Causal Inference using Gaussian Processes
with Structured Latent Confounders

Sam Witty [1] **Kenta Takatsu** [1] **David Jensen** [1] **Vikash Mansinghka** [2]

Abstract

Latent confounders—unobserved variables that influence both treatment and outcome—can bias estimates of causal effects. In some cases, these confounders are shared across observations, e.g. all students taking a course are influenced by the course's difficulty in addition to any educational interventions they receive individually. This paper shows how to semiparametrically model latent confounders that have this structure and thereby improve estimates of causal effects. The key innovations are a hierarchical Bayesian model, *Gaussian processes with structured latent confounders* (GP-SLC), and a Monte Carlo inference algorithm for this model based on elliptical slice sampling. GP-SLC provides principled Bayesian uncertainty estimates of individual treatment effect with minimal assumptions about the functional forms relating confounders, covariates, treatment, and outcome. Finally, this paper shows GP-SLC is competitive with or more accurate than widely used causal inference techniques on three benchmark datasets, including the Infant Health and Development Program and a dataset showing the effect of changing temperatures on state-wide energy consumption across New England.

1. Introduction

Multiple causal models can be *observationally equivalent*, i.e., they induce the same likelihoods for observed data, while producing different estimates of the effects of an intervention. This observational equivalance between causal models is the basis for the colloquial expression "correlation does not imply causation." Distinguishing between causal models, and estimating the effects of interventions, requires untestable assumptions about causal structure.

One such common assumption is *unconfoundedness* (Imbens & Rubin, 2015), i.e., that there exist no latent variables that influence both treatment and outcome. This assumption enables the unique identification of interventional distributions from the joint distribution over observed variables (Pearl, 2009) and reduces causal inference to probabilistic estimation. Unfortunately, assuming unconfoundedness is often unreasonable in real observational settings (Shadish et al., 2008). However, it may be more reasonable to assume uncounfoundedness for a subset of data instances that are known to share a common structure.

For example, suppose an educator proposes a new policy of holding back poor performing kindergarten students (Hong & Raudenbush, 2006; Hong & Yu, 2008) with the intention of increasing their future academic performance. To estimate the effect of this policy change, they gather data on student retention and education outcomes from a national database. Here, the unconfoundedness assumption is not justified, as the schools' retention policies are likely to be influenced by local economic conditions, which may also influence student outcomes through other causal mechanisms, such as the availability of educational resources. However, the assumption may be justified when considering only students within a particular school, as this subset of students are similarly influenced by local economic conditions. In other words, statistical relationships within a school are less likely to be biased by latent confounders than are statistical relationships across the entire population.

In this paper, we present Gaussian processes with structured latent confounders (GP-SLC), a novel Bayesian nonparametric approach to causal inference with hierarchical data. The key innovation behind GP-SLC is to place Gaussian process priors over functions in a hierarchical structural causal model, bringing the flexibility of Gaussian process models to a wide variety of practical causal inference techniques. GP-SLC naturally handles binary and continuous treatments and requires minimal assumptions about functional relationships between latent confounders, observed covariates, treatment, and outcomes. See Figure 1 for an overview on how GP-SLC estimates counterfactual outcomes from data.

Symbol	Description	Entity
$\mathbf{u}_o$	Confounders	Object
$\mathbf{x}_i$	Covariates	Instance
$\mathbf{t}_i$	Treatment	Instance
$\mathbf{y}_i$	Outcome	Instance

(a) Variable descriptions.

(b) Causal graph for GP-SLC.

$$f_u \sim GP(0, k_u) \qquad f_x \sim GP(0, k_x)$$
$$f_t \sim GP(0, k_t) \qquad f_y \sim GP(0, k_y)$$
$$\mathbf{u}_{o=1\ldots N_O} = f_u(\boldsymbol{\epsilon}_{u_o})$$
$$\mathbf{x}_{i=1\ldots N_I} = f_x(\mathbf{u}_{o=Pa(i)}, \boldsymbol{\epsilon}_{x_i})$$
$$\mathbf{t}_{i=1\ldots N_I} = f_t(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \boldsymbol{\epsilon}_{t_i})$$
$$\mathbf{y}_{i=1\ldots N_I} = f_y(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \mathbf{t}_i, \boldsymbol{\epsilon}_{t_i})$$

(c) Prior and causal functions for GP-SLC.

(d) Example grounding of the structural causal model in (b) and (c). Latent confounders are shared within objects.

(e) Treatment, covariates, and inferred object-level confounders for instances in (d). Color = o. Size = $\mathbf{x}$.

(f) Kernel covariance matrix over observed ($\mathbf{y}_i$) and counterfactual ($\mathbf{y}_{1,t_*}$) outcomes for instances in (e). Dark > light.
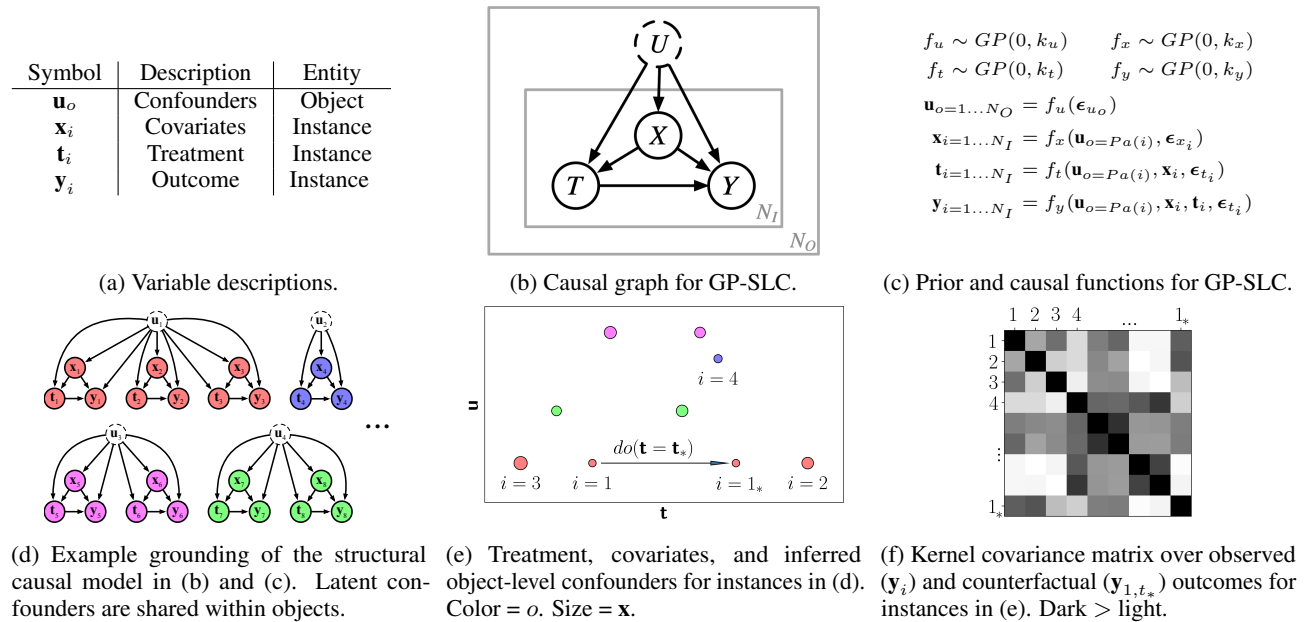
Figure 1: **Model summary.** GP-SLC (a-c) is a Gaussian process model for causal inference in settings where object-level latent confounders, $\mathbf{u}$, influence instance-level observed covariates, $\mathbf{x}$, treatment, $\mathbf{t}$, and outcome, $\mathbf{y}$, random variables. For a given grounding (d), the outcome kernel function, k_y, applied to treatment, covariates, and inferred confounders (e) induces the covariance between observed and counterfactual outcomes (f). Instances belonging to the same object always have the same inferred latent $\mathbf{u}$. In this example, the counterfactual outcome $\mathbf{y}_{1,t_*}$ has high covariance with factual outcomes $\mathbf{y}_1$ and $\mathbf{y}_2$. $\mathbf{y}_{1,t_*}$ has low, but non-zero, covariance with $\mathbf{y}_4$ because $\mathbf{u}_{Pa(1)} \not\approx \mathbf{u}_{Pa(4)}$, despite the fact that $\mathbf{t}_* \approx \mathbf{t}_4$ and $\mathbf{x}_1 \approx \mathbf{x}_4$.

