{X}, K , function $\tilde{g} = \mu \circ g$, truncation parameters ϵ, δ

. Algorithm 3 summarizes this algorithm.

1. Apply SVD to \mathbf{X} and write $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$, where $\Sigma = \text{diag}(\sigma_i)$ and $\sigma_1 \geq \dots \geq \sigma_J$.
2. Let $\mathbf{X}_{\tilde{K}} = \sum_{k=1}^{\tilde{K}} \sigma_k \mathbf{u}_k \mathbf{v}_k^\top$, where $\tilde{K} := \max\{K + 1, \arg \max_k \{\sigma_k \geq 1.01\sqrt{N}\}\}$.
3. Define $\widehat{\mathbf{X}}_{\tilde{K}}$ by truncating $\mathbf{X}_{\tilde{K}}$ to the range of responses, at level ϵ .
4. Define $\widehat{\mathbf{L}}$ by letting $\widehat{l}(i, j) = \tilde{g}^{-1}(\widehat{y}_{\tilde{K}}(i, j))$.
5. Let $\widehat{\mathbf{L}}_0$ be the centered version of $\widehat{\mathbf{L}}$, that is, $\widehat{l}_0(i, j) = \widehat{l}(i, j) - \frac{1}{N} \sum_{k=1}^N \widehat{l}(k, j)$.
6. Apply SVD to $\widehat{\mathbf{L}}_0$ and write its rank- K approximation as $\widehat{\mathbf{L}}_0 \approx \widehat{\mathbf{U}}\widehat{\Sigma}\widehat{\mathbf{V}}$.
7. Let $\tilde{\mathbf{V}}$ be the rotated version of $\widehat{\mathbf{V}}$ according to the Varimax criteria.
8. Entrywise threshold $\tilde{\mathbf{V}}$ at δ to induce sparsity, and flip the sign of each column so that all columns have positive mean. Let $\widehat{\mathbf{Q}}$ be the estimated sparsity pattern.
9. Estimate the centered \mathbf{Z}_0 by $\widehat{\mathbf{Z}}_0 := \widehat{\mathbf{L}}_0 \tilde{\mathbf{V}} (\tilde{\mathbf{V}}^\top \tilde{\mathbf{V}})^{-1}$, and estimate \mathbf{Z} by reading off the signs: $\widehat{Z}(i, k) = \mathbb{1}(Z_0(i, k) > 0)$.
10. Let $\widehat{\mathbf{Z}}_{\text{long}} := [\mathbf{1}, \widehat{\mathbf{Z}}]$. Estimate \mathbf{B} by $\widehat{\mathbf{B}} := C_g ((\widehat{\mathbf{Z}}_{\text{long}}^\top \widehat{\mathbf{Z}}_{\text{long}})^{-1} \widehat{\mathbf{Z}}_{\text{long}}^\top \widehat{\mathbf{L}}_0) \cdot \widehat{\mathbf{G}}$, where \cdot is the element-wise product and C_g is a positive constant.
11. Define $\widehat{\mathbf{R}} = \mathbf{X} - \widehat{\mathbf{Z}}_{\text{long}} \widehat{\mathbf{X}}^\top$ and estimate γ_j by $\widehat{\gamma}_j = \frac{1}{N} \sum_{i=1}^N \widehat{r}(i, j)^2$.

Output: $\widehat{\boldsymbol{\rho}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{Z}}$.

We explain more on the truncation details in Step 3 by considering specific response

types. For Normal responses, the original sample space is \mathbb{R} and the truncation (Steps 1-4 in Algorithm 3) may be omitted. For Binary responses, we set

$$\widehat{x}_K(i, j) = \begin{cases} \epsilon, & \text{if } x_K(i, j) = 0, \\ 1 - \epsilon, & \text{if } x_K(i, j) = 1. \end{cases}$$

For Poisson responses, we set

$$\widehat{x}_K(i, j) = \begin{cases} \epsilon, & \text{if } x_K(i, j) < \epsilon, \\ x_K(i, j), & \text{otherwise.} \end{cases}$$

In terms of implementing the method, we follow the suggestions of Zhang et al. (2020) with $\epsilon = 10^{-4}$.

