A Multi-Task Gaussian Process Model for Inferring Time-Varying Treatment Effects in Panel Data

Yehu Chen **Annamaria Prati** **Jacob Montgomery** **Roman Garnett**
Washington University in St. Louis

Abstract

We introduce a Bayesian multi-task Gaussian process model for estimating treatment effects from panel data, where an intervention outside the observer's control influences a subset of the observed units. Our model encodes structured temporal dynamics both within and across the treatment and control groups and incorporates a flexible prior for the evolution of treatment effects over time. These innovations aid in inferring posteriors for dynamic treatment effects that encode our uncertainty about the likely trajectories of units in the absence of treatment. We also discuss the asymptotic properties of the joint posterior over counterfactual outcomes and treatment effects, which exhibits intuitive behavior in the large-sample limit. In experiments on both synthetic and real data, our approach performs no worse than existing methods and significantly better when standard assumptions are violated.

1 INTRODUCTION

Across the natural, health, and social sciences, researchers frequently wish to estimate the effects of interventions on downstream outcomes based on "natural experiments." In the typical case, outcomes are collected from multiple units over time in a panel structure. At some fixed point, a subset of units receives a "treatment," whereas others do not; in a natural experiment, this intervention is assumed to be outside the observer's control. The inferential goal is to estimate the effect of the intervention on the outcomes of the treated individuals from observations of the system both pre- and post-treatment. Such designs are crucial to areas of research as disparate as economics, political science, epidemiology, biology, agriculture, and medicine (Schlenker and Roberts, 2006; Angrist and Pischke, 2008; Abadie et al., 2010; Schulam and Saria, 2016). Although direct intervention and treatment randomization would be ideal, this may not always be possible due to ethical issues, undue cost, or other factors. Analyzing natural experiments provides an alternative that can offer insight in spite of this limitation.

A common inferential framework in this setting is the *potential outcomes* model (Rubin, 2005). Here, the treatment effect is modeled as the difference between the *observed* outcomes for a given treatment status and the *counterfactual* outcomes that would have arisen for those same units under a different status. Since counterfactual outcomes are never observed, treatment effects cannot be directly computed. Instead, counterfactual values must be inferred from the observed outcomes combined with reasonable assumptions about the data generating process (DGP).

Creating accurate estimates of treatment effects in this setting presents multiple challenges. First, since the intervention is typically not randomized, the treatment and control groups differ at baseline in both observed and unobserved factors. Second, unobserved time-varying confounders may affect some or all units. Finally, the treatment effect itself may be non-constant over multiple post-treatment periods. To address these issues, we introduce MGP-PANEL, a flexible Bayesian model for inferring time-varying treatment effects from panel data.

Our model has two critical innovations. First, we propose a hierarchical multi-task Gaussian process prior that encodes (1) smooth trends for the treatment and control groups that may be correlated, and (2) smooth unit-level deviations from group trends. Importantly, we do *not* require that group trends will move in lock step in the absence of treatment. Instead, we make the more realistic assumption that these trends may be correlated. When these correlations are imperfect, it results in more uncertainty about the treatment effect that propagates naturally into posterior the distribution. Our Bayesian framework also allows us to derive a full posterior belief over the correlation between the control and treatment group, giving us further insight into the system.

Second, we place a flexible nonparametric prior on the temporal evolution of treatment effects, which provides a com-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

promise between the more rigid parametric models (Xu et al., 2016; Soleimani et al., 2017) or the completely agnostic priors (Arbour et al., 2021) in previous approaches to this problem in the Bayesian literature. This is important because in most settings we would not expect the causal effect of an intervention to swing wildly from time period to time period, but also would not always have a strong prior belief about the structure of individual-level treatment effects to design an insightful parametric representation.

These two innovations allow us to infer the counterfactual outcomes in the treated group, thus treatment effects, from pre- and post-treatment observations both accurately and with calibrated uncertainty. We may also represent the complete model as a single GP with a modest number of hyperparameters, facilitating efficient computation using off-the-shelf software. In all, our model offers a flexible Bayesian approach to inferring dynamic treatment effects that relaxes some standard assumptions while providing superior finite-sample performance in terms of coverage and accuracy.

2 PROBLEM SETUP

Consider panel data of N units that are repeatedly observed over time range 0 to T. We focus on the classic setting of a binary treatment that partitions the units into a *control* group and a *treatment* group. We can use g to index each group, where $g = 1$ for units that were treated and 0 otherwise. We assume that the treatment is applied to the treatment group at the same time $1 < T_0 < T$, so the treatment assignments for each unit–time pair is $D_i(t) = 1$ if $g = 1$ (treatment group) and $t > T_0$ (post-treatment), and $D_i(t) = 0$ otherwise.

We follow the standard stable unit treatment value assumption (SUTVA) (Rubin, 1980), which requires that units are unaffected by treatment assignments for other units. Hence, there are only two potential outcomes: the outcome under treatment $Y^{(1)}$, and the outcome under control $Y^{(0)}$. Each unit is also associated with a vector of observed covariates, x_i. The quantity of interest is the *average treatment effect on the treated* (ATT), which is the expected difference between treated and untreated outcomes in the treated group:

$$\delta(t) = \mathbb{E}\big[Y_i^{(1)}(t) - Y_i^{(0)}(t) \mid g_i = 1\big]; \quad t > T_0. \tag{1}$$

Note that $\delta(t)$ may vary as a function of time.

The fundamental problem is that we observe either $Y^{(1)}$ or $Y^{(0)}$ but not both simultaneously, so $\delta(t)$ cannot be inferred without additional assumptions. Standard assumptions include unconfoundedness (no unobserved confounders), overlap (non-zero chance for any treatment assignment) (Rosenbaum and Rubin, 1983; Imbens, 2004) and no anticipation Athey and Imbens (2022) (no treatment effect prior intervention). With these assumptions, model-based approaches to estimating (1) generally incorporate time and covariates into a regression model of the form:

$$Y_i(x_i, t) = h(x_i) + \gamma_g(t) + u_i(t) + \delta(t) \cdot D_i(t) + \varepsilon_{it}. \tag{2}$$

Here $Y_i(x_i, t)$ is the observed outcome, $h(x_i)$ reflects the mapping between observed covariates and baseline outcomes, $\gamma_g(t)$ is the latent trend for group g in which unit i belongs to in the absence of treatment, $u_i(t)$ is the unit-level deviation from the trend and $\delta(t) \cdot D_i(t)$ represents the ATT. Finally, ε_{it} represents exogenous errors.

The general regression model in (2) thus effectively decomposes panel observations into five parts: a covariate-effect component that maps from observed covariates to the outcomes, a group-time component modeling trends for the groups unrelated to treatment, an individual-time component modeling deviations from group trends for specific units, a treatment-effect process, and noise. We will adopt this general model construction, and will design carefully chosen priors for each component. We will contrast our approach with alternatives in the literature in Section 4.

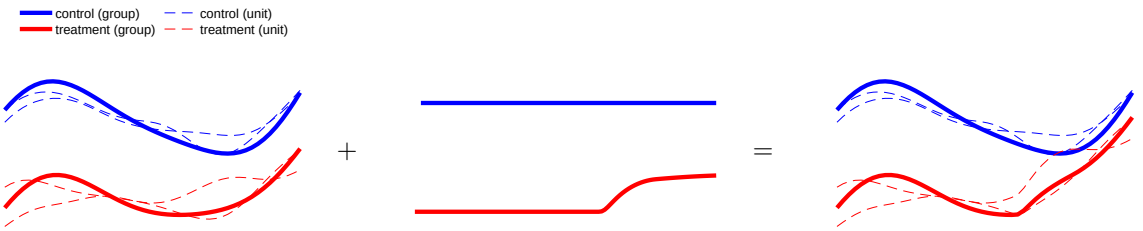
3 MGP-PANEL

A Gaussian process (GP) is an infinite-dimensional analog of the multivariate normal distribution appropriate for modeling functions with structured correlation (Rasmussen and Williams, 2006). Let $\mathcal{X}$ be an arbitrary domain, and let $f\colon \mathcal{X} \to \mathbb{R}$ be a function to be inferred on $\mathcal{X}$. A GP on f is specified by its first two moments, a mean function $\mu(x)$ and a covariance function $K(x, x')$.The process is then defined by its finite-dimensional distributions, which are Gaussian with moments derived by pointwise evaluation of the mean and covariance functions. By construction, a GP enjoys properties such as closure under addition and affine transformation. Moreover, a GP on f allows for exact inference from noisy observations corrupted by additive Gaussian noise; the posterior is a GP with moments that can be computed in closed form (Rasmussen and Williams, 2006).

A multi-task GP (Bonilla et al., 2008) is an extension of a GP suitable for modeling multiple jointly correlated functions. The model is specified by *shared* mean and covariance functions over the values of multiple functions, which, when correlations across the functions are nontrivial, allows us to perform inference of one function from observations of another. This will be a key component of our model.

3.1 Multi-task GP Model for Panel Data

We now outline MGP-PANEL, a multi-task GP model for panel data that encodes correlated group trends, unit-level deviations from these trends, and dynamic treatment effects. The key components of MGP-PANEL are illustrated at a high level in Figure 1. First, we model outcomes in treatment and control groups as having *correlated* (but not necessarily perfectly equal) group trends in the absence of treatment



(a) Correlated group trends. (b) Treatment effect trends. (c) Observed trends.

Figure 1: Illustration of key components in MGP-PANEL. In the left panel, the bold solid curves represent the coupled group trends shared across groups and the dashed curves represent unit variations within groups. The middle panel shows the smooth, time-varying effects of the intervention on the treated units (note we assume no effect on the controls). In the right panel, the bold solid curves represent the observed group trends after absorbing the actual effects.

(left, solid); within each group, unit trajectories are then modeled as potentially having smooth deviations from these trends (left, dashed). We then model the *average* treatment effects associated with our measurements to be identically zero in the control group and in the treatment group pre-treatment, then to dynamically evolve in the treatment group post-treatment (middle). The expectation of the observed data is then the sum of the underlying trends and the average treatment effect (right).

To realize this model practically, we adopt the general regression model in (2) and will construct structured GP priors for the covariate-effect process h, the group trends γ, the unit-level deviations u, and the treatment-effect process δ. We will not comment extensively on the covariate-effect process, as established techniques can be used there (Hill, 2011; Alaa and van der Schaar, 2017; Yoon et al., 2018). We will simply assume that an appropriate GP prior

$$h(x) \sim \mathcal{GP}\big(\mu_x(x), K_x(x, x')\big) \tag{3}$$

may be chosen and move on to the temporal components of MGP-PANEL, where its novelty primarily lies.

Group-level process. Our modeling of group trends is inspired by the *parallel trends assumption* (PTA) foundational to difference-in-differences estimators, which assumes that the treatment and control groups share *identical* temporal evolution up to a constant additive shift (see Section 4 for details). In the interest of flexibility and relaxing this potentially strong assumption, we instead assume the presence of *nontrivial* (but not necessarily perfect) correlation between control and treatment trends. Let $\gamma_0(t)$ denote the latent trend shared by units in the control group, and let $\gamma_1(t)$ denote the corresponding trend for units in the treatment group. We model these trends *jointly* with a multi-task GP:

$$[\gamma_0, \gamma_1] \sim \mathcal{GP}(\mu_\gamma, K_\gamma). \tag{4}$$

We take the mean function μ_γ to be constant for each group (but not necessarily equal across groups). In our experiments, we had success using a simple *separable* covariance to model further structure in the trends:

$$K_\gamma\big([t, g], [t', g']\big) = K_{\text{time}}(t, t') \cdot K_{\text{task}}(g, g'), \tag{5}$$

where the temporal component K_{time} was a squared-exponential (SE) kernel and the group component K_{task} was a 2×2 matrix with unit diagonal and value $\rho \in [-1, 1]$ off diagonal encoding the degree of cross-group correlation controlling how "parallel" the coupled group trends are.

