The relative value of interventional and observational samples in Bayesian Causal Linear Gaussian Models

Valentinian Lungu[1] **Anish Dhir**[2] **Mark van der Wilk**[3] **Ioannis Kontoyiannis**[1]

[1]Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge
[2]Imperial College London
[3]University of Oxford

Abstract

We investigate the asymptotic properties of Bayesian bivariate causal discovery for Gaussian Linear Structural Equation Models (SEMs) with heteroscedastic noise. We demonstrate that with purely observational data, the posterior distribution over the models fails to consistently identify the true causal structure—a consequence of the fundamental non-identifiability within the Markov Equivalence Class. Specifically, if the true generating mechanism corresponds to a connected graph ($A \to B$ or $B \to A$), the asymptotic behavior of the posterior is given by the ratio between the prior on the true model and the push-forward prior of the alternative. In contrast, for the independence model, we establish that the posterior concentrates at a stochastic polynomial rate of $O_p(n^{-1/2})$.
To resolve this non-identifiability, we incorporate m interventional samples and characterize the concentration rates as a function of the observational-to-total sample ratio, η. We identify a sharp *concentration dichotomy*: while the independence graph maintains a polynomial $O_p(N^{-1/2})$ rate (where $N = n + m$), connected graphs undergo a phase transition to exponentially fast convergence. This highlights an *exponential* relative importance between the two data types, as altering the amount of one data type directly changes the exponent governing the concentration speed. We derive explicit formulae for the exponential decay rates and provide precise conditions under which mixing observational and interventional data optimizes concentration speed. Finally, our theoretical findings are validated through empirical simulations in Bayesian Gaussian equivalent (BGe)-style prior specifications offering a principled foundation for experimental design in Bayesian causal discovery.

1 INTRODUCTION

Over the last several decades, there has been a significant surge of interest in developing robust and scalable algorithms for detecting hidden dependencies between variables. In the causal modeling framework, these variables are typically represented as nodes in a Directed Acyclic Graph (DAG), where directed edges represent causal mechanisms. The researcher's objective is to perform *structure learning*—identifying the graph that best explains the observed data. This problem of causal graph identification is foundational across diverse domains, including the mapping of gene regulatory networks Lozano et al. [2019], the evaluation of clinical treatments Wu et al. [2024], and the optimization of advertising strategies Van den Broeck et al. [2018].

There is a rich set of causal discovery tools. These include constraint-based methods like the PC algorithm Spirtes and Glymour [1991], score-based optimization approaches Peters and Bühlmann [2014], Meinshausen and Bühlmann [2006], Zheng et al. [2018], and Bayesian methodologies Annadani et al. [2023], Cundy et al. [2021], Friedman and Koller [2003], Annadani et al. [2021], Deleu et al. [2022], Eaton and Murphy [2007], Cao et al. [2019]. More recently, meta-learning frameworks based on Neural Processes have also been explored in Dhir et al. [2025]and Dhir et al. [2026], offering new avenues for amortized causal inference. While optimization-based methods typically yield a point estimate of the causal graph, Bayesian/amortized approaches are usually preferred because they provide a posterior distribution over the graph space. This uncertainty quantification is particularly vital in safety-critical domains, where distinguishing between strong causal evidence and mere lack of data is essential for robust decision-making. Although these works show good prediction power, a deep theoretical approach is still missing.

In this work, we focus on the Bayesian theoretical treatment of *Gaussian Linear Structural Equation Models (SEMs)*. These linear models have been rigorously studied from both frequentist Peters and Bühlmann [2014], Pearl [2009] and

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4067–4099.

Bayesian Heckerman et al. [2006] perspectives. Notably, Peters and Bühlmann [2014] demonstrated that if Gaussian errors are homoscedastic (sharing a common variance) or if variances are known, the causal structure is identifiable from purely observational data. Conversely, under unknown heteroscedastic noise, the model is identifiable only up to its *Markov Equivalence Class (MEC)*. From a Bayesian standpoint, Heckerman et al. [2006] introduced a specific class of priors that satisfy *score equivalence*, ensuring the posterior assigns equal mass to all structures within the same MEC. This is commonly known as the *Bayesian Gaussian equivalent (BGe)* score.

This paper presents a two-fold investigation into posterior behavior in non-identifiable and interventional settings. First, we examine the bivariate Gaussian SEM with heteroscedastic noise. We demonstrate that when a causal edge exists between two nodes, the posterior of the true model fails to concentrate to 1 as the sample size n grows large. Instead, the posterior limit is determined by the ratio between the prior on the true model and the *push-forward prior* of the alternative model. This result generalizes the findings by Dhir et al. [2024] and Geiger and Heckerman [2002], which showed that a ratio of 1 leads to an uninformative posterior split between the two causal directions. In contrast, when the independence model is the true generating mechanism, we find that the posterior concentrates to 1 at a stochastic rate of $O_p(1/\sqrt{n})$.

To resolve such non-identifiability, researchers often leverage *hard interventions* to distinguish between direct causation and latent confounding Pearl [2009]. However, in many practical scenarios—such as gene knockout experiments or clinical trials—interventions are costly or technically difficult, while observational data is abundant. This creates a critical trade-off: determining how many expensive interventional samples are required to augment cheap observational data to achieve identifiability. By forcing a variable to a fixed value, an intervention "changes" the graph, severing its natural dependencies and isolating its downstream effects.

The second part of this work analyzes the scenario where the researcher has access to m interventional samples alongside n observational samples. While the sample complexity of mixing observational and interventional data has been recently studied as a *closeness testing* problem for discrete Acharya et al. [2023] and continuous Jamshidi et al. [2025] data, we provide the first comprehensive theoretical investigation into *posterior concentration* as a function of the observational sample ratio $\eta = n/N$, where $N = n+m$. We identify a second dichotomy in concentration speeds similar to the one observed in the homoscedastic setting described by Lungu et al. [2025]: for the connected graphs, the posterior concentrates to 1 at an *exponential rate*, with an exponent characterized by a weighted sum of *relative entropies* (Kullback-Leibler divergences) between the observational and interventional densities. This result implies an *exponential relative dependence* between the two data regimes; specifically, the sample composition directly determines the exponent of the convergence rate, such that even marginal shifts in the data mixture result in exponential-scale changes to the speed of identification. Our analysis distinguishes between interventions on the cause versus on the effect. While intervening on the effect intuitively benefits from a mix of data, one might assume that intervening on the cause renders observational data redundant. Counter-intuitively, we show that for *moderate SNR regimes and conservative interventions*, observational samples remain important. In the latter setting, the interventional signal alone may be insufficient to distinguish the causal mechanism from the noise floor; observational data helps by allowing the posterior to calibrate the baseline variance and more effectively "learn" the structure. In contrast to this exponential regime, we establish that for the independence model, the posterior convergence undergoes a phase transition, slowing to a stochastic polynomial rate of $O_p(1/\sqrt{N})$.

2 RELATED WORK

The problem of identifying generating structures from observational data has been a central focus of the SEM literature. Early constraint-based approaches, such as the PC and FCI algorithms Spirtes et al. [2000], leverage conditional independence tests to recover the causal graph. However, because these methods rely on the faithfulness assumption, they are generally limited to recovering the Markov Equivalence Class (MEC) rather than the unique Directed Acyclic Graph (DAG). Along the same lines, score-based methods were developed to search for graphs that maximize a scoring function, often derived from parametric assumptions Chickering [2002], Heckerman et al. [2006].

In the linear-Gaussian setting, the Bayesian Gaussian equivalent (BGe) score is a standard choice Geiger and Heckerman [2021]. While computationally attractive, linear-Gaussian models suffer from a fundamental limitation known as the "Gaussian Trap": the covariance matrix of the observed variables generally does not contain sufficient information to distinguish between Markov equivalent graphs (e.g., $A \to B$ vs. $B \to A$) Shimizu et al. [2006]. Exceptions exist, such as the result by Peters and Bühlmann [2014], which proves identifiability in the special case where all error terms share equal variance. This problem is, however, an artifact of the distributional assumption. The introduction of the Linear Non-Gaussian Acyclic Model (LiNGAM) by Shimizu et al. [2006] demonstrated that if error terms are non-Gaussian, the full causal structure is identifiable using Independent Component Analysis (ICA) and the Darmois-Skitovich theorem. Building on this, Hoyer and Hyttinen [2012] proposed Bayesian LiNGAM to combine the identifiability of non-Gaussian models with the uncertainty quantification of the

Bayesian paradigm. Rather than outputting a single point-estimate graph, Bayesian LiNGAM computes the posterior probability over DAGs, using non-Gaussian likelihoods to break the symmetry of the MEC.

While non-Gaussianity provides a statistical route to identifiability, interventions offer a physical route. Interventional data fundamentally alters the search space by refining the MEC into the smaller Interventional Markov Equivalence Class (I-MEC) Hauser and Bühlmann [2012]. Interventions are typically categorized as "hard" or "soft." A hard intervention, denoted by $do(A = a)$, forces a variable to a specific value, effectively removing all incoming edges to A. In contrast, a soft (or parametric) intervention modifies the conditional distribution of A—for instance, by shifting the noise mean or variance—without changing the dependency on its parents. Hauser and Bühlmann [2012] characterized the I-MEC for hard interventions, showing that as the set of intervention targets expands, the I-MEC shrinks, eventually collapsing to the unique true DAG.

Recent interest has shifted focus from asymptotic identifiability to finite-sample complexity, addressing the practical trade-offs between data types. Acharya et al. [2023] investigate the sample complexity of distinguishing cause from effect in discrete bivariate settings. Their work establishes tight bounds on the number of interventional samples (m) needed as a function of available observational samples (n).

In the continuous domain, Jamshidi et al. [2025] address the problem of detecting causal direction/hidden confounding from interventional and observational data. They frame it as a non-parametric closeness testing problem. Estimating divergence measures like the KL divergence in continuous, high-dimensional spaces is notoriously difficult due to estimation bias. To mitigate this, the authors propose an estimator based on the Von Mises expansion, establishing the first sample complexity guarantees for distinguishing cause from effect in continuous, non-linear systems with potential hidden confounders.

3 NON-IDENTIFIABLE SETTING

In this work, we consider the 2-nodes linear structural equations model

$$X = AX + \epsilon, \tag{1}$$

where $X = [X(1), X(2)]^\intercal \in \mathbb{R}^2$ is a 2-dimensional vector, and $\epsilon = [\epsilon(1), \epsilon(2)]^\intercal \in \mathbb{R}^2$ is joint Gaussian noise with zero-mean and diagonal covariance matrix $\Sigma = Diag(\tau_1^2, \tau_2^2)$. The matrix A is described in terms of its structure S and weight w, where S is a binary matrix with entries $S_{ij} = \mathbb{I}\{A_{ij} \neq 0\}$. Naturally, we refer to S as being the structure matrix and the parameters $\theta := [w, \tau_1^2, \tau_2^2]^\intercal \in \mathbb{R} \times (0, \infty)^2 =: \Theta$ as the associated parameters.

Throughout the article we assume that S is the corresponding adjacency matrix of a DAG on the set $\{1, 2\}$. Since we consider only 2 nodes, S can take values only in the following set $\mathcal{S} := \{S^1, S^2, S^3\}$ where

$$S^1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad S^2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \text{and} \quad S^3 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

For simplicity, we write $A = m(S, w)$ for the map and denote $P_{m(S,\theta)}$ as the law of $X = (I - A^\intercal)^{-1}\epsilon$ with the covariance matrix given by the last two components of θ. The corresponding density for $P_{m(S,\theta)}$ is denoted by $f(x \mid \theta, S)$ and the log-likelihood by $\ell(\theta \mid S) := \log f(x \mid \theta, S)$ with $x = [x(1), x(2)]^\intercal$ for all $S \in \mathcal{S}$.

In the case of n independent and identically distributed (i.i.d.) observations, $\mathcal{D}_{\text{obs}} := (X_i)_{i=1}^n$, generated from (1), we denote the joint likelihood by $f_n(\mathcal{D}_{\text{obs}} \mid \theta, S)$ and by $\ell_n(\theta \mid S)$ the corresponding joint log-likelihood.

For each structure matrix S^i, and the corresponding parameters θ^i we define the model $\mathcal{M}^i$ as follows

$$\mathcal{M}^i = \{\theta^i, S^i\}, \text{ for } i \in \{1, 2, 3\}.$$

3.1 FREQUENTIST APPROACH

Lemma 3.1. *The two structures, S^1 and S^2 are non-identifiable, i.e. for any choice of parameters $\theta^1 \in \Theta$, there exists $\theta^2 \in \Theta$ such that the likelihoods of the two models are the same, i.e. $f(x \mid \theta^1, S^1) = f(x \mid \theta^2, S^2)$ for any $x = [x(1), x(2)] \in \mathbb{R}^2$.*

We denote by $\gamma : \Theta \to \Theta$ the map defined by the parameter transformation described in the proof (see Appendix A.1) of the above lemma . Specifically, if $\theta^1 = [w, \tau_1^2, \tau_2^2]^\intercal$, then the mapping $\theta^2 = \gamma(\theta^1)$ is given by

$$\gamma(w, \tau_1^2, \tau_2^2) = \left[\frac{w\tau_2^2}{w^2\tau_2^2 + \tau_1^2},\ w^2\tau_2^2 + \tau_1^2,\ \frac{\tau_1^2\tau_2^2}{w^2\tau_2^2 + \tau_1^2}\right]^\intercal \tag{2}$$

This mapping formally establishes the non-identifiability result. Since $\gamma(\theta^1)$ is a continuous transformation, and we have shown that for any choice of parameters θ^1 there exists a distinct set of parameters $\theta^2 = \gamma(\theta^1)$ such that $f(x \mid S^1, \theta^1) = f(x \mid S^2, \theta^2)$ for any $x \in \mathbb{R}^2$, the two structures S^1 and S^2 are non-identifiable.

Given n i.i.d. observations $\mathcal{D}_{\text{obs}} := (X_i)_{1 \leq i \leq n}$ generated from (1), we calculate the maximum likelihood estimators (MLEs) for the three structures considered: $\hat{\theta}^1$ for S^1 (where $X(1) \to X(2)$), $\hat{\theta}^2$ for S^2 (where $X(2) \to X(1)$) and $\hat{\theta}^3$ for the independence model S^3. Lemma A.1 establishes that $\hat{\theta}_2 = \gamma(\hat{\theta}_1)$, confirming again that S^1 and S^2 are observationally equivalent. This symmetry implies that the causal

structures are non-identifiable through likelihood maximization alone, as both models achieve the same global maximum likelihood.

3.2 BAYESIAN APPROACH

We adopt a Bayesian framework to perform causal structure inference. While S^1 and S^2 remain observationally non-identifiable, we characterize the asymptotic behavior of the posterior distribution and discuss the impact of the prior distributions.

We define a hierarchical model over the set of causal structures by first assigning a uniform prior distribution over the model space, $\pi(S) = \frac{1}{3}$, for $S \in \mathcal{S}$. Conditional on the structure S, we assign a prior density $\pi(\theta \mid S)$ to the associated parameter vector θ. For the purpose of this general formulation, we leave the functional form of $\pi(\theta \mid S)$ unspecified for now.

Assuming access to $\mathcal{D}_{\text{obs}} = (X_i)_{1 \leq i \leq n}$, the posterior over the structures is $S \in \mathcal{S}$

$$\pi(S \mid \mathcal{D}_{\text{obs}}) = \frac{\int_{\theta \in \Theta} f(\mathcal{D}_{\text{obs}} \mid S, \theta) \pi(\theta \mid S) d\theta}{\sum_{j=1}^{3} \int_{\theta \in \Theta} f(\mathcal{D}_{\text{obs}} \mid S^j, \theta) \pi(\theta \mid S^j) d\theta}. \tag{3}$$

The marginal likelihood, $\int_{\theta \in \Theta} f(\mathcal{D}_{\text{obs}} \mid S, \theta) \pi(\theta \mid S) d\theta$, is inherently sensitive to the choice of prior $\pi(\theta \mid S)$. However, for sufficiently regular priors, the integral is well-approximated via Laplace's method as $n \to \infty$. Due to space constraints, we defer the formal regularity conditions to the Appendix in Lemma A.3 where we verify that our setting satisfies the Laplace regularity requirements described by [Kass et al., 1990].

Theorem 3.1: Let $\mathcal{D}_{\text{obs}} = (X_i)_{i=1}^n$ be an i.i.d. sample generated from (1). Assume the priors $\pi(\theta \mid S)$ are four times differentiable for all $S \in \mathcal{S}$. Under the true generating model $\mathcal{M}^1 = \{S^{1*}, \theta^*\}$, where $X \sim P^X_{m(S^{1*}, \theta^*)}$ with $\theta^* = [w^*, \tau_1^{*2}, \tau_2^{*2}]^\intercal$ and $w^* \neq 0$, the posterior probability satisfies, as $n \to \infty$,

$$\pi(S^{1*} \mid \mathcal{D}_{\text{obs}}) \xrightarrow{a.s.} \frac{1}{1 + \frac{\tau_2^{*2}}{w^{*2}\tau_2^{*2} + \tau_1^{*2}} \frac{\pi(\theta^* \mid S^2)}{\pi(\theta^* \mid S^{1*})}}. \tag{4}$$

Moreover, we note that

$$\frac{\tau_2^{*2}}{w^{*2}\tau_2^{*2} + \tau_1^{*2}} \cdot \frac{\pi(\theta^* \mid S^2)}{\pi(\theta^* \mid S^1)} = \frac{\gamma^{-1} \# \pi(\theta^* \mid S^2)}{\pi(\theta^* \mid S^1)}$$

with $\gamma^{-1} \# \pi(\cdot \mid S^2)$ being the push-forward measure of the alternative prior (for S^2) under the transformation defined in (2).

Remark 1: Assume the conditions in Theorem 3.1 hold. If θ^* are the true parameters under S^1, then $\gamma(\theta^*)$ are the corresponding parameters under S^2 that yield the same likelihood. If S^2 is considered the true generating mechanism, one can show similarly that as $n \to \infty$

$$\pi(S_2 \mid \mathcal{D}_{\text{obs}}) \xrightarrow{a.s.} \frac{1}{1 + \frac{w^{*2}\tau_2^{*2} + \tau_1^{*2}}{\tau_2^{*2}} \cdot \frac{\pi(\theta^* \mid S_1)}{\pi(\theta^* \mid S_2)}},$$

which implies that $\pi(S^1 \mid \mathcal{D}_{\text{obs}}) + \pi(S^2 \mid \mathcal{D}_{\text{obs}}) \xrightarrow{a.s.} 1$.

Remark 2: Although structures S^1 and S^2 are observationally non-identifiable, the posterior probability of structure S^1 does not necessarily converge to $1/2$. This might suggest the existence of a decision rule for detecting the true generating structure; for instance, one could select S^1 if its posterior probability exceeds $1/2$, and S^2 otherwise. At first glance, this appears to contradict the non-identifiability of the model. However, we note that the true generating parameters θ^* are unknown to the observer. Consequently, without knowledge of the specific parameter values, one cannot determine whether the threshold for the posterior ratio should be set above or below $1/2$ to correctly identify the true graph.

Remark 3: This result also shows that the posterior odds ratio converges to a constant value which is dependent only on the priors placed on the two structures S^1 and S^2. It also confirms the result of Dhir et al. [2024] which shows theoretically that if the posterior of two models are equal, then the selected priors must satisfy the following relation $\pi(\theta^* \mid S^{1*}) = \gamma^{-1} \# \pi(\theta^* \mid S^2)$. Moreover, we note that this ratio is also central to the derivation of the Bayesian Gaussian equivalent (BGe) score in the case of two nodes which was studied in Geiger and Heckerman [1998]. In fact, they show that if the global parameter independence (or the independent causal mechanism – ICM) assumption holds, i.e. the two priors factorise as follows $\pi(w, \tau_1^2, \tau_2^2 \mid S^1) = \pi(w, \tau_1^2 \mid S^1)\pi(\tau_2^2 \mid S_1)$ and $\pi(w, \tau_1^2, \tau_2^2 \mid S^2) = \pi(w, \tau_2^2 \mid S^2)\pi(\tau_1^2 \mid S_2)$, then the ratio is one

$$\frac{\gamma^{-1} \# \pi(\theta \mid S^2)}{\pi(\theta \mid S^1)} = 1$$

if and only if $[w, \tau_1^2, \tau_2^2]^\intercal$ follow a normal-Wishart prior under the two models S^1 and S^2. This case is further discussed in Section 5.2.