S.9.3 Simulation Setup Details

In all simulations, we set \mathbf{Q} and \mathbf{p} as follows. The measurement matrix \mathbf{Q} takes the form

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}' \\ \mathbf{I}_K \\ \mathbf{I}_K \end{pmatrix}, \quad \mathbf{Q}'_1 = \begin{pmatrix} 1 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 1 \end{pmatrix}_{K \times K}, \quad \mathbf{Q}'_2 = \begin{pmatrix} 1 & 1 & 1 & & 0 \\ 1 & \ddots & \ddots & \ddots & \\ 1 & \ddots & \ddots & \ddots & 1 \\ & \ddots & \ddots & \ddots & 1 \\ 0 & & 1 & 1 & 1 \end{pmatrix}_{K \times K}.$$

We consider two banded choices for the submatrix \mathbf{Q}' : \mathbf{Q}'_1 and \mathbf{Q}'_2 , and denote the corresponding full matrices by \mathbf{Q}_1 , \mathbf{Q}_2 respectively. Both choices satisfy the identifiability conditions in Theorem S.1.

Given a DAG on the latent variables, we define the distribution \mathbf{p} so as to yield balanced conditional probabilities and avoid degeneracy. If a child has a single parent, then when the

parent equals 1 the Bernoulli parameter of the child is drawn uniformly from $[0.3, 0.35]$ or from $[0.65, 0.7]$ with equal probability, and the parameter for parent 0 is set to be the complement. If a child has two parents, we consider the four parent configurations $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. For configuration $(0, 0)$ we draw the Bernoulli parameter independently from $[0.2, 0.25]$, and for $(1, 1)$ we draw it from $[0.6, 0.65]$. For the mixed configurations $(0, 1)$ and $(1, 0)$, we draw one parameter from $[0.35, 0.4]$ and the other from $[0.77, 0.82]$.

We consider three parametric families for the observed layer: Gaussian, Poisson, and Bernoulli for continuous, count, and binary data, respectively. This allows us to assess the robustness of our method under both continuous and discrete measurement models. Because these families live on different scales, we specify different values for the regression parameters β :

$$\beta_{j,0} = \begin{cases} -1, & \text{Gaussian;} \\ 1, & \text{Poisson/Bernoulli;} \end{cases}$$

$$\beta_{j,k} = \begin{cases} \frac{3}{\sum_{k'=1}^K q_{j,k'}} \mathbf{1}(q_{j,k} = 1), & \forall j \in [J], k \in [K], \quad \text{Gaussian;} \\ \frac{2}{\sum_{k'=1}^K q_{j,k'}} \mathbf{1}(q_{j,k} = 1), & \forall j \in [J], k \in [K], \quad \text{Poisson/Bernoulli.} \end{cases}$$

The variance parameter is fixed to be $\gamma_j = \sigma_j^2 = 1$ for all j .

S.9.4 Choice of $f(N)$

Recalling our algorithmic procedure, we obtain a $f(N) \times K$ matrix by sampling from $\hat{\mathbf{p}}$. Although we have specified the admissible range for $f(N)$, the exact choice within this range remains to be determined. In our simulations, balancing computational efficiency and performance, we record the recovered DAGs when $f(N) = N, 2N, 3N$. In these three cases, the matrices on which GES is applied are denoted as $\hat{\mathbf{Z}}_1, \hat{\mathbf{Z}}_2, \hat{\mathbf{Z}}_3$. We record the averaged SHD (Structural Hamming Distance) for different scenarios in Table S.3

