Bayesian Hierarchical Invariant Prediction

Francisco Madaleno FMFSA@DTU.DK
Pernille Julie Viuff Sand
Francisco C. Pereira
Department of Technology, Management and Economics, Danish Technical University

Sergio Hernan Garrido Mejia
Max Planck Institute for Intelligent Systems

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

We propose Bayesian Hierarchical Invariant Prediction (BHIP) reframing Invariant Causal Prediction (ICP) through the lens of Hierarchical Bayes. We leverage the hierarchical structure to explicitly test invariance of causal mechanisms under heterogeneous data, resulting in improved computational scalability for a larger number of predictors compared to ICP. Moreover, given its Bayesian nature BHIP enables the use of prior information. We evaluate BHIP on both synthetic and real-world datasets, demonstrating its potential as an alternative inference method to ICP and related methods.

Keywords: causality, invariant causal prediction, hierarchical bayes

1. Introduction

Heterogeneous data is ubiquitous in real-world applications, spanning fields such as medicine, transportation, and environmental science. The critical observation of this paper is that the heterogeneous data forming the basis of Bayesian Hierarchical Models (BHM) corresponds to "environment data" described in Invariant Causal Prediction (ICP) (Peters et al., 2016). Both perspectives acknowledge that real-world data often arises from distinct underlying processes or 'environments,' underscoring the need for robust models that account for these differences and ultimately improve out-of-distribution performance. For example, data collected across hospitals, regions, or time periods often exhibit distribution shifts: patient populations and treatment protocols may differ across clinics, and travel behavior may differ across cities or seasons.

Without specific structural assumptions observational data can only allow inferring a Directed Acyclic Graph (DAG) model up to the Markov Equivalence Class (Pearl, 2000). ICP identifies causal predictors by leveraging invariance properties across environments, exhaustively employing conditional independence tests to find sets of predictors for which the target variable is independent of the environment when conditioned on those predictors. As most statistical methods, ICP is sample-size sensitive, but also computationally complex since it involves testing for invariance across all possible subsets of predictors, 2^d where d is the number of predictors. Additionally, ICP is designed to be conservative and minimize Type I errors which often leads to low power (Peters et al., 2016).

Addressing these limitations, we introduce Bayesian Hierarchical Invariant Prediction (BHIP), which models the variability of predictors' effects across environments, quantifies uncertainty, and allows for the incorporation of priors (which we explore in Appendix E with the inclusion of sparsity inducing priors), enhancing flexibility and overall performance compared to frequentist approaches like ICP. While BHIP utilizes established Bayesian techniques, namely hierarchical

modeling, its novelty lies not in the individual components themselves, but in their specific synthesis and application to probabilistically test causal invariance for parent discovery. Specifically, this work contributes a unified probabilistic framework for invariant prediction. Leveraging its hierarchical structure and Bayesian nature, it simultaneously accounts for environmental heterogeneity, allows variable selection (e.g., through sparsity-inducing priors), quantifies uncertainty in estimated effects and, crucially, enables probabilistic invariance testing directly within the model's posterior analysis.

By repurposing the standard hierarchical modeling structure and combining it with specific posterior decision rules, BHIP directly identifies invariant parents in heterogeneous data. This targeted and integrated Bayesian approach represents a significant step beyond simply applying standard Bayesian methods or relying solely on frequentist invariance testing. Code is available at GitHub repository [1].

1.1. Related Work

The problem of **causal feature selection** or causal discovery on a specific target variable has been studied as an alternative to the more difficult problem of full causal discovery. Under Pearl's view of Causality (Pearl, 2000), ICP laid down the foundations for causal feature selection using the idea of invariance. What differentiated ICP from other methods for causal discovery using both interventional and observational data is that ICP does not require the analyst to know where the interventions are performed, but instead only to know from which setting a particular data point comes. This powerful idea was later extended to non-linear models (Heinze-Deml et al., 2018), dynamical systems (Pfister et al., 2019a), time series data (Pfister et al., 2019b), spatio-temporal data (Christiansen et al., 2022) and different outcome models (Kook et al., 2024). Moreover, the invariant principle has been also used together with adjacent machine learning methods: active learning (Gamella and Heinze-Deml, 2020), to find the causal features with experimental data efficiently; and reinforcement learning (Saengkyongam et al., 2023), for policy learning.

Bayesian inference has also been used for causal discovery, (Heckerman et al., 2006) being one of the earliest examples with Monte Carlo inference on causal graphs, all the way to recent advances, such as (Hägele et al., 2023) who use differentiable methods to infer causal structure in a Bayesian way. The theoretical foundations of the Bayesian method have also helped gain insights in causality using the invariance principle in exchangeable data (Guo et al., 2024a,b), and understanding causality in the context of hierarchical models (Weinstein and Blei, 2024).

The relation between hierarchical Bayes and causal discovery is long-standing. (Gelman and Imbens, 2013) talk about the differences between asking about effects of causes (causal inference) and causes of effects (causal discovery) and mention how bayesian, and specifically hierarchical, models can be used for the latter (they present (Manton et al., 1989) as an example).

Concurrently, Wu et al. (2025) develop a Bayesian invariant prediction model that, alike this work, casts ICP-style invariance as posterior inference over invariant features, but differs by encoding the invariant set through a latent feature-selection variable in a joint generative model with consistency and variational-inference guarantees, rather than via hierarchical partial pooling on regression coefficients as in BHIP.

1. https://github.com/fmfsa/bhip

2. Background

2.1. Graphical Models and Causal Inference

In this paper we use SCMs (Peters et al., 2017) to represent causal relations.

Definition 1 (Structural Causal Model (SCM)) *A $D-$dimensional SCM is a tuple $\mathcal{M} := (\mathbb{S}, P_\varepsilon)$ consisting of a set $\mathbb{S}$ of structural assignments*

$$X_d := f_d(\text{PA}(X_d), \varepsilon_d)), \text{ for } d = 1, \ldots, D, \tag{1}$$

where $\text{PA}(X_d) \subseteq \{X_1, \ldots, X_D\} \setminus \{X_d\}$ are called the parents or causes of X_d, and a joint distribution $P_\varepsilon = P_{\varepsilon_1,\ldots,\varepsilon_D}$ over noise variables that we assume to be independent.

We can obtain a causal graph $\mathcal{G}$ from the SCM by drawing a vertex for each X_d and a directed edge from each vertex $X_i \in \text{PA}(X_d)$ to X_d. We assume the obtained causal graph is a DAG. Likewise, even though the causal mechanisms are deterministic functions of its inputs, we obtain a distribution of each of the variables of our system, P_{X_d}, by considering the pushforward distribution of $\text{PA}(X_d)$ and ε_d. We exploit the independent (Janzing and Schölkopf, 2010) and invariant (Aldrich, 1989; Peters et al., 2016) nature of causal mechanisms to identify the causes of a target variable, Y, using heterogeneous data.

2.2. Invariant Causal Prediction

Traditionally, Machine Learning methods assume data is *independent and identically distributed (i.i.d)*, unless specified by a particular structure like time series or graph models. Researchers in causality have found ways of exploiting heterogeneous data to do causal discovery (Cooper and Yoo, 1999; Mooij et al., 2020; Brouillard et al., 2020; Peters et al., 2016). In particular, ICP aims to estimate a target's set of causal parents by leveraging the property that the conditional of Y given its direct causes $\text{PA}(Y)$ remains invariant across heterogeneous data collection, assuming the causal mechanism of Y is not affected (Peters et al., 2016).

Throughout this work, the discrete heterogeneous contexts are called *environments* $e \in \mathcal{E}$, where $\mathcal{E}$ is an index set, each consist of i.i.d data of a *target* variable of interest Y and D covariates (or predictors) $\mathbf{X} = (X_1, \ldots, X_D)$, that is,

$$(Y^e, \mathbf{X}^e) \text{ for } e \in \mathcal{E}. \tag{2}$$

ICP is based on a fundamental invariance assumption, along with the notion of an *invariant set of predictors*, formulated as follows.

Assumption 1 *(Invariance) Given environments $\mathcal{E}$, for a subset of indices $S \subseteq \{1, \ldots, D\}$, there exists a subset X_S of covariates such that,*

$$Y^e \mid (X_S^e = x) \overset{d}{=} Y^f \mid (X_S^f = x), \tag{3}$$

for all $e, f \in \mathcal{E}$ and all x, that is, the conditionals of Y given X_S are equal in distribution in all environments.

Any subset $S \subseteq \{1, \ldots, D\}$ for which Assumption 1 holds, is called *invariant with respect to* $\mathcal{E}$ and the set of covariates X_S are denoted ***invariant predictors***. The invariance assumption can be equipped with more structure, assuming a system induced by an SCM:

Assumption 2 *(Structural invariance) For an SCM and environments, $\mathcal{E}$, the structural equation for Y remains the same across $\mathcal{E}$, and the distribution of X is allowed to change. That is, for all $e \in \mathcal{E}$,*

$$X^e \sim P_X^e, \tag{4}$$
$$Y^e = f_Y(X^e_{\text{PA}(Y)}, \varepsilon_Y), \quad \varepsilon \sim P_{\varepsilon_Y}, \quad \varepsilon_Y \perp\!\!\!\perp X^e_{\text{PA}(Y)}, \tag{5}$$

X^e represents the covariate X in the environment $e \in \mathcal{E}$, P_ε remains the same and P_X^e can differ.

Under this structure, the set of direct causes $\text{PA}(Y)$ satisfy invariance (Bühlmann, 2018), further explained in Appendix B. ICP tests the invariance of all possible covariate subsets S. The set of *identifiable causal predictors* $\hat{S}$ is the intersection of the subsets that pass as invariant. The main theorem in (Peters et al., 2016) states that with a controllable coverage probability, the method recovers the set of true causal parents, that is, with desired $\alpha \in (0, 1)$, we have $\mathbb{P}[\hat{S} \subseteq \text{PA}(Y)] \geq 1 - \alpha$.

The intersection controls against Type I errors and often leads to a conservative estimate, given that in the case where the amount of heterogeneity is insufficient, the intersection is $\hat{S} = \emptyset$. In Section 3.2 we introduce a Bayesian test for invariance based on the BHIP that serves as an alternative to the test proposed in (Peters et al., 2016).

2.3. Bayesian Hierarchical Models

Hierarchical models estimate posterior distributions, accounting for structured heterogeneity (Gelman and Hill, 2007). With heterogeneous data as in Equation (2), a hierarchical model contains both parameters related to variation within each environment, termed local-level parameters or environmental-specific β^e for all $e \in \mathcal{E}$, and parameters related to variation between environments, termed global parameters ϕ.