Unit-level deviations. We also allow smooth unit-level deviations from the group trends as a potential part of the data generating process. These unit-level deviations, $\{u_i(t)\}$, are i.i.d. and share a centered GP prior:

$$u_i(t) \sim \mathcal{GP}(0, K_u). \tag{6}$$

In our studies, we again chose the SE kernel for K_u.

Treatment-effect process. Treatment effects often exhibit temporal correlations. However, existing models usually either ignore any correlations (Arbour et al., 2021) or address them with *ad hoc* parametric models (Xu et al., 2016; Soleimani et al., 2017). In MGP-PANEL we model the temporal evolution of ATT with a separate GP prior:

$$\delta(t) \sim \mathcal{GP}(0, K_\delta). \tag{7}$$

In general, this prior can be constructed to flexibly reflect any prior beliefs, including those used in previous work. However, we outline one possible construction that may be of interest in a variety of applications: a process that is identically zero prior to treatment, then smoothly deviates from zero with increasing scale post-treatment. We may construct such a process by scaling an appropriate stationary covariance function K (such as the SE kernel) to realize the desired shape. For instance, to model a gradually increase from zero at intervention time T_0 to a full unit scale at some later time T_1, we could define a scaling function $s(t)$ by

$s(t) = 0$ for $t < T_0$, $s(t) = 1$ for $t > T_1$, and

$$s(t) = \left[1 + \exp\left(\tfrac{T_1 - T_0}{t - T_0} - \tfrac{T_1 - T_0}{T_1 - t}\right)\right]^{-1} \quad (8)$$

otherwise. This worked well in our experiments, but we emphasize again that other choices are available.

Full model. Our full model is then the sum of the individual components in (3), (4), (6) and (7), which are assumed to be independent, coupled with i.i.d. additive Gaussian noise. As the model is simply a sum of independent GPs, the induced observation model is $y \sim \mathcal{GP}(\mu_y, K_y)$, where

$$\mu_y = \mu_x + \mu_\gamma; \quad K_y = K_x + K_\gamma + K_u + K_\delta + \sigma^2_{\text{noise}}\mathbb{I}. \quad (9)$$

In this work, we embrace the idea of fully Bayesian inference, which addresses model uncertainty by marginalizing over the hyperparameters of our mean and covariance functions.[1] Our inference framework is similar to the one in Xu et al. (2016), but tailored for our setting with coupled group trends. In addition, our framework reduces to an alternative framework in (Arbour et al., 2021) where the GP prior on ATT is assumed to be infinitely wide and hence incorporates no temporal correlation. We note that efficient scaling of GP inference is an issue that has received an enormous amount of attention, but can be resolved from prudent use of approximate GP inference such as inducing point methods (Titsias, 2009) and variational inference (Hensman et al., 2015).

We stress that there is a great deal of flexibility in specifying the components of our model and that the choices we made here are not canonical. In the Supplement we show that our model performs well even when these components are misspecified, including under violations in the noise model, the degree of smoothness in temporal trends, and with completely uncorrelated group trends. Further, we show that in such circumstances standard methods for Bayesian model selection facilitate diagnosing these issues and adjusting the model accordingly with standard tools from the GP literature (e.g., alternative kernels, non-normal error, etc.). This stands in stark contrast to the performance of existing methods, which fail almost completely when their assumptions (such as PTA) are violated (see Table 2).

4 DISCUSSION AND RELATED WORK

The presentation in Equation (2) is a general statement for the problem and highlights the key challenge. To succeed, we must not only account for variation in the observed responses $\{Y_i(t)\}$, but also deconstruct that variation to isolate the ATT δ from the covariate-effect process h, the group-level trends γ, unit deviations u, and noise.

A variety of machine learning methods (Hill, 2011; Johansson et al., 2016; Alaa et al., 2017; Atan et al., 2018; Yoon et al., 2018; Wager and Athey, 2018; Chen et al., 2019; Künzel et al., 2019; Curth and van der Schaar, 2021) have been applied to the covariate-effect component, including multi-task GP models, to estimate causal effects in cross-sectional data (Flaxman et al., 2015; Alaa and van der Schaar, 2017, 2018; Aglietti et al., 2020; Witty et al., 2020). However, the greater challenge in this setting is disentangling the temporal group-level and unit-level trends from the treatment effects. At best, assuming an accurate covariate model, the data itself is still constructed from three additive components plus exogenous errors, $Y_i(t) = \gamma_g(t) + u_i(t) + \delta(t) \cdot D_i(t) + \varepsilon_{it}$. Attributing variation in $\{Y_i(t)\}$s to any one component necessarily requires further assumptions.

Common solutions leverage assumptions about the DGP to infer counterfactuals for trends in treated units from data from the pre-treatment period for both groups and data from the post-treatment period in the control group. Examples include the family of *synthetic control* (SC) models (Abadie and Gardeazabal, 2003; Abadie et al., 2015; Ben-Michael et al., 2021; Arkhangelsky et al., 2021; Chernozhukov et al., 2021), matching procedures (Diamond and Sekhon, 2013; Imai et al., 2021), matrix completion method (Athey et al., 2021) and variations on two-way or interactive fixed effects models (Bai, 2009; Xu, 2017; De Chaisemartin and d'Haultfoeuille, 2020; Imai and Kim, 2021).

There are two notable limitations of this line of work. First, these approaches largely assume little or (more typically) no temporal structure. This includes unit-level deviations, group-level trends, and the treatment effects. For instance, the two-way fixed effects (2FE) model includes indicator variables for each unit and time period, implicitly encoding the assumption that temporal shocks are implausibly *independent* from time period to time period. Likewise, these models typically assume that the treatment effect evolves with absolutely no implied structure. This is justified as an extension of the popular *difference-in-differences* (DID) method (Card and Krueger, 1994; Angrist and Pischke, 2008) originally designed for a two-period panel, but is harder to rationalize when observations are densely sampled over many time periods. Pang et al. (2021) proposed a Bayesian dynamic multilevel latent factor model that at least incorporates an autoregressive component over temporal trends; however, the order of autoregression is often taken to be relatively minor. Meanwhile, other related work (Athey and Imbens, 2022) assumes that treatment effects are *constant* over time, perhaps encoding too *much* stability.

Second, these models typically invoke strong assumptions about counterfactual trends among treated units. The most extreme is the *parallel trends assumption* (PTA), which specifies that the unmeasured temporal shocks are *identical* in expectation for the treatment and control groups in *every* time period. Notably, assuming zero-mean noise, the PTA

[1] A detailed description of our inference procedure and the hyperpriors used in our simulation and case studies can be found in the Supplement.

implies that, for any time periods t and t':

$$\mathbb{E}\big[(\gamma_g(t') + u_i(t')) - (\gamma_g(t) + u_i(t)) \mid D_i(t) = 1\big] = \mathbb{E}\big[(\gamma_g(t') + u_i(t')) - (\gamma_g(t) + u_i(t)) \mid D_i(t) = 0\big]. \quad (10)$$

Although the PTA is defended as it allows nonparametric identification, the assumption is strong, relies on limiting properties rarely relevant with modest numbers of units (see our example below), and is difficult to justify in most applied settings.

A less strict model-based approach is the interactive fixed effects (IFE) model, which allows for the heterogeneous impact of common temporal shocks on cross sections of the population via a low-rank product of latent fixed effects (Bai, 2009; Xu, 2017; Pang et al., 2021). The idea is to use the pre-treatment period to identify units that respond to shocks similarly, estimating causal effects by assuming this latent structure holds for all time periods. SC models rely on similar strategy of using pre-treatment periods to create a synthetic counter-factual where the PTA holds, which Xu (2017) connects directly back to IFE.

In contrast, a family of approaches built on Bayesian nonparametric models (primarily GP) have been proposed to better encode structure in temporal trends and treatment effects (Shi et al., 2012; Xu et al., 2016; Soleimani et al., 2017; Moraffah et al., 2021; Arbour et al., 2021). Particularly relevant is Shi et al. (2012), which proposes a mixed-effects GP functional regression (ME-GPFR) model to predict dose-response curves. Similarly, Xu et al. (2016) estimated individual treatment-response curves by independently modeling each unit's trend. These assume that these trends are independent across units, failing to account for commonalities *within* groups. Some assume trends to be the same across groups (Shi et al., 2012; Alaa and van der Schaar, 2017) while others do not model group trends at all (Xu et al., 2016; Soleimani et al., 2017).

Our approach tackles the basic problem in a novel way by leveraging the reasonable assumption that the contaminating processes $u_i(t)$ and $\gamma_g(t)$ are smooth over time, in a manner that can be captured by a GP.[2] Further, *a priori* we allow the group trends to be correlated with each other (or not) to varying degrees. As we detail in the next section and the Supplement, these two assumptions together allow us to generate posterior beliefs about the treatment effect that reflect our remaining uncertainty about group-level and unit-level confounding. The joint posterior over counterfactual outcomes and treatment effects also exhibits intuitive behavior in the large-sample limit.

In essence, when group-level trends are tightly linked and unit-level deviations are small, the posterior for $\delta(t)$ will be precise. However, when either or both of these conditions fail, the posterior will correctly reflect this uncertainty. Thus, the idea is that it is often not possible to remove the contaminating influence of $\gamma_g(t)$ and $u_i(t)$ without making inappropriately strong assumptions (e.g., the PTA). Still, we can use data from the pre-treatment period and the post-treatment control group to make reasonable predictions for both. Our remaining uncertainty about these trends is then reflected in the posterior for $\delta(t)$.

The work most closely related to our own is Arbour et al. (2021), which provides a multi-task GP model for individualized time trends with a multi-task GP. Arbour et al. (2021), however, rely on an intrinsic coregionalization model, which is conceptually an analogue to the IFE approach, assuming no prior knowledge on the treatment effects (such as smoothness and monotonicity) and estimating the average treatment effect with a trivial estimator that uses point estimates of γ and h, potentially ignoring a great deal of uncertainty. Further, they develop a model in the context where a single unit is treated, as is common in the SC approach (Brodersen et al., 2015). Our approach may be seen as more appropriate for a canonical setting where a *group* of units is treated and we wish to encode *a priori* intuition about the shape of the treatment effects over time.

5 POSTERIOR ANALYSIS

One major question regarding our modeling scheme is how the posterior distribution over the treatment effect and counterfactual trends in the treatment group evolves with repeated observations of the system, and in particular under what conditions we may recover the treatment effect in the large-sample limit. We conduct a thorough analysis of the asymptotic behavior of the posterior distribution of our model in the Supplement through the lens of a simplified model whose behavior nonetheless gives considerable insight. As conditioning on additional observations never increases uncertainty of GP posterior, we conduct this analysis by looking at one post-treatment period at a time, assuming that ρ has been inferred from pre-treatment observations. Specifically, we derive the posterior correlation between the post-treatment treated counterfactual outcome and treatment effect, as well as upper bounds for the posterior variance of both quantities and their sum.