2. Background

2.1. Object Conditioning

Recent work has studied how the analytical procedure of partitioning data based on a known object hierarchy (e.g. students belonging to the same school) relates to the syntax and semantics of causal graphical models (Jensen et al., 2019). This work concludes that conditioning on the identify of objects (referred to as *object conditioning*) is distinct from existing notions of conditioning on the values of variables. Importantly, object conditioning constrains a set of latent variables to be identical across a set of instances, but does not constrain the particular value of those variables. Furthermore, the statistical implications of object conditioning differ from those of variable conditioning in that object conditioning does not induce collider bias when variables on the object are caused jointly by treatment and outcome.

Partitioning hierarchical data in this way is the key analytical procedure for a variety of practical causal inference techniques, including within-subjects designs (Loftus & Masson, 1994), difference-in-differences designs (Shadish et al., 2008), longitudinal studies (Liang & Zeger, 1986), twin studies (Boomsma et al., 2002), and multi-level-modeling (Gelman & Hill, 2006). As in the student retention example, these techniques take advantage of background

knowledge about which instances (students) belong to which objects (schools) to mitigate the biasing effects of latent confounders. However, these methods typically rely on simple parametric assumptions, such as linear functional dependencies. These parametric assumptions are often unjustified in real domains, leading to poor estimates of causal effect.

We employ the idea of object conditioning directly in the GP-SLC model, constraining the joint distribution over individuals' latent confounders instead of treating object identity as a covariate in and of itself. By explicitly performing inference over object-level latent confounders, GP-SLC's estimates of counterfactual outcomes in one object are informed by observed outcomes in another. Sharing information between objects in this way is particularly valuable when each object contains few observed instances, as we show in Section 6.

2.2. Causal Inference with Latent Confounders

Latent confounders—unobserved variables that cause both treatment and outcome—bias estimates of treatment effect. However, this bias can be adjusted for with additional background knowledge, such as that a latent confounder influences an observed proxy variable (Kuroki & Pearl, 2014; Miao et al., 2018). Similarly, recent work indicates that latent confounders can be adjusted for if they cause multiple candidate treatment variables (Wang & Blei, 2019).

GP-SLC is similar to these approaches, in that it leverages additional background knowledge to adjust for latent confounders. However, unlike prior work using generative models for causal inference with latent confounders, it leverages known hierarchical structure to identify causal effects.

2.3. Gaussian Process Models

Gaussian process models are a flexible technique for probabilistic modeling. Specifically, a Gaussian process is a distribution over deterministic functions, $\mathbf{y} = f(\mathbf{x})$, $f \sim GP(m, k)$, which is fully specified by its mean function, $m(\mathbf{x})$ and covariance function, $k(\mathbf{x}, \mathbf{x}')$, which we will refer to as the kernel function (Rasmussen, 2003). By definition, any finite collection of draws from a Gaussian process prior are jointly Gaussian distributed, $Y \sim \mathcal{N}(\mu, \Sigma)$, where $\mu_i = m(\mathbf{x}_i)$ and $\Sigma_{i,i'} = k(\mathbf{x}_i, \mathbf{x}_{i'})$. We denote such covariance matrices as $K(X, X)$, where X and Y are matrices of all $\mathbf{x}_i$ and $\mathbf{y}_i$ respectively. It is common to set the prior mean function to $m(\mathbf{x}) = 0$, which we do in GP-SLC.

This identity is useful for two reasons: (i) it provides an explicit likelihood, which can be used to perform inference over latent variables (Lawrence, 2004; Titsias & Lawrence, 2010); and (ii) it enables closed-form out-of-sample probabilistic prediction (Rasmussen, 2003). We take advantage of both of these characteristics in GP-SLC, performing approximate inference over latent confounders in Section 4.1 and predicting counterfactual outcomes in Section 4.2.

2.4. Structural Causal Models

GP-SLC can be thought of as a Bayesian semiparametric prior over functions in a structural causal model (SCM). SCMs provide a syntax and semantics for reasoning about interventional and counterfactual distributions in a system of random variables (Pearl, 2009). Counterfactuals (Pearl, 2011)—answers to what-if questions—are expressed in probability notation as $P(\mathbf{y}_{\mathbf{t}_*}|\mathbf{y}, \mathbf{t})$, where $\mathbf{y}_{\mathbf{t}_*}$ is the counterfactual outcome under intervention $do(\mathbf{t} = \mathbf{t}_*)$, $\mathbf{y}$ is the observed outcome, and $\mathbf{t}$ is the observed treatment. In our education example, we may be interested in the counterfactual, "given that the student was not retained in kindergarten and they performed poorly in high school, how would they have performed if they had been retained?" Answering these kinds of counterfactual queries involves: (i) computing the conditional distribution over latent variables and exogenous noise given observed evidence; (ii) applying the intervention to the structural causal model; and (iii) recomputing the distribution over the outcome variable(s) using the modified structural causal model. We apply this procedure to estimate counterfactual outcomes using GP-SLC in Section 3.

