Using Multitask Gaussian Processes to estimate the effect of a targeted effort to remove firearms[*]

David Arbour, Eli Ben-Michael, Avi Feller, Alex Franks, and Steven Raphael

September 27, 2021

Abstract

In 2006 California implemented the Armed and Prohibited Persons System (APPS), a monitoring program that identifies known firearm owners who become prohibited from owning a firearm, and then subsequently attempts to retrieve all prohibited weapons from such individuals. Our goal is to assess the effect of APPS on California murder rates using annual crime data from both California and comparison states. To do so, we adapt a non-parametric Bayesian approach, multitask Gaussian Processes (MTGPs), which allows for a flexible and parsimonious panel data model that nests many existing approaches and allows for direct control over both dependence across time and dependence across units, as well as natural uncertainty quantification. We extend this approach to incorporate non-Normal outcomes, auxiliary covariates, and multiple outcome series, which are all important in our application. We also show that this approach has attractive Frequentist properties, including a representation as a weighting estimator with separate weights over units and time periods. Applying this approach, we find large effects of the APPS program on gun-related homicides.

1 Introduction

Every year, thousands of people in the United States are murdered with a firearm. In 2017 alone, nearly 40,000 people in the U.S. died from a gun-related injury, including deaths by suicide and murder (Gramlich, 2019). Addressing gun violence remains one of the most pressing public safety issues in the United States.

In 2006 California introduced an innovative program designed to curb gun violence, known as the Armed and Prohibited Persons System (APPS). This monitoring program identifies known firearm owners who become prohibited from owning a firearm and then subsequently attempts to retrieve all prohibited weapons. Through this program, state law enforcement has made hundreds of thousands of contacts with prohibited individuals and has removed tens of thousands of firearms from their possession, often permanently. See Ben-Michael et al. (2021) for detailed discussion.

Our goal is to estimate the impact of APPS on the number of murders in California using annual, state-level data. Unfortunately, simple approaches are unlikely to perform well here; for example, the assumptions underlying the workhorse "difference-in-differences" regression approach do not hold in this application. In one recent analysis, Ben-Michael et al. (2021) instead consider a variant of the synthetic control method (Abadie et al., 2010). This approach, however, fails to address many important features of the problem, such as multiple outcome series, and does not lead to convincing uncertainty quantification in this case.

In this paper, we adapt a non-parametric Bayesian approach, *multitask Gaussian Processes* (MTGPs), to estimate the causal effect of APPS. Originally developed for the setting with multiple outcomes rather than multiple units (Bonilla et al., 2008; Álvarez, 2017), Multitask GPs separately parameterize dependence across time and dependence across units. This allows for a flexible and parsimonious panel data model that nests many existing approaches: Our primary model is a natural generalization of low-rank factor models commonly used in panel data settings (e.g., Xu, 2017; Athey et al., 2021) with an additional prior that encourages smoothness in the underlying factors. We also take advantage of the large literature on the statistical and computational properties of Gaussian Processes (Williams and Rasmussen, 2006; Gelman et al., 2013) to extend this model. Immediate extensions include: allowing for a count (Poisson) observational model; allowing for multiple outcome series; incorporating a mean model; and incorporating auxiliary covariates.

This flexibility, however, also comes with a cost. In settings such as ours, with relatively few units and (pre-treatment) time periods, estimates can be especially sensitive to the prior distributions. Thus, careful model checking is an important feature of the proposed workflow. In particular, we re-frame many standard panel data diagnostics in terms of standard Bayesian workflows, such as posterior predictive checks. This avoids many of the issues with, e.g., inflated Type I error due to pre-testing for parallel trends (Roth, 2019).

In addition to diagnostics, the Bayesian approach also yields coherent uncertainty quantification, appropriately propagating uncertainty from the different sources of information in the model. In

particular, the specific model we consider automatically produces uncertainty intervals that grow over time post-treatment. Surprisingly, this last point is a departure from many existing approaches, which sometimes have constant uncertainty intervals post-treatment. We also report many flexible summaries of the posterior distribution, allowing us to obtain posterior distributions for estimands like the benefit-cost ratio, a key quantity in policy analysis.

Finally, we adapt results from Kanagawa et al. (2018) showing that the Gaussian Process can be represented as a weighting estimator. We then use this representation to show that the overall weights decompose into separate unit and time weights. As a result, the proposed MTGP approach has the form of a doubly-weighted average over units and time periods, similar to recent proposals in the panel data literature (Arkhangelsky et al., 2019; Ben-Michael et al., 2021a).

Applying these ideas to the impact of APPS yields large, negative effect estimates on gun-related homicides in California. Our preferred Poisson model estimates an impact of several hundred prevented homicides per year, suggesting very large benefits from this program. In contrast, using the same methodology, we estimate that there is effectively no impact of APPS on non-gun-related homicides, as expected.

As we discuss in Section 2.3, our proposed approach combines the growing literature on Gaussian Processes for causal inference (see Oganisian and Roy, 2020) with the robust literature on estimating causal effects for a single treated unit with panel data (see Samartsidis et al., 2019). Of particular relevance are Modi and Seljak (2019) and Carlson (2020), who also consider GPs in this context, albeit with different structures. More broadly, our proposed framework is a natural (Bayesian) generalization of recent proposals for low-rank factor models with panel data (e.g., Xu, 2017; Athey et al., 2021); see also recent extensions from Pang et al. (2020).

Our paper proceeds as follows. In Section 1.1 we give a brief overview of the APPS program and other institutional details. In Section 2 we describe the underlying causal problem and review related work. In Section 3 we review single- and multi-task Gaussian Processes and apply these ideas to our application setting. In Section 4 we represent the MTGP approach as a weighting estimator. In Section 5 we report model diagnostics and the estimated impact of APPS under different models. Finally, in Section 6 we discuss open questions and possible extensions. The Appendix includes additional computational details, diagnostics, and results.

1.1 Armed and Prohibited Persons System

The Armed and Prohibited Persons System is a law enforcement program that proactively removes firearms from known gun owners who become prohibited from owning a gun. We provide a brief overview here; see Ben-Michael et al. (2021) for further details. The program has two parts: a state-wide database and an enforcement operation. While the state has maintained the APPS database since the mid 1990s, the enforcement operation did not begin until December 2006 (California Department of Justice, 2015).
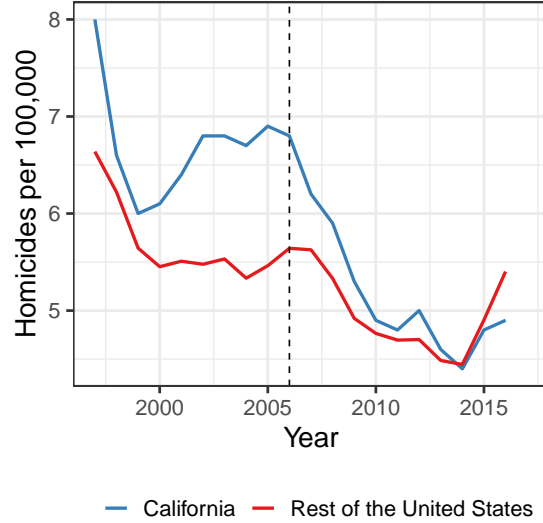
Figure 1: Homicide rate per 100,000 people in California and the rest of the United States. The dotted line is 2006, the year the APPS program was launched.

First, the APPS database lists all known gunowners in the State of California, building on sales records and voluntary gun registrations. The known population of gunowners in the APPS database has grown substantially over time, from under 900,000 in 2006 to over 2.5 million in 2019, or roughly 6 percent of the state's resident population. This database is cross-referenced daily against other statewide databases that record events that trigger a firearm prohibition, including the state's criminal history system, mental health reporting system, restraining and protective order system, and a wanted persons system. When the cross-reference uncovers a triggering event for an individual in APPS, the case is analyzed by state criminal intelligence specialists to determine whether the person is actually prohibited. If the person is deemed prohibited, the APPS database notes a change in status from a lawful to a prohibited gun owner and the case is referred to the state Bureau of Firearms (BOF) for an enforcement action. Among the 2.5 million individuals in the APPS database in early 2019, approximately 23,000 were armed and prohibited with cases at various stages of investigation (California Department of Justice, 2019).

Using the APPS database, a group of analysts and state law enforcement officers then directly enforce state firearm restrictions for gun owners who become prohibited. Typically, these officers attempt to contact the individual at their last known address and request to perform a consensual search of the premises. If refused, officers will return with a search warrant. Often, officers only recover some of the firearms associated with the individual in question — but frequently discover unregistered weapons along with the registered firearms. Following the firearm seizures, the individuals are then removed from the APPS database. Between 2007 and 2017, the BOF directly retrieved nearly 32,000 firearms from prohibited individuals. Interviews with BOF officials suggest

that law enforcement officers at the county and local levels retrieved substantially more firearms related to APPS seizures during this period, though exact numbers are not available.

Figure 1 shows the homicide rate per 100,000 people in California versus the rest of the United States (with state-level estimates weighted by population); the dotted line indicates 2006, the year the APPS program was first launched (in December). The figure shows several striking patterns. First, the overall murder rate in the United States declined over this period, with a large increase in the final years of the panel. Second, the murder rate in California was substantially higher than the national average in the late 1990s and mid 2000s, but eventually converged to the rest of the country by the mid 2010s.

The methodological question is whether and how we can use this panel data structure to estimate the impact of APPS on the murder rate. Furthermore, we would like to separate out the effect on *gun related* and *non-gun related* homicides. The APPS program should primarily impact gun-related homicides by removing firearms from a high-risk population, though substitution effects may lead to changes in non-gun-related homicides. Therefore our methodological goal also includes estimating effects on *multiple outcomes* simultaneously.

Unfortunately, relatively simple approaches are unlikely to perform well here. For example, the workhorse "difference-in-differences" regression model assumes *parallel trends*: in the absence of the intervention, the differences between the murder rates in California and the rest of the country are constant over time. This is clearly violated in Figure 1, with much wider differences in the mid 2000s than in the late 1990s.

In a recent paper, Ben-Michael et al. (2021) instead consider a variant of the *synthetic control method* (SCM) to estimate this effect, constructing a weighted average of other states ("synthetic control"), which closely matches California's pre-treatment murder rate. Ben-Michael et al. (2021) then compare the observed murder rate in California to that in "synthetic California," finding a reduction of around 0.8 homicides per 100,000 people, roughly 10 percent below pre-intervention baseline.

While a promising initial analysis, several open questions remain. First, the estimated SCM weights on control units are sensitive to modeling choices (even if the overall SCM estimate is relatively stable), with little guidance on how to choose among them. Second, Ben-Michael et al. (2021) propose a preliminary SCM analysis using multiple outcomes series: gun-related, non-gun-related, and overall homicide rates. However, the methodology for panel data with multiple outcomes is under-developed, and it is unclear how the approach used in Ben-Michael et al. (2021) can be applied more generally. Relatedly, the number of homicides per state are inherently count outcomes, a fact ignored in existing SCM analyses. Finally, quantifying uncertainty with synthetic controls is challenging. Ben-Michael et al. (2021) consider both placebo and conformal inference approaches (Abadie et al., 2010; Chernozhukov et al., 2021), which rest on versions of exchangeability either over units or over time. These assumptions are difficult to justify in our setting. Moreover, the

uncertainty intervals from the conformal inference approach are constant over time even though, intuitively, we anticipate far greater uncertainty for the counterfactual murder rate in 2017 than immediately after the implementation of APPS in 2007. These concerns motivate our alternative approach, which we turn to next.

2 Setup and Background

2.1 Problem setup

We now describe the standard panel data setting with $i = 1, \ldots, N$ units observed for $t = 1, \ldots, T$ time periods. In our application, we observe $N = 50$ states for $T = 20$ years. Let W_i be an indicator that unit i is treated at time $T_0 < T$ where units with $W_i = 0$ never receive the treatment. While our setup is more general, we restrict our attention to the case where a single unit receives treatment, and follow the convention that this is the first one, $W_1 = 1$. The remaining $N_0 = N - 1$ units are untreated. In our application, the treatment of interest is the APPS program, treated at time $T_0 = 10$; California is the treated unit and the remaining states are possible controls.

We describe our problem of interest in the potential outcomes framework. We assume no interference between units and stable treatment, which allows us to write the potential outcomes for unit i in time t under control and treatment as $Y_{it}(0)$ and $Y_{it}(1)$ respectively.[1] Our primary outcome of interest is the overall murder rate; we consider multiple outcomes in Section 3.3.3. For ease of exposition, we initially ignore the presence of background covariates from our analyses; we discuss extensions to incorporate them in Section 3.3.1.

Since the first unit is treated at time T_0, the observed outcomes Y_{it} are therefore:

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } W_i = 0 \text{ or } t \leq T_0 \\ Y_{it}(1) & \text{if } W_i = 1 \text{ and } t > T_0 \end{cases}.$$

The goal is to estimate the causal effect for the treated unit at each post-treatment time $t > T_0$, $\tau_t = Y_{1t}(1) - Y_{1t}(0) = Y_{1t} - Y_{1t}(0)$, as well as the average treatment effect after time T_0,

$$\tau = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \tau_t.$$

Since we observe the treated outcome $Y_{1t} = Y_{1t}(1)$ for the treated unit at $t > T_0$, the challenge is to impute the missing potential outcome for the treated unit at time $t > T_0$. Note that we adopt a finite sample causal inference framework here: our goal is to estimate the impact for California, rather than for a hypothetical population (see Imbens and Rubin, 2015).