Model	\mathbf{Q}	$\widehat{\mathbf{Z}}$	Bernoulli			Poisson			Lognormal		
			3000	5000	7000	3000	5000	7000	3000	5000	7000
Chain-10	\mathbf{Q}_1	$\widehat{\mathbf{Z}}_1$	5.55	3.294	2.938	2.248	1.362	0.638	0.412	0.254	0.122
		$\widehat{\mathbf{Z}}_2$	5.966	4.92	4.88	2.284	1.986	1.206	0.992	0.49	0.33
		$\widehat{\mathbf{Z}}_3$	7.852	6.91	7.344	3.518	3.178	2.542	1.82	1.246	0.95
	\mathbf{Q}_2	$\widehat{\mathbf{Z}}_1$	5.664	3.258	3.042	2.174	0.704	0.488	0.406	0.208	0.146
		$\widehat{\mathbf{Z}}_2$	6.042	4.784	5.314	2.24	1.306	1.048	0.894	0.562	0.328
		$\widehat{\mathbf{Z}}_3$	8.058	7.016	7.54	3.534	2.53	2.272	1.75	1.286	1.026
Tree-10	\mathbf{Q}_1	$\widehat{\mathbf{Z}}_1$	4.372	2.308	1.308	1.85	1.692	1.36	0.94	0.51	0.308
		$\widehat{\mathbf{Z}}_2$	5.024	3.674	3.068	3.09	2.656	2.66	1.528	0.788	0.344
		$\widehat{\mathbf{Z}}_3$	7.07	5.566	5.106	4.622	4.518	3.982	2.388	1.32	1.032
	\mathbf{Q}_2	$\widehat{\mathbf{Z}}_1$	4.3	1.936	1.26	2.676	1.556	1.146	1.14	0.618	0.346
		$\widehat{\mathbf{Z}}_2$	5.152	3.346	2.936	3.62	2.41	2.218	1.582	0.878	0.496
		$\widehat{\mathbf{Z}}_3$	7.26	5.368	4.964	5.22	4.136	3.706	2.66	1.578	0.882
Model-7	\mathbf{Q}_1	$\widehat{\mathbf{Z}}_1$	7.692	6.334	5.798	5.838	5.422	4.892	0.196	0.042	0
		$\widehat{\mathbf{Z}}_2$	7.352	6.274	6.132	5.278	4.754	4.786	0.06	0.036	0.014
		$\widehat{\mathbf{Z}}_3$	7.806	6.976	6.994	5.18	5.362	6.288	0.144	0.1	0.102
	\mathbf{Q}_2	$\widehat{\mathbf{Z}}_1$	7.554	6.304	5.68	5.848	5.526	5.082	0.262	0.054	0.004
		$\widehat{\mathbf{Z}}_2$	7.312	6.238	6.022	5.39	5.048	5.124	0.1	0.036	0.036
		$\widehat{\mathbf{Z}}_3$	7.714	6.83	6.87	5.148	5.452	6.552	0.188	0.098	0.084
Model-8	\mathbf{Q}_1	$\widehat{\mathbf{Z}}_1$	4.336	2.682	2.19	2.106	1.916	1.878	0.132	0.048	0
		$\widehat{\mathbf{Z}}_2$	5.012	3.498	3.336	2.356	2.238	2.742	0.172	0.084	0.058
		$\widehat{\mathbf{Z}}_3$	6.354	5.108	5.532	3.17	3.292	4.336	0.416	0.248	0.22
	\mathbf{Q}_2	$\widehat{\mathbf{Z}}_1$	4.342	2.72	2.374	2.264	1.9	1.782	0.202	0.052	0.002
		$\widehat{\mathbf{Z}}_2$	4.992	3.416	3.29	2.516	2.388	2.588	0.15	0.07	0.036
		$\widehat{\mathbf{Z}}_3$	6.264	4.952	5.454	3.336	3.808	4.17	0.498	0.302	0.25
Model-13	\mathbf{Q}_1	$\widehat{\mathbf{Z}}_1$	22.37	16.454	14.062	24.65	16.472	14.162	3.206	1.646	1.008
		$\widehat{\mathbf{Z}}_2$	23.274	17.152	15.544	24	16.106	15.258	3.138	1.788	1.17
		$\widehat{\mathbf{Z}}_3$	25.134	19.594	18.24	25.16	18.706	18.732	3.89	2.554	1.728
	\mathbf{Q}_2	$\widehat{\mathbf{Z}}_1$	22.252	16.29	14.606	25.134	15.626	12.934	3.032	1.872	0.994
		$\widehat{\mathbf{Z}}_2$	23.348	16.816	15.622	24.686	15.162	14.258	2.852	2.082	1.158
		$\widehat{\mathbf{Z}}_3$	25.262	19.334	18.68	25.612	17.592	17.46	3.624	2.696	1.68