3. Gaussian Processes with Structured Latent Confounders

Consider the common scenario where there are N_O object-level latent confounders ($\mathbf{u}_o \in \mathbb{R}^{N_U}$) that influence N_I instances of observed treatment ($\mathbf{t}_i \in \mathbb{R}$), covariates ($\mathbf{x}_i \in \mathbb{R}^{N_X}$), and outcomes ($\mathbf{y}_i \in \mathbb{R}$). We can describe this scenario as a structural causal model, where the particular functions relating $\mathbf{u}$, $\mathbf{x}$, $\mathbf{t}$, and $\mathbf{y}$ are given by the following:

$$
\begin{aligned}
\mathbf{u}_{o=1...N_O} &= f_u(\boldsymbol{\epsilon}_{u_o}) \\
\mathbf{x}_{i=1...N_I} &= f_x(\mathbf{u}_{o=Pa(i)}, \boldsymbol{\epsilon}_{x_i}) \\
\mathbf{t}_{i=1...N_I} &= f_t(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \boldsymbol{\epsilon}_{t_i}) \\
\mathbf{y}_{i=1...N_I} &= f_y(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \mathbf{t}_i, \boldsymbol{\epsilon}_{y_i}).
\end{aligned}
\tag{1}
$$

If all instances belong to the same object ($N_O = 1$) the structural causal model in Equation 1 reduces to the standard propositional case and the latent $\mathbf{u}$ will not bias estimated counterfactual outcomes. However, if we wish to estimate counterfactual outcomes using instances from multiple objects ($N_I > N_O > 1$), $\mathbf{u}$'s influence on $\mathbf{t}$ and $\mathbf{y}$ would appear to render counterfactual queries nonparametrically unidentifiable (Pearl, 2009). Rather than be fully nonparametric, GP-SLC instead places a Gaussian process prior over each function in the structural causal model in Equation 1, with kernel functions k_x, k_t, and k_y respectively as follows:

$$
f_x \sim GP(0, k_x) \quad f_t \sim GP(0, k_t) \quad f_y \sim GP(0, k_y).
$$

The particular choice of each kernel function plays an important role in the prior over functions, and by extension the conditional distribution over counterfactual outcomes. We use a radial basis function (RBF) kernel with automatic relevance determination (ARD) (Neal, 2012) and additive Gaussian exogenous noise for each Gaussian process prior. Each kernel is parameterized by a set of kernel lengthscales, λ, scaling factors, σ^2, and exogenous noise variances σ_ϵ^2. We assume f_u is the identity function. We refer to the noise-free component of each kernel function as k', e.g. $k_t = k'_t([\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i], [\mathbf{u}_{o'=Pa(i')}, \mathbf{x}_{i'}]) + \sigma_{\epsilon_y}^2 \delta_{i,i'}$, where $\sigma_{\epsilon_y}^2$ is the exogenous noise variance, $\delta_{i,i'}$ is the Dirac-delta function at $i' = i$, and k'_t is the ARD kernel. See the supplementary materials for detailed descriptions of these kernels.

In addition to placing Gaussian process priors on the functions in the structural causal model in Equation 1, we also place inverse-gamma priors, $P(\theta) = \gamma^{-1}(\theta; \alpha_\theta, \beta_\theta)$ on each $\theta \in \Theta$, where Θ is the set of all kernel lengthscales, scaling factors, and exogenous noise variances. In Section 4.1 we show how to perform approximate posterior inference on Θ.

3.1. Conditional Density

As f_y, f_t, and f_x are all drawn from Gaussian process priors, $P(Y|T, X, U, \Theta), P(T|X, U, \Theta)$, and $P(X|U, \Theta)$

Algorithm 1 Individual Treatment Effect Estimation

1: **Input:**
2:     Intervention assignment: $\mathbf{t}_*$
3:     Data: Y, T, X
4:     Prior hyperparameters: $\alpha_{\theta \in \Theta}, \beta_{\theta \in \Theta}$
5:     Inference parameters: $N_{\text{Outer}}, N_{\text{MH}}, N_{\text{ES}}, \text{drift}_{\theta \in \Theta}$
6: **Procedure:**
7:     $\theta \sim \gamma^{-1}(\alpha_\theta, \beta_\theta), \forall \theta \in \Theta$ ▷ Prior sample
8:     $\mathbf{u}_o \sim \mathcal{N}(0, \sigma_U^2 I_{N_U}), \forall o = 1...N_O$ ▷ Prior sample
9:     $ITESamples \leftarrow \{\}$
10:    **for** $l = 1$ **to** n_{Outer} **do**
11:        $\Theta \leftarrow \text{HyperparameterUpdate}(...)$ ▷ Algorithm 2
12:        $U \leftarrow \text{ConfounderUpdate}(...)$ ▷ Algorithm 3
13:        $\mathbf{w}_i \leftarrow [\mathbf{t}_i, \mathbf{x}_i, \mathbf{u}_{o=pa(i)}], \forall i \in 1...N_I$
14:        $\mathbf{w}_{i,*} \leftarrow [\mathbf{t}_*, \mathbf{x}_i, \mathbf{u}_{o=pa(i)}], \forall i \in 1...N_I$
15:        $\mu_{ITE} \leftarrow (K'(W, W_*)\text{-}K'(W, W))K(W, W)^{-1}Y$
16:        $ITE \sim \mathcal{N}(\mu_{ITE}, \Sigma_{ITE})$ ▷ See Supplement for Σ_{ITE}
17:        $ITESamples \leftarrow ITE \cup ITESamples$
18:    **return** $ITESamples$

---

are all multivariate Gaussian distributions with mean zero and covariance given by their respective kernel covariance matrices. For example, $P(T|X, U, \Theta) = \mathcal{N}(T; 0, K_t([U, X], [U, X]))$, where $K_t([U, X], [U, X])_{i,i'} = k_t([\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i], [\mathbf{u}_{o'=Pa(i')}, \mathbf{x}_{i'}])$. As $\mathbf{u}_o$ is given by the identity function of exogenous Gaussian noise, $P(\mathbf{u}_o|\Theta) = \mathcal{N}(\mathbf{u}_o; 0, \sigma_{\epsilon_u}^2 I_{N_U})$. Therefore, the joint density is given by the following, which we use in Algorithms 2 and 3:

$$P(Y, T, X, U, \Theta) = P(Y|T, X, U, \Theta)P(X|U, \Theta)$$
$$P(T|X, U, \Theta) \prod_{o=1...N_O} P(\mathbf{u}_o|\Theta)P(\Theta).$$

By placing Gaussian process priors over each function in the hierarchical structural model, we encode our assumptions about which configurations of observed and latent variables are reasonable a-priori. Using a radial basis function kernel, we assume that if two objects have similar object-level latent confounders, they are likely to induce similar distributions over observed covariates, treatment, and outcome. Placing higher density on smooth structural causal functions in this way enables inference over object-level confounders.