---

[1] While this is not always the case, in our application, it is reasonable to assume that any state could launch an APPS-style program at any point. Thus, both potential outcomes could reasonably exist for all units and all time periods.

2.2 Bayesian causal inference

We adopt a Bayesian approach to estimating causal quantities. There are many subtle issues that arise in Bayesian causal inference that are not central to our discussion. For instance, identification in the Bayesian approach is conceptually distinct from identification in non-Bayesian causal inference: if the prior distribution is proper then the posterior distribution will also be proper, regardless of whether the parameters in the likelihood are fully or partially identified (see Gustafson, 2010; Imbens and Rubin, 2015). See Ding and Li (2018) and Oganisian and Roy (2020) for additional background. Pang et al. (2020) and Menchetti and Bojinov (2020) give overviews of Bayesian causal inference for panel or time series data.

At a high level, Bayesian causal inference views missing potential outcomes as unobserved random variables to be imputed (Rubin, 1978; Ding and Li, 2018). We therefore break the problem into two parts: first, estimate a Bayesian model using the observed data; second, use this model to obtain the posterior (predictive) distribution for the missing potential outcomes. Specifically, let $\mathbf{Y}^{\text{mis}} = \{Y_{it}(0)\}_{i=1,t>T_0} \in \mathbb{R}^{T_0}$ be the missing potential outcomes for the treated unit after T_0, and let $\mathbf{O}^{\text{obs}} = (\{Y_{it}(0)\}_{i>1}, \{Y_{it}(0)\}_{i=1,t\leq T_0}, \{Y_{it}(1)\}_{i=1,t>T_0}, \mathbf{W})$ be the set of observed data, including all control outcomes, the treated unit's pre-treatment outcomes, the treated unit's observed outcomes post-treatment, and the set of treatment assignments. We introduce auxiliary covariates in Section 3.3.1.

The goal is to impute the missing potential outcomes, $\mathbf{Y}^{\text{mis}}$, by drawing from the posterior predictive distribution, $\mathbb{P}(\mathbf{Y}^{\text{mis}}|\theta, \mathbf{O}^{\text{obs}})$, governed by a model parameter θ. We can then estimate $\hat{\tau}_t = Y_{1t} - Y_{1t}^*(0)$ at post-treatment time $t > T_0$, where $Y_{1t}^*(0)$ is a draw from the posterior predictive distribution for $Y_{1t}(0)$ for unit 1 at time t. To do so, we follow Richardson et al. (2011) and use the fact that:

$$\mathbb{P}\{\mathbf{Y}^{\text{mis}}, \theta \mid \mathbf{O}^{\text{obs}}\} = \mathbb{P}\{\theta \mid \mathbf{O}^{\text{obs}}\} \cdot \mathbb{P}\{\mathbf{Y}^{\text{mis}} \mid \theta, \mathbf{O}^{\text{obs}}\},$$

where $\mathbb{P}\{\mathbf{Y}^{\text{mis}}, \theta|\mathbf{O}^{\text{obs}}\}$ is the posterior distribution of our model parameters and the missing potential outcomes given observed data. Now, to obtain the target posterior predictive distribution, we first specify a prior distribution $\mathbb{P}(\theta)$ and use Bayes rule to obtain the posterior distribution $\mathbb{P}(\theta|\mathbf{O}^{\text{obs}})$. With this posterior over the model parameters, we can repeatedly draw samples of $\mathbf{Y}^{\text{mis}}$, yielding the joint posterior on the left-hand side of the above expression. Obtaining the target posterior predictive distribution is then a matter of rearranging the above expression to isolate $\mathbb{P}(\mathbf{Y}^{\text{mis}}|\theta, \mathbf{O}^{\text{obs}})$. Intuitively, this approach rests entirely on the form of the model and the estimated model parameters. See Pang et al. (2020) for an alternative framing that explicitly states sufficient conditions in the form of identifying assumptions.

Formally, we assume that units and time periods are exchangeable conditional on the model parameters θ, which have a prior distribution $p(\theta)$. Following Richardson et al. (2011), we partition

the parameter space into θ^m, which governs the marginal potential outcome distribution, and θ^a, which governs the association between $Y_{it}(0)$ and $Y_{it}(1)$. In principle, we could treat θ^a as a sensitivity parameter; for the purposes of this paper, we assume that the potential outcomes are conditionally independent given data and model parameters, allowing us to ignore θ^a. The posterior for θ^m is then:

$$\mathbb{P}\{\theta^m \mid \mathbf{O}^{\text{obs}}\} \propto p(\theta) \prod_{(i,t) \in \mathcal{C}} \mathbb{P}(Y_{it}(0) \mid \theta^m, W_{it}),$$

where $\mathcal{C}$ denotes the "control" unit and time periods: $t = 1, \ldots, T_0$ for $i = 1$ and $t = 1, \ldots, T$ for other units. Let $Y_{1t}^*(0)$ denote the posterior predictive distribution over the missing potential outcomes, $Y_{1t}^*(0) = \mathbb{P}\{\mathbf{Y}^{\text{mis}} \mid \theta, \mathbf{O}^{\text{obs}}\}$. The estimated treatment effect at time t is then the observed treated potential outcome minus the predicted missing potential outcome at post-treatment time $t > T_0$, $\hat{\tau}_t = Y_{1t} - Y_{1t}^*(0)$, where repeatedly drawing $Y_{1t}^*(0)$ from the posterior predictive distribution propagates the uncertainty.

A key consequence of this setup is that we never use post-treatment data from the treated unit in our modeling (or, equivalently, we have an arbitrarily flexible model for the observed treated data). Rather, our focus is on modeling the observed control potential outcomes. And by imputing $Y_{it}^*(0)$ from the posterior *predictive* distribution, our inference is inherently finite sample, allowing us to directly reason about the effect of APPS on California's murder rate (Imbens and Rubin, 2015).

Finally, while the setup here is quite general, in practice we restrict ourselves to a structure where the control potential outcomes are a *model component* plus additive noise.

Assumption 1 (Additive, separable error structure). The control potential outcomes, $Y_{it}(0)$, are generated as:

$$Y_{it}(0) = \text{model}_{it}(\theta) + \varepsilon_{it} \qquad \mathbb{E}[\varepsilon_{it}] = 0$$

where $\text{model}_{it}(\theta)$ is the outcome model for unit i at time t, governed by parameters θ; and ε_{it} is mean-zero noise.

This assumption, sometimes referred to *strict exogeneity* in the panel data setting (see, e.g., Imai and Kim, 2019), places no restrictions on the model terms and θ. This allows for *non-parametric* modelling of the outcomes where θ is potentially infinitely dimensional — but does not allow the errors to depend on treatment assignment. In other words, given the model and the possibly infinitely many parameters θ, the error distribution for the treated unit's (control) potential outcomes are exchangeable with other units' errors. As we discuss in Section 6, we can relax this restriction within the fully Bayesian workflow, for instance, via sensitivity analysis (Franks et al., 2019).

7

2.3 Related work

Our paper bridges several robust literatures. First, we build on the many existing methods for estimating causal effects for a single treated unit with panel data; see Samartsidis et al. (2019) for a recent review. Within this, there are several important threads. Most directly relevant are recent papers on Bayesian implementations of (or alternatives to) the synthetic control method, including (Tuomaala, 2019; Kim et al., 2020; Pang et al., 2020; Pinkney, 2021). There are also several recent papers that directly estimate factor models for causal effects (e.g., Xu, 2017; Athey et al., 2021), as well as approaches that directly address multiple outcomes (Samartsidis et al., 2020). Finally, in an important addition, Brodersen et al. (2015) propose a Bayesian structural time series model for estimating causal effects, focused on a single treated series; see Menchetti and Bojinov (2020) for a recent extension. As we discuss below, we incorporate many of the novel ideas in these papers, including many of the prior choices. However, our proposed MTGP framework is typically more general, and, we believe, more conducive to estimating causal effects with panel data.

Next, there is a small but growing set of papers on the use of Gaussian Processes for estimating causal effects. To date, nearly all of these papers have focused on the cross-sectional setting (Alaa and van der Schaar, 2017; Schulam and Saria, 2017; Ray and van der Vaart, 2018; Huang et al., 2019; Branson et al., 2019; Witty et al., 2020; Ren et al., 2021). See Oganisian and Roy (2020) for a recent review.

We are aware of very few papers specifically using GPs to estimate causal effects in panel data settings, although, as we discuss in Section 4, many existing estimators can be written as special cases of this approach.[2] Two working papers are especially relevant for our work. The first is recent work from Modi and Seljak (2019), who also propose a GP approach for estimating causal effects in this setting. Their main proposal, however, is very different than ours, with a focus on GP for the frequency domain. Thus, we view our model as a useful complement to theirs. A second recent proposal is Carlson (2020), who also uses a GP approach but does not exploit the multitask structure. Finally, while not explicitly causal, of particular relevance to our approach is Flaxman et al. (2015), who propose a hierarchical GP model that is similar to what we discuss here.

3 Multitask Gaussian Processes for Causal Inference

We now give a high-level description of Multitask Gaussian Processes for causal effects with panel data, beginning with the simpler setting of single-task GPs. We discuss specific model implementation choices in Appendix A.

---

[2]See also Karch et al. (2018), who use GPs for latent growth curve modeling in psychometrics, and Huang et al. (2015), who instead focus on causal discovery.

3.1 Review: Single Task Gaussian Processes

We begin by reviewing the setup and properties of a Single-Task GP, which focuses solely on the treated unit; see Williams and Rasmussen (2006) for a textbook discussion. While the original GP setup is quite general, to fix ideas we initially focus on applying GPs to the *interrupted time series* or *horizontal regression* setting, in which we use the pre-treatment outcomes for California to forecast post-treatment outcomes in the absence of the intervention. Specifically, we initially model the (control) murder rate for California at time t, $Y_{1t}(0)$, as the sum of a model component f_{1t} and a mean-zero, independent noise component (as in Assumption 1):

$$Y_{1t}(0) = f_{1t} + \varepsilon_{1t}$$
$$\mathbf{f}_1 \sim \mathcal{GP}(0, k(t, t'))$$
$$\varepsilon_{1t} \sim N(0, \sigma^2)$$

where $\mathcal{GP}(\cdot)$ is a Gaussian Process prior and $\mathbf{f}_1 = (f_{11}, \ldots, f_{1T})$ is the vector of model components for the treated unit. The key idea is that the GP prior over $\mathbf{f}_1$ incorporates smoothness over time via a *kernel function* $k(t, t')$, where similar values of t imply larger covariances. Here we use the squared exponential kernel:

$$k_{\text{time}}(t, t') = \exp\left(-\frac{\|t - t'\|^2}{2\rho}\right),$$

where ρ is the length scale.

The choice of kernel function plays an important role in defining the behavior of the model. For the family of kernels we consider, the relation between the time difference $|t - t'|$ and the kernel value $k(t, t')$ is central in determining how we expect the model to behave over time. If $k(t, t')$ is large even for far-apart time periods, then the model components $f_{1t}, f_{1t'}$ are assumed to be highly correlated, which corresponds to an assumption that the underlying model is very smooth over time. Conversely, if $k(t, t')$ is low even for close time periods, then the model components are assumed to be close to independent, which corresponds to assuming that the model varies strongly over time. The length scale hyper-parameter ρ explicitly controls the level of smoothness in the model components. Many other kernels are possible and appropriate for other settings, such as periodic kernels; see Williams and Rasmussen (2006) for a textbook discussion.

To write the MTGP model in matrix notation, let $\mathbf{K}_{\text{time}} \in \mathbb{R}^{T \times T}$ be the corresponding *time kernel matrix*, where $\mathbf{K}_{\text{time}, tt'} = k_{\text{time}}(t, t')$. Then, we can write $\mathbf{f}_1 \sim \mathcal{GP}(0, k(t, t'))$ as:

$$\mathbf{f}_1 \sim \text{MVN}(0, \mathbf{K}_{\text{time}}).$$

Because the model is a multivariate Normal, the conditional distributions have a convenient form. Specifically, as in Section 2.2, we can partition the vector of control potential outcomes for the treated unit into the observed and missing components: $\mathbf{Y}_1^{\text{obs}} = (Y_{11}^{\text{obs}}, Y_{12}^{\text{obs}}, \ldots, Y_{1T_0}^{\text{obs}})$, and

$\mathbf{Y}^{\text{mis}} = (Y_{1T_0+1}(0), Y_{1T_0+2}(0), \ldots, Y_{1T}(0))$. Suppressing the superscript *time* to reduce clutter, we can then partition this multivariate Normal model as:

$$\begin{pmatrix} \mathbf{Y}^{\text{mis}} \\ \mathbf{Y}_1^{\text{obs}} \end{pmatrix} \ \sim \ \text{MVN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{\text{mis}} & \mathbf{K}_{\text{mis,obs}} \\ \mathbf{K}_{\text{obs,mis}} & \mathbf{K}_{\text{obs}} + \sigma^2 \mathbb{I}_{T_0} \end{pmatrix} \right], \tag{1}$$

where $\mathbf{K}_{\text{obs}} \in \mathbb{R}^{T_0 \times T_0}$ is the kernel matrix for observed points, $\mathbf{K}_{\text{mis}} \in \mathbb{R}^{(T-T_0) \times (T-T_0)}$ is the kernel matrix for unobserved points, and $\mathbf{K}_{\text{obs,mis}} \in \mathbb{R}^{T_0 \times (T-T_0)}$ is a rectangular matrix of kernel evaluations for pre- and post-treatment times, whose t, t' element is the kernel value $k(t, T_0 + t')$. We can then write the the posterior predictive distribution for California's outcomes at future time points, $t > T_0$, as:

$$\mathbf{Y}^{\text{mis}} | \mathbf{Y}_1^{\text{obs}} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} = \mathbf{K}_{\text{mis,obs}}(\mathbf{K}_{\text{obs}} + \sigma^2 \mathbb{I})^{-1} \mathbf{Y}_1^{\text{obs}} \tag{2}$$
$$\boldsymbol{\Sigma} = \mathbf{K}_{\text{mis}} - \mathbf{K}_{\text{mis,obs}}(\mathbf{K}_{\text{obs}} + \sigma^2 \mathbb{I})^{-1} \mathbf{K}_{\text{obs,mis}}. \tag{3}$$

As we discuss in Section 4, the posterior mean is a linear combination of the pre-treatment observations, $\mathbf{Y}_1^{\text{obs}}$.