Next, we consider the scenario where the true generating mechanism is the independence model S^3. In this case, the following theorem demonstrates that the posterior probability of the true model concentrates to 1 asymptotically. Notably, however, this convergence is not almost sure (as observed in the connected cases), but is instead characterized by a rate of $O_p(n^{-1/2})$, where $O_p(\cdot)$ denotes boundedness in probability. This suggests a significant slow-down in concentration compared to the exponential rates typical of identifiable causal directions derived in Lungu et al. [2025]or in the next sections.

Theorem 3.2: Assume the conditions from Theorem 3.1. Under the true generating model $\mathcal{M}^3 = \{S^{3*}, \theta^*\}$, where

$X \sim P^X_{m(S^{3*},\theta^*)}$ with $\theta^* = [0, \tau_1^{*2}, \tau_2^{*2}]^\intercal$, the posterior probability satisfies, as $n \to \infty$,

$$\sqrt{n}(1 - \pi(S^{3*} \mid \mathcal{D}_{\text{obs}})) = O_p(1).$$

Remark 4: The above result indicates that the posterior recovers the true generating structure asymptotically. Furthermore, one can derive the asymptotic distribution of the log-posterior odds from our proof. For S^i with $i \in \{1, 2\}$, then

$$2 \log \left(\sqrt{n} \frac{\pi(S^i \mid \mathcal{D}_{\text{obs}})}{\pi(S^{3*} \mid \mathcal{D}_{\text{obs}})} \right) + \log \frac{\pi(\theta^* \mid S^i)}{\pi(\theta^* \mid S^{3*})} \xrightarrow{d} \chi^2_1, \quad (5)$$

where $\theta^* = [0, \tau_1^{*2}, \tau_2^{*2}]^\intercal$, $\xrightarrow{d}$ denotes convergence in distribution and χ^2_1 is a chi-squared distribution with one degree of freedom.

4 IDENTIFIABILITY VIA INTERVENTIONAL DATA

As established in the previous section, the posterior distribution over the model space fails to concentrate on the true generating structure when restricting observations to the observational regime, owing to the Markov equivalence of structures S^1 and S^2. To resolve this non-identifiability problem, we extend the data setting to include interventional data.

We assume access to a heterogeneous dataset consisting of an observational component, $\mathcal{D}_{\text{obs}} := (X_i)_{i=1}^n$ with $X \overset{\text{i.i.d.}}{\sim} P^X_{m(S,\theta)}$, and an interventional component, $\mathcal{D}_{\text{int}} := (Y_j)_{j=1}^m$. To formalize the latter, we use the $do(\cdot)$ operator introduced in Pearl [2009]. An intervention $do(X(j) = y)$ corresponds to a structural modification of the data-generating process wherein the value of node j is fixed to y, effectively removing all incoming causal edges to j while preserving the structural equations of its descendants.

We consider a fixed interventional design targeting the second variable. Let the intervention be defined as $do(X(2) = y)$ for a fixed $y \in \mathbb{R}$. The samples in the interventional dataset are denoted by $Y_i = [Y_i(1), y]^\intercal$ for $i = 1, \ldots, m$. The marginal distribution of the non-intervened variable $Y(1)$ allows us to discriminate between the two connected causal structures, S^1 and S^2.

If the true generating mechanism is S^1 (implying the causal direction $X(2) \to X(1)$), the intervention on $X(2)$ propagates to $X(1)$. Specifically, for parameters $\theta^* = [w^*, \tau_1^{*2}, \tau_2^{*2}]^\intercal$

$$Y(1) \mid \{do(X(2) = y), \theta^*, S^1\} \sim N(w^* y, \tau_1^{*2}). \quad (6)$$

If the true mechanism is S^2 (where $X(1) \to X(2)$) or the independence model S^3, the intervention on $X(2)$ breaks the causal dependency (in the case of S^2) or no dependency exists (in S^3). In both cases, the distribution of $Y(1)$ remains governed by its marginal parameters

$$Y(1) \mid \{do(X(2) = y), \theta^*, S^i\} \sim N(0, \tau_1^{*2}), \quad (7)$$

for $i \in \{2, 3\}$. This distributional difference renders S^1 and S^2 distinguishable. We denote the law of $Y(1) \mid \{do(X(2) = y), \theta, S\}$ by $P^Y_{m(S,\theta)}$.

To analyze the asymptotic behavior of the posterior $\pi(S \mid \mathcal{D}_{\text{mix}})$ with $\mathcal{D}_{\text{mix}} := \mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{int}}$, we characterize the data regime by the fraction of observational samples. To formalize this setting, let N be the total number of samples, with n_N observational and m_N interventional samples. Assuming both sequences, $\{n_N\}_N$ and $\{m_N\}_N$, increase with N, we define the asymptotic observational ratio $\eta \in (0, 1)$ as

$$\eta_N := \frac{n_N}{N}, \quad \text{where} \quad \lim_{N \to \infty} \eta_N = \eta. \quad (8)$$

The following analysis investigates the concentration rate of the posterior as a function of the mixing parameter η and the true generating mechanism. First, we consider the connected cases, i.e. S^1 and S^2, and show that in this setting, the posterior concentrates almost surely to 1 at an exponential rate.

Theorem 4.1: Let $\mathcal{D}_{\text{obs}} = (X_i)_{i=1}^{n_N}$ be an i.i.d. observational sample generated from (1), and $\mathcal{D}_{\text{int}} = (Y_i)_{i=1}^{m_N}$ be an i.i.d. interventional sample generated under the intervention $do(X(2) = y)$. Assume the priors $\pi(\theta \mid S)$ are four times differentiable for all $S \in \mathcal{S}$. Under the true generating model $\mathcal{M}^1 = \{S^{1*}, \theta^*\}$, where $X \sim P^X_{m(S^{1*},\theta^*)}$ and $Y \sim P^Y_{m(S^{1*},\theta^*)}$ with $\theta^* = [w^*, \tau_1^{*2}, \tau_2^{*2}]^\intercal$, the posterior probability satisfies, as $N \to \infty$,

$$\frac{1}{N} \log \left(\frac{1}{\pi(S^{1*} \mid \mathcal{D}_{\text{mix}})} - 1 \right) \xrightarrow{a.s.} -D_{12}(\eta),$$

where $D_{12}(\eta)$ is defined as

$$D_{12}(\eta) := \frac{1}{2} \log \left(\frac{\eta \sigma^2_{1,X} + (1 - \eta)\sigma^2_{1,Y}}{(\sigma^2_{1,X})^\eta (\tau_1^{*2})^{\bar{\eta}}} \right), \quad (9)$$

with $\sigma^2_{1,X} := w^{*2}\tau_2^{*2} + \tau_1^{*2}$ and $\sigma^2_{1,Y} := w^{*2} y^2 + \tau_1^{*2}$.

Conversely, if the true generating model is $\mathcal{M}^2 = \{\theta^*, S^{2*}\}$, then the posterior concentrates to 1, as $N \to \infty$

$$\frac{1}{N} \log \left(\frac{1}{\pi(S^{2*} \mid \mathcal{D}_{\text{mix}})} - 1 \right) \xrightarrow{a.s.} -D_{21}(\eta),$$

with

$$D_{21}(\eta) = \frac{1}{2} \log \left(1 - \frac{\eta^2 w^{*2} \tau_1^{*2}}{\eta \sigma^2_{2,Y} + \bar{\eta} y^2} \right) \left(\frac{\sigma^2_{2,Y}}{\tau_2^{*2}} \right)^{\frac{\eta}{2}}, \quad (10)$$

with $\sigma^2_{2,Y} = w^{*2}\tau_1^{*2} + \tau_2^{*2}$.

Lemmas B.1 and B.2 prove that $D_{12}(\eta)$ and $D_{21}(\eta)$ are concave as a function of η and provide conditions for the optimal mixing parameter η. The implications of the results are further discussed in the the next section.

Analogous to Theorem 3.2, the posterior achieves a stochastic polynomial concentration rate of $O_p(N^{-1/2})$. We formalize this below.

Theorem 4.2: Assume the conditions in Theorem E.1 hold. If the true generating model is $\mathcal{M}^3 = \{\theta^*, S^{3*}\}$ where $X \sim P^X_{m(S^{3*},\theta^*)}$ and $Y \sim P^Y_{m(S^{3*},\theta^*)}$ with $\theta^* := [0, \tau_1^{*2}, \tau_2^{*2}]^\intercal$, the posterior probability satisfies, as $N \to \infty$,

$$\sqrt{N}(1 - \pi(S^{3*} \mid \mathcal{D}_{\text{mix}})) = O_p(1).$$

Remark 5: We believe that for $d > 2$, two mechanisms govern the behavior of the posterior:

1. **Interventional Markov Equivalence Classes (I-MECs):** Following Hauser and Bühlmann [2012], the size of the equivalence class is reduced through interventions on specific nodes. The identifiability structure is thus constrained by the specific I-MEC determined by the intervention set.
2. **Graph Maximality:** As discussed in Lungu et al. [2025], a graph is *maximal* if no edges can be added without violating the acyclicity constraint.

For better understanding, let I denote the resulting I-MEC and G^* the true generating structure. We believe the following qualitative behaviors occur in all dimensions $d \geq 2$:

- **Case 1: G^* is maximal.** In this scenario, all graphs in I are also maximal. We expect the posterior to concentrate exponentially fast on I, with the exponent being a weighted combination of relative entropies, analogous to Theorem 4.1. However, if I is not a singleton, the posterior fails to concentrate on G^* specifically; instead, the posterior ratios between G^* and the remaining graphs in I would converge to constants, similarly to the result in Theorem 3.1.
- **Case 2: G^* is non-maximal.** Here, we expect the posterior to concentrate on I at a polynomial stochastic rate of order $1/\sqrt{n}$. This comes from the fact that for any non-maximal graph in I, there exists at least one larger graph containing an additional edge (and thus an extra parameter). While the likelihoods may be identical, the posterior will favor the more parsimonious model due to the "Bayesian Occam's razor" phenomenon.

Formally proving these conjectures for $d > 2$ would require a separate, technical and rather lengthy theoretical analysis.

5 INTERPRETATION AND NUMERICAL STUDY

In this section we investigate the behavior of the asymptotic convergence rates as a function of the mixing parameter η. Furthermore, we illustrate the theoretical findings through an empirical study on a 2-node causal model with a prior structure closely related to the one used for the Bayesian Gaussian equivalent (BGe) models.

5.1 ANALYSIS OF CONVERGENCE RATES

Theorem E.1 established that if the true generating mechanism is S^1, the posterior probability of the true model concentrates to 1 at an exponential rate determined by D_{12}. Specifically, the error term behaves as $\pi(S^{1*} \mid \mathcal{D}_{\text{mix}}) \approx 1 - e^{-ND_{12}}$, where $N = n_N + m_N$ is the total sample size.

We first examine the rate $D_{12}(\eta)$ as a function of the sample size mixing parameter η. Lemma B.1 shows that the exponent is strictly concave and the condition (35) reveals an interesting phenomenon. If $y^2 > \tau_2^{*2}$, the inequality does not hold (the LHS becomes negative) and the exponent is strictly decreasing, suggesting that asymptotically, observational data does not contribute to the optimal rate.

In the case where $y^2 < \tau_2^{*2}$, the behavior depends on the Signal-to-Noise Ratio (SNR). For low SNR (small $|w^*|$), the linear and logarithmic terms are approximately equal; and hence D_{12} remains strictly decreasing. Similarly, for large SNR (large $|w^*|$), the inequality is violated because the logarithmic term (RHS) grows unbounded while the linear term (LHS) saturates. Therefore, optimal mixing occurs only for intermediate SNR with $y^2 < \tau_2^{*2}$. In this case, the posterior uses the observational data to better *learn* w and hence a combination of interventional and observational data yields a faster concentration rate.

Figure 1 (left) illustrates the behavior of D_{12}. This confirms the insight that in the limit $\eta \to 1$ (purely observational data), the problem becomes non-identifiable, and the exponential driving force disappears. Moreover, this figure shows that two choices of the hyper-parameters which give different behaviors of the rate speed as a function of η; for the blue line, mixing can improve the speed, while for the red dotted line, the observational samples affect the concentration speed.

Next, we analyze the rate $D_{21}(\eta)$ which governs convergence when the true model is the reverse causal structure S^2. In Lemma B.2, we show that the exponent is strictly concave in η and always achieves a maximum inside $(0, 1)$. Unlike $D_{12}(\eta)$, Figure 1(right) demonstrates that $D_{21}(\eta)$ vanishes in both extremes: as $\eta \to 1$ and as $\eta \to 0$.

- As $\eta \to 1$ (observational limit), S^1 and S^2 become observationally equivalent, rendering $D_{21}(\eta) \to 0$.

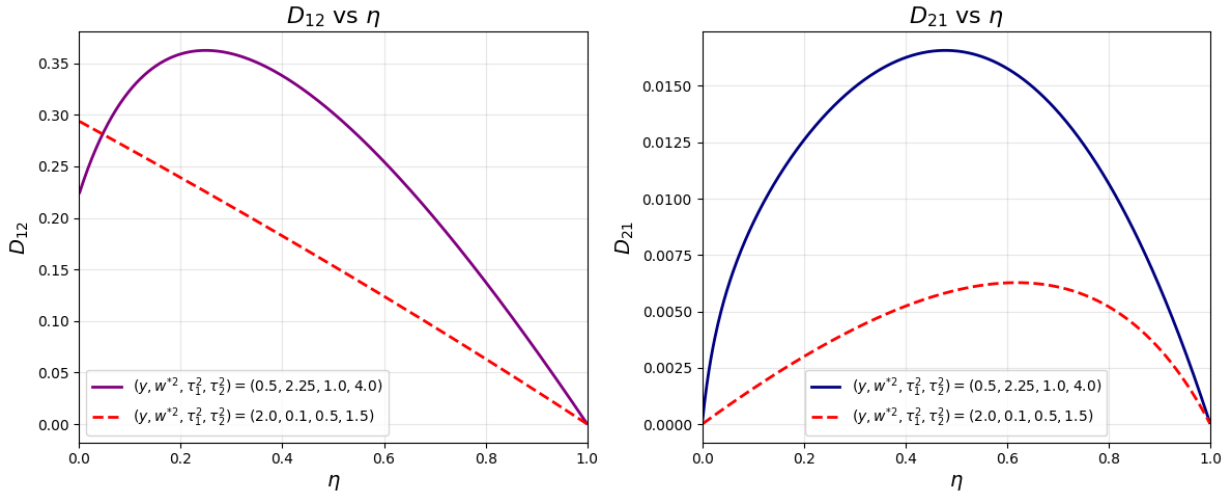

Figure 1: Comparison of the rates D_{12} (left) and D_{21} (right) as a function of η for different choices of $[w^*, \tau_1^{*2}, \tau_2^{*2}]^\intercal$.

- As $\eta \to 0$ (interventional limit), the intervention $do(Y(2) = y)$ renders $Y(1)$ independent of y in both S^2 and S^3. Since both models predict $Y(1) \sim N(0, \tau_1^{*2})$, they become indistinguishable, and the rate again drops to zero.

This highlights an important trade-off: observational data is required to distinguish S^2 from the independence model S^3, while interventional data is required to distinguish S^2 from S^1.

5.2 APPLICATION TO BGE - TYPE PRIORS

We consider the following hierarchical model

$$S \sim \text{Unif}\{S^1, S^2, S^3\}; \quad \tau_j^2 \mid S^i \sim \text{IG}(\alpha_{i,j}, \beta), \\ w \mid (S^i, \tau_i^2) \sim N(0, \lambda\tau_i^2)^{\mathbb{I}(i \neq 3)} \tag{11}$$

where $\mathbb{I}$ denotes the indicator function, $\beta, \lambda, \alpha_{i,j} \in (0, \infty)$ for all $i \in \{1, 2, 3\}$, $j \in \{1, 2\}$ and $\text{IG}(\alpha, \beta)$ is the Inverse-Gamma distribution with parameters α and β. We first investigate the case where only observational data is available.

5.2.1 Observational Data

For this choice of priors, the posterior distribution over the model structures can be calculated in closed form. The exact formula is given in Appendix H.1.

Remark 6: Under the specific hyper-parameter choice $\alpha_1 = \alpha_4 = \alpha$, $\alpha_2 = \alpha_3 = \alpha - \frac{1}{2}$, and $2\beta = \lambda$, we have $S_1^{X\lambda} = S_1^{X\beta}$ and $S_2^{X\lambda} = S_2^{X\beta}$, where these quantities are defined in (68). Consequently, the posterior distribution, $\left[\pi(S^1 \mid \mathcal{D}_{\text{obs}}), \pi(S^2 \mid \mathcal{D}_{\text{obs}}), \pi(S^3 \mid \mathcal{D}_{\text{obs}})\right]^\intercal$ simplifies to

$$\frac{1}{c}\left[\Delta^{-\nu/2}, \Delta^{-\nu/2}, (S_1^{X\beta})^{\frac{1-\nu}{2}}(S_2^{X\beta})^{-\frac{\nu}{2}}\right]^\intercal, \tag{12}$$

where $\Delta = S_1^{X\beta} S_2^{X\beta} - (S_{12}^X)^2$. where c is a normalizing constant and $\nu = 2\alpha + n$. This choice of priors places equal mass on models within the same Markov equivalence class. Geiger and Heckerman [1998] demonstrates that these are the unique priors satisfying this property.

Empirically, Figure 6 in the Appendix demonstrates that when S^1 is the true mechanism, the posterior odds between the non-identifiable models S^1 and S^2 do not concentrate at 1, but rather at a constant determined by the priors. Furthermore, we note that with this choice of hyper-parameters the Bayes factor concentrates at a value less than 1, suggesting that the Bayesian approach favors the incorrect model S^2 in this scenario.

Next, to empirically verify the convergence behavior of the independence model S^3, we plot the posterior probability of the true structure in Figure 2. We observe that the posterior mass concentrates at 1 at a rate of $O_p(n^{-1/2})$, as established in Theorem 3.2. The minor fluctuations observed at larger n are consistent with this stochastic bound; unlike the connected cases which exhibit almost sure concentration, the independence case is characterized by this slower, probabilistic rate.

To verify the distributional convergence of the log-posterior ratio to a χ_1^2 random variable (see Remark 4), we evaluate the *augmented posterior odds* defined in (5). The experimental results presented in Figure 3 empirically confirm that these statistics converge in distribution to χ_1^2, aligning with our theoretical predictions.

5.2.2 Observational and Interventional Data

We now add m interventional datapoints which come from the distribution of $\cdot \mid \{\text{do}(Y(2) = y), S\}$. Again, the posterior over the structures has a closed form solution and the

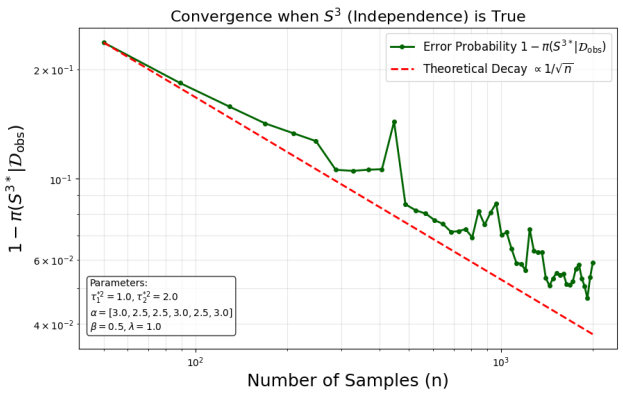

Figure 2: Concentration speed (in log scale) of the posterior $\pi(S^{3*} \mid \mathcal{D}_{\text{obs}})$ to 1 when the true generating model is S^3.

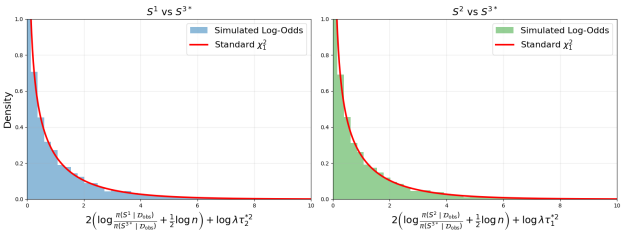

Figure 3: Augmented posterior odds showing convergence to a χ^2_1 distribution when the true generating model is S^{3*}.