4. Estimating Treatment Effects

In this section we describe how to estimate the *individual treatment effect*, $ITE_{i,t_*} = \mathbf{y}_{i,t_*} - \mathbf{y}_i$, the difference between observed and counterfactual outcomes for the ith instance. Standard aggregate measures of causal effect, such as the *sample average treatment effect*, $SATE_{t_*} = \frac{1}{N_I} \sum_i ITE_{i,t_*}$, can be derived from the individual treatment effect. We use ITE_{t_*} to denote the vector of individual treatment effects for the intervention $do(\mathbf{t}_i = \mathbf{t}_*)$ applied uniformly to each

instance i, although the estimation procedure can be easily applied to any arbitrary set of intervention assignments.

First, note that when exogenous noise is additive in f_y, i.e $f_y(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \mathbf{t}_i, \epsilon_{t_i}) = f'_y(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \mathbf{t}_i) + \epsilon_{t_i}$, as in the GP-SLC model, individual treatment effect is given by the difference between noise-free functions $ITE_{i,t_*} = f'_y(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \mathbf{t}_*) - f'_y(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \mathbf{t}_i)$. We denote the outcome of these noise-free functions as $\mathbf{y}'_{i,t_*}$ and $\mathbf{y}'_i$, and the vector of outcomes as Y'_{t_*} and Y' respectively.[1] As $U \cup X$ blocks all backdoor paths from T to Y, we have that the distribution over individual treatment effects is given by the following expression (Pearl, 2009):

$$P(ITE_{t_*}|Y, T, X) = P(Y'_{t_*} - Y'|Y, T, X)$$
$$= \int P(Y'_* - Y'|T_*, Y, T, X, U, \Theta)P(U, \Theta|Y, T, X)dUd\Theta.$$

This equation directly informs our hybrid procedure for estimating counterfactual outcomes shown in Algorithm 1, (i) generate approximate samples from the posterior $\hat{U}, \hat{\Theta} \sim P(U, \Theta|Y, T, X)$ and (ii) for each posterior sample compute the conditional distribution $(Y'_* - Y'|T_*, Y, T, X, \hat{U}, \hat{\Theta})$ in closed-form, taking advantage of Gaussian closure under conditioning and subtraction. As the posterior distribution $P(U, \Theta|Y, T, X)$ is intractable for non-trival kernels, we turn to Monte Carlo approximate inference techniques.

4.1. Approximate Inference: U and Θ

Because we assume that our structural functions were drawn from Gaussian Processes, which provide a closed-form expression for the conditional density of the data, we are able to use standard likelihood-based approximate inference techniques. In our experiments, we approximate this posterior distribution using elliptical slice sampling (Murray et al., 2010) for the latent confounder, U, and random walk Metropolis Hastings (Hastings, 1970) on all kernel hyperparameters and exogenous noise variances, Θ. Psuedo-code implementations are presented in Algorithms 2 and 3.

4.2. Exact Inference: $Y'_* - Y'$

To estimate $P(Y'_* - Y'|T_*, Y, T, X, U, \Theta)$, we extend the Gaussian process model over in-sample and out-of-sample outcomes (Rasmussen, 2003). Using the shorthand $\mathbf{w}_i = [\mathbf{t}_i, \mathbf{x}_i, \mathbf{u}_{o=pa(i)}]$ and $\mathbf{w}_{i,*} = [\mathbf{t}_*, \mathbf{x}_i, \mathbf{u}_{o=pa(i)}]$, the joint distribution over observed outcomes, Y, noise-free outcomes for each observed instance, Y', and noise-free counterfactual outcomes, Y'_* conditioned on observed treatments, T, covariates, X, inferred confounders, U, and kernel hyperparameters, Θ, is Gaussian distributed as follows, where

---

[1]Noise-free prediction is often denoted as f in Gaussian process regression models. We avoid this notation to avoid confusion with functions in the structural causal model.

Algorithm 2 Hyperparameter Update - Random Walk MH

1: **Input:**
2:     Data: Y, T, X
3:     Posterior sample: U, Θ
4:     Prior hyperparameters: $\alpha_{\theta \in \Theta}, \beta_{\theta \in \Theta}$
5:     Inference parameters: $\text{N}_{\text{MH}}, \text{drift}_{\theta \in \Theta}$
6: **Procedure:**
7:     **for** $j = 1$ **to** N_{MH} **do**
8:       **for** $\theta \in \Theta$ **do**
9:         $\alpha_{\theta'} \leftarrow \theta^2 / \text{drift}_\theta$
10:         $\beta_{\theta'} \leftarrow \theta(\alpha_{\theta'} - 1)$
11:         $\theta' \sim \gamma^{-1}(\alpha_{\theta'}, \beta_{\theta'})$
12:         $\alpha_{\theta''} \leftarrow \theta'^2 / \text{drift}_{\theta'}$
13:         $\beta_{\theta''} \leftarrow \theta'(\alpha_{\theta''} - 1)$
14:         $\Theta' \leftarrow \Theta \setminus \theta \cup \theta'$
15:         $A \leftarrow \dfrac{P(Y, T, X, U, \Theta')}{P(Y, T, X, U, \Theta)} \dfrac{\gamma^{-1}(\theta'; \alpha_{\theta'}, \beta_{\theta'})}{\gamma^{-1}(\theta; \alpha_{\theta''}, \beta_{\theta''})}$
16:         $\eta \sim \text{Uniform}(0, 1)$
17:         **if** $\eta > \min(A, 1)$ **then**
18:           $\Theta = \Theta'$
19:     **return** Θ

Algorithm 3 Confounder Update - Elliptical Slice Sampling

1: **Input:**
2:     Data: Y, T, X
3:     Posterior sample: U, Θ
4:     Inference parameter: N_{ES}
5: **Procedure:**
6:     **for** $j = 1$ **to** N_{ES} **do**
7:       **for** $k = 1$ **to** N_{U} **do**
8:         done $\leftarrow$ False
9:         $\nu \sim \mathcal{N}(0, \sigma_{\epsilon_U}^2 I)$
10:         $y \sim \text{Uniform}(0, P(Y, T, X, U, \Theta))$
11:         $\phi \sim \text{Uniform}(0, 2\pi)$
12:         $[\phi_{\min}, \phi_{\max}] \leftarrow [\phi - 2\pi, \phi]$
13:         **while** not done **do**
14:           $U'_{*,k} \leftarrow U_{*,k} \cos \phi + \nu \sin \phi$
15:           **if** $P(Y, T, X, U, \Theta) > y$ **then**
16:             $\mathbf{u}_k \leftarrow \mathbf{u}'_i$
17:             done $\leftarrow$ True
18:           **else**
19:             **if** $\phi < 0$ **then** $\phi_{\min} \leftarrow \phi$ **else** $\phi_{\max} \leftarrow \phi$
20:             $\phi \sim \text{Uniform}(\phi_{\min}, \phi_{\max})$
21:     **return** U'

