BAYESIAN INVARIANCE MODELING OF MULTI-ENVIRONMENT DATA

Luhuan Wu
Columbia University
New York, NY 10025, USA
lw2827@columbia.edu

Mingzhang Yin
University of Florida
Gainesville, FL 32611, USA
mingzhang.yin@warrington.ufl.edu

Yixin Wang
University of Michigan
Ann Arbor, MI 48104, USA
yixinw@umich.edu

John P. Cunningham & David M. Blei
Columbia University
New York, NY 10025, USA
{jpc2181, david.blei}@columbia.edu

ABSTRACT

Peters et al. (2016) introduced the problem of invariant modeling. In this problem, we observe feature/outcome data from multiple environments and our goal is to identify a set of *invariant features*, those that maintain a stable predictive relationship with the outcome. Identifying such features is important for robust generalization to new environments and for uncovering causal mechanisms. While previous methods primarily tackle this problem through hypothesis testing or regularized optimization, we take a Bayesian approach. We develop a probabilistic model of multi-environment data where the indices of the invariant features are encoded as a latent variable. Under the data-generating assumptions as Peters et al. (2016), we show that posterior inference in our model targets the true invariant features. We prove that this posterior is consistent and we provide theoretical results about the posterior contraction rate. In particular, we show that, under a certain metric, greater heterogeneity among environments leads to a faster contraction of the posterior. When the number of features is large, we design an efficient variational inference algorithm to approximate the posterior. In both simulations and real-world data, we show that Bayesian invariance is more accurate and scalable than existing approaches.

1 INTRODUCTION

An important goal of statistics is to identify features that have stable effects on the outcome of interests across varying settings, which is crucial for building robust prediction models and uncovering causal mechanisms. To achieve this goal, Peters et al. (2016) leverages the invariance idea when given access to data collected from multiple environment. In this setting, each environment $e \in \mathcal{E}$ indexes a unique joint distribution $p_e(x, y)$ of the features $x \in \mathbb{R}^p$ and the outcome y, where p is the feature dimension and $\mathcal{E}$ is the set of all environments of interest. They assume the existence of a subset of features $x^* \subset x$ such that the conditional distribution $p_e(y \mid x^*)$ remains the same across environments. We call these features x^* the *invariant features*.

To identify these features, previous methods primarily focus on enforcing some invariance property on the *conditional* model of y given candidate invariant features. For example, Peters et al. (2016) proposed a hypothesis testing framework, and Fan et al. (2023) formulated a linear regression objective with a regularization term on the model residuals. However, these approaches lack a principled characterization of the underlying data generative process. Moreover, they often struggle to scale to high-dimensional settings.

In this work, we propose an alternative perspective by modeling the *joint* distribution of features and outcome through a generative process that incorporates the invariance assumption. In our data generative process, we begin with a latent variable $z \in \{0, 1\}^p$ that selects a candidate set of invariant

features $x^z := \{x^{(i)} : z^{(i)} = 1\}$ associated with a prior $p(z)$. Given z, each environment's data x, y is drawn from an environment-specific likelihood model that encodes an invariance property. Specifically, the conditional distribution of y given selected features x^z is assumed to be invariant across environments, while the distribution of the selected features x^z and that of the de-selected features $x^{-z} := \{x^{(i)} : z^{(i)} = 0\}$ conditioned on others are allowed to vary.

We estimate the invariant features through posterior inference on the latent variable z. While our model involves the joint distribution of x, y, we derive a simplified posterior expression that only requires conditional modeling of y given x^z within each environment and across pooled environments. This simplification enhances the computational efficiency while maintaining theoretical rigor.

Our theoretical results show that the posterior distribution is consistent. Moreover, we characterize the posterior contraction rate and reveal factors that influence the rate including the prior specification and the heterogeneity level of environments.

In high-dimensional settings, we develop a variational inference algorithm to approximate the posterior of z. By optimizing a variational objective, our approach avoids the exponential complexity of exact inference and previous methods.

We validate our theoretical findings through extensive simulations. Additionally, We compare our method against existing works in both simulated and real-world gene-perturbation datasets (Kemmeren et al., 2014). Our results show that our approach, particularly when combined with variational inference, outperforms existing methods in inference accuracy and scalability across various tasks. For a detailed discussion of related works, including comparisons to our approach, see Appendix A.

We summarize our contributions in the following:

1. We propose a Bayesian model to infer invariant features from multi-environment data.
2. We establish theoretical guarantees on posterior consistency and contraction rates.
3. We develop a scalable variational inference algorithm that overcomes the exponential complexity of previous methods in high-dimensional settings.
4. Our method with variational inference outperforms existing approaches in accuracy and scalability across simulation and real-world studies.

2 METHOD

2.1 DATA FROM MULTIPLE ENVIRONMENTS

The data $\mathcal{D} := \{\{x_{ei}, y_{ei}\}_{i=1}^{n_e}\}_{e=1}^{E}$ is organized into multiple *environments*. Each environment e indexes a distinct data distribution $p_e(x, y)$, and $x_{ei}, y_{ei} \overset{i.i.d.}{\sim} p_e(x, y), \forall e$. Data from different environments are drawn independently. For a collection of environments $\mathcal{E}$ of interests, we make the following assumption,

Assumption 1 (Invariance)**.** *$\exists z^* \in \{0, 1\}^p$ such that $p_e(y \mid x^{z^*})$ is invariant with respect to $e \in \mathcal{E}$.*

We call z^* the *(true) invariant feature selector* and x^{z^*} the *(true) invariant features*. As a consequence of Assumption 1 and the chain rule, each environment's joint distribution factorizes,

$$p_e(x, y) = p_e(x^{z^*}) p_*(y \mid x^{z^*}) p_e(x^{-z^*} \mid x^{z^*}, y) \tag{1}$$

where $p_*(y \mid x^{z^*}) \equiv p_e(y \mid x^{z^*})$ denotes the invariant conditional distribution that is independent with the environment e, $\forall e \in \mathcal{E}$.

In reality, we do not observe the invariant feature selector z^*. Our goal in this work is to infer z^* from the observed multi-environment data.

2.2 A BAYESIAN MODEL FOR INFERRING INVARIANT FEATURES

We propose a Bayesian model to infer z^*. Our model posits a prior $p(z)$ over all candidate invariant feature selectors z in $\{0, 1\}^p$. Given multi-environment data $\mathcal{D}$, we will show that the posterior distribution of z from our model centers on the true value z^*.

For now, we assume the per-environment joint distributions $\{p_e(x, y)\}_e$ are known, as in practice, estimating $p_e(x, y)$ is feasible with a sufficiently large dataset. Consequently, for a candidate invariant feature selector z, we observe the marginal $p_e(x^z)$, conditionals $p_e(y|x^z)$, and $p_e(x^{-z} \mid x^z, y)$, which are derived from $p_e(x, y)$. We call $p_e(y|x^z)$ the *per-environment conditional* of y given a candidate set of invariant features x^z, as it is the conditional within each environment. With the same z, we also define the *pooled conditional* as the following

Definition 1 (Pooled conditional)**.**

$$p(y \mid x^z) := \frac{\iint_{\mathcal{E}} p(e)p_e(x, y)\,\mathrm{d}e\,\mathrm{d}x^{-z}}{\int_{\mathcal{E}} p(e)p_e(x^z)\,\mathrm{d}e}, \quad p(e) \propto 1. \tag{2}$$

The pooled conditional $p(y \mid x^z)$ represents the conditional distribution of y given x^z when we pool all environments in $\mathcal{E}$. Importantly, the pooled conditional for $z = z^*$ recovers the invariant conditional distribution, that is,

Proposition 1. *The pooled conditional matches the local conditional, i.e.* $p(y|x^z) = p_e(y|x^z)$, $\forall e \in \mathcal{E}$*, only for* $z = z^*$ *with* z^* *any value satisfying Assumption 1.*

The proof is included in Appendix B.1.

Consequently, incorporating the pooled conditional into our model will prove useful to show that, with the correct $z = z^*$, the model recovers the true data generating process in Section 2.1.

Our Bayesian model. With the above ingredients in place, our model is the following

- Draw invariant feature selector $z \sim p(z)$.
- For each environment $e = 1, \cdots, E \in \mathcal{E}$:
  For each observation $i = 1, \cdots, n_e$:
  1. Draw the invariant features $x_{ei}^z \sim p_e(x_{ei}^z)$.
  2. Draw outcome $y_{ei} \mid x_{ei}^z \sim p(y_{ei} \mid x_{ei}^z)$.
  3. Draw the other features $x_{ei}^{-z} \mid x_{ei}^z, y_{ei} \sim p_e(x_{ei}^{-z} \mid x_{ei}^z, y_{ei})$.

Why does this model help us estimate the true z^*? Consider any z. The true data generating distribution for y given x^z in step 2 is the per-environment conditional $p_e(y|x^z)$, which may change with e. Therefore, the pooled conditional $p(y \mid x^z)$ cannot match all $p_e(y \mid x^z)$, leading to a low data likelihood. But for $z = z^*$, $p(y \mid x^{z^*})$ matches all $p_e(y \mid x^{z^*})$ by Proposition 1, leading to a high data likelihood. This intuition is made precise in the following posterior inference outcome.

Exact posterior inference. We derive a simplified posterior expression which only involves the prior and the ratio of pooled conditionals to local conditionals. We first re-write the joint distribution

$$p(z, \mathcal{D}) = p(z) \prod_{e=1}^{E} \prod_{i=1}^{n_e} p_e(x_{ei}^z) p(y_{ei} \mid x_{ei}^z) p_e(x_{ei}^{-z} \mid x_{ei}^z, y_{ei}) \tag{3}$$

$$= p(z) \prod_{e=1}^{E} \prod_{i=1}^{n_e} p_e(x_{ei}^z) p(y_{ei} \mid x_{ei}^z) p_e(x_{ei}^{-z} \mid x_{ei}^z, y_{ei}) \frac{p_e(y_{ei} \mid x_{ei}^z)}{p_e(y_{ei} \mid x_{ei}^z)}. \tag{4}$$

In eq. (4), each $p_e(x_{ei}^z) p_e(y_{ei} \mid x_{ei}^z) p_e(x_{ei}^{-z} \mid x_{ei}^z, y_{ei}) = p_e(x_{ei}, y_{ei})$. So,

$$p(z, \mathcal{D}) = p(z) \prod_{e=1}^{E} \prod_{i=1}^{n_e} p_e(x_{ei}, y_{ei}) \frac{p(y_{ei} \mid x_{ei}^z)}{p_e(y_{ei} \mid x_{ei}^z)}. \tag{5}$$

Because $p_e(x_{ei}, y_{ei})$ in eq. (5) is a constant factor with respect to z, we have the posterior

$$p(z \mid \mathcal{D}) \propto \underbrace{p(z)}_{\text{prior}} \underbrace{\prod_{e=1}^{E} \prod_{i=1}^{n_e} \frac{p(y_{ei} \mid x_{ei}^z)}{p_e(y_{ei} \mid x_{ei}^z)}}_{\text{likelihood ratio}}. \tag{6}$$

This expression only involves the prior and the likelihood ratio $\Lambda(\mathcal{D} \mid z) := \prod_{e=1}^{E} \prod_{i=1}^{n_e} \frac{p(y_{ei}|x_{ei}^z)}{p_e(y_{ei}|x_{ei}^z)}$ between the pooled and per-environment conditionals. It further provides insights into the limiting behavior of the posterior distribution for theoretical analysis. With sufficient amount of data, the posterior is dominated by the likelihood ratio $\Lambda(\mathcal{D} \mid z)$. For arbitrary z, the term $p_e(y_{ei} \mid x_{ei}^z)$ in the denominator of $\Lambda(\mathcal{D} \mid z)$ tends to be no smaller than the term $p(y_{ei} \mid x_{ei}^z)$ in the numerator, since $p_e(y_{ei} \mid x_{ei}^z)$ is the true data distribution. But for $z = z^*$, they are equal by Proposition 1. Consequently, the posterior will concentrate on z^* provided that its prior $p(z^*) > 0$. We formalize this reasoning in Section 3, where we establish posterior consistency and contraction rates.

In practice, we do not know per-environment conditionals $p_e(y \mid x^z)$ and pooled conditionals $p(y \mid x^z)$. We estimate these distributions from data with maximum likelihood. See Appendix B.2.

Variational inference. Computing the exact posterior in Equation (6) requires enumerating 2^p possible values of z, which is intractable for high-dimensional settings. To address this challenge, we develop a variational inference algorithm (Blei et al., 2017) with the following variational posterior

$$q_\phi(z) \coloneqq \prod_{j=1}^{p} \text{Bernoulli}\left(z^{(j)} \mid \text{sigmoid}(\phi^{(j)})\right) \tag{7}$$

where $\text{sigmoid}(\cdot) \coloneqq \frac{\exp(\cdot)}{1+\exp(\cdot)}$, and $\phi \in \mathbb{R}^p$ is the variational parameter. We optimize q_ϕ by maximizing the evidence lower bound. See details in Appendix B.3.

3 THEORY

We establish the main theoretical results: (1) the posterior distribution $p(z \mid \mathcal{D})$ concentrates around the true value z^* as either the number of environments E or the number of observations per environment n tends to infinity, under suitable assumptions; and (2) the posterior contracts exponentially in n and E, with a rate that improves under stronger prior knowledge or greater environment heterogeneity. These results apply when the true data distributions $p_e(x, y)$ are available. At the end of this section, we discuss how these results extend to settings where the data distributions are estimated from finite data, and how they are modified when certain assumptions are relaxed for practical considerations.