Table S.3: SHD.

One can find that although $\widehat{\mathbf{Z}}_1$ achieves the best results in most cases, $\widehat{\mathbf{Z}}_2$ and $\widehat{\mathbf{Z}}_3$ can occasionally outperform it. Therefore, we regard $\frac{f(N)}{N}$ as a tuning parameter, and the choice

of its optimal value remains an open question. At this stage, we recommend using $\widehat{\mathbf{Z}}_1$.

S.9.5 Q-matrix for the TIMSS data

This subsection provides the Q-matrix used in Section 6.1; see Table S.4.

Table S.4: Q-matrix for the TIMSS 2019 math assessment booklet 14.

Item ID	Number	Algebra	Geometry	Data and Prob.	Knowing	Applying	Reasoning
1	1	0	0	0	1	0	0
2	1	0	0	0	0	1	0
3	1	0	0	0	0	1	0
4	1	0	0	0	1	0	0
5	0	1	0	0	0	1	0
6	0	1	0	0	1	0	0
7	0	1	0	0	0	0	1
8	0	1	0	0	0	1	0
9	0	0	1	0	0	0	1
10	0	0	1	0	0	1	0
11	0	0	1	0	0	1	0
12	0	0	0	1	0	1	0
13	0	0	0	1	0	1	0
14	1	0	0	0	1	0	0
15	1	0	0	0	1	0	0
16	1	0	0	0	0	0	1
17	1	0	0	0	0	1	0
18	0	1	0	0	1	0	0
19	0	1	0	0	1	0	0
20	0	1	0	0	0	1	0
21	0	1	0	0	0	1	0
22	0	1	0	0	0	1	0
23	0	0	1	0	1	0	0
24	0	0	1	0	0	0	1
25	0	0	1	0	0	0	1
26	0	0	1	0	0	1	0
27	0	0	0	1	0	1	0
28	0	0	0	1	0	1	0
29	0	0	0	1	0	0	1

S.9.6 Seesaw Image Generator and Preprocessing Details

For each sample $i \in [10000]$, we draw (Z_{1i}, Z_{2i}) independently as Bernoulli(0.5). Then Z_{3i} is generated from a stochastic seesaw-response rule

$$Z_{3i} \sim \text{Bernoulli}(p_3(Z_{1i}, Z_{2i})), \quad p_3(1, 1) = 0.8, \quad p_3(z_1, z_2) = 0.2 \text{ for } (z_1, z_2) \neq (1, 1).$$

The occluded-ball visibility Z_{4i} is generated conditionally on Z_{3i} as

$$Z_{4i} \sim \text{Bernoulli}(p_4(Z_{3i})), \quad p_4(1) = 0.99, \quad p_4(0) = 0.$$

Given a latent configuration (z_1, z_2, z_3, z_4) , the generator renders a 256×256 RGB scene and converts it to grayscale. The scene consists of a fixed pivot and a rigid plank (the seesaw) that rotates by an angle

$$\alpha(z_3) = \begin{cases} -\alpha_0, & z_3 = 1 \text{ (left side up),} \\ +\alpha_0, & z_3 = 0 \text{ (left side down),} \end{cases} \quad \text{with } \alpha_0 = 25^\circ.$$

A left ball is placed on top of the plank near its left endpoint. The two tray balls (z_1 and z_2) lie on two well-separated slots of a horizontal forked tray whose position is fixed in the image, so the tray does not rotate with the seesaw.