$K(W, W) = K'(W, W) + \sigma_Y^2 I_{N_I}$ and $K'(W, W)$ is the kernel matrix of k'_y given Θ:

$$\left(\begin{bmatrix} Y \\ Y' \\ Y'_* \end{bmatrix} \Big| T_*, T, X, U, \Theta \right)$$
$$\sim \mathcal{N} \left(0, \begin{bmatrix} K(W, W) & K'(W, W) & K'(W, W_*) \\ K'(W, W) & K'(W, W) & K'(W, W_*) \\ K'(W_*, W) & K'(W_*, W) & K'(W_*, W_*) \end{bmatrix} \right).$$

As Gaussian distributions are closed under conditioning and subtraction, we have that $(Y'_* - Y' | T_*, Y, T, X, U, \Theta)$ is also jointly Gaussian distributed as follows, where $\mu_{ITE} = (K'(W, W_*) - K'(W, W))K(W, W)^{-1}Y$:

$$(Y'_* - Y' | T_*, Y, T, X, U, \Theta) \sim \mathcal{N}(\mu_{ITE}, \Sigma_{ITE}) \qquad (2)$$

See the supplementary materials for details and for a closed-form expression for Σ_{ITE}.

5. Asymptotic Posterior Consistency

In the special case where each RBF kernel in the GP-SLC model is replaced with a linear kernel, $k(A, A') = A \cdot A'$, shared confounding among instances enables asymptotically consistent estimates of individual treatment effect. This is contrasted with the propositional setting (i.e. $N_O = N_I$) which does not lead to asymptotically consistent counterfactual estimation. Informally, a continuous random variable ψ is asymptotically consistent if its posterior $P(\psi | data)$ ap-

proaches a Dirac-delta distribution at some point ψ', regardless of the prior $P(\psi)$. We present proofs of Proposition 5.1 and Theorems 5.2 and 5.3 in the supplementary materials.

Proposition 5.1. *When $N_O = N_I$, ITE_{t_*} is not asymptotically consistent $\forall t_* \in \mathbb{R}$.*

Theorem 5.2. *Assume there exists an object o that is the parent of n instances, $I' = \{i'_1, ..., i'_n\}$. Then ITE_{t_*} is asymptotically consistent as n approaches $\infty, \forall t_* \in \mathbb{R}$.*

Theorem 5.3. *Assume there exist n objects $\mathbb{O} = \{o_1, ..., o_n\}$, each of which are the unique parents of $k \geq 2$ instances $I'_o = \{i'_{o,1}, ..., i'_{o,k_o}\}$. Then ITE_{t_*} is asymptotically consistent as n approaches $\infty, \forall t_* \in \mathbb{R}$.*

6. Experiments

Unlike associational models, which can be evaluated using accuracy on held-out test data, causal models produce predictions about unobserved *counterfactual* distributions. As a result, effective evaluation of causal models requires different methods (Gentzel et al., 2019). We evaluate the GP-SLC model using three benchmarks with known counterfactual outcomes. In Section 6.1, we evaluate GP-SLC using a fully synthetic hierarchical data generating process. In Section 6.2 we modify the Infant Health and Development Program (IHDP) benchmark (Hill, 2011) to include hierarchical structure and latent confounders. In Section 6.3 we introduce and evaluate on a new benchmark task for observational causal inference with hierarchical data, predicting the effect of changes in temperature on state-wide

(a) Original data. (b) Unbiased sampling. (c) Biased sampling. (d) Energy consumption (GWh)

(e) Mean squared error in estimated sample average treatment effect.
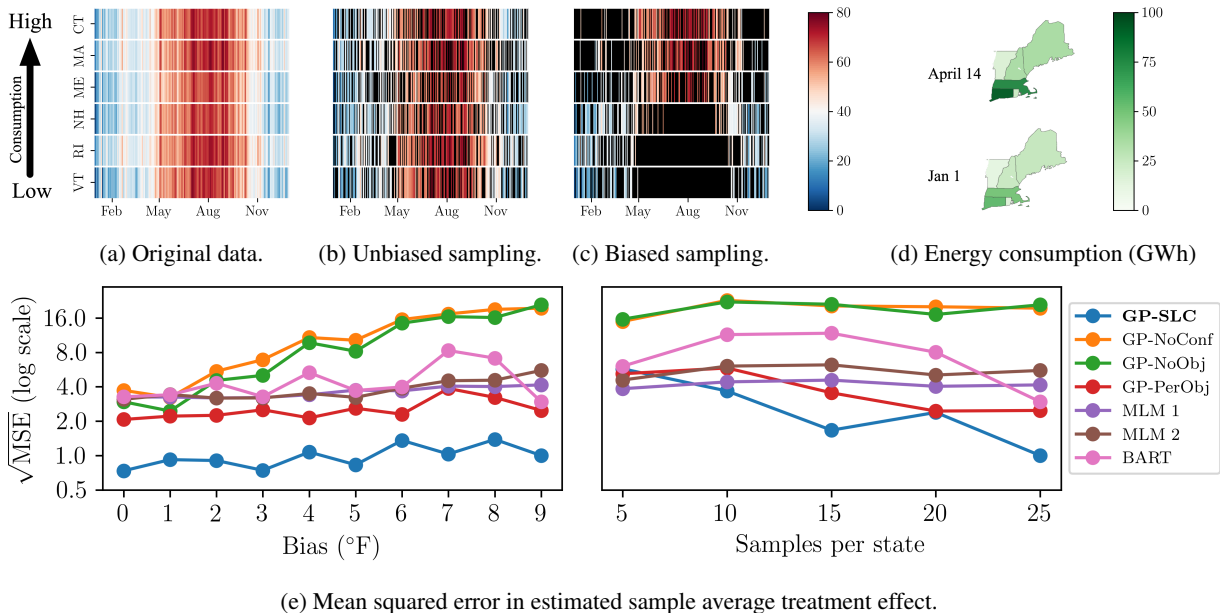
Figure 2: **Process and results for New England energy consumption benchmark.** We sample hotter days with higher probability for states with higher daily energy consumption (a-d). Sampling in this way simulates confounding, creating an observational relationship (consumption is signicantly higher in hotter days) that differs from the causal relationship (low or high temperature causes a moderate increase in energy consumption). GP-SLC (this paper) produces accurate estimates of counterfactual outcomes, despite this confounding bias (e). For baselines that ignore hierarchical structure (GP-NoObj and GP-NoConf), accuracy decreases significantly with increasing confounding bias. Results are normalized by the $\sqrt{\text{MSE}}$ of the GP-SLC model with bias $= 9°F$ and 25 samples per state.

electric energy consumption in New England (NEEC).

We implement the GP-SLC model using Gen (Cusumano-Towner et al., 2019), a probabilistic programming language with programmable inference. Except where otherwise specified we set $N_U = 3$ and $\alpha_\theta = \beta_\theta = 4$ for each inverse gamma prior over kernel hyperparameters and exogenous noise variance. We estimate individual treatment effects using Algorithm 1, with $N_{\text{Outer}} = 5000$, $N_{\text{MH}} = 3$, $N_{\text{ES}} = 5$, and $\text{drift}_\theta = 0.5, \forall \theta \in \Theta$.