To summarize, with repeated pre- and post-treatment observations: (1) uncertainty in the control trend collapses both pre- and post-treatment, (2) uncertainty in the treatment trend collapses pre-treatment, (3) when the control and treatment trends are *perfectly* correlated ($\rho = 1$, analogous to the PTA) uncertainty in the treatment effect and counterfactual trend post-treatment collapses, and (4) when the treatment and control group trends are *not* perfectly correlated, uncertainty in these components falls asymptotically to some limiting (but nonzero) value decreasing in $|\rho|$.

The nature of the posterior uncertainty is largely consistent with the properties of classical estimators, which also rely

[2] See Hainmueller et al. (2014, p. 148) for a justification of smoothness assumptions in the social sciences.

Table 1: Averaged performance scores of MGP-PANEL compared to baseline estimators in the setting of perfectly correlated group trends and time-invariant unit variations. Standard errors are shown in parentheses to the precision of last significant digits. Bold numbers indicate results that are significantly better in paired t-tests and italic numbers indicate results that are not significantly worse than the bold numbers at $\alpha = 0.05$ level. MGP-PANEL does significantly better than all other estimators in RMSE and LL but slightly worse in coverage.

	model						
measure	ours	GSC	2FE	CMGP	DM-LFM	ICM	LTR
RMSE	**0.0027**(3)	0.0042(1)	0.0041(1)	0.0090(7)	0.0051(2)	0.0051(6)	0.0720(124)
coverage	*0.924*(33)	0.898(14)	*0.916*(13)	0.700(53)	**0.938**(12)	0.776(34)	0.310(69)
LL	**4.50**(12)	4.00(4)	4.03(4)	2.79(20)	3.96(4)	3.34(18)	−31.5(12.9)

Table 2: Averaged performance scores of MGP-PANEL compared to baseline estimators under different correlation values. In the case of imperfectly correlated group trends and time-variant unit variations, MGP-PANEL does significantly better than all other estimators in paired t-tests at $\alpha = 0.05$ level.

	ours	GSC	2FE	CMGP	DM-LFM	ICM	LTR
ρ	RMSE						
0.1	**0.0242**(27)	0.1640(184)	0.1109(98)	0.0945(99)	0.0993(97)	0.0853(67)	0.0566(78)
0.5	**0.0229**(26)	0.1220(131)	0.0787(69)	0.0738(84)	0.0968(98)	0.0810(72)	0.0547(73)
0.9	**0.0171**(19)	0.0567(59)	0.0342(31)	0.0360(40)	0.0554(53)	0.0470(34)	0.0674(114)
ρ	coverage						
0.1	**0.802**(61)	0.332(52)	0.140(41)	0.216(38)	0.570(61)	0.444(66)	0.358(63)
0.5	**0.816**(55)	0.306(50)	0.156(38)	0.222(45)	0.488(66)	0.404(68)	0.358(69)
0.9	**0.802**(59)	0.354(49)	0.320(43)	0.284(51)	0.402(58)	0.420(58)	0.226(48)
ρ	LL						
0.1	**2.19**(19)	−22.4(11.8)	−206(20)	−44.9(9.8)	−3.23(1.83)	−2.07(91)	−41.3(23.3)
0.5	**2.23**(20)	−14.1(5.1)	−53.4(10.9)	−35.1(6.4)	−4.22(1.63)	−4.00(2.01)	−20.4(6.7)
0.9	**2.55**(19)	−6.03(2.00)	−8.28(2.29)	−13.4(3.5)	−3.17(1.00)	−1.62(0.77)	−24.5(6.5)

on the parallel trends assumption to ensure identification of the treatment effect. However, we regard the behavior in (4) to be a useful *feature* of our modeling approach. In a Bayesian analysis, one can derive the posterior distribution of the coupling parameter ρ to question whether perfect correlation is indeed a plausible assumption given the observed data (see our case study), rather than insisting on it even when implausible. The partial identification and unavoidable residual uncertainty resulting in the $|\rho| < 1$ case is simply a natural consequence of a model that can relax and challenge the assumption of perfectly correlated trends.

6 EXPERIMENTS

In this section, we provide a simulation study to compare MGP-PANEL with several existing estimators and move on to describe a real-world case study. We also complete an ablation study evaluating the effectiveness of each novel aspect of our model and an additional simulation study evaluating the robustness of MGP-PANEL to model misspecification (see Supplement for details).

6.1 Simulation Study

We consider two synthetic settings in our simulations. In the first, the group trends are perfectly correlated and the unit-level deviations are time-invariant, rendering two-way fixed effects the "correct" model. In the second setting, the group trends are weakly or moderately correlated and the unit-level deviations are time-dependent. We expect that our model should do no worse than other estimators in the first setting, but much better in the second one.

We follow Xu (2017) for designing our DGP. In both settings, we considered 10 treatment units and 20 control units. We let the number of pre-treatment periods be $T_0 = 30$ and the total number of periods be $T = 50$. The outcomes were generated from two time-dependent covariates, a group-level time-trend process, a unit-level deviation process, a grand mean response, an effect process and white noise:

$$y_i(t) = x_{i,1} + 3x_{i,2} + \gamma_g(t) + u_i(t) + \delta(t) \cdot D_i(t) + 0.5 + \varepsilon_{it}.$$

The covariates $x_{i,1}$ and $x_{i,2}$ were drawn i.i.d. from $\mathcal{N}(0, 0.5^2)$. The group-level time-trends $\gamma_0(t)$ and $\gamma_1(t)$ were jointly drawn from a multi-task GP with a zero mean, SE kernel with length scale of 7, output scale of 0.1, and

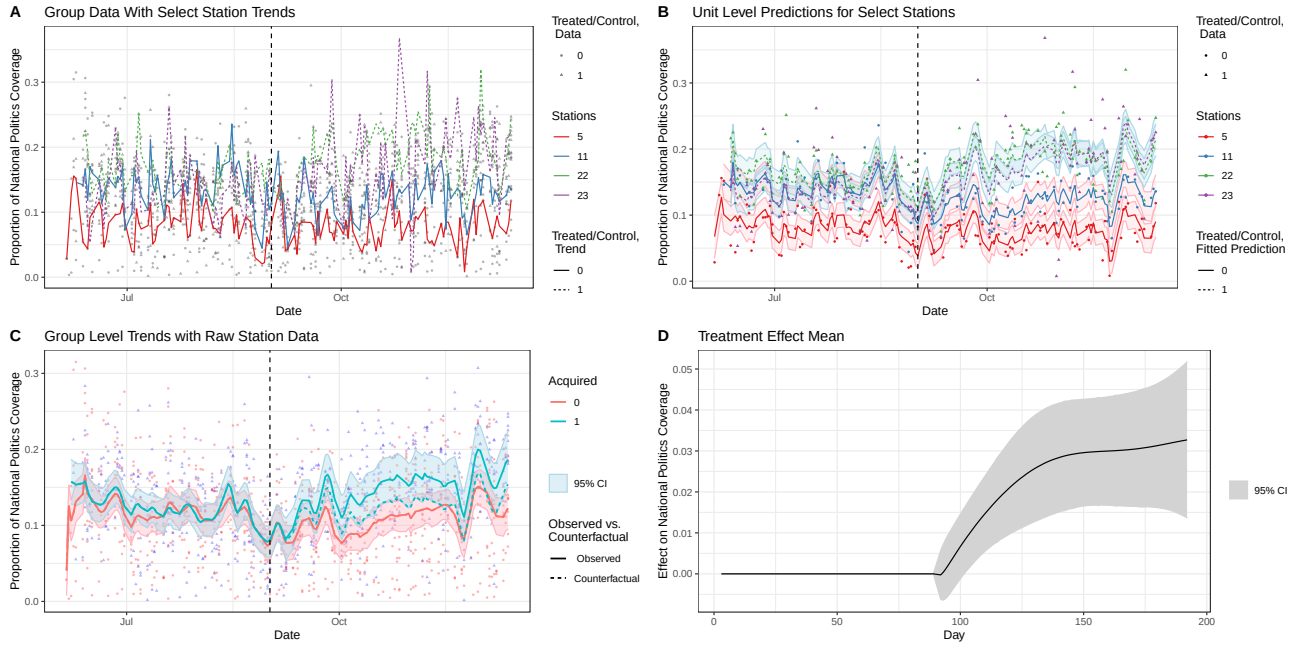


Figure 2: Our posterior beliefs for the local news case study. Panel A shows the raw station data over time, with the outlined trends of two selected control stations (solid) and two treated stations (dashed); Panel B shows those four stations' predicted trends with 95% credible intervals and the raw data as points; Panel C shows the fitted group level trends with 95% credible intervals (solid), the modeled counterfactual trend of the treated group (dashed green line), and the raw data as points; Panel D shows the modeled average treatment effect on the treated with a 95% credible interval.

correlation parameter ρ. When included, the unit deviations $u_i(t)$s were i.i.d. draws from a single-task GP with zero mean and SE kernel with length scale of ℓ and output scale of 0.02. To simulate smoothly converging treatment effects $\delta(t)$, we used an effect process with a scaled SE kernel with length scale of 30 and output scale of 0.1, masked by the scaling function (8). We first conditioned this effect process prior on two data points $(T_0, 0)$ and (T, τ), and then used the posterior mean as the true effect. We fixed the eventual full effect size at T, τ, to 0.1. The error terms ε_{it} were drawn i.i.d. from $\mathcal{N}(0, 0.01^2)$. We fixed the grand mean response to be 0.5. We averaged the results of 25 repeated experiments with different random seeds.

Setting 1 had perfectly correlated group trends and time-invariant unit deviations. We sampled a group trend from a single GP shared across both groups (equivalent to $\rho = 1$), and sampled constant unit-level effects from $\mathcal{N}(0, 0.02^2)$ (equivalent to $\ell = \infty$).

Setting 2 had imperfectly correlated group trends and time-dependent unit variations with $\ell = 21$. We considered a variety of settings from weakly correlated to highly correlated ($\rho = 0.1, 0.5, 0.9$).

Baselines. We include the standard two-way fixed effects (2FE), generalized synthetic control (GSC in (Xu, 2017)), causal multi-task GP (CMGP in (Alaa and van der Schaar, 2017)), dynamic multilevel latent factor (DM-LFM in (Pang et al., 2021)), intrinsic coregionalization (ICM in (Arbour et al., 2021)) and longitudinal treatment response (LTR in (Xu et al., 2016)) models as baselines (see Supplement for implementation details). Both 2FE and GSC ignore temporal/cross-sectional correlations in the underlying time trends and any prior knowledge in the treatment effect, while DM-LFM captures temporal but not cross-sectional correlations in temporal trends using auto-regressive models. LTR includes dynamic structure in treatment effects, but assumes complete independence on the underlying trends, while CMGP assumes perfect dependence. Among these baselines, ICM is the most similar to MGP-PANEL, but still ignores temporal structure in the treatment effect.

Results. We evaluated performance on estimating the ATT in the simulation study using three metrics. The *root mean squared error* (RMSE) measures the quality of point estimates (from the posterior mean) of the estimated effects. The 95% *coverage rate* (coverage) measures the credibility of the claimed 95% confidence intervals. We defined "coverage" as the frequency of actual true effects falling into the claimed 95% credible (confidence) intervals. We also considered the average log predictive likelihood (LL) of the true effects in the posterior.