Across images, we introduce mild nuisance variation through small i.i.d. positional jitter in the seesaw subsystem. In particular, for each rendered scene we perturb the pivot and plank center by independent uniform offsets in $[-\delta, \delta]$ along each axis, with $\delta = 0.001$ in normalized coordinates, and we apply an additional small shared offset to the seesaw-side balls. This prevents the images from being identical templates within the same latent state and makes the learning task more challenging, while preserving the intended semantics.

The fourth ball is positioned at the left ball’s canonical “down” location (the $z_3 = 0$

geometry) and is drawn before the left ball so that, when $z_3 = 0$, the left ball occludes it. When $z_3 = 1$, the left ball moves upward and the fourth ball may become visible if $z_4 = 1$.

Let $X_i^{\text{original}} \in \{0, 1, \dots, 255\}^{256 \times 256}$ denote the grayscale image for sample i . We then construct an inverted binary mask at the original resolution by thresholding at 80: pixels with grayscale value greater than 80 are set to 1, and dark pixels (the balls) are set to 0. We resize this binary mask to 96×96 using nearest-neighbor interpolation to preserve sharp object boundaries, and denote the result by $X_i^{\text{mask}} \in \{0, 1\}^{96 \times 96}$. Under this convention, ball pixels are coded as 0 and non-ball pixels are coded as 1.

Finally, we produce the 16×16 pooled representation $X_i^{\text{pooled}} \in \{0, 1\}^{16 \times 16}$ by applying min-pooling over non-overlapping blocks of the 96×96 mask, yielding 256 binary features per image. Because the mask is inverted, a pooled entry equals 0 if at least one ball pixel appears in that spatial cell, and equals 1 otherwise.

