Causal Inference using Gaussian Processes with Structured Latent Confounders: Supplementary Materials

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

1. Kernel Functions

In this section, we present a detailed definition of each of the kernel functions used in GP-SLC:

$$k'_{x_k}([\mathbf{u}_{o=Pa(i)}], [\mathbf{u}_{o'=Pa(i')}]) = \sigma_x^2 \exp\left[-\sum_j^{N_U} \frac{(\mathbf{u}_{o,j} - \mathbf{u}_{o',j})^2}{\lambda_{ux_{j,k}}}\right]$$

$$k'_t([\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i], [\mathbf{u}_{o'=Pa(i')}, \mathbf{x}_{i'}]) = \sigma_t^2 \exp\left[-\sum_j^{N_U} \frac{(\mathbf{u}_{o,j} - \mathbf{u}_{o',j})^2}{\lambda_{ut_j}} - \sum_k^{N_X} \frac{(\mathbf{x}_{i,k} - \mathbf{x}_{i',k})^2}{\lambda_{xt_k}}\right]$$

$$k'_y([\mathbf{u}_{o=Pa(i)}, \mathbf{x}_i, \mathbf{t}_i], [\mathbf{u}_{o'=Pa(i')}, \mathbf{x}_{i'}, \mathbf{t}_{i'}]) = \sigma_y^2 \exp\left[-\sum_j^{N_U} \frac{(\mathbf{u}_{o,j} - \mathbf{u}_{o',j})^2}{\lambda_{uy_j}} - \sum_k^{N_X} \frac{(\mathbf{x}_{i,k} - \mathbf{x}_{i',k})^2}{\lambda_{xy_k}} - \frac{(\mathbf{t}_i - \mathbf{t}_{i'})^2}{\lambda_{ty}}\right].$$

where λ_* is a lengthscale hyperparameter and defined for each dimension of corresponding variables. Here, each dimension of $\mathbf{x}$ is generated independently given $\mathbf{u}$, and k'_{x_k} refers to the kernel function for the kth dimension of x. Intuitively, each kernel lengthscale determines the relative strength of influence of each variable's parents in Equation 1. For example, if $\lambda_{ty} >> \lambda_{xy_{i=1\ldots N_X}}$, the covariance between instances (or counterfactuals) with similar treatments will be greater than the covariance between instances with similar covariates.

2. Exact Inference: $Y'_* - Y'$ Details

Here we provide additional details on how to compute GP-SLC's conditional distribution over individual treatment effects. Given the expression for $\left(\begin{bmatrix} Y \\ Y' \\ Y'_* \end{bmatrix} \middle| T_*, T, X, U, \Theta\right)$ in Section 4.2, conditioning on Y yields the following:

$$\left(\begin{bmatrix} Y' \\ Y'_* \end{bmatrix} \middle| T_*, Y, T, X, U, \Theta\right) \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix}\right)$$

where,

$$\mu_1 = K'(W,W)K(W,W)^{-1}Y \qquad \mu_2 = K'(W,W_*)K(W,W)^{-1}Y$$
$$\Sigma_{1,1} = K'(W,W) - K'(W,W)K(W,W)^{-1}K'(W,W) \qquad \Sigma_{1,2} = K'(W,W_*) - K'(W,W)K(W,W)^{-1}K'(W,W_*)$$
$$\Sigma_{2,1} = K'(W_*,W) - K'(W_*,W)K(W,W)^{-1}K'(W,W) \qquad \Sigma_{2,2} = K'(W_*,W_*) - K'(W_*,W)K(W,W)^{-1}K(W,W_*)$$

[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, United States [2]Massachusetts Institute of Technology, Cambridge, United States. Correspondence to: Sam Witty <switty@cs.umass.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

As the difference of variables that are jointly Gaussian is Gaussian, we have that $(Y'_* - Y'|T_*, X, T, Y, U, \Theta) \sim \mathcal{N}(\mu_{ITE}, \Sigma_{ITE})$, where $\mu_{ITE} = \mu_2 - \mu_1$ and $\Sigma_{ITE} = \Sigma_{1,1} - \Sigma_{1,2} - \Sigma_{2,1} + \Sigma_{2,2}$.