Table 1 shows the averaged performance scores of our model compared to baseline estimators under Setting 1. Standard errors are shown in parentheses to the precision of last significant digit. Bold numbers indicate the best performance

and italic numbers indicate results that are *not significantly worse* than the bold numbers under a paired t-test with $\alpha = 0.05$. MGP-PANEL does better than other estimators in RMSE (significantly so) and LL when the 2FE model model is "correct," although the coverage rate is slightly, but not significantly, lower than DM-LFM. The improvement in RMSE and LL can be ascribed to our modeling of correlations in the effect process.

Table 2 shows the average performance measures for MGP-PANEL and the baseline methods under Setting 2. MGP-PANEL has significantly lower RMSE, higher coverage rate, and higher LL scores for all levels of correlation in the parallel"ish" trend setting.

Ablation study. In the Supplement we present an ablation study where we compare the performance of MGP-PANEL in Setting 2 after "stripping away" various components of our model. The two most important components of our model responsible for its success appear to be, in order, (1) the modeling of nonlinear group trends and (2) modeling group trends as correlated rather than perfectly parallel. Again, by learning the correlation in the non-linear baseline time trends across groups from data rather than assuming full independence (uncorrelated) or dependency (perfectly correlated), MGP-PANEL can infer the counterfactual outcomes in the treatment group with more fidelity and accuracy.

We also include an additional simulation evaluating the robustness of MGP-PANEL when the kernel or the observation model is mis-specified or when data are scarce. We show that in most cases MGP-PANEL is robust under mild-to-moderate mis-specification, which can be avoided with simple model selection prior to inference.

6.2 Case Studies

Finally, we present a case study using real-world data to demonstrate the flexibility of MGP-PANEL (but not accuracy, as ground truth is not available). The LocalNews data consists of proportions of national news coverage in 2.5 minute segments from 25 television stations for six months (Martin and McCrain, 2019a,b). In September 2017, the U.S. media conglomerate Sinclair Broadcasting Group acquired 11 television stations (the "treated" group) across seven media markets. The 14 non-acquired stations in those same markets serve as the control group. The original paper hypothesized that Sinclair would nationalize news coverage as a cost-cutting measure, since national stories are not specific to local stations. The authors used the LocalNews data to test their hypothesis with a DID design.

The results are shown in Figure 2 (see Supplement for additional model details). We find that, on average, acquisition by Sinclair increased the proportion of national news coverage by 3.27%. This estimate is more than double the estimate of 1.4% reported by Martin and McCrain (2019b),

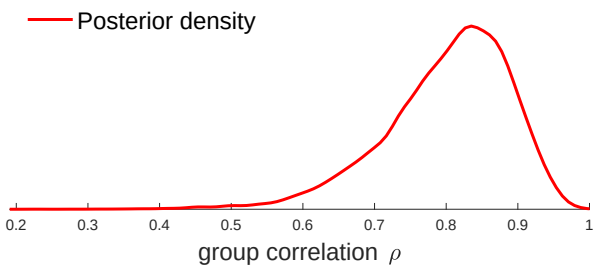


Figure 3: Posterior density of task correlation parameter in local news data. The x-axis is truncated due to minimal mass for $\rho < 0.2$. The posterior mode is $\rho = 0.835$, indicating highly but not perfectly correlated inter-group trends.

which assumes time-invariant effects in post-treatment periods. We show the posterior distribution of the task correlation parameter for the LocalNews data in Figure (3). The posterior mode is $\rho = 0.835$, indicating that the data supports the hypothesis of highly but not perfectly correlated inter-group trends. Therefore the assumption of parallel"ish" trends is better supported by the data, and our model can draw useful conclusions under this relaxation.

In the Supplement we examine how well baseline methods analyze the Localnews data and emphasize that MGP-PANEL learns a more reasonable delayed treatment effect due to its tailored prior. In addition, MGP-PANEL has more precise estimation with the lowest uncertainty over time. In the Supplement we also provide another case study using data from the War in Afghanistan to illustrate the incorporation of a non-Gaussian (Poisson) likelihood into our model.

7 CONCLUSION

We proposed MGP-PANEL, a novel multi-task GP model for inferring time-varying effects in panel data. MGP-PANEL (1) encodes structured temporal correlations in baseline trends across groups and across individuals, and (2) includes a flexible nonparametric prior on the temporal evolution of treatment effects. Our Bayesian causal inference framework can infer posteriors for dynamic treatment effects, while reflecting remaining uncertainty about unobserved counterfactual trends. We show this by analyzing the asymptotic properties of the joint posterior of the treatment effect, which exhibits intuitive behavior in the limit. Experiments show that the MGP-PANEL approach does no worse than existing methods and far better when the standard assumptions are violated. In the Supplement, we show the model is robust to mild mis-pecifications, and also demonstrate its flexibility in a case study when applied to data with a non-Gaussian likelihood.

We can anticipate several potential adjustments to our model for practitioners. First, although we focus on time-varying

effects, our model can be easily adapted to infer conditional effects by including observed characteristics in the treatment effect model. Second, although we use squared exponential kernels throughout this study, researchers may easily defend against mis-specification by including model selection in the inference procedure to determine the optimal kernels, or marginalizing over a pool of pre-selected kernels. Third, our model can be extended to the staggered adoption setting where interventions are correlated with observed confounding variables and thus vary from unit to unit. Finally, our model may further consider categorical and/or continuous treatments by adjusting the task-level kernel.

Acknowledgements

We thank anonymous reviewers and members of WashU Political Science Data Lab, as well as seminar participants at the 39th annual meeting of the Society for Political Methodology (PolMeth 2022) for many valuable feedbacks. We are specifically grateful to Yiqing Xu (Stanford University), Christopher Lucas (Washington University in St. Louis) and Ted Enamorado (Washington University in St. Louis) for their insightful comments and suggestions, and Joshua McCrain (The University of Utah) and Gregory Martin (Stanford University) for sharing data. YC and RG were supported by the National Science Foundation (NSF) under award number IIS–1845434.

References

A. Abadie and J. Gardeazabal. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132, 2003.

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490): 493–505, 2010.

A. Abadie, A. Diamond, and J. Hainmueller. Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510, 2015.

V. Aglietti, T. Damoulas, M. Álvarez, and J. González. Multi-task Causal Learning with Gaussian Processes. *Advances in Neural Information Processing Systems*, 33: 6293–6304, 2020.

A. M. Alaa and M. van der Schaar. Bayesian Inference of Individualized Treatment Effects Using Multi-task Gaussian Processes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

A. M. Alaa and M. van der Schaar. Bayesian Nonparametric Causal Inference: Information Rates and Learning Algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

A. M. Alaa, M. Weisz, and M. Van Der Schaar. Deep Counterfactual Networks with Propensity-Dropout. *arXiv preprint arXiv:1706.05966*, 2017.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ, 2008.

D. Arbour, E. Ben-Michael, A. Feller, A. Franks, and S. Raphael. Using Multitask Gaussian Processes to estimate the effect of a targeted effort to remove firearms. *arXiv preprint arXiv:2110.07006*, 2021.

D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. Synthetic Difference-in-Differences. *American Economic Review*, 111(12):4088–4118, 2021.

O. Atan, J. Jordon, and M. Van der Schaar. Deep-Treat: Learning Optimal Personalized Treatments From Observational Data Using Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

S. Athey and G. W. Imbens. Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. *Journal of Econometrics*, 226(1):62–79, 2022.

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

J. Bai. Panel Data Models With Interactive Fixed Effects. *Econometrica*, 77(4):1229–1279, 2009.

A. Beath, F. Christia, and R. Enikolopov. Can Development Programs Counter Insurgencies?: Evidence from a Field Experiment in Afghanistan, December 2017. MIT Political Science Department Research Paper No. 2011-14. Available at SSRN: `https://papers.ssrn.com/abstract=1809677`.

R. M. Bell and D. F. McCaffrey. Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples. *Survey Methodology*, 28(2):169–181, 2002.

E. Ben-Michael, A. Feller, and J. Rothstein. The Augmented Synthetic Control Method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021.

E. V. Bonilla, K. Chai, and C. Williams. Multi-task Gaussian Process Prediction. In J. Platt and D. Koller and Y. Singer and S. Roweis, editor, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. Inferring Causal Impact Using Bayesian Structural Time-Series Models. *The Annals of Applied Statistics*, 9 (1):247–274, 2015.

D. Card and A. B. Krueger. Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4):772–84, 1994.

P. Chen, W. Dong, X. Lu, U. Kaymak, K. He, and Z. Huang. Deep Representation Learning for Individualized Treatment Effect Estimation Using Electronic Health Records. *Journal of Biomedical Informatics*, 100:103303, 2019.

V. Chernozhukov, K. Wüthrich, and Y. Zhu. An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021.

T. Choi and M. J. Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2007.01.004.

A. Curth and M. van der Schaar. Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

C. De Chaisemartin and X. d'Haultfoeuille. Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–96, 2020.

A. Diamond and J. S. Sekhon. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

S. R. Flaxman, D. B. Neill, and A. J. Smola. Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–23, 2015.

J. Geweke. Comment: Inference and Prediction in the Presence of Uncertainty and Determinism. *Statistical Science*, 7(1):94–101, 1992.

J. Hainmueller, D. J. Hopkins, and T. Yamamoto. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, 22(1):1–30, 2014. doi: 10.1093/pan/mpt024.

J. Hensman, A. Matthews, and Z. Ghahramani. Scalable Variational Gaussian Process Classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015.

J. L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

K. Imai and I. S. Kim. On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis*, 29(3):405–415, 2021.

K. Imai, I. S. Kim, and E. H. Wang. Matching Methods for Causal Inference with Time-Series Cross-Sectional Data. *American Journal of Political Science*, 2021.

G. W. Imbens. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

F. Johansson, U. Shalit, and D. Sontag. Learning Representations for Counterfactual Inference. In *International Conference on Machine Learning*, pages 3020–3029. PMLR, 2016.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

G. J. Martin and J. McCrain. Replication Data for: Local News and National Politics, 2019a. URL `https://doi.org/10.7910/DVN/G3X4EW`.

G. J. Martin and J. McCrain. Local News and National Politics. *American Political Science Review*, 113(2):372–384, 2019b.

R. Moraffah, P. Sheth, M. Karami, A. Bhattacharya, Q. Wang, A. Tahir, A. Raglin, and H. Liu. Causal Inference for Time Series Analysis: Problems, Methods and Evaluation. *Knowledge and Information Systems*, pages 1–45, 2021.

X. Pang, L. Liu, and Y. Xu. A Bayesian Alternative to Synthetic Control for Comparative Case Studies. *Political Analysis*, pages 1—20, 2021.

J. Perlez. Musharraf Quits as Pakistan's President. *New York Times*, 2008.

C. E. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

D. Reynolds. Gaussian Mixture Models. In S. Z. Li and A. Jain, editors, *Encyclopedia of Biometrics*, pages 659–663. Springer, Boston, MA, 2009.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

D. B. Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

W. Schlenker and M. J. Roberts. Nonlinear Effects of Weather on Corn Yields. *Review of Agricultural Economics*, 28(3):391–398, 2006.

P. Schulam and S. Saria. Integrative Analysis Using Coupled Latent Variable Models for Individualizing Prognoses. *The Journal of Machine Learning Research*, 17(1):8244–8278, 2016.

J. Shi, B. Wang, E. Will, and R. West. Mixed-effects Gaussian Process Functional Regression Models with Application to Dose–Response Curve Prediction. *Statistics in Medicine*, 31(26):3165–3177, 2012.