Therefore, we obtain

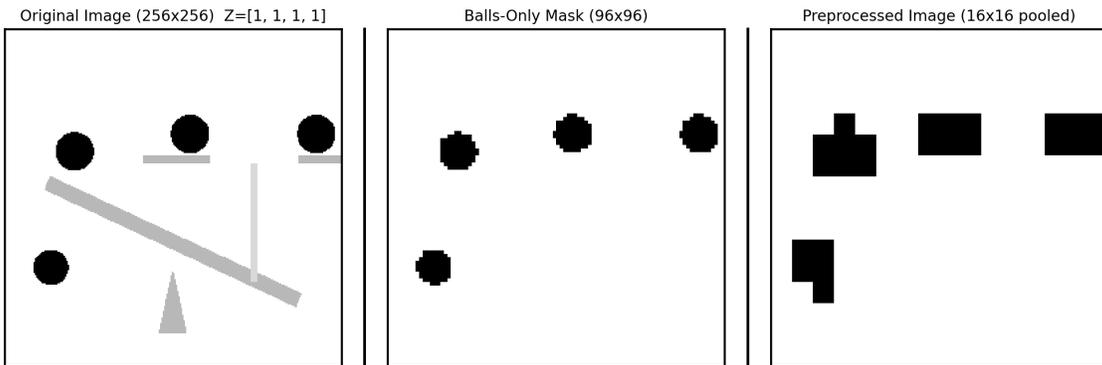
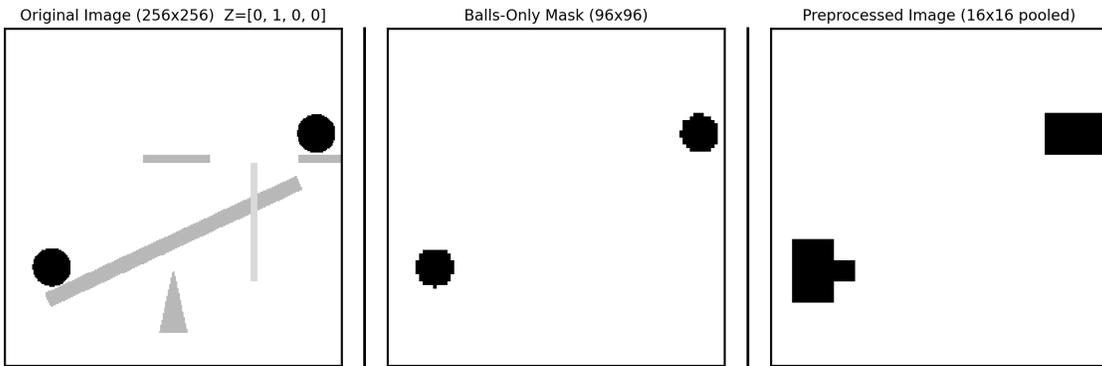
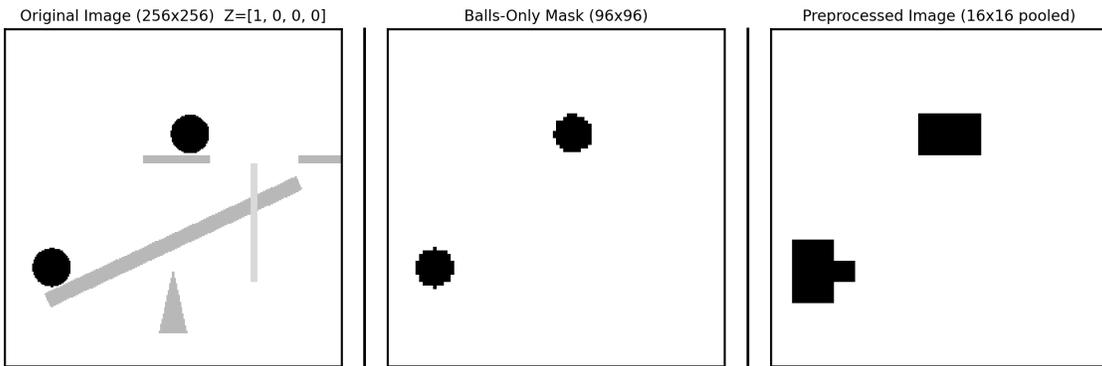
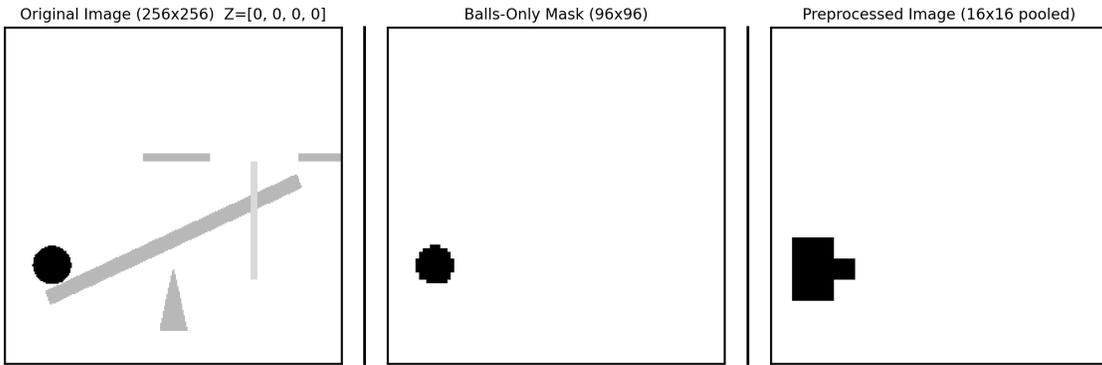
$$\mathbf{X}_{\text{original}} \in \{0, \dots, 255\}^{10000 \times 256^2}, \quad \mathbf{X}_{\text{mask}} \in \{0, 1\}^{10000 \times 96^2},$$