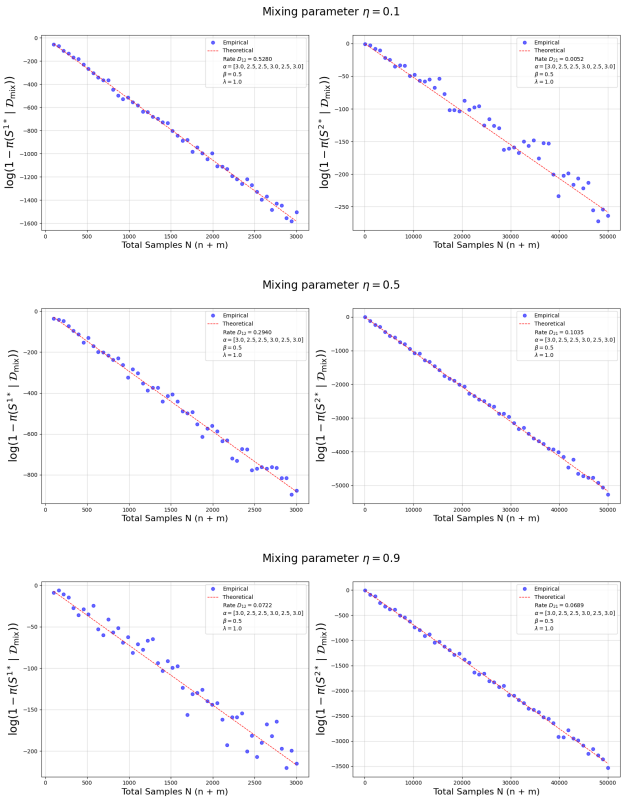

Figure 4: Concentration speed of the posterior to 1 when the true model is S^1 (left) or S^2 (right).

exact calculations can be found in Theorem H.2

Figure 4 illustrates the posterior's exponential concentration to 1 when the data is generated by either model S^1 (left) or S^2 (right). For all experiments we used the BGe priors, i.e. $\alpha_1 = \alpha_4 = 3.0$ and $\alpha_2 = \alpha_3 = 2.5$ with $2\beta = \lambda = 1.0$, which places equal mass on models S^1 and S^2 when only observational samples are available. As expected, the closeness to the theoretical prediction varies with the mixing parameter η. For structure S^1, this correspondence degrades as η increases from 0.1 to 0.9. This behavior is strongly connected to the results in Section 5.1, where the exponent D_{12} approaches zero as the proportion of observational samples increases, effectively nullifying the exponential decrease. One can see that for the hyperparameters used, the theoretical exponent decreases from 0.53 for $\eta = 0.1$ to 0.30 for $\eta = 0.5$ and finally to 0.07 for $\eta = 0.9$. In contrast, for structure S^2, the empirical results track the predicted exponential decay most accurately at intermediate values of η, with performance degrading at the extremes, again in line with the results of Section 5.1.

Finally, we show the posterior concentration speed to 1 for S^3 in Figure 5. As previously observed, the concentration speed follows the $\sqrt{N}$ line, with the fluctuations around given by the stochastic bound.

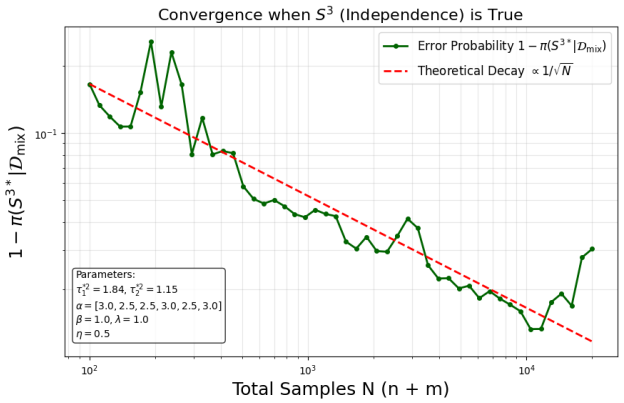

Figure 5: Concentration speed (in log scale) of the posterior $\pi(S^{3*} \mid \mathcal{D}_{\text{mix}})$ to 1 when the true generating model is S^3.

6 CONCLUSIONS

We presented the first formal characterization of Bayesian posterior concentration rates for bivariate Gaussian Linear SEMs. With purely observational data, the posterior probability of a connected true structure fails to concentrate to 1, instead converging to a fixed constant determined by the prior specification. For the independence model, the posterior identifies the structure at a stochastic rate of $O_p(n^{-1/2})$.

Our analysis of the mixed-data regime revealed a sharp concentration dichotomy: the introduction of interventional samples triggers a phase transition from polynomial to exponentially fast convergence for connected graphs. We derived explicit formulae for these concentration exponents and identified conditions on the observational-to-total sample ratio, η, where a blend of data sources gives an increase in the concentration speed. We validated our theoretical results in BGe-type priors.

While we focused on the linear case, the identified concentration dichotomy suggests that the interplay between observational calibration and interventional signal is a fundamental feature of causal discovery. Extending these asymptotic bounds to non-linear mechanisms and high-dimensional settings remains a promising avenue for future research.

References

J. Acharya, S. Bhadane, A. Bhattacharyya, S. Kandasamy, and Z. Sun. Sample complexity of distinguishing cause from effect. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10487–10504. PMLR, 25–27 Apr 2023.

Y. Annadani, J. Rothfuss, A. Lacoste, N. Scherrer, A. Goyal, Y. Bengio, and S. Bauer. Variational causal networks: Approximate Bayesian inference over causal structures. *arXiv e-prints*, 2106.07635 [cs.LG], June 2021.

Y. Annadani, N. Pawlowski, J. Jennings, S. Bauer, C. Zhang, and W. Gong. BayesDAG: Gradient-based posterior inference for causal discovery. *NeurIPS*, 36:1738–1763, 2023.

X. Cao, K. Khare, and M. Ghosh. Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Annals of Statistics*, 47(1):319–348, 2019.

D.M. Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2002.

C. Cundy, A. Grover, and S. Ermon. BCD Nets: Scalable variational approaches for Bayesian causal discovery. *NeurIPS*, 34:7095–7110, 2021.

T. Deleu, A. Góis, C. Emezue, M. Rankawat, S. Lacoste-Julien, S. Bauer, and Y. Bengio. Bayesian structure learning with generative flow networks. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180, pages 518–528, 2022.

A. Dhir, S. Power, and M. van der Wilk. Bivariate causal discovery using bayesian model selection, 2024. URL https://arxiv.org/abs/2306.02931.

A. Dhir, M. Ashman, J. Requeima, and M. van der Wilk. A meta-learning approach to bayesian causal discovery, 2025. URL https://arxiv.org/abs/2412.16577.

A. Dhir, C. Diaconu, V. Lungu, J. Requeima, R. E. Turner, and M. van der Wilk. Estimating interventional distributions with uncertain causal graphs through meta-learning, 2026. URL https://arxiv.org/abs/2507.05526.

D. Eaton and K. Murphy. Bayesian structure learning using dynamic programming and MCMC. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 101–108, 2007.

N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003.

D. Geiger and D. Heckerman. A characterization of the bivariate wishart distribution. *Probability and Mathematical Statistics*, 18(1):119–131, 1998.

D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412 – 1440, 2002. doi: 10.1214/aos/1035844981. URL https://doi.org/10.1214/aos/1035844981.

D. Geiger and D. Heckerman. Learning gaussian networks, 2021. URL https://arxiv.org/abs/1302.6808.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs, 2012. URL https://arxiv.org/abs/1104.2808.

D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In *Computation, Causation, and Discovery*, pages 141–166. AAAI Press, 2006.

P.O. Hoyer and A. Hyttinen. Bayesian discovery of linear acyclic causal models, 2012. URL https://arxiv.org/abs/1205.2641.

F. Jamshidi, S. Akbari, and N. Kiyavash. Sample complexity of nonparametric closeness testing for continuous distributions and its application to causal discovery with hidden confounding, 2025. URL https://arxiv.org/abs/2503.07475.

R. E. Kass, L. Tierney, and J. B. Kadane. The validity of posterior expansions based on Laplace's method. In S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, editors, *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor*

of George A. Barnard, pages 473–488. North-Holland, 1990. URL https://api.semanticscholar.org/CorpusID:116768493.

G.L. Lozano, J.I. Bravo, M.F. Garavito Diago, H.B. Park, A. Hurley, S.N. Peterson, E.V. Stabb, J.M. Crawford, N.A. Broderick, and J. Handelsman. Introducing THOR, a model microbiome for genetic dissection of community behavior. *mBio*, 10(2):e02846–18, 2019. doi: 10.1128/mBio.02846-18. URL https://journals.asm.org/doi/abs/10.1128/mBio.02846-18.

V. Lungu, J. Shaska, I. Kontoyiannis, and U. Mitra. Bayesian causal discovery: Posterior concentration and optimal detection, 2025.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

J. Pearl. *Causality*. Cambridge University Press, 2009.

J. Peters and P. Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

P. Spirtes and C.N. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, Cambridge, MA, 2000.

E. Van den Broeck, K. Poels, and M. Walrave. An experimental study on the effect of ad placement, product involvement and motives on Facebook ad avoidance. *Telematics and Informatics*, 35(2):470–479, 2018. ISSN 0736-5853. doi: https://doi.org/10.1016/j.tele.2018.01.006. URL https://www.sciencedirect.com/science/article/pii/S0736585317307645.

X. Wu, S. Peng, J. Li, J. Zhang, Q. Sun, W. Li, Q. Qian, Y. Liu, and Y. Guo. Causal inference in the medical domain: A survey. *Applied Intelligence*, 54(6):4911–4934, 2024.

X. Zheng, B. Aragam, P.K. Ravikumar, and E.P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *NeurIPS*, 31:9492–9503, 2018.

The relative value of interventional and observational samples in Bayesian Causal Linear Gaussian Models (Supplementary Material)

Valentinian Lungu[1] **Anish Dhir**[2] **Mark van der Wilk**[3] **Ioannis Kontoyiannis**[1]

[1]Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge
[2]Imperial College London
[3]University of Oxford

A AUXILIARLY LEMMAS

A.1 PROOF OF LEMMA 3.1

Rearraging (1), we obtain $X = (I - A)^{-1}\epsilon$, and we note that the inverse of $(I - A)$ always exists. Then, under structure S^1, we have

$$\{X \mid \theta^1, S^1\} \sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}w^2\tau_2^2 + \tau_1^2 & w\tau_2^2\\ w\tau_2^2 & \tau_2^2\end{bmatrix}\right), \text{ for any } \theta^1 := [w, \tau_1^2, \tau_2^2]^\intercal \in \Theta.$$

It is easy to check that if the parameters in the second structure, S^2, are chosen as follows $\theta^2 = \left[\frac{w\tau_2^2}{w^2\tau_2^2+\tau_1^2}, w^2\tau_2^2 + \tau_1^2, \frac{\tau_1^2\tau_2^2}{w^2\tau_2^2+\tau_1^2}\right]^\intercal \in \Theta$, then the distribution of X is

$$\{X \mid S^2, \theta_2\} \sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}w^2\tau_2^2 + \tau_1^2 & w\tau_2^2\\ w\tau_2^2 & \tau_2^2\end{bmatrix}\right),$$

and therefore, the two structures are non-identifiable. □

Lemma A.1. *If $\mathcal{D}_{obs} := (X_i)_{1\le i\le n}$ are n i.i.d. observations generated by* (1)*, then the maximum likelihood estimators for each structure are*

$$\hat{\theta}^1 := [\hat{\theta} \mid S^1] = \left[\frac{S_{12}^X}{S_2^X}, \frac{S_1^X S_2^X - (S_{12}^X)^2}{nS_2^X}, \frac{S_2^X}{n}\right]^\intercal, \tag{13}$$

$$\hat{\theta}^2 := [\hat{\theta} \mid S^2] = \left[\frac{S_{12}^X}{S_1^X}, \frac{S_1^X}{n}, \frac{S_1^X S_2^X - (S_{12}^X)^2}{nS_1^X}\right]^\intercal, \tag{14}$$

$$\hat{\theta}^3 := [\hat{\theta} \mid S^3] = \left[0, \frac{S_1^X}{n}, \frac{S_2^X}{n}\right]^\intercal, \tag{15}$$

where

$$S_{12}^X := \sum_{i=1}^n X_i(1)X_i(2), \quad S_1^X := \sum_{i=1}^n X_i(1)^2, \quad S_2^X := \sum_{i=1}^n X_i(2)^2. \tag{16}$$

Proof. Let $\ell_n(\theta \mid S^1) := \sum_{i=1}^n \log f(X_i \mid S^1, \theta)$ denote the log-likelihood function for structure S^1. By factorizing the joint density as $f(x \mid S^1, \theta) = f(x(1), x(2) \mid S^1, \theta) = f(x(1) \mid x(2), w, \tau_1^2) f(x(2) \mid \tau_2^2)$, and letting $\phi_v(\cdot)$ denote the

density of a centered Gaussian with variance v, we obtain

$$\begin{aligned}\ell_n(\theta \mid S^1) &= \sum_{i=1}^{n} \Big[\log \phi_{\tau_1^2}(X_i(1) - wX_i(2)) + \log \phi_{\tau_2^2}(X_i(2))\Big] \\ &= -\frac{n}{2}\log(2\pi\tau_1^2) - \sum_{i=1}^{n} \frac{(X_i(1) - wX_i(2))^2}{2\tau_1^2} - \frac{n}{2}\log(2\pi\tau_2^2) - \sum_{i=1}^{n} \frac{X_i(2)^2}{2\tau_2^2} \\ &= -n\log(2\pi) - \frac{n}{2}\log(\tau_1^2\tau_2^2) - \frac{S_1^X - 2wS_{12}^X + w^2 S_2^X}{2\tau_1^2} - \frac{S_2^X}{2\tau_2^2}.\end{aligned}$$

The maximum likelihood estimates are obtained by solving the first-order conditions $\nabla_\theta \ell_n(\theta \mid S^1) = 0$. Differentiating with respect to w, we find:

$$\frac{\partial \ell_n(\theta \mid S^1)}{\partial w} = -\frac{1}{2\tau_1^2}(-2S_{12}^X + 2wS_2^X) = 0 \implies \hat{w} = \frac{S_{12}^X}{S_2^X}.$$

Substituting $\hat{w}$ and differentiating with respect to the variance parameters yields the standard variance estimators. It is easy to compute the second derivatives and show that the optimum is in fact unique. The MLEs for models S^2 and S^3 follow analogously. □

Lemma A.2. *Let $\mathcal{D}_{obs} := (X_i)_{i=1}^n$ be n i.i.d. observations generated by* (1)*. Then, as $n \to \infty$*

$$\begin{aligned}S_1^X &= n\mathbb{E}[X(1)^2] + O\left(\sqrt{n\log\log n}\right) \text{ a.s.} \\ S_2^X &= n\mathbb{E}[X(2)^2] + O\left(\sqrt{n\log\log n}\right) \text{ a.s.}, \\ \frac{S_1^X S_2^X - (S_{12}^X)^2}{S_2^X} &= n\frac{\mathbb{E}[X(1)^2]\mathbb{E}[X(2)^2] - (\mathbb{E}[X(1)X(2)])^2}{\mathbb{E}[X(2)^2]} + O\left(\sqrt{n\log\log n}\right) \text{ a.s.}\end{aligned}$$

with S_1^X, S_2^X and S_{12}^X defined in (16)*.*

Proof. Rearranging (1), we have $X = (I - A)^{-1}\epsilon$. Since $(I - A)$ is invertible, the covariance matrix of X is given by $\text{Var}(X) = (I - A)^{-1}\Sigma(I - A)^{-1}$. Since the second and fourth moments are finite: $\mathbb{E}[X(i)^2] < \infty$ and $\mathbb{E}[X(i)^4] < \infty$ for $i \in \{1, 2\}$. Then, by the Cauchy-Schwarz inequality, $\mathbb{E}[|X(1)X(2)|] \leq \sqrt{\mathbb{E}[X(1)^2]\mathbb{E}[X(2)^2]} < \infty$.

Applying the Law of the Iterated Logarithm to $S_1^X = \sum_{i=1}^n X_i(1)^2$, $S_2^X = \sum_{i=1}^n X_i(2)^2$, and $S_{12}^X = \sum_{i=1}^n X_i(1)X_i(2)$, we obtain the following almost sure results

$$\begin{aligned}S_1^X &= n\mathbb{E}[X(1)^2] + O\left(\sqrt{n\log\log n}\right) \text{ a.s.} \\ S_2^X &= n\mathbb{E}[X(2)^2] + O\left(\sqrt{n\log\log n}\right) \text{ a.s.}, \\ S_{12}^X &= n\mathbb{E}[X(1)X(2)] + O\left(\sqrt{n\log\log n}\right) \text{ a.s.}\end{aligned}$$

We define the function $g(x, y, z) = \frac{xy - z^2}{y}$. Note that since X follows a non-degenerate Gaussian distribution, $\mathbb{E}[X(1)^2] > 0$ and $\mathbb{E}[X(2)^2] > 0$. Since the function is continuously differentiable, using the above almost sure expansions we have

$$\begin{aligned}\frac{S_1^X S_2^X - (S_{12}^X)^2}{nS_2^X} &= g\left(\frac{S_1^X}{n}, \frac{S_2^X}{n}, \frac{S_{12}^X}{n}\right) \\ &= \frac{\mathbb{E}[X(1)^2]\mathbb{E}[X(2)^2] - (\mathbb{E}[X(1)X(2)])^2}{\mathbb{E}[X(2)^2]} + O\left(\sqrt{\frac{\log\log n}{n}}\right).\end{aligned}$$

□

Lemma A.3. *Let $B_\delta(\theta)$ denote the open ball of radius δ centered at $\theta \in \Theta := \mathbb{R} \setminus \{0\} \times \mathbb{R}_+^2$. For any $S \in \mathcal{S}$ and $\theta \in \Theta$, let $\mathcal{D}_{obs} = (X_i)_{1 \le i \le n}$ be n i.i.d. observations generated by* (1) *with law $P^X_{m(S,\theta)}$ and density $f(x \mid S, \theta)$. The following four conditions hold*

(i) for all $x_1^n = (X_i)_{i=1}^n$ with $x_i \in \mathbb{R}^2$ and $\theta \in \Theta$, $\prod_{i=1}^n f(x_i \mid S, \theta) > 0$. Moreover, for all x_1^n, $\ell_n(\theta \mid S) = \sum_{i=1}^n \log f(x_i \mid S, \theta)$ is six times continuously differentiable in θ;

(ii) for any $\theta_0 \in \Theta$ there exist $\epsilon > 0$ and $M < \infty$ such that $B_\epsilon(\theta_0) \subseteq \Theta$ and for $1 \le j_1, \ldots, j_d \le 3$ with $d \le 6$,

$$\limsup_{n\to\infty} \sup\{n^{-1} |\partial_{j_1} \ldots \partial_{j_d} \ell_n(\theta \mid S)| : \theta \in B_\epsilon(\theta_0)\} < M,$$

with $P^X_{m(S,\theta_0)}$-probability one;

(iii) for any $\theta_0 \in \Theta$ and some $\epsilon > 0$,

$$\limsup_{n\to\infty} \sup\{\det(n^{-1} \nabla^2_\theta \ell_n(\theta \mid S)) : \theta \in B_\epsilon(\theta_0)\} < 0$$

with $P^X_{m(S,\theta_0)}$-probability one;

(iv) for any $\theta_0 \in \Theta$ and any $\delta > 0$,

$$\limsup_{n\to\infty} \sup\{n^{-1}[\ell_n(\theta \mid S) - \ell_n(\theta_0 \mid S)] : \theta \in \Theta \setminus B_\delta(\theta_0)\} < 0$$

with $P^X_{m(S,\theta_0)}$-probability one.

Hence, the models are Laplace regular as per Kass et al. [1990].

Proof. We check the four conditions presented in Kass et al. [1990] which allow us to approximate the marginal likelihoods using Laplace's method. Without loss of generality we assume $S = S^1$. The other remaining cases follow along the same lines. Under S^1, the log-likelihood $\ell_n(\theta \mid S^1)$ as a function of the MLE $\hat{\theta}^1$ is

$$\ell_n(\theta \mid S^1) = \log = -n \log 2\pi - \frac{n}{2} \log \tau_1^2 \tau_2^2 - \frac{n}{2} \cdot \frac{(w - \hat{w})^2 \hat{\tau}_2^2}{\tau_1^2} - \frac{n}{2} \cdot \frac{\hat{\tau}_1^2}{\tau_1^2} - \frac{n}{2} \cdot \frac{\hat{\tau}_2^2}{\tau_2^2}. \tag{17}$$

(i) We note that $\ell_n(\theta \mid S^1)$ is 6-times continuously differentiable. Moreover, since the observations are assumed to be independent, the joint density function for x_1^n is $\prod_{i=1}^n f(x_i \mid S, \theta)$ which is non-negative because each term in the product is strictly positive,

$$f(x_i \mid S^1, \theta) = \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left\{ -\frac{(x_i(1) - w x_i(2))^2}{2\tau_1^2} \right\} \frac{1}{\sqrt{2\pi\tau_2^2}} \exp\left\{ -\frac{x_i(2)^2}{2\tau_2^2} \right\} > 0.$$

(ii) The log-likelihood function $\ell_n(\theta \mid S_1)$ decomposes into a sum of terms that are either quadratic polynomials in the structural parameter w or rational functions of the variance parameters τ_1^2, τ_2^2.
Specifically, partial derivatives with respect to w of order $d \ge 3$ vanish and hence are bounded. Partial derivatives with respect to the variances involve terms of the form τ^{-k}, which are C^∞ smooth and bounded on any compact domain where the variances are bounded away from zero (i.e., $\tau_j^2 \ge \delta > 0$). Consequently, all mixed partial derivatives of order $d \ge 3$, when normalized by n^{-1}, are uniformly bounded in a neighborhood of the true parameter, satisfying the necessary condition.