H. Soleimani, A. Subbaswamy, and S. Saria. Treatment-Response Models for Counterfactual Reasoning with Continuous-time, Continuous-valued Interventions. *arXiv preprint arXiv:1704.02038*, 2017.

M. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Artificial Intelligence and Statistics*, pages 567–574. PMLR, 2009.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

S. Witty, K. Takatsu, D. Jensen, and V. Mansinghka. Causal Inference Using Gaussian Processes with Structured Latent Confounders. In *International Conference on Machine Learning*, pages 10313–10323. PMLR, 2020.

Y. Xu. Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*, 25(1):57–76, 2017.

Y. Xu, Y. Xu, and S. Saria. A Bayesian Nonparametic Approach for Estimating Individualized Treatment-Response Curves. In *Machine Learning for Healthcare Conference*, pages 282–300. PMLR, 2016.

J. Yoon, J. Jordon, and M. Van Der Schaar. GANITE: Estimation of Individualized Treatment Effects Using Generative Adversarial Nets. In *International Conference on Learning Representations*, 2018.

A POSTERIOR ANALYSIS: MATHEMATICAL DETAILS

We address potential identification issues in MGP-PANEL by analyzing its posterior consistency. We show that the uncertainty in the sum of the post-treatment treated group counterfactual plus treatment effect will shrink to 0 at certain rate, while the uncertainty in the expected post-treatment treated counterfactual and that in the treatment effect will shrink to minimal values depending on group correlation ρ. As conditioning on additional observations never increases uncertainty of GP posterior, we derive this analysis by looking at one post-treatment period at a time, assuming that ρ has been inferred from pre-treatment observations. Choi and Schervish (2007) shows that under certain conditions, GPs can serve as universal approximators and consistently estimate (in terms of posterior contraction in the large-scale limit) continuous regression functions even if the function itself is not sampled from the GP prior used to model it.

Formally for any post-treatment time t, denote the post-treatment treated counterfactual outcome as $\gamma_1(t)$, post-treatment control counterfactual outcome as $\gamma_0(t)$ and treatment effect as $\delta(t)$. Let the prior variance of the treatment effect be denoted by $\sigma^2 = K_\delta(t,t)$. Suppose we have noisy observations of n treated and n control units. By MGP-PANEL, we have a joint normal prior on $[\gamma_0, \gamma_1]$ where the marginalized distribution $\gamma_0 \sim \mathcal{N}(0, \sigma_\gamma^2)$, $\gamma_1 \sim \mathcal{N}(0, \sigma_\gamma^2)$ and $\text{cov}(\gamma_0, \gamma_1) = \rho\sigma_\gamma^2$, a normal prior on unit deviation $\mathcal{N}(0, \sigma_u^2)$ and a normal prior on the treatment effect $\delta \sim \mathcal{N}(0, \sigma^2)$. For mathematical convenience, we scale the prior variances to $\sigma_\gamma^2 = 1$ while white noise has variance σ_{noise}^2. The factual/counterfactual and effect are independent so $\text{cov}(\gamma_0, \delta) = \text{cov}(\gamma_1, \delta) = 0$. To simplify, let $s^2 = \sigma_u^2 + \sigma_{\text{noise}}^2$. Hence, we can compute the posterior correlation between $\gamma_1(t)$ and δ, as well as upper bounds for the posterior variance of $\gamma_1(t) + \delta$, $\gamma_1(t)$ and δ as:

$$\text{var}_{\text{post}}\big(\delta(t) + \gamma_1(t)\big) \leq \frac{s^2(1+\sigma^2)}{n + n\sigma^2 + s^2} \tag{11}$$

$$\text{cor}_{\text{post}}\big(\delta(t), \gamma_1(t)\big) = -\frac{n\sigma^2}{\sqrt{\sigma^2(n+s^2)(n\sigma^2+s^2)}} \tag{12}$$

$$\text{var}_{\text{post}}\big(\gamma_1(t)\big) \leq \frac{n(1-\rho^2)+s^2}{n+s^2} \tag{13}$$

$$\text{var}_{\text{post}}\big(\delta(t)\big) \leq \frac{(s^2)(1+\sigma^2)}{n+n\sigma^2+s^2} + \frac{n(1-\rho^2)+s^2}{n+s^2} \tag{14}$$

Note that t is arbitrary post-treatment time period so the above results hold for the entire post-treatment time series. Assume we can achieve the large-sample limit. We first observe that the treatment factual outcome can be *exactly* identified, as the posterior variance of $\gamma_1(t) + \delta(t)$ (24) will shrink to zero in the large-sample limit ($n \to \infty$). The joint posterior over treated counterfactual outcomes and treatment effects also becomes increasingly negatively correlated to -1 as $n \to \infty$. Hence, as long as one hypothesizes a counterfactual or a treatment effect with no uncertainty, the posterior on the other will collapse to a Dirac delta function. Moreover, (32) shows that the treated counterfactual outcome is *partially* identified up to an upper bound depending on the inter-group correlation parameter ρ, and can be *exactly* identified if trends were perfectly correlated (parallel) because this upper bound will shrink to zero. Finally, the treatment effect is also *partially* identified up to the same upper bound and can be *exactly* identified if trends were perfectly correlated (parallel).

A.1 Joint Posterior of Treatment Effect and Post-treatment Treated Counterfactual

Suppose we do not directly observe either γ_1 or δ but rather n corrupted sum $y_i = \gamma_1 + \delta + u_i + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0, \sigma_{\text{noise}}^2)$, where u_i indicates unit deviations at time t. Again to simplify, let $s^2 = \sigma_u^2 + \sigma_{\text{noise}}^2$. Let $\mathbf{1}_n$ denote the $n \times 1$ all-one vector and $\mathbb{I}_n$ denote the $n \times n$ identity matrix, where the $n \times n$ all-one matrix can be represented by $\mathbf{1}_n\mathbf{1}_n^T$. Let $\mathbf{y}$ collects all $y_1, \ldots, y_n$, then $[\delta, \gamma_1, \mathbf{y}]^T$ has the joint distribution of

$$\begin{bmatrix} \delta \\ \gamma_1 \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \sigma^2\mathbf{1}_n^T \\ 0 & 1 & \mathbf{1}_n^T \\ \sigma^2\mathbf{1}_n & \mathbf{1}_n & (1+\sigma^2)\mathbf{1}_n\mathbf{1}_n^T + s^2\mathbb{I}_n \end{bmatrix} \right), \tag{15}$$

where $\otimes$ is the Kronecker product. We appeal to Woodbury matrix identity to derive the posterior variance of the conditional distribution $p(\delta, \gamma_1 \mid \mathbf{y})$ as

$$\text{var}[\delta, \gamma_1 \mid \mathbf{y}] = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \left(\mathbf{1}_n^T \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}\right)\left((1+\sigma^2)\mathbf{1}_n\mathbf{1}_n^T + s^2\mathbb{I}_n\right)^{-1}\left(\mathbf{1}_n \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}^T\right) \tag{16}$$

$$= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{s^2}\left(\mathbf{1}_n^T \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}\right)\left(\frac{1+\sigma^2}{s^2}\mathbf{1}_n\mathbf{1}_n^T + \mathbb{I}_n\right)^{-1}\left(\mathbf{1}_n \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}^T\right) \tag{17}$$

$$= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{s^2}\left(\mathbf{1}_n^T \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}\right)\left(\mathbb{I}_n - \frac{1+\sigma^2}{A}\mathbf{1}_n\mathbf{1}_n^T\right)\left(\mathbf{1}_n \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}^T\right) \tag{18}$$

$$= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \frac{n}{A}\begin{bmatrix} \sigma^4 & \sigma^2 \\ \sigma^2 & 1 \end{bmatrix} \tag{19}$$

$$= \frac{1}{A}\begin{bmatrix} \sigma^2(n+s^2) & -n\sigma^2 \\ -n\sigma^2 & n\sigma^2 + s^2 \end{bmatrix} \tag{20}$$

where $A = n + n\sigma^2 + s^2$. Hence we can compute the posterior correlation between δ and γ_1 as

$$\text{cor}(\delta, \gamma_1) = -\frac{n\sigma^2}{\sqrt{\sigma^2(n+s^2)(n\sigma^2+s^2)}} \tag{21}$$

$$= -\frac{1}{\sqrt{(1+s^2/n)(1+s^2/(n\sigma^2))}} \to -1 \text{ if } n \to \infty \tag{22}$$

We can also compute the posterior variance on $\delta + \gamma_1$ as

$$\text{var}(\delta + \gamma_1) = \frac{\sigma^2(n+s^2) - 2n\sigma^2 + (n\sigma^2 + s^2)}{n + n\sigma^2 + s^2} \tag{23}$$

$$= \frac{s^2(1+\sigma^2)}{n + n\sigma^2 + s^2} \to 0 \text{ if } n \to \infty \tag{24}$$

We can see that although δ and γ_1 are uncorrelated a priori, the posterior correlation will approach to negative one in the limit of infinitely many data ($n \to 0$). In addition, the posterior variance of their sum will also approach to zero in the limit of infinitely many data ($n \to 0$). Hence, if we hypothesize a counterfactual or a treatment effect, then the posterior on the other will collapse.

A.2 Posterior of Post-treatment Control Factual

Suppose we have n corrupted post-treatment control observations $y_i = \gamma_0 + u_i + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$, where u_i indicates unit deviations at time t. Again, denote $s^2 = \sigma_u^2 + \sigma^2_{\text{noise}}$. As GP posterior variance does not increase with more observations, we could be the posterior of γ_0 conditioning on all pre- and post-treatment observations by the posterior variance of $p\big(\gamma_0 \mid y_1, \ldots, y_n\big)$

$$\text{var}[\gamma_0 \mid \mathbb{Y}_{\text{obs}}] \leq \text{var}[\gamma_0 \mid y_1, \ldots, y_n] \tag{25}$$

$$= 1 - \mathbf{1}_n^T\Big(\mathbf{1}_n\mathbf{1}_n^T + s^2\mathbb{I}_n\Big)^{-1}\mathbf{1}_n \tag{26}$$

$$= \frac{s^2}{n + s^2} \to 0 \text{ if } n \to \infty \tag{27}$$

Hence, post-treatment control factual is identifiable as its posterior variance will shrink to 0 with infinitely many data ($n \to \infty$).

A.3 Posterior of Post-treatment Treated Counterfactual

Suppose we have n noisy post-treatment control observations $y_0^{(i)} = \gamma_0 + u_i + \varepsilon_i$, with white noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$ and group correlation parameter $\rho = \text{cov}(\gamma_1, \gamma_0)$. Again u_is are unit deviations at time t and $s^2 = \sigma_u^2 + \sigma^2_{\text{noise}}$. Let $\mathbf{y}_0$ collects all $y_0^{(1)}, \ldots, y_0^{(n)}$, then the variance of $p\big(\gamma_1 \mid \mathbb{Y}_{\text{obs}}\big)$ is bounded by variance of $p(\gamma_1 \mid \mathbf{y}_0)$. We can write the joint covariance matrix of $[\gamma_1, \mathbf{y}_0]$ as

$$\text{cov}\begin{bmatrix} \gamma_1 \\ \mathbf{y}_0 \end{bmatrix} = \begin{bmatrix} 1 & \rho\mathbf{1}_n^T \\ \rho\mathbf{1}_n & \mathbf{1}_n\mathbf{1}_n^T + s^2\mathbb{I}_n \end{bmatrix} \tag{28}$$

Hence we can derive the posterior variance of $p(\gamma_1 \mid \mathbf{y}_0)$ as

$$\begin{aligned}
\text{var}\big[\gamma_1 \mid \mathbb{Y}_{\text{obs}}\big] &\leq \text{var}\big[\gamma_1 \mid \mathbf{y}_0\big] && (29)\\
&= 1 - \rho \mathbf{1}_n^T \Big(\mathbf{1}_n \mathbf{1}_n^T + s^2 \mathbb{I}_n\Big)^{-1} \rho \mathbf{1}_n^T && (30)\\
&= 1 - \rho^2 \frac{n}{n + s^2} && (31)\\
&= \frac{n(1-\rho^2) + s^2}{n + s^2} \to 1 - \rho^2 \text{ if } n \to \infty && (32)
\end{aligned}$$

Hence, the post-treatment treated counterfactual is *partially* identified with an upper bound $1 - \rho^2$ on its variance in the limit of infinitely many data ($n \to \infty$). In the case of perfectly correlated group trends ($\rho = 1$), the post-treatment treated counterfactual can be *exactly* identified.