Without loss of generality, we assume the per-environment sample size is the same, i.e. $n_e = n$, and write the observed data as $\mathcal{D}_{n,E} = \{\{x_{ei}, y_{ei}\}_{i=1}^{n}\}_{e\in\mathcal{E}}$. We use $\xrightarrow{P}$ to denote convergence in probability. All relevant assumptions and proofs are included in Appendix C.

Theorem 1 (Posterior consistency)**.** *Given Assumptions 1 and 2, as $n \to \infty$ (fixing E) or as $E \to \infty$ (fixing n), we have*

- *posterior mode consistency, i.e. $z_{n,E} := \arg\max_z p(z \mid \mathcal{D}_{n,E}) \xrightarrow{P} z^*$, and*
- *posterior consistency at z^*, i.e. $p(z^* \mid \mathcal{D}_{n,E}) \xrightarrow{P} 1$.*

Theorem 1 shows that our Bayesian model is faithful: the posterior will concentrate on the invariant z^* as we observe more environments and/or more data points per environment.

Theorem 2 (Posterior contraction rate)**.** *Given Assumptions 1, 2 and 3, there exists a sequence $\epsilon_{n,E} = O(R \cdot (|supp(p(z))| - 1) \cdot e^{-\kappa\mu_{\min}\cdot nE})$ such that*

$$P\left(TV\Big(p(z \mid \mathcal{D}_{n,E}), \delta_{z^*}(z)\Big) > \epsilon_{n,E}\right) \to 0, \tag{8}$$

as $n \to \infty$ or $E \to \infty$, where

$$R := \max_{z\neq z^*} p(z)/p(z^*), \tag{9}$$

$$\mu_{\min} : \min_{z\neq z^*, z\in supp(p(z))} \mathbb{E}_{p(e)p_e(x^z)} KL(p_e(y \mid x^z) \| p(y \mid x^z)), \tag{10}$$

and κ is a fixed value in $(0, 1)$.

Theorem 2 demonstrates the exponential contraction of the posterior at z^* with respect to n and E. Additionally, it reveals key factors that influence the rate: (1) The prior. when the prior assigns less mass to non-invariant zs, the resulting R value is smaller, leading to a faster contraction rate. Additionally, if the prior is supported on fewer non-invariant zs, the factor $(|\text{supp}\,(p(z))\,| - 1)$ in the rate expression is smaller, also resulting in faster contraction. (2) The heterogeneity of environments encoded through $\mu_{\min}$. When environments $p_e(x, y)$ are more similar, we expect the per-environment conditionals $p_e(y \mid x^z)$ to be also more similar. Consequently, these conditionals are closer to the pooled $p(y \mid x^z)$, resulting in a smaller $\mu_{\min}$ and a slower contraction. Conversely, when environments are more dissimilar, $\mu_{\min}$ tends to be larger and the posterior contracts faster.

Extension to the finite-sample case. In practice, we do not observe the true distributions but only finite samples from each environment. In this setting, we estimate the relevant distributions from data and compute the posterior as described in Equation (16). The key insight is that, when each environment has a sufficiently large number of samples, these estimated distributions become accurate, and the same theoretical guarantees—posterior consistency and contraction—continue to hold under mild conditions.

We analyze two asymptotic regimes. First, when the environment size E is fixed and per-environment sample size n grows to infinity, the estimated pooled and local conditionals converge to their true counterparts under consistency assumptions, and our theoretical results extend naturally. Second, when n is fixed and E grows, consistency for the pooled conditional models can still be achieved. However, the estimated local models remain biased due to limited data per environment. In this case, we require an *additional* assumption to ensure that the true invariant feature can still be identified despite the estimation bias. When n is sufficiently large, the bias becomes negligible, and the assumption is expected to hold. Formal statements and proofs are provided in Appendix C.3.

Discussion on relaxations of conditions. When exact invariance i.e. Assumption 1, does not hold, the posterior would concentrate on the most approximately invariant selector, defined as $\arg\min_{z \in \text{supp}(p(z))} \mathbb{E}_{p(e)p_e(x^z)}\text{KL}(p_e(y \mid x^z) \| p(y \mid x^z))$, minimizing the discrepancy between per-environment and pooled conditionals. When there are multiple minimizers, invalidating Assumption 2, the posterior would concentrate on all these values, with densities proportional to their prior support. See a rigorous analysis in Appendix C.4.

4 Experiments

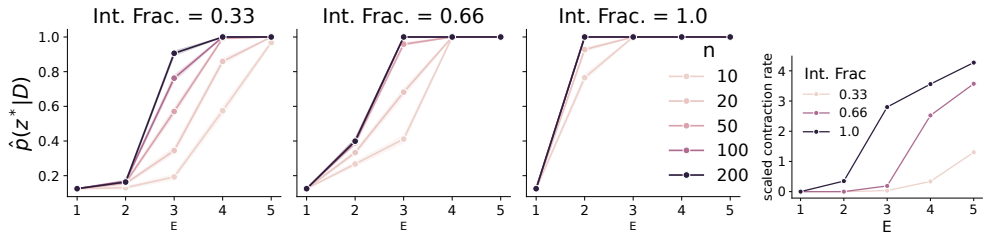
We verify our theory and compare our method to existing approaches in simulations and real-world datasets. All experiments use a linear Gaussian model for the conditional relationship between outcome y and selected features x^z. Additional details and results are included in Appendix E.1.

4.1 Simulation study

Empirical verification of theory. We study the empirical behavior of the posterior in simulations with $p = 3$ features. We vary n from 10 to 200, E from 1 to 5, and intervention fraction (controlling the diversity of environments) from 0.33 to 1. Figure 1a shows the posterior value at the true z^*, $\hat{p}(z^*|\mathcal{D})$. We observe that for each level of intervention fraction, $\hat{p}(z^*|\mathcal{D})$ converges to 1 with increasing n or E, corroborating our theory on posterior consistency. Furthermore, at a higher level of intervention fraction, the empirical rate at which $\hat{p}(z^*|\mathcal{D})$ converges to 1 is faster.

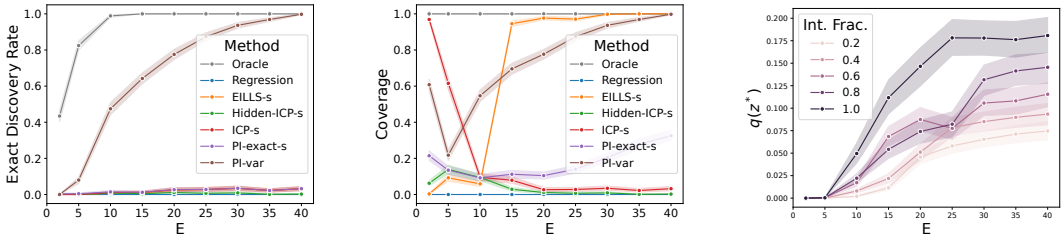
Figure 1b plots the theoretical posterior contraction rate against the number of environments E under different levels of intervention fraction. The rate is plotted in log scale and normalized by a factor of $1/nE$. We observe that the contraction rate increases with a greater number of environments and a higher intervention fraction. In either case, the observed environments become more heterogeneous. Notably, the theoretical rate shown in Figure 1b aligns with the empirical trends in Figure 1a.

Comparison to other methods. In simulations with $p = 450$ features, we compare our method, using both exact inference (*PI-exact-s*) and variational inference (*PI-var*), against existing approaches: (1) *Oracle* – linear regression on the true invariant features; (2) *Regression* – linear regression on all features; (3) *ICP-s* (Peters et al., 2016); (4) *Hidden-ICP-s* (Rothenhäusler et al.,



(a) Posterior value at true z^* (b) Posterior contraction rate

Figure 1: Empirical verification of theory in simulation with $p = 3$ features. **Left**: The posterior value at the invariant z^* increases with larger value of n, E, and intervention fraction. Results are averaged over 1,000 simulations with 95% confidence bands. **Right**: The theoretical posterior contraction rate increases with more environments, as well as higher intervention fraction.



(a) Exact discovery rate (b) Coverage (c) Variational posterior at z^*

Figure 2: Comparison to other methods in simulation with $p = 450$ features. **Left:** the exact discovery rate of Oracle and PI-var converges to 1 as E increases, while the rate for other methods remain low. **Middle:** Oracle maintains near-perfect coverage, and the coverage of PI-var and EILLS-s approaches to 1 as E increases; the remaining methods show low coverage. **Right** variational posterior value at z^* by PI-var increases with larger E and intervention fraction. Results are averaged over $n = 50, 200, 500$, each with 400 simulations. Error bars indicate 95% confidence intervals.

2019); and (5) *EILLS-s* (Fan et al., 2023). PI-exact-s, ICP-s, Hidden-ICP-s, and EILLS-s require a screening step to pre-select 10 features.

Figure 2 displays the results of (i) *exact discovery rate* – the average occurrence where the inferred invariant features recover the true invariant features, and (ii) *coverage* – the average occurrence where the inferred invariant features form a subset of the true invariant features.

Figure 2a shows that Oracle and PI-var achieve an exact discovery rate approaching 1 as E increases, while other methods remain near zero due to failures in pre-screening. Figure 2b demonstrates that PI-var's coverage improves with more environments, approaching Oracle's near-perfect coverage. Notably, EILLS achieves high coverage at large E by predicting an empty invariant feature set, while other methods maintain low coverage. Figure 2c illustrates the variational posterior density $q(z^*)$ from PI-var, which increases with E and intervention fraction.

These results show that PI-var successfully recovers invariant features in high-dimensional settings given sufficient data, whereas other methods struggle due to challenges in pre-screening.

4.2 Gene Perturbation Study

We analyze a large-scale yeast gene perturbation dataset (Kemmeren et al., 2014), which includes mRNA expression for 6,170 genes. It consists of 262 observational samples ($e = 1$, no gene deletion) and 1,479 interventional samples ($e = 2$, each with a unique gene deletion). For each target gene, we infer invariant feature genes and validate them following Peters et al. (2016) by checking whether their deletion significantly changes the target gene expression. In addition to the methods introduced in Section 4.1, we include Lasso (Tibshirani, 1996) for comparison. All experiment details and additional results are included in Appendix E.2.

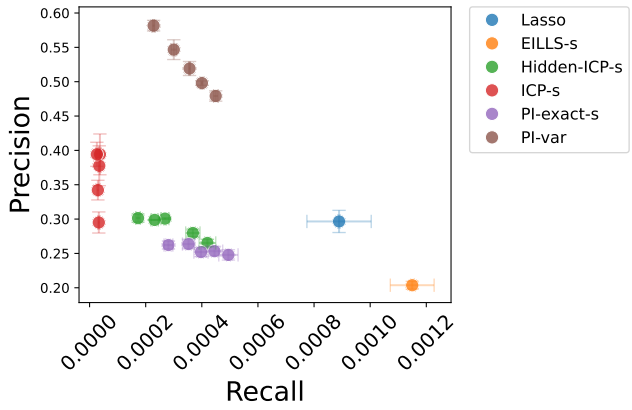


Figure 3: Gene perturbation study: precision v.s. recall. Each dot is the average result over 3 random seeds, with error bar indicating 2 standard errors. The color corresponds to a specific method, and multiple dots for the same method represent different hyperparameter settings. We observe that PI-var achieves a favorable trade-off between precision and recall: it has the highest precision among all methods, with a recall higher than ICP-s and comparable to PI-exact-s and Hidden-ICP-s.

To compare the performance among the above methods, we evaluate *precision* and *recall*. We check the effect of a predicted invariant feature gene on the target gene through the validation set. However, since the validation set is limited, not all predictions can be checked. Consequently, we define *precision* as the ratio of the number of the true predicted effects to the total number of predicted effects that can be checked through the validation set. Similarly, *recall* is defined as the ratio of the number of the true predicted effects to the total number of true effects that can checked.

We present the results of precision and recall in Figure 3, excluding Marginal as it is an outlier. Different methods demonstrate different trade-offs between precision and recall. The summary is that (i) PI-var is the most accurate method (with highest precision) and has a moderate recall; (ii) ICP-s is the second most accurate but also the most conservertive as indicated by the lowest recall; (iii) Other invariance inference methods either struggle with a relatively low precision or a relatively low recall; (iv) Marginal, which is not plotted in the figure, has the lowest precision of 0.11, but the highest recall of 0.87.

We comment on the low recall values across all methods except for Marginal. For methods that require screening, this low recall rate is expected as their predictions are limited within a maximum of 10 screened features. The recall of PI-var, which does not require screening, is also low. A possible explanation is that the true posterior given only two environments may be multi-modal, which the mean-field variational distribution may struggle to capture. Other factors include violations of invariance assumption or model mis-specifications. Lastly, inferring invariant features is only a proxy for detecting gene perturbation effect – it is possible that a non-invariant feature gene could still have an effect on the target gene when it is perturbed, which would not be captured by the invariance inference methods.

5 Discussion

This paper presents a Bayesian model to infer invariant features from multi-environment data. We establish theoretical guarantees on posterior consistency and contraction rates. In high-dimensional settings, we develop a scalable variational inference algorithm that mitigates the exponential complexity of previous approaches. We validate our method in both synthetic and real data studies, demonstrating its superior performance compared to existing approaches.

Our current method has some limitations: (1) model estimation bias from finite data may obscure posterior inference results, and (2) the mean-field variational family is restrictive. Future work will explore a full Bayesian model to better quantify model parameter uncertainty and develop more expressive variational family to improve inference fidelity.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Jianqing Fan, Cong Fang, Yihong Gu, and Tong Zhang. Environment invariant linear least squares. *arXiv preprint arXiv:2303.03092*, 2023.