In BHMs, we impose a distribution on the latent global parameter ϕ; the joint probability over the data and model parameters is then given by the following factorization:

$$P(Y^e, \mathbf{X}^e, \boldsymbol{\beta}^e, \phi) = P(\phi) \prod_{e \in \mathcal{E}} P(\beta^e \mid \phi) P(Y^e \mid \mathbf{X}^e, \beta^e). \tag{6}$$

Bayesian models integrate domain knowledge through priors. With Bayesian inference of the global parameter distribution jointly with local level parameters it is possible to analyze the strength of pooling consistent with the observed data (Betancourt, 2020). We reinterpret this strength of pooling forming an invariant perspective to estimate ICP with a BHM: if, for a specific covariate, the global and local parameters are different from zero and 'close' (see Section 3.2 for a precise definition) to each other, then we can think of them as invariant.

3. Bayesian Hierarchical Invariant Prediction

Consider a dataset comprising of data from a set of environments $\mathcal{E}$ as in Equation (2) induced by an SCM where Assumption 2 holds. The heterogeneity is modeled in a BHM allowing global

parameters to be random while simultaneously modeling individual-level effects. The goal of BHIP is to identify covariates where the individual effects on Y are non-zero and remain the same across environments; that is, the relevant parameters that remain invariant. In addition to Assumption 2, BHIP relies on a set of assumptions common in ICP and Bayesian modeling. A detailed list can be found in Appendix A.

3.1. Probabilistic Model

First, we specify the hierarchical model where we can draw inferences about the influence of each covariate on the target and how it varies between environments. Assume the data generation process of Y^e for a given environment $e \in \mathcal{E}$ is governed by environment-specific parameters $\boldsymbol{\beta}^e = (\beta_1^e, \ldots, \beta_D^e)$, explaining the effects of covariates $\boldsymbol{X}^e = (X_1^e, \ldots, X_D^e)$ on the target Y^e,

$$Y^e \mid \boldsymbol{X}^e, \boldsymbol{\beta}^e \sim P(Y^e \mid \boldsymbol{X}^e, \boldsymbol{\beta}^e), \text{ for } e \in \mathcal{E}. \tag{7}$$

For each covariate index $d \in \{1, \ldots, D\}$, it is assumed that a common global-level distribution with parameters ϕ_d generates a local-level parameter β_d^e with regards to each environment,

$$\beta_d^e \mid \phi_d \sim P(\beta_d^e \mid \phi_d), \text{ for } e \in \mathcal{E}. \tag{8}$$

To obtain a hierarchical model we treat the global parameters as unknown and impose a (hyper)prior distribution on the global parameters:

$$\phi_d \sim P(\phi_d), \text{ for } d = 1, \ldots, D. \tag{9}$$

As an example, consider the following **normal linear hierarchical model**. We assume a Gaussian likelihood for y_n^e. For the local parameters $\boldsymbol{\beta}^e$, we assume each predictor's coefficient β_d^e is drawn from a Gaussian distribution governed by predictor-specific global parameters: a mean μ_d and a scale τ_d. These form the global vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_D)$ [2]. The generative process of BHIP model represented in Figure 1 as a Probabilistic Graphical Model (PGM) can then be summarized as follows:

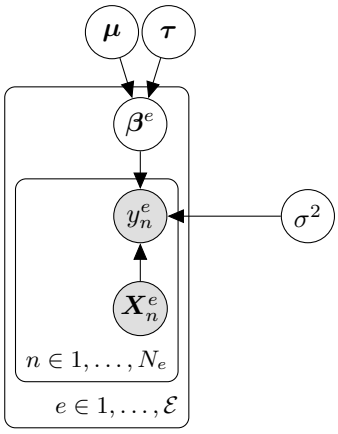


Figure 1: Bayesian Hierarchical PGM

1. Draw global parameters:
   - $\boldsymbol{\mu} \sim \mathcal{N}(\mu_0, \Sigma_0)$
   - $\boldsymbol{\tau} \sim$ Half-Cauchy(σ_0)
2. For each environment $e \in \{1, \ldots, E\}$:
   (a) Draw local parameters:
   - $\boldsymbol{\beta^e} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\tau})$

   i. For each observation $n \in \{1, \ldots, N_e\}$:
   - Draw $y_n^e \sim \mathcal{N}(\boldsymbol{X}_n^e \boldsymbol{\beta}^e, \sigma^2)$

2. If y is not continuous, the likelihood in step 2.(a)i. should be adjusted, for example, for categorical y, use $y_n^e \sim$ Categorical$(\text{softmax}(\boldsymbol{X}_n^e \boldsymbol{\beta}^e))$.

3.2. Inference and Testing Procedures

This section describes how we identify invariant predictors within the BHIP framework using our BHM. Our definition of invariance is based upon on a predictor's global and local parameters being both credibly non-zero and 'close' to each other.

To assess this relationship formally, we employ specific Bayesian statistical tests. This approach provides a more integrated and potentially more efficient way to assess invariance directly from estimated parameter relationships, offering a distinct advantage over ICP that requires exhaustive conditional independence tests. We fit the BHM to obtain posterior distributions for the parameters (See D for details), and then apply our proposed tests to the posterior samples for each predictor to classify them as invariant based on our predefined criteria. We employ two key statistical tests:

HDI+ROPE Decision Rule is used for the 'effect non-zero test,' which determines whether a predictor consistently influences the target across environments. It combines the Highest Density Interval (HDI) with the Region of Practical Equivalence (ROPE) (Kruschke, 2018). ROPE is an interval around zero representing negligible effects. A predictor is considered relevant if its HDI lies substantially outside this region. For a parameter θ, the procedure follows these steps:

1. Define the ROPE as $[-\epsilon, \epsilon]$, where $\epsilon = 0.1 \cdot \hat{s}$ and $\hat{s}$ is the sample standard deviation of the posterior distribution of θ.
2. Compute the HDI. Calculate a standard HDI (e.g., a 95% HDI) from the posterior distribution. This gives a specific interval $[\theta_{\text{lower}}, \theta_{\text{upper}}]$, which represents the most credible values for the parameter.
3. Compare the HDI to the ROPE. The position of the computed HDI relative to the ROPE determines the outcome of the test based on the following decision rule:
    - Rejected: The effect is considered statistically significant if the entire HDI is outside the ROPE.
    - Accepted: The effect is considered practically equivalent to zero if the entire HDI is inside the ROPE.
    - Undecided: The test is inconclusive if the HDI and the ROPE partially overlap.

The result from this procedure is a categorical decision ('Rejected', 'Accepted', or 'Undecided') about the predictor's effect. We also compute the largest probability mass outside of ROPE, which is a value in $[0, 1]$. Where higher values represent higher confidence in rejecting the hypothesis that the predictor does not have an effect on the target variable.

Pooling Factor quantifies information sharing (pooling) across environments in a BHM, as proposed by (Gelman and Pardoe, 2006). Define for each environment and each covariate X_d the error term δ_d^e, which is the difference between the global mean and the local parameter, that is $\beta_d^e = \mu_d + \delta_d^e$. It is defined as,

$$\gamma_d = 1 - \frac{\text{Var}_{e \in \mathcal{E}} \left[\mathbb{E} \left[\delta_d^e \right] \right]}{\mathbb{E} \left[\text{Var}_{e \in \mathcal{E}} \left[\delta_d^e \right] \right]}. \tag{10}$$

The denominator, $\mathbb{E}\left[\text{Var}_{e \in \mathcal{E}}(\delta_d^e)\right]$, is the expected variance in the deviations from the environment-level effect and the global parameter. This is the unexplained component of the variability in the

β^e's. Interpreting the pooling factor γ_d: values close to 1 signal strong invariance, suggesting the variable is a potential invariant predictor if its effect is non-zero. The lower the value of γ_d (i.e., the further it is below 1), the higher the indicated heterogeneity, suggesting the variable is less likely to be invariant.

In an applied setting, we rely on both the HDI+ROPE decision rule and the pooling factor test together. Requiring a predictor to pass both tests embodies the core idea that invariant parameters must have credibly non-zero effects across different environments.

Under the invariance assumption and regularity conditions we can guarantee that as we collect more data and environments, the pooling factor converges to 1 for those variables that are causal parents. The full proof of our main result can be found in Appendix C.1).

Theorem 1 *(Asymptotic behavior of pooling factor) Suppose the invariance assumption (Assumption 1) and the Bernstein von Mises holds for a BHIP model, then*

$$\gamma_d \xrightarrow{P} \begin{cases} 1 & \text{if } \beta_d^{e*} = \beta_d^{e'*} \text{ for all } e, e' \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

4. Experiments and Results

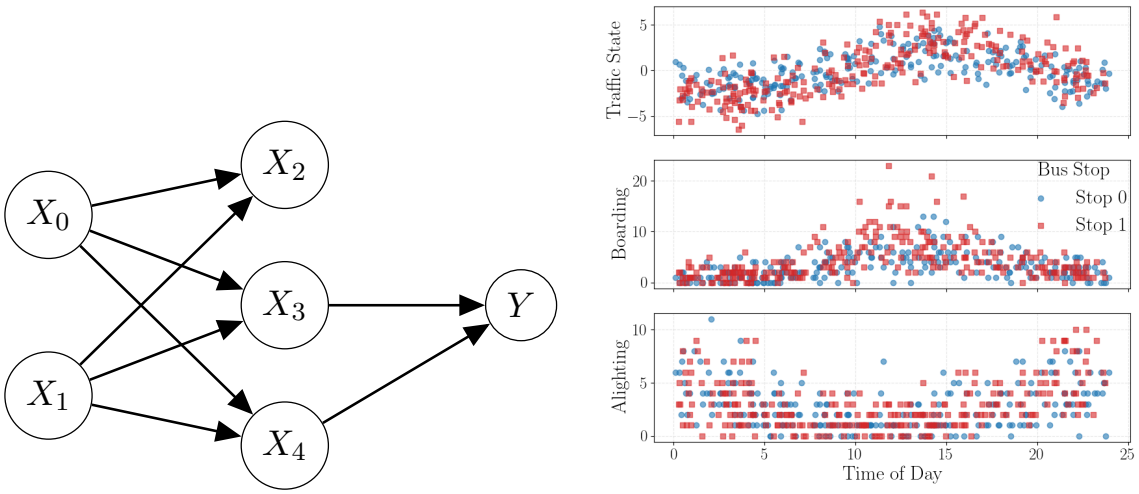
This section presents several experiments to evaluate the performance and characteristics of BHIP. In Section 4.1 we introduce the Bus Dwelling Problem, a controlled setting where we define the SCM and its corresponding DAG, allowing us to assess BHIP's ability to recover causal relationships and compare its performance against ICP. Section 4.2 focuses on the educational attainment dataset (Rouse, 1995) that was used to introduce ICP (Peters et al., 2016). In Section 4.3, we provide an analysis of the computational scalability of BHIP compared to ICP, demonstrating a key advantage for problems with many predictors. After establishing these properties, in Section 4.4.1 and Section 4.4.2 we benchmark BHIP against other invariant prediction methods, including the concurrent work of Wu et al. (2025), on the low-dimensional synthetic and gene perturbation setups introduced in that work. Further experiments were conducted and can be found in Appendices H and J. The former focuses on a synthetic problem with many different configurations for a quantitative analysis of BHIP's performance, while the latter focuses on a transport-related problem with real data, where BHIP is applied to infer causal predictors in the choice of mode of transport. All experiments in this work were run on a Threadripper 2950X (40M Cache, 3.4 GHz base) CPU, 16 cores, 128 GB RAM.