A.4 Posterior of Treatment Effect

Suppose we have n noisy post-treatment control observations $y_0^{(i)} = \gamma_0 + u_i + \varepsilon_i$ and n noisy post-treatment control observations $y_1^{(i)} = \gamma_1 + v_i + \delta + \varepsilon_i$ with white noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$, where u_i and v_i are unit deviations at time t and $s^2 = \sigma_u^2 + \sigma^2_{\text{noise}}$, respectively. Suppose the group correlation parameter is $\rho = \text{cov}(\gamma_1, \gamma_0)$. Using Eq. 24 and 32, we can derive an upper bound of posterior variance of δ as

$$\begin{aligned}
\text{var}[\delta \mid \mathbb{Y}_{\text{obs}}] &= \text{var}[\delta + \gamma_1 \mid \mathbb{Y}_{\text{obs}}] + \text{var}[\gamma_1 \mid \mathbb{Y}_{\text{obs}}] && (33)\\
&\leq \frac{s^2(1+\sigma^2)}{n + n\sigma^2 + s^2} + \frac{n(1-\rho^2) + s^2}{n + s^2} \to 1 - \rho^2 \text{ if } n \to \infty && (34)
\end{aligned}$$

Hence, the treatment effect is *partially* identified with an upper bound $1 - \rho^2$ on its variance in the limit of infinitely many data ($n \to \infty$). In the case of perfectly correlated group trends ($\rho = 1$), the treatment effect can be *exactly* identified as $1 - \rho^2$ will be zero.

B ADDITIONAL MODEL DETAILS

We briefly provide additional details about the individual components of our model and the hyperpriors used in Eq. (9). We put a multi-task GP prior with constant mean and SE kernel for the group trends $\{\gamma_g\}$. We place a shared GP prior with zero mean and SE kernel on all the unit deviations $\{u_i\}$. We put a GP prior with linear mean and SE kernel on the effects from covariates h. We place a GP prior with zero mean and SE kernel but scaled smoothly from zero at the time of intervention T_0 to a fixed output scale at some later time $T_1 = T_0 + \Delta T$ for the treatment effect process δ. To enforce the full effect time T_1 is always no earlier than intervention time T_0, we require $\Delta T \geq 0$. Observation noise is assumed to be i.i.d. Gaussian. Table 3 summarizes all the model hyperparameters.

Table 3: Table of notations.

Component	Prior	Hyperparameter
Group trends $\gamma_g(t)$	constant $\mu_g(t)$ $K_\gamma(t,t') \cdot K_{\text{task}}(m,m')$	c_γ ℓ_γ, λ_γ, ρ
Unit deviation $u_i(t)$	zero $\mu_u(t) = 0$ $K_u(t,t')$	N/A ℓ_u, λ_u
Covariate $h(x)$	linear $\mu_x(x)$ $K_x(x,x')$	slope a_x ℓ_x, λ_x
Treatment effect $\delta(t)$	zero $\mu_\delta(t) = 0$ $K_\delta(t,t')$	N/A ℓ_δ, λ_δ, ΔT
General noise ε	$\mathcal{N}(0, \sigma^2_{\text{noise}})$	σ_{noise}

In our implementation we transformed some hyperparameters to allow unconstrained optimization sampling. In particular, all length/output scale hyperparameters were parameterized by their log, the correlation parameter was parameterized by an inverse sigmoid (the inverse cumulative normal distribution), and ΔT was left untransformed.

C INFERENCE

In this section, we present a Bayesian causal inference framework that derives the posterior on the evolution of ATT $\delta(t)$, from observed potential outcomes. This framework is similar to the one proposed by Xu et al. (2016), but tailored for our setting with expected group trends.

Denote the observed outcomes under different treatment assignments $\mathbb{Y}_{\text{obs}}^{(1)} = \{Y_i^{(1)}(t) \mid D_i(t) = 1\}$ and $\mathbb{Y}_{\text{obs}}^{(0)} = \{Y_i^{(0)}(t) \mid D_i(t) = 0\}$, and define $\mathbb{Y}_{\text{obs}} = \mathbb{Y}_{\text{obs}}^{(1)} \cup \mathbb{Y}_{\text{obs}}^{(0)}$. Assume we have a Gaussian process prior on the treatment effect $p(\delta)$ and a GP model for controlled potential outcomes $p\big(\mathbb{Y}^{(0)}\big)$, which are connected via the treated potential outcomes $\mathbb{Y}^{(1)} = \mathbb{Y}^{(0)} + \delta$. Since the effects are independent of controlled potential outcomes, $\mathbb{Y}^{(1)}$ has an induced GP prior of $p\big(\mathbb{Y}^{(1)}\big)$. Collecting all prior parameters into θ, the posterior inference of δ can be derived by conditioning on observed treated and controlled potential outcomes $\mathbb{Y}_{\text{obs}}$:

$$p\big(\delta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}, \theta\big) \propto p\big(\mathbb{Y}_{\text{obs}}^{(1)} \mid \mathbb{Y}^{(0)}, \delta, \theta\big)\, p\big(\mathbb{Y}_{\text{obs}}^{(0)} \mid \mathbb{Y}^{(0)}, \theta\big)\, p(\delta \mid \theta). \tag{35}$$

In this work, we embrace the idea of *fully* Bayesian inference, which addresses model uncertainty by marginalizing over the hyperparameters of our mean and covariance functions. Let $\boldsymbol{\theta}$ denote the set of hyperparameters in our model. We put mildly informative priors on $\boldsymbol{\theta}$. The hyperparameter-marginal posterior is then:

$$p\big(\delta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}\big) = \int p\big(\delta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}, \theta\big)\, p\big(\theta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}, \delta\big)\, \mathrm{d}\theta. \tag{36}$$

Unfortunately, this integral is intractable, so we must resort to approximation or sampling. Here we used Hamiltonian Markov chain Monte Carlo sampling. Given a set of K hyperparameter samples from the posterior $\{\boldsymbol{\theta}_k\} \sim p\big(\theta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}, \delta\big)$, the marginalized effect posterior $p\big(\delta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}\big)$ can be approximated with a Gaussian process mixture model Reynolds (2009). If desired, we may approximate this GP mixture with a single Gaussian process via moment matching.

Note that our proposed framework also reduces to an alternative Bayesian causal inference framework in Arbour et al. Arbour et al. (2021) if we assign an infinitely wide GP prior on δ. The alternative framework transforms the effect estimation into an imputation problem of the unobserved post-treatment counterfactuals for treatment group $\mathbb{Y}_{\text{mis}}^{(0)} = \{Y_i^{(0)}(t) \mid D_i(t) = 1\}$, and then uses the difference between observed $\mathbb{Y}_{\text{obs}}^{(1)}$ and imputed $\mathbb{Y}_{\text{mis}}^{(0)}$ as an estimation for δ. While this alternative framework allows infinite flexibility for the treatment effects, it ignores any prior knowledge on δ, such as their dynamic structure or effect size. We evaluated this model in the experiments in the main text.

D HYPERPRIORS

We place mildly informative priors on the hyperparameters in both simulation and case studies, and then marginalize over the hyperparameters by sampling from their posterior given the data.

The sampling is done via Hamiltonian Markov chain Monte Carlo.[3] We sampled five chains using a random restart around the MAP estimator for initialization. Specifically, we initialized each chain by perturbing the MAP hyperparamters (in the transformed space) with an additive Gaussian jitter with standard deviation equal to 0.1. For each chain, 3000 samples were collected after a burn-in of 1000 samples, which we found to be typically sufficient.

D.1 Simulation Studies

The data generating process of the simulation studies is described in main text. We fixed the prior mean for group trends as the empirical observation mean $c_\gamma = \mathbb{E}[\mathbf{y}]$. We do not put hyperpriors on the slope a_x in the prior mean for h. Hyperpriors

[3]We use the `hmcSampler` function from the *Statistics and Machine Learning Toolbox* in Matlab R2019b. The leapfrog step size, number of leapfrog integration steps and mass vector are automatically tuned using the `tuneSampler` function.

for the remaining hyperparameters are listed below:

$$\begin{aligned}
\ell_\gamma &\sim \text{Gamma}(10, 2) & \lambda_\gamma &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) \\
\rho &\sim \text{Uniform}(-1, 1) & \ell_u &\sim \text{Gamma}(2, 10) \\
\lambda_u &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) & \ell_x &\sim \text{Gamma}(10, 2) \\
\lambda_x &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) & \Delta T &\sim \text{Gamma}(10, 2) \\
\ell_\delta &\sim \text{Gamma}(10, 3) & \lambda_\delta &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) \\
\varepsilon &\sim \text{SmoothUniform}(e^{-4}, e^{-1})
\end{aligned}$$

Here, for all the length scales and general noise we used a modified uniform distribution with rapidly decaying but smooth tails that is differentiable everywhere; this ensures the gradient is informative outside the chosen range. The upper bound is set very generously at $e^{-1} \approx 0.368$, which is much larger than the total variation of outcomes in the simulation and case studies.

D.2 LocalNews

The model for LocalNews data is almost the same as the one for the simulation studies, but the covariates are replaced with day and "day of the week" effects. These effects account for daily variations in the news coverage trends and additional indexes are used for indicating each day and weekday. We put Gaussian priors on those effects and mild hyperpriors on the hyperparameters.

$$\begin{aligned}
Y_i(t) &= \gamma_g(t) + u_i(t) + \text{day} + \text{day-of-week} + \delta(t) + \varepsilon \\
\text{day} &\sim \mathcal{N}(0, \sigma^2_{\text{day}}) \\
\text{day-of-week} &\sim \mathcal{N}(0, \sigma^2_{\text{day-of-week}})
\end{aligned}$$

Since Gaussian distributions are closed under addition, we could marginalize over these day effects and absorb them into the full model covariance.

$$K_y = K_u + K_\gamma + K_\delta + \sigma^2_{\text{day}} + \sigma^2_{\text{day-of-week}} + \sigma^2_{\text{noise}}$$

The hyperpriors for LocalNews model are summarized below:

$$\begin{aligned}
\ell_\gamma &\sim \text{Gamma}(10, 5) & \lambda_\gamma &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) \\
\rho &\sim \text{Uniform}(-1, 1) & \ell_u &\sim \text{Gamma}(10, 5) \\
\lambda_u &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) & \sigma_{\text{day-of-week}} &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) \\
\sigma_{\text{day}} &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) & \Delta T &\sim \text{Gamma}(2, 5) \\
\ell_\delta &\sim \text{Gamma}(10, 5) & \lambda_\delta &\sim \text{SmoothUniform}(e^{-4}, e^{-1}) \\
\varepsilon &\sim \text{SmoothUniform}(e^{-4}, e^{-1})
\end{aligned}$$

E BASELINE IMPLEMENTATION DETAILS

Here we provide implementation details for all of the alternative methods.