3.2 Multitask Gaussian Processes

We now turn to modeling the murder rate for all states jointly, rather than modeling California's outcome series alone. The primary change is that each observation is now a unit-time pair (i, t), so the kernel measuring similarity between points is doubly-indexed: $k(\ (i, t), (i', t')\ )$. The *multitask* GP framework, originally developed for the setting with a single unit but many outcomes (Goovaerts et al., 1997; Bonilla et al., 2008), is a natural approach for capturing similarity across multiple dimensions. As in the single-task GP setting and Assumption 1, we assume that the control potential outcomes consist of a model component plus additive (Normal) noise:

$$Y_{it}(0) = f_{it} + \varepsilon_{it}$$
$$\varepsilon \sim N(0, \sigma^2),$$

where we want to flexibly model f_{it}. While this framework is quite general, we focus on two common simplifications: allowing for separable unit and time covariances, and imposing a low-rank structure on the unit covariance.

Separable unit and time kernels. The first key idea is to decompose the GP kernel into separate kernels that capture similarity across time (the "time covariance") and similarity across

units (the "unit covariance"):

$$k(\,(i,t),(i',t')\,) = k_{\text{unit}}(i,i') \times k_{\text{time}}(t,t'), \tag{4}$$

where k_{unit} defines similarity between the N units and k_{time} defines similarity between the T time periods. This is a strong restriction that implies that the covariance across units, k_{unit}, is constant in time. (Conversely, this implies that the time covariance k_{time} is constant across units.) We discuss some approaches for relaxing this restriction in Section 6.

Low-rank task covariance. Even with the separable kernel restriction, the resulting model is overparameterized. In particular, the similarity between the T time points, $k_{\text{time}}(t,t')$, has a natural parameterization in terms of the difference in time, $|t - t'|$, which is a scalar. By contrast, the similarity between the N units, $k_{\text{unit}}(i,i')$, involves $(N-1)^2$ comparisons. Following Goovaerts et al. (1997) and Bonilla et al. (2008), we therefore simplify the problem using the *intrinsic coregionalization model* (ICM), which imposes a low-rank assumption on the unit covariance k_{unit} (see also Flaxman et al., 2015, Section 4.2). While initially proposed in terms of restrictions on the matrices, we can equivalently write the ICM model as an assumption that the outcome model for each of the $i = 1, \ldots, N$, units is a unit-specific linear combination of J latent draws from a Gaussian process, i.e.,

$$f_{it} = \sum_{j=1}^{J} \beta_{ij} u_{jt}, \tag{5}$$

where each latent $\mathbf{u}_j$ is itself a GP with kernel k_{time},[3] and where each unit has its own set of unit-specific factor loadings $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iJ})$ with corresponding hyper-parameters, $\boldsymbol{\beta}_i \sim G(\cdot)$. Thus, the underlying model components $\mathbf{u}_1, \ldots, \mathbf{u}_J$ are shared across units and induce a dependence across them. The number of shared latent functions J corresponds to the rank of the unit kernel matrix $\mathbf{K}_{\text{unit}}$. This is a key hyper-parameter governing the overall complexity of the model; a higher rank J will lead to a more flexible model that potentially has the risk of overfitting. We discuss choosing the rank J via Bayesian model criticism in Section 5.1.

The resulting model is then:

$$Y_{it}(0) = f_{it} + \varepsilon_{it}$$

$$f_{it} = \sum_{j=1}^{J} \beta_{ij} u_{jt}$$

$$\mathbf{u}_j \sim \mathcal{GP}(0, k_{\text{time}}(t,t')),$$

---

[3]We can similarly write this restriction as $k_{\text{unit}}(i,i') = \langle \boldsymbol{\beta}_i, \boldsymbol{\beta}_{i'} \rangle$, i.e., a linear kernel between weights for units i and j over the J latent functions inferred from the data.

where $\varepsilon_{it} \sim N(0, \sigma^2)$. We estimate this model using the probabilistic programming language, Stan (Stan Development Team, 2021); see Appendix A for hyperprior distributions and additional implementation details.

As above, we can re-write this model in matrix notation. Define the unit covariance matrix as $\mathbf{K}_{\text{unit}} \in \mathbb{R}^{N \times N}$ and the time covariance matrix as $\mathbf{K}_{\text{time}} \in \mathbb{R}^{T \times T}$. Then we can represent the decomposition in Equation (4) as the *Kronecker product* between the two covariances:

$$\mathbf{K} \equiv \mathbf{K}_{\text{unit}} \otimes \mathbf{K}_{\text{time}} \in \mathbb{R}^{NT \times NT},$$

where any one entry of $\mathbf{K}^*$ can be written as the product of the task and time covariance, as in Equation (4).

We can similarly divide this (admittedly unwieldy) kernel matrix into a matrix for observed unit-time pairs, $\mathbf{K}_{\text{obs}} \in \mathbb{R}^{((n-1)T+T_0) \times ((n-1)T+T_0)}$, a matrix for the unobserved counterfactual post-treatment outcomes for the treated unit $\mathbf{K}_{\text{mis}} \in \mathbb{R}^{(T-T_0) \times (T-T_0)}$, and a rectangular matrix for the kernel between the outcomes for the observed unit-times and the post-treatment outcomes for the treated unit $\mathbf{K}_{\text{obs,mis}} \in \mathbb{R}^{((n-1)T+T_0) \times (T-T_0)}$. With this setup, the joint distribution of the missing post-treatment outcomes for the treated unit and the observed untreated outcomes are multivariate Normal:

$$\begin{pmatrix} \mathbf{Y}^{\text{mis}} \\ \mathbf{Y}^{\text{obs}} \end{pmatrix} \sim \text{MVN} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{\text{mis}} & \mathbf{K}_{\text{mis,obs}} \\ \mathbf{K}_{\text{obs,mis}} & \mathbf{K}_{\text{obs}} + \sigma^2 \mathbb{I} \end{pmatrix} \right].$$

The resulting conditional mean and variance have the same form as in Equations (2) and (3).

As we discuss in Section 4 below, there are two important special cases of this model, which correspond to setting either the time or unit covariances to be the identity. Setting the *time* covariance matrix to the identity, $\mathbf{K}_{\text{time}} = \mathbb{I}_T$, corresponds to a model with no smoothness in the underlying latent factors, $\mathbf{u}$, allowing the function value to change arbitrarily between time points. The resulting approach is a (Bayesian) linear factor model (e.g., Xu, 2017; Athey et al., 2021; Samartsidis et al., 2020). In Section 4 below, we further connect this to "vertical regression," which finds a set of weights over units that optimize some imbalance criterion (Doudchenko and Imbens, 2016). On the other hand, setting the *unit* covariance matrix to the identity, $\mathbf{K}_{\text{unit}} = \mathbb{I}_N$, corresponds to a model in which there is no correlation in the underlying process for each unit, and each model component is an independent GP. The model therefore reduces back to the single-task GP discussed in Section 3.1 above. Finally, while we focus on the ICM model with separate time and unit covariances, there is a large literature on more elaborate generalizations, such as the *semiparametric latent factor model* and the *linear model of coregionalization*; see Álvarez (2017) for a review.

3.3 Extensions

Building on the broad GP literature, we can immediately extend the MTGP model above to better match our application. We focus on three main extensions here: incorporating a mean model; allowing for a non-Gaussian likelihood; and modeling multiple outcomes simultaneously.

3.3.1 Incorporating a mean model

Thus far, we have focused on mean-zero Gaussian Processes, $Y_{it}(0) = f_{it} + \varepsilon_{it}$. We can extend this to incorporate a mean model, m_{it}:

$$Y_{it}(0) = \underbrace{m_{it}}_{\text{mean model}} + \underbrace{f_{it}}_{\text{MTGP}} + \underbrace{\varepsilon_{it}}_{\text{noise}}.$$

A particularly natural model for the panel data setting is to set m_{it} to have unit- and time-specific intercepts:

$$m_{it} = \mu + \text{unit}_i + \text{time}_t$$
$$\text{unit}_i \sim N(0, \sigma_{\text{unit}}^2)$$
$$\textbf{time} \sim \mathcal{GP}(0, \textbf{K}_{\text{global}})$$

where μ is a global intercept, unit_i are the unit-specific intercepts, and time_t are the time-specific intercepts. Here we place independent Normal priors on the unit-specific intercepts,[4] and a global Gaussian Process prior on the time-specific intercepts.

The distinction between the mean function and the covariance matrix in the MTGP is somewhat artificial, and is typically chosen for clarity and to facilitate the choice of prior (Kanagawa et al., 2018). To see this, consider a simplified MTGP in which the mean model only includes unit-specific intercepts, $m_{it} = \text{unit}_i$. This is equivalent to a mean-zero MTGP with kernel function:

$$k(\ (i,t),(i',t')\ ) = k_{\text{delta}}(i,i')\ +\ k_{\text{unit}}(i,i')\ \times\ k_{\text{time}}(t,t'),$$

where k_{delta} is the *delta* kernel:

$$k_{\text{delta}}(i,i') = \begin{cases} 1 & \text{if}\ \ i = i' \\ 0 & \text{otherwise} \end{cases}.$$

Including the unit-specific intercepts in the mean function, however, allows us to directly specify reasonable priors and makes the role of these intercepts more transparent.

---

[4]With independent priors, these are often referred to as *fixed effects* in the panel data literature. We could instead model *random effects* using a hierarchical prior. See Hazlett and Wainstein (2020) for discussion.

Finally, we can immediately extend the mean model to incorporate (time invariant) auxiliary covariates, $X_i \in \mathbb{R}^p$:[5]

$$m_{it} = \mu + \text{unit}_i + \text{time}_t + \eta' X_i,$$

where η is a vector of regression coefficients. As with the unit-specific intercepts, we could instead include this as part of the GP kernel, for example by incorporating it into the unit covariance.

3.3.2 Non-Gaussian Likelihoods and heteroskedasticity

Another natural GP extension is to allow for non-Gaussian observation models, which is essential for well calibrated uncertainty quantification. Analogous to using alternative link functions with generalized linear models (Gelman et al., 2013, Ch. 21.3), we can model the control potential outcomes as:

$$Y_{it}(0) \sim g^{-1}(f_{it}),$$

where $g^{-1}(\cdot)$ is an appropriate link function and f_{it} is the latent GP factor. As above, we could also incorporate a mean model, $Y_{it}(0) \sim g^{-1}(m_{it} + f_{it})$. See Naish-Guzman and Holden (2008); Hensman et al. (2015) for discussions of related computational issues.

We highlight two link functions here. First, the outcome in our application is the number of murders, which is more naturally modeled via a Poisson (or, perhaps, Negative Binomial) link. A Poisson model also captures the restriction that the variance in the murder rate increases with the mean murder rate. We use this as our primary model below. Alternatively, we could allow for a "robust" Gaussian Process by allowing for t-distributed rather than Normal errors; see Jylänki et al. (2011). In our application, the estimates using t-distributed errors are nearly identical to those using Normal errors.

Finally, an important practical feature of our application is that populations vary considerably across states. Thus, we expect that the underlying variability in the murder rate will be much larger in, say Vermont or Wyoming than in California or Texas. With a Gaussian likelihood, we can parametrize this as:

$$Y_{it}(0) \sim N\left(f_{it}, \frac{\sigma^2}{N_{it}}\right)$$

where N_{it} is the population for state i at time t. We can similarly allow for a population offset in a Poisson model.[6] Allowing for heteroskedasticity by size is an important departure from many Frequentist panel data methods used in this setting; see Samartsidis et al. (2019).

---

[5]With additional restrictions (e.g., "exogeneity"), we can extend this to include time-varying covariates as well. See Imai and Kim (2019) for discussion.