Ragnar Frisch, Trygve Haavelmo, Tjalling C. Koopmans, and Jan Tinbergen. *Autonomy of Economic Relations*. Memorandum fra Universitets Socialøkonomiske Institutt. Universitets Socialøkonomiske Institutt, Oslo, Norway, 1948.

Yihong Gu, Cong Fang, Yang Xu, Zijian Guo, and Jianqing Fan. Fundamental computational limits in pursuing invariant causal prediction and invariance-guided regularization. *arXiv preprint arXiv:2501.17354*, 2025.

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

Kevin D. Hoover. Causality in economics and econometrics. In Steven N. Durlauf and Lawrence E. Blume (eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, UK, 2nd edition, 2008.

Patrick Kemmeren, Katrin Sameith, Loes AL Van De Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

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.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Zhenyu Wang, Yifan Hu, Peter Bühlmann, and Zijian Guo. Causal invariance learning via efficient optimization of a nonconvex objective. *arXiv preprint arXiv:2412.11850*, 2024.

Mingzhang Yin, Nhat Ho, Bowei Yan, Xiaoning Qian, and Mingyuan Zhou. Probabilistic best subset selection via gradient-based optimization. *arXiv preprint arXiv:2006.06448*, 2020.

A Related works

The idea of invariance closely relates to the independent causal mechanism in causal inference literature (Peters et al., 2017; Schölkopf et al., 2021), which states that the causal mechanism $p(y|x)$ and the cause distribution $p(x)$ are independent under the causal structure $x \to y$. This property implies the possibility of intervening on $p(x)$ while keeping $p(y|x)$ invariant. Invariance was also studied in econometrics under the terms autonomy and modularity (Frisch et al., 1948; Hoover, 2008), and was later introduced to computer science research as stable and autonomous parent-child relationships in a causal graph (Pearl, 2009, p. 22). For a historical overview of the invariance idea, please refer to Peters et al. (2017, Chapter 2.2) and Bühlmann (2020).

Peters et al. (2016) is the seminal work that formalized the concept of invariant predictions in machine learning models, and established its connection to causal inference. They assumed that the distribution of the outcome given a subset of features remains the same across environments, and under certain conditions, these features are the direct causes of the outcome. They developed a hypothesis testing approach for identifying such features in linear models, which requires an exhaustive search over exponentially many candidates. Heinze-Deml et al. (2018) later extended this framework to non-linear settings. Rothenhäusler et al. (2019) proposed an invariance notion based on the inner product between features and residuals in linear models, linking it to causal inference under additive interventions. Their estimator solves a modified linear system using differences of Gram matrices across environments, which poses some invertibility condition. Fan et al. (2023) incorporated residual-based invariance as a regularization term in linear regression but faced the limitation of exponential feature subset enumeration. However, their approach is limited by the need to enumerate all possible feature subsets exponentially. Wang et al. (2024) established a theoretical connection between invariant prediction models and the true causal outcome model, proposing a computationally efficient algorithm for causal discovery. Gu et al. (2025) further showed that solving for exact invariance is NP-hard and introduced an objective that interpolates between exact invariance and predictive performance. In contrast, our method focuses on the (approximate) pursuit of invariance without relaxing the objective. In this sense, our goal is most similar to Peters et al. (2016) and Fan et al. (2023).

Invariance can also be studied in general representation learning contexts: consider a potentially non-linear function $h(\cdot)$ such that $p(y \mid h(x))$ or $\mathbb{E}[y \mid h(x)]$ is invariant across environment. For example, Arjovsky et al. (2019) considers learning invariant representations through neural networks such that the empirical risk of the prediction is invariant across environments. The feature selection setting considered in this work is a special case where $h(\cdot)$ is a binary mask.

Broadly speaking, invariance is a useful tool to improve the generalization of machine learning models when given access to multi-environment data. There are other objectives that can be enforced on models to achieve similar goals. For example, Group DRO (Sagawa et al., 2019) focuses on optimizing the worst-group performance when provided with data from multiple groups. We do not compare to these methods in the paper as they rely on fundamentally different assumptions.

B Additional details for Section 2

B.1 Proof of Proposition 1

Proof Substituting the factorization of $p_e(x, y)$ in eq. (1) into $p(y \mid x^z)$ defined in eq. (2), we have for $z = z^*$ that

$$p(y \mid x^{z^*}) = \frac{\iint p(e)p_e(x^{z^*})p_*(y \mid x^{z^*})p_e(x^{-z^*} \mid x^{z^*}, y)\,\mathrm{d}e\,\mathrm{d}x^{-z^*}}{\int p(e)p_e(x^{z^*})\,\mathrm{d}e} \tag{11}$$

$$= p_*(y \mid x^{z^*})\frac{\iint p(e)p_e(x^{z^*})p_e(x^{-z^*} \mid x^{z^*}, y)\,\mathrm{d}e\,\mathrm{d}x^{-z^*}}{\int p(e)p_e(x^{z^*})\,\mathrm{d}e} \tag{12}$$

$$= p_*(y \mid x^{z^*}) = p_e(y \mid x^{z^*}) \quad \forall e. \tag{13}$$

On the other hand, for any z such that $p(y \mid x^z) = p_e(y \mid x^z) \forall e \in \mathcal{E}$, $p_e(y \mid x^z)$ is invariant across environments and hence z satisfies Assumption 1. ■

B.2 MODEL ESTIMATION

Let $\mathcal{P}_{y|x^z}$ be a class of conditional models indexed by z. Given the dataset $\mathcal{D}$, we estimate per-environment and pooled conditionals with maximum likelihood,

$$\hat{p}_e(y \mid x^z) := \underset{\tilde{p}(y|x^z)\in\mathcal{P}_{y|x^z}}{\arg\max} \sum_{i=1}^{n_e} \log \tilde{p}(y_{ei} \mid x_{ei}^z), \qquad \forall e, z \tag{14}$$

$$\hat{p}(y \mid x^z) := \underset{\tilde{p}(y|x^z)\in\mathcal{P}_{y|x^z}}{\arg\max} \sum_{e=1}^{E}\sum_{i=1}^{n_e} \log \tilde{p}(y_{ei} \mid x_{ei}^z), \qquad \forall z, \tag{15}$$

We write $\hat{p}(z \mid \mathcal{D})$ for the posterior distribution given estimated conditionals $\hat{p}_e(y \mid x^z)$ and $\hat{p}(y \mid x^z)$, that is,

$$\hat{p}(z \mid \mathcal{D}) \propto p(z) \prod_{e\in\mathcal{E}} \prod_{i=1}^{n_e} \frac{\hat{p}(y_{ei} \mid x_{ei}^z)}{\hat{p}_e(y_{ei} \mid x_{ei}^z)}. \tag{16}$$

We summarize the exact inference procedure with estimated conditionals in Algorithm 1, termed as *PI-exact*. This algorithm is exhaustive: it sweeps over all candidates z in the prior support and for each z it estimates the corresponding pooled and per-environment conditionals, and their likelihood ratio; finally it combines and normalizes the results over all zs to get the posterior.

B.3 VARIATIONAL INFERENCE

We optimize the variational parameter ϕ by maximizing the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\mathcal{D}, \phi) = \mathbb{E}_{q_\phi(z)} \left[\log p(z, \mathcal{D}) - \log q_\phi(z)\right] \tag{17}$$

$$= \mathbb{E}_{q_\phi(z)} \left[\log p(z) + \left(\sum_{e=1}^{E}\sum_{i=1}^{n_e} \log p(y_{ei} \mid x_{ei}^z) - p_e(y_{ei} \mid x_{ei}^z)\right) - \log q_\phi(z)\right] + C. \tag{18}$$

where we substitute the joint $p(z, \mathcal{D})$ with the expression from eq. (5), and $C = \sum_{e=1}^{E}\sum_{i=1}^{n_e} \log p(x_{ei}, y_{ei})$ is a term independent of the variational parameter ϕ.

In practice, we estimate the pooled $\hat{p}(y_{ei} \mid x_{ei}^z)$ and per-environment conditionals $\hat{p}_e(y_{ei} \mid x_{ei}^z)$ from data, and optimize the following objective:

$$l(\mathcal{D}, \phi) = \mathbb{E}_{q_\phi(z)} \left[f(z, \phi, \mathcal{D})\right], \tag{19}$$

$$\text{where } f(z, \phi, \mathcal{D}) = \log p(z) + \left(\sum_{e=1}^{E}\sum_{i=1}^{n_e} \log \hat{p}(y_{ei} \mid x_{ei}^z) - \hat{p}_e(y_{ei} \mid x_{ei}^z)\right) - \log q_\phi(z). \tag{20}$$

We summarize the variational inference algorithm, *PI-var*, in Algorithm 2. PI-var optimizes the variational objective $l(\mathcal{D}, \phi)$ with stochastic gradient ascent. It requires a stochastic gradient estimator $g(\cdot)$ that produces a random, unbiased estimate of the gradient $\nabla_\phi l(\mathcal{D}, \phi)$. And we use the average of multiple Monte Carlo samples to reduce the variance of gradient estimation. After T iterations of gradient ascent updates, the algorithm returns $q_{\phi_T}(z)$ with the optimized parameter ϕ_T as the approximate posterior distribution of z.

Gradient estimation. We discuss the choice of the stochastic gradient estimator. In this work, we use the U2G gradient estimator summarized in Algorithm 3 in the appendix, which provides an unbiased estimate of the true gradient (Yin et al., 2020).

In general, estimating the gradient of an expectation over discrete variables, including the form of $l(\mathcal{D}, z)$ in eq. (19), is an active research area. In addition to the U2G estimator explored in this work, there are other methods that can provide unbiased or low-variance gradient estimation. See a comprehensive review in Mohamed et al. (2020).

Algorithm 1: PI-exact: exact posterior inference for inferring invariant features

Input: Dataset $\mathcal{D} = \{(x_{ei}, y_{ei})\}$
Output: Posterior distribution $\hat{p}(z \mid \mathcal{D})$
for z *in support of* $p(z)$ **do**
  Estimate per-environment conditionals $\{\hat{p}_e(y \mid x^z)\}_{e=1}^E$ and pooled conditional $\hat{p}(y \mid x^z)$ by eqs. (14) and (15) using data $\mathcal{D}$
  Compute the likelihood ratio $\hat{\Lambda}(z) \leftarrow \sum_{e=1}^E \sum_{i=1}^{n_e} \frac{\hat{p}(y_{ei}^z | x_{ei}^z)}{\hat{p}_e(y_{ei}^z | x_{ei}^e)}$
Compute the posterior $\hat{p}(z \mid \mathcal{D}) \propto p(z)\hat{\Lambda}(z)$

Algorithm 2: PI-var: variational inference for inferring invariant features

Input: Dataset $\mathcal{D}$, stochastic gradient estimator g, objective function f in eq. (20), number of optimization iterations T, number of samples for gradient estimation M, learning rate scheduler $r(\cdot)$
Output: Variational posterior $q_{\phi_T}(z)$
Initialize variational parameter ϕ_0
for $t = 1$ **to** T **do**
  Compute M gradient estimates: $\hat{g}_m \leftarrow g(f, \mathcal{D}, \phi_{t-1})$ for $m = 1, \cdots, M$
  Update variational parameters: $\phi_t \leftarrow \phi_{t-1} + r(t) * \frac{1}{M} \sum_{m=1}^M \hat{g}_m$

Complexity. The complexity of PI-var is $O(T \cdot M \cdot c(\mathcal{D}, p))$ for T optimization iterations and M stochastic gradient samples, in contrast to $(2^p \cdot c(\mathcal{D}, p))$ of PI-exact under an uninformative prior. Similarly, with an informative prior that limits the number of features within $p_{max} \ll p$, VI's complexity can be reduced to $O(T \cdot M \cdot c(\mathcal{D}, p_{max}))$.

C Additional Theory and Proofs

In this section we provide additional theoretical results and proofs that are in complement to Section 3.

Without loss of generality, we assume the per-environment sample size is the same, denoted by $n_e = n$. We explicitly write the observed data as $\mathcal{D}_{n,E} = \{\{x_{ei}, y_{ei}\}_{i=1}^n\}_{e \in \mathcal{E}}$. The notation $\xrightarrow{P}$ indicates convergence in probability. Unless otherwise specified, we consider convergence over the probability space defined by: (i) E random environments sampled from $p(e)$, and (ii) n random data points sampled from the corresponding $p_e(x, y)$ for $e = 1, \cdots, E$, where $p(e)$ is a uniform distribution over all environments $\mathcal{E}$. We adopt the notation $\forall_p e$ to mean "for $p(e)$-almost every e", $\forall_p z$ to mean "for $p(z)$-almost every z, and supp$(p(z))$ for the support of $p(z)$.

C.1 Preliminary lemmas

We first introduce two lemmas that are helpful for proofs of the main theorems.

Lemma 1. *Let $\{X_n\}$ be a sequence of random variables where each $X_n \in \mathbb{R}^p$ and p is a fixed number. Assume $X_n \xrightarrow{P} \mu$ for some constant $\mu \in \mathbb{R}^p$, and define $\mathcal{I}^* := \{i : \mu_i = \mu_{\max}\}$ where $\mu_{\max}$ is value of the largest component(s) of μ and therefore $\mathcal{I}^*$ is the set of corresponding indices. Let $i_n^* := \arg\max_i X_n^{(i)}$ denote the index of the maximal component of X_n. Then we have*

$$P(i_n^* \in \mathcal{I}^*) \to 1, \text{ as } n \to \infty. \tag{21}$$

As a special case, when $\mathcal{I}^* = \{i^* : \mu_{i^*} > \mu_j \forall j \neq i^*\}$ is a singleton, we have

$$i_n^* \xrightarrow{P} i^*, \text{ as } n \to \infty. \tag{22}$$

Proof By continuous mapping theorem, $\forall j \notin \mathcal{I}^*$, as $n \to \infty$,

$$\max_{i \in \mathcal{I}^*} X_n^{(i)} - X_n^{(j)} \xrightarrow{P} \mu_{\max} - \mu^{(j)} > 0. \tag{23}$$

Hence $P(\max_{i\in\mathcal{I}^*} X_n^{(i)} - X_n^{(j)} < 0) \to 0$ as $n \to \infty$.