1. The 2FE model was implemented using the standard approach by introducing time/unit indicators into the regression model as fixed effects, but used the ordinary least square estimator with robust standard errors (Bell and McCaffrey, 2002) to account for noise heteroscedasticity among units. OLS with the robust standard error is similar to standard OLS, but assumes the noise may be heteroscedastic, such that the noise matrix is clustered as a blocked matrix, in our case, by unit. The implementation relied on the `lm_robust` function from `estimatr` library in R. We used the default setting of HC2 type.
2. The GSC model was implemented using software[4] provided by Xu (2017). We set the number of maximal factors to be 10 and allow the built-in cross validation procedure to select the optimal number of factors. We impose unit and day fixed effects besides interactive effects.

[4]Code available on `https://github.com/xuyiqing/gsynth`

Table 4: Performance measures of MGP-PANEL compared to ablated models across selected settings of correlation parameters (standard errors are shown in parentheses). Among all models, MGP-PANEL has the highest coverage and LL scores in all settings. Although the MAP estimator has lower RMSE than MGP-PANEL, the differences are not significant in a paired t-test.

		model							
	ρ	MGP-PANEL	MAP	naïve	uncorr effect	uncorr trend	no unit trend	BLR	perfect corr
RMSE	0.1	*0.0242*	**0.0210**	0.0619	0.0592	0.0765	0.0392	0.1120	0.0773
		(0.0027)	(0.0034)	(0.0080)	(0.0074)	(0.0075)	(0.0034)	(0.0096)	(0.0096)
	0.5	*0.0229*	**0.0202**	0.0561	0.0527	0.0656	0.0365	0.0782	0.0561
		(0.0027)	(0.0030)	(0.0073)	(0.0068)	(0.0078)	(0.0036)	(0.0068)	(0.0066)
	0.9	*0.0171*	**0.0163**	0.0276	0.0250	0.0335	0.0245	0.0341	0.0281
		(0.0020)	(0.0017)	(0.0034)	(0.0034)	(0.0048)	(0.0029)	(0.0031)	(0.0036)
coverage	0.1	**0.802**	0.614	0.556	*0.606*	0.266	0.626	0.060	0.222
		(0.062)	(0.072)	(0.065)	(0.064)	(0.055)	(0.078)	(0.022)	(0.050)
	0.5	**0.816**	0.596	0.550	0.594	0.230	0.656	0.080	0.268
		(0.055)	(0.074)	(0.065)	(0.065)	(0.057)	(0.073)	(0.027)	(0.052)
	0.9	**0.802**	0.582	0.630	*0.710*	0.314	*0.738*	0.186	0.248
		(0.059)	(0.068)	(0.057)	(0.060)	(0.052)	(0.068)	(0.035)	(0.045)
LL	0.1	**2.19**	*0.761*	−2.260	−1.800	-33.2	0.789	−468	−119
		(0.19)	(0.816)	(1.532)	(1.428)	(13.2)	(0.379)	(74)	(57)
	0.5	**2.23**	0.841	−2.020	−1.420	-38.6	0.830	−226	−68.3
		(0.20)	(0.649)	(1.438)	(1.358)	(20.8)	(0.407)	(40)	(32)
	0.9	**2.55**	1.170	−0.002	*0.686*	-14.8	1.700	−39.8	−23.5
		(0.19)	(0.532)	(0.932)	(0.918)	(4.2)	(0.310)	(8.1)	(7.4)

3. The CMGP model was implemented using the software[5] provided by Alaa and van der Schaar (2017). Since CMGP is not designed for time series data, we incorporated time as an additional feature into the GP kernel to account for the effect of time. Time was also inserted as a feature when computing multiple-periods treatment effects. We averaged estimated treatment effects across units, since CMGP outputs *individualized* treatment effects.

4. The DM-LFM model was implemented using the `bpCausal` library in R provided by Pang et al. (2021). We allowed the time-varying parameters and factors to be 1-order autocorrelated. We set the number of maximal factors to be 10, where the optimal number is determined by build-in hierachical shrinkage priors for factor selection procedure. We used the default number of burn-in and runs for MCMC sampling. We also imposed unit and day fixed effects in addition to interactive effects. We manually checked Geweke's convergence diagnostics Geweke (1992) to ensure convergence.

5. The ICM model was implemented using the Matlab gpml software [6] following the setup in Arbour et al. (2021). We used independent standard normal priors for the scalar coefficients, and set the number of latent processes to be 5. The details for MCMC sampling are the same with MGP-PANEL to ensure comparability.

F ABLATION STUDY

We conducted an ablation study to demonstrate the effectiveness of each part of our model by ablating several key components in MGP-PANEL separately. We restricted this study to Setting 2.

The **maximum a posteriori** (MAP) estimator restricts inference to point estimations rather than a fully Bayesian inference to show the benefit of model averaging to infer treatment effects. MAP uses the same model as MGP-PANEL but fixes hyperparameters to maximum a posteriori estimation.

The **naïve** estimator shows the value of including dynamics in treatment effect structures into the model. Specifically, naïve only extrapolates the post-treatment periods for the treated group after conditioning on the pre-treatment observations for the treated group and the pre- and post-treatment observations for the control group, and then subtracts the observed outputs

[5]Code available on `https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/causal_multitask_gaussian_processes_ite`

[6]See `https://gaussianprocess.org/gpml/code/matlab/doc/index.html`

from the posterior predictions in the post-treatment periods for the treated group as estimations of treatment effects. This estimator is equivalent to ours except that the GP prior on ATT is assumed to be infinitely wide (Arbour et al., 2021). Hence, naïve estimator allows ATT to be arbitrary but incorporates no temporal correlation.

The **uncorr effect** and **uncorr trend** models do not impose a smooth GP prior on treatment effects and on group trends, demonstrating the regularization effect of GP in modeling temporal relations of either the effect or time trends. Specifically, the "uncorr effect" and "uncorr trend" estimators have almost the same model as MGP-PANEL, but separately fix the length scales for the effect and group trends to be zero so that the treatment effects/time trends are uncorrelated.

The **no unit trends** estimator deletes the unit deviation component to illustrate the effectiveness of allowing unit deviations from group trends.

The **BLR** estimator is the Bayesian version of the 2FE model, which speaks to the advantage of using non-linear GP for modeling time trends. To ensure comparability, we use the following set of priors, which are similar to the GP priors in MAP:

$$
\begin{aligned}
\gamma_g(t) &= c_1 \\
u_i(t) &= c_2 \\
h(x) &= ax \\
c_1 &\sim \mathcal{N}\big(\mathbb{E}[\mathbf{y}], \lambda_\gamma^{*\,2}\big) \\
c_2 &\sim \mathcal{N}\big(0, \lambda_u^{*\,2}\big) \\
a &\sim \mathcal{N}\big(a_x^*, \lambda_x^{*\,2}\big) \\
\delta(t) &\sim \mathcal{N}\big(\mathbf{0}, \lambda_u^{*\,2}\mathbb{I}\big)
\end{aligned}
$$

The **perfect corr** estimator forces the correlation parameter ρ to be 1, inducing perfect parallel trends. By imposing a perfect correlation between group trends, "perfect corr" tests the necessity of weakening the parallel trends assumption.

Table 4 shows the averaged RMSE, 95% confidence interval coverage rate, and log likelihood scores of MGP-PANEL and ablated models for different levels of correlation. Among all models, MGP-PANEL has the highest COVERAGE and LL scores in all settings, but the MAP estimator has the lowest RMSE. A reason for this may be that MGP-PANEL is designed for a better coverage rate and log likelihood due to the fully Bayesian inference that absorbs the model and hyperparameter uncertainty. Note that although the MAP estimator has a better RMSE than the MGP-PANEL estimator, the differences are not significant in a paired t-test.

Several implications could be drawn from Table 4. First, the **BLR** and **perfect corr** estimators have the worst performance among all models, emphasizing that inferring time trends correctly is the most critical aspect of accurately estimating treatment effects using time series. Second, the **uncorr trend** model has better performance than the **BLR** and **perfect corr** estimators but is still much worse than the other models, indicating the vital role of regularization in modeling time effects when temporal structure exists. Finally, the **naïve**, **uncorr trend** and **no unit trends** estimators perform are just slightly worse than MGP-PANEL and MAP models, encouraging practitioners to further regularize treatment effects, fully make use of post-treatment data and take into consideration unit-level heterogeneity in groups whenever possible.

G ADDITIONAL SIMULATION

We conducted additional simulations to examine whether our proposed model is correctly specified and whether it is robust to model misspecification. Accordingly, we conducted additional simulation experiments assessing the performance of our approach in settings where the kernel and/or the observation model is misspecified. Performance scores of MGP-PANEL compared to baseline estimators under different data generating processes (DGP) averaged across different random seeds are reported below, including using a student-t noise model with degree of freedom equal 4 (non normal error), modeling non smooth time trends using GP with Matérn kernel (non smooth trends), observing one unit per group (few units) and modeling group trends using independent GP (independent GP trends). Lower RMSE, higher LL and COVERAGE scores closer to the theoretical value 0.95 indicate better performance.

Table 5: Performance scores of MGP-PANEL compared to baseline estimators under different data generating process (DGP) averaged across different random seeds, including using a student-t noise model with degree of freedom equal 4 (non normal error), modeling non smooth time trends using GP with Matérn kernel (non smooth trends), observing one unit per group (few units) and modeling group trends using independent GP (independent GP trends). Lower RMSE, higher LL and COVERAGE scores closer to the theoretical value 0.95 indicate better performance. Performance scores of GSC and 2FE under few units setting are not available due to bugs in released code from original authors.