We compare the GP-SLC model against six baselines: a GP regression model that ignores latent confounding variables (GP-NoConf), a GP-SLC model where each instance is incorrectly assigned a single object (GP-NoObj), a seperate GP regression model for each object (GP-PerObj), Bayesian additive regression trees (BART) (Hill, 2011), a random slope and intercepts linear model (MLM 1), and a random intercepts linear model (MLM 2) (Gelman, 2006). The Gaussian process baselines are ablations of the full GP-SLC model, and use the same kernels, priors over hyperparameters, and inference scheme. The BART baseline uses the object identifier, o, as an additional covariate. See the supplementary materials for additional details on baselines.

We use two evaluation metrics to evaluate GP-SLC and baselines, *mean squared error of the sample average treat-*

ment effect, $\text{MSE} = \mathbb{E}_{t_*}[(SATE^*_{t_*} - SATE_{t_*})^2]$, and *precision in estimation of heterogenous effect* (Hill, 2011), $\text{PEHE} = \mathbb{E}_{t_*}[\sum_i^{N_i} (ITE^*_{i,t_*} - ITE_{i,t_*})^2/N_i]$, where ITE^*_{i,t_*} and $SATE^*_{T_*}$ are the actual effects and ITE_{i,t_*} and $SATE_{t_*}$ are the predicted effects. For the synthetic benchmark, we average over 100 regular intervals between the 5th and 95th percentile of treatment assignment in the observational data. For the NEEC benchmask, we average over $\{30, 30.1, ..., 70°F\}$.

6.1. Synthetic Data

We evaluate GP-SLC and various baselines on two synthetic datasets with hierarchically structured latent confounders, one with additive and one with multiplicative treatment and outcome functions. Both synthetic datasets are generated using three dimensional object-level confounders for 20 objects, each of which contains 10 instances. Observed instance-level covariates are generated as a linear function of object-level Gaussian distributed latent confounders. Details for synthetic treatment and outcome functions are presented in the supplementary materials, and evaluation results are shown in Table 1. GP-SLC consistently matches and exceeds the counterfactual prediction performance of the six baselines on synthetic data. Baselines that ignore object
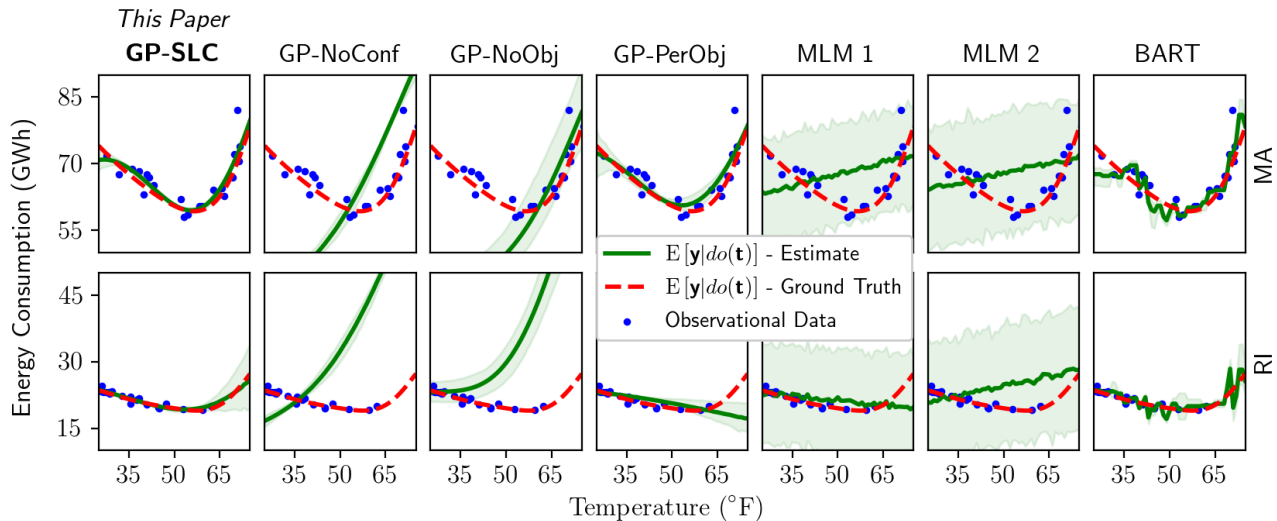
Figure 3: **Comparison among methods on the New England energy consumption benchmark.** Above are GP-SLC and all baselines' effect estimates on the NEEC benchmark with bias $= 9°F$ and 25 samples per state. Green shaded regions indicate 90% credible intervals. GP-SLC effectively recovers the effect of temperature on energy consumption, despite the latent confounding introduced by biased sampling. The best performing baseline, GP-PerObj, produces poor estimates of the effect of high temperatures in Rhode Island.

structure (GP-NoConf and GP-NoObj) produce the least accurate counterfactual predictions.

In addition to the synthetic experiments presented in Table 1, we tested the behaviour of GP-SLC using two alternative synthetic data generating processes. On the first, a linear structural data generating process with shared confounding, GP-SLC produces comparable estimates to the multi-level model baselines. On the second, in which each object shares a common effect of treatment and outcome rather than a common cause, GP-SLC is not susceptible to collider bias (Berkson, 1946; Elwert & Winship, 2014). This empirical finding is consistent with recent theory on object conditioning (Jensen et al., 2019).

Model	Additive		Multiplicative	
	$\sqrt{\text{PEHE}}$	$\sqrt{\text{MSE}}$	$\sqrt{\text{PEHE}}$	$\sqrt{\text{MSE}}$
GP-SLC	**1.0**	**1.0**	**1.0**	1.0
GP-NoConf	21.3	25.3	4.2	7.6
GP-NoObj	22.2	27.0	4.5	8.1
GP-PerObj	3.7	3.4	1.1	**0.9**
MLM1	1.2	1.02	2.4	2.9
MLM2	1.3	1.6	4.4	9.3
BART	8.5	10.7	2.6	4.3

Table 1: Results on synthetic data with additive and multiplicative nonlinear treatment and outcome functions. Scores are normalized by the score of GP-SLC. Lower is better.