4.1. Case Study: The Bus Dwelling Problem

The Bus Dwelling Problem is a controlled experiment where we model the time a bus dwells at different bus stops as functions of multiple factors: time of the day X_0, day of the week X_1, traffic conditions X_2, and the number of boarding X_3 and alighting passengers X_4. The DAG that represents our SCM is represented by Figure 2(*a*). Unlike real-world datasets, this setup allows us to define the ground-truth causal relationships and systematically evaluate how well BHIP recovers them.

4.1.1. DATA GENERATION PROCESS

The dwelling time Y is directly influenced by covariates such as the number of boarding passengers X_3, and the number of alighting passengers X_4. This choice is consistent with empirical findings



(*a*) Synthetic bus dwelling problem causal graph. (*b*) Bus dwelling predictors time series.

Figure 2: Bus dwelling problem: (a) causal graph and (b) predictor time series examples.

that identify passenger boarding and alighting as the primary determinants of bus stop dwell time Levinson (1983); Dueker et al. (2004), and here we adopt this simplified specification. The generative model follows:

$$Y = f(X_3, X_4) + \epsilon, \tag{12}$$

where f represents an underlying causal function, and ϵ is an independent noise term. Importantly, we ensure that X_2 (traffic state) does not directly cause Y, allowing us to test BHIP's ability to avoid spurious correlations. Realistic temporal variations and passenger counts were simulated using modulated sinusoidal and Poisson processes, respectively, and are represented in Figure 2(*b*). Full generation details are in Appendix F.

4.1.2. RESULTS

We apply both BHIP and ICP to infer the set of causal predictors. Since both methods rely on tests across different environments, we define environments based on the different bus stops, that have different data generative models but the same underlying SCM.

Regarding the model setup for the experiments, we fix the observation noise variance at $\sigma^2 = 1.0$. We assign standard priors to the global parameters, specifically $\mu \sim \mathcal{N}(0, 1)$ and $\tau \sim$ Half-Cauchy(1), and local parameters $\beta^e \sim \mathcal{N}(\mu, \tau)$. We try two different experiments, one with the number of data points per bus stop $N = 100$ and another one with $N = 500$. The key question we evaluate is: can BHIP correctly recover X_3 and X_4 as causal predictors while excluding X_0, X_1 and X_2?

Table 1: BHIP and ICP results

X	$N = 100$		$N = 500$	
	ICP	BHIP	ICP	BHIP
X_0	-	-	-	-
X_1	-	-	-	-
X_2	-	-	-	-
X_3	✓	✓	✓	✓
X_4	-	-	-	✓

Using non-centered parameterization (see Appendix D), BHIP without the inclusion of informative priors identified predictors X_3 and X_4 as having significant invariant effects. For $N = 500$,

X_3 showed 100% HDI outside ROPE for local and 95% for global parameters ($\gamma_3 = 0.96$), while X_4 had 100% local and 90% global HDI outside ROPE ($\gamma_4 = 0.97$). Other predictors consistently showed HDIs below 65% outside ROPE, indicating BHIP successfully identified the true causal predictors.

Table 1 compares BHIP (with a decision rule: HDI out of ROPE $> 85\%$, $\gamma_d > 0.85$) against ICP ($\alpha = 0.05$). Covariates detected as invariant causal predictors are marked '✓'. In contrast to BHIP, ICP identified X_3 but consistently missed X_4, likely due to its conservative nature (type I errors) (Peters et al., 2016). Overall, BHIP demonstrated improved power in identifying both true causal predictors (X_3, X_4) while rejecting non-causal ones in this setting.

4.2. Case study: Educational attainment

We use the educational attainment dataset (Rouse, 1995), which contains information on 4739 students in 1,100 high schools in the USA. The dataset includes 13 covariates such as gender, ethnicity, standardized test scores, parental education levels, family income, and the binary target variable that indicates whether a student attained a bachelor's degree or higher, corresponding to at least 16 years of education. To introduce heterogeneity, we follow the environmental split based on the distance to the nearest 4-year college, a variable assumed to have no direct causal effect on the target. Students who live closer than the median form one environment, while those living farther than the median form the other.

4.2.1. METHODOLOGY

We implement the BHIP framework using a hierarchical logistic regression model. The model includes environment-specific effects for each predictor while pooling information across environments through global parameters. We compare the results with those obtained using the ICP method.

For both methods, most predictors are encoded as dummy variables (e.g., *fcollege_yes* indicates if the father attained a college degree). The primary goal is to identify invariant predictors with non-zero effects across environments and quantify uncertainty in their contributions.

4.2.2. RESULTS

We applied BHIP on the educational attainment dataset, with some results shown in Figure 3 (additional figures in Appendices G). While direct comparison to ICP's results is for intuition due to lack of ground truth, we hereby show how BHIP offers a richer analysis.

BHIP identified achievement *score* as a strong invariant predictor (Figure 3), with 95% HDI outside ROPE and $\gamma = 1$, aligning with ICP findings. Father's education (*fcollege_yes*) local parameters do not lie completely outside the ROPE in any environment but showed a positive, largely invariant effect ($\gamma = 0.92$), again similar to ICP's finding for the inverse predictor. Notably, BHIP identified *income_low* as a negative invariant predictor ($\gamma = 0.99$), an effect not detected by ICP. Other predictors like *region_west* and ethnicity showed more nuanced, non-invariant effects (HDIs overlapping ROPE), consistent with ICP's exclusions. While both methods identified *score* and father's education, BHIP's hierarchical modeling provided a richer analysis of environment-specific effects and uncertainty quantification, demonstrating its flexibility.

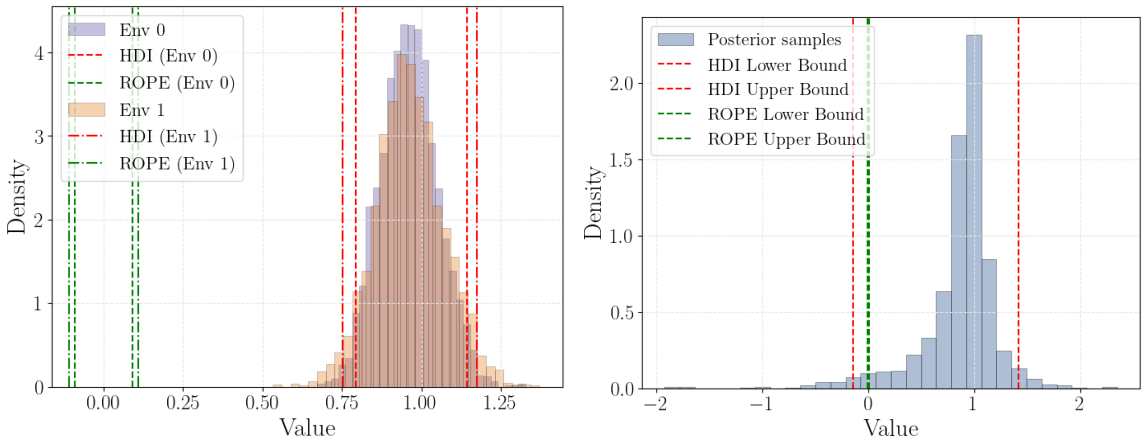


Figure 3: HDI + ROPE for local parameters left (Beta score) and global parameters right (mu score) of predictor: `score`. Pooling factor of $\gamma_d = 1$.

4.3. Computational Complexity Study

We generated 100 random DAGs for configurations with the number of nodes N ranging from 3 to 20. For simplicity and focus on the algorithmic scaling with N, each problem instance used 200 samples per environment and 2 environments. Figure 4(*a*) displays the distribution of the computational times. As hypothesized based on the algorithmic differences, Figure 4(*a*) shows that the computational time for ICP grows exponentially with the number of nodes. In contrast, BHIP's runtime increases at a much slower rate. This empirical result strongly supports the notion that leveraging the hierarchical Bayesian framework to assess invariance through parameter relationships, rather than relying on exhaustive conditional independence testing across all predictor subsets, allows BHIP to remain computationally viable for problems involving a larger number of potential causal predictors.

4.4. Benchmark against invariant prediction methods

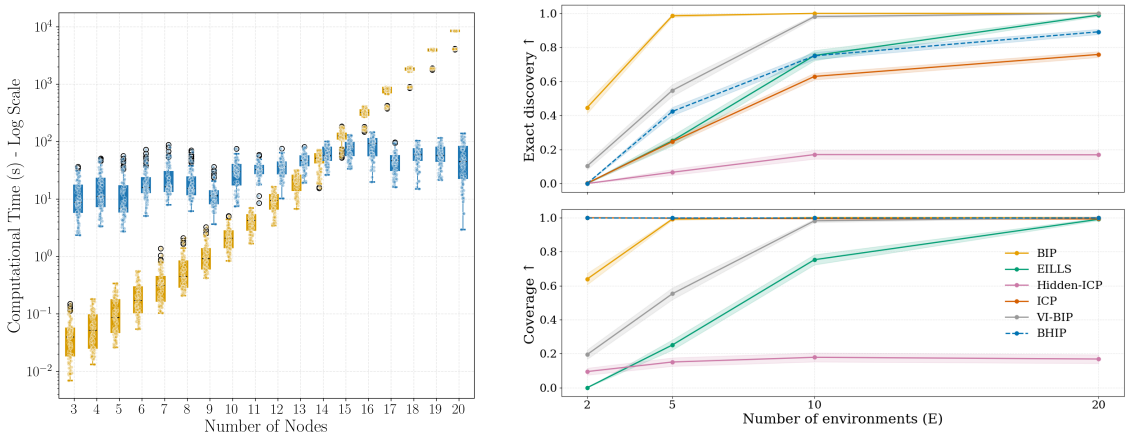
To position BHIP among recently proposed invariant prediction approaches, we adopt the experimental setups of Wu et al. (2025), who introduce Bayesian Invariance Modeling of Multi-Environment Data (BIP) and its variational approximation VI-BIP. In particular, we (i) replicate their low-dimensional linear-Gaussian synthetic study, where the true invariant feature set is known, and (ii) reuse the large-scale yeast gene perturbation dataset originally collected by Kemmeren et al. (2014) and previously analyzed in the context of invariant prediction by Peters et al. (2016); Meinshausen et al. (2016). In these experiments, in addition to BHIP (ran with HDI–ROPE decision rule and pooling factors of 0.95), BIP and VI-BIP (ran with the recommended hyperparameters), we include ICP (Peters et al., 2016), which we run with significance level $\alpha = 0.05$ for its multiple hypothesis testing procedure; Hidden-ICP (Rothenhäusler et al., 2019), a relaxed invariance extension of ICP that we also run with $\alpha = 0.05$; and EILLS (Fan et al., 2024), a linear regression method with an invariance regularization term across environments, for which we follow the default implementation and set the regularization strength to $\gamma = 36$.