(iii) The Hessian is

$$H(\theta) := H(w, \tau_1^2, \tau_2^2) = \begin{bmatrix} -\frac{n\hat{\tau}_2^2}{\tau_1^2} & \frac{n\hat{\tau}_2^2(w-\hat{w})}{(\tau_1^2)^2} & 0 \\ \frac{n\hat{\tau}_2^2(w-\hat{w})}{(\tau_1^2)^2} & \frac{n}{2(\tau_1^2)^2} - \frac{n}{(\tau_1^2)^3}\left[\hat{\tau}_1^2 + \hat{\tau}_2^2(w-\hat{w})^2\right] & 0 \\ 0 & 0 & \frac{n}{2(\tau_2^2)^2} - \frac{n\hat{\tau}_2^2}{(\tau_2^2)^3} \end{bmatrix} \tag{18}$$

with the determinant of the normalized Hessian being

$$h_n(\theta) := \det\left(\frac{1}{n}H(\theta)\right) = -\frac{\hat{\tau}_2^2}{4\tau_1^6\tau_2^4}\Big(1-\frac{2\hat{\tau}_2^2}{\tau_2^2}\Big)\Big(1-\frac{2\hat{\tau}_1^2}{\tau_1^2}\Big). \tag{19}$$

Let $\theta_0 := \big(w_0, \tau_{1,0}^2, \tau_{2,0}^2\big) \in \Theta$ be the true generating parameter and $d := \min\{\tau_{1,0}^2, \tau_{2,0}^2\} > 0$. Consider the closed ball $\bar{B}_{\epsilon_1}(\theta)$ with $\epsilon_1 = d/2$. Since both variances are bounded away from 0, $h_n(\theta)$ is continuous and well-defined on $\bar{B}_{\epsilon_1}(\theta_0)$. Furthermore, both estimators $\hat{\tau_1}^2$ and $\hat{\tau_2}^2$ are consistent per (13) and Lemma A.1,

$$\hat{\tau}_1^2 \xrightarrow{a.s.} \tau_{1,0}^2 \quad \text{and} \quad \hat{\tau}_2^2 \xrightarrow{a.s.} \tau_{2,0}^2,$$

because $\mathbb{E}[X(1)^2] = w_0^2\tau_{2,0}^2 + \tau_{1,0}^2$ and $\mathbb{E}[X(2)^2] = \tau_{2,0}^2$. Therefore, because $h_n(\theta)$ is a continuous function of the estimators, by the continuous mapping theorem, it converges uniformly and almost surely over $\bar{B}_{\epsilon_1}(\theta_0)$

$$h_n(\theta) \xrightarrow{a.s.} h_\infty(\theta) := -\frac{\tau_{2,0}^2}{4\tau_1^6\tau_2^4}\left(1-\frac{2\tau_{2,0}^2}{\tau_2^2}\right)\left(1-\frac{2\tau_{1,0}^2}{\tau_1^2}\right),$$

with $h_\infty(\theta_0) = -1/(4\tau_{1,0}^6\tau_{2,0}^2) =: -C < 0$. Since $h_\infty(\theta)$ is continuous on $\bar{B}_{\epsilon_1}(\theta_0)$, there exist $\epsilon^* < \epsilon_1$ such that

$$h_\infty(\theta) \le -\frac{C}{2} < 0, \text{ for all } \theta \in B_{\epsilon^*}(\theta_0),$$

where $B_{\epsilon^*}(\theta_0)$ is an open ball centered at θ_0. Since uniform convergence on $\bar{B}_{\epsilon_1}(\theta_0)$ implies uniform convergence on its subset $B_{\epsilon^*}(\theta_0)$, we can pass the limit through the supremum

$$\lim_{n\to\infty} \sup_{\theta\in B_{\epsilon^*}(\theta_0)} h_n(\theta) = \sup_{\theta\in B_{\epsilon^*}(\theta_0)} h_\infty(\theta) \le -\frac{C}{2} < 0 \quad \text{a.s.}$$

Because this limit exists and is strictly negative, the limit superior is also strictly negative, yielding the desired result.

(iv) Condition (iv) requires the global consistency of the MLE. For simplicity, we consider only the S^1 case; the other cases follow analogously. Let $\hat{\theta}^1 = [\hat{w}, \hat{\tau}_1^2, \hat{\tau}_2^2]^\top$ denote the MLE under model S^1. Using the expansion in (17), the normalized difference between the log-likelihood at a parameter $\theta = [w, \tau_1^2, \tau_2^2]^\top$ and at the MLE is given by

$$g_n(\theta) := \frac{1}{n}\Big[\ell_n(\theta \mid S^1) - \ell_n(\hat{\theta}^1 \mid S^1)\Big] = -\frac{1}{2}\left(\frac{(w-\hat{w})^2\hat{\tau}_2^2}{\tau_1^2} + \phi\left(\frac{\hat{\tau}_1^2}{\tau_1^2}\right) + \phi\left(\frac{\hat{\tau}_2^2}{\tau_2^2}\right)\right), \tag{20}$$

where we define $\phi(u) = u - 1 - \log u$ for $u > 0$. It is a standard result that $\phi(u) \ge 0$ for all $u > 0$, with equality holding if and only if $u = 1$. Furthermore, $\phi(u)$ is strictly decreasing on $(0, 1)$, strictly increasing on $(1, \infty)$, and satisfies $\lim_{u\to 0}\phi(u) = \lim_{u\to\infty}\phi(u) = \infty$. Because each term inside the bracket is non-negative, $g_n(\theta) \le 0$ globally, with equality strictly at $\theta = \hat{\theta}^1$.

The primary technical challenge is that the parameter space $\Theta = \mathbb{R}\setminus\{0\} \times (0,\infty)^2$ is not compact. We therefore proceed by partitioning $\Theta \setminus B_\delta(\theta_0)$ into a non-compact region and a compact region.

By Lemma A.2, the MLE is consistent, i.e. $\hat{\theta}^1 \xrightarrow{a.s.} \theta_0$. Consequently, the sequence $\hat{\theta}^1$ is eventually bounded almost surely. There exist positive constants c, C such that for n sufficiently large, we have $\hat{\tau}_1^2, \hat{\tau}_2^2 \in [c, C]$ and $|\hat{w}| \le C$.

Fix an arbitrary constant $M > 0$. We construct a compact set $K := [-W, W] \times [l, L] \times [l, L]$ containing $B_\delta(\theta_0)$ by choosing the positive constants W, l, L such that

$$L > C \quad \text{and} \quad \phi\left(\frac{C}{L}\right) > 2M \tag{21}$$

$$l < c \quad \text{and} \quad \phi\left(\frac{c}{l}\right) > 2M \tag{22}$$

$$W > C + \sqrt{\frac{2ML}{c}} \tag{23}$$

Conditions (21) and (22) are trivially satisfied by taking L sufficiently large and l sufficiently small, due to the behavior of ϕ when $u \to 0$ and $u \to \infty$. Next, we demonstrate that for any n large enough and any $\theta \in \Theta \setminus K$, we have $g_n(\theta) < -M$. We evaluate the following three cases

- **Case 1:** If $\tau_i^2 > L$ for at least one $i \in \{1,2\}$, then $\frac{\hat{\tau}_i^2}{\tau_i^2} < \frac{C}{L} < 1$. Because ϕ is strictly decreasing on $(0,1)$, and the condition in (21) implies $\phi\left(\frac{\hat{\tau}_i^2}{\tau_i^2}\right) > \phi\left(\frac{C}{L}\right) > 2M$. This ensures $g_n(\theta) < -M$.
- **Case 2:** If $\tau_i^2 < l$ for at least one $i \in \{1,2\}$, then $\frac{\hat{\tau}_i^2}{\tau_i^2} > \frac{c}{l} > 1$. Because ϕ is strictly increasing on $(1,\infty)$, and the condition in (22) implies $\phi\left(\frac{\hat{\tau}_i^2}{\tau_i^2}\right) > \phi\left(\frac{c}{l}\right) > 2M$. This ensures $g_n(\theta) < -M$.
- **Case 3:** Assume θ is not caught by Cases 1 or 2, which guarantees $\tau_1^2 \leq L$. Because $\theta \notin K$, it must be that $|w| > W$. We then have $\frac{(w-\hat{w})^2\hat{\tau}_2^2}{\tau_1^2} \geq \frac{(|w|-|\hat{w}|)^2 c}{L} \geq \frac{(W-C)^2 c}{L}$. By the condition in (23), this term is strictly greater than $2M$, ensuring $g_n(\theta) < -M$.

Thus, we have established that $\sup_{\theta\in\Theta\setminus K} g_n(\theta) \leq -M$ eventually almost surely.

We now restrict our attention to the compact set, $K \setminus B_\delta(\theta_0)$, bounded away from the true parameter. Because $\hat{\theta}^1 \xrightarrow{a.s.} \theta_0$ and $g_n(\theta)$ is continuous in both θ and $\hat{\theta}^1$, the continuous mapping theorem yields pointwise convergence

$$g_n(\theta) \xrightarrow{a.s.} g_\infty(\theta) = -\frac{1}{2}\left(\frac{(w-w_0)^2\tau_{2,0}^2}{\tau_1^2} + \phi\left(\frac{\tau_{1,0}^2}{\tau_1^2}\right) + \phi\left(\frac{\tau_{2,0}^2}{\tau_2^2}\right)\right).$$

Moreover, since $K \setminus B_\delta(\theta_0)$ is compact, the above convergence implies uniform converge as well, i.e.

$$\sup_{\theta\in K\setminus B_\delta(\theta_0)} |g_n(\theta) - g_\infty(\theta)| \xrightarrow{a.s.} 0.$$

Notice that $g_\infty(\theta) \leq 0$ everywhere, with $g_\infty(\theta) = 0$ if and only if $\theta = \theta_0$. Because $\theta_0 \notin K \setminus B_\delta(\theta_0)$ and the set is compact, $g_\infty(\theta)$ must attain a strictly negative maximum on this set. Let $\sup_{\theta\in K\setminus B_\delta(\theta_0)} g_\infty(\theta) = -M'$ with $M' > 0$. Therefore

$$\begin{aligned}
\sup_{\theta\in K\setminus B_\delta(\theta_0)} g_n(\theta) &= \sup_{\theta\in K\setminus B_\delta(\theta_0)} (g_n(\theta) - g_\infty(\theta) + g_\infty(\theta)) \\
&\leq \sup_{\theta\in K\setminus B_\delta(\theta_0)} g_\infty(\theta) + |g_n(\theta) - g_\infty(\theta)| \\
&\leq \sup_{\theta\in K\setminus B_\delta(\theta_0)} g_\infty(\theta) + \sup_{\theta\in K\setminus B_\delta(\theta_0)} |g_n(\theta) - g_\infty(\theta)| \leq -M' + \frac{M'}{2} = -\frac{M'}{2},
\end{aligned}$$

eventually almost surely. Combining the bounds from the two regions, we bound the supremum over the entire set $\Theta \setminus B_\delta(\theta_0)$

$$\sup_{\theta\in\Theta\setminus B_\delta(\theta_0)} g_n(\theta) = \max\left\{\sup_{\theta\in K\setminus B_\delta(\theta_0)} g_n(\theta), \sup_{\theta\in\Theta\setminus K} g_n(\theta)\right\} \leq \max\{-M'/2, -M\} < 0.$$

Taking the limit supremum yields

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta\setminus B_\delta(\theta_0)} g_n(\theta) < 0 \quad \text{a.s.}$$

Finally, observe that $\frac{1}{n}\left[\ell_n(\theta \mid S^1) - \ell_n(\theta_0 \mid S^1)\right] = g_n(\theta) - g_n(\theta_0)$. Because $g_n(\theta_0) \xrightarrow{a.s.} g_\infty(\theta_0) = 0$, the limit supremum of the difference is identical to the limit supremum of $g_n(\theta)$. Therefore

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta\setminus B_\delta(\theta_0)} \frac{1}{n}\left[\ell_n(\theta \mid S^1) - \ell_n(\theta_0 \mid S^1)\right] < 0,$$

with $P^X_{m(S,\theta_0)}$-probability one, completing the proof.

Since all these conditions are fulfilled, then by Theorem 7 in Kass et al. [1990] the sequence of log-likelihood functions is Laplace regular with $P^X_{m(S,\theta_0)}-$probability one, for any $\theta_0 \in \Theta$. □

Lemma A.4. *Let $\mathcal{D}_{obs} := (X_i)_{i=1}^n$ be n i.i.d. observational samples generated by* (1)*, and let $\mathcal{D}_{int} := (Y_j)_{1\le j\le m}$ be m i.i.d. interventional samples drawn from the post-intervention distribution induced by an intervention $do(Y(2) = y)$. The maximum likelihood estimators for each structure are*

$$\hat{\theta}^1 := [\hat{\theta} \mid S^1] = \left[\frac{S_{12}^X + S_{12}^Y}{S_2^X + S_2^Y}, \frac{(S_1^X + S_1^Y)(S_2^X + S_2^Y) - (S_{12}^X + S_{12}^Y)^2}{(n+m)(S_2^X + S_2^Y)}, \frac{S_2^X}{n}\right]^\intercal, \tag{24}$$

$$\hat{\theta}^2 := [\hat{\theta} \mid S^2] = \left[\frac{S_{12}^X}{S_1^X}, \frac{S_1^X + S_1^Y}{n+m}, \frac{S_1^X S_2^X - S_{12}^{X2}}{nS_1^X}\right]^\intercal, \tag{25}$$

$$\hat{\theta}^3 := [\hat{\theta} \mid S^3] = \left[0, \frac{S_1^X + S_1^Y}{n+m}, \frac{S_2^X}{n}\right]^\intercal, \tag{26}$$

where

$$S_{12}^X := \sum_{i=1}^n X_i(1)X_i(2), \quad S_1^X := \sum_{i=1}^n X_i(1)^2, \quad S_2^X := \sum_{i=1}^n X_i(2)^2, \tag{27}$$

$$S_{12}^Y := y\sum_{i=1}^m Y_i(1), \quad S_1^Y := \sum_{i=1}^n Y_i(1)^2, \quad S_2^Y := my^2. \tag{28}$$

Proof. First, we consider the S^1 structure. The log-likelihood for this case is

$$\begin{aligned}
\ell_{n,m}(\theta \mid S^1) :&= \sum_{i=1}^n \log f_{\text{obs}}(X_i \mid \theta, S^1) + \sum_{j=1}^m \log f_{\text{int}}(Y_j \mid \theta, S^1) \\
&= \sum_{i=1}^n \Big[\log \phi_{\tau_1^2}(X_i(1) - wX_i(2))^2 + \log \phi_{\tau_2^2}(X_i(2))\Big] + \sum_{j=1}^m \log_{\tau_1^2}(Y_i(1) - wy) \\
&= -\Big(n + \frac{m}{2}\Big)\log 2\pi - \frac{n+m}{2}\log\tau_1^2 - \frac{n}{2}\log\tau_2^2 - \frac{S_2^X}{2\tau_2^2} \\
&- \frac{(S_1^X + S_1^Y) - 2w(S_{12}^X + S_{12}^Y) + w^2(S_2^Y + S_2^X)}{2\tau_1^2},
\end{aligned} \tag{29}$$

where f_{obs} and f_{int} are the densities of the observational and interventional samples, respectively. The maximum likelihood estimates are obtained by solving the first-order conditions $\nabla_\theta \ell_{n,m}(\theta \mid S^1) = 0$. Differentiating with respect to w, we find:

$$\frac{\partial \ell_{n,m}(\theta \mid S^1)}{\partial w} = -\frac{1}{2\tau_1^2}\Big[-2(S_{12}^X + S_{12}^Y) + 2w(S_2^X + S_2^Y)\Big] = 0 \implies \hat{w} = \frac{S_{12}^X + S_{12}^Y}{S_2^X + S_2^Y}.$$

Substituting $\hat{w}$ and differentiating with respect to the variance parameters yields the variance estimators. Moreover, one can easily check that $\ell_{n,m}(\theta \mid S^1)$ is strictly concave in θ by showing that the Hessian is negative definite. Therefore, the maximum is unique. The MLEs for models S^2 and S^3 follow analogously. □

Lemma A.5. *Let $\{n_N\}_{N\geq 1}$ and $\{m_N\}_{N\geq 1}$ be two increasing sequences such that $N = n_N + m_N$, and define $\eta_N := \frac{n_N}{N}$, with $\lim_{N\to\infty} \eta_N = \eta \in (0,1)$. Let $\mathcal{D}_{obs} := (X_i)_{1\leq i\leq n_N}$ be n_N i.i.d. observational samples generated by* (1)*, and let $\mathcal{D}_{int} := (Y_j)_{1\leq j\leq m_N}$ be m_N i.i.d. interventional samples drawn from the post-intervention distribution induced by an intervention $do(Y(2) = y)$. Then, as $N \to \infty$, with $\bar{\eta} = 1 - \eta$*

$$\frac{S_2^X}{n_N} \xrightarrow{a.s.} \mathbb{E}[X(2)^2], \tag{30}$$

$$\frac{S_{12}^X + S_{12}^Y}{S_2^X + S_2^Y} \xrightarrow{a.s.} \frac{\eta\mathbb{E}[X(1)X(2)] + \bar{\eta}y\mathbb{E}[Y(1)]}{\eta\mathbb{E}[X(2)^2] + \bar{\eta}y^2}, \tag{31}$$

$$\frac{(S_2^X + S_2^Y)(S_1^X + S_1^Y) - (S_{12}^X + S_{12}^Y)^2}{N(S_2^Y + S_2^X)} \xrightarrow{a.s.} \frac{C}{\eta\mathbb{E}[X(2)^2] + \bar{\eta}y^2}, \tag{32}$$

where S_{12}^X, S_{12}^Y, S_1^Y and S_2^Y are defined in (28) *and*

$$C = (\eta\mathbb{E}[X(1)^2] + \bar{\eta}\mathbb{E}[Y(1)^2])(\eta\mathbb{E}[X(2)^2] + \bar{\eta}y^2) - (\eta\mathbb{E}[X(1)X(2)] + \bar{\eta}y\mathbb{E}[Y(1)])^2.$$

Proof. Since the second moments are finite, $\mathbb{E}[X(1)^2] < \infty$, $\mathbb{E}[X(2)^2] < \infty$ and $\mathbb{E}[Y(1)^2] < \infty$, then, by the Cauchy-Schwarz inequality, $\mathbb{E}[|X(1)X(2)|] \leq \sqrt{\mathbb{E}[X(1)^2]\mathbb{E}[X(2)^2]} < \infty$ and $\mathbb{E}[|Y(1)|] < \infty$.