	MGP-PANEL	GSC	2FE	CMGP	DM-LFM	ICM	LTR
DGP				RMSE			
Non normal error	*0.0270(21)*	0.0583(19)	0.0522(22)	**0.0221(16)**	0.0540(17)	0.0500(17)	0.0359(51)
Non smooth trends	**0.0157(22)**	0.0424(13)	0.0413(13)	0.0235(10)	0.0406(22)	0.0368(22)	0.0471(47)
Few units	**0.0328(41)**	N/A	N/A	*0.0477(51)*	0.1039(49)	0.1035(45)	*0.0326(37)*
Independent GP trends	**0.0298(45)**	0.1034(43)	0.0767(70)	*0.0433(50)*	0.0606(29)	0.0497(24)	*0.0378(36)*
DGP				COVERAGE			
Non normal error	*0.935*(25)	*0.985*(31)	**0.945**(7)	1.000(0)	*0.975*(11)	0.758(15)	0.665(91)
Non smooth trends	*0.905*(40)	**0.950**(7)	*0.935*(7)	0.990(5)	*0.980*(11)	0.785(28)	0.400(32)
Few units	*0.970*(13)	N/A	N/A	1.000(0)	1.000(0)	*0.930(13)*	**0.955**(19)
Independent GP trends	0.855(27)	*0.930(22)*	0.645(54)	0.800(47)	**0.950**(20)	0.725(42)	0.555(87)
DGP				LL			
Non normal error	*2.062(93)*	1.415(25)	1.525(45)	**2.198(31)**	1.486(33)	0.570(208)	−1.602(1.003)
Non smooth trends	**2.338(167)**	1.672(28)	1.741(30)	*2.310*(34)	1.772(50)	1.475(124)	−1.999(0.854)
Few units	**1.826(65)**	N/A	N/A	1.173(41)	0.779(42)	0.798(41)	*1.658(75)*
Independent GP trends	**1.468(164)**	0.738(79)	1.086(96)	*1.353(270)*	*1.312(53)*	0.949(160)	0.579(73)

The above table shows that our proposed method is robust under mild-to-moderate misspecification. Our proposed method still outperforms the baselines under almost all DGP settings, and only has slightly RMSE and lower log likelihood than CMGP under non normal noise setting. Although the performance of our model does occasionally break down under wild misspecification (e.g., modeling extremely heavy-tailed noise as Gaussian or no correlation in the group trends), we want to stress that the particular modeling choices of (squared exponential kernel, Gaussian noise) we made in our experiments are by no means set in stone, nor do we necessarily recommend their use "off the shelf" without validation. As in any task, we'd recommend a practitioner invest time in model selection prior to inference following standard practice to avoid finding themselves with a wildly misspecified model. As another example, in our simulation modeling heavy-tailed Student t noise as Gaussian, a model combining a GP with a heavy tailed (rather than Gaussian) error model is overwhelmingly preferred to the same model with Gaussian noise (BIC = -863 vs -1014). Thus with some prudent modeling the breakdown could have been avoided entirely.

Table 6: Demonstration of model selection when the true noise model is student-t distributed or the group trends are generated from Matérn kernels. Lower Bayesian information criterion (BIC) is preferred.

DGP	Student-t distributed noise		group trends with Matérn kernels	
model	Student-t noise	Gaussian noise	Matérn kernel	Gaussian kernel
BIC	**−1014**	−863	**−847**	−832

H CASE STUDY: COMPARISON WITH BASELINES

We also examine how well the baseline methods do in the case study and why our method is favored. In general, our method is designed to infer time-varying treatment effects and can also incorporate prior belief on the treatment effects such as smoothness or if they are instantaneous. However, the treatment effects in the original case study is assumed to be constant over post-treatment periods and is estimated by a standard two-way fixed effect model. Although the baselines methods in this paper can accommodate time-varying effects, they tend to ignore any prior belief that sometimes is reasonable in

real-world data.

Table 7: Comparison of estimated averaged treatment effects and uncertainties between MGP-PANEL and other baselines at various post-treatment time periods in the applied study. Amongst all models, MGP-PANEL has the lowest uncertainty on ATE and learned a minor instantaneous treatment effect.

	MGP-PANEL	GSC	2FE	DM-LFM	CMGP	ICM	LTR
Post-treatment time				ATE			
after two weeks	0.36%	2.93%	2.30%	2.65%	3.24%	1.36%	-0.46%
after six weeks	2.83%	5.10%	4.00%	4.40%	4.23%	1.26%	2.86%
after twelve weeks	3.19%	5.14%	3.30%	4.52%	4.21%	4.05%	3.89%
Post-treatment time				STD			
after two weeks	0.46%	2.34%	1.55%	1.39%	1.68%	4.82%	0.96%
after six weeks	0.67%	2.41%	1.82%	1.44%	1.81%	4.76%	1.06%
after twelve weeks	0.81%	2.96%	1.65%	1.45%	1.81%	5.27%	1.00%

Here we show the estimated treatment effects of baseline methods. Since there is no ground truth for computing performance measure such as RMSE or log likelihood, we report the estimated ATEs and uncertainty, averaged over the first, third and last two weeks. We make two observations: 1) Our proposed method has more precise estimation as we have the lowest uncertainty over time, because the baseline methods do not model temporal correlations of the treatment effects and apply no smooth conditions; and 2) all baseline models learned a sharp, immediate effect. We do not find this to be substantively realistic. We posit that the nationalization of local news would have a delayed effect as it takes time for stations to adapt programs, reports, and restructure their journalistic practices, which our model can accommodate.

I CASE STUDY: NON-GAUSSIAN LIKELIHOODS

In this section we provide an example using SIGACTS data that highlights how the model can be extended to accommodate non-Gaussian likelihoods.

The SIGACTS data is collected by the U.S. military, reporting details such as dates, locations, and categories of violence (e.g. direct or indirect fire) for significant actions (SIGACTS). Our SIGACTS data is for the War in Afghanistan from January 1, 2007 to December 31, 2008, where actions are daily counts. We use the SIGACTS data to show how MGP-PANEL performs with non-Gaussian observation likelihoods.

We specifically examine an increase in conflict along the Afghanistan–Pakistan border in 2008. In 2008, President Pervez Musharraf of Pakistan, a U.S. ally, became embroiled in a corruption scandal, culminating in a coalition government agreeing to impeach him on August 11, 2008. Musharraf resigned on August 18 Perlez (2008). We assume that the increased violence started on Aug 12, 2008 (the day after the government agreed to impeach Musharraf) and compared the aggregated number of reported direct fire events along the border to the number of events inland. We hypothesize that, after announcing plans to impeach the president, the coalition sought to establish their independence from the West by violently confronting the U.S. and ISAF forces along the Afghanistan–Pakistan border, so we should observe more direct fire events in the border region.

The number of reported direct fires can be naturally modeled as count data using the Poisson likelihood, where the log of the rate parameter across time is specified by a latent Gaussian process. This specification is referred to as latent Gaussian models in the literature Rue et al. (2009). Hence, the treatment effect can be interpreted as the increase in the *ratio* of the densities of direct fire between the two regions. Although the Poisson likelihood does not allow exact inference, it is bell-shaped and can be approximated using Laplace's method.

$$y(t) \sim \text{Poisson}\big(\lambda(t)\big) \tag{37}$$

$$\log\big(\lambda(t)\big) \sim \mathcal{GP}(\mu_y, K_y) \tag{38}$$

The model we used for this study is similar to that described above. The key difference for the SIGACTS model is its Poisson likelihood, so we do not have the general noise component. In addition, we remove the unit deviation component since there

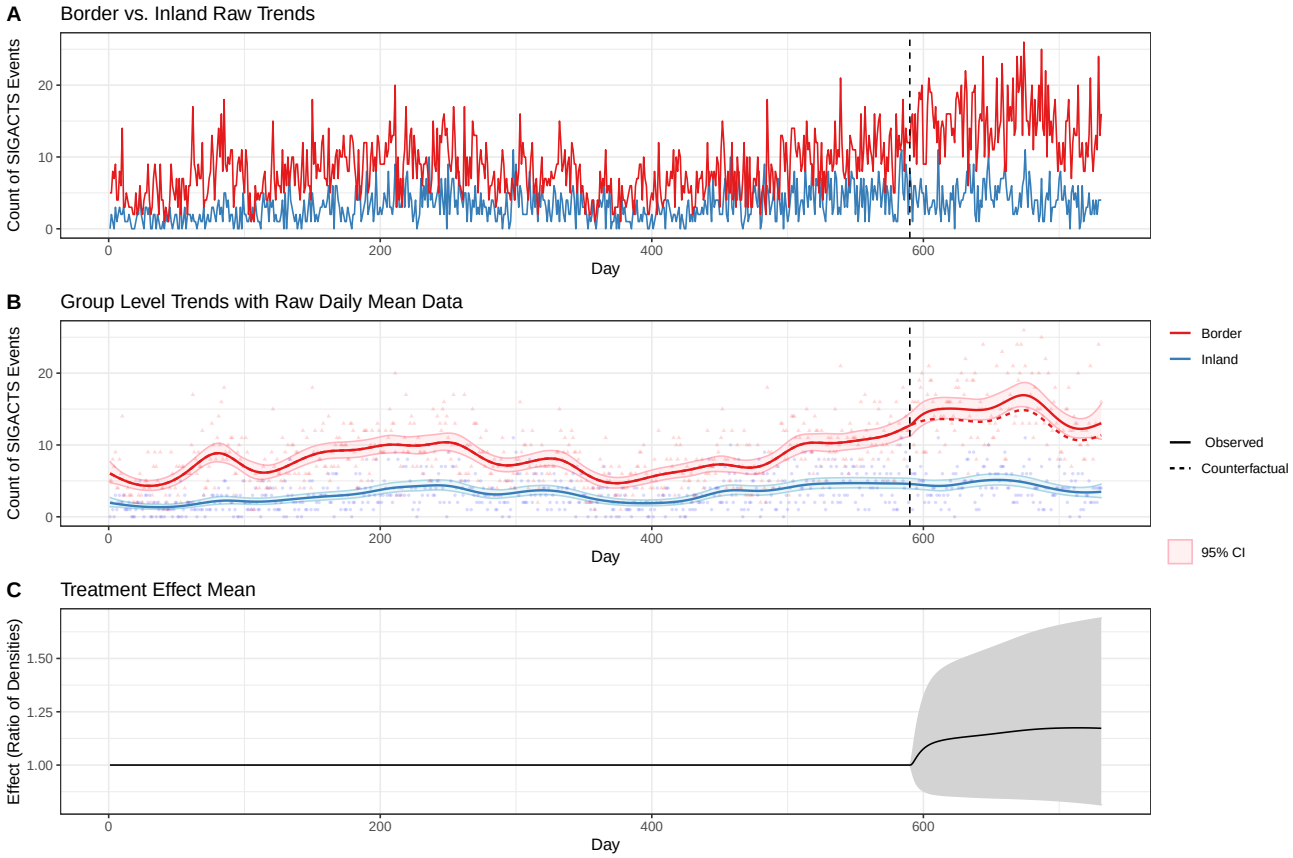


Figure 4: Panel A shows the trends of the raw group counts over time, where the border province count (treatment) is in red and the inland province count (control) is in blue. Panel B shows the fitted group level trends with 95% credible intervals (solid lines and shaded regions), the modeled counterfactual trend of the treated group (dashed red line), and the raw group counts as points. Panel C shows the posterior of the average treatment effect on the treated interpreted as a ratio of the density of direct fire on the border to the density of direct fire inland with a 95% credible interval.

are only two time series. The hyperpriors for the SIGACTS model are summarized below:

$$\begin{aligned}
\ell_\gamma &\sim \text{Gamma}(10, 8) \\
\rho &\sim \text{Uniform}(-1, 1) \\
\Delta T &\sim \text{Gamma}(3, 10) \\
\ell_\delta &\sim \text{Gamma}(10, 8) \\
\lambda_\delta &\sim \text{SmoothUniform}(e^{-4}, e^{-1})
\end{aligned}$$

In Figure 4, Panel A shows the trends of the raw group-day counts over time, where the border province counts (treatment) are in red and the inland province counts (control) are in blue. Panel B shows the fitted group level trends with 95% credible intervals (solid lines and shaded regions), the modeled counterfactual trend of the treated group (dashed red line), and the raw group counts as points. Panel C shows the posterior of the average treatment effect on the treated interpreted as a ratio of the density of direct fire on the border to the density of direct fire inland, which, on average, is 1.18 although the credible interval does include 1. This result is consistent with studies in international politics on SIGACTS and U.S.-led coalition forces in Afghanistan Beath et al. (2017), but also illustrates that the added structure of the model by no means ensures that we will recover large (low variance) treatment effect estimates.