4.4.1. SYNTHETIC BENCHMARK

Following Wu et al. (2025), we generate multi-environment linear-Gaussian data with $p = 10$ predictors. We vary the number of environments $E \in \{2, 5, 10, 20\}$ and the per-environment sample size, and evaluate all methods using the exact discovery rate (probability of recovering the invariant set exactly) and coverage (probability that the estimated invariant set is a subset of the true set).

To maintain consistency with the original benchmark by Wu et al. (2025), we focus our main presentation on exact discovery and coverage. However, to better understand traditional metrics results including precision, recall, and F1 scores are available in Appendix I.

Figure 4(*b*) summarizes the results. As E increases, BHIP approaches perfect exact discovery, closely tracking BIP and VI-BIP and substantially outperforming ICP and Hidden-ICP, which remain either overly conservative (high coverage but low exact discovery for ICP) or less accurate overall. BHIP matched ICP with coverage of 1 irrespective of E, outperforming BIP and VI-BIP. Furthermore, while EILLS performs well when the number of environments is high, it exhibits low coverage in low-environment experiments, which, as detailed in the appendix, suggests low precision under those conditions.



(*a*) Computational time distribution for ICP vs. BHIP across varying numbers of nodes.

(*b*) Exact discovery (above) and coverage (below) as a function of the number of environments E for BHIP and competing invariant prediction methods.

Figure 4: (a) Computational cost comparison of ICP and BHIP; (b) low-dimensional synthetic benchmark of BHIP and competing invariant prediction methods.

4.4.2. GENE PERTURBATION BENCHMARK

Finally, we consider the gene perturbation dataset from large-scale yeast gene deletion experiments (Kemmeren et al., 2014). The data comprise genome-wide mRNA expression levels measured under an observational condition and under single-gene deletion interventions. Following past works, we focus on a benchmark set of target genes and, for each target, compare the invariant feature genes inferred by different methods. As in Wu et al. (2025), ICP-s first applies a Lasso-based pre-screening that retains 10 candidate predictors per target before running ICP with significance level $\alpha = 0.01$;

for BHIP we use the same screening scores but keep the top 200 candidates per target to avoid prohibitively large, memory-wise, Monte Carlo fits over all genes.

Table 2 summarizes the inferred invariant feature genes for the 10 benchmark targets. Overall, BHIP broadly agrees with the previously validated effects: selected features mostly coincide with genes identified by ICP-s and/or VI-BIP and reported as true effects in Meinshausen et al. (2016). For some targets BHIP returns an empty invariant set, behaving more conservatively than VI-BIP, whereas for others it proposes slightly larger invariant sets that extend the ICP-s predictions. Taken together, these results indicate that BHIP can recover the core validated regulatory relationships in this benchmark while remaining relatively conservative in high-dimensional regions.

Table 2: Inferred invariant feature gene(s) for benchmark target genes in the yeast perturbation dataset. ∅ denotes that no invariant feature was selected. Genes in blue are validated to have significant effects on the corresponding target gene

Target gene	Inferred invariant feature gene(s)			
	Meinshausen et al. (2016)	ICP-s ($\alpha = 0.01$)	VI-BIP ($t = 0.5$)	BHIP
YMR103C	YMR104C	YMR104C	YMR104C, YHR209W	YMR104C
YMR321C	YPL273W	YPL273W	YPL273W	YPL273W
YCL042W	YCL040W	YCL040W	YCL040W	YCL040W
YLL020C	YLL019C	YLL019C	YLL019C	YLL019C
YPL240C	YMR186W	YMR186W	YJL077C, YMR186W	YMR186W, YOL121C, YLL045C, YMR143W
YBR126C	YDR074W	∅	YGR008C, YDR074W, YKL035W	∅
YMR173W-A	YMR173W	YMR173W, YOL100W	YMR173W	YMR173W
YGR264C	YGR162W	∅	∅	∅
YJL077C	YOR027W	∅	YLL026W, YOR027W, YFL010W-A	YOR027W, YDR214W
YLR170C	YJL115W	YDR322C-A, YDR180W, YJL184W, YLR438C-A	∅	∅

In conclusion, the results reinforce the flexibility of BHIP in identifying invariant predictors while quantifying their uncertainty. Unlike ICP, BHIP provides a probabilistic perspective that is crucial for real-world applications where heterogeneity and data limitations are prevalent and allows for the use of priors whether to incorporate domain knowledge or regularization.

5. Discussion and Conclusion

Our study presents the BHIP framework, a novel approach for identifying invariant predictors across heterogeneous environments. While BHIP builds upon established techniques such as BHMs and invariance testing, its novelty lies in its specific synthesis and targeted application to probabilistically test causal invariance for parent discovery.

Despite its significant strengths, BHIP also presents some limitations. As a Bayesian method relying on MCMC for inference, it incurs a higher computational overhead cost compared to frequentist alternatives like ICP. It also assumes that the defined environments capture meaningful heterogeneity reflective of underlying causal structure, an aspect that may not always hold perfectly in real-world applications. Moreover, while offering flexibility, the effectiveness of BHIP depends on appropriate model specification, selection of priors, and careful interpretation of the rich posterior

outputs, requiring a degree of understanding of BHMs. Nonetheless, our analysis demonstrated that BHIP effectively identifies strong invariant predictors while comprehensively quantifying the uncertainty of their effects and invariance. This enables a more nuanced and informative analysis of predictors than a simple binary classification approach. Furthermore, while ICP is limited by the curse of dimensionality, BHIP scales more effectively and remains computationally tractable even with a significantly larger number of predictors p.

BHIP's capacity for probabilistic invariance testing, rigorous uncertainty quantification, and identification of robust causal predictors represents a significant step forward for causal discovery in complex, real-world scenarios. Future research could explore several directions. One avenue is developing more computationally efficient inference algorithms and extending the framework to capture more intricate forms of heterogeneity and non-linear causal relationships. Another is investigating how an active learning approach could leverage the extra information from BHIP to perform interventions that generate ideal environments. Overall, BHIP contributes a principled, probabilistic approach to discovering invariant causal relationships across diverse environments.

Acknowledgments

We thank Luhuan Wu for the facilitation of code and experiments relative to BIP. The work presented in this article is supported by Novo Nordisk Foundation grant NNF23OC0085356.

References

John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

Michael Betancourt. Hierarchical modeling. November 2020. URL https://betanalpha.github.io/assets/case_studies/hierarchical_modeling.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. 2018. URL https://arxiv.org/abs/1810.09538.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 2018. URL https://api.semanticscholar.org/CorpusID:88523994.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

George Casella and Edward I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46 (3):167–174, 08 1992.

Hjalmar Christiansen and Britt Zöega Skougaard. Documentation of the danish national travel survey. 2015.

Rune Christiansen, Matthias Baumann, Tobias Kuemmerle, Miguel D Mahecha, and Jonas Peters. Toward causal inference for spatio-temporal data: Conflict and forest loss in colombia. *Journal of the American Statistical Association*, 117(538):591–601, 2022.

Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 116–125, 1999.

Kenneth J. Dueker, Thomas J. Kimpel, James G. Strathman, and Steve Callas. Determinants of bus dwell time. *Journal of Public Transportation*, 7(1):21–40, 2004.

Jianqing Fan, Cong Fang, Yihong Gu, and Tong Zhang. Environment invariant linear least squares. *The Annals of Statistics*, 52(5):2268–2292, 2024.

Roy Frostig, Matthew Johnson, and Chris Leary. Compiling machine learning programs via high-level tracing. 2018. URL https://mlsys.org/Conferences/doc/2018/146.pdf.

Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 33:15464–15475, 2020.

Juan L. Gamella, Armeen Taeb, Christina Heinze-Deml, and Peter Bühlmann. Characterization and greedy learning of gaussian structural causal models under unknown interventions, 2025.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, volume Analytical methods for social research. Cambridge University Press, New York, 2007.

Andrew Gelman and Guido Imbens. Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research, 2013.

Andrew Gelman and Iain Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48:241–251, 2006.

Edward I. George and Robert E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2), 1997.

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *Advances in Neural Information Processing Systems*, 36, 2024a.

Siyuan Guo, Chi Zhang, Karthika Mohan, Ferenc Huszár, and Bernhard Schölkopf. Do finetti: On causal effects for exchangeable data. *arXiv preprint arXiv:2405.18836*, 2024b.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. Bacadi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 1411–1436. PMLR, 2023.

David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. *Innovations in Machine Learning: Theory and Applications*, pages 1–28, 2006.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Patrick Kemmeren, Katrin Sameith, Loes A. L. van de Pasch, Joris J. Benschop, Tineke L. Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W. Ko, Sebastiaan van Heesch, Mehdi M. Kashani, Giannis Ampatziadis-Michailidis, Mariel O. Brok, Nathalie A. C. H. Brabers, Anthony J. Miles, Diane Bouwmeester, Sander R. van Hooff, Harm van Bakel, Erik Sluiters, Linda V. Bakker, Berend Snel, Philip Lijnzaad, Dik van Leenen, Marian J. A. Groot Koerkamp, and Frank C. P. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014. doi: 10.1016/j.cell.2014.02.054.

Lucas Kook, Sorawit Saengkyongam, Anton Rask Lundborg, Torsten Hothorn, and Jonas Peters. Model-based causal feature selection for general response types. *Journal of the American Statistical Association*, pages 1–12, 2024.

John K. Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1:270–280, 2018.

Herbert S. Levinson. Analyzing transit travel time performance. *Transportation Research Record*, 915(1):1–6, 1983.

Kenneth G Manton, Max A Woodbury, Eric Stallard, Wilson B Riggan, John P Creason, and Alvin C Pellom. Empirical bayes procedures for stabilizing maps of us cancer mortality rates. *Journal of the American Statistical Association*, 84(407):637–650, 1989.