Consequently, as $n \to \infty$,

$$P(i_n^* \notin \mathcal{I}^*) = P(\exists j \notin \mathcal{I}^*, \max_{i\in\mathcal{I}^*} X_n^{(i)} < X_n^{(j)}) \tag{24}$$

$$\leq \sum_{j\notin\mathcal{I}^*} P(\max_{i\in\mathcal{I}^*} X_n^{(i)} - X_n^{(j)} < 0) \to 0, \tag{25}$$

and therefore

$$P(i_n^* \in \mathcal{I}^*) \to 1. \tag{26}$$

■

Lemma 2. *Let $\{X_n\}$ be a sequence of random variables such that its sample average $\frac{S_n}{n}$ converges to μ in probability, where $\mu < 0$, and $S_n := \sum_{i=1}^n X_i$. Then we have*

$$\exp\{S_n\} \xrightarrow{P} 0. \tag{27}$$

Proof $\forall\epsilon > 0, \forall\delta > 0$, it suffices to find an $N = N(\epsilon, \delta)$ such that for $n > N$,

$$P(\exp\{S_n\} < \epsilon) = P(\frac{S_n}{n} < \frac{1}{n}\log\epsilon) > 1 - \delta. \tag{28}$$

Since $\frac{S_n}{n} \xrightarrow{P} \mu$, for $\epsilon' = -\frac{\mu}{2}$ and δ given above, $\exists N' = N'(\epsilon', \delta)$ such that for $n > N'$,

$$P(\frac{S_n}{n} < \mu + \epsilon') > 1 - \delta. \tag{29}$$

We choose $N = \max(N', \frac{1}{\mu+\epsilon'}\log\epsilon)$. For $n > N$, we have

$$\frac{1}{n}\log\epsilon > \mu + \epsilon', \tag{30}$$

and therefore

$$P(\frac{S_n}{n} < \frac{1}{n}\log\epsilon) > P(\frac{S_n}{n} < \mu + \epsilon') > 1 - \delta. \tag{31}$$

■

C.2 Assumptions and proofs for theorems in Section 3

Assumption 2 (Prior positivity and uniqueness of invariant z^*)**.** *The existence of z^* in Assumption 1 is unique and the prior value $p(z^*) > 0$.*

C.2.1 Proof of Theorem 1 on posterior consistency

Proof $\forall z$, the posterior density at z can be written as follows

$$p(z \mid \mathcal{D}_{n,E}) = \frac{p(z)\exp\{-S_{n,E}(z)\}}{\sum_{z'} p(z')\exp\{-S_{n,E}(z')\}} \tag{32}$$

where $S_{n,E}(z)$ is the sum of log conditional likelihood ratio given z,

$$S_{n,E}(z) := \sum_{i=1}^{n}\sum_{e=1}^{E} l_{ei}(z), \tag{33}$$

with $l_{ei}(z)$ the per-data point log likelihood ratio given z.

$$l_{ei}(z) := \log p_e(y_{ei} \mid x_{ei}^z) - \log p(y_{ei} \mid x_{ei}^z). \tag{34}$$

We divide the analysis for two cases, infinite n (fixing E) and infinite E (fixing n).

Case 1: infinite n. We first consider the case of holding E fixed and letting $n \to \infty$. We interpret $S_{n,E}(z)$ as the sum of i.i.d. variables of the form $\sum_{e=1}^{E} l_{e,i}(z)$, indexed by i for $i = 1, \cdots, n$, each with mean $E\mu(z)$ where

$$\begin{aligned} \mu(z) &:= \mathbb{E}_{p(e)p_e(x^z)} \mathrm{KL}\left[p_e(y \mid x^z) \,\|\, p(y \mid x^z)\right] \\ &= \mathbb{E}_{p(e)} \mathbb{E}_{p_e(x,y)}\left[\log p_e(y \mid x^z) - \log p(y \mid x^z)\right], \end{aligned} \tag{35}$$

with $\mu(z) \geq 0$ and $\mu(z^*) = 0$. By Law of Large Numbers (LLN), as $n \to \infty$, we have

$$\frac{1}{n} S_{n,E}(z) \xrightarrow{P} E\mu(z). \tag{36}$$

Applying Lemma 1 to eq. (36) and by Assumption 2, we have

$$\arg\max_z -\frac{1}{n} S_{n,E}(z) \xrightarrow{P} \arg\max_z -E\mu(z) = z^*. \tag{37}$$

Consequently, the posterior mode is consistent:

$$z_{n,E} := \arg\max_z p(z \mid \mathcal{D}_{n,E}) \tag{38}$$

$$= \arg\max_z \log p(z) - S_{n,E}(z) \tag{39}$$

$$= \arg\max_z \frac{1}{n} \log p(z) - \frac{1}{n} \sum_{i=1}^{n} S_{n,E}(z) \tag{40}$$

$$\xrightarrow{P} z^*, \tag{41}$$

where we note that $\log p(z)$ in eq. (40) is bounded (since the support $\{0,1\}^p$ is finite-dimensional) and so $\frac{1}{n} \log p(z) \xrightarrow{P} 0, \forall z$.

Next, we prove the posterior consistency at z^*. By Assumption 2, $p(z^*) > 0$. Therefore, dividing both the numerator and denominator on the RHS of Equation (32) by $p(z^*) \exp\{-S_{n,E}(z^*)\}$, we obtain the posterior density at z^* as follows

$$p(z^* \mid \mathcal{D}_{n,E}) = \frac{1}{1 + \sum_{z \neq z^*} p(z)/p(z^*) \cdot \exp\{S_{n,E}(z^*) - S_{n,E}(z)\}} \tag{42}$$

To show $p(z^* \mid \mathcal{D}_{n,E}\}) \xrightarrow{P} 1$ with respect to n, it suffices to show that $\forall_p z \neq z^*$,

$$\exp\{S_{n,E}(z^*) - S_{n,E}(z)\} \xrightarrow{P} 0 \tag{43}$$

as $n \to \infty$.

By eqs. (36) and (35) and continuous mapping theorem, $\forall_p z \neq z^*$, as $n \to \infty$

$$\frac{1}{n}\left(S_{n,E}(z^*) - S_{n,E}(z)\right) \xrightarrow{P} E(\mu(z^*) - \mu(z)) < 0. \tag{44}$$

By applying Lemma 2 to eq. (44), eq. (43) holds. Hence we conclude the proof for posterior consistency at z^* with respect to n.

Case 2: infinite E. The above proof extends similarly to the case when n is held fixed while $E \to \infty$. In this setting, we interpret $S_{n,E}$ as defined in eq. (33), as the sum of i.i.d. variables of the form $\sum_{i=1}^{n} l_{e,i}(z)$, indexed by e over $e = 1, \cdots, E$, each with mean as $n\mu(z)$.

■

C.2.2 Proof of Theorem 2 on posterior contraction rate

Next, we characterize the posterior contraction rate. We consider the total variational distance $\mathrm{TV}(\cdot, \cdot)$ between the posterior and the Dirac measure centered at the true z^*,

$$\mathrm{TV}\Big(p(z \mid \mathcal{D}_{n,E}), \delta_{z^*}(z)\Big) = \frac{1}{2}\left(\sum_{z \neq z^*} |p(z \mid \mathcal{D}_{n,E}) - 0| + |p(z^* | \mathcal{D}_{n,E}) - 1|\right) = 1 - p(z^* | \mathcal{D}_{n,E}). \tag{45}$$

We make an additional assumption:

Assumption 3 (Finite variance of log likelihood ratio)**.** $\forall_p z$*, the log likelihood ratio* $\log p_e(y \mid x^z) - \log p(y \mid x^z)$ *has finite variance, that is,*

$$v(z) := \mathit{Var}_{p(e)p_e(y,x^z)}\left([\log p_e(y \mid x^z) - \log p(y \mid x^z)]\right) < \infty. \tag{46}$$

Incorporating the above assumption, we derive the following posterior contraction rate.

Proof We retain the notation $S_{n,E}(z)$ and $l_{ei}(z)$ as defined in eqs. (33) and (34), respectively.

Using the expression in Equation (42), we lower bound the posterior at z^*,

$$p(z^* \mid \mathcal{D}_{n,E}) \geq \frac{1}{1 + R\sum_{z \neq z^*} \exp\{S_{n,E}(z^*) - S_{n,E}(z)\}}. \tag{47}$$

where we recall that $R := \max_{z \neq z^*} p(z)/p(z^*)$.

We structure the remaining proof to show that

(i) the term $[S_{n,E}(z^*) - S_{n,E}(z)]/(nE)$ in the denominator of eq. (47) concentrates at its expected value, $-\mu(z)$, for large n or E;

(ii) consequently, the rate at which $p(z^* \mid \mathcal{D}_{n,E})$ converges to 1 depends on $\mu_{\min}$, the smallest $\mu(z)$ among all z satisfying $z \neq z^*$ and $p(z) > 0$.

We first show the concentration behavior of $S_{n,E}(z^*) - S_{n,E}(z)$. For any $z \neq z^*$ and any $k > 0$, by Chebyshev's inequality we have

$$P\bigg(\mid S_{n,E}(z^*) - S_{n,E}(z) - \tilde{\mu}_{n,E}(z)| \geq k \bigg) \leq \frac{\tilde{\sigma}^2_{n,E}(z)}{k^2} \tag{48}$$

where $\tilde{\mu}_{n,E}(z)$ and $\tilde{\sigma}^2_{n,E}(z)$ are the mean and variance of $S_{n,E}(z^*) - S_{n,E}(z)$,

$$\tilde{\mu}_{n,E}(z) := \mathbb{E}\left[\sum_{i=1}^{n}\sum_{e=1}^{E}\left(l_{e,i}(z^*) - l_{e,i}(z)\right)\Big| z\right] = nE(\mu(z^*) - \mu(z)) = -nE\mu(z) \tag{49}$$

$$\tilde{\sigma}^2_{n,E}(z) := \mathrm{Var}\left[\sum_{i=1}^{n}\sum_{e=1}^{E}\left(l_{e,i}(z^*) - l_{e,i}(z)\right)\Big| z\right] = nEv(z). \tag{50}$$

For any k, taking $k = (1-\kappa)|\tilde{\mu}_{e,E}(z)| = (1-\kappa)nE\mu(z)$, with a fixed value $\kappa \in (0,1)$, and substituting the expressions for the mean and variance from eqs. (49) and (50) into eq. (48), we obtain

$$P\bigg(\mid S_{n,E}(z^*) - S_{n,E}(z) + nE\mu(z)| \geq (1-\kappa)nE\mu(z) \bigg) \leq \frac{v(z)}{nE(1-\kappa)^2\mu(z)^2}. \tag{51}$$

By Assumption 2 $\mu(z) > 0$, and by Assumption 3 $v(z)$ is finite for any $z \neq z^*$. Therefore, the RHS of eq. (51) is well-defined.

Let $\mathcal{A}_{n,E}$ be the event that $\exists z \neq z^*$ such that $|S_{n,E}(z^*) - S_{n,E}(z) + nE\mu(z)| \geq (1-\kappa)nE\mu(z)$, and $\mathcal{A}^c_{n,E}$ denotes the complement of $\mathcal{A}_{n,E}$.

By definition of $\mathcal{A}_{n,E}$ and the subadditivity of probability measure, we have

$$\begin{aligned} P(\mathcal{A}_{n,E}) &\leq \sum_{z \neq z^*} P\bigg(|S_{n,E}(z^*) - S_{n,E}(z) + nE\mu(z)| \geq (1-\kappa)nE\mu(z)\bigg) \\ &\leq \sum_{z \neq z^*} \frac{v(z)}{nE(1-\kappa)^2\mu(z)^2} \to 0 \end{aligned} \tag{52}$$

as either $n \to \infty$ or $E \to \infty$.

Next, we define the rate $\epsilon_{n,E}$ as follows

$$\begin{aligned}\epsilon_{n,E} &:= 1 - \frac{1}{1 + R\sum_{z\neq z^*}\exp\{-\kappa n E\mu(z)\}} \\ &= \frac{R\sum_{z\neq z^*}\exp\{-\kappa n E\mu(z)\}}{1 + R\sum_{z\neq z^*}\exp\{-\kappa n E\mu(z)\}} \\ &= \frac{1}{1 + \Big(R\sum_{z\neq z^*}\exp\{-\kappa n E\mu(z)\}\Big)^{-1}} \\ &\leq \frac{1}{1 + \Big(R\cdot(|\text{supp}\,(p(z))\,| - 1)\cdot\exp\{-\kappa n E\mu_{\min}\}\Big)^{-1}}\end{aligned} \tag{53}$$

where we recall $\mu_{\min} := \min_{z\neq z^*, z\in\text{supp}(p(z))}\mu(z)$.