[6]More broadly, we could incorporate much more complex noise models, such as an additional GP on the noise term (e.g., Naish-Guzman and Holden, 2008; Hensman et al., 2015). This is a challenging model to fit given our limited sample size.

3.3.3 Multiple outcomes

Finally, we can extend the MTGP approach to allow for multiple correlated outcomes, following the large literature on multi-output GPs (Álvarez, 2017); see also Samartsidis et al. (2020) for a non-GP setting. This is particularly important for our application because we are interested in assessing the impact of APPS on both gun- and non-gun-related homicides.

To do so, we can write each data point as a unit-time-outcome triple, (i, t, j), with control potential outcome, $Y_{itj}(0)$. While many models are possible, we focus on a simple version with a shared factor structure across outcomes, which we can write as a multivariate Normal observation model with common covariance across units and time:

$$\mathbf{Y}_{it}(0) \sim \text{MVN}(f_{it}, \Sigma).$$

Equivalently, we can write this as an MTGP where the resulting kernel decomposes into separate unit, time, and outcome kernels:

$$k(\ (i, t, j), (i', t', j')\ ) = k_{\text{unit}}(i, i')\ \times\ k_{\text{time}}(t, t')\ \times\ k_{\text{outcome}}(j, j').$$

Similarly, we can write this as a Kronecker structured overall kernel $\mathbf{K} = \mathbf{K}_{\text{unit}} \otimes \mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{outcome}}$, as in Flaxman et al. (2015).

4 MTGP as a weighting estimator

Our discussion thus far has been entirely Bayesian. Gaussian Processes, however, are fundamentally linked with (Frequentist) kernel ridge regression; a large literature exploits this connection to describe the Frequentist property of GPs; see Kanagawa et al. (2018) for a recent review. Here we adapt those results to our panel data setting. We first represent the MTGP estimate as a weighting estimator with separate unit and time weights, and give error bounds under fairly general conditions on the latent (noiseless) outcomes. We then describe several special cases to better illustrate this result.

4.1 Weighting representation

We now show that the MTGP estimate with separable unit and time kernels has a corresponding representation as a weighting estimator with separate unit and time weights. This setup differs from our fully Bayesian formulation above; to make this connection with weighting, we instead consider a Frequentist setup that assumes a known kernel. Specifically, conditional on hyperparameters, the MTGP estimate can be written as the solution to a constrained optimization problem that minimizes the worst-case mean-square error across the "noiseless" latent functions, f. This representation exploits the connection between Gaussian Processes and kernel ridge regression

(Kanagawa et al., 2018); see Hazlett and Xu (2018) for additional discussion of kernel weighting methods for panel data.

To set up the problem, we again assume that (control) potential outcomes are the sum of structural and error components, $Y_{it}(0) = f_{it} + \varepsilon_{it}$, where ε_{it} is mean-zero noise. We assume that f is "well behaved" in the sense of belonging to a Reproducing Kernel Hilbert Space (RKHS); this setup is quite flexible and incorporates a wide range of function classes (see, e.g., Wainwright, 2019). We focus on the ICM kernel, $\mathbf{K} = \mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}}$, where, as in other Frequentist analyses, the kernel is assumed known.

We can then represent the MTGP as a weighting estimator. We consider two basic forms, corresponding to the weight-space and function-space interpretations of Gaussian processes (Williams and Rasmussen (2006), Ch. 2). We first show that the MTGP minimizes the worst-case imbalance in the noiseless latent functions f. We then show the equivalent formulation in terms of imbalance on the observed outcomes Y.

Proposition 4.1. Let $Y_{it}(0) = f_{it} + \varepsilon_{it}$, where f is a fixed function. Let $\mathcal{H}_k$ be the RKHS for kernel k with Hilbert-Schmidt norm, $\|f\|_{\mathcal{H}_k}^2 = k(f, f)$. Then let f be contained in the unit ball of the reproducing Hilbert space, $\|f\|_{\mathcal{H}_k} \leq 1$, where $k(\ (i,t),(i',t')\ ) = k_{\text{unit}}(i, i') \times k_{\text{time}}(t, t')$, and ε_{it} are iid mean-zero random variables with observed variance $\sigma^2 > 0$. Further, model the time covariance $\mathbf{K}_{\text{time}}$ with a squared exponential kernel with lengthscale ρ, and model the unit covariance $\mathbf{K}_{\text{unit}} = \beta\beta^\top$. Finally, let $\gamma^{(i)} \in \mathbb{R}^N$ and $\lambda^{(t)} \in \mathbb{R}^T$ be the vectors of unit and time weights for target unit i at time t, and let γ and λ, respectively, be the set of weights across all targets. Then:

(a) The unit and time weights, γ and λ, have the following closed form:

$$\gamma^{(i)} \otimes \lambda^{(t)} = \left(\mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}} + \sigma^2 \mathbb{I}\right)^{-1} \left(k_{\text{time}}(t, \cdot) \otimes k_{\text{unit}}(i, \cdot)\right). \tag{6}$$

The posterior (predictive) mean estimate for target observation $(1, t^*) \notin \mathcal{C}$, for $t^* > T_0$ is:

$$\hat{\mu}_{1t^*} = \sum_{(i,t) \in \mathcal{C}} \gamma_i^{(1)} \lambda_t^{(t^*)} Y_{it}.$$

(b) The unit and time weights also solve the following, equivalent optimization problems, in terms of (1) the weight-space formulation:

$$\min_{\gamma, \lambda} \ \sup_{f \in \mathcal{H}_k} \ \sum_{(i,t) \in \mathcal{C}} \left(f_{it} - \sum_{(i',t') \in \mathcal{C}} \gamma_{i'}^{(i)} \lambda_{t'}^{(t)} f_{i't'} \right)^2 \ + \ \sigma^2 \|\gamma\|_2^2 \|\lambda\|_2^2, \tag{7}$$

16

and (2) the function-space formulation:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\xi}} \quad \sum_{(i,t)\in\mathcal{C}} \left(Y_{it} - \sum_{(i',t')\in\mathcal{C}} \underbrace{\alpha_{i'}^{(i)} k_{\text{unit}}(i,i')}_{\gamma_{i'}^{(i)}} \cdot \underbrace{\xi_{t'}^{(t)} k_{\text{time}}(t,t')}_{\lambda_{t'}^{(t)}} Y_{i't'} \right)^2 \quad + \quad \sigma^2 \left(\boldsymbol{\xi} \otimes \boldsymbol{\alpha}\right)^{\top} \mathbf{K} \left(\boldsymbol{\xi} \otimes \boldsymbol{\alpha}\right),$$

(8)

where $\mathbf{K} = \mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}}$ and where $\alpha^{(i)} \in \mathbb{R}^N$ and $\xi^{(t)} \in \mathbb{R}^T$ are the coefficient vectors for target unit i at time t, with $\gamma_{i'}^{(i)} = \alpha_{i'}^{(i)} k_{\text{unit}}(i,i')$ and $\lambda_{t'}^{(t)} = \xi_{t'}^{(t)} k_{\text{time}}(t,t')$, and with corresponding sets $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$.

(c) This estimate has the following out-of-training-sample estimation error for the structural component of the missing potential outcome f_{1t^*},

$$|f_{1t^*} - \hat{\mu}_{1t^*}| \leq \left[\underbrace{\left(k_{\text{time}}(t^*,\cdot)^{\top} \lambda^{(t^*)} \otimes k_{\text{unit}}(1,\cdot)^{\top} \gamma^{(1)} \right)}_{\text{error for } f_{1t^*}} + \underbrace{\sigma^2 \left\| \gamma^{(1)} \right\|_2^2 \left\| \lambda^{(t^*)} \right\|_2^2}_{\text{irreducible noise}} \right]^{1/2}.$$

(9)

Proposition 4.1 begins with the algebraic result that the posterior mean estimate of the MTGP can be written as a weighting estimator, and that these overall weights separate into unit- and time-weights. This separation does not hold in general, but, in our setting with an ICM kernel, follows immediately from the restriction that the unit and time covariances are separable (Bonilla et al., 2008). As a result, we can view our proposed MTGP approach as a doubly-weighted panel data estimator, similar to recent proposals from e.g. Ben-Michael et al. (2021a) and Arkhangelsky et al. (2019) that also impute the counterfactual as a double weighted average. Importantly, the MTGP implicitly estimates unit and time weights *simultaneously*, while previous proposals construct them separately.

Proposition 4.1 then leverages Frequentist results on GPs to provide additional intuition for these weights (see Kanagawa et al., 2018). First, Equation (7) shows that the MTGP weights minimize the error on the noiseless outcome functions, subject to regularization on the weights. Importantly, this formulation does not rely on parametric assumptions, only that the noiseless outcomes, f_{it}, are relatively "well behaved" in the sense of belonging to an RKHS. Equivalently, Equation (8) shows that the weights can be seen as the solution to an equivalent optimization problem in terms of minimizing imbalance in the observed outcomes directly, but subject to regularization on the "coefficients" α and ξ, rather than on the weights themselves. Finally, building off this, Equation (9) shows that the error bound depends directly on the distance between the control observations, $\mathcal{C}$, and the unobserved point, $(1, t^*)$. Specifically, with a squared exponential kernel, $k_{\text{time}}(t^*, \cdot)$, the bound is increasing for $t^* > T_0$. See Fiedler et al. (2021) for recent generalizations

of such bounds under model mis-specification.

Outcome model. We can incorporate a (prior) outcome model into the posterior mean estimate, echoing several recent proposals that combine outcome modeling and weighting in the panel data setting (see, e.g. Ben-Michael et al., 2021a; Ferman and Pinto, 2021; Arkhangelsky and Imbens, 2021):

$$
\hat{\mu}_{1t^*}^{\text{aug}} = \sum_{(i,t)\in\mathcal{C}} \gamma_i^{(1)}\lambda_t^{(t^*)}Y_{it} + \underbrace{\left(m_{1t^*} - \sum_{(i,t)\in\mathcal{C}} \gamma_i^{(1)}\lambda_t^{(t^*)}m_{it}\right)}_{\text{estimated bias}} \tag{10}
$$

$$
= m_{1t^*} + \underbrace{\sum_{(i,t)\in\mathcal{C}} \gamma_i^{(1)}\lambda_t^{(t^*)}(Y_{it} - m_{it})}_{\text{re-weighted residuals}}, \tag{11}
$$

where m_{it} is the outcome (prior mean) model for unit i at time t. Equation (10) has a form similar to bias correction for matching (Rubin, 1973), which corrects the doubly-weighted average by an estimate of the difference in the outcome model m_{it}. Equation (11) instead has the form of augmented inverse propensity score weighting (Robins, 1997), which re-weights the outcome model residuals.

4.2 Special cases

To help build intuition for the weighting representation, we consider two special cases: (1) the unit kernel $\mathbf{K}_{\text{unit}}$ is the identity; and (2) the time kernel $\mathbf{K}_{\text{time}}$ is the identity matrix.

Identity unit covariance. First, let the unit covariance $\mathbf{K}_{\text{unit}}$ be the identity matrix. We can view this as the setting where there is no correlation between the underlying processes for each unit, so each model component is an independent GP. In this setting, there is nothing to be learned from the comparison units, and the MTGP estimate only compares to the pre-treatment outcomes of the treated unit. In particular, following the setup in Proposition 4.1, if we set the unit covariance matrix $\mathbf{K}_{\text{unit}}$ to be the identity, then the weights for post-treatment time period $t^* > T_0$ are

$$
\lambda^{(t^*)} = (\mathbf{K}_{\text{time}} + \sigma\mathbb{I}_{T_0})^{-1}k_{\text{time}}(t^*, \cdot),
$$

with posterior mean estimate:

$$
\hat{\mu}_{1t^*}^{\text{horiz}} = \sum_{t=1}^{T_0} \lambda_t^{(t^*)}Y_{1t}.
$$

This is a *horizontal regression* formulation (see Athey et al., 2021). As these weights ignore the outcomes for the comparison units, we can view this case as fitting N separate single-task GP

18

models, one for each unit. Thus, when the unit covariance is the identity matrix, this problem ignores the outcomes for comparison units and reduces to the single-task GP for the treated unit discussed in Section 3.1.

Identity time kernel. Next, let the time kernel $\mathbf{K}_{\text{time}}$ be the identity matrix. This corresponds to the setting where there is no smoothness at all in the underlying latent functions, so the function value can change wildly between time periods. In this case, it is impossible to use outcomes from different time periods to inform our estimates, and so only the unit weights remain. Again following the setup in Proposition 4.1, if the time covariance matrix $\mathbf{K}_{\text{time}}$ is the identity, then the weights for the treated unit are:
$$\gamma^{(1)} = (\mathbf{K}_{\text{unit}} + \sigma \mathbb{I}_N)^{-1} k_{\text{unit}}(1, \cdot),$$
with posterior mean estimate:
$$\hat{\mu}_{1t^*}^{\text{vert}} = \sum_{i=2}^{N} \gamma^{(1)} Y_{it^*}.$$

This is a *vertical regression* setup (see Doudchenko and Imbens, 2016), and corresponds to the implicit weights of a linear factor model or interactive fixed effects model (Gobillon and Magnac, 2016; Xu, 2017; Athey et al., 2021). The estimator only uses the learned unit covariance to implicitly construct weights over comparison units.