Applying the strong law of large numbers to $S_2^X = \sum_{i=1}^{n_N} X_i(2)^2$, we obtain almost sure convergence, $\frac{S_1^X}{n_N} \xrightarrow{a.s.} \mathbb{E}[X(1)^2]$. Next, for the second estimator we have

$$\frac{S_{12}^X + S_{12}^Y}{S_2^X + S_2^Y} = \frac{\eta_N \frac{1}{n_N}\sum_{i=1}^{n_N} X_i(1)X_i(2) + \bar{\eta}_N \sum_{j=1}^{m_N} yY_j(1)}{\eta_N \frac{1}{n_N}\sum_{i=1}^{n_N} X_i(2)^2 + \bar{\eta}_N y^2}.$$

Using the strong law of large numbers, the continuous mapping theorem, and the fact that $\eta_N \to \eta \in (0,1)$, the required result follows. The result for the third estimator is obtained by defining the continuous function $g(x,y,z) = \frac{xy-z^2}{y}$ and following the steps taken in the proof of Lemma A.2. □

Lemma A.6. *Let $B_\delta(\theta)$ denote the open ball of radius δ centered at $\theta \in \Theta := \mathbb{R}\setminus\{0\}\times\mathbb{R}_+^2$. Let $\{n_N\}_{N\geq 1}$ and $\{m_N\}_{N\geq 1}$ be two increasing sequences such that $N = n_N + m_N$, and define $\eta_N := \frac{n_N}{N}$, with $\lim_{N\to\infty}\eta_N = \eta \in (0,1)$. For any $S \in \mathcal{S}$ and $\theta \in \Theta$, let $\mathcal{D}_{obs} := (X_i)_{1\leq i\leq n_N}$ be n_N i.i.d. observational samples generated by* (1) *with density $f_{obs}(x \mid S,\theta)$, and let $\mathcal{D}_{int} := (Y_j)_{1\leq j\leq m_N}$ be m_N i.i.d. interventional samples drawn from the post-intervention distribution induced by an intervention $do(Y(2) = y)$ with density $f_{int}(x \mid S,\theta)$. The following four conditions hold*

(i) for all $z_1^N = ((x_i)_{1\leq i\leq n_N}, (y_j)_{1\leq j\leq m_N})$ with $x_i \in \mathbb{R}^2$ and $\theta \in \Theta$, $\prod_{i=1}^{n_N} f_{obs}(x_i \mid S,\theta)\prod_{j=1}^{m_N} f_{int}(y_j \mid S,\theta) > 0$. Moreover, for all z_1^N, $\ell_{n_N,m_N}(\theta \mid S) = \sum_{i=1}^{n_N}\log f_{obs}(x_i \mid S,\theta) + \sum_{j=1}^{m_N}\log f_{int}(y_j \mid S,\theta)$ is six times continuously differentiable in θ;

(ii) for any $\theta_0 \in \Theta$ there exist $\epsilon > 0$ and $M < \infty$ such that $B_\epsilon(\theta_0) \subseteq \Theta$ and for $1 \leq j_1,\ldots,j_d \leq m$ with $d \leq 6$,

$$\limsup_{N\to\infty}\sup\{N^{-1}|\partial_{j_1}\ldots\partial_{j_d}\ell_{n_N,m_N}(\theta \mid S)| : \theta \in B_\epsilon(\theta_0)\} < M,$$

with $P_{m(S,\theta_0)}$-probability one;

(iii) for any $\theta_0 \in \Theta$ and some $\epsilon > 0$,

$$\limsup_{N\to\infty}\sup\{\det(N^{-1}\nabla_\theta^2\ell_{n_N,m_N}(\theta \mid S)) : \theta \in B_\epsilon(\theta_0)\} < 0$$

with $P_{m(S,\theta_0)}$-probability one;

(iv) for any $\theta_0 \in \Theta$ and any $\delta > 0$,

$$\limsup_{N\to\infty}\sup\{N^{-1}[\ell_{n_N,m_N}(\theta \mid S) - \ell_{n_N,m_N}(\theta_0 \mid S)] : \theta \in \Theta\setminus B_\delta(\theta_0)\} < 0$$

with $P_{m(S,\theta_0)}$-probability one.

Hence, the models are Laplace regular as per Kass et al. [1990].

Proof. The proof follows the same steps as in the proof of Lemma A.3. For simplicity, we assume $S = S^1$. The remaining cases follow analogously. Using the expression for the ML estimator, $\hat{\theta}^1$, derived in Lemma A.4 and (29), the log-likelihood is

$$\begin{aligned}\ell_{n_N,m_N}(\theta \mid S^1) &= -\left(n_N + \frac{m_N}{2}\right)\log 2\pi - \frac{n_N}{2}\log(\tau_1^2\tau_2^2) - \frac{m_N}{2}\log\tau_1^2 - \frac{(n_N\hat{\tau}_2^2 + m_N y^2)(w-\hat{w})^2}{2\tau_1^2} \\ &\qquad - \frac{(n_N+m_N)\hat{\tau}_1^2}{2\tau_1^2} - \frac{n_N\hat{\tau}_2^2}{2\tau_2^2} \\ &= -N\Bigg\{\left(\eta_N + \frac{\bar{\eta}_N}{2}\right)\log 2\pi - \frac{\eta_N}{2}\log(\tau_1^2\tau_2^2) - \frac{\bar{\eta}_N}{2}\log\tau_1^2 \\ &\qquad - \frac{(\eta_N\hat{\tau}_2^2 + \bar{\eta}_N y^2)(w-\hat{w})^2}{2\tau_1^2} - \frac{\hat{\tau}_1^2}{2\tau_1^2} - \frac{\eta_N\hat{\tau}_2^2}{2\tau_2^2}\Bigg\}, \end{aligned} \tag{33}$$

where $\bar{\eta}_N = 1 - \eta_N$.

(i)-(ii) Similarly to the proof of LemmaA.3, the first two conditions are easily verifiable; i.e. $\ell_{n_N,m_N}(\theta \mid S^1)$ is 6-times differentiable and $f(z_1^N \mid S^1, \theta) > 0$ for all $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{int}}$ and $\theta \in \Theta$. Moreover, $\ell_{n_N,m_N}(\theta \mid S^1)$ is composed of terms that are either quadratic polynomials in w and rational functions of the variance parameters τ_1^2 and τ_2^2 and hence the partial derivatives are bounded in a vicinity of θ_0.

(iii) The Hessian is

$$H(\theta) := H(w, \tau_1^2, \tau_2^2) = N\begin{bmatrix} -\frac{(\eta_N\hat{\tau}_2^2+\bar{\eta}_N y^2)}{\tau_1^2} & \frac{(\eta_N\hat{\tau}_2^2+\bar{\eta}_N y^2)(w-\hat{w})}{(\tau_1^2)^2} & 0 \\ \frac{(\eta_N\hat{\tau}_2^2+\bar{\eta}_N y^2)(w-\hat{w})}{(\tau_1^2)^2} & \frac{1}{2(\tau_1^2)^2} - \frac{\left[\hat{\tau}_1^2+(\eta_N\hat{\tau}_2^2+\bar{\eta}_N y^2)(w-\hat{w})^2\right]}{(\tau_1^2)^3} & 0 \\ 0 & 0 & \frac{\eta_N}{2(\tau_2^2)^2} - \frac{\eta_N\hat{\tau}_2^2}{(\tau_2^2)^3} \end{bmatrix} \tag{34}$$

with the determinant of the normalized Hessian being

$$h_N(\theta) := \det\left(\frac{1}{N}H(\theta)\right) = -\frac{\eta_N(\eta_N\hat{\tau}_2^2 + \bar{\eta}_N y^2)}{4\tau_1^6\tau_2^4}\left(1 - \frac{2\hat{\tau}_2^2}{\tau_2^2}\right)\left(1 - \frac{2\hat{\tau}_1^2}{\tau_1^2}\right).$$

Let $\theta_0 := \left(w_0, \tau_{1,0}^2, \tau_{2,0}^2\right) \in \Theta$ be the true generating parameter. Furthermore, we note that both estimators ${\hat{\tau}_1}^2$ and ${\hat{\tau}_2}^2$ are consistent per (24) and Lemma A.5,

$$\hat{\tau}_1^2 \xrightarrow{a.s.} \tau_{1,0}^2 \quad \text{and} \quad \hat{\tau}_2^2 \xrightarrow{a.s.} \tau_{2,0}^2,$$

because $\mathbb{E}[X(1)^2] = w_0^2\tau_{2,0}^2 + \tau_{1,0}^2$, $\mathbb{E}[X(2)^2] = \tau_{2,0}^2$, $\mathbb{E}[Y(1)^2] = w_0^2 y^2 + \tau_{1,0}^2$, $\mathbb{E}[X(1)X(2)] = w_0\tau_{2,0}^2$ and $\mathbb{E}[Y(1)Y(2)] = w_0 y^2$. Since the estimators are consistent and the function $h_N(\theta)$ resembles the definition of $h_n(\theta)$ from (19), the desired result follows from the same arguments presented in the proof of condition (iii) in Lemma A.3.

(iv) Using the expansion in (33), the normalized difference between the log-likelihood at a parameter $\theta = [w, \tau_1^2, \tau_2^2]^\top$ and at the MLE is given by

$$\begin{aligned} g_N(\theta) &:= \frac{1}{N}\left[\ell_{n_N,m_N}(\theta \mid S^1) - \ell_{n_N,m_N}(\hat{\theta}^1 \mid S^1)\right] \\ &= -\frac{1}{2}\left[\frac{(\eta_N\hat{\tau}_2^2 + \bar{\eta}_N y^2)(w-\hat{w})^2}{\tau_1^2} + \phi\left(\frac{\hat{\tau}_1^2}{\tau_1^2}\right) + \eta_N\phi\left(\frac{\hat{\tau}_2^2}{\tau_2^2}\right)\right], \end{aligned}$$

where we define $\phi(u) = u - 1 - \log u$ for $u > 0$. The function $g_N(\theta)$ is structurally analogous to $g_n(\theta)$ defined in (20), differing only by the existence of the η_N for which $\eta_N \to \eta$ as $N \to \infty$. Consequently, the result follows from the arguments presented in the proof of condition (iv) in Lemma A.3.

□

Lemma A.7 (Useful integral)**.** *For any $\nu > 0$ and $A, B, C \in \mathbb{R}$ such that $CA^2 > B$ we have*

$$\int_{-\infty}^{\infty} \frac{dx}{(A^2x^2 - 2Bx + C)^{\frac{\nu+1}{2}}} = \sqrt{\pi}\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})} \times \frac{(A^2)^{\frac{\nu-1}{2}}}{(CA^2 - B^2)^{\frac{\nu}{2}}},$$

where $\Gamma(\cdot)$ is the usual Gamma-function.

PROOF: Completing the square in the dominator we have

$$\int_{-\infty}^{\infty} \frac{dx}{(A^2x^2 - 2Bx + C)^{\frac{\nu+1}{2}}} = \left(\frac{A^2}{CA^2 - B^2}\right)^{\frac{\nu+1}{2}} \int_{-\infty}^{\infty} \frac{dx}{\left(\frac{A^4}{CA^2-B^2}\left(x - \frac{B}{A^2}\right)^2 + 1\right)^{\frac{\nu+1}{2}}}.$$

Next, we recall the probability density function (pdf) of a shifted and scaled version of the standard student-t distribution (i.e. $Z = \mu + \sigma T$, with $T \sim t(\nu)$)

$$f_Z(z) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi}}\left(1 + \frac{1}{\nu}\left(\frac{z-\mu}{\sigma^2}\right)^2\right)^{-\frac{\nu+1}{2}}.$$

Then, matching the terms gives the desired result. □

B PROOFS FOR CONCAVITY OF THE CONCENTRATION RATES

Lemma B.1. *The exponent $D_{12}(\eta)$ defined in* (51) *is positive and strictly concave for $\eta \in (0,1)$ if $\tau_2^{*2} \neq y^2 > 0$. Moreover, if*

$$\frac{w^{*2}(\tau_2^{*2} - y^2)}{w^{*2}y^2 + \tau_1^{*2}} > \log\left(1 + \frac{w^{*2}\tau_2^{*2}}{\tau_1^{*2}}\right), \tag{35}$$

$D_{12}(\eta)$ achieves a unique maximum for some $\eta^* \in (0,1)$. Otherwise, $D_{12}(\eta)$ is strictly decreasing for $\eta \in (0,1)$.

Proof. The explicit form derived is

$$D_{12}(\eta) = \frac{1}{2}\log\left(\eta\sigma_{1,X}^2 + \bar{\eta}\sigma_{1,Y}^2\right) - \frac{\eta}{2}\log\sigma_{1,X}^2 - \frac{\bar{\eta}}{2}\log\tau_1^{*2},$$

where $\sigma_{1,X}^2$ and $\sigma_{1,Y}^2$ are defined in (51) and $\eta \in (0,1)$. We investigate the analytic properties of this exponent by computing the first two derivatives with respect to η

$$\frac{dD_{12}(\eta)}{d\eta} = \frac{\sigma_{1,X}^2 - \sigma_{1,Y}^2}{2[\eta(\sigma_{1,X}^2 - \sigma_{1,Y}^2) + \sigma_{1,Y}^2]} - \frac{1}{2}\log\sigma_{1,X}^2 + \frac{1}{2}\log\tau_1^{*2},$$
$$\frac{d^2D_{12}(\eta)}{d\eta^2} = -\frac{(\sigma_{1,X}^2 - \sigma_{1,Y}^2)^2}{2(\eta\sigma_{1,X}^2 + \bar{\eta}\sigma_{1,Y}^2)^2}.$$

Observing the second derivative, we conclude that the function is strictly concave provided $\sigma_{1,X}^2 \neq \sigma_{1,Y}^2$ (or equivalently, $y^2 \neq \tau_2^{*2}$). Furthermore, we note the boundary conditions: $\lim_{\eta\to 0} D_{12}(\eta) = \frac{1}{2}\log\left(1 + \frac{w^{*2}y^2}{\tau_1^{*2}}\right) > 0$ for all $y \neq 0$, and $\lim_{\eta\to 1} D_{12}(\eta) = 0$.

Consequently, if the first derivative evaluated at $\eta = 0$ is non-positive, the exponent is strictly decreasing in η, implying that the maximal rate is achieved at $\eta^* = 0$. Conversely, if

$$\left.\frac{dD_{12}(\eta)}{d\eta}\right|_{\eta=0} > 0 \iff \frac{w^{*2}(\tau_2^{*2} - y^2)}{2(w^{*2}y^2 + \tau_1^{*2})} > \frac{1}{2}\log\left(1 + \frac{w^{*2}\tau_2^{*2}}{\tau_1^{*2}}\right), \tag{36}$$

then $D_{12}(\eta)$ achieves a maximum at an interior point $\eta^* \in (0,1)$. □

Lemma B.2. *The exponent $D_{21}(\eta)$ defined in* (59) *is positive and strictly concave for $\eta \in (0,1)$. $D_{21}(\eta)$ achieves a unique maximum for $\eta^* \in (0,1)$.*

Proof. The rate is given by:

$$D_{21}(\eta) = \frac{1}{2}\log\left(1 - \frac{\eta^2 w^{*2}\tau_1^{*2}}{\eta(w^{*2}\tau_1^{*2} + \tau_2^{*2}) + \bar{\eta}y^2}\right) + \frac{\eta}{2}\log\left(1 + \frac{w^{*2}\tau_1^{*2}}{\tau_2^{*2}}\right). \tag{37}$$

One can easily show that

$$\frac{d^2}{d\eta^2}\left(\frac{\eta^2 w^{*2}\tau_1^{*2}}{\eta(w^{*2}\tau_1^{*2} + \tau_2^{*2} - y^2) + y^2}\right) = \frac{2w^{*2}\tau_1^{*2}y^4}{[(w^{*2}\tau_1^{*2} + \tau_2^{*2} - y^2)\eta + y^2]^3} > 0,$$

which indicates that the inner term in the first logarithm is strictly convex. This, combined with the fact that the second term in D_{21} is linear in η (hence concave) and the function $u \mapsto \log(1-u)$ is concave and decreasing for $u \in (0,1)$, shows that $D_{21}(\eta)$ is strictly concave. Moreover, since the function is continuous in η and

$$\lim_{\eta\to 0^+} D_{21}(\eta) = 0, \text{ and } \lim_{\eta\to 1^-} D_{21}(\eta) = 0,$$

then the maximum is achieved for some $\eta^* \in (0,1)$. □

C PROOF OF THEOREM 3.1

Proof. The marginal likelihood of the data $\mathcal{D}_{\text{obs}}$ conditioned on model S^1 is given by the integral of the likelihood weighted by the prior

$$f_n(\mathcal{D}_{\text{obs}} \mid S^{1*}) := \int_\Theta f_n(\mathcal{D}_{\text{obs}} \mid S^{1*}, \theta)\pi(\theta|S^1)\, d\theta = \int_\Theta \exp\left(\ell_n(\theta \mid S^{1*})\right) \pi(\theta \mid S^{1*})\, d\theta, \tag{38}$$

where $f_n(\mathcal{D}_{\text{obs}} \mid S^{1*}, \theta)$ denotes the joint likelihood function with parameters θ for structure S^1 and $\ell_n(\theta \mid S^1) := \sum_{i=1}^n \log f(X_i \mid S^1, \theta)$.

Since the models considered satisfy the Laplace regularity conditions (Lemma A.3), the log-marginal likelihood can be approximated as follows

$$\begin{aligned}\log f_n(\mathcal{D}_{\text{obs}} \mid S^{1*}) = \ell_n(\hat{\theta}^1 \mid S^{1*}) - \frac{3}{2}\log n + \frac{3}{2}\log 2\pi \\ - \frac{1}{2}\log\det\left(-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}^1 \mid S^{1*})\right) + \log\pi(\hat{\theta}^1 \mid S^{1*}) + O(n^{-1}) \text{ a.s.}\end{aligned} \tag{39}$$

Next, we denote the empirical Fisher information matrix by $D_n^1(\theta) := -\frac{1}{n}\nabla^2\ell_n(\theta \mid S^1) = -\frac{1}{n}\sum_{i=1}^n \nabla^2 \log f(X_i \mid \theta, S^1)$ for any $\theta \in \Theta$, and the expected Fisher information matrix $I_1(\theta) = \mathbb{E}[-\nabla^2 \log f(X \mid \theta, S^1)] = \text{Diag}\left(\frac{\tau_2^2}{\tau_1^2}, \frac{1}{2\tau_1^4}, \frac{1}{2\tau_2^4}\right)$ is the Fisher information matrix under structure S^1 (parameterized by $\theta = [w, \tau_1^2, \tau_2^2]^\intercal$) where $X \sim P^X_{m(S^1,\theta)}$. We decompose the difference between the observed and expected information at the MLE as follows

$$\begin{aligned}D_n^1(\hat{\theta}^1) - I_1(\hat{\theta}^1) &= [D_n^1(\hat{\theta}^1) - D_n^1(\theta^*)] + [D_n^1(\theta^*) - I_1(\theta^*)] + [I_1(\theta^*) - I_1(\hat{\theta}^1)] \\ &=: T_{n,1} + T_{n,2} + T_{n,3}\end{aligned}$$

where $T_{n,1} = D_n^1(\hat{\theta}^1) - D_n^1(\theta^*)$, $T_{n,2} = D_n^1(\theta^*) - I_1(\theta^*)$ and $T_{n,3} = I_1(\theta^*) - I_1(\hat{\theta}^1)$. Since $\|\hat{\theta}^1 - \theta^*\|_2 = O\left(\sqrt{\frac{\log\log n}{n}}\right)$ a.s. (from Lemma A.2) and $D_n^1(\theta)$, $I_1(\theta)$ are continuously differentiable in θ, then by the Mean Value Theorem, we have $T_{n,1} = O\left(\sqrt{\frac{\log\log n}{n}}\right)$ and $T_{n,3} = O\left(\sqrt{\frac{\log\log n}{n}}\right)$ a.s. Then, the Law of the Iterated Logarithm (LIL) yields $T_{n,2} = O\left(\sqrt{\frac{\log\log n}{n}}\right)$ a.s. for $n \to \infty$. Therefore,

$$D_n^1(\hat{\theta}^1) - I_1(\hat{\theta}^1) = O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.},$$

which, combined with the fact that the log-determinant is a continuously differentiable function, gives

$$\log\det\left(-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}^1 \mid S^{1*})\right) = \log\det(I_1(\hat{\theta}^1)) + O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.} \tag{40}$$

Therefore, substituting this into (39), the marginal likelihood is approximated as follows for $n \to \infty$:

$$\begin{aligned}\log f_n(\mathcal{D}_{\text{obs}} \mid S^{1*}) = \ell_n(\hat{\theta}^1 \mid S^{1*}) - \frac{3}{2}\log n + \frac{3}{2}\log 2\pi \\ - \frac{1}{2}\log\det(I_1(\hat{\theta}^1)) + \log\pi(\hat{\theta}^1 \mid S^{1*}) + O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.}\end{aligned} \tag{41}$$

Similarly, for the reverse causal model S^2 and the independence model S^3, we have as $n \to \infty$

$$\log f_n(\mathcal{D}_{\text{obs}} \mid S^2) = \ell_n(\hat{\theta}^2 \mid S^2) - \frac{3}{2}\log n - \frac{1}{2}\log\det(I_2(\hat{\theta}^2)) + \log \pi(\hat{\theta}^2 \mid S^2) + \frac{3}{2}\log 2\pi + O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.}, \tag{42}$$

$$\log f_n(\mathcal{D}_{\text{obs}} \mid S^3) = \ell_n(\hat{\theta}^3 \mid S^3) - \frac{2}{2}\log n - \frac{1}{2}\log\det(I_3(\hat{\theta}^3)) + \log \pi(\hat{\theta}^3 \mid S^3) + \log 2\pi + O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.} \tag{43}$$

The Fisher information matrix for S^2, $I_2(\theta)$, is $I_2(\theta) := I_2\left([w, \tau_1^2, \tau_2^2]^\intercal\right) = Diag\left(\frac{\tau_1^2}{\tau_2^2}, \frac{1}{2\tau_2^4}, \frac{1}{2\tau_1^4}\right)$. For the independence model S^3, the parameter space is reduced to the two variances ($w = 0$), yielding the 2×2 information matrix: $I_3(\theta) := I_3\left([0, \tau_1^2, \tau_2^2]^\intercal\right) = Diag\left(\frac{1}{2\tau_1^4}, \frac{1}{2\tau_2^4}\right)$.