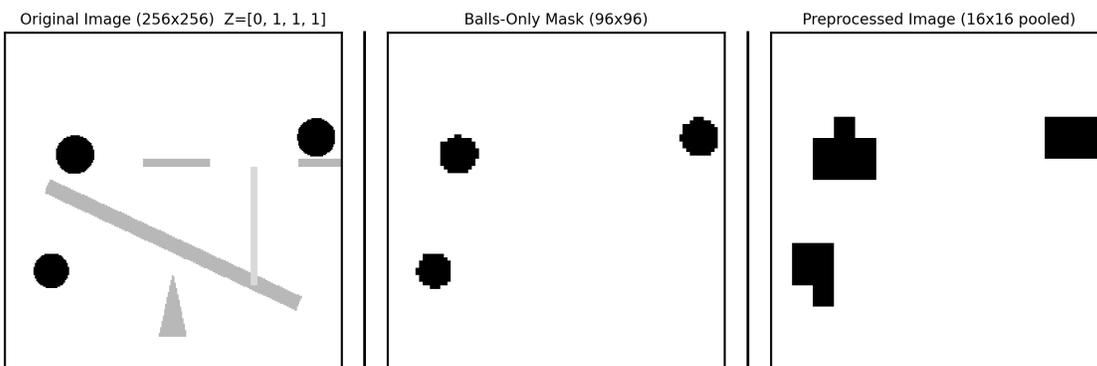
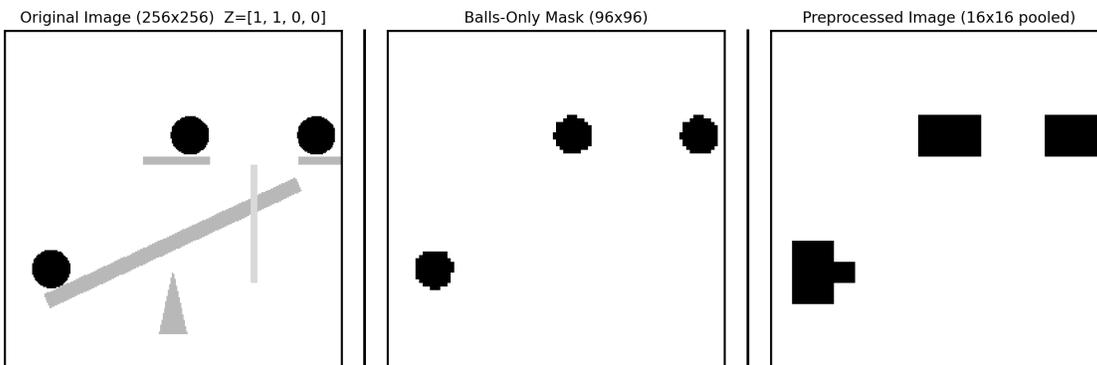
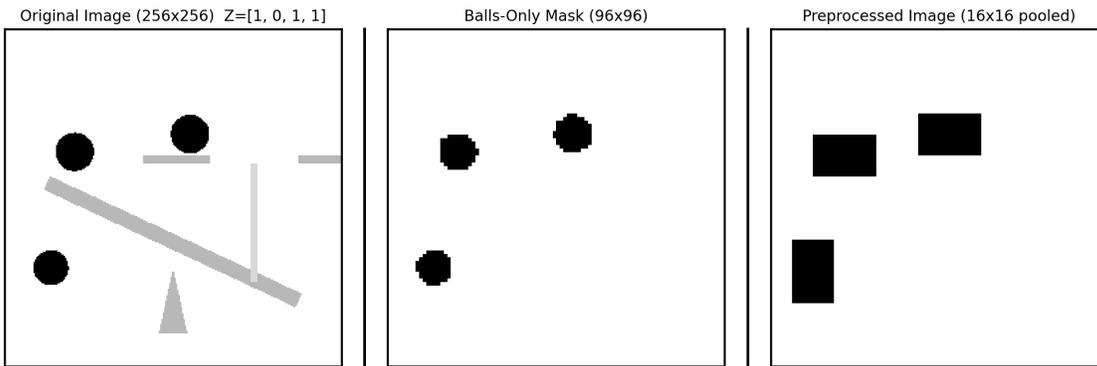
$$\mathbf{X}_{\text{pooled}} \in \{0, 1\}^{10000 \times 256}, \quad \mathbf{Z} \in \{0, 1\}^{10000 \times 4},$$

where each row corresponds to one rendered scene and its associated latent state. Example images are shown below. In each figure, the single row displays (from left to right) the original image, the balls-only mask, and the pooled representation. We fit our model using the pooled data $\mathbf{X}_{\text{pooled}}$.