5   Model diagnostics and estimated impact for APPS

We now use the MTGP framework to estimate the impact of APPS on homicides. We begin by using posterior predictive checks to guide the choice of model. We then estimate the impacts both on overall homicides and separately on gun- and non-gun-related homicides.

5.1   Posterior Predictive Checks and other model diagnostics

Diagnostics and other forms of model criticism are important components of the Bayesian workflow, helping researchers diagnose and compare the fit of the proposed models (Gelman et al., 2013). We begin with *posterior predictive checks* (PPCs). First, we choose a test statistic, $T(\text{data}, \theta)$, which reflects a relevant aspect of model fit and is a function of both data and model parameters. We then compare the posterior predictive distribution of this statistic, $T(\text{data}_{\text{rep}}, \theta)$, to the value of the test statistic for the observed data, $T(\text{data}_{\text{obs}}, \theta)$, over draws of θ.

PPCs play an especially important role in panel data settings like ours, where it is often difficult to decide between modeling approaches given limited data. In particular, posterior predictive checks avoid many of the inferential issues associated with, for example, testing for parallel trends prior to estimating a difference-in-differences model (Roth, 2019). See Liu et al. (2020) for a review of Frequentist approaches to model checks for panel data, including equivalence and placebo tests.

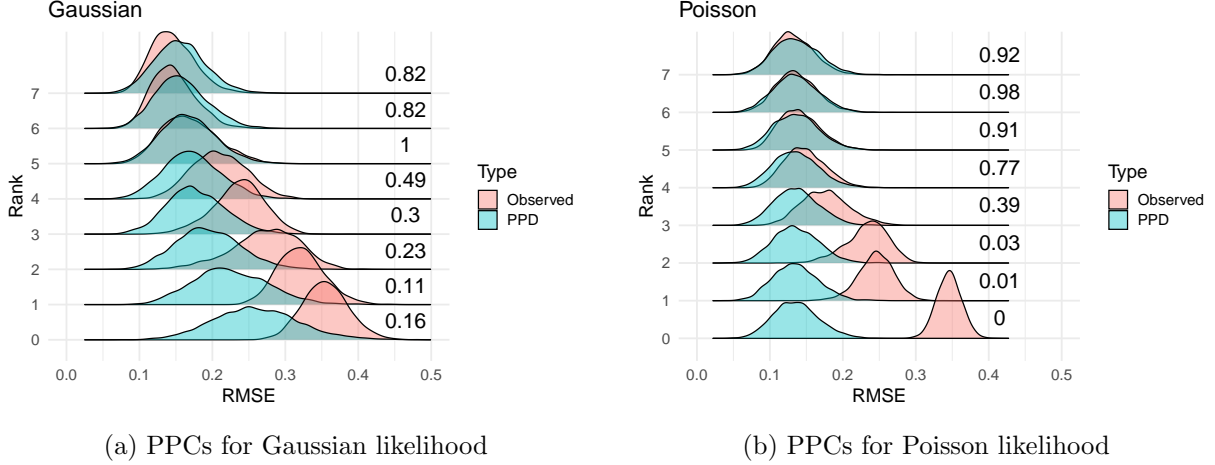(a) PPCs for Gaussian likelihood

(b) PPCs for Poisson likelihood

Figure 2: Posterior Predictive Checks (PPCs) for the pre-treatment RMSE for California. Numbers indicate posterior predictive p-value associated with each rank, i.e., $\mathbb{P}\left(T(\text{data}_{\text{rep}}, \theta) > T(\text{data}_{\text{obs}}, \theta) \mid \text{data}\right)$.

We focus on two common measures of model fit: the overall pre-treatment fit and the pre-treatment imbalance at each time point. First, we evaluate the pre-treatment fit for California by comparing the posterior predictive distribution of the root mean squared error (RMSE),

$$T(\mathbf{Y}^*, \theta) = \frac{1}{T_0} \sqrt{\sum_{t=1}^{T_0} (Y_{1t}^* - f_{1t})^2},$$

to the observed RMSE, $T(\mathbf{Y}, \theta) = \frac{1}{T_0} \sqrt{\sum_{t=1}^{T_0} (Y_{1t} - f_{1t})^2}$, where Y_{1t}^* is a posterior predictive draw of the overall murder rate for California at (pre-treatment) time t. Figure 2 shows the distributions for both Gaussian and Poisson observation models with ranks $J = 0, \ldots, 7$. For each combination, we also compute the posterior predictive p-value, the fraction of posterior predictive test statistics that are larger than the observed test statistic. For both Gaussian and Poisson models, the observed RMSE is larger than we would expect for ranks 0 through 3, with minimal discrepancies by rank 5 for both models. The Poisson model has slightly lower RMSE, suggesting that Poisson with rank 5 is a reasonable initial choice here.

Figure 3 shows the imbalance at each pre-treatment time point for the Poisson MTGP model with rank 5. The upper pane shows the observed data (in red) as well as posterior predictive draws from the model on the scale of the original outcome (homicides per 100,000). The lower pane instead shows the "gap" between the outcome and the posterior mean, $T(\text{data}, \theta) = Y_{1t} - \hat{\mu}_{1t}$.[7] Both plots show little difference between the observed data and the posterior predictive distribution,

---

[7]This is the Bayesian analog to estimates of placebo impacts that are standard in Frequentist panel data studies, although we compare the observed estimates to the posterior predictive distribution rather than to zero.
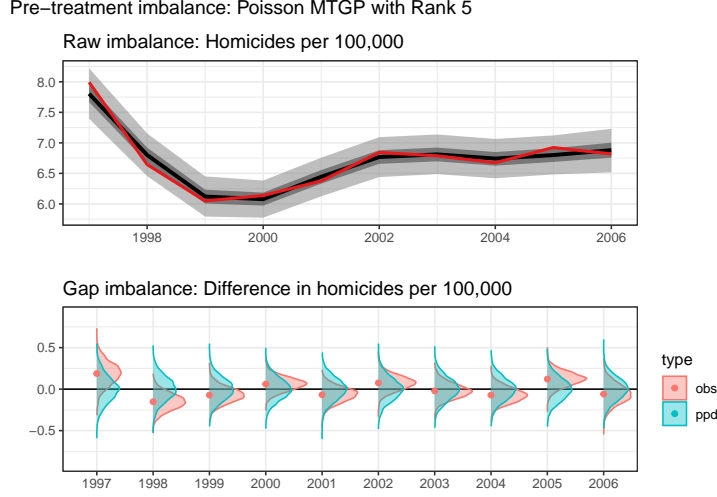
Figure 3: Pre-treatment imbalance for Poisson model with Rank 5. Observed data is in red.

suggesting reasonable model fit along the dimensions we investigate here.

Finally, the Appendix includes several additional model diagnostics. Appendix Figure C.2 shows the coverage of the 50% and 95% posterior predictive intervals for pre-treatment outcomes across all 50 states, computed as $\frac{1}{NT} \sum_{i,t} I(L_{it}^q < Y_{it} < U_{it}^q)$ where L_{it}^q and U_{it}^q are the lower and upper endpoints of the credible region with probability q for unit i at time t. The Poisson intervals have close to nominal coverage for ranks 4 and 5, but under-cover for lower ranks and over-cover for higher ranks. By contrast, the corresponding Gaussian intervals over-cover across all ranks, with especially poor coverage for the 50% intervals for higher ranks, leading further credence to the Poisson model over the Gaussian model. This difference is likely due to the fact that the sampling variance increases with the mean in the Poisson model but is constant in the mean and across units for the Gaussian model.

In Appendix Figures C.3 and C.4 we show the inferred Gaussian Process for the time-specific intercept and the estimated factors from the rank 5 Poisson model next to the factor loadings for California. In Appendix Figures C.5 and C.6, we show the inferred time and unit weights for California in the first post-treatment period using a rank 5 Gaussian multitask Gaussian process without the global time series trend GP. Each of the time and unit weights are obtained by marginalizing over the non-relevant component (unit, or time, respectively) for each sample from the MCMC sampler.

5.2   Impact of APPS on homicides

We now turn to estimating the impact of APPS on homicides in California. We begin with the overall impact and then disaggregate by gun and non-gun-related homicides.
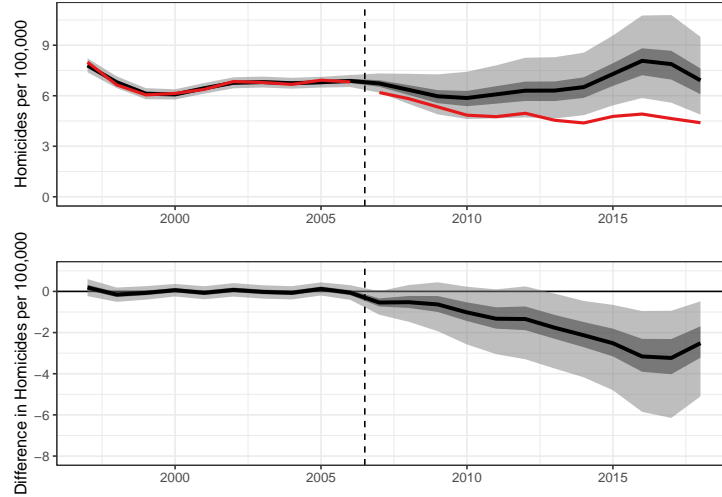
Figure 4: Impact of APPS on homicides per 100,000 in California by year. The observed data is show in red. The black line is the posterior mean estimate; the dark and light gray regions are, respectively, the 50% and 95% posterior predictive intervals.

Overall impact of APPS. Based on our model diagnostics above, we focus on the MTGP estimated with rank $J = 5$, Poisson observation model, and unit- and time-specific intercepts. Figure 4 shows the estimated impact of APPS on homicide rates in California over time.[8] The upper pane shows the posterior predictive distribution from the MTGP, with the posterior mean as well as 50% and 95% posterior predictive intervals. The red line shows the observed murder rate for California. The estimates pre-treatment repeat the pre-treatment imbalance checks in Figure 3; the counterfactual estimates post-treatment are consistently higher than the observed murder rate. The lower pane shows this difference explicitly, where zero is no impact. These estimates suggest that APPS reduced murders in the state, with both impacts and uncertainty growing over time.

Figure 5 shows the posterior estimates for the average annual impact. The posterior mean estimate is roughly 1.4 murders avoided per 100,000 people per year, with a 95% credible interval of $(0.5, 3.6)$ murders per 100,000 people. To contextualize these estimates, California's annual murder rate in 2004 and 2005 was 6.8 per 100,000. Focusing on the smaller end of the 95% credible interval suggests a decline of at least 8 percent from baseline. We can also compare the annual number of murders prevented by this system to expenditures for APPS investigative teams. In 2018, California's population was 39.6 million. An increment in the murder rate of

---

[8]Appendix Figure C.10 shows the corresponding estimates after adjusting for auxiliary covariates and several measures of crime, all for 2005. The auxiliary covariates are: the prison incarceration rate, the age distribution (percent 0-17, 18-24, 25-44, 45-64, 65 and older), the percent Black, the unemployment rate, the poverty rate, and log median income. The crime measures, normalized by population, are: overall violent crime, rape, robbery, assault, property crime, burglary, larceny, and motor/vehicular crime. As the results are largely unchanged, we focus on the unadjusted estimate here.

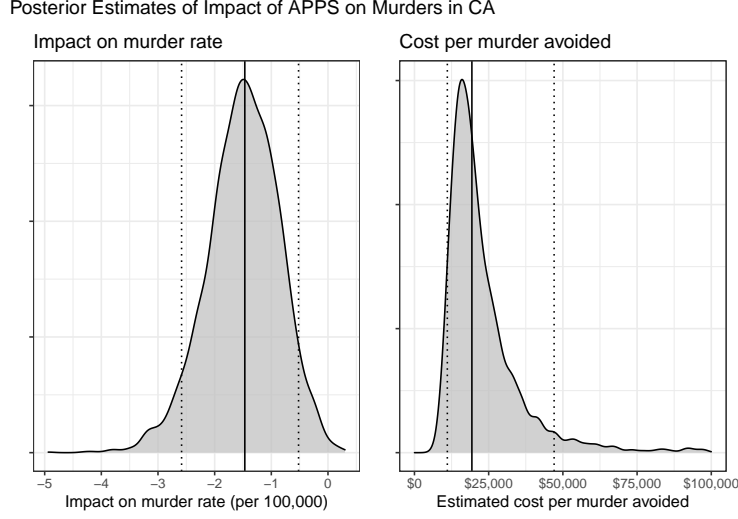Posterior Estimates of Impact of APPS on Murders in CA

Figure 5: (left) Posterior distribution of the average impact of APPS on homicides per 100,000 in California for 2007 to 2017; (right) Posterior distribution of the cost per murder avoided, with a state population of 39.6 million and annual budget expenditure of \$11.3 million. The solid line denotes the posterior mean; the dashed lines denote 95% credible intervals.