6.2. Infant Health and Development Program

The IHDP benchmark (Hill, 2011) uses real data for treatments (whether a child recieves high-quality child care and home visits from a trained provider) and covariates (birth weight, head circumference, etc.) from the 1992 Infant Health and Development Program (Ramey et al., 1992) with a synthetic nonlinear outcome function. We modify the IHDP benchmark to simulate hierarchically structured data by randomly duplicating 30% of the data instances and reassigning the duplicate's treatment assignment to be the opposite of the original instance. In order to introduce variation between duplicated instances, we add noise to each individuals' continuous covariates from a $\mathcal{N}(0, \sigma_j^2)$, where σ_j^2 is 5% of the jth covariate's marginal variance. We obscure the remaining 15 categorical covariates, representing object-level latent confounding. Even though the 15 categorical covariates are obscured from the GP-SLC model, they are identical across duplicates, unlike the observed covariates. We then generate observed and counterfactual outcomes using the benchmark synthetic outcome function, applied to treatment, modified covariates, and latent confounders. In this setting, $Pa(i) = Pa(i')$ if instance i is a duplicate of instance i' or vice versa. Although each duplicate's treatment assignment is deterministic, the overall relationship between treatment and outcome is still confounded, as we only duplicate a subset of the original instances.

For the IHDP benchmark, which has binary treatment variables, we modify the GP-SLC model by replacing the ex-

Model	Control		Treated	
	$\sqrt{\text{PEHE}}$	$\sqrt{\text{MSE}}$	$\sqrt{\text{PEHE}}$	$\sqrt{\text{MSE}}$
GP-SLC	**1.0**	**1.0**	1.0	1.0
GP-NoConf	1.03	1.07	1.04	0.94
GP-NoObj	1.11	1.02	**0.82**	1.08
MLM1	68.3	33.2	106.7	1028.4
MLM2	73.3	389.1	45.8	63.2
BART	3.7	1.1	2.4	**0.33**
BALReg	5.1	82.7	1.9	0.5
BALNN	2.1	7.0	1.7	4.5
TMLE	n/a	209.8	n/a	12.2
IPTW	n/a	50.6	n/a	90.5

Table 2: Results on the modified infant health and development program benchmark, shown seperately for treated and untreated individuals. Scores are normalized by the score of GP-SLC. TMLE and IPTW do not estimate individual treatment effects. Lower is better.

pression $\mathbf{t}_i = f_t(\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \boldsymbol{\epsilon}_{t_i})$ with the expressions $\hat{\mathbf{t}}_i = f_{\hat{\mathbf{t}}}(\mathbf{u}_{o=Pa(i)}, X_i, \boldsymbol{\epsilon}_{\hat{t}_i})$ and $\mathbf{t}_i \sim \text{Bernoulli}(expit(\hat{\mathbf{t}}_i))$. In this setting, we use elliptical slice sampling to approximate the latent logit probability of treatment, $\hat{\mathbf{t}}$.

Given the small size of each object, we omit the GP-PerObj baseline model from this evaluation. As the IHDP benchmark includes binary treatment variables we compared against four additional baselines: balanced linear regression (BalReg) and balanced neural nets (BALNN) (Johansson et al., 2016), targeted maximum likelihood estimation with the superlearner (TMLE) (Van der Laan et al., 2007), and inverse probability of treatment weighting with logistic regression (IPTW) (Imbens & Rubin, 2015).

Results of the IHDP evaluation are presented in Table 2. GP-SLC matches and exceeds the performance of other baselines when predicting the effect of assigning treatment to individuals who were previously untreated. In this setting, the linear models (MLM 1 and MLM 2) produce the least accurate counterfactual predictions.

6.3. New England Energy Consumption

We introduce a new benchmark for estimating heterogenous effects in hierarchically structured settings, predicting the effect of changing temperature on state-wide electric energy consumption in New England. Unlike the evaluation in Section 6.2, which includes real treatments, covariates, and confounders and a synthetic outcome function, the New England energy consumption (NEEC) benchmark preserves outcome functions from real quasi-experimental data, and uses biased sampling to induce confounding. Specifically, we generate data for the NEEC benchmark task using the New England Independent Service Operator's public records

on hourly dry-bulb temperature and state-wide energy consumption for the 2018 calendar year (ISO New England, 2018), which we then aggregate into daily averages.

While the marginal distribution over daily average temperature is nearly identical across states in the original dataset, the causal relationship between temperature and energy consumption differs across states, likely due to differences in population density, and commercial/industrial activity. To introduce confounding, we systematically sample days (instances) from states (objects) based on the state's typical energy consumption, including hotter days with higher probability for high consuming states. Specifically, we use importance resampling with a target distribution over Farenheit temperatures $T \sim \mathcal{N}(45 + bias \cdot s_o, 15)$, where $s_{CT} = 3, s_{MA} = 2, s_{ME} = 1, s_{NH} = -1, s_{RI} = -2, s_{VT} = -3$. An example of this sampling with bias $= 9$ is shown in Figure 2 (a-c). Biased sampling in this way introduces a statistical dependency across the dataset (consumption is significantly higher in hotter days), that differs from the causal relationship (low or high temperature causes a moderate increase in energy consumption). This approach of sampling quasi-experimental data to simulate confounding is an emerging standard in causal inference evaluation (Gentzel et al., 2019) although existing benchmarks are not hierarchically structured. Figure 2 (a-d) shows an example of this sampling process for the NEEC benchmark.

Sampling in this way does not provide instance-level counterfactual outcomes. Instead, we estimate the sample-average ground truth counterfactual outcome by fitting a Gaussian process regression model for each state, using treatments and outcomes from the entire calendar year.

Figure 2e shows the models' performances with varying degree of confounding and sample sizes, and Figure 3 shows the estimated and actual effect of temperate on electric energy consumption for two of the six states. Despite the induced confounding, GP-SLC consistently produces accurate estimates of causal effect. The baselines that ignore

Model	CT	MA	ME	NH	RI	VT
GP-SLC	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	1.0
GP-NoConf	13.2	13	31.5	41.6	47.4	14.9
GP-NoObj	19.1	14	26.8	36.2	48	16.5
GP-PerObj	1.6	1.3	5.2	9.7	6.5	**0.7**
MLM1	6.9	5	25.0	5	5.1	**0.7**
MLM2	6.4	4.9	39.4	6.3	9.9	3.8
BART	4.1	2.1	13.3	3.6	3.3	2.4

Table 3: $\sqrt{\text{MSE}}$ for the New England energy consumption benchmark, with bias $= 9°F$ and 25 samples per state. Lower is better. Scores are normalized by GP-SLC's score for the same state.

confounding (GP-NoConf and GP-NoObj) perform poorly as the degree of confounding increases, incorrectly attributing sample-wide association as indicative of causal effect. The linear multi-level models (MLM 1 and MLM 2) are not biased by confounding, but produce poor estimates due to their restrictive parametric assumptions. The remaining two baselines (GP-PerObj and BART) produce more accurate estimates than the other four baselines, but still overfit.