3. Asymptotic Posterior Consistency

Here we provide proofs for Proposition 5.1 and Theorems 5.2 and 5.3. The analysis in this section follows the setup presented in (D'Amour, 2019), with the inclusion of shared latent confounding amongst individual instances. We omit covariates X from this analysis and assume that $N_U = 1$ for brevity without loss of generality. Note that these theoretical results also hold for the random intercepts multilevel model (Gelman, 2006).

3.1. Setup

Assuming linear kernels and additive Gaussian exogenous noise, we can equivalently rewrite the GP-SLC model as follows. This equivalent structural causal model is parameterized by latent variables $\alpha, \beta, \tau \in \mathbb{R}$ and $\sigma_U^2, \sigma_T^2, \sigma_Y^2 \in \mathbb{R}^+$. For all $o \in 1, ..., N_O$ and $i \in 1, ..., N_I$, we have that:

$$\begin{aligned} \boldsymbol{\epsilon}_{u_o} &\sim \mathcal{N}(0, \sigma_U^2) \quad & \mathbf{u}_o &= \boldsymbol{\epsilon}_{u_o} \\ \boldsymbol{\epsilon}_{t_i} &\sim \mathcal{N}(0, \sigma_T^2) \quad & \mathbf{t}_i &= \alpha \mathbf{u}_{o=Pa(i)} + \boldsymbol{\epsilon}_{t_i} \\ \boldsymbol{\epsilon}_{y_i} &\sim \mathcal{N}(0, \sigma_Y^2) \quad & \mathbf{y}_i &= \beta \mathbf{t}_i + \tau \mathbf{u}_{o=Pa(i)} + \boldsymbol{\epsilon}_{y_i}. \end{aligned}$$

In this setting, estimating individual treatment effect reduces to estimating β, as $\mathbf{y}_{i,\mathbf{t}_*} - \mathbf{y}_i = \beta(\mathbf{t}_* - \mathbf{t}_i)$. We make the following observations.

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

For a detailed proof of Proposition 5.1, see Proposition 1 in (D'Amour, 2019). In summary, they show that given any set of latent parameters $\Theta = (\alpha, \beta, \tau, \sigma_U^2, \sigma_T^2, \sigma_Y^2)$, there exists an alternative set of parameters Θ' such that $P(T, Y|\Theta) = P(T, Y|\Theta')$ and $\beta \neq \beta'$. In other words, the structural causal model forms a linear system of equations that is rank-deficient. The set of parameters that satisfy this condition construct an *ignorance region*.

Extending their results to the Bayesian setting, we have that for any two sets of parameters Θ and Θ' on the same ignorance region, the posterior odds ratio reduces to the prior odds ratio, $\frac{P(\Theta|T,Y)}{P(\Theta'|T,Y)} = \frac{P(\Theta)P(T,Y|\Theta)}{P(\Theta')P(T,Y|\Theta')} = \frac{P(\Theta)}{P(\Theta')}$. By definition, Θ is not asymptotically consistent, as the posterior $P(\Theta|T, Y)$ depends on the prior $P(\Theta)$. The problem of *asymptotic consistency* can be mitigated when $N_O < N_I$.

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 $\infty, \forall t_* \in \mathbb{R}$.*

Proof. For all $i' \in I'$, we have that $\mathbf{y}_{i'} = \beta \mathbf{t}_{i'} + C + \boldsymbol{\epsilon}_{y_{i'}}$ for some constant $C \in \mathbb{R}$. Therefore, the covariance between T and Y in I' is uniquely given by β, i.e. $cov(\mathbf{t}_{i' \in I'}, \mathbf{y}_{i' \in I'}) = \beta$. Estimating the covariance of a bivariate normal has a unique maximum likelihood solution. Therefore, by the Bernstein-von Mises Theorem (Doob, 1949) we have that the posterior over β, and thus ITE_{t_*}, is asymptotically consistent as n approach ∞. □

Theorem 5.3 *Assume there exists 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 ∞.*