0.5 corresponds to nearly 200 murders avoided per year. Total funding for APPS investigation teams in fiscal year 2017-2018 stood at \$11.3 million (Petek, 2019). For this range of estimates of homicide prevented, a (conservative) back-of-the-envelope calculation suggests expenditures of around \$50,000 per murder prevented. This value is considerably lower than conventional value-of-a-statistical-life (VSL) estimates, which tend towards \$10 million per life saved (Heaton, 2010; Domínguez and Raphael, 2015).[9]

Separate impacts on gun- and non-gun-related homicides. Substantively, an important check is to estimate separate impacts of APPS on gun- and non-gun-related. In particular, we expect the impact of APPS to be entirely on gun-related homicides, with negligible impacts on non-gun homicides.[10] Following Ben-Michael et al. (2021), we also apply our approach to a different data source that disaggregates overall homicides by whether a firearm was used (Kaplan, 2019).[11]

---

[9]Given the pain and anguish suffered by murder victims and their loved ones, the VSL estimate greatly undervalues the true costs of homicide. To be sure, the actual costs of the program are likely higher as other units within the California Department of Justice, as well as other law enforcement agencies throughout the state, contribute resources to the enforcement efforts carried out via the APPS program. Nonetheless, this program has a very high benefit-cost ratio based these back-of-the-envelope calculations.

[10]In fact, if there is substitution across weapons associated with relative cost and availability among those determined to commit murder, we might hypothesize that APPS may have increased homicide rates where firearms are not used.

[11]In particular, we use the Kaplan (2019) concatenation of the UCR Supplementary Homicide Files, restricted to our study period (1997 through 2016). We first use the incident level data to calculate totals by state and year of the number of murders involving a firearm and the number of murders where a gun is not used. We then use these totals in conjunction with state level population counts to tabulate rates per 100,000. In general, gun murder rates
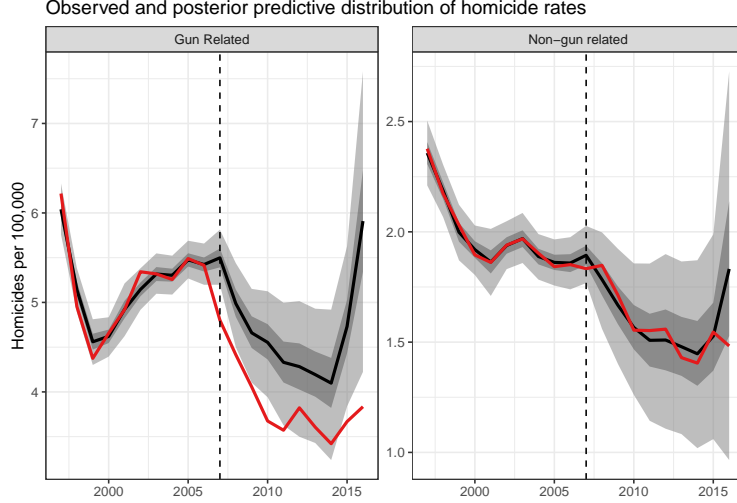
Figure 6: Impact of APPS on gun- and non-gun-related homicides per 100,000 in California by year. The observed data is show in red. The black line is the posterior mean estimate; the dark and light gray regions are, respectively, the 50% and 95% posterior predictive intervals.

Figure 6 shows estimates for gun- and non-gun-related homicides, estimated via a joint multitask GP. Fitting both outcomes jointly is advantageous since these two outcomes are positively correlated due to shared latent factors. (See Appendix Figure C.11, for the posterior predictive correlation between gun and non-gun related homicides for California). While estimates are derived from a different data source than used above, the impacts for gun-related homicides are nearly identical to the estimated overall impacts in Figure 5. In contrast, the estimated impact of APPS on non-gun-related homicides is essentially zero.

6   Discussion

This paper uses a multitask Gaussian Processes framework in a panel data setting to estimate the causal effect of an innovative gun control program in California. MTGPs find a natural middle ground between pure "horizontal" time series forecasting models, based on extrapolating the treated unit's outcomes series alone, and pure "vertical" regression models, based on a weighted average of control units' outcomes. Further, MTGPs automatically quantify the relevant sources of uncertainty, providing a unified Bayesian framework in a setting where uncertainty quantification is notoriously challenging. They also allow for extensive model diagnostics via posterior predictive checks. Building on a robust GP literature, there are also many immediate extensions to the standard MTGP setup, including allowing for multiple outcomes and incorporating a mean model. Using this flexible approach, we find large effects of APPS on homicides in California, with program

---

are two to three times that of non-gun murder rates. See Ben-Michael et al. (2021) for additional discussion.

benefits far exceeding costs under typical assumptions.

There are several promising directions for future research. First, there are many existing results for Gaussian Processes that would apply immediately in other settings but are not as relevant in our application. For instance, standard panel data models face a range of practical challenges when the data structure deviates from a so-called "balanced panel," either due to irregularly sampled observations over time or due to missing data (see Imai and Kim, 2019). The MTGP framework naturally accommodates both irregularly sampled time points and missing values, and, unlike most Frequentist approaches, will automatically propagate the corresponding uncertainty.

We can also extend this approach to allow for multiple treated units that adopt treatment over time, also known as staggered adoption (e.g., Ben-Michael et al., 2021b). First, we can immediately extend to the case where multiple treated units adopt treatment at the *same* time (i.e., simultaneous adoption) by changing the target to the *average* of the treated units. However, the extension to the case when units adopt over time (i.e., staggered adoption) is slightly more complicated. For instance, we could adapt the model to have a separate kernel for each treatment time, possibly including a hierarchical structure on the kernels (Flaxman et al., 2015).

Next, we can develop a framework for assessing sensitivity to departures from key modeling assumptions for the MTGP approach, especially the restriction that the time and unit kernels are separable. While there are many possible generalizations (see, for example Álvarez, 2017), these are inherently under-specified in settings like ours with relatively few units and time periods. Thus, a sensitivity analysis approach in the spirit of Franks et al. (2019) is a natural avenue for appropriately quantifying uncertainty in this case.

Finally, our proposed MTGP approach implicitly puts the same weight on the error for California as on other states, even though California is the target unit. We can explore alternative approaches for parameterizing similarity across units and time, such as by imposing an "arrowhead" structure on the overall covariance, or by learning a richer dependence structure across units; see, for example, Li and Kontar (2020).

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). 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.

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

Álvarez, M. (2017). Multi-output gaussian processes. Gaussian Process Summer School 2017.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2019). Synthetic difference in differences. Technical report, National Bureau of Economic Research.

Arkhangelsky, D. and G. W. Imbens (2021). Double-robust identification for causal panel data models. Technical report, National Bureau of Economic Research.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 1–15.

Ben-Michael, E., A. Feller, and S. Raphael (2021). The Effect of a Targeted Effort to Remove Firearms from Prohibited Persons on State Murder Rates.

Ben-Michael, E., A. Feller, and J. Rothstein (2021a). The augmented synthetic control method. *Journal of the American Statistical Association* (just-accepted), 1–34.

Ben-Michael, E., A. Feller, and J. Rothstein (2021b). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society (Series B)*.

Bonilla, E. V., K. M. Chai, and C. Williams (2008). Multi-task gaussian process prediction. In *Advances in Neural Information Processing systems*, pp. 153–160.

Branson, Z., M. Rischard, L. Bornn, and L. W. Miratrix (2019). A nonparametric bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference 202*, 14–30.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott (2015). Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics 9*, 247–274.

California Department of Justice (2015). Legislative Report Number One – Calendar Year 2014: Armed and Prohibited Persons System. Sacramento, CA.

California Department of Justice (2019). Legislative Report Number One – Armed and Prohibited Persons System SB140 Legislative Report Number Five – Calendar Year 2018. Sacramento, CA.

Carlson, D. (2020). Estimating a counter-factual with uncertainty through gaussian process projection.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 1–16.

Ding, P. and F. Li (2018). Causal inference: A missing data perspective. *Statistical Science 33*(2), 214–237.

Domínguez, P. and S. Raphael (2015). The role of the cost-of-crime literature in bridging the gap between social science research and policy making: Potentials and limitations. *Criminology & Public Policy 14*(4), 589–632.

Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Ferman, B. and C. Pinto (2021). Synthetic controls with imperfect pre-treatment fit. *Quantitative Economics*.

Fiedler, C., C. W. Scherer, and S. Trimpe (2021). Practical and rigorous uncertainty bounds for gaussian process regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 35, pp. 7439–7447.

Flaxman, S., A. Wilson, D. Neill, H. Nickisch, and A. Smola (2015). Fast kronecker inference in gaussian processes with non-gaussian likelihoods. In *International Conference on Machine Learning*, pp. 607–616. PMLR.

Franks, A., A. D'Amour, and A. Feller (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

Gobillon, L. and T. Magnac (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics 98*(3), 535–551.

Goovaerts, P. et al. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

Gramlich, J. (2019). What the Data Says About Gun Deaths in the U.S. Pew Research Center https://www.pewresearch.org/fact-tank/2019/08/16/what-the-data-says-about-gun-deaths-in-the-u-s/, Accessed on October 10, 2019.

Gustafson, P. (2010). Bayesian inference for partially identified models. *The International Journal of Biostatistics 6*(2).

Hazlett, C. and L. Wainstein (2020). Understanding, choosing, and unifying multilevel and fixed effect approaches. *Political Analysis*, 1–20.

Hazlett, C. and Y. Xu (2018). Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data. *Available at SSRN 3214231*.

Heaton, P. (2010). *Hidden in plain sight: What cost-of-crime research can tell us about investing in police*, Volume 279. Rand Corporation.

Hensman, J., A. Matthews, and Z. Ghahramani (2015). Scalable variational gaussian process classification.

Huang, B., C. Chen, and J. Liu (2019). Gpmatch: A bayesian doubly robust approach to causal inference with gaussian process covariance function as a matching tool. *arXiv preprint arXiv:1901.10359*.

Huang, B., K. Zhang, and B. Schölkopf (2015). Identification of time-dependent causal model: A gaussian process treatment. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Imai, K. and I. S. Kim (2019). On the use of two-way fixed effects regression models for causal inference with panel data.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jylänki, P., J. Vanhatalo, and A. Vehtari (2011). Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research 12*(Nov), 3227–3257.

Kanagawa, M., P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Kaplan, J. (2019). Jacob Kaplan's Concatenated Files: Uniform Crime Reporting (UCR) Program Data: Supplementary Homicide Reports, 1976-2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-07-15. https://doi.org/10.3886/E100699V7.

Karch, J. D., A. M. Brandmaier, and M. C. Voelkle (2018). Gaussian process panel modeling-kernel-based longitudinal modeling.

Kim, S., C. Lee, and S. Gupta (2020). Bayesian synthetic control methods. *Journal of Marketing Research 57*(5), 831–852.

Li, M. and R. Kontar (2020). On negative transfer and structure of latent functions in multi-output gaussian processes. *arXiv preprint arXiv:2004.02382*.

Liu, L., Y. Wang, and Y. Xu (2020). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *Available at SSRN 3555463*.

Menchetti, F. and I. Bojinov (2020). *Estimating causal effects in the presence of partial interference using multivariate Bayesian structural time series models*.

Modi, C. and U. Seljak (2019). Generative learning of counterfactual for synthetic control applications in econometrics. *arXiv preprint arXiv:1910.07178*.

Naish-Guzman, A. and S. Holden (2008). The generalized fitc approximation. In *Advances in neural information processing systems*, pp. 1057–1064.

Oganisian, A. and J. A. Roy (2020). A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches. *arXiv preprint arXiv:2004.07375*.

Pang, X., L. Liu, and Y. Xu (2020). A bayesian alternative to synthetic control for comparative case studies. *Available at SSRN*.

Petek, G. (2019). The 2019-2020 budget: Analysis of the governor's criminal justice proposals. Sacramento, CA.

Pinkney, S. (2021). An improved and extended bayesian synthetic control. *arXiv preprint arXiv:2103.16244*.

Ray, K. and A. van der Vaart (2018). Semiparametric bayesian causal inference using gaussian process priors. *arXiv preprint arXiv:1808.04246*.

Ren, B., X. Wu, D. Braun, N. Pillai, and F. Dominici (2021). Bayesian modeling for exposure response curve via gaussian processes: Causal effects of exposure to air pollution on health outcomes. *arXiv preprint arXiv:2105.03454*.

Richardson, T. S., R. J. Evans, and J. M. Robins (2011). Transparent parameterizations of models for potential outcomes. *Bayesian Statistics 9*, 569–610.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pp. 69–117. Springer.

Roth, J. (2019). Pre-test with caution: Event-study estimates after testing for parallel trends. *Department of Economics, Harvard University, Unpublished manuscript*.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 159–183.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58.

Samartsidis, P., S. R. Seaman, S. Montagna, A. Charlett, M. Hickman, and D. D. Angelis (2020). A bayesian multivariate factor analysis model for evaluating an intervention by using observational time series data on multiple outcomes.

Samartsidis, P., S. R. Seaman, A. M. Presanis, M. Hickman, D. De Angelis, et al. (2019). Assessing the causal effect of binary interventions from observational panel data with few treated units. *Statistical Science 34*(3), 486–503.

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

Solin, A. and S. Särkkä (2020). Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing 30*(2), 419–446.

Stan Development Team (2021). Stan Modeling Language Users Guide and Reference Manual. Version 2.26.

Tuomaala, E. (2019). The bayesian synthetic control: Improved counterfactual estimation in the social sciences through probabilistic modeling. *arXiv preprint arXiv:1910.06106*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian processes for machine learning*, Volume 2. MIT press Cambridge, MA.