The rate $\epsilon_{n,E}$ can be asymptotically upper bounded as either $n\to\infty$ or $E\to\infty$:

$$\epsilon_{n,E} = O\Big(R\cdot(|\text{supp}\,(p(z))\,| - 1)\cdot\exp\{-nE\kappa\mu_{\min}\}\Big). \tag{54}$$

Using the inequality that lower bounds $p(z^*\mid\mathcal{D}_{n,E})$ from eq. (47), we have

$$P\left(p\Big(z^*\mid\mathcal{D}_{n,E}\Big) < 1-\epsilon_{n,E}\right) \leq P\left(\frac{1}{1 + R\sum_{z\neq z^*}\exp\{S_{n,E}(z^*) - S_{n,E}(z)\}} < 1-\epsilon_{n,E}\right) \tag{55}$$

$$= P\left(\frac{1}{1 + R\sum_{z\neq z^*}\exp\{S_{n,E}(z^*) - S_{n,E}(z)\}} < 1-\epsilon_{n,E}\mid\mathcal{A}^c_{n,E}\right)P(\mathcal{A}^c_{n,E}) \tag{56}$$

$$+ P\left(\frac{1}{1 + R\sum_{z\neq z^*}\exp\{S_{n,E}(z^*) - S_{n,E}(z)\}} < 1-\epsilon_{n,E}\mid\mathcal{A}_{n,E}\right)P(\mathcal{A}_{n,E}) \tag{57}$$

where eqs. (56) and (57) follow from the law of total probability.

The event $\mathcal{A}^c_{n,E}$ implies that $\forall z\neq z^*$,

$$|S_{n,E}(z^*) - S_{n,E}(z) + nE\mu(z)| < (1-\kappa)nE\mu(z) \tag{58}$$

$$\Rightarrow S_{n,E}(z^*) - S_{n,E}(z) < -\kappa n E\mu(z). \tag{59}$$

Therefore, given $\mathcal{A}^c_{n,E}$,

$$\frac{1}{1 + R\sum_{z\neq z^*}\exp\{S_{n,E}(z^*) - S_{n,E}(z)\}} > \frac{1}{1 + R\sum_{z\neq z^*}\exp\{-\kappa n E\mu(z)\}} \tag{60}$$

where the RHS equals $1-\epsilon_{n,E}$ by the definition of $\epsilon_{n,E}$ in eq. (53).

Hence

$$P\left(\frac{1}{1 + R\sum_{z\neq z^*}\exp\{S_{n,E}(z^*) - S_{n,E}(z)\}} < 1-\epsilon_{n,E}\mid\mathcal{A}^c_{n,E}\right) = 0. \tag{61}$$

Substituting eq. (61) into eq. (56) and recall the convergence of $P(\mathcal{A}_{n,E})$ from eq. (52), we obtain

$$P\left(p\Big(z^*\mid\mathcal{D}_{n,E}\Big) < 1-\epsilon_{n,E}\right) \leq 0\cdot P(\mathcal{A}^c_{n,E}) + 1\cdot P(\mathcal{A}_{n,E}) = P(\mathcal{A}_{n,E}) \to 0 \tag{62}$$

as $n\to\infty$ or $E\to\infty$.

Finally, we recall that the total variational distance $\text{TV}\Big(p(z\mid\mathcal{D}_{n,E}), \delta_{z^*}(z)\Big) = 1 - p(z^*|\mathcal{D}_{n,E})$ from eq. (45), which reduces the expression in eq. (62) to the desired result:

$$P\left(\text{TV}\Big(p(z\mid\mathcal{D}_{n,E}), \delta_{z^*}(z)\Big) > \epsilon_{n,E}\right) \to 0, \qquad \text{as } n\to\infty \text{ or } E\to\infty. \tag{63}$$

■

C.3 RESULTS FOR THE ESTIMATED POSTERIOR

Next we extend the analysis to the practical scenario where the true distributions $\{p_e(x, y)\}$ are unknown and we estimate the relevant conditionals from observed data. Specifically, we derive asymptotic results for the estimated $\hat{p}(z \mid \mathcal{D}_{n,E})$ as defined in Equation (16).

We first define the *population* version of the estimated pooled and per-environment conditionals:

$$\bar{p}(y \mid x^z) := \arg\max_{\tilde{p} \in \mathcal{P}_{y|x^z}} \mathbb{E}_{p(e)p_e(x,y)} \left[\log \tilde{p}(y \mid x^z)\right], \quad \forall z \tag{64}$$

$$\bar{p}_e(y \mid x^z) := \arg\max_{\tilde{p} \in \mathcal{P}_{y|x^z}} \mathbb{E}_{p_e(x,y)} \left[\log \tilde{p}(y \mid x^z)\right], \forall z \forall e. \tag{65}$$

These population conditional models do not necessarily coincide with the true conditionals $p(y \mid x^z)$ and $p_e(y \mid x^z)$, unless the model class $\mathcal{P}_{y|x^z}$ is correctly specified and the maximum likelihood estimation is consistent.

Correspondingly, we define

$$\bar{\mu}(z) := \mathbb{E}_{p(e)} \mathbb{E}_{p_e(x,y)} \left[\log \bar{p}_e(y \mid x^z) - \log \bar{p}(y \mid x^z)\right]. \tag{66}$$

which measures the discrepancy between the per-environment population conditional and the pooled population conditional.

We divide the remaining analysis into two parts: the limiting behavior with respect to n and the limiting behavior with respect to E.

C.3.1 LIMITING BEHAVIOR WITH RESPECT TO n.

We make the following assumptions.

Assumption 4 (Model well-specification for the invariant z^*)**.** *$p(y \mid x^{z^*}) \in \mathcal{P}_{y|x^{z^*}}$, which also implies $p_e(y \mid x^{z^*}) \in \mathcal{P}_{y|x^{z^*}}$ since $p_e(y \mid x^{z^*}) = p(y \mid x^{z^*}), \forall e$.*

Assumption 5 (Regularity condition on the modeling class)**.** *For any $z \neq z^*$ with $p(z) > 0$ and $\forall_p e$,*

$$\mathbb{E}_{p_e(x,y)} \left[\log \bar{p}_e(y \mid x^z)\right] > \mathbb{E}_{p_e(x,y)} \left[\log \bar{p}(y \mid x^z)\right]. \tag{67}$$

Assumption 6 (Likelihood consistency of pooled conditional with respect to n)**.** *$\forall_p z$, as $n \to \infty$,*

$$\frac{1}{n}\frac{1}{E} \sum_{i=1}^{n} \sum_{e=1}^{E} \log \hat{p}(y_{ei} \mid x_{ei}^z) \xrightarrow{P} \mathbb{E}_{p(e)p_e(x^z,y)} \left[\log \bar{p}(y \mid x^z)\right]. \tag{68}$$

Assumption 7 (Likelihood consistency of per-environment conditional with respect to n)**.** *$\forall e \in \mathcal{E}$ and $\forall_p z$, as $n \to \infty$,*

$$\frac{1}{n} \sum_{i=1}^{n} \log \hat{p}_e(y_{ei} \mid x_{ei}^z) \xrightarrow{P} \mathbb{E}_{p_e(x^z,y)} \left[\log \bar{p}_e(y \mid x^z)\right]. \tag{69}$$

Given these relaxed assumptions, we can extend **??** to the following one:

Theorem 3 (Posterior consistency with respect to n given estimated models)**.** *Given Assumptions 1 and 4 to 7 and that $p(z^*) > 0$, fixing E and as $n \to \infty$ we have*

1. *posterior mode consistency , i.e. $\hat{z}_{n,E} := \arg\max_z \hat{p}(z \mid \mathcal{D}_{n,E}) \xrightarrow{P} z^*$*
2. *posterior consistency at z^*, i.e. $\hat{p}(z^* \mid \mathcal{D}_{n,E}) \xrightarrow{P} 1$.*

Proof The finite-sample posterior can be written as

$$\hat{p}(z \mid \mathcal{D}_{n,E}) = \frac{p(z) \exp\{-\hat{S}_{n,E}(z)\}}{\sum_z p(z) \exp\{-\hat{S}_{n,E}(z)\}} \tag{70}$$

where $\hat{S}_{n,E}(z)$ is the sum of log likelihood ratios under estimated conditional models corresponding to z,

$$\hat{S}_{n,E}(z) := \sum_{i=1}^{n}\sum_{e=1}^{E} \log \hat{p}_e(y_{ei} \mid x_{ei}^z) - \log \hat{p}(y_{ei} \mid x_{ei}^z). \tag{71}$$

By the likelihood consistency in Assumptions 6 and 7 and continuous mapping theorem, fixing E and as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{S}_{n,E}(z) \xrightarrow{P} E\bar{\mu}(z), \tag{72}$$

where we recall $\bar{\mu}(z)$ as defined in eq. (66).

Similarly to the proof of Theorem 1, it suffices to show that z^* is the unique minimizer of $\bar{\mu}(z)$ with positive prior mass. This holds because, by Assumption 4, $\bar{\mu}(z^*) = 0$, and by Assumption 5, $\bar{\mu}(z) > 0$ for $z \neq z^*$.

■

Next, we establish the contraction rate result with respect to n. Similarly to Assumption 3, we impose a condition on the finite variance of the log ratios of population conditionals.

Assumption 8 (Finite variance of log likelihood ratio for population conditionals)**.** *$\forall_p z$, the log likelihood ratio $\log p_e(y \mid x^z) - \log \bar{p}(y \mid x^z)$ has finite variance, that is,*

$$\bar{v}(z) := \mathit{Var}_{p(e)p_e(y,x^z)}\left([\log \bar{p}_e(y \mid x^z) - \log \bar{p}(y \mid x^z)]\right) < \infty. \tag{73}$$

Taking into account this assumption, we derive the following contraction rate result.

Theorem 4 (Contraction rate of estimated posterior with respect to n)**.** *Given Assumptions 1 and 4 to 8 and that $p(z^*) > 0$, for any fixed E, there exists a sequence $\epsilon_{n,E} = O(R \cdot (|\mathit{supp}(p(z))| - 1) \cdot e^{-\kappa n E \bar{\mu}_{\min}})$ such that*

$$P\left(\mathit{TV}\Big(p(z \mid \mathcal{D}_{n,E}), \delta_{z^*}(z)\Big) > \epsilon_{n,E}\right) \to 0 \tag{74}$$

as $n \to \infty$, where R is the same as defined in eq. (9)*,*

$$\bar{\mu}_{\min} := \min_{z \neq z^*, z \in \mathit{supp}(p(z))} \bar{\mu}(z) = \min_{z \neq z^*, z \in \mathit{supp}(p(z))} \mathbb{E}_{p(e)p_e(x^z)} \mathit{KL}(\bar{p}_e(y \mid x^z) \| \bar{p}(y \mid x^z)), \tag{75}$$

and κ is a fixed value in $(0, 1)$.

Proof Similarly to the derivation of the lower bound of the true posterior in eq. (47), we have the following lower bound on the finite-sample posterior value at z^*:

$$\hat{p}(z^* \mid \mathcal{D}_{n,E}) \geq \frac{1}{1 + R\sum_{z \neq z^*} \exp\{\hat{S}_{n,E}(z^*) - \hat{S}_{n,E}(z)\}}. \tag{76}$$

where we recall the definition of $\hat{S}_{n,E}(z)$ from eq. (71).

We decompose $\hat{S}_{n,E}(z)$ as follows:

$$\hat{S}_{n,E}(z) = \bar{S}_{n,E}(z) + B_{n,E}(z) \tag{77}$$

where

$$\bar{S}_{n,E}(z) := \sum_{e=1}^{E}\sum_{i=1}^{n_e} \log \bar{p}_e(y_{ei} \mid x_{ei}^z) - \log \bar{p}(y_{ei} \mid x_{ei}^z), \tag{78}$$

$$B_{n,E}(z) := B_{n,E}^{(1)}(z) - B_{n,E}^{(2)}(z), \tag{79}$$

$$\text{with } B_{n,E}^{(1)}(z) := \sum_{e=1}^{E}\sum_{i=1}^{n} \left[\log \hat{p}_e(y_{ei} \mid x_{ei}^z) - \log \bar{p}_e(y_{ei} \mid x_{ei}^z)\right], \tag{80}$$

$$B_{n,E}^{(2)}(z) := \sum_{e=1}^{E}\sum_{i=1}^{n} \left[\log \hat{p}(y_{ei} \mid x_{ei}^z) - \log \bar{p}(y_{ei} \mid x_{ei}^z)\right]. \tag{81}$$

We define the rate

$$\epsilon_{n,E} := 1 - \frac{1}{1 + R\sum_{z \neq z'} \exp\{-\kappa n E \bar{\mu}(z)\}} \tag{82}$$

$$= O(R \cdot (|\text{supp}(p(z))| - 1) \cdot e^{-\kappa n E \bar{\mu}_{\min}}), \tag{83}$$

Following a similar analysis in the proof of Theorem 2, one can show that

$$P\left(\text{TV}\Big(\hat{p}(z \mid \mathcal{D}_{n,E}), \delta_{z^*}(z)\Big) > \epsilon_{n,E}\right) \leq P(\hat{\mathcal{A}}_{n,E}), \tag{84}$$

where $\kappa \in (0,1)$ is a fixed constant, and $\hat{\mathcal{A}}_{n,E}$ is the event that $\exists z \neq z^*$ such that $|\bar{S}_{n,E}(z^*) - \bar{S}_{n,E}(z) + nE\mu(z)| \geq \frac{1-\kappa}{2} nE\mu(z)$, or $|B_{n,E}(z^*) - B_{n,E}(z)| \geq \frac{1-\kappa}{2} nE\mu(z)$.