Nicolai Meinshausen, Alain Hauser, Joris M. Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–32, 1988.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

Omiros Papaspiliopoulos and Gareth Roberts. Non-centered parameterisations for hierarchical models and data augmentation. *Bayesian Statistics*, 7:307–326, 01 2003.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3 (none), January 2009. ISSN 1935-7516. doi: 10.1214/09-SS057. URL https://projecteuclid.org/journals/statistics-surveys/volume-3/issue-none/Causal-inference-in-statistics-An-overview/10.1214/09-SS057.full.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Niklas Pfister, Stefan Bauer, and Jonas Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019a.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019b.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal dantzig. *The Annals of Statistics*, 47(3):1688–1722, 2019. doi: 10.1214/18-AOS1736.

Cecilia E. Rouse. Democratization or diversion? the effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 13:217–224, 1995.

Sorawit Saengkyongam, Nikolaj Thams, Jonas Peters, and Niklas Pfister. Invariant policy learning: A causal perspective. *IEEE transactions on pattern analysis and machine intelligence*, 45(7): 8606–8620, 2023.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 1993.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Eli N. Weinstein and David M. Blei. Hierarchical causal models. *arXiv preprint arXiv:2401.05330*, 2024. URL https://arxiv.org/abs/2401.05330.

Luhuan Wu, Mingzhang Yin, Yixin Wang, John P. Cunningham, and David M. Blei. Bayesian invariance modeling of multi-environment data, 2025. URL https://arxiv.org/abs/2506.22675.

Supplementary Material

Table 3: Notation table

Sets, indices, and environments	
D	Number of variables (excluding the target Y).
$\mathcal{E}$	Index set of environments; $e \in \mathcal{E}$ indexes an environment.
N_e	Number of samples in environment e; $N = \sum_{e \in \mathcal{E}} N_e$.
n	Index for observations within an environment, $n = 1, \ldots, N_e$.
$S \subseteq \{1, \ldots, D\}$	Subset of predictor indices; X_S denotes the corresponding covariates.
Random variables and data	
$\mathbf{X} = (X_1, \ldots, X_D)$	Predictors / covariates.
Y	Target variable.
$(Y^e, \mathbf{X}^e)$	Data in environment e (cf. Eq. (2)).
y_n^e, X_n^e	n-th observation of target and covariates in environment e.
SCM and graphical model	
$\mathcal{M} = (\mathbb{S}, P_\varepsilon)$	Structural Causal Model (SCM).
$X_d := f_d(\mathrm{PA}(X_d), \varepsilon_d)$	Structural assignment for variable X_d.
$\mathcal{G}$	DAG induced by the SCM over $(\mathbf{X}, Y)$.
$\mathrm{PA}(Y)$	Set of (direct) causal parents of Y.
Distributions	
$\mathcal{N}(\cdot, \cdot)$, $\mathsf{Cauchy}^+(\cdot, \cdot)$	Normal and half-Cauchy distributions.
Hierarchical model parameters (BHIP)	
$\boldsymbol{\beta}^e = (\beta_1^e, \ldots, \beta_D^e)$	Environment-specific regression coefficients.
ϕ_d	Global (population) parameter(s) governing β_d^e (generic notation).
μ_d, τ_d	Mean and standard deviation hyperparameters for β_d^e (e.g., $\beta_d^e \sim \mathcal{N}(\mu_d, \tau_d^2)$).
σ^2	Observation noise variance in Gaussian likelihoods.
δ_d^e	Deviation from the global mean: $\delta_d^e = \beta_d^e - \mu_d$.
Bayesian testing / decision quantities	
γ_d	Pooling factor for predictor d: $\gamma_d = 1 - \dfrac{\mathrm{Var}_e[\,\mathbb{E}[\delta_d^e]\,]}{\mathbb{E}[\,\mathrm{Var}_e(\delta_d^e)\,]}$.
$\mathrm{HDI}_q(\theta)$	$q\%$ Highest Density Interval of a posterior for parameter θ.
ROPE	Region of Practical Equivalence around 0 (e.g., $[-\epsilon, \epsilon]$).
$p_{\mathrm{out}}(\theta)$	Posterior mass of θ outside ROPE (used to judge non-zero effects).
Sparsity priors	
λ_d	Local shrinkage (horseshoe) for predictor d; $\lambda_d \sim \mathsf{Cauchy}^+(0, 1)$.
τ	Global shrinkage (horseshoe); $\tau \sim \mathsf{Cauchy}^+(0, 1)$.
z_d	Spike-and-slab inclusion indicator for predictor d ($z_d \in \{0, 1\}$).
π	Prior inclusion probability for spike-and-slab ($z_d \sim \mathrm{Bernoulli}(\pi)$).

This supplementary material is organized as follows. Table 3 summarizes the notation used throughout the paper and Appendix A states the key causal and statistical assumptions underlying BHIP. Appendices B and C provide theoretical support for the invariance principle and the behavior of the pooling factor. Appendix D describes the non-centered parameterization and inference scheme, while Appendix E details the use of sparse priors. Appendices F–J contain additional experimental details and results for the Bus Dwelling Problem, the educational attainment study, the large-scale synthetic simulations, and the TU Danish Travel Survey.

Appendix A. Key Assumptions

This section outlines the key assumptions required for the BHIP framework. These assumptions stem from those commonly made in Causal Discovery, while also including assumptions specific to our Bayesian modeling approach.

BHIP relies on the following assumptions:

Causal and Graphical Model Assumptions

These assumptions are standard in many causal discovery methods, including ICP.

- **Acyclicity:** The underlying causal system relating the variables $\mathbf{X}$ and Y can be represented by a DAG.
- **SCM Representation:** The data is assumed to be generated by an underlying SCM, where each variable is a function of its direct causes (parents) and an independent noise term (Pearl, 2009).
- **Faithfulness**: The conditional independence relationships observed in the data from each environment precisely correspond to the d-separation properties in the true causal graph, and vice versa (Spirtes et al., 1993).
- **Causal Sufficiency (relative to $\mathbf{X}, Y$):** There is no unobserved confounding between the predictor variables $\mathbf{X}$ and the target variable Y that would create spurious invariant associations across the observed environmental changes.

Invariance Assumptions

These assumptions are central to the principle of identifying causal relationships by leveraging heterogeneity across environments, as in ICP.

- **Heterogeneous Environments:** The data is observed across distinct environments $e \in \mathcal{E}$, where the distributions of the predictor variables $P(\mathbf{X}^e)$ are allowed to change.
- **Structural Invariance (Assumption 2):** The conditional distribution of the target variable given its true causal parents, $P(Y|PA(Y))$, remains invariant across all environments $e \in \mathcal{E}$. The distribution of the noise term ε_Y for the target variable is also assumed to be the same across environments.

BHIP Statistical Model and Inference Assumptions

These assumptions pertain specifically to the statistical modeling choices and inference procedure employed by BHIP.

- **Correct Likelihood & Functional Form:** The chosen likelihood function (e.g., Gaussian for continuous Y, Logistic for binary Y) and the functional form of the relationship between Y and its predictors within each environment (e.g., linear $Y \approx \mathbf{X}\beta^e$) accurately model $P(Y|\mathbf{X}, \beta^e)$. While the model is presented with a linear form, the framework can be extended to handle non-linear relationships, similar to extensions in the frequentist ICP (Heinze-Deml et al., 2018).

- **Correct Hierarchical Structure:** The assumption that the environment-specific coefficients β_d^e for a given predictor d are drawn from a common distribution (e.g., Normal(μ_d, τ_d^2)) across environments adequately captures the underlying structure of parameter variation and sharing.

- **Reasonable Priors:** The prior distributions selected for the model parameters are chosen such that they allow the posterior distribution to concentrate correctly on the true parameter values, particularly in the large sample limit. Sparsity-inducing priors (like Horseshoe or Spike-and-Slab) are assumed to appropriately reflect beliefs about the sparsity of the true parent set and facilitate variable selection.

- **Independent Noise:** The noise term ε_Y is assumed to be independent across individual observations and across environments (conditional on the parent variables).

- **MCMC Convergence:** The Markov Chain Monte Carlo (MCMC) algorithm used for posterior inference is assumed to have converged, providing samples that accurately represent the true posterior distribution.

Assumptions for Asymptotic Guarantees

These conditions are typically required for theoretical results regarding the asymptotic behavior and consistency of Bayesian estimators and variable selection procedures.

- **Standard Regularity Conditions:** Standard technical conditions necessary for Bayesian asymptotic theorems (such as the Bernstein-von Mises theorem or results on posterior consistency for hierarchical and sparse models) are assumed to hold for the specific model and data.

- **Sufficient Heterogeneity:** The changes in the predictor distributions $P(\mathbf{X}^e)$ across environments must be sufficiently diverse and informative to allow the true invariant parent set S^* to be reliably distinguished from non-invariant predictors.

- **Appropriate Thresholds:** The decision thresholds used for variable selection based on posterior summaries (e.g., ROPE sizes for HDI tests, the threshold value for the pooling factor) are assumed to be set at appropriate values relative to the scale of the true effects and the rate of posterior concentration.

Note on **Identifiability**:

Given the aforementioned assumptions the true invariant parent set PA(Y) is identifiable from the data. BHIP's statistical framework is designed to leverage this identifiability property and recover PA(Y) from the data by identifying variables whose relationship with Y is invariant across the observed environments.

Appendix B. Invariance Proposition and Proof

Full proposition and proof (Bühlmann, 2018):

Proposition 1 *Assume an SCM as in Assumption 2 and let the set of environments $\mathcal{E}$ be such that Assumption 2 holds. Then, the set of direct causes $\mathrm{PA}(Y)$ is invariant with respect to $\mathcal{E}$ and Assumption 1 holds.*

Proof In Assumption 2 it is seen that the conditional distribution of Y^e given $X^e_{\mathrm{PA}(Y)}$ is fully determined by the structural equation f_Y and the noise distribution $F_{\mathcal{E}}$, both remaining the same for all $e \in \mathcal{E}$, thus Assumption 1 is satisfied and $\mathrm{PA}(Y)$ is invariant with respect to $\mathcal{E}$. ■

Appendix C. Asymptotic Guarantees

BHIP's ability to recover the true invariant parent set $\mathrm{PA}(Y)$ asymptotically (as sample size $N \to \infty$) is motivated by Bayesian posterior consistency. While BHIP's specific posterior-based decision rules differ from ICP's frequentist tests, we argue for its asymptotic correctness. Under standard regularity conditions, Bayesian posterior distributions concentrate around true parameter values, often approximating normality centered at efficient estimators (as suggested by the Bernstein-von Mises (BvM) theorem (Vaart, 1998)). We expect this concentration for BHIP's hierarchical parameters $(\mu_d, \tau_d, \beta_d^e)$. Furthermore, the use of appropriate sparsity priors yields posterior consistency for variable selection in high-dimensional regression settings. This expected posterior concentration implies specific asymptotic behavior for BHIP's tests:

- **HDI+ROPE Tests:** For true parents $d \in \mathrm{PA}(Y)$ (with non-zero effects), the concentrating posteriors for μ_d and β_d^e will eventually place the HDIs entirely outside a fixed ROPE, leading to inclusion with probability approaching 1. For non-parents with zero effects, the HDIs will concentrate within the ROPE, ensuring exclusion.
- **Pooling Factor:** For true parents $d \in \mathrm{PA}(Y)$, invariance (Assumption 1) implies identical β_d^e. Posterior concentration on this common value should lead the pooling factor γ_d posterior to concentrate near 1. For non-invariant non-parents, differing true β_d^e will keep the pooling factor below asymptotically. See Proof C.1 below.