S.10 Connections with Existing Studies

We emphasize that our statistical specification of causal mechanisms is deliberately broad and flexible. For the causal structure among latent variables, we allow for arbitrarily complex dependencies among the latent factors. The only structural limitation is that the latent variables are taken to be discrete. Far from being restrictive, this choice has proven espe-





cially fruitful. On the theoretical side, assuming discrete latent variables allows us to employ powerful identifiability results from mixture and latent class models, thereby greatly facilitating more rigorous analysis (Teicher, 1967; Yakowitz and Spragins, 1968; Allman et al., 2008). On the modeling side, discrete latent hierarchies have formed the basis of influential architectures in machine learning, such as deep Boltzmann machines (DBMs) (Salakhutdinov and Hinton, 2009), deep belief networks (DBNs) (Hinton et al., 2006). DBMs and DBNs in particular were originally designed with multiple binary latent layers (Salakhutdinov, 2015). Moreover, a collection of K binary latent variables yields 2^K possible latent configurations, so that a relatively small number of latent factors can encode a combinatorially large family of data-generating regimes. This corresponds to a distributed representation in the usual sense of representation learning: each regime is encoded by a pattern of activations across multiple latent units, rather than a single categorical latent with 2^K levels. These examples show that discrete latents are sufficiently rich to capture complex data distributions while yielding parsimonious representations and tractable identifiability analysis.

For the links between latent and observable variables, we adopt a specification in the spirit of all-effect general-response CDMs (GR-CDMs) (Liu et al., 2025). Broadly speaking, cognitive diagnosis models (CDMs) are latent variable models in which each subject possesses a vector of binary latent attributes indicating mastery versus non-mastery on a collection of skills, and each item is designed to depend only on a specified subset of these attributes, typically encoded by a binary design matrix. Our CDM-style measurement layer is motivated by the remarkable success of CDMs in educational measurement, where they have proven to be powerful tools for modeling multidimensional discrete skills with both mature identifiability theory (Lee and Gu, 2024) and rich representational capacity. In the binary-response case, our parameterization naturally subsumes well-known models such as the additive CDM (ACDM) and the generalized DINA (G-DINA) (de la Torre, 2011), while for continuous responses it also includes extensions such as the continuous DINA (cDINA)

model under positive outcomes (Minchen et al., 2017). In this way, the same decomposition provides a unified framework that accommodates polytomous, continuous, and mixed responses, while allowing higher-order interactions when supported by data.

Prashant et al. (2025) investigate causal discovery in hierarchical latent-variable models whose observed and latent variables are all modeled as continuous. While the functional relationships among variables are quite flexible, their identifiability theory imposes a strong structural restriction on the latent DAG: latent variables are partitioned into hierarchical layers and edges are permitted only across layers. Consequently, the recoverable graphs are essentially concatenations of bipartite graphs between successive layers. By contrast, our framework allows arbitrary latent DAGs and provides a unified treatment of both continuous and discrete observed variables.

Recent work by Dong et al. (2026) also studies a score-based greedy search procedure for partially observed causal models. Their theory is developed for a linear-Gaussian latent-variable SEM, where the greedy search compares maximized scores that depend on the observed covariance matrix Σ_X . In our framework, the latent variables are discrete and the likelihood of X is not determined by Σ_X alone. Therefore, the covariance-based Gaussian scoring theory of Dong et al. (2026) is not applicable to our setting.

Kivva et al. (2021) consider a general latent-causal setting and reduce recovery to a mixture-oracle problem. But as already discussed in Section 5, this generality comes at the price of not performing well in all settings.