Wilson, A. G., C. Dann, and H. Nickisch (2015). Thoughts on massively scalable gaussian processes. *arXiv preprint arXiv:1511.01870*.

Witty, S., K. Takatsu, D. Jensen, and V. Mansinghka (2020, 13–18 Jul). Causal inference using Gaussian processes with structured latent confounders. In H. D. III and A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *Proceedings of Machine Learning Research*, pp. 10313–10323. PMLR.

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

A Estimation and implementation choices

A.1 Single outcome

Both the Gaussian and Poisson model share the following three components to model the latent process

1. Each unit, i, is given an independent offset drawn from a standard normal distribution, $\nu \sim \mathcal{N}(0,1)$, providing a unit-specific intercept.

2. The global trend is modeled as a single-task Gaussian process with lengthscale ρ_{global},

$$\rho_{\text{global}} \sim \text{InvGamma}(5,5)$$
$$g \sim \mathcal{N}(0, \mathbf{K}_{\text{global}}; \rho_{\text{global}})$$

where $\mathbf{K}_{\text{global}}$ is a covariance matrix with each entry given by the squared exponential kernel with lengthscale, ρ, i.e. $\mathbf{K}_{\text{global}}(i,j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\rho}\right)$.

3. Individual components are modeled with a multitask Gaussian process with a rank k task-covariance matrix,

$$\rho_{\text{time}} \sim \text{InvGamma}(5,5)$$
$$\alpha_{\text{time}} \sim \mathcal{N}(0,1)$$
$$\beta \sim \mathcal{N}(0, \mathbf{I}_k)$$
$$\mathbf{K}_{\text{unit}} = \beta\beta^T$$
$$\text{vec}(\mathbf{f}) \sim \mathcal{N}(0, \mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}}; \rho_{\text{time}}, \alpha_{\text{time}})$$

4. The variance for a Normal outcome is drawn from $\sigma^2 \sim \mathcal{N}(0,1)$

With $\mathbf{K}_{\text{time}}$ is a squared exponential kernel of the same form as $\mathbf{K}_{\text{global}}$. For the Gaussian link, a shared standard deviation is drawn, $\sigma \sim \mathcal{N}(0,1)$, and the observed control outcomes for each unit, i, are modeled as

$$\frac{\mathbf{Y}_i}{\mathbf{N}_i} \sim \mathcal{N}\left(\nu_i + \mathbf{f}_i + g, \frac{\sigma^2}{\sqrt{\mathbf{N}_i}}\right),$$

where $\mathbf{N}$ is a matrix containing the population for each state for all time periods. Similarly, for the Poisson link function the observed untreated outcomes for each unit is modeled as

$$\mathbf{Y_i} \sim \text{Pois}\left(\mathbf{N}_i + \exp\left(\nu_i + \mathbf{f}_i + g\right)\right),$$

where N_i is an offset.

A.2 Multiple outcomes

The multi-outcome models analogously use the following components,

1. The hyperparameters of the global and individual kernels are drawn as

$$\rho_{\text{global}} \sim \text{InvGamma}(5, 5)$$
$$\alpha_{\text{global}} \sim \mathcal{N}(0, 1)$$

2. The global time trend for each outcome is drawn as

$$g \sim \mathcal{N}(0, \mathbf{K}_{\text{global}}; \rho_{\text{global}}, \alpha_{\text{global}})$$

3. Individual components are modeled as

$$\rho_{\text{time}} \sim \text{InvGamma}(5, 5)$$
$$\beta \sim \mathcal{N}(0, \mathbf{I}_k)$$
$$\mathbf{K}_{\text{unit}} = \beta\beta^T$$

4. Each unit is given an independent intercept for each outcome drawn from a standard normal distribution $\nu \sim \mathcal{N}(0, 1)$.

5. The covariance between outcomes is modeled as

$$\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I})$$
$$\mathbf{K}_{\text{outcome}} := \mathbf{Z}\mathbf{Z}^T$$

and the latent function is drawn as

$$\text{vec}(\mathbf{f}) \sim \mathcal{N}(0, \mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}} \otimes \mathbf{K}_{\text{outcomes}}; \rho_{\text{time}})$$

6. The variance for each Normal outcome is drawn from $\sigma^2 \sim \mathcal{N}(0, 1)$

Finally the probability of each outcome is modeled as

$$\frac{\mathbf{Y}_{i,j}}{\mathbf{N}_i} \sim \mathcal{N}(\nu_{i,j} + \mathbf{f}_{i,j} + g_j, \frac{\sigma_j^2}{\sqrt{\mathbf{N}_i}})$$

for the normal outcome for unit i and outcome j, and

$$\mathbf{Y}_{i,j} \sim \text{Pois}\left(\mathbf{N}_i + \exp \nu_{i,j} + \mathbf{f}_{i,j} + g_j\right)$$

for the Poisson.

A.3   Additional computational details

Naive estimation of the latent variable MTGP formulation used in our model has $\mathcal{O}(NT^3)$ complexity due of the cost of taking the Cholesky decomposition of $\mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}}$, with an additional $\mathcal{O}(NT^2)$ complexity from multiplying the Cholesky decomposition with an NT length draw from

a standard normal. This can be reduced by utilizing properties of the Kronecker product, namely that

$$\text{chol}\left(\mathbf{K}_{\text{unit}} \otimes \mathbf{K}_{\text{time}}\right) \text{vec}(\mathbf{Z}) = \left(\text{chol}(\mathbf{K}_{\text{unit}}) \otimes \text{chol}(\mathbf{K}_{\text{time}})\right) \text{vec}(\mathbf{Z})$$

$$= \left(\text{chol}(\mathbf{K}_{\text{unit}}) \otimes \text{chol}(\mathbf{K}_{\text{time}})\right) \text{vec}(\mathbf{Z}) = \text{vec}\left(\text{chol}(\mathbf{K}_{\text{time}})^{\top} \mathbf{Z} \text{chol}(\mathbf{K}_{\text{unit}})\right)$$

Which avoids performing Cholesky decomposition on the combined matrix $\mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}}$, and instead entails taking separate decompositions of the matrices individually which results in a complexity of $\mathcal{O}\left(\max(T^3 + N^3)\right)$. We reduce complexity further by taking advantage of the form of $\mathbf{K}_{\text{unit}} = \beta\beta^{\top}$, where $\beta \in \mathcal{R}^{N \times k}$ which, when k is sufficiently small, is substantially smaller than $\mathcal{R}^{N \times N}$. The sampling procedure then consists of drawing $\mathbf{Z} \in \mathbb{R}^{N,k}$, and taking $\text{vec}\left(\text{chol}(\mathbf{K}_{\text{time}})\mathbf{Z}\beta^{\top}\right)$. The final complexity with this low rank parameterization is then $\mathcal{O}\left(\max\left(T^3 + Nk\right)\right)$ which provides substantial advantage when N is sufficiently large. Finally, we note that in cases where T is large, a number of methods can be used to improve computational efficiency (see Solin and Särkkä, 2020; Wilson et al., 2015; Flaxman et al., 2015).

B Proofs

Proposition 4.1. Let $Y_{it}(0) = f_{it} + \varepsilon_{it}$, where f is a fixed function. Let $\mathcal{H}_k$ be the RKHS for kernel k with Hilbert-Schmidt norm, $\|f\|_{\mathcal{H}_k}^2 = k(f, f)$. Then let f be contained in the unit ball of the reproducing Hilbert space, $\|f\|_{\mathcal{H}_k} \leq 1$, where $k((i, t), (i', t')) = k_{\text{unit}}(i, i') \times k_{\text{time}}(t, t')$, and ε_{it} are iid mean-zero random variables with observed variance $\sigma^2 > 0$. Further, model the time covariance $\mathbf{K}_{\text{time}}$ with a squared exponential kernel with lengthscale ρ, and model the unit covariance $\mathbf{K}_{\text{unit}} = \beta\beta^\top$. Finally, let $\gamma^{(i)} \in \mathbb{R}^N$ and $\lambda^{(t)} \in \mathbb{R}^T$ be the vectors of unit and time weights for target unit i at time t, and let $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, respectively, be the set of weights across all targets. Then:

(a) The unit and time weights, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, have the following closed form:

$$\gamma^{(i)} \otimes \lambda^{(t)} = \left(\mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}} + \sigma^2 \mathbb{I}\right)^{-1} \left(k_{\text{time}}(t, \cdot) \otimes k_{\text{unit}}(i, \cdot)\right). \tag{6}$$

The posterior (predictive) mean estimate for target observation $(1, t^*) \notin \mathcal{C}$, for $t^* > T_0$ is:

$$\hat{\mu}_{1t^*} = \sum_{(i,t) \in \mathcal{C}} \gamma_i^{(1)} \lambda_t^{(t^*)} Y_{it}.$$

(b) The unit and time weights also solve the following, equivalent optimization problems, in terms of (1) the weight-space formulation:

$$\min_{\boldsymbol{\gamma}, \boldsymbol{\lambda}} \ \sup_{f \in \mathcal{H}_k} \ \sum_{(i,t) \in \mathcal{C}} \left(f_{it} - \sum_{(i',t') \in \mathcal{C}} \gamma_{i'}^{(i)} \lambda_{t'}^{(t)} f_{i't'}\right)^2 \ + \ \sigma^2 \|\boldsymbol{\gamma}\|_2^2 \|\boldsymbol{\lambda}\|_2^2, \tag{7}$$

and (2) the function-space formulation:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\xi}} \ \sum_{(i,t) \in \mathcal{C}} \left(Y_{it} - \sum_{(i',t') \in \mathcal{C}} \underbrace{\alpha_{i'}^{(i)} k_{\text{unit}}(i, i')}_{\gamma_{i'}^{(i)}} \cdot \underbrace{\xi_{t'}^{(t)} k_{\text{time}}(t, t')}_{\lambda_{t'}^{(t)}} Y_{i't'}\right)^2 \ + \ \sigma^2 \left(\boldsymbol{\xi} \otimes \boldsymbol{\alpha}\right)^\top \mathbf{K} \left(\boldsymbol{\xi} \otimes \boldsymbol{\alpha}\right), \tag{8}$$

where $\mathbf{K} = \mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}}$ and where $\alpha^{(i)} \in \mathbb{R}^N$ and $\xi^{(t)} \in \mathbb{R}^T$ are the coefficient vectors for target unit i at time t, with $\gamma_{i'}^{(i)} = \alpha_{i'}^{(i)} k_{\text{unit}}(i, i')$ and $\lambda_{t'}^{(t)} = \xi_{t'}^{(t)} k_{\text{time}}(t, t')$, and with corresponding sets $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$.

(c) This estimate has the following out-of-training-sample estimation error for the structural component of the missing potential outcome f_{1t^*},

$$|f_{1t^*} - \hat{\mu}_{1t^*}| \leq \left[\underbrace{\left(k_{\text{time}}(t^*, \cdot)^\top \lambda^{(t^*)} \otimes k_{\text{unit}}(1, \cdot)^\top \gamma^{(1)}\right)}_{\text{error for } f_{1t^*}} \ + \ \underbrace{\sigma^2 \left\|\gamma^{(1)}\right\|_2^2 \left\|\lambda^{(t^*)}\right\|_2^2}_{\text{irreducible noise}}\right]^{1/2}. \tag{9}$$

34

Proof. The proof largely follows from propositions in Kanagawa et al. (2018), presented here for the sake of completeness.

Part (a). The form of the weights follows directly from the conditional expectation of the multivariate normal in Equation (2). Specifically, the posterior (predictive) mean of the MTGP for unit i at time t is:

$$\begin{aligned}
\hat{\mu}_{it} &= (\mathbf{K}_{\text{obs}} + \sigma^2 \mathbb{I})^{-1} k((i,t), \cdot) \mathbf{Y}_1 \\
&= \underbrace{\left(\mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}} + \sigma^2 \mathbb{I} \right)^{-1} \left(k_{\text{time}}(t, \cdot) \otimes k_{\text{unit}}(i, \cdot) \right)}_{\boldsymbol{w}} \mathbf{Y}_1 \\
&= \sum_{(i', t') \in \mathcal{C}} w_{i't'}^{(i,i)} Y_{i't'},
\end{aligned}$$

for general weights $w^{(i,t)} \in \mathbb{R}^{N \times T}$ for target observation (i, t).

The separation of the overall weights w into the product of unit and time weights γ and λ follows from fundamental properties of kernels. Specifically, the form of the ICM kernel is given by the product of its constituent components, i.e. $k_{\text{time}}(t, t') k_{\text{unit}}(i, i')$. This resulting kernel is equivalent to

$$k_{\text{time}}(t, t') k_{\text{unit}}(i, i') = \mathbb{E}_{f \sim \mathcal{N}(0, \mathbf{K}_{\text{time}})} \left[f(t)^\top f(t') \right] \mathbb{E}_{g \sim \mathcal{N}(0, \mathbf{K}_{\text{unit}})} \left[g(i)^\top g(i') \right].$$

Finally, considering the full covariance we have

$$\mathbf{K}_{\text{time}} \otimes \mathbf{K}_{\text{unit}} = (f^\top f) \otimes (g^\top g) = (f \otimes g)^\top (f \otimes g)$$

which we can see is the product of independent functions, with equivalent representations given in terms of functions of the respective basis representations, $f = \lambda^\top \phi(t)$, and $g = \gamma^\top \psi(i)$.