The combination of asymptotic posterior concentration, consistent sparsity priors, and the logic of the invariance tests provides strong theoretical motivation for BHIP's ability to correctly identify $\mathrm{PA}(Y)$ asymptotically under the stated assumptions.

C.1. Pooling Factor Consistency

Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from a parametric model $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$ and the true parameter $\theta^* \in \Theta$.

For a Bayesian approach, the parameter θ is treated as random and encodes our prior beliefs about the value of the parameter in a distribution π. The setup is:

$$\begin{aligned} \theta &\sim \pi, \\ \{X_1, \ldots, X_n\} \mid \theta &\sim P_\theta. \end{aligned}$$

Using Bayes' Theorem the posterior distribution of θ can be computed as:

$$p(\theta \mid X_1, \ldots, X_n) = \frac{\mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)}{\int_\theta \mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)d\theta} \propto \mathcal{L}(\theta)\pi(\theta). \tag{13}$$

The Bernstein–von Mises (BvM) theorem guarantees us that in fixed-dimensional problems, under the assumption that the prior is continuous and strictly positive in a neighborhood around θ^*, the posterior distribution converges in total-variation distance to a Gaussian distribution centered around the Maximum Likelihood Estimator (MLE) $\hat{\theta}_n$, that is,

$$\left\| p(\theta \mid X_1, \ldots, X_n) - \mathcal{N}\left(\hat{\theta}_n, \frac{1}{n}\mathcal{I}(\hat{\theta}_n)^{-1}\right) \right\|_{\text{TV}} \to 0, \tag{14}$$

where $\mathcal{I}(\hat{\theta}_n)$ is the Fisher information matrix around the MLE estimate. Assuming a regular parametric model and that the BvM theorem holds, the posterior distribution converges in probability to a Gaussian distribution around the true parameter with rapid shrinkage, that is, as $n \to \infty$

$$p_n(\theta \mid X^{(n)}) \xrightarrow{\text{TV}} \mathcal{N}\left(\hat{\theta}_n,\ \frac{1}{n}I^{-1}(\theta^*)\right) \tag{15}$$

which implies that the posterior is a consistent estimator for θ^* as

$$\mathbb{E}[\theta \mid X^{(n)}] = \int \theta dp_n(\theta \mid X^{(n)}) \xrightarrow{P} \theta^* \tag{16}$$

with variance shrinking at a rate of $1/n$

$$\text{Var}(\theta \mid X^{(n)}) \xrightarrow{P} \frac{1}{n}I^{-1}(\theta^*) \tag{17}$$

hence as $n \to \infty$, we have

$$\text{Var}(\theta \mid X^{(n)}) \xrightarrow{P} 0 \tag{18}$$

Consider the hierarchical Bayesian BHIP model where for each covariate X_d, $d = 1, \ldots, D$ we have:

$$\begin{aligned}
\phi &\sim \pi && \text{(global parameter)} \\
\beta^e \mid \phi &\sim p(\beta^e \mid \phi) && \text{(parameter in environment e)} \\
\mathbf{X}^e \mid \beta^e &\sim p(\mathbf{X}^e \mid \beta^e) && \text{(observed data for environment e),}
\end{aligned}$$

where $e \in \mathcal{E}$ is environment with $|\mathcal{E}| = E$, $\mathbf{X}^e$ is the observed data and we denote the size N^e and $N = \sum_{e\in\mathcal{E}} N^e$.

The **Pooling Factor** quantifies information sharing (pooling) across environments in a BHM, (Gelman and Pardoe, 2006). Define for each environment and each covariate X_d the error term δ_d^e, which is the difference between the global and the local parameters, that is $\beta_d^e = \phi_d + \delta_d^e$. For each predictor X_d, the pooling factor is defined as,

$$\gamma_d = 1 - \frac{\text{Var}_{e\in\mathcal{E}}\left[\mathbb{E}\left[\delta_d^e\right]\right]}{\mathbb{E}\left[\text{Var}_{e\in\mathcal{E}}\left[\delta_d^e\right]\right]}. \tag{19}$$

Lemma 2 *Assume the BvM theorem holds for both ϕ and β^e for all $e \in \mathcal{E}$, then their deviations $\beta^e - \phi$ concentrates around the true difference $\beta^{e^*} - \phi^*$ with uncertainty shrinking to zero as $N^e, E \to \infty$.*

Proof *Assume the BvM theorem holds for both ϕ and β^e (we need sufficient data per environment $N^e \to \infty$ and sufficient number of environments $E \to \infty$, regular parametric model, etc), then the posterior distributions are asymptotically normal:*

$$p(\phi \mid \mathbf{X}) \to \mathcal{N}\left(\hat{\phi}, \frac{1}{E \cdot N^e} I_{\phi}^{-1}(\phi^*)\right) \tag{20}$$

$$p(\beta^e \mid \mathbf{X}^e) \to \mathcal{N}\left(\hat{\beta}^e, \frac{1}{N^e} I_{\beta^e}^{-1}(\beta^{e^*})\right) \tag{21}$$

With $\hat{\phi}, \hat{\beta}^e$ being consistent estimators for the true parameters ϕ^*, β^{e^*} and $I_{\phi}(\phi^*)$ the Fisher information for ϕ, based on the marginal model $p(\mathbf{X} \mid \phi)$ and $I_{\beta}(\beta^{e^*})$ the Fisher information for β^e, from the environment-specific likelihood $p(\mathbf{X}^e \mid \beta^e)$.

Under BvM the posterior distributions of $\hat{\phi}$ and $\hat{\beta}^e$ each become independent Gaussians in the limit, so their difference is also approximately Gaussian, with variance equal to the sum of variances, that is

$$p(\beta^e - \phi \mid x) \to \mathcal{N}\left(\hat{\beta}^e - \hat{\phi},\ \frac{1}{N^e}\left(I_{\beta^e}^{-1} + \frac{1}{E} I_{\phi}^{-1}\right)\right). \tag{22}$$

Thus, as $N^e, E \to \infty$ the posterior for the difference between local and global parameters converges to a degenerate distribution at the true difference $\beta^{e^*} - \phi^*$.

Lemma 3 (Asymptotic behavior of the posterior deviation ratio) *Assume a BHIP model. Under regularity conditions and the BvM theorem applied to ϕ and β^e then the posterior deviation ratio converges to*

$$R := \frac{\mathrm{Var}_e\left(\mathbb{E}[\delta^e \mid X]\right)}{\mathbb{E}\left[\mathrm{Var}_e(\delta^e) \mid X\right]} \xrightarrow{P} \begin{cases} 1 & \text{if } \mathrm{Var}_e(\beta^{e*} - \phi^*) > 0 \\ 0 & \text{if } \beta^{e*} = \phi^* \text{ for all } e \end{cases} \tag{23}$$

as $N \to \infty$ and optionally as $E \to \infty$.

Proof *The numerator $\mathrm{Var}_e\left(\mathbb{E}[\delta^e \mid X]\right)$ is the posterior-estimated between-environment variance in deviations and the denominator is the total posterior variance across environments. From the law of total variance we have that*

$$\mathbb{E}\left[\mathrm{Var}_e(\beta^e - \phi) \mid X\right] = \mathbb{E}_e\left[\mathrm{Var}_e(\beta^e - \phi) \mid X\right] + \mathrm{Var}_e\left(\mathbb{E}[\beta^e - \phi \mid X]\right)$$

where $\mathbb{E}_e\left[\mathrm{Var}_e(\beta^e - \phi) \mid X\right]$ is the posterior within-environment variance, so the ratio R is the proportion of posterior variability in the deviations δ^e that comes from differences in posterior means across environments.

Under regularity conditions and the BvM theorem applied to ϕ and β^e, then as for Lemma 2, the deviation is a consistent estimator with shrinking uncertainty, so

$$\mathbb{E}[\beta^e - \phi \mid X] \xrightarrow{P} \beta^{e*} - \phi^* \tag{24}$$

$$\mathrm{Var}(\beta^e - \phi \mid X) \xrightarrow{P} 0 \tag{25}$$

as $N \to \infty$ and optionally $E \to \infty$. Consequently:

$$\mathrm{Var}_e\left(\mathbb{E}[\beta^e - \phi \mid X]\right) \xrightarrow{P} \mathrm{Var}_e\left(\beta^{e*} - \phi^*\right) \tag{26}$$

$$\mathbb{E}_e\left[\mathrm{Var}(\beta^e - \phi \mid X)\right] \xrightarrow{P} 0. \tag{27}$$

Lets consider the limit of the ratio for two cases:

$$\lim_{N,E\to\infty} R = \frac{\mathrm{Var}_e(\beta^{e*} - \phi^*)}{\mathrm{Var}_e(\beta^{e*} - \phi^*) + \lim_{N,E\to\infty} \mathbb{E}_e\left[\mathrm{Var}(\beta^e - \phi \mid x)\right]}$$

(i) Heterogeneous true effects:
There exist environments e, e' such that $\beta^{e*} \neq \beta^{e'*}$, and $\mathrm{Var}_e(\beta^{e*} - \phi^*) > 0$ which implies that

$$\lim_{N,E\to\infty} R = 1 \tag{28}$$

(ii) Homogeneous true effects
For all $e \in \mathcal{E}$ we have that $\beta^{e*} = \phi^*$ so $\mathrm{Var}_e(\beta^{e*} - \phi^*) = 0$ then,

$$\lim_{N,E\to\infty} R = 0 \tag{29}$$

Hence the posterior deviation ratio, converges as $N \to \infty$ and optionally as $E \to \infty$ to

$$R = \frac{\mathrm{Var}_e\left(\mathbb{E}[\beta^e - \phi \mid X]\right)}{\mathbb{E}\left[\mathrm{Var}_e(\beta^e - \phi \mid X)\right]} \xrightarrow{P} \begin{cases} 1 & \text{if } \mathrm{Var}_e(\beta^{e*} - \phi^*) > 0 \\ 0 & \text{if } \beta^{e*} = \phi^* \text{ for all } e \end{cases} \tag{30}$$

Our main result arises as a Corollary to Theorem 3:

Theorem *(Asymptotic behavior of pooling factor) Suppose the invariance assumption (Assumption 1) and the Bernstein von Mises holds for a BHIP model, then*

$$\gamma_d = 1 - \frac{\mathrm{Var}_{e\in\mathcal{E}}\left[\mathbb{E}\left[\delta_d^e\right]\right]}{\mathbb{E}\left[\mathrm{Var}_{e\in\mathcal{E}}\left[\delta_d^e\right]\right]} \xrightarrow{P} \begin{cases} 1 & \text{if } \beta_d^{e*} = \beta_d^{e'*} \text{ for all } e, e' \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$$

Appendix D. Non-centered parameterization and inference

Bayesian inference often uses *Markov Chain Monte Carlo (MCMC)* methods to explore the posterior distribution of the model. When sampling from a hierarchical parameter space, the convergence of MCMC methods can depend crucially on the parameterization of unknown quantities (Papaspiliopoulos and Roberts, 2003).