Proof. For all $o \in \mathbb{O}$, $j \in \{1, ..., k_o\}$ let $\mathbf{t}'_{i'_{o,j}} = \mathbf{t}_{i'_{o,j}} - \bar{\mathbf{t}}_o$ and $\mathbf{y}'_{i'_{o,j}} = \mathbf{y}_{i'_{o,j}} - \bar{\mathbf{y}}_o$, where $\bar{\mathbf{t}}_o = \sum_j \mathbf{t}_{i'_{o,j}}/k_o$ and $\bar{\mathbf{y}}_o = \sum_j \mathbf{y}_{i'_{o,j}}/k_o$, i.e., the sample average over all instances that share a parent object. Therefore, $\mathbf{t}'_{i'_{o,j}} = \alpha \mathbf{u}_o + \boldsymbol{\epsilon}_{t_{i'_{o,j}}} - \sum_j (\alpha \mathbf{u}_o + \boldsymbol{\epsilon}_{t_{i'_{o,j}}})/k_o = \boldsymbol{\epsilon}_{t_{i'_{o,j}}} - \sum_j \boldsymbol{\epsilon}_{t_{i'_{o,j}}}/k_o$ and $\mathbf{y}'_{i'_{o,j}} = \beta(\alpha \mathbf{u}_o + \boldsymbol{\epsilon}_{t_{i'_{o,j}}}) + \tau \mathbf{u}_o + \boldsymbol{\epsilon}_{y_{i'_{o,j}}} - \sum_j (\beta(\alpha \mathbf{u}_o + \boldsymbol{\epsilon}_{t_{i'_{o,j}}}) + \tau \mathbf{u}_o + \boldsymbol{\epsilon}_{y_{i'_{o,j}}})/k_o = \beta \mathbf{t}'_{i'_{o,j}} + \boldsymbol{\epsilon}_{y_{i'_{o,j}}} - \sum_i \boldsymbol{\epsilon}_{y_{i'_{o,j}}}/k_o$. As $\boldsymbol{\epsilon}_{y_{i'_{o,j}}}$ is independent of $\mathbf{t}'_{i'_{o,j}}$, we have that the covariance between $\mathbf{t}'_{i'_{o,j}}$ and $\mathbf{y}'_{i'_{o,j}}$ is equal to β. Therefore, the problem of estimating β reduces to estimating the covariance of a bivariate normal distribution, $P(T', Y')$, which has a unique maximum likelihood solution. As in the proof of Theorem 5.2, by the Bernstein-von Mises Theorem (Doob, 1949) we have that the estimate of β, and thus ITE_{t_*}, is asymptotically consistent as n approach ∞. □

4. Bayesian Linear Multilevel Model Baseline

One of the baselines we use in the experiments is Bayesian linear multilevel models (Gelman, 2006). We implement two multilevel models, which introduce varying degrees of shared parameters across objects. The first multilevel model, also known as a random slope and intercepts model, (MLM 1) fits the observations using the following structural equations.

$$\begin{aligned}
\sigma_y^2 &\sim \gamma^{-1}(\alpha_{\sigma_y}, \beta_{\sigma_y}) \\
\alpha &\sim \mathcal{N}(\mu_\alpha, \Sigma_\alpha) \\
\beta_o &\sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \text{ for } o = 1 \dots N_O \\
\eta_o &\sim \mathcal{N}(\mu_\eta, \sigma_\eta^2) \text{ for } o = 1 \dots N_O \\
\mathbf{y}_i &\sim \mathcal{N}(\beta_{o=Pa(i)}\mathbf{t}_i + \alpha^T \mathbf{x}_i + \eta_{o=Pa(i)}, \sigma_y^2)
\end{aligned}$$

This model allows varying intercepts η and treatment effect β across objects while assuming α is held constant across objects.

The second multilevel model, also known as the random intercepts model, (MLM 2) fits the observations using the following structural equations.

$$\begin{aligned}
\sigma_y^2 &\sim \gamma^{-1}(\alpha_{\sigma_y}, \beta_{\sigma_y}) \\
\alpha &\sim \mathcal{N}(\mu_\alpha, \Sigma_\alpha) \\
\beta &\sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \\
\eta_o &\sim \mathcal{N}(\mu_\eta, \sigma_\eta^2) \text{ for } o = 1 \dots N_O \\
\mathbf{y}_i &\sim \mathcal{N}(\beta\mathbf{t}_i + \alpha^T \mathbf{x}_i + \eta_{o=Pa(i)}, \sigma_y^2)
\end{aligned}$$

This model allows varying intercepts η across objects while assuming α and β are held constant across objects. We implement both models in Gen (Cusumano-Towner et al., 2019). For both models, we use $\alpha_{\sigma_y} = 4.0, \beta_{\sigma_y} = 4.0, \mu_{(\cdot)} = 0, \sigma_\alpha^2 = 3.0, \sigma_\beta^2 = 1.0$, and $\sigma_\eta^2 = 10.0$ as priors.