We first analyze the log-marginal likelihood ratio between the observationally equivalent models S^1 and S^2. Assuming the true generating structure is S^{1*} with parameter θ^*, we note that $\ell_n(\hat{\theta}^1 \mid S^1) = \ell_n(\hat{\theta}^2 \mid S^2)$ and the dimension penalty terms are equal. Thus, as $n \to \infty$

$$\log \frac{f_n(\mathcal{D}_{\text{obs}} \mid S^{1*})}{f_n(\mathcal{D}_{\text{obs}} \mid S^2)} = \frac{1}{2}\log \frac{\det(I_2(\hat{\theta}^2))}{\det(I_1(\hat{\theta}^1))} + \log \frac{\pi(\hat{\theta}^1 \mid S^{1*})}{\pi(\hat{\theta}^2 \mid S^2)} + O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.} \tag{44}$$

From Lemma A.2, $\hat{\theta}^1 \xrightarrow{a.s.} \theta^*$, $\hat{\theta}^2 \xrightarrow{a.s.} \gamma(\theta^*)$ and $\hat{\theta}^3 \xrightarrow{a.s.} [0, w^{*2}\tau_2^{*2} + \tau_1^{*2}, \tau_2^{*2}]^\intercal =: \theta^3_\infty$. The Fisher information matrices satisfy the transformation rule $I_1(\theta) = J_\gamma(\theta)^\top I_2(\gamma(\theta)) J_\gamma(\theta)$, with $J_\gamma(\theta)$ being the Jacobian of the $\gamma-$ transformation,

$$J_\gamma(\theta) = \begin{bmatrix} \frac{\tau_2^2(\tau_1^2 - w^2\tau_2^2)}{D^2} & -\frac{w\tau_2^2}{D^2} & \frac{w\tau_1^2}{D^2} \\ 2w\tau_2^2 & 1 & w^2 \\ -\frac{2w\tau_1^2(\tau_2^2)^2}{D^2} & \frac{w^2(\tau_2^2)^2}{D^2} & \frac{(\tau_1^2)^2}{D^2} \end{bmatrix}, \tag{45}$$

where $D = w^2\tau_2^2 + \tau_1^2$ and the determinant of $J_\gamma(\theta)$ is $\det(J_\gamma(\theta)) = \frac{\tau_2^2}{w^2\tau_2^2 + \tau_1^2}$. This implies that $\det(I_1(\theta)) = \det(I_2(\gamma(\theta))) \cdot (\det J_\gamma(\theta))^2$ which gives

$$\begin{aligned} \log \frac{f_n(\mathcal{D}_{\text{obs}} \mid S^{1*})}{f_n(\mathcal{D}_{\text{obs}} | S^2)} &\xrightarrow{a.s.} \frac{1}{2}\log\left(\frac{\det(I_2(\gamma(\theta^*)))}{\det(I_2(\gamma(\theta^*)))(\det J_\gamma(\theta^*))^2}\right) + \log\frac{\pi(\theta^* \mid S^{1*})}{\pi(\gamma(\theta^*) \mid S^2)} \\ &= \log\left(\frac{\pi(\theta^* \mid S^{1*})}{\pi(\gamma(\theta^*) \mid S^2)|\det J_\gamma(\theta^*)|}\right) \\ &= \log\left(\frac{\pi(\theta^* \mid S^{1*})}{\pi(\gamma(\theta^*) \mid S^2)} \cdot \frac{w^{*2}\tau_2^{*2} + \tau_1^{*2}}{\tau_2^{*2}}\right) \text{ a.s.} \end{aligned} \tag{46}$$

This is precisely the log-ratio of the prior density, $\pi(\theta^* \mid S^{1*})$ to the push-forward of the alternative prior, $\gamma^{-1}\#\pi(\theta^* \mid S^2)$. For the incorrect model S^3, the normalized log marginal likelihood ratio is

$$\frac{1}{n}\log\frac{f_n(\mathcal{D}_{\text{obs}} \mid S^3)}{f_n(\mathcal{D}_{\text{obs}} \mid S^{1*})} = \frac{1}{n}\left(\ell_n(\hat{\theta}^3 \mid S^3) - \ell_n(\hat{\theta}^1 \mid S^{1*})\right) + O\left(\sqrt{\frac{\log\log n}{n^{3/2}}}\right). \text{ a.s.}$$

Since the log-likelihood functions are continuous in $\theta \in \Theta$, by the continuous mapping theorem and the strong law of large numbers, the normalized log-marginal likelihood ratio converges almost surely to the relative entropy between $P^X_{m(S^{1*}, \theta^*)}$ and $P^X_{m(S^3, \theta^3_\infty)}$

$$\frac{1}{n}\log\frac{f_n(\mathcal{D}_{\text{obs}} \mid S^3)}{f_n(\mathcal{D}_{\text{obs}} \mid S^{1*})} \xrightarrow{a.s.} \mathbb{E}[\log f(X \mid \theta^3_\infty, S^3)] - \mathbb{E}[\log f(X \mid \theta^*, S^{1*})]$$
$$= -D(P^X_{m(S^{1*},\theta^*)} \| P^X_{m(S^3,\theta^3_\infty)}) = -\frac{1}{2}\log\left(1 + \frac{w^{*2}\tau_2^{*2}}{\tau_1^{*2}}\right) < 0$$

where $X \sim P^X_{m(S^{1*},\theta^*)}$, $\theta^3_\infty = [0, w^{*2}\tau_2^{*2} + \tau_1^{*2}, \tau_2^{*2}]^\intercal$ are the limiting MLE parameters under the "wrong" structure S^3 and

$$P^X_{m(S^{1*},\theta^*)} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} w^2\tau_2^{*2} + \tau_1^{*2} & w^*\tau_2^{*2} \\ w^*\tau_2^{*2} & \tau_2^{*2}\end{bmatrix}\right), P^X_{m(S^3,\theta^3_\infty)} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} w^2\tau_2^{*2} + \tau_1^{*2} & 0 \\ 0 & \tau_2^{*2}\end{bmatrix}\right).$$

Therefore, by the continuous mapping theorem,

$$\frac{f_n(\mathcal{D}_{\text{obs}} \mid S^3)}{f_n(\mathcal{D}_{\text{obs}} \mid S^{1*})} \xrightarrow{a.s.} 0.$$

Finally, by the application of the continuous mapping theorem the posterior probability of S^1 converges to

$$\pi(S^{1*} \mid \mathcal{D}_{\text{obs}}) = \frac{1}{1 + \frac{f_n(\mathcal{D}_{\text{obs}}|S^2)}{f_n(\mathcal{D}_{\text{obs}}|S^{1*})} + \frac{f_n(\mathcal{D}_{\text{obs}}|S^3)}{f_n(\mathcal{D}_{\text{obs}}|S^{1*})}} \xrightarrow{a.s.} \frac{1}{1 + \frac{\gamma^{-1}\#\pi(\theta^*|S^2)}{\pi(\theta^*|S^{1*})}}.$$

□

D PROOF OF THEOREM 3.2

Proof. Following the result of Lemma A.2, for $\mathcal{M}^3 = \{\theta^*, S^{3*}\}$ with $\theta^* := [0, \tau_1^{*2}, \tau_2^{*2}]^\intercal$, the corresponding MLEs converge almost surely as follows

$$\hat{\theta}^1, \hat{\theta}^2, \hat{\theta}^3 \xrightarrow{a.s.} [0, \tau_1^{*2}, \tau_2^{*2}]^\intercal,$$

because $\mathbb{E}[X(1)X(2)] = \mathbb{E}[X(1)]\mathbb{E}[X(2)] = 0$, $\mathbb{E}[X(1)^2] = \tau_1^{*2}$ and $\mathbb{E}[X(2)^2] = \tau_2^{*2}$. Using the Laplace approximation (Lemma A.3), we expand the log-marginal likelihoods. Note that model S^3 has dimension $d_3 = 2$, while S^1 and S^2 have dimension $d_1 = d_2 = 3$. The Bayes factor between the complex models (S^1, S^2) and the true simple model (S^3) is determined by the difference in dimensions (see the asymptotic expansions in (41),(42) and (43))

$$\log\frac{f_n(\mathcal{D}_{\text{obs}} \mid S^1)}{f_n(\mathcal{D}_{\text{obs}} \mid S^{3*})} = \ell_n(\hat{\theta}^1 \mid S^1) - \ell_n(\hat{\theta}^3 \mid S^3) - \frac{3-2}{2}\log n + O(1) \text{ a.s.} \tag{47}$$

Exponentiating this, we analyze the behavior of the likelihood ratio scaled by $\sqrt{n}$

$$\sqrt{n}\frac{f_n(\mathcal{D}_{\text{obs}} \mid S^1)}{f_n(\mathcal{D}_{\text{obs}} \mid S^{3*})} = \exp\left(\ell_n(\hat{\theta}^1 \mid S^1) - \ell_n(\hat{\theta}^3 \mid S^3)\right) \cdot n^{1/2} \cdot n^{-1/2} \cdot e^{O_p(1)} = \lambda_1,$$

with $\lambda_1 = O_p(1)$ as we show next. Using the definitions in (16), twice the log-likelihood difference is

$$2(\ell_n(\hat{\theta}^1 \mid S^1) - \ell_n(\hat{\theta}^3 \mid S^{3*})) = \frac{2n}{2}\log\left(1 - \frac{(S^X_{12})^2}{S^X_1 S^X_2}\right) = n\log(1 - r_n^2).$$

where for simplicity we denote $r_n = \frac{S^X_{12}}{\sqrt{S^X_1 S^X_2}}$. By the strong law of large numbers and (A.2), we have

$$\frac{1}{n}S^X_1 \xrightarrow{a.s.} \tau_1^{*2}, \quad \frac{1}{n}S^X_2 \xrightarrow{a.s.} \tau_2^{*2} \tag{48}$$

and by the central limit theorem

$$\frac{1}{\sqrt{n}}S^X_{12} \xrightarrow{d} N(0, \tau_1^{*2}\tau_2^{*2}). \tag{49}$$

Therefore, using the continuous mapping theorem and then Slutsky's lemma, we obtain that $\sqrt{n}r_n \xrightarrow{d} N(0,1)$. To find the distribution of the statistics $n\log(1-r_n^2)$ we perform a Taylor expansion of the function $f(t) = \log(1-t)$ around $t = 0$, substituting $t = r_n^2$

$$n\log(1-r_n^2) = -nr_n^2 - \frac{n(r_n^2)^2}{2} - O(nr_n^6).$$

We established previously that $nr_n^2 \xrightarrow{d} \chi_1^2$. Therefore, $-nr_n^2 \xrightarrow{d} -\chi_1^2$. We can rewrite $\frac{nr_n^4}{2}$ as $\frac{1}{2n}(nr_n^2)^2$. Since nr_n^2 converges to a random variable (is $O_p(1)$) and $\frac{1}{2n} \to 0$, the product converges to 0 in probability, i.e. $\frac{nr_n^4}{2} \xrightarrow{p} 0$. Therefore, by Slutsky's lemma, the limit is determined entirely by the leading term

$$n\log(1-r_n^2) \xrightarrow{d} -\chi_1^2. \tag{50}$$

Therefore, $2(\ell_n(\hat{\theta}^1 \mid S^1) - \ell_n(\hat{\theta}^3 \mid S^{3*}))$ converges in distribution to a χ_1^2 random variable and hence it is bounded in probability. Similarly, for S^2, we have $\sqrt{n}\frac{f(\mathcal{D}_{\text{obs}}|S^1)}{f(\mathcal{D}_{\text{obs}}|S^{3*})} = \lambda_2$ with $\lambda_2 = O_p(1)$. The posterior probability of S^3 is

$$\pi(S^{3*} \mid \mathcal{D}_{\text{obs}}) = \frac{1}{1 + \frac{f_n(\mathcal{D}_{\text{obs}}|S^1)}{f_n(\mathcal{D}_{\text{obs}}|S^{3*})} + \frac{f_n(\mathcal{D}_{\text{obs}}|S^2)}{f_n(\mathcal{D}_{\text{obs}}|S^{3*})}}.$$

For simplicity, let $C_n := \frac{f_n(\mathcal{D}_{\text{obs}}|S^1)}{f_n(\mathcal{D}_{\text{obs}}|S^3)} + \frac{f_n(\mathcal{D}_{\text{obs}}|S^2)}{f_n(\mathcal{D}_{\text{obs}}|S^3)}$ and $Z_n := \frac{1}{1+C_n}$. Since $\sqrt{n}C_n = O_p(1)$, it implies $C_n = O_p(n^{-1/2})$, and thus $C_n \xrightarrow{p} 0$. By the continuous mapping theorem, $Z_n \to 1$. To analyze the rate of convergence to 1, we investigate the behaviour of the scaled difference $\sqrt{n}(Z_n - 1)$. Expanding the difference, we obtain

$$\sqrt{n}(Z_n - 1) = \sqrt{n}\Big(\frac{1}{1+C_n} - 1\Big) = -\sqrt{n}C_n \times \frac{1}{1+C_n}.$$

Since $\sqrt{n}C_n \xrightarrow{d} \lambda_1 + \lambda_2$ and $C_n \xrightarrow{p} 0$ by Slutsky's lemma we obtain that

$$\sqrt{n}(Z_n - 1) \xrightarrow{d} -(\lambda_1 + \lambda_2).$$

Therefore, the posterior mass concentrates to 1 with an error of order $O_p(n^{-1/2})$. □

E PROOF OF THEOREM E.1: S^{1*} THE CORRECT MODEL

Theorem E.1 (Posterior concentration for S^{1*})**:** Let $\mathcal{D}_{\text{obs}} = (X_i)_{i=1}^{n_N}$ be an i.i.d. observational sample generated from (1), and $\mathcal{D}_{\text{int}} = (Y_i)_{i=1}^{m_N}$ be an i.i.d. interventional sample generated under the intervention $do(X(2) = y)$. Assume the priors $\pi(\theta \mid S)$ are four times differentiable for all $S \in \mathcal{S}$. Under the true generating model $\mathcal{M}^1 = \{S^{1*}, \theta^*\}$, where $X \sim P^X_{m(S^{1*},\theta^*)}$ and $Y \sim P^Y_{m(S^{1*},\theta^*)}$ with $\theta^* = [w^*, \tau_1^{*2}, \tau_2^{*2}]^\intercal$, the posterior probability satisfies, as $N \to \infty$,

$$\frac{1}{N}\log\left(\frac{1}{\pi(S^{1*} \mid \mathcal{D}_{\text{mix}})} - 1\right) \xrightarrow{a.s.} -D_{12}(\eta),$$

where $D_{12}(\eta)$ is defined as

$$D_{12}(\eta) := \frac{1}{2}\log\left(\frac{\eta\sigma_{1,X}^2 + (1-\eta)\sigma_{1,Y}^2}{(\sigma_{1,X}^2)^\eta(\tau_1^{*2})^{\bar{\eta}}}\right), \tag{51}$$

with $\sigma_{1,X}^2 := w^{*2}\tau_2^{*2} + \tau_1^{*2}$ and $\sigma_{1,Y}^2 := w^{*2}y^2 + \tau_1^{*2}$.

Proof. Given that the covariance matrix for X exists and $y \in \mathbb{R}$, the second moments defined in Lemma A.4 are finite. Let $\bar{\eta} := 1 - \eta$. Following the results of Lemmas A.4 and A.5, under the first model, S^{1*}, the MLE for the first model is consistent

$$\hat{w}^1 = \frac{S_{12}^X + S_{12}^Y}{S_2^X + S_2^Y} \xrightarrow{a.s.} \frac{\eta\mathbb{E}[X(1)X(2)] + \bar{\eta}y\mathbb{E}[Y(1)]}{\eta\mathbb{E}[X(2)^2] + \bar{\eta}y^2} = \frac{\eta w^*\tau_2^{*2} + \bar{\eta}w^*y^2}{\eta\tau_2^{*2} + \bar{\eta}y^2} = w^*.$$

Similarly, $\{\hat{\tau}_1^2 \mid S^{1*}\} \xrightarrow{a.s.} \tau_1^{*2}$ and $\{\hat{\tau}_2^2 \mid S^{1*}\} \xrightarrow{a.s.} \tau_2^{*2}$. For the second structure S^2, assuming S^{1*} is true, the MLE $\hat{\theta}^2$ converge to a pseudo-true vector $\hat{\theta}^2 \xrightarrow{a.s.} \theta_\infty^2$ with components

$$\begin{aligned}
\frac{S_{12}^X}{S_1^X} &\xrightarrow{a.s} \frac{w^* \tau_2^{*2}}{w^{*2}\tau_2^{*2} + \tau_1^{*2}}, \\
\frac{S_1^X + S_1^Y}{N} &\xrightarrow{a.s.} \eta(w^{*2}\tau_2^{*2} + \tau_1^{*2}) + \bar{\eta}(w^{*2}y^2 + \tau_1^{*2}), \\
\frac{S_1^X S_2^X - (S_{12}^X)^2}{n_N S_1^X} &\xrightarrow{a.s.} \frac{\tau_1^{*2}\tau_2^{*2}}{w^{*2}\tau_2^{*2} + \tau_1^{*2}}.
\end{aligned}$$

We can explicitly decompose the limit θ_∞^2 into the observational non-identifiable component $\gamma(\theta^*)$ and a shift term induced by the intervention

$$\theta_\infty^2 = [w_{\infty|S^2}, \tau_{1,\infty|S^2}^2, \tau_{2,\infty|S^2}^2] = \gamma(\theta^*) + [0, \bar{\eta} w^{*2}(y^2 - \tau_2^{*2}), 0]^\intercal. \tag{52}$$

The non-zero middle term, $\bar{\eta} w^{*2}(y^2 - \tau_2^{*2})$, represents the discrepancy between the natural variance of the cause and the fixed interventional value. This term breaks the observational symmetry provided $y^2 \neq \tau_2^{*2}$. Finally, for the third structure S^3 (independence, where $w = 0$), the estimator for w is constrained to zero, and the variance parameters converge to the marginal variances of the data mixture. Specifically, we have

$$\hat{\theta}^3 \xrightarrow{a.s.} \theta_\infty^3 = [w_{\infty|S^3}, \tau_{1,\infty|S^3}^2, \tau_{2,\infty|S^3}^2] = [0, \eta(w^{*2}\tau_2^{*2} + \tau_1^{*2}) + \bar{\eta}(w^{*2}y^2 + \tau_1^{*2}), \tau_2^{*2}]^\intercal. \tag{53}$$

The log-marginal likelihood of the mixed dataset $\mathcal{D}_{\text{mix}}$ under S^{1*} is

$$\begin{aligned}
\log f_N(\mathcal{D}_{\text{mix}} \mid S^{1*}) &:= \log \int_\Theta \exp(\ell_{n_N, m_N}(\theta \mid S^{1*}))\pi(\theta \mid S^{1*}) d\theta \\
&= \log \int_\Theta \exp\left(\ell_{n_N}(\theta \mid S^{1*}) + \ell_{m_N}(\theta \mid S^{1*})\right) \pi(\theta \mid S^{1*})\, d\theta,
\end{aligned}$$

where we define $\ell_{n_N}(\theta \mid S)$ and $\ell_{m_N}(\theta \mid S)$ as being the corresponding log-likelihoods under structure S of the observational, $\mathcal{D}_{\text{obs}}$, and interventional datasets, $\mathcal{D}_{\text{int}}$, respectively

$$\ell_{n_N}(\theta \mid S) := \sum_{i=1}^{n_N} \log f_{\text{obs}}(X_i \mid \theta, S) \quad \text{and} \quad \ell_{m_N}(\theta \mid S) := \sum_{j=1}^{m_N} \log f_{\text{int}}(Y_j \mid \theta, S). \tag{54}$$