It suffices to show that $P(\hat{\mathcal{A}}_{n,E}) \to 0$ as $n \to \infty$. By definition of $\hat{\mathcal{A}}_{n,E}$ and subadditivity of probability measure, we have

$$\begin{aligned} P(\hat{\mathcal{A}}_{n,E}) \leq \sum_{z \neq z^*} P\left(|\bar{S}_{n,E}(z^*) - \bar{S}_{n,E}(z) + nE\bar{\mu}(z)| \geq \frac{1-\kappa}{2} nE\bar{\mu}(z)\right) \\ + \sum_{z \neq z^*} P\left(|B_{n,E}(z^*) - B_{n,E}(z)| \geq \frac{1-\kappa}{2} nE\bar{\mu}(z)\right). \end{aligned} \tag{85}$$

By Chebyshev's inequality, $\forall z \neq z^*$ as $n \to \infty$,

$$P\left(|S_{n,E}(z^*) - S_{n,E}(z) + nE\bar{\mu}(z)| \geq \frac{1-\kappa}{2} nE\bar{\mu}(z)\right) \leq \frac{\bar{v}(z)}{nE(1-\kappa)^2 \bar{\mu}(z)^2} \to 0, \tag{86}$$

where the RHS of the above inequality is valid because by Assumptions 4 and 5 $\bar{\mu}(z) > 0$ and by Assumption 8 $\bar{v}(z)$ is finite.

By Assumptions 6 and 7 and LLN, $\forall z \neq z^*$ as $n \to \infty$,

$$\frac{B_{n,E}(z)}{n} \xrightarrow{P} 0, \tag{87}$$

and consequently

$$P\left(|B_{n,E}(z^*) - B_{n,E}(z)| \geq nE\frac{1-\kappa}{2}|\bar{\mu}(z)|\right) \to 0. \tag{88}$$

Substituting the results from eqs. (86) and (88) into eq. (85), we have $P(\hat{\mathcal{A}}_{n,E}) \to 0$ as $n \to \infty$, which concludes the proof. ■

C.3.2 Limiting behavior with respect to E

We make the following assumptions.

Assumption 9 (Likelihood consistency of pooled conditional with respect to E)**.** *$\forall_p z$, as $E \to \infty$,*

$$\frac{1}{n}\frac{1}{E}\sum_{i=1}^{n}\sum_{e=1}^{E} \log \hat{p}(y_{ei} \mid x_{ei}^z) \xrightarrow{P} \mathbb{E}_{p(e)p_e(x^z,y)}\left[\log \bar{p}(y \mid x^z)\right]. \tag{89}$$

We note that we view each size-n dataset $\{x_{ei}, y_{ei}\}_{i=1}^n$ indexed by e as an i.i.d. draw from the distribution $\int p(e) \prod_{i=1}^n p_e(x,y) de$. When $E \to \infty$, we expect the collection of data $\{\{x_{ei}, y_{ei}\}_{i=1}^n\}_{e=1}^E$ provides a consistent estimate of the population conditional model $\bar{p}(y \mid x^z)$. Assumption 9 further requires that the average log likelihood value of the estimator models converges to the expected log likelihood value given the population model.

Assumption 10 (Prior positivity and uniqueness of invariant z^* under estimation bias)**.** $p(z^*) > 0$*, and* z^* *is the unique minimizer of* $\bar{\mu}_n(z)$ *among* $z \in$ *supp* $(p(z))$ *where*

$$\bar{\mu}_n(z) := \underbrace{\mathbb{E}_{p(e)p_e(x^z)} KL(\bar{p}_e(y \mid x^z) \| \bar{p}(y \mid x^z))}_{:=\textit{population model discrepancy } \bar{\mu}(z)} + \underbrace{\frac{1}{n} \mathbb{E}_{p(e)\prod_{i=1}^n p_e(x_i,y_i)} \sum_{i=1}^{n} [\log \hat{p}_e^z(y_i; x_i^z) - \log \bar{p}_e(y_i \mid x_i^z)]}_{:=\textit{estimation bias of per-environment conditionals } \bar{b}_n(z)}, \tag{90}$$

While Assumption 9 is a reasonable assumption, Assumption 10 is non-trivial. When the conditional model class is well-specified or regular, i.e. Assumptions 4 and 5 hold, z^* minimizes $\bar{\mu}(z)$. Furthermore, if the estimation bias $\bar{b}_n(z)$ is well-controlled for all z (which is the non-trivial part particularly when the per-environment sample size n is small), one can expect that z^* also minimizes $\bar{\mu}_n(z)$, thereby satisfying Assumption 10.

With the above assumptions, we extend **??** to the following result.

Theorem 5 (Posterior consistency with respect to E given estimated conditionals)**.** *Given Assumptions 1, 9 and 10, for any fixed* n*, as* $E \to \infty$*, we have*

1. *posterior mode consistency, i.e.* $\hat{z}_{n,E} \xrightarrow{P} z^*$
2. *posterior consistency at* z^**, i.e.* $\hat{p}(z^* \mid \mathcal{D}_{n,E}) \xrightarrow{P} 1$.

Proof The proof follows the same structure as that for the infinite-n case in Theorem 3. It suffices to show that

$$\arg\max_z -\frac{1}{E} \hat{S}_{n,E}(z) \xrightarrow{P} 0 \text{ as } E \to \infty,$$

after which the remainder of the proof is similar to the previous analysis, and is therefore omitted.

Recall $\bar{S}_{n,E}(z) = \sum_{e=1}^{E} \sum_{i=1}^{n} \log \bar{p}_e(y_{ei} \mid x_{ei}^z) - \log \bar{p}(y_{ei} \mid x_{ei}^z)$ defined in eq. (78). Given a fixed value of z, we view $\bar{S}_{n,E}(z)$ as a sum of i.i.d. variables of the form $\sum_{i=1}^{n} \log \bar{p}_e(y_{ei} \mid x_{ei}^z) - \log \bar{p}(y_{ei} \mid x_{ei}^z)$ indexed by e, each with mean $n\bar{\mu}(z)$ where $\bar{\mu}(z) := \mathbb{E}_{p(e)} \mathbb{E}_{p_e(x,y)} [\log \bar{p}_e(y \mid x^z) - \log \bar{p}(y \mid x^z)]$ is introduced in eq. (66). By LLN, as $E \to \infty$,

$$\frac{1}{E} \bar{S}_{n,E}(z) \xrightarrow{P} n\bar{\mu}(z). \tag{91}$$

Similarly, we have that

$$\frac{1}{E} B_{n,E}^{(1)}(z) = \frac{1}{E} \sum_{e=1}^{E} \sum_{i=1}^{n} [\log \hat{p}_e(y_{ei} \mid x_{ei}^z) - \log \bar{p}_e(y_{ei} \mid x_{ei}^z)] \xrightarrow{P} n\bar{b}_n(z), \tag{92}$$

where we recall $B_{n,E}^{(1)}(z)$ from eq. (80), and $\bar{b}_n(z)$ is the estimation bias of per-environment conditionals from eq. (90).

By Assumption 9, as $E \to \infty$, the estimation bias of the pooled conditional goes to 0, that is,

$$\frac{1}{E} B_{n,E}^{(2)}(z) = \frac{1}{E} \sum_{e=1}^{E} \sum_{i=1}^{n} [\log \hat{p}(y_{ei} \mid x_{ei}^z) - \log \bar{p}(y_{ei} \mid x_{ei}^z)] \xrightarrow{P} 0. \tag{93}$$

where $B_{n,E}^{(2)}$ is first introduced in eq. (81).

Combining eqs. (91) to (93), as $E \to \infty$,

$$\frac{1}{E} \hat{S}_{n,E}(z) = \frac{1}{E} \left(\bar{S}_{n,E}(z) + B_{n,E}^{(1)}(z) - B_{n,E}^{(2)}(z) \right) \xrightarrow{P} n(\bar{\mu}(z) + \bar{b}_n(z)) = n\bar{\mu}_n(z). \tag{94}$$

Applying Lemma 1 to eq. (94), as $E \to \infty$,

$$\arg\max_z -\frac{1}{E} \hat{S}_{n,E}(z) \xrightarrow{P} \arg\max_z -n\bar{\mu}_n(z) = z^*. \tag{95}$$

where the last equality follows from Assumption 10. ■

To establish the contraction rate result for the finite-sample posterior, we assume the following condition:

Assumption 11 (Finite variance of estimation bias in per-environment models)**.** *For any fixed n, the variance of the estimation bias of per-environment conditionals*

$$\bar{v}_n(z) := \mathit{Var}_{p(e)\prod_{i=1}^n p_e(y_i|x_i^z)}\left[\sum_{i=1}^n (\log \hat{p}_e(y_i \mid x_i^z) - \log \bar{p}_e(y_i \mid x_i^z))\right]$$

exists and is finite.

Theorem 6 (Contraction rate of estimated posterior with respect to E)**.** *Given Assumptions 1 and 8 to 11, for any fixed n, there exists a sequence $\epsilon_{n,E} = O(R \cdot (|supp(p(z))| - 1) \cdot e^{-\kappa n E \bar{\mu}_{\min}})$ such that*

$$P\left(\mathit{TV}\Big(p(z \mid \mathcal{D}_{n,E}), \delta_{z^*}(z)\Big)\epsilon_{n,E}\right) \to 0 \tag{96}$$

as $E \to \infty$ with a fixed E, where R and $\bar{\mu}_{\min}$ are defined in eqs. (9) *and* (75) *respectively, and κ is a fixed value in* $(0, 1)$.

Proof The finite-sample posterior at z^* is

$$\hat{p}(z^* \mid \mathcal{D}_{n,E}) \propto \frac{1}{1 + \sum_{z \neq z'} p(z)/p(z^*) \cdot \exp\{d_{S_{n,E}}(z^*, z) + d_{B^{(1)}_{n,E}}(z^*, z) - d_{B^{(2)}_{n,E}}(z^*, z)\}}. \tag{97}$$

where

$$\begin{aligned} d_{S_{n,E}}(z^*, z) &:= \bar{S}_{n,E}(z^*) - \bar{S}_{n,E}(z), \\ d_{B^{(1)}_{n,E}}(z^*, z) &:= B^{(1)}_{n,E}(z^*) - B^{(1)}_{n,E}(z), \\ d_{B^{(2)}_{n,E}}(z^*, z) &:= B^{(2)}_{n,E}(z^*) - B^{(2)}_{n,E}(z), \end{aligned}$$

and we recall $\bar{S}_{n,E}(z), B^1_{n,E}(z), B^2_{n,E}(z)$ from eqs. (78), (80) and (81).

We first analyze the concentration behavior of the term $d_{S_{n,E}}(z^*, z) + d_{B^{(1)}_{n,E}}(z^*, z)$ in eq. (97). By Chebyshev's inequality, $\forall z \neq z^*$,

$$\begin{aligned} &P\left(|d_{S_{n,E}}(z^*, z) + d_{B^{(1)}_{n,E}}(z^*, z) - nE d_{\bar{\mu}_n}(z^*, z)| \geq \frac{1-\kappa}{2} nE |d_{\mu_n}(z^*, z)|\right) \\ &\leq \frac{n\bar{v}(z) + \bar{v}_n(z)}{En^2 \cdot (1-\kappa)^2/4 \cdot d_{\bar{\mu}_n}(z^*, z)^2} \to 0, \text{ as } E \to \infty. \end{aligned} \tag{98}$$

where $d_{\bar{\mu}_n} := \bar{\mu}_n(z^*) - \bar{\mu}_n(z) < 0$ by Assumption 10, and $\bar{v}(z)$ and $\bar{v}_n(z)$ exist and are finite quantities by Assumptions 8 and 11.

We then characterize the concentration behavior of $d_{B^{(2)}_{n,E}}(z)$ in eq. (97). By Assumption 9, $\forall_p z$, as $E \to \infty$,

$$\frac{B^{(2)}_{n,E}(z)}{E} \xrightarrow{P} 0, \tag{99}$$

and consequently

$$\frac{1}{E} d_{B^{(2)}_{n,E}}(z^*, z) := \frac{1}{E}\left(B^{(2)}_{n,E}(z^*) - B^{(2)}_{n,E}(z)\right) \to 0. \tag{100}$$

Therefore,

$$P\left(|d_{B^{(2)}_{n,E}}(z^*, z)| \geq nE \frac{1-\kappa}{2} |d_{\bar{\mu}_n}(z^*, z)|\right) \to 0. \tag{101}$$

We define $\hat{\mathcal{A}}_{n,E}$ be the event that $\exists z \neq z^*$ such that $|d_{S_{n,E}}(z^*, z) + d_{B^{(1)}_{n,E}}(z^*, z) - nEd_{\bar{\mu}_n}(z^*, z)| \geq \frac{1-\kappa}{2} nE|d_{\mu_n}(z, z^*)|$, or $|B^{(2)}_{n,E}(z^*) - B^{(2)}_{n,E}(z)| \geq \frac{1-\kappa}{2} nE|d_{\bar{\mu}_n(z^*,z)}|$.

Following a similar analysis in the proof of Theorem 4, we can show that

$$P\left(\text{TV}\Big(p(z \mid \mathcal{D}_{n,E}), \delta_{z^*}(z)\Big) > \epsilon_{n,E}\right) = P\left(\hat{p}\Big(z^* \mid \mathcal{D}_{n,E}\Big) < 1 - \epsilon_{n,E}\right) \leq P(\hat{\mathcal{A}}_{n,E}) \quad (102)$$

It suffices to show that $P(\hat{\mathcal{A}}_{n,E}) \to 0$ as $E \to \infty$. This convergence follows directly from eqs. (98) and (101), thereby concluding the proof. ■

C.4 Discussion

We discuss adjustments to the posterior consistency results when certain conditions change.

We begin with the case where the true distributions $\{p_e(x, y)\}$ are available.