A typical parameterization of the hierarchical model is the *non-centered parameterization*, capturing the interdependent relationship between the global and local latent parameters, and can be useful in moderating degeneracies inherent to sampling from a hierarchical latent parameters space (Betancourt, 2020).

The normal probability density functions are closed under translation and scaling, so in a normal hierarchical model, the non-centered parameterization generates the local-level parameters from a parameterization of an independent standard normally distributed parameter $\eta^e \sim \mathcal{N}(0,1)$, where β^e then is deterministically reconstructed as

$$\beta^e = \mu + \tau \cdot \eta^e \sim \mathcal{N}(\mu, \tau) \tag{31}$$

We implement the BHM with the aforementioned non-centered parameterization using NumPyro (Phan et al., 2019) and No-U-turn Hamiltonian Monte Carlo sampler (NUTS) (Hoffman and Gelman, 2014). NumPyro provides a backend for Pyro, developed by (Bingham et al., 2018), powered by JAX, developed by (Frostig et al., 2018), and is thus a lightweight probabilistic programming library. To efficiently sample both continuous and discrete parameters when using the spike-and-slab prior, we combine NUTS with a Gibbs sampler (Casella and George, 1992), which iteratively samples inclusion indicators z_d and coefficients β_d^e. This improves convergence and computational efficiency.

Appendix E. Sparse Priors: Horseshoe and Spike-and-Slab

Using a BHM enables us to take advantage of adjusting the priors of our model, for instance, to encourage sparsity (George and McCulloch, 1997), improving the flexibility of this work. We consider two widely used priors in Bayesian variable selection to illustrate the flexibility of BHIP: the horseshoe prior and the spike-and-slab prior.

Horseshoe Prior (Carvalho et al., 2010), is a continuous shrinkage prior particularly well suited for situations where a few predictors have strong effects while many are close to zero. It is defined as:

$$\beta_d^e \sim \mathcal{N}(0, \lambda_d^2 \tau^2), \tag{32}$$
$$\lambda_d \sim \text{Cauchy}^+(0,1), \tag{33}$$
$$\tau \sim \text{Cauchy}^+(0,1) \tag{34}$$

The shrinkage parameter λ_d allows large effects to escape strong shrinkage while small effects are pushed towards zero. The global scale τ controls overall sparsity, ensuring that only a few predictors significantly influence Y.

Spike-and-Slab Prior (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005) is a discrete mixture model that explicitly models sparsity. It assumes that each coefficient β_d is drawn from either a spike distribution represented by a Dirac delta function δ_0 centered around zero (for irrelevant predictors) or a slab distribution, allowing nonzero values, following a Normal distribution (for relevant predictors):

$$z_d \sim \text{Bernoulli}(\pi), \tag{35}$$
$$\beta_d^e \mid z_d = z_d \mathcal{N}(\mu_d, \tau_d^2) + (1 - z_d)\delta_0, \tag{36}$$
$$\tau_d \sim \text{Cauchy}^+(0,1) \tag{37}$$

The binary inclusion variable z_d determines whether the predictor is relevant ($z_d = 1$) or not ($z_d = 0$). The prior probability π controls the expected proportion of relevant predictors. In practice, when strict adherence to theoretical conditions is paramount, one can use a continuous spike-and-slab prior (Ishwaran and Rao, 2005, Example 2). These priors achieve the same goal by replacing the point mass with a continuous, highly concentrated distribution (for example, a Gaussian with near-zero variance), thereby satisfying the necessary regularity conditions. This is, in fact, what was used in all our results since it allows for more efficient MCMC sampling.

E.1. Sparse Priors on Bus Dwelling Problem

Table 4 compares BHIP (with sparse priors and decision rule: HDI out of ROPE $> 85\%$, $\gamma_d > 0.85$) against ICP ($\alpha = 0.05$). Covariates detected as invariant causal predictors are marked '✓'. BHIP with **spike-and-slab priors** showed increasing confidence in selecting true predictors X_3 and X_4 (inclusion probabilities $z_3, z_4 \to 1$) as N increased from 100 to 500, while correctly excluding others. Similarly, the **horseshoe prior** yielded higher shrinkage parameters (λ_d) for X_3 and X_4 with $N = 500$, indicating stronger effects, compared to negligible effects for X_0, X_1, X_2.

Table 4: Sparse priors and ICP results

X	$N = 100$				$N = 500$			
	z_d	λ_d	ICP	BHIP	z_d	λ_d	ICP	BHIP
X_0	0.03	0.50	-	-	0.00	0.43	-	-
X_1	0.00	0.24	-	-	0.00	0.16	-	-
X_2	0.09	0.63	-	-	0.00	0.37	-	-
X_3	0.89	7.32	✓	✓	1.00	12.50	✓	✓
X_4	0.57	2.36	-	-	1.00	4.64	-	✓

E.2. Sparse Priors on School Attainment Dataset

With sparse priors similar results are achieved as can be seen on Table 5, z_d values show that predictors *score* and *fcollege_yes* are selected. Similarly, then using the horseshoe prior, the two predictors that have the strongest effects are again *score* and *fcollege_yes*.

Table 5: Sparse priors results

Predictor	z_d	λ_d
score	1.00	3.40
unemp	0.00	0.42
wage	0.00	0.59
tuition	0.00	0.42
gender_male	0.00	0.71
ethnicity_hispanic	0.00	2.75
ethnicity_other	0.00	0.73
fcollege_yes	1.00	3.91
mcollege_yes	0.00	2.66
home_yes	0.00	3.01
urban_yes	0.00	0.83
income_low	0.00	0.75
region_west	0.00	0.82

Appendix F. Bus Dwelling Problem: Data Generation

To introduce realistic temporal variations, we employ sinusoidal functions for both daily and weekly patterns: The daily pattern captures the daily fluctuations in boarding and alighting passengers over the course of a day as well as the traffic state. It is parameterized by the peak hour, which determines the time of day when passenger activity is highest and the amplitude which controls the magnitude of daily variation. The weekly pattern reflects weekly variations in traffic conditions. It is parameterized by the peak day which specifies the day of the week with the highest traffic impact and the amplitude that controls the magnitude of weekly variation. These patterns

allow us to simulate realistic fluctuations in passenger activity and traffic conditions over time. Each bus stop in our experiment is characterized by specific parameters that influence dwell time:

Traffic State, X_2: Simulates traffic conditions at a given bus stop, including daily and weekly fluctuations. It does not directly cause dwell time.

Boarding, X_3, and **Alighting,** X_4, **Passengers**: Modeled as Poisson random variables, these parameters capture the stochastic nature of passenger arrivals and departures at a bus stop. Daily and weekly patterns modulate their base rates and peak times.

On a randomized setup, the aforementioned predictors can be represented by the data represented in Figure 2(*b*) with $N = 500$ for each bus stop. Using the parameters described above, dwell time Y is generated as a function of boarding passengers X_3 and alighting passengers X_4. The function $f(X_3, X_4)$ incorporates coefficients that scale with the number of boarding and alighting passengers. The noise term ϵ represents independent noise in dwell time, capturing unpredictable factors not accounted for by the model.

This setup ensures that we can systematically evaluate BHIP's ability to identify and distinguish causal factors influencing bus dwell time from observational noise and spurious correlations.

Appendix G. Educational Attainment Additional Results

This section serves to present additional figures of the Bayesian Hierarchical inference on the educational attainment case study.

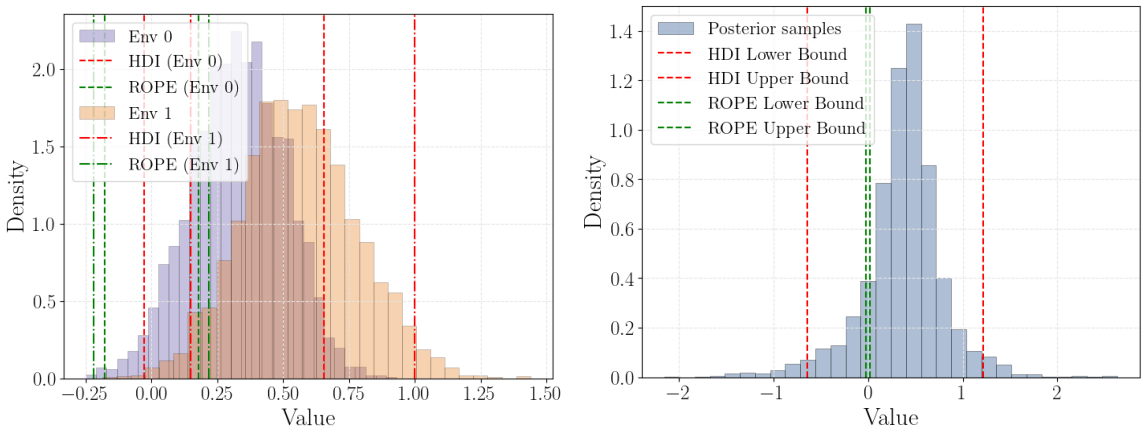


Figure 5: HDI + ROPE for local parameters left (Beta score) and global parameters right (mu score) of predictor: `fcollege_yes`. Pooling factor of $\gamma_d = 0.92$.