Since the models satisfy the Laplace regularity conditions described in Lemma A.6, the marginal likelihood can be approximated as follows as $N \to \infty$,

$$\begin{aligned}
\log f_N(\mathcal{D}_{\text{mix}} \mid S^{1*}) = &\left(\ell_{n_N}(\hat{\theta}^1 \mid S^{1*}) + \ell_{m_N}(\hat{\theta}^1 \mid S^{1*})\right) + \frac{3}{2}\log 2\pi \\
&- \frac{1}{2}\log\det\left(\nabla^2 \ell_{n_N}(\hat{\theta}^1 \mid S^{1*}) + \nabla^2 \ell_{m_N}(\hat{\theta}^1 \mid S^{1*})\right) \\
&+ \log\left(\pi(\hat{\theta}^1 \mid S^{1*}) + O\left(\frac{1}{N}\right)\right) \text{ a.s.}
\end{aligned} \tag{55}$$

Similarly for the remaining models S^k ($k = 2, 3$) we have

$$\begin{aligned}
\log f_N(\mathcal{D}_{\text{mix}} \mid S^k) = &\left(\ell_{n_N}(\hat{\theta}^k \mid S^k) + \ell_{m_N}(\hat{\theta}^k \mid S^k)\right) + \frac{d_k}{2}\log 2\pi \\
&- \frac{1}{2}\log\det\left(\nabla^2 \ell_{n_N}(\hat{\theta}^k \mid S^k) + \nabla^2 \ell_{m_N}(\hat{\theta}^k \mid S^k)\right) \\
&+ \log\left(\pi(\hat{\theta}^k \mid S^k) + O\left(\frac{1}{N}\right)\right) \text{ a.s.}
\end{aligned}$$

where d_k denotes the corresponding dimensions of the parameter spaces, i.e. $d_2 = 3$ and $d_3 = 2$ (there are only 2 parameters for the independence model because w is set to 0). Normalizing by $1/N$ the log-marginal likelihood ratio of S^{1*} and S^2 we obtain, as $N \to \infty$

$$\frac{1}{N}\log\frac{f_N(\mathcal{D}_{\text{mix}} \mid S^{1*})}{f_N(\mathcal{D}_{\text{mix}} \mid S^2)} = \eta_N\left[\frac{1}{n_N}\left(\ell_{n_N}(\hat{\theta}^1 \mid S^{1*}) - \ell_{n_N}(\hat{\theta}^2 \mid S^2)\right)\right] + (1-\eta_N)\left[\frac{1}{m_N}\left(\ell_{m_N}(\hat{\theta}^1 \mid S^{1*}) - \ell_{m_N}(\hat{\theta}^2 \mid S^2)\right)\right] + O\left(\frac{\log N}{N}\right) \text{ a.s.} \tag{56}$$

Since both $\ell_n(\theta \mid S)$ and $\ell_m(\theta \mid S)$ are continuous for all $\theta \in \Theta$, then, by the continuous mapping theorem and the strong law of large numbers, the log-marginal likelihood odds converge to a weighted sum of relative entropies

$$\frac{1}{N}\log\frac{f_N(\mathcal{D}_{\text{mix}} \mid S^{1*})}{f_N(\mathcal{D}_{\text{mix}} \mid S^2)} \xrightarrow{a.s.} \eta D(P^X_{m(\theta^*,S^{1*})} \| P^X_{m(\theta^2_\infty,S^2)}) + \bar{\eta} D(P^Y_{m(\theta^*,S^{1*})} \| P^Y_{m(\theta^2_\infty,S^2)}) =: D_{12}, \tag{57}$$

and similarly

$$\frac{1}{N}\log\frac{f_N(\mathcal{D}_{\text{mix}} \mid S^{1*})}{f_N(\mathcal{D}_{\text{mix}} \mid S^3)} \xrightarrow{a.s.} \eta D(P^X_{m(\theta^*,S^{1*})} \| P^X_{m(\theta^3_\infty,S^3)}) + \bar{\eta} D(P^Y_{m(\theta^*,S^{1*})} \| P^Y_{m(\theta^3_\infty,S^3)}) =: D_{13}, \tag{58}$$

where for the first structure we have

$$P^X_{m(\theta^*,S^{1*})} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}w^{*2}\tau_2^{*2}+\tau_1^{*2} & w^*\tau_2^{*2}\\ w^*\tau_2^{*2} & \tau_2^{*2}\end{bmatrix}\right),$$
$$P^Y_{m(\theta^*,S^{1*})} = N\left(w^*y, \tau_1^{*2}\right),$$

for the second structure,

$$P^X_{m(\theta^*,S^{2*})} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{1,\infty|S^2} & w_{\infty|S^2}\tau^2_{1,\infty|S^2}\\ w_{\infty|S^2}\tau^2_{1,\infty|S^2} & w^2_{\infty|S^2}\tau^2_{1,\infty|S^2}+\tau^2_{2,\infty|S^2}\end{bmatrix}\right),$$
$$P^Y_{m(\theta^*,S^{2*})} = N\left(0, \tau_1^{*2}\right),$$

and for the third structure, the limiting distributions are

$$P^X_{m(\theta^3_\infty,S^3)} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{1,\infty|S^3} & 0\\ 0 & \tau^2_{1,\infty|S^3}\end{bmatrix}\right),$$
$$P^Y_{m(\theta^3_\infty,S^3)} = N\left(0, \tau^2_{1,\infty|S^3}\right).$$

with θ^2_∞, θ^3_∞ defined in (52)- (53) and D_{12} in (51). Next, computing the Gaussian relative entropies, the difference between the two exponents is

$$D_{12} - D_{13} = \frac{1}{2}\log\tau_1^{*2} - \frac{1-\eta}{2}\log\tau_1^{*2} - \frac{1-\eta}{2}\log(w^{*2}\tau_2^{*2}+\tau_1^{*2}) = -\frac{\eta}{2}\log\frac{w^{*2}\tau_2^{*2}+\tau_1^{*2}}{\tau_1^{*2}} < 0,$$

Since the posterior probability of S^{1*} is given by the following formula,

$$\pi(S^{1*} \mid \mathcal{D}_{\text{mix}}) = \frac{1}{1 + \frac{f_N(\mathcal{D}_{\text{mix}}|S^2)}{f_N(\mathcal{D}_{\text{mix}}|S^{1*})} + \frac{f_N(\mathcal{D}_{\text{mix}}|S^3)}{f_N(\mathcal{D}_{\text{mix}}|S^{1*})}}$$

and D_{12} and D_{13} are strictly positive (within each model, the parameters are identifiable) with $D_{12} < D_{13}$ the log-posterior ratio for the true model behaves as follows

$$\frac{1}{N} \log \left(\frac{1}{\pi(S^{1*} \mid \mathcal{D}_{\text{mix}})} - 1 \right) \xrightarrow{a.s.} -\min(D_{12}, D_{13}) = -D_{12}.$$

Thus, $\pi(S^{1*} \mid \mathcal{D}_{\text{mix}}) \xrightarrow{a.s.} 1$. □

F PROOF OF THEOREM E.1: S^{2*} THE CORRECT MODEL

Theorem F.1: Let $\mathcal{D}_{\text{obs}} = (X_i)_{i=1}^{n_N}$ be an i.i.d. observational sample generated from (1), and $\mathcal{D}_{\text{int}} = (Y_i)_{i=1}^{m_N}$ be an i.i.d. interventional sample generated under the intervention $do(X(2) = y)$. Assume the priors $\pi(\theta \mid S)$ are four times differentiable for all $S \in \mathcal{S}$. Under the true generating model $\mathcal{M}^2 = \{S^{2*}, \theta^*\}$, where $X \sim P^X_{m(S^{2*},\theta^*)}$ and $Y \sim P^Y_{m(S^{2*},\theta^*)}$ with $\theta^* = [w^*, \tau_1^{*2}, \tau_2^{*2}]^\intercal$, the posterior probability satisfies, as $N \to \infty$,

$$\frac{1}{N} \log \left(\frac{1}{\pi(S^{2*} \mid \mathcal{D}_{\text{mix}})} - 1 \right) \xrightarrow{a.s.} -D_{21}(\eta),$$

where $D_{21}(\eta)$ is defined as

$$D_{21}(\eta) := \frac{1}{2} \log \left[\left(1 - \frac{\eta^2 w^{*2} \tau_1^{*2}}{\eta \sigma^2_{2,Y} + \bar{\eta} y^2} \right) \left(\frac{\sigma^2_{2,Y}}{\tau_2^{*2}} \right)^{\eta} \right], \tag{59}$$

with the interventional variance given by $\sigma^2_{2,Y} := w^{*2}\tau_1^{*2} + \tau_2^{*2}$.

Proof. We proceed analogously to the proof of Theorem E.1. Given that the covariance matrix of X exists and the intervention value $y \in \mathbb{R}$ is fixed, all relevant second moments are finite. Let $\bar{\eta} := 1 - \eta$.

Under the true causal structure $X(1) \to X(2)$, the intervention on the child node $X(2)$ does not alter the marginal distribution of the parent $X(1)$. Therefore, $\mathbb{E}[Y(1)Y(2)] = \mathbb{E}[X(1)]y = 0$ and $\mathbb{E}[Y(1)^2] = \tau_1^{*2}$. Furthermore, $\mathbb{E}[X(1)^2] = \tau_1^{*2}$, $\mathbb{E}[X(2)^2] = w^{*2}\tau_1^{*2} + \tau_2^{*2}$ and $\mathbb{E}[Y(1)^2] = \tau_1^{*2}$. We establish the consistency of the MLEs under the true generating model S^{2*}. Applying the results from Lemma A.4 and A.5 we obtain

$$\hat{w} \xrightarrow{a.s.} \frac{\eta w^* \tau_1^{*2}}{\eta \tau_1^{*2} + \bar{\eta} \tau_1^{*2}} = w^*. \tag{60}$$

Similarly, the variance estimators satisfy

$$\hat{\tau}_1^2 = \frac{S_1^X + S_1^Y}{N} \xrightarrow{a.s.} \eta \tau_1^{*2} + \bar{\eta} \tau_1^{*2} = \tau_1^{*2},$$
$$\hat{\tau}_2^2 = \frac{S_2^X + S_2^Y - \hat{w}^2 (S_{12}^X + S_{12}^Y)}{N} \xrightarrow{a.s.} \tau_2^{*2}$$

which implies that $\hat{\theta}^2 \xrightarrow{a.s.} \theta^*$. Next, we analyze the convergence of the MLE for the incorrect reverse model S^1. Given the data generating process S^2, the estimators converge to a pseudo-true parameter vector $\hat{\theta}^1 \xrightarrow{a.s.} \theta^1_\infty$

$$\begin{aligned} \theta^1_\infty &:= [w_{\infty|S^1}, \tau^2_{1,\infty|S^1}, \tau^2_{2,\infty|S^1}]^\intercal \\ &= \left[\frac{\eta w^* \tau_1^{*2}}{\eta(w^{*2}\tau_1^{*2} + \tau_2^{*2}) + \bar{\eta} y^2}, \quad \frac{\eta \tau_1^{*2} \tau_2^{*2} + \eta \bar{\eta} w^{*2} \tau_1^{*4} + \bar{\eta} y^2 \tau_1^{*2}}{\eta(w^{*2}\tau_1^{*2} + \tau_2^{*2}) + \bar{\eta} y^2}, \quad w^{*2}\tau_1^{*2} + \tau_2^{*2} \right]^\intercal. \end{aligned} \tag{61}$$

Notably, unlike the purely observational case, the presence of interventional data ($\eta < 1$) affects both the estimated weight and the variance of the first node, breaking the symmetry with $\gamma^{-1}(\theta^*)$. For $\eta \to 1$, the above expression recovers the observational limit. For the independence model S^3, the MLE converges to:

$$\hat{\theta}^3 \xrightarrow{a.s.} [0, \tau_1^{*2}, w^{*2}\tau_1^{*2} + \tau_2^{*2}]^\intercal = [0, \tau^2_{1,\infty|S^3}, \tau^2_{2,\infty|S^3}]^\intercal =: \theta^3_\infty. \tag{62}$$

Following the same steps as in (56) – (58), the normalized log-marginal likelihood ratio converge almost surely to the weighted relative entropies

$$\frac{1}{N}\log\frac{f_N(\mathcal{D}_{\text{mix}} \mid S^{2*})}{f_N(\mathcal{D}_{\text{mix}} \mid S^1)} \xrightarrow{a.s.} \eta D(P^X_{m(\theta^*,S^{2*})} \| P^X_{m(\theta^1_\infty,S^1)}) + \bar{\eta} D(P^Y_{m(\theta^*,S^{2*})} \| P^Y_{m(\theta^1_\infty,S^1)}) =: D_{21},$$
$$\frac{1}{N}\log\frac{f_N(\mathcal{D}_{\text{mix}} \mid S^{2*})}{f_N(\mathcal{D}_{\text{mix}} \mid S^3)} \xrightarrow{a.s.} \eta D(P^X_{m(\theta^*,S^{2*})} \| P^X_{m(\theta^3_\infty,S^3)}) + \bar{\eta} D(P^Y_{m(\theta^*,S^{2*})} \| P^Y_{m(\theta^3_\infty,S^3)}) =: D_{23},$$

where for the first structure, the asymptotic distributions are

$$P^X_{m(\theta^1_\infty,S^1)} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} w^2_{\infty|S^1}\tau^2_{2,\infty|S^1} + \tau^2_{1,\infty|S^1} & w_{\infty|S^1}\tau^2_{2,\infty|S^1} \\ w_{\infty|S^1}\tau^2_{2,\infty|S^1} & \tau^2_{2,\infty|S^1}\end{bmatrix}\right),$$
$$P^Y_{m(\theta^1_\infty,S^1)} = N\left(w_{\infty|S^1} y, \tau^2_{1,\infty|S^1}\right),$$

for the second structure, we have

$$P^X_{m(\theta^*,S^{2*})} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} \tau^{*2}_1 & w^*\tau^{*2}_2 \\ w^*\tau^{*2}_2 & w^{*2}\tau^{*2}_1 + \tau^{*2}_2\end{bmatrix}\right),$$
$$P^Y_{m(\theta^*,S^{2*})} = N\left(0, \tau^{*2}_1\right),$$

and for the third structure, the distributions are given by

$$P^X_{m(\theta^3_\infty,S^3)} = N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} \tau^2_{1,\infty|S^3} & 0 \\ 0 & \tau^2_{1,\infty|S^3}\end{bmatrix}\right),$$
$$P^Y_{m(\theta^3_\infty,S^3)} = N\left(0, \tau^2_{1,\infty|S^3}\right).$$

with θ^1_∞ and θ^3_∞ defined in (61) and (62), respectively. Evaluating the Gaussian relative entropies yields

$$D_{21} = \frac{1}{2}\log\left(1 - \frac{\eta^2 w^{*2}\tau^{*2}_1}{\eta(w^{*2}\tau^{*2}_1 + \tau^{*2}_2) + \bar{\eta}y^2}\right) + \frac{\eta}{2}\log\left(1 + \frac{w^{*2}\tau^{*2}_1}{\tau^{*2}_2}\right),$$
$$D_{23} = \frac{\eta}{2}\log\left(1 + \frac{w^{*2}\tau^{*2}_1}{\tau^{*2}_2}\right).$$

Comparing the two rates, we observe that

$$D_{21} = D_{23} + \frac{1}{2}\log\left(1 - \frac{\eta^2 w^{*2}\tau^{*2}_1}{\eta(w^{*2}\tau^{*2}_1 + \tau^{*2}_2) + \bar{\eta}y^2}\right).$$

Since the term inside the logarithm is strictly less than 1 because $\eta^2 w^{*2}\tau^{*2}_1 < \eta w^{*2}\tau^{*2}_1$ for any $\eta \in (0,1)$, the log term is negative, implying $D_{21} < D_{23}$. This indicates that the reverse model S^1 is "closer" to the true distribution than the independence model S^3. Therefore, for any $\eta > 0$, the posterior $\pi(S^{2*} \mid \mathcal{D}_{\text{mix}})$ concentrates to 1 at an exponential rate given by the minimum relative entropy

$$\frac{1}{N}\log\left(\frac{1}{\pi(S^{2*} \mid \mathcal{D}_{\text{mix}})} - 1\right) \xrightarrow{a.s.} -\min(D_{21}, D_{23}) = -D_{21}.$$

□

G PROOF OF THEOREM 4.2

Proof. If S^{3*} is the true generating model, then it is easy to see from Lemmas A.4 and A.5 that the MLEs of all three models converge almost surely to the same limit

$$\hat{\theta}^1, \hat{\theta}^2, \hat{\theta}^3 \xrightarrow{a.s.} [0, \tau_1^{*2}, \tau_2^{*2}]^\intercal = \theta^*$$

because $\frac{1}{n} S_{12}^X \xrightarrow{a.s.} \mathbb{E}[X(1)X(2)] = \mathbb{E}[X(1)]\mathbb{E}[X(2)] = 0$ and $\frac{1}{n} S_{12}^Y \xrightarrow{a.s.} y\mathbb{E}[Y(1)] = 0$. Since the models are all Laplace regular from Lemma A.6, the marginal likelihoods can be approximated for $N \to \infty$ as in (55)

$$\begin{aligned}\log f_N(\mathcal{D}_{\text{mix}} \mid S^1) = &\left(\ell_{n_N}(\hat{\theta}^1 \mid S^1) + \ell_{m_N}(\hat{\theta}^1 \mid S^1)\right) + \frac{3}{2}\log 2\pi \\ &- \frac{1}{2}\log\det\left(\nabla^2\ell_{n_N}(\hat{\theta}^1 \mid S^{1*}) + \nabla^2\ell_{m_N}(\hat{\theta}^1 \mid S^1)\right) + \log\left(\pi(\hat{\theta}^1 \mid S^1) + O\left(\frac{1}{N}\right)\right) a.s.\end{aligned}$$

Next, by the continuous mapping theorem and the strong law of large numbers, we have that the two Hessians weighted by $\frac{1}{n_N}$ and $\frac{1}{m_N}$, respectively, converge almost surely to their respective negative Fisher information matrices as $N \to \infty$, i.e.

$$\begin{aligned}-\frac{1}{n_N}\nabla^2\ell_n(\theta \mid S^1) &\xrightarrow{a.s.} Diag\left(\frac{\tau_2^2}{\tau_1^2}, \frac{1}{2\tau_1^4}, \frac{1}{2\tau_2^4}\right) =: I_1^X([w, \tau_1^2, \tau_2^2]^\intercal) \\ -\frac{1}{m_N}\nabla^2\ell_m(\theta \mid S^1) &\xrightarrow{a.s.} Diag\left(\frac{y^2}{\tau_1^2}, \frac{1}{2\tau_1^4}, 0\right) =: I_1^Y([w, \tau_1^2, \tau_2^2]^\intercal)\end{aligned}$$

For completeness, we provide the Fisher information matrices for the remaining two models

$$I_2^X([w, \tau_1^2, \tau_2^2]^\intercal) := Diag\left(\frac{\tau_1^2}{\tau_2^2}, \frac{1}{2\tau_2^4}, \frac{1}{2\tau_1^4}\right) \quad \text{and} \quad I_2^Y([w, \tau_1^2, \tau_2^2]^\intercal) := Diag\left(0, \frac{1}{2\tau_2^4}, 0\right)$$

for the second structure S^2 and

$$I_3^X([0, \tau_1^2, \tau_2^2]^\intercal) := Diag\left(\frac{1}{2\tau_1^4}, \frac{1}{2\tau_2^4}\right) \quad \text{and} \quad I_3^Y([0, \tau_1^2, \tau_2^2]^\intercal) := \left(\frac{1}{2\tau_1^4}, 0\right)$$

for the third structure, S^3. Therefore, by the continuous mapping theorem

$$\begin{aligned}\frac{1}{N^3}\det\left(\nabla^2\ell_{n_N}(\hat{\theta}^1 \mid S^1) + \nabla^2\ell_{m_N}(\hat{\theta}^1 \mid S^1)\right) &= \det\left(\eta_N\frac{1}{n_N}\nabla^2\ell_{n_N}(\hat{\theta}^1 \mid S^1) + \bar{\eta}_N\frac{1}{m_N}\nabla^2\ell_{m_N}(\hat{\theta}^1 \mid S^1)\right) \\ &\xrightarrow{a.s.} \det\left(\eta I_1^X(\theta^*) + \bar{\eta} I_1^Y(\theta^*)\right) = \frac{\eta(\eta\tau_2^{*2} + \bar{\eta}y^2)}{4\tau_1^{*6}\tau_2^{*4}},\end{aligned} \tag{63}$$

$$\begin{aligned}\frac{1}{N^3}\det\left(\nabla^2\ell_{n_N}(\hat{\theta}^2 \mid S^2) + \nabla^2\ell_{m_N}(\hat{\theta}^2 \mid S^2)\right) &= \det\left(\eta_N\frac{1}{n_N}\nabla^2\ell_{n_N}(\hat{\theta}^2 \mid S^2) + \bar{\eta}_N\frac{1}{m_N}\nabla^2\ell_{m_N}(\hat{\theta}^2 \mid S^2)\right) \\ &\xrightarrow{a.s.} \det\left(\eta I_2^X(\theta^*) + \bar{\eta} I_2^Y(\theta^*)\right) = \frac{\eta^2}{4\tau_2^{*6}\tau_1^{*2}}\end{aligned} \tag{64}$$

and for the third model, S^{3*}

$$\begin{aligned}\frac{1}{N^2}\det\left(\nabla^2\ell_{n_N}(\hat{\theta}^3 \mid S^{3*}) + \nabla^2\ell_{m_N}(\hat{\theta}^3 \mid S^{3*})\right) &= \det\left(\eta_N\frac{1}{n_N}\nabla^2\ell_{n_N}(\hat{\theta}^3 \mid S^{3*}) + \bar{\eta}_N\frac{1}{m_N}\nabla^2\ell_{m_N}(\hat{\theta}^3 \mid S^{3*})\right) \\ &\xrightarrow{a.s.} \det\left(\eta I_3^X(\theta^*) + \bar{\eta} I_3^Y(\theta^*)\right) = \frac{\bar{\eta}}{4\tau_2^{*4}\tau_1^{*4}}.\end{aligned} \tag{65}$$