Violation of the invariance assumption. When Assumption 1 is violated, the posterior would instead concentrate on an "approximately invariant" feature selector – one that minimizes an invariance measure $\mu(z)$ among all z values:

$$\mu(z) := \mathbb{E}_{p(e)p_e(x^z)} \text{KL}\left[p_e(y \mid x^z) \,\|\, p(y \mid x^z)\right]. \quad (103)$$

The invariance measure $\mu(z)$ quantifies the expected KL divergence between the per-environment conditional and the pooled conditional given z, and therefore is non-negative for all z. A lower value of $\mu(z)$ indicates less discrepancy between per-environment conditionals and the pooled conditional, meaning that z is closer to resulting in an invariant model. In particular, when Assumption 1 holds, the invariant z^* satisfies $\mu(z^*) = 0$.

Violation of prior positivity for the invariant z^*. When the prior $p(z)$ assigns zero mass to z^* which invalidates part of Assumption 2, the posterior would concentrate on an approximately invariant feature selector within the prior support, i.e. $\arg\min_{z \in \text{supp}(p(z))} \mu(z)$.

Violation of Assumption 2 If there are multiple invariant feature selectors – violating the uniqueness condition in Assumption 2 – the posterior would concentrate on all invariant feature selectors, with density proportional to their prior values.

Next, we examine the impact of model misspecification on the results for estimated posterior given finite data.

Model misspecification. We let $\{\bar{p}_e(y \mid x^z)\}_{e,z}$ and $\{\bar{p}(y \mid x^z)\}_z$ be the best-fitting conditional models within the chosen model class. When the model class is misspecified, the posterior distribution under these fitted models will concentrate on the feature selector that minimizes the following invariance measure:

$$\bar{\mu}(z) := \mathbb{E}_{p(e)p_e(x,y)} \left[\log \bar{p}_e(y \mid x^z) - \log \bar{p}(y \mid x^z)\right]. \quad (104)$$

This measure $\bar{\mu}(z)$ quantifies the average discrepancy between the best-fitting per-environment and pooled conditionals over the multi-environment data.

The minimizer of $\bar{\mu}(z)$ is not necessarily the true invariant z^*, which minimizes $\mu(z)$ defined in eq. (103). Therefore, for posterior inference to remain reliable, it is crucial to ensure that our choice of conditional model class is reasonably specified:

(i) we can model the true invariant model accurately such that $\bar{\mu}(z^*) \approx 0$;

(ii) for remaining $z \neq z^*$, the conditional model class must be expressive enough to better fit the per-environment data than the pooled data such that $\bar{\mu}(z) > 0$.

When both conditions (i) and (ii) are satisfied, z^* can be expected to minimize $\bar{\mu}(z)$.

Finally, we comment on the limiting behavior with respect to n conditioned on a fixed set of environments.

Conditioning on the fixed set of environments. The above analysis takes into account of the randomness in sampling E environments from all potential environments $\mathcal{E}$. If we instead conditioned on a fixed set of observed environments, denoted as $\mathcal{E}_{obs}$ – the scenario encountered in practice – we can restrict the environments of interests $\mathcal{E}$ to $\mathcal{E}_{obs}$. When all relevevant assumptions hold for $\mathcal{E}_{obs}$, the consistency and contraction rate results would still hold.

Practically speaking, when $\mathcal{E}_{obs}$ includes only a few environments, it is possible that the uniqueness assumption in Assumption 2 does not hold, i.e. there exists multiple invariant feature selectors within $\mathcal{E}_{obs}$. Thus, the posterior distribution will concentrate on all these solutions when $n \to \infty$, as above discussion indicates. That said, adding new environments to $\mathcal{E}_{obs}$ can help in reducing the number of invariant feature selectors within the expanded $\mathcal{E}_{obs}$, resulting in fewer modes in the posterior. In this sense, the uncertainty will shrink with respect to the number of observed environments.

Theoretical justification. To justify the above claims, we prove the following theorem generalizes the consistency result in Theorem 1 to more general conditions.

Theorem 7 (Generalized posterior consistency)**.** *Let $\mathcal{Z}'$ be the set of minimizer(s) of $\mu(z)$ as defined in eq.* (35)*, and assume that $p(Z^*) > 0$. Then as $n \to \infty$ (fixing E) or $E \to \infty$ (fixing n), we have*

(i) $P(\hat{z}_{n,E} \in \mathcal{Z}') \to 1$,

(ii) $p(z^* \mid \mathcal{D}_{n,E}) \xrightarrow{P} \frac{p(z^*)}{p(Z^*)}$.

Proof (Sketch) We adapt the proof from Theorem 1. The proof to (i) follows similarly which we skip here.

To prove (ii), we note that

$$p(z^* \mid \mathcal{D}_{n,E}) = \frac{p(z^*)}{\sum_{z \in \mathcal{Z}'} p(z) I_1(z) + \sum_{z \notin \mathcal{Z}'} p(z) I_2(z)} \tag{105}$$

where

$$I_1(z) = \exp\{S_{n,E}(z^*) - S_{n,E}(z)\} \tag{106}$$
$$I_2(z) = \exp\{S_{n,E}(z^*) - S_{n,E}(z)\}. \tag{107}$$

Following the proof of Theorem 1, we know that $I_1(z) = 1$ for $z \in \mathcal{Z}'$, and $I_2(z) \xrightarrow{P} 0$ as n or E goes to ∞, for $z \notin \mathcal{Z}'$

Therefore, as $n \to \infty$ or $E \to \infty$,

$$p(z^* \mid \mathcal{D}_{n,E}) \xrightarrow{P} \frac{p(z^*)}{p(\mathcal{Z}')}. \tag{108}$$

■

The above results apply to the case when the true distributions $\{p_e(x, y)\}_e$ are known. Similar results apply when estimated models are used but additional conditions on model well-specification and estimation bias must be imposed. For example, when the models are not correctly specified, the above results need to be adjusted for the best-fitting model within the chosen model class. Moreover, in the case of finite per-environment sample n, the estimation bias must be adequately controlled to ensure posterior consistency.

D VARIATIONAL INFERENCE

D.1 GRADIENT ESTIMATORS FOR BINARY LATENT VARIABLES

We use the following algorithm to update the variational parameters of binary latent variables.

Algorithm 3: U2G gradient estimation (Yin et al., 2020)

Input: Objective function f, dataset $\mathcal{D}$, variational parameter ϕ
Output: A random unbiased estimate $\hat{g}$ of $\nabla_\phi \mathbb{E}_{q_\phi(z)}\left[f(z, \mathcal{D}, \phi)\right]$

Draw $u \sim \prod_{i=1}^{p} \text{Uniform}[0,1]$
$z_1 \leftarrow \mathbf{1}\left[u > 1 - \sigma(\phi)\right], z_2 \leftarrow \mathbf{1}\left[u < \sigma(\phi)\right]$
$\hat{g} \leftarrow \frac{1}{2}\sigma(|\phi|) \cdot \left[f(z_1, \mathcal{D}, \phi) - f(z_2, \mathcal{D}, \phi)\right] \cdot (z_1 - z_2)$

D.2 Optimization details

For all experiments, we use $M = 10$ Monte Carlo samples for the gradient estimation, the learning rate of 1.0 and the cyclic learning rate scheduler.

In simulation experiments, the variational parameter ϕ is initialized with $\sigma(0.4)$. In the gene data experiments, the variational parameter ϕ is initialized with $\sigma(0.05)$,

In practice, the initialization value of ϕ can be tuned using the training-set ELBO, and by checking whether the resulting optimization is stable. We found that making sure the expected number of invariant features under the variational posterior at the initialization is smaller than the maximum number of invariant features under the prior to be a good practice.

E Experiment details

E.1 Synthetic data study

E.1.1 Data generation details

We first describe the generic data generative process in details.

1. **The factorization of the joint distribution.** We first uniformly draw a permutation π over $[1, \cdots, p+1]$, which is used to re-arrange $(x^{(1:p)}, y)$ to obtain t: $t^{(i)} = x^{(\pi(i))}$ if $\pi(i) \neq p+1$ and $t^{(i)} = y$ otherwise. [1]

   The permutation π defines a factorization of any joint distribution $p(x^{(1:p)}, y)$:

   $$p(x^{(1:p)}, y) = \prod_{i=1}^{p+1} p(t^{(i)} \mid t^{(1:i-1)}), \tag{109}$$

   where we define $t^{(1:0)}$ to be an empty conditioning set.

   We let each conditional density $p(t^{(i)} \mid t^{(1:i-1)})$ to be a linear Gaussian model, i.e. $p(t^{(i)} \mid t^{(1:i-1)}) = \mathcal{N}(t^{(i)} \mid \beta t^{(1:i-1)} + \beta_0, \sigma^2)$ for some β, β_0, σ^2.

   We next specify the joint distribution $p_e(x, y)$ for different environment e that factorizes according to π. It suffices to specify the linear Gaussian model parameters for each conditional density.

2. **Observational environment** $e = 1$**.** For $i = 1, \cdots, p+1$:
   - sample the intercept parameter from $\mathcal{N}(0, 1)$,
   - sample the variance parameter from Uniform $\left([\sigma^2_{\min}, \sigma^2_{\max}]\right)$, where $\sigma^2_{\min}$ and $\sigma^2_{\max}$ are lower and upper bounds on the variance parameter,
   - sample each coefficient value independently:
     - if $\pi(i) = p+1$, i.e. $t^{(i)} = y$, first sample its absolute value from Uniform $([lb, ub])$ and then assign a random sign;
     - if $\pi(i) \neq p+1$ i.e. $t^{(i)} = x^{(\pi(i))}$, (i) with probability $actprob \in [0,1]$ first sample its absolute value from Uniform $([lb, ub])$ and then assign a random sign, (ii) otherwise set to 0.

[1] If there are pre-specified lower bound $p_{\min}$ and upper bound $p_{\max}$ on the number of the true invariant features $|z^*|$, we keep uniformly sampling π until we get a π such that $p_{\min} < \pi^{-1}(p+1) \leq p_{\max} + 1$.

3. **Interventional Environments** ($e = 2, \cdots, E$) Based on the observational environment $e = 1$, we first randomly draw a fraction $intfrac_e$ of *features* $x'_e \subset x$ to be intervened on. Then, for $i = 1, \cdots, p + 1$:
   - If $t^{(i)} \notin x'_e$, set $p_e(t^{(i)} \mid t^{(1:i-1)}) = p_1(t^{(i)} \mid t^{(1:i-1)})$
   - Otherwise, draw new parameters for $p_e(t^{(i)} \mid t^{(1:i-1)})$
     - sample intercept parameter by first sampling its absolute value from $\mathcal{N}(m_e, 1)$ and assigning to it a random sign, for some $m_e \in \mathbb{R}$ specific to the environment e.
     - sample variance parameter from Uniform $\left([\sigma^2_{e,\min}, \sigma^2_{e,\max}]\right)$, where $\sigma^2_{e,\min}, \sigma^2_{e,\max}$ are the lower and upper bounds of variance parameters specific to the environment e.
     - sample each coefficient value independently:
       - with probability $changeprob_e$, copy the corresponding coefficient from $e = 1$ case
       - with probability $1 - changeprob_e$, (i) with probability $actprob$ sample its absolute value from Uniform $([lb_e, ub_e])$, assign a random sign and (ii) otherwise set to 0.
4. **Draw data** Finally, we can draw n data points $\{x_{ei}, y_{ei}\}_{i=1}^n$ from each environment e independently, for $e = 1, \cdots, E$.

We next describe the choices of number of data points per environment n, the number of environments E, the number of features p, and other hyperparameters in different experiments.

E.1.2 Empirical verification of theory

We set $p = 3$, the maximum number of the true invariant features $p_{\max} = 3$, and the minimum number $p_{\min} = 1$. For $e = 1$, the lower bound of the absolute coefficient value $lb = 0.5$, upper bound $ub = 2$; $actprob = 1.0$; noise level lower bound $\sigma_{\min} = 0.1$ and upper bound $\sigma_{\max} = 0.2$. For $e = 2, \cdots, E$,

E.1.3 Method details

We run the following methods for comparison:

1. Oracle: Linear regression of y against x^{z^*}, the true invariant features, with parameters estimated using the least square method on pooled data. When $p+1 > nE$, the least square solution is infeasible, and we instead use Lasso, with 10-fold cross-validation to select the ℓ_1-regularization hyperparameter. We return the feature dimensions whose coefficients are significant at the $\alpha = 0.05$ level. If Lasso is used, we return the dimensions of the top 10 features by absolute coefficient values.
2. PI-exact: Our method with exact inference as described in Algorithm 1. Posterior mode is used for computing relevant metrics.
3. PI-var: Our method with variational inference as described in Algorithm 2. Posterior mode is used for computing relevant metrics. See optimization details in Appendix D.2.
4. Regression: Similar to Oracle, but linear regression is performed on all features rather than restricting to the true invariant features.
5. ICP: The Invariant Causal Prediction method proposed by Peters et al. (2016). We use the significance level $\alpha = 0.05$ for its multiple hypothesis testing procedure.
6. Hidden-ICP: an extension of ICP proposed by Rothenhäusler et al. (2019) that enables faster inference and permits the existence of hidden variable. We run the method at the significance level $\alpha = 0.05$.
7. EILLS: A linear regression model proposed by Fan et al. (2023) that introduces a regularization term to encourage invariant predictions among environments. Following the default choice in their original codebase, the regularization penalty is set to 36.