5. Synthetic Experiments

We examine the finite-sample behavior of the GP-SLC model using two synthetic datasets that match GP-SLC's assumptions about the existence of *object-level* latent confounders (U) that simultaneously influence *instance-level* observed treatments (T), covariates (X), and outcomes (Y). The following structural equations summarize the data generating process:

$$\begin{aligned}
W_j &\sim \mathcal{N}(0, 1I_3) \text{ for } j = 1, 2, 3 \\
\mathbf{u}_o &\sim \mathcal{N}(0, 0.5I_3) \text{ for } o = 1 \dots N_O \\
\mathbf{x}_i &= W \cdot \mathbf{u}_{o=pa(i)} + \boldsymbol{\epsilon}_{x_i} \text{ where } \boldsymbol{\epsilon}_{x_i} \sim \mathcal{N}(0, 0.5I_3) \text{ for } i = 1...N_I \\
\mathbf{t}_i &= g_t(\mathbf{x}_i, \mathbf{u}_{o=pa(i)}) + \boldsymbol{\epsilon}_{t_i} \text{ where } \boldsymbol{\epsilon}_{t_i} \sim \mathcal{N}(0, 0.5) \text{ for } i = 1...N_I \\
\mathbf{y}_i &= g_y(\mathbf{t}_i, \mathbf{x}_i, \mathbf{u}_{o=pa(i)}) + \boldsymbol{\epsilon}_{y_i} \text{ where } \boldsymbol{\epsilon}_{y_i} \sim \mathcal{N}(0, 0.5) \text{ for } i = 1...N_I
\end{aligned}$$

First, we draw $\mathbf{u}$ from a multivariate Gaussian distribution. Then, we generate covariates $\mathbf{x}$ as linear combinations of $\mathbf{u}$ with additive exogenous noise. We generate treatments $\mathbf{t}$ as a function (g_t) of $\mathbf{x}$ and $\mathbf{u}$ with additive noise. Finally, we generate outcome $\mathbf{y}$ as a function (g_y) of $\mathbf{x}$, $\mathbf{t}$, and $\mathbf{u}$ with additive noise. For multi-dimensional variables, $\mathbf{x}$ and $\mathbf{u}$, we first apply the nonlinear function to each dimension of $\mathbf{x}$ and $\mathbf{u}$, then we aggregate them by summing across dimensions.

The nonlinear treatment and outcome functions are shown in Table 1.

Dataset	$g_t(\mathbf{x}, \mathbf{u})$	$g_y(\mathbf{t}, \mathbf{x}, \mathbf{u})$
Additive	$\sum_j \mathbf{x}_{*,j} \sin(\mathbf{x}_{*,j}) - \sum_j \mathbf{u}_{*,j} \sin(\mathbf{u}_{*,j})$	$\mathbf{t}\sin(2\mathbf{t}) + \sum_j \mathbf{x}_{*,j} \sin(\mathbf{x}_{*,j}) + 3\sum_j \mathbf{u}_{*,j} \sin(\mathbf{u}_{*,j})$
Multiplicative	$\frac{1}{10}(\sum_j \mathbf{x}_{*,j} \sin(\mathbf{x}_{*,j}))(\sum_j \mathbf{u}_{*,j} \sin(\mathbf{u}_{*,j}))$	$\frac{1}{10}(\mathbf{t}\sin(2\mathbf{t}))(\sum_j \mathbf{x}_{*,j} \sin(\mathbf{x}_{*,j}))(\sum_j \mathbf{u}_{*,j} \sin(\mathbf{u}_{*,j}))$

Table 1. The functional form of T and Y for 2 synthetic datasets with continuous treatments and nonlinear outcome functions.

References

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.

D'Amour, A. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In The 22nd International Conference on Artificial Intelligence and Statistics, pp. 3478–3486, 2019.

Doob, J. L. Application of the theory of martingales. Le calcul des probabilites et ses applications, pp. 23–27, 1949.

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