6.4. Limitations

Despite the fact that GP-SLC produces state-of-the-art counterfactual predictions on most of our synthetic and semisynthetic benchmarks, it tends to underestimate the uncertainty in these estimates. In other words, the posterior density on the ground-truth counterfactual is sometimes low, despite the fact that the mean estimate is close to the ground-truth relative to the baselines. We suspect that this is partially attributable to inaccuracies resulting from our approximate inference procedure (Algorithms 2 and 3). Alternative approximate inference schemes, such as using our current approach as a rejuvenation move in a sequential Monte Carlo (SMC) algorithm (Doucet et al., 2001), may resolve these inaccuracies. This kind of SMC-based inference procedure may also help GP-SLC scale to problems with more covariates and objects than we explore in this paper.

Our empirical study focusses on data generating processes that satisfy GP-SLC's implicit semiparametric assumptions; (i) covariates for individuals belonging to the same object are marginally Gaussian distributed, and (ii) exogenous noise is additive and Gaussian. The effect of these modeling assumptions on counterfactual prediction and estimates of effect strength needs additional empirical characterization, ideally via large-scale synthetic experiments (where ground truth is known and robustness to modeling bias can be qualitatively studied).

7. Related Work

Leveraging hierarchical structure is well-established as a technique for adjusting for latent confounding (Gelman, 2006; Gelman & Hill, 2006; Hong & Raudenbush, 2006). Using Gaussian processes for causal inference is also well-established (Alaa & van der Schaar, 2017; 2018; Silva & Gramacy, 2010; Schulam & Saria, 2017; Zhang et al., 2010), as is the use of generative model approaches to adjust for latent confounders given restrictions on structure (Miao et al., 2018; Louizos et al., 2017; Tran & Blei, 2018; Wang & Blei, 2019). To the best of our knowledge, GP-SLC is the first semiparametric generative modeling approach that leverages hierarchical structure to adjust for latent confounders.

GP-SLC is one of many recent techniques (Shalit et al., 2017; Johansson et al., 2016) for estimating individual-level treatment effects. Prior work focusses on the propositional setting under *strong ignorability*, i.e. with no latent confounders. We focus on the hierarchical setting in which latent confounders are shared across multiple instances.

Recent work (Schulam & Saria, 2017) has used Gaussian process models for causal inference in temporal settings, which assumes unconfoundedness and that the outcome is smooth with respect to time and covariates. GP-SLC allows for the existence of object-level latent confounders, and instead assumes that the outcome is smooth with respect to treatment assignment, covariates, and latent confounders. Longitudinal data analysis is closely related to the hierarchical settings we consider in this work: measurements (instances) of individuals (objects) are repeated over a period of time. Extending GP-SLC to the setting where latent confounders are not shared across instances, but instead change over time, is an exciting area of future work.

GP-SLC is most similar to (Alaa & van der Schaar, 2017), in that their approach also uses GP models to estimate individual treatment effects. However, GP-SLC: (i) handles hierarchical latent confounders by first performing inference over object-level latent variables; (ii) accounts for the covariance between noise-free factual and counterfactual outcomes (see Σ_{12} and Σ_{21} in the Supplementary materials); and (iii) uses a Monte Carlo algorithm for inference that yields quantified uncertainty estimates. Their approach could be applied in hierarchical settings by treating the object identifier o as a categorical covariate and using a delta kernel to construct the outcome kernel covariance matrix. This is identical to the GP-PerObj baseline, except that GP-PerObj does not share inferred kernel hyperparameters across objects.

8. Conclusions

This paper presents GP-SLC, a Gaussian process model for causal inference with hierarchically structured latent confounders. In Section 6, we show that, compared to widely used alternatives, GP-SLC produces more accurate estimates of causal effect in realistic sparse observational settings where strong prior knowledge about structure can inform causal estimates. The hierarchical structure we exploit in this paper is one of many kinds of structural background knowledge that could improve causal estimates, and developing techniques to exploit such knowledge is an important area of future work. Extending GP-SLC to handle large observational datasets (Cao, 2018; Quiñonero-Candela & Rasmussen, 2005) or to leverage experimental evidence (Witty et al., 2019) are also exciting areas of future work.

Acknowledgments

References

Alaa, A. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 3424–3432, 2017.

Alaa, A. and van der Schaar, M. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

Berkson, J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

Boomsma, D., Busjahn, A., and Peltonen, L. Classical twin studies and beyond. *Nature Reviews Genetics*, 3(11):872–882, 2002.

Cao, Y. *Scaling Gaussian Processes*. PhD thesis, University of Toronto (Canada), 2018.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., and Mansinghka, V. K. Gen: A general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 221–236. ACM, 2019.

Doucet, A., De Freitas, N., and Gordon, N. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pp. 3–14. Springer, 2001.

Elwert, F. and Winship, C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, 40:31–53, 2014.

Gelman, A. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Gentzel, A., Garant, D., and Jensen, D. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems*, pp. 11717–11727, 2019.

Hastings, K. *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*. Oxford University Press, 1970.

Hill, J. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hong, G. and Raudenbush, S. Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101(475):901–910, 2006.

Hong, G. and Yu, B. Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, 44(2):407, 2008.

Imbens, G. and Rubin, D. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

ISO New England. Energy, load, and demand reports. https://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/zone-info, 2018.

Jensen, D., Burroni, J., and Rattigan, M. Object conditioning for causal inference. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2019.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Lawrence, N. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, pp. 329–336, 2004.

Liang, K.-Y. and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Loftus, G. and Masson, M. Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, 1(4):476–490, 1994.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 08 2018.

Murray, I., Prescott Adams, R., and MacKay, D. J. Elliptical slice sampling. 2010.

Neal, R. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012.

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

Pearl, J. The algorithmization of counterfactuals. *Annals of Mathematics and Artificial Intelligence*, 61(1):29, 2011.

Quiñonero-Candela, J. and Rasmussen, C. E. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

Ramey, C. T., Bryant, D. M., Wasik, B. H., Sparling, J. J., Fendt, K. H., and La Vange, L. M. Infant health and development program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics*, 89(3):454–465, 1992.

Rasmussen, C. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer, 2003.

Schulam, P. and Saria, S. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pp. 1697–1708, 2017.

Shadish, W., Clark, M., and Steiner, P. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484): 1334–1344, 2008.

Shalit, U., Johansson, F., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3076–3085. JMLR.org, 2017.

Silva, R. and Gramacy, R. B. Gaussian process structural equation models with latent variables. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 537–545, 2010.

Titsias, M. and Lawrence, N. Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 844–851, 2010.

Tran, D. and Blei, D. M. Implicit causal models for genome-wide association studies. In *International Conference on Learning Representations*, 2018.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Wang, Y. and Blei, D. M. The blessings of multiple causes. *Journal of the American Statistical Association*, pp. 1–71, 2019.

Witty, S., Lew, A., Jensen, D., and Mansinghka, V. Bayesian causal inference via probabilistic program synthesis. *arXiv preprint*, arXiv:1910.14124, 2019.

Zhang, K., Schölkopf, B., and Janzing, D. Invariant Gaussian process latent variable models and application in causal discovery. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.