Notably, PI-exact, ICP and EILLS require enumeration of all candidate z values, which is only feasible when p is small. Additionally, parameter estimation for Hidden-ICP is intractable when

the number of features exceed the number of observations. Therefore, for $p = 450$, we apply a screening step to reduce to 10 features with highest coefficient magnitudes from Regression, and then apply these methods to the screened features. We call the modified versions *PI-exact-s*, *ICP-s*, *Hidden-ICP-s*, and *EILLS-s*.

PI-var does not require enumeration of z values, and is therefore directly applicable to high-dimensional settings. However, we find that using a uniform prior over the full space of $\{0,1\}^p$ can lead to unstable optimization process when p is large. Consequently, we set the prior $p(z)$ in PI-var to be uniform over $\{0,1\}^p$ when $p = 10$, and to be uniformly distributed over zs with $\|z\|_0 \leq 20$ when $p = 450$.

E.1.4 SIMULATION DETAILS

We describe the process to create a random set of joint distributions $\{p_e(x,y)\}_{e=1}^E$ given p and E. The reason to differentiate between various p is to control the scale of parameter values to prevent the sampled x, y values from exploding when p increases.

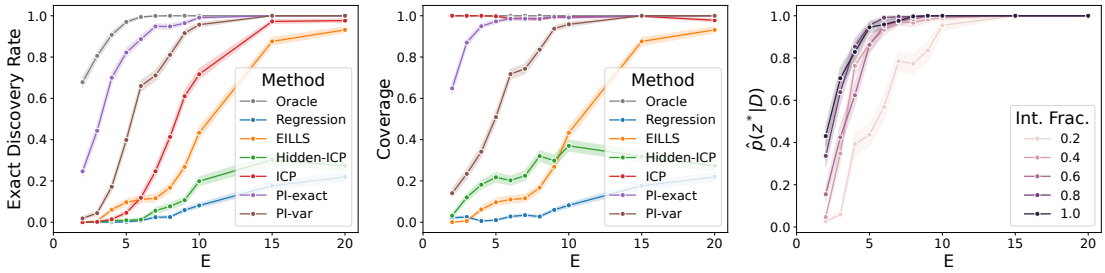
1. When $p = 10$,
   (a) $actprob \sim$ Uniform $(\{0.6, 0.7, 0.8, 0.9\})$
   (b) For $e = 1$: (i) coefficient lower bound $lb = 1$; (ii) coefficient upper bound $ub = 2.1$, (iii) noise level lower bound $\sigma_{\min} = 0.1$; (iv) noise level upper bound $\sigma^2_{\max} = 0.2$.
   (c) Separately for $e = 2, \cdots, E$: for the intervened variable (i) mean of the absolute value of the intercept $m_e \sim$ Uniform $([0,1])$; (ii) coefficient lower bound $lb_e = (m_e + 0.01) * 2$; (iii) coefficient upper bound $ub_e = (m_e + 0.5) * 2$; (iv) noise level lower bound $\sigma_{e,\min} \sim$ Uniform $([0.1, 0.2])$; (v) noise level upper bound $\sigma_{e,\max} = \sigma_{e,\min} +$ Uniform $([0.1, 0.5])$; (vi)the probability of changing the coefficient $changeprob_e \sim$ Uniform $([0.1, 0.3])$; and (vii) the fraction of features to be intervened on $intfrac_e \sim$ Uniform $([0.5, 1])$
2. When $p = 60$,
   (a) $actprob \sim$ Uniform $(\{0.1, 0.2, 0.3, 0.4\})$
   (b) For $e = 1$: same as $p = 10$ case
   (c) Separately for $e = 2, \cdots, E$: for the intervened variable (i) mean of the absolute value of the intercept $m_e \sim$ Uniform $([0,1])$; (ii) coefficient lower bound $lb_e = m_e + 0.01$; (iii) coefficient upper bound $ub_e = (m_e + 0.01) * 1.5$; other hyperparameters are chosen in the same way as $p = 10$ case
3. When $p = 450$,
   (a) $actprob \sim$ Uniform $(\{0.1, 0.15, 0.2, 0.25\})$
   (b) For $e = 1$: same as $p = 10$ case
   (c) Separately for $e = 2, \cdots, E$: for the intervened variable (i) mean of the absolute value of the intercept $m_e \sim$ Uniform $([0, 0.4])$; (ii) coefficient lower bound $lb_e = m_e + 0.01$; (iii) coefficient upper bound $ub_e = (m_e + 0.01) * 1.5$; other hyperparameters are chosen in the same way as $p = 10$ case

We set the maximum number of the true invariant features $p_{\max} = 5$ when $p = 10$, and $p_{\max} = 10$ otherwise. For all cases of p we set the minimum number of the true invariant features $p_{\min} = 1$.

E.1.5 ADDITIONAL RESULTS

Results for $p = 10$ case. Figure 4 displays the results for $p = 10$ case. In Figure 4a we observe that the exact discovery rate of all methods improves with more environments. In particular, all methods except Regression and Hidden-ICP have an exact discovery rate close to 1 for large $E = 20$, demonstrating their effectiveness with sufficiently many environments. The exact discovery rate of Oracle has the fastest convergence to 1, followed by PI-exact and then PI-var.

In the coverage results shown in Figure 4b, we observe an upward tread for all methods as E increases. Both Oracle and ICP consistently maintain a coverage close to 1 across different values of E. Notably, when E is small, ICP returns empty prediction for most of times, leading to a high



(a) Exact discovery rate (b) Coverage (c) Exact posterior at z^*

Figure 4: Synthetic study: comparison to other methods with $p = 10$ features. **Left panel:** As E increases, the exact discovery rate of all methods increases; PI-exact converges to 1 faster than other methods except for Oralce, followed by PI-var. **Middle panel:** The coverage of Oracle and ICP remains close to 1 across different E; the coverage of PI-exact, PI-var and ELLIS converges to 1 as E increases; other methods have relatively low converge. **Right panel:** The posterior value at z^* from PI-exact increases with larger E and higher intervention fraction, demonstrating posterior consistency. All results are averaged over cases with $n = 50, 200, 500$, each comprising 400 random simulations. Error bars indicate 95% confidence intervals.

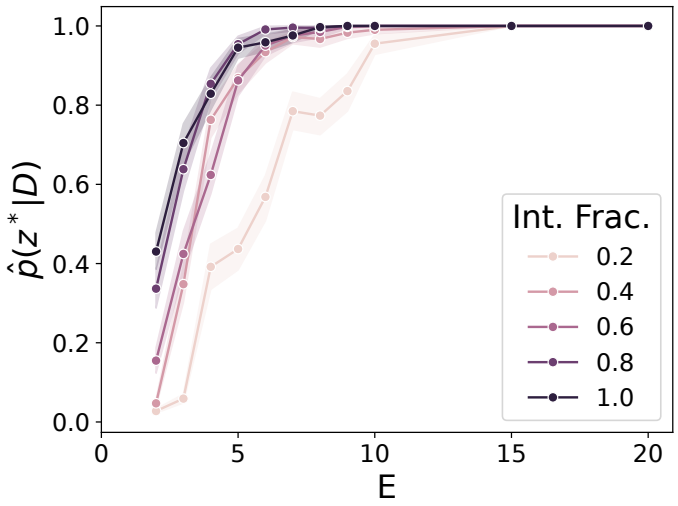


Figure 5: Synthetic study with $p = 10$ features. Variational posterior at z^* v.s. number of environments E under different fraction of intervention features. Similar to Figure 4c, we observe that the variational posterior also concentrates at z^* with increasing number of environments, and the convergence is faster when the fraction of intervention features is higher. However, the convergence of the variational posterior is slower than that of the exact posterior as in Figure 4c.

coverage but low exact discovery rate. The coverage of our methods, both PI-exact and PI-var, converges to a near perfect coverage when $E = 20$.

Figure 4c displays the posterior value at the true z^* given by PI-exact. The posterior value converges to 1 with more environments. Additionally, the convergence is faster with higher level of intervention fraction. This result is consistent with the theoretical contraction rates. We include the variational posterior value at z^* from PI-var in Figure 5 in the appendix, where we observe a similar trend as in the exact inference, but with a slower convergence rate compared to the exact posterior.

E.2 Gene perturbation study

We include the full experiment setup and additional results for the gene experiment in **??**.

Following prior work (Peters et al., 2016), we treat the target gene as outcome y and the remaining $p = 6,169$ genes as features x. Our goal is to find a subset of feature genes $x^{z^*} \subset x$ that invariantly predict y across environments. If the value of an invariant feature $x' \in x^{z^*}$ is perturbed, the invariance of the predictive model implies that the value of y is expected to change in the new environment created by the perturbation, which allows us to predict a *gene perturbation effect*. For non-invariant feature $x' \notin x^{z^*}$, we do not make such prediction. Peters et al. (2016) also provides a justificiation through a causal perspective, where the invariant feature genes are viewed as the direct causes of the target gene.

We set up the training and validation procedure largely following Peters et al. (2016). The interventional samples from $e = 2$ are randomly divided into 5 blocks at the beginning. We then sweep through 6,170 genes, treating each as the target gene, with the remaining genes as features. For each target gene, we construct a training set that includes all observational samples from $e = 1$ and 4 out of 5 blocks of interventional samples from $e = 2$, reserving the remaining block for validation. Notably, if the training set contains a sample corresponding to the perturbation of the target gene, we exclude it as we want to make sure the distribution of the target gene is not affected. This process is repeated five times, with each block of interventional samples held out once, ensuring that each gene perturbation result is used for validation at least once. When a method infers some set of invariant features, we check through the validation set that whether perturbing these features leads to a significant change in the expression level of the target gene.

We run the following methods:

1. Marginal: we pool training data over the two environments and retain all features that have a correlation with the outcome y at significance level α/p with $\alpha = 0.01$.
2. Lasso: we run Lasso on the pooled training data where the regularization penalty is selected from $\{0.1, 0.05, 0.01, 0.005, 0.001\}$ with 5-fold cross-validation and retain up to 10 features with the highest non-zero absolute coefficient values.
3. EILLS-s: we run EILLS on the screened features with an invariance regularization strength of 36.
4. Hidden-ICP-s: We run Hidden-ICP on the screened covaraites at significance level α, where we vary $\alpha \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$;
5. ICP-s: we run ICP on the screened features at significance level α/n_{int} where we vary $\alpha \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, and the factor $n_{int} = 1,479$ is used for Bonferroni correction, as suggested by Peters et al. (2016).
6. PI-exact-s: we run PI with exact inference on the screened features, and retain the features whose marginal posterior probability is above a threshold t where we vary $t \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.
7. PI-var: we run PI with variational inference on the full set of features with a sparse prior uniform over $\{z \in \{0,1\}^p : \|z\|_0 \leq 200\}$, and rxetain the features whose marginal posterior probability is above a threshold t where we vary $t \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. See the optimization details in Appendix D.2.

For the same reason as in the synthetic study, EILLS-s, ICP-s, PI-exact-s and Hidden-ICP-s all use a screening step. The screened features are initially selected using the features returned by Lasso. If Lasso does not return any features, namely, all the Lasso coefficients are 0, we instead select up to the top 10 features with largest absolute correlation values from Marginal. For all methods, the linear Gaussian family is used for specifying the prediction models.

As PI-var and ICP-s achieve the highest precision among the methods, we make a more detailed comparison between them.

Comparison between PI-var and ICP-s. We examine the consistency of predictions made by PI-var (with marginal probability threshold $t = 0.5$) and ICP-s (with a significance level $\alpha = 0.01$). For any fixed target gene, we take the intersection of predicted invariance features across 3 random seeds for each method. We find that around 72% (23 out of 32) of the predictions made by ICP-s are also identified by PI-var, suggesting a certain level of consistency. Moreover, PI-var is less conservative than ICP-s, making a total of 371 true predictions – about 11.6 times that of ICP-s.

Target gene	Inferred invariant feature gene(s)		
	Meinshausen et al. (2016)	ICP-s ($\alpha = 0.01$)	PI-var ($t = 0.5$)
YMR103C	YMR104C	YMR104C	YMR104C, YHR209W*
YMR321C	YPL273W	YPL273W	YPL273W
YCL042W	YCL040W	YCL040W	YCL040W
YLL020C	YLL019C	YLL019C	YLL019C
YPL240C	YMR186W	YMR186W	YJL077C, YMR186W
YBR126C	YDR074W		YGR008C*, YKL035W*, YDR074W
YMR173W-A	YMR173W	YMR173W	YMR173W
YGR264C	YGR162W		
YJL077C	YOR027W		YLL026W*, YOR027W, YFL010W-A*
YLR170C	YJL115W$^\times$		

Table 1: Comparison of predictions based on ICP from previous findings (Meinshausen et al., 2016) (left column) to predictions from our implementation of ICP-s (middle column) and PI-var (right column). We observe consistency in most of their predictions on the 10 target genes selected by Meinshausen et al. (2016). Genes highlighted in blue are validated to have significant effects on the corresponding target gene; genes marked with a superscript * cannot be checked given existing data, while those with a superscript $^\times$ can be checked but do not have a significant effect; blank means no invariant features are predicted by the method.

We also compare our findings to previous findings in Meinshausen et al. (2016) which is also based on ICP. They use ICP at significance level $\alpha = 0.01$ after a Lasso pre-screening step, followed by a stable selection procedure over 50 randomly bootstrapped dataset [2]. The comparison results are summarized in Table 1. We observe that most of the true effects predicted by PI-var, ICP-s, and Meinshausen et al. (2016) are the same. Moreover, PI-var infers more invariant feature genes; However, not all of these predictions can be checked with existing experimental data.

[2]In addition, the dataset used in Meinshausen et al. (2016) only has 160 observational samples, whereas the updated version that we use has 262 observational samples, which can be downloaded here `https://deleteome.holstegelab.nl/`.