Appendix H. Synthetic Results

To rigorously validate BHIP's capability to recover the invariant parental set for each node, we conducted a significantly expanded suite of simulation experiments comparing its performance against ICP. The experimental design encompassed 18 distinct configurations, systematically varying the number of nodes $N \in \{4, 5, 6\}$, the number of samples per environment $S \in \{50, 500, 2000\}$, and the number of environments $E \in \{2, 3\}$. Each setup included one observational environment

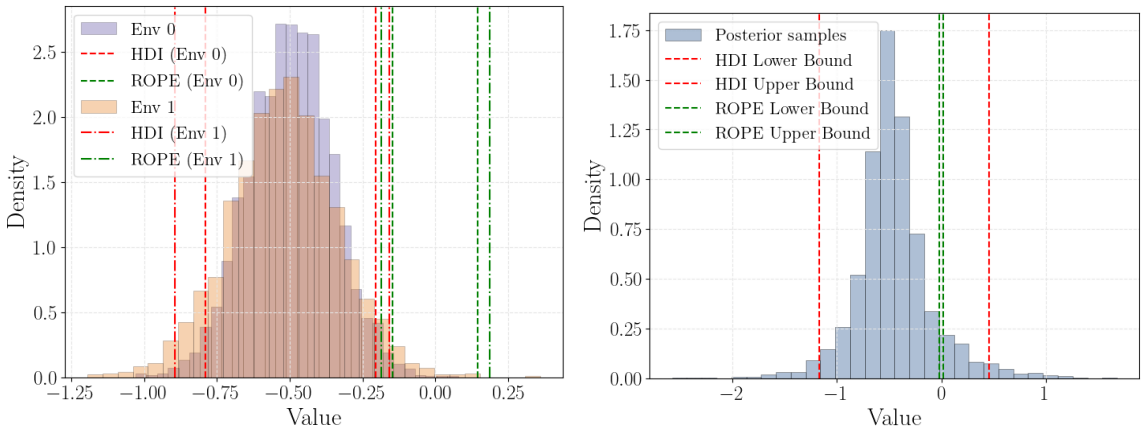


Figure 6: HDI + ROPE for local parameters left (Beta score) and global parameters right (mu score) of predictor: `income_low`. Pooling factor of $\gamma_d = 0.99$.

and $E - 1$ interventional environments, where each intervention consisted of a distinct single-node manipulation.

For each of the 18 configurations (N, S, E), we generated 1000 random DAGs. Corresponding SCMs were generated based on Linear Gaussian Additive Noise Models (LGANMs). Following the setup, we employed noise variance, drawn from a $U(0, 0.3)$ distribution, and causal effects, drawn from a $U(1, 5)$ distribution (Gamella et al., 2025).

For BHIP's evaluation specific decision rules were employed. To assess parameter stability across environments, we utilized the HDI+ROPE method both on local and on global parameters. The ROPE was defined symmetrically around zero as the interval $[-0.1\hat{s}, +0.1\hat{s}]$, where $\hat{s}$ represents the sample standard deviation of the target variable. A parameter stability test was considered 'passed' if the posterior probability of the parameter value falling outside this ROPE was greater than 0.95. This is equivalent to ensuring that 95% of the posterior HDI mass lies outside the ROPE. Separately, a decision rule based on a pooling factor was employed, requiring this factor to exceed a threshold of 0.85 to pass. The final acceptance of a candidate parent set in these simulations required satisfying criteria based on both the HDI+ROPE rule and the pooling factor rule. For comparison, ICP was evaluated using a standard significance level of $\alpha = 0.05$.

Performance was quantified by evaluating the recovery of the true parental set for each node, using the F1 score, Recall, Precision, and Specificity metrics. The reported results for each configuration represent the average performance across the 1000 independent runs. The comparative performance metrics for BHIP and ICP across all configurations are presented in Table 6 and several key observations can be made:

BHIP generally outperforms ICP in terms of F1 score, Recall, and Precision across a majority of the configurations (14 out of 18). The advantage of BHIP is particularly pronounced in scenarios with larger sample sizes ($S = 500$ or $S = 2000$). As expected, performance for both methods improves significantly with increasing sample size per environment. BHIP's relative advantage over ICP often widens as S increases, suggesting BHIP benefits more effectively from larger datasets within each environmental context. Increasing the number of environments from E=2 (one observational, one interventional) to E=3 (one observational, two distinct interventions) generally boosts performance for both methods.

In summary, the simulation results indicate that BHIP offers a substantial improvement over ICP for invariant parent set recovery under a wide range of conditions, particularly demonstrating robustness with fewer environmental interventions $E = 2$ and leveraging increased sample sizes S more effectively.

Table 6: Comparison of BHIP vs ICP Metrics

Config	f1_score		Recall		Precision	
	BHIP	ICP	BHIP	ICP	BHIP	ICP
N=4,S=50,E=2	**0.2003**	0.1538	**0.1898**	0.1391	**0.2225**	0.1857
N=4,S=50,E=3	0.3830	**0.4536**	0.3693	**0.4295**	0.4110	**0.5035**
N=4,S=500,E=2	**0.3755**	0.1659	**0.3600**	0.1525	**0.4123**	0.1959
N=4,S=500,E=3	**0.5621**	0.5035	**0.5397**	0.4837	**0.6088**	0.5438
N=4,S=2000,E=2	**0.4948**	0.1928	**0.4790**	0.1762	**0.5338**	0.2277
N=4,S=2000,E=3	**0.6411**	0.5310	**0.6282**	0.5132	**0.6695**	0.5680
N=5,S=50,E=2	0.0732	**0.0830**	0.0675	**0.0711**	0.0860	**0.1123**
N=5,S=50,E=3	0.2313	**0.2436**	0.2176	**0.2177**	0.2615	**0.3029**
N=5,S=500,E=2	**0.2258**	0.0907	**0.2122**	0.0786	**0.2560**	0.1171
N=5,S=500,E=3	**0.3690**	0.2673	**0.3548**	0.2478	**0.4053**	0.3090
N=5,S=2000,E=2	**0.3073**	0.1139	**0.2948**	0.1018	**0.3399**	0.1404
N=5,S=2000,E=3	**0.4574**	0.3145	**0.4444**	0.2898	**0.4967**	0.3692
N=6,S=50,E=2	0.0241	**0.0302**	0.0219	**0.0241**	0.0290	**0.0470**
N=6,S=50,E=3	**0.1595**	0.1234	**0.1512**	0.1035	**0.1793**	0.1717
N=6,S=500,E=2	**0.1113**	0.0419	**0.1024**	0.0346	**0.1325**	0.0596
N=6,S=500,E=3	**0.2529**	0.1671	**0.2424**	0.1454	**0.2866**	0.2182
N=6,S=2000,E=2	**0.1924**	0.0661	**0.1777**	0.0560	**0.2306**	0.0903
N=6,S=2000,E=3	**0.3383**	0.2250	**0.3254**	0.2000	**0.3812**	0.2813

Appendix I. Synthetic Benchmark Extension

Following the experimental setup in Section 4.4.1, this section provides an extended evaluation of the linear-Gaussian synthetic benchmark using traditional machine learning classification metrics. Figure 7 illustrates the Precision, Recall, and F1-score for BHIP and the competing invariant prediction methods across a varying number of environments ($E \in \{2, 5, 10, 20\}$). By analyzing these metrics, we can better contextualize the exact discovery and coverage results presented in the main text.

Both BHIP and ICP maintain a perfect precision of 1.0 across all evaluated environments. This confirms the Coverage result, mirroring the strict Type I error control of ICP. Conversely, methods like EILLS and Hidden-ICP exhibit significantly lower precision in the low-environment regime ($E = 2$ and $E = 5$), explaining their lower Coverage despite having high recall.

While ICP is highly precise, it is also conservative as reflected in its lower recall compared to the BHIP. BHIP recall starts conservatively at $E = 2$ but climbs steeply as E increases, efficiently leveraging the added environmental heterogeneity to identify the true invariant set and surpassing ICP.

Table 7: Sparse priors variable selection results for the TU dataset

Predictor	z_d	λ_d
RespHasBicycle	0.82	3.92
HousehNumcars	0.62	2.69
IncHouseh	0.05	0.46
RespAgeSimple	0.14	0.88
RespSex	0.15	0.56
DiaryDaytype	0.14	0.47
RespEdulevel	0.01	0.76

As seen in the rightmost panel, BHIP provides a good trade-off. BHIP's F1-score significantly outperforms traditional ICP and Hidden-ICP, and it closely tracks the state-of-the-art performance of BIP and VI-BIP. This demonstrates BHIP's capacity to maintain the rigorous Type I error control while overcoming their power deficiencies (recall) through hierarchical Bayes.

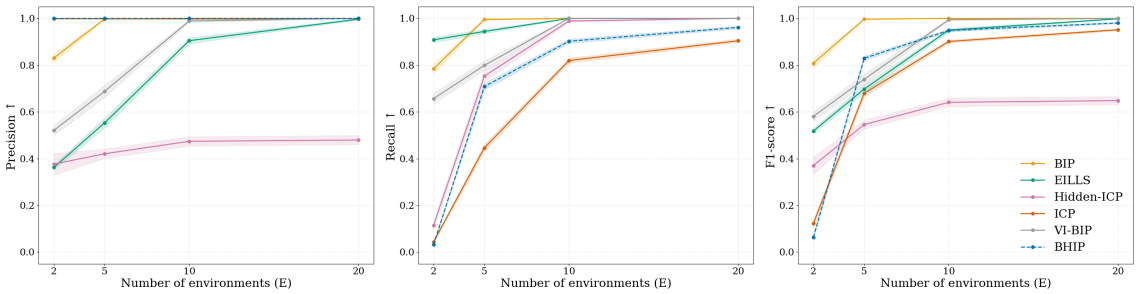


Figure 7: Extended low-dimensional synthetic benchmark. Precision, Recall, and F1-score shown as a function of the number of environments E.

Appendix J. TU Dataset Results

BHIP was also applied to the TU Danish Travel Survey Dataset which surveys the transport behaviour of Danish people residing in Denmark (Christiansen and Skougaard, 2015). The research question can be described as: what are the causal predictors of the primary mode of transportation choice?

After some data cleaning and feature selection, the predictors considered the respondent's age, sex, education level, purpose of trip as well as if the respondent has a bicycle. The data was split in two environments depending on the correspondent's distance from work to home. Additionally, the number of cars and the income of the respondent's household are also considered.

Both the possession of a bicycle as well as the number of cars of the respondent's household are predictors with a considerable effect on the primary transport mode of choice of the respondents. Furthermore, RespHasBicycle has a pooling factor of 0.98 and HousehNumcars has a pooling factor of 0.99. When considering sparse priors these results are validated and both spike-and-slab as well as horseshoe prior models select these variables as the most relevant, as seen on Table 7.