Similar to the proof of Theorem 3.2, the maximum log-likelihood difference between models S^1 and S^3 is

$$\ell_{n_N, m_N}(\hat{\theta}^3 \mid S^{3*}) - \ell_{n,m}(\hat{\theta}^1 \mid S^1) = \frac{n_N + m_N}{2}\log\left(1 - \frac{(S_{12}^X + S_{12}^Y)^2}{(S_1^X + S_1^Y)(S_2^X + S_2^Y)}\right) + \frac{n_N}{2}\log\left(1 - \frac{(S_{12}^X)^2}{S_1^X S_2^X}\right) \tag{66}$$

with both terms in the sum converging in distribution to a χ^2_1, i.e.

$$-(n_N + m_N)\log\left(1 - \frac{(S^X_{12} + S^Y_{12})^2}{(S^X_1 + S^Y_1)(S^X_2 + S^Y_2)}\right) \xrightarrow{d} \chi^2_1$$

and

$$-n_N \log\left(1 - \frac{(S^X_{12})^2}{S^X_1 S^X_2}\right) \xrightarrow{d} \chi^2_1$$

as per the same steps, (48)– (50), in the proof of Theorem 3.2. Next, the log-likelihood difference converges in distribution to a weighted χ^2_1 distribution, and hence is bounded in probability, i.e. it is $O_p(1)$. Similarly, the log-likelihood difference between models S^2 and S^{3*} is expressed as

$$2\left(\ell_{n_N,m_N}(\hat{\theta}^3 \mid S^3) - \ell_{n_N,m_N}(\hat{\theta}^2 \mid S^2)\right) = \frac{n_N}{2}\log\left(1 - \frac{(S^X_{12})^2}{S^X_1 S^X_2}\right) \xrightarrow{d} -\frac{1}{2}\chi^2_1, \tag{67}$$

which again is bounded in probability, $O_p(1)$. Then, using (66)–(67), the weighted log-marginal likelihood odds between models S^1 vs S^{3*} and S^2 vs S^{3*} are

$$\sqrt{N}\log\frac{f_N(\mathcal{D}_{\text{mix}} \mid S^3)}{f_N(\mathcal{D}_{\text{mix}} \mid S^1)} \xrightarrow{d} \Lambda_1, \text{ and } \sqrt{N}\log\frac{f_N(\mathcal{D}_{\text{mix}} \mid S^3)}{f_N(\mathcal{D}_{\text{mix}} \mid S^2)} \xrightarrow{d} \Lambda_2$$

with Λ_1 and Λ_2 being both $O_p(1)$, bounded in probability. Therefore, the posterior probability of the third model, S^{3*}, concentrates to 1 asymptotically at a rate of $O_p(1/\sqrt{N})$

$$\sqrt{N}(\pi(S^{3*} \mid \mathcal{D}_{\text{mix}}) - 1) \xrightarrow{d} -(\Lambda_1 + \Lambda_2).$$

□

H EXACT CALCULATION OF THE POSTERIOR

Theorem H.1 (Posterior distribution for BGe-type priors)**:** Let $\mathcal{D}_{\text{obs}} = (X_i)_{i=1}^n$ be and i.i.d. observational sample generated by (1). Given the hierarchical priors defined in (11), the posterior distribution over the structures, $[\pi(S^1 \mid \mathcal{D}_{\text{obs}}), \pi(S^2 \mid \mathcal{D}_{\text{obs}}), \pi(S^3 \mid \mathcal{D}_{\text{obs}})]^\intercal$, is

$$\left[\begin{array}{l}
\frac{(2\beta)^{\alpha_1+\alpha_2}\Gamma(\alpha_2+\frac{n}{2})\Gamma(\alpha_1+\frac{n}{2})}{\sqrt{\eta}\pi^n\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot \frac{(S_2^{X\lambda})^{\alpha_1+(n-1)/2}}{(S_2^{X\beta})^{\alpha_2+n/2}} \cdot \frac{1}{(S_1^{X\beta}S_2^{X\lambda} - (S^X_{12})^2)^{\alpha_1+n/2}}, \\
\frac{(2\beta)^{\alpha_3+\alpha_4}\Gamma(\alpha_3+\frac{n}{2})\Gamma(\alpha_4+\frac{n}{2})}{\sqrt{\eta}\pi^n\Gamma(\alpha_3)\Gamma(\alpha_4)} \cdot \frac{(S_1^{X\lambda})^{\alpha_4+(n-1)/2}}{(S_1^{X\beta})^{\alpha_3+n/2}} \cdot \frac{1}{(S_1^{X\lambda}S_2^{X\beta} - (S^X_{12})^2)^{\alpha_4+n/2}}, \\
\frac{(2\beta)^{\alpha_5+\alpha_6}\Gamma(\alpha_5+\frac{n}{2})\Gamma(\alpha_6+\frac{n}{2})}{\pi^n\Gamma(\alpha_5)\Gamma(\alpha_6)} \cdot \frac{1}{(S_1^{X\beta})^{\alpha_5+n/2}(S_2^{X\beta})^{\alpha_6+n/2}}
\end{array}\right]^\intercal,$$

where

$$S_i^{X\lambda} := S_i^X + \frac{1}{\lambda}, \quad S_i^{X\beta} := S_i^X + 2\beta \quad \text{for } i \in \{1,2\}, \tag{68}$$

and S_1^X, S_2^X, and S_{12}^X are defined in (27).

PROOF. From Bayes' rule, the marginal posterior distribution is

$$\pi(S^1 \mid \mathcal{D}_{\text{obs}}) \propto \int f_n(\mathcal{D}_{\text{obs}} \mid S^1, w, \tau_1^2, \tau_2^2) f(w \mid \tau_1^2) f(\tau_1^2) f(\tau_2^2)\, d(\tau_1^2)\, d(\tau_2^2)\, dw$$

which yields

$$\begin{aligned}
\pi(S^1 \mid \mathcal{D}_{\text{obs}}) &\propto \int f_n(\mathcal{D}_{\text{obs}} \mid S^1, w, \tau_1^2, \tau_2^2) f(w \mid \tau_1^2) f(\tau_1^2) f(\tau_2^2)\, d(\tau_1^2)\, d(\tau_2^2)\, dw \\
&= \int \frac{1}{\left(\sqrt{2\pi\tau_1^2}\right)^n} \exp\left(-\frac{1}{2\tau_1^2}\sum_{i=1}^{n}(X_i(1) - wX_i(2))^2\right) \\
&\quad \times \frac{1}{\left(\sqrt{2\pi\tau_2^2}\right)^n} \exp\left(-\frac{1}{2\tau_2^2} S_2^X\right) \\
&\quad \times \frac{1}{\sqrt{2\pi\lambda\tau_1^2}} \exp\left(-\frac{w^2}{2\lambda\tau_1^2}\right) \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} (\tau_1^2)^{-\alpha_1-1} \exp\left(-\frac{\beta}{\tau_1^2}\right) \\
&\quad \times (\tau_2^2)^{-\alpha_2-1} \exp\left(-\frac{\beta}{\tau_2^2}\right) d(\tau_1^2)\, d(\tau_2^2)\, dw \\
&= \frac{1}{\sqrt{\lambda}} \frac{1}{(2\pi)^{n+1/2}} \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_2 + \frac{n}{2})}{\left(\beta + \frac{S_2^X}{2}\right)^{\alpha_2+n/2}} \\
&\quad \times \int_{-\infty}^{\infty} \frac{\Gamma(\alpha_1 + \frac{n+1}{2})}{\left(\beta + \frac{\sum_{i=1}^n (X_1^i - wX_2^i)^2}{2} + \frac{w^2}{2\lambda}\right)^{\alpha_1 + \frac{n+1}{2}}} dw.
\end{aligned}$$

Using the result of Lemma A.7 we obtain

$$\begin{aligned}
\pi(S^1 \mid \mathcal{D}_{\text{obs}}) &\propto \frac{(2\beta)^{\alpha_1+\alpha_2}\Gamma(\alpha_2 + \frac{n}{2})\Gamma(\alpha_1 + \frac{n}{2})}{\sqrt{\lambda}\pi^n\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot \frac{(1/\lambda + S_2^X)^{\alpha_1+(n-1)/2}}{(2\beta + S_2^X)^{\alpha_2+n/2}} \\
&\quad \cdot \frac{1}{[(2\beta + S_1^X)(1/\lambda + S_2^X) - (S_{12}^X)^2]^{\alpha_1+n/2}} \qquad (69) \\
&= \frac{(2\beta)^{\alpha_1+\alpha_2}\Gamma(\alpha_2 + \frac{n}{2})\Gamma(\alpha_1 + \frac{n}{2})}{\sqrt{\lambda}\pi^n\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot \frac{(S_2^{X\lambda})^{\alpha_1+(n-1)/2}}{(S_2^{X\beta})^{\alpha_2+n/2}} \\
&\quad \cdot \frac{1}{(S_1^{X\beta} S_2^{X\lambda} - (S_{12}^X)^2)^{\alpha_1+n/2}}, \qquad (70)
\end{aligned}$$

Similarly, the posterior probability of the second causal structure, S^2, is proportional to

$$\begin{aligned}
\pi(S^2 \mid \mathcal{D}_{\text{obs}}) &\propto \frac{(2\beta)^{\alpha_3+\alpha_4}\Gamma(\alpha_3 + \frac{n}{2})\Gamma(\alpha_4 + \frac{n}{2})}{\sqrt{\lambda}\pi^n\Gamma(\alpha_3)\Gamma(\alpha_4)} \cdot \frac{(S_1^{X\lambda})^{\alpha_4+(n-1)/2}}{(S_1^{X\beta})^{\alpha_3+n/2}} \\
&\quad \cdot \frac{1}{(S_1^{X\lambda} S_2^{X\beta} - (S_{12}^X)^2)^{\alpha_4+n/2}}, \qquad (71)
\end{aligned}$$

and

$$\begin{aligned}
\pi(S^3 \mid \mathcal{D}_{\text{obs}}) &\propto \int f_n(\mathcal{D}_{\text{obs}} \mid S^3, \tau_1^2, \tau_2^2) f(\tau_1^2) f(\tau_2^2)\, d(\tau_1^2)\, d(\tau_2^2) \\
&= \int \frac{1}{(\sqrt{2\pi\tau_1^2})^n} \exp\left(-\frac{1}{2\tau_1^2} S_1^X\right) \frac{1}{(\sqrt{2\pi\tau_2^2})^n} \exp\left(-\frac{1}{2\tau_2^2} S_2^X\right) \\
&\quad \times \frac{\beta^{\alpha_5+\alpha_6}}{\Gamma(\alpha_5)\Gamma(\alpha_6)} (\tau_1^2)^{-\alpha_5-1} \exp\left(-\frac{\beta}{\tau_1^2}\right) (\tau_2^2)^{-\alpha_6-1} \exp\left(-\frac{\beta}{\tau_2^2}\right) d(\tau_1^2)\, d(\tau_2^2) \\
&= \frac{(2\beta)^{\alpha_5+\alpha_6}\Gamma(\alpha_5 + \frac{n}{2})\Gamma(\alpha_6 + \frac{n}{2})}{\pi^n\Gamma(\alpha_5)\Gamma(\alpha_6)} \cdot \frac{1}{(S_1^{X\beta})^{\alpha_5+n/2}(S_2^{X\beta})^{\alpha_6+n/2}}. \qquad (72)
\end{aligned}$$

□

Theorem H.2 (Posterior distribution for BGe-type priors)**:** Let $\mathcal{D}_{\text{obs}} = (X_i)_{i=1}^n$ be an i.i.d. observational sample, $\mathcal{D}_{\text{int}} = (Y_i)_{i=1}^m$ be an i.i.d. interventional sample generated under the intervention $do(Y(2) = y)$ for a fixed $y \in \mathbb{R}$, and let

$\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{int}}$. Given the hierarchical priors defined in (11), the posterior distribution over the structures, $[\pi(S^1 \mid \mathcal{D}_{\text{mix}}), \pi(S^2 \mid \mathcal{D}_{\text{mix}}), \pi(S^3 \mid \mathcal{D}_{\text{mix}})]^\intercal$, is given by

$$\frac{1}{c}\Bigg[\frac{(2\beta)^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\frac{m+n}{2})\Gamma(\alpha_2+\frac{n}{2})}{\sqrt{\eta}\pi^{n+\frac{m+1}{2}}\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{(S_2^{X\lambda}+S_2^Y)^{\alpha_1+\frac{m+n-1}{2}}}{[(S_1^{X\beta}+S_1^Y)(S_2^{X\lambda}+S_2^Y)-(S_{12}^X+S_{12}^Y)^2]^{\alpha_1+\frac{m+n}{2}}(S_2^{X\beta})^{\alpha_2+\frac{n}{2}}},$$
$$\frac{(2\beta)^{\alpha_3+\alpha_4}\Gamma(\alpha_3+\frac{m+n}{2})\Gamma(\alpha_4+\frac{n}{2})}{\sqrt{\eta}\pi^{n+\frac{m+1}{2}}\Gamma(\alpha_3)\Gamma(\alpha_4)}\frac{(S_1^{X\lambda})^{\alpha_4+\frac{n-1}{2}}}{(S_1^{X\lambda}S_2^{X\beta}-(S_{12}^X)^2)^{\alpha_4+\frac{n}{2}}(S_1^{X\beta}+S_1^Y)^{\alpha_3+\frac{m+n}{2}}},$$
$$\frac{(2\beta)^{\alpha_5+\alpha_6}\Gamma(\alpha_5+\frac{m+n}{2})\Gamma(\alpha_6+\frac{n}{2})}{\pi^{n+\frac{m+1}{2}}\Gamma(\alpha_5)\Gamma(\alpha_6)}\frac{1}{(S_1^{X\beta}+S_1^Y)^{\alpha_5+\frac{m+n}{2}}(S_2^{X\beta})^{\alpha_6+\frac{n}{2}}}\Bigg], \tag{73}$$

where c is the normalization constant, $S_i^{X\beta}, S_i^{X\lambda}$ are defined in (68) and S_i^Y, S_{12}^Y are defined in (28) with $i \in \{1,2\}$.

PROOF. From Bayes' rule, the marginal posterior distribution, $\pi(S^1 \mid \mathcal{D}_{\text{mix}})$ is proportional to

$$\int f_n(\mathcal{D}_{\text{obs}} \mid S^1, w, \tau_1^2, \tau_2^2) f_m(\mathcal{D}_{\text{int}} \mid S^1, w, \tau_1^2, \tau_2^2) f(w \mid \tau_1^2) f(\tau_1^2) f(\tau_2^2)\, d(\tau_1^2)\, d(\tau_2^2)\, dw$$

which yields

$$\begin{aligned}
\pi(S^1 \mid \mathcal{D}_{\text{mix}}) \propto \int & \frac{1}{\left(\sqrt{2\pi\tau_1^2}\right)^n}\exp\left(-\frac{1}{2\tau_1^2}\sum_{i=1}^n (X_i(1)-wX_i(2))^2\right)\frac{1}{\left(\sqrt{2\pi\tau_2^2}\right)^n}\exp\left(-\frac{1}{2\tau_2^2}S_2^X\right)\\
&\times\frac{1}{\left(\sqrt{2\pi\tau_1^2}\right)^m}\exp\left(-\frac{1}{2\tau_1^2}\sum_{i=1}^n (Y_i(1))-wy)^2\right)\\
&\times\frac{1}{\sqrt{2\pi\lambda\tau_1^2}}\exp\left(-\frac{w^2}{2\lambda\tau_1^2}\right)\frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}(\tau_1^2)^{-\alpha_1-1}\exp\left(-\frac{\beta}{\tau_1^2}\right)\\
&\times(\tau_2^2)^{-\alpha_2-1}\exp\left(-\frac{\beta}{\tau_2^2}\right)\,d(\tau_1^2)\,d(\tau_2^2)\,dw\\
=\ & \frac{1}{\sqrt{\lambda}}\frac{1}{(2\pi)^{n+(m+1)/2}}\frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{\Gamma(\alpha_2+\frac{n}{2})}{\left(\beta+\frac{S_2^X}{2}\right)^{\alpha_2+n/2}}\\
&\times\int_{-\infty}^{\infty}\frac{\Gamma(\alpha_1+\frac{n+1}{2})}{\left(\beta+\frac{\sum_{i=1}^n(X_i(1)-wX_i(2))^2+\sum_{j=1}^m(Y_j(1)-wy)^2}{2}+\frac{w^2}{2\lambda}\right)^{\alpha_1+\frac{n+1}{2}}}\,dw.
\end{aligned}$$

Then, using the result of Lemma A.7 we obtain the desired result. The results for the second, S^2, and third model, S^3 follow analogously. □

I FURTHER PLOTS

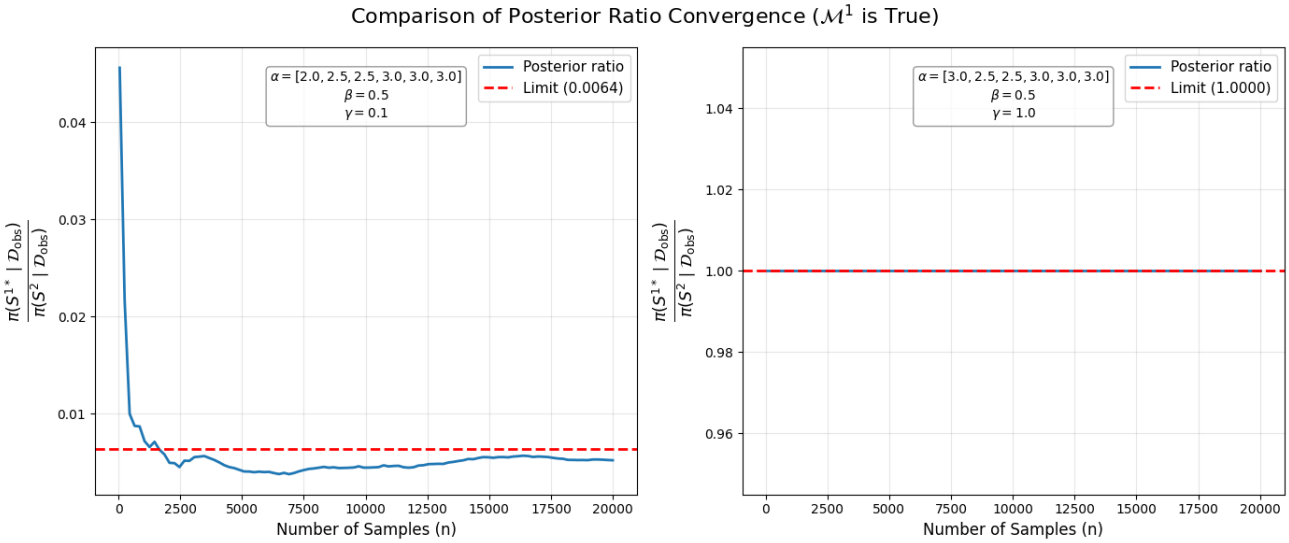

Figure 6: Posterior ratio $\frac{\pi(S^1|\mathcal{D}_{\text{obs}})}{\pi(S^2|\mathcal{D}_{\text{obs}})}$ as a function of sample size n when the true model is S^1. The hyper-parameters in the right plot correspond to the BGe prior.