Part (b). This is immediate from proposition 3.12 (equation 43) of Kanagawa et al. (2018) after applying decomposing the weights into the time and unit components.

Part (c). We start with the following result from Kanagawa.

Proposition B.1. (Kanagawa et al., 2018) Assume that $Y_i = f_i + \varepsilon_i$ where ε is independent zero-mean noise, and f is a fixed function contained in the reproducing Hilbert space implied by the kernel k. For any previously unobserved point, denoted t^*, we have

$$\sqrt{\sum_{i=1}^{T} k(t^*, \cdot)_i w_i^{(t^*)} + \sigma^2} = \sup_{f \in \mathcal{H}_k} \left(f_{t^*} - \sum_{i=1}^{T} w_i^{(t^*)} f_i \right),$$

where i indexes over the T observed points, and $w = (\boldsymbol{K} + \sigma \mathbb{I})^{-1} k(t^*, \cdot)$.

The error bounds follow by applying this result, noting that the supremum forms an upper bound, rearranging terms, and applying the separation of weights shown earlier. $\qquad \square$
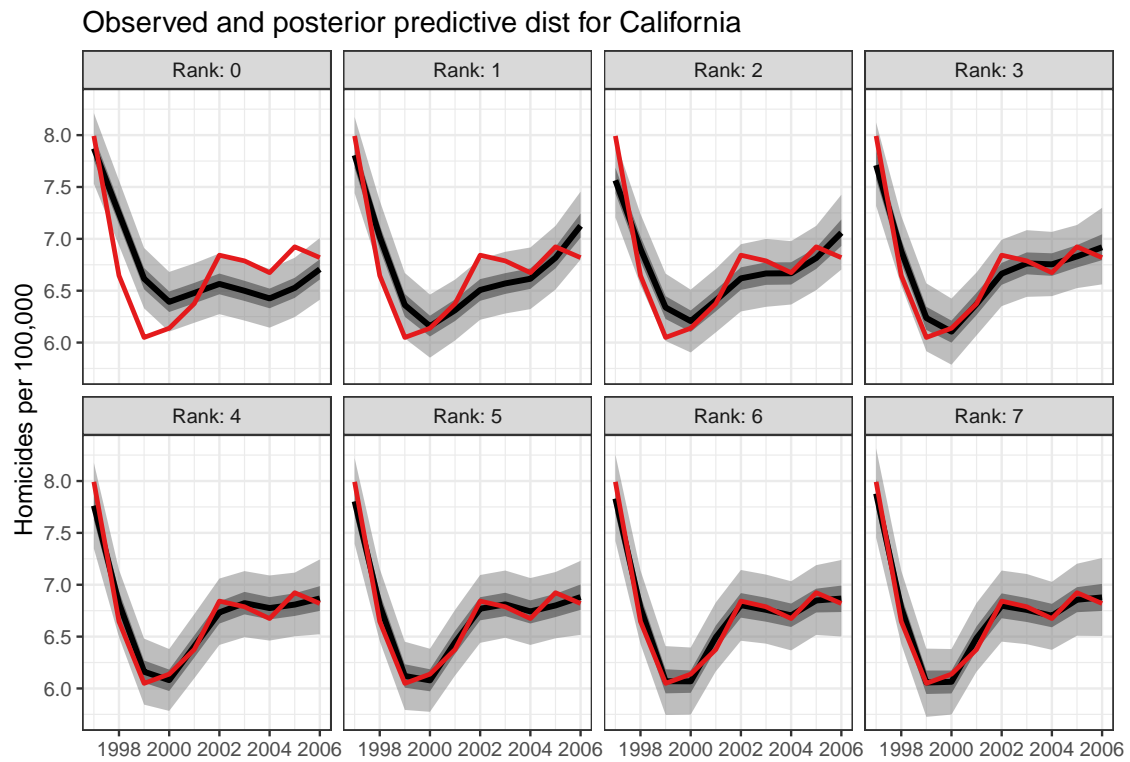
C Additional results

C.1 Additional diagnostics

Observed and posterior predictive dist for California

Figure C.1: Pre-treatment imbalance for Poisson model. 50% interval and 95% interval plotted. Observed data is in red.
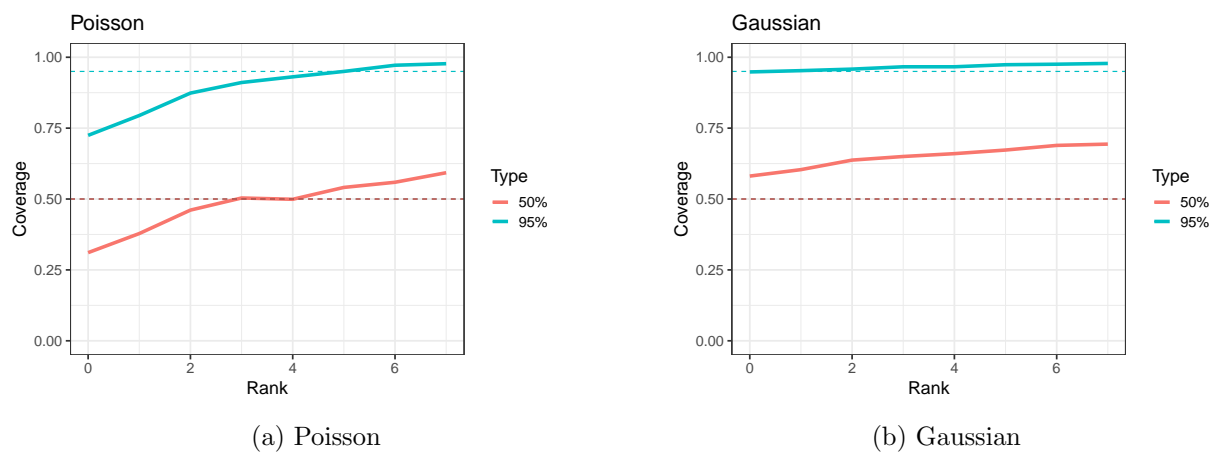
(a) Poisson

(b) Gaussian

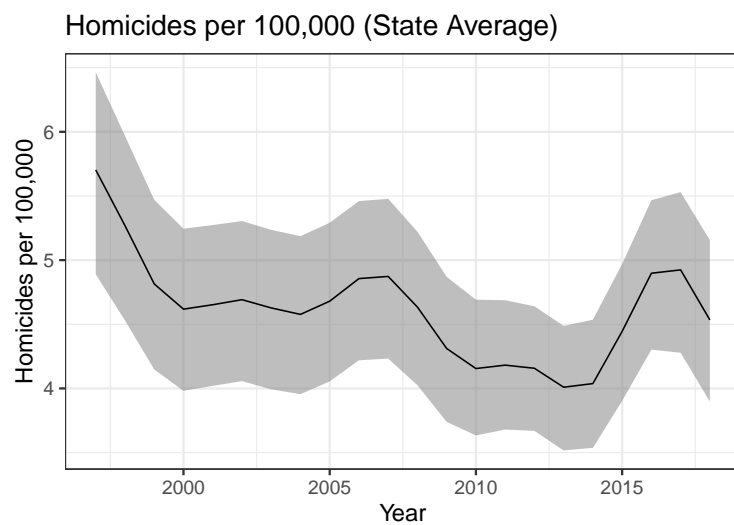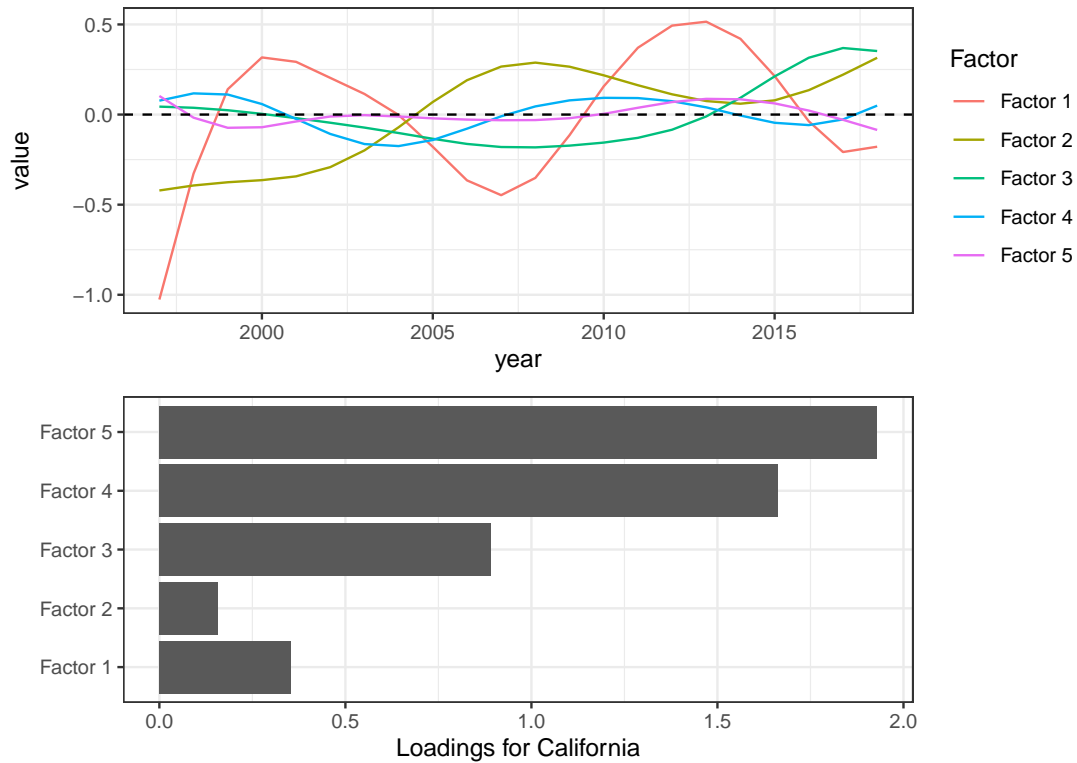Figure C.2: Coverage vs rank for Poisson and Gaussian models

Figure C.3: Global Gaussian Process Mean

Figure C.4: Inferred latent factors (top) and loadings for California. Note that the latent factors are only identifiable up to rotation. Here, we order the factors according to the fraction of variance explained across outcomes from all states, and we choose the sign of the factors so that the loadings are all positive for California. All the estimated factors are smooth over time due to the kernel, $\mathbf{K}_{\text{time}}$.
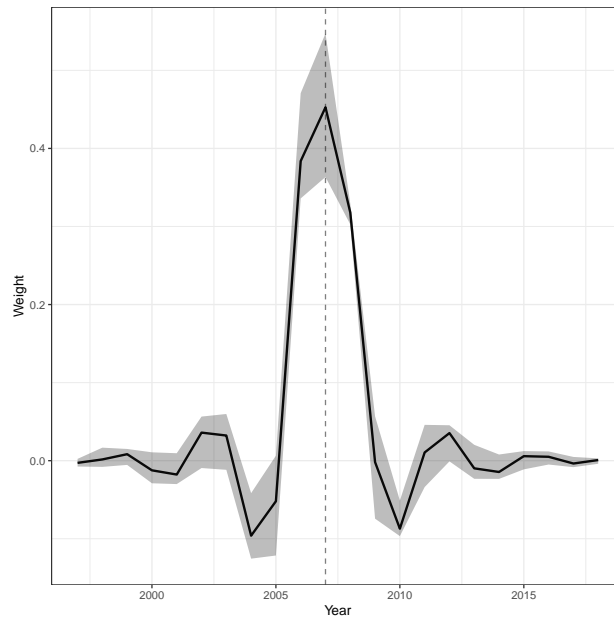
Figure C.5: Inferred time weights for California.
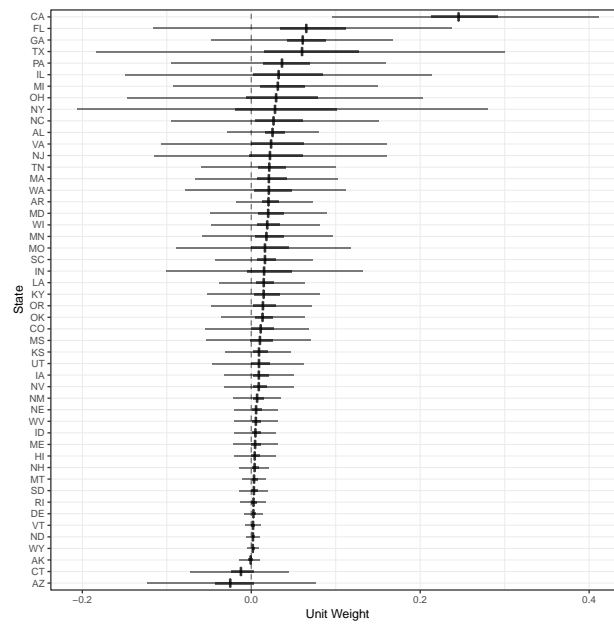
Figure C.6: Inferred unit weights for California.

C.2 Additional impact estimates

Observed and posterior predictive distribution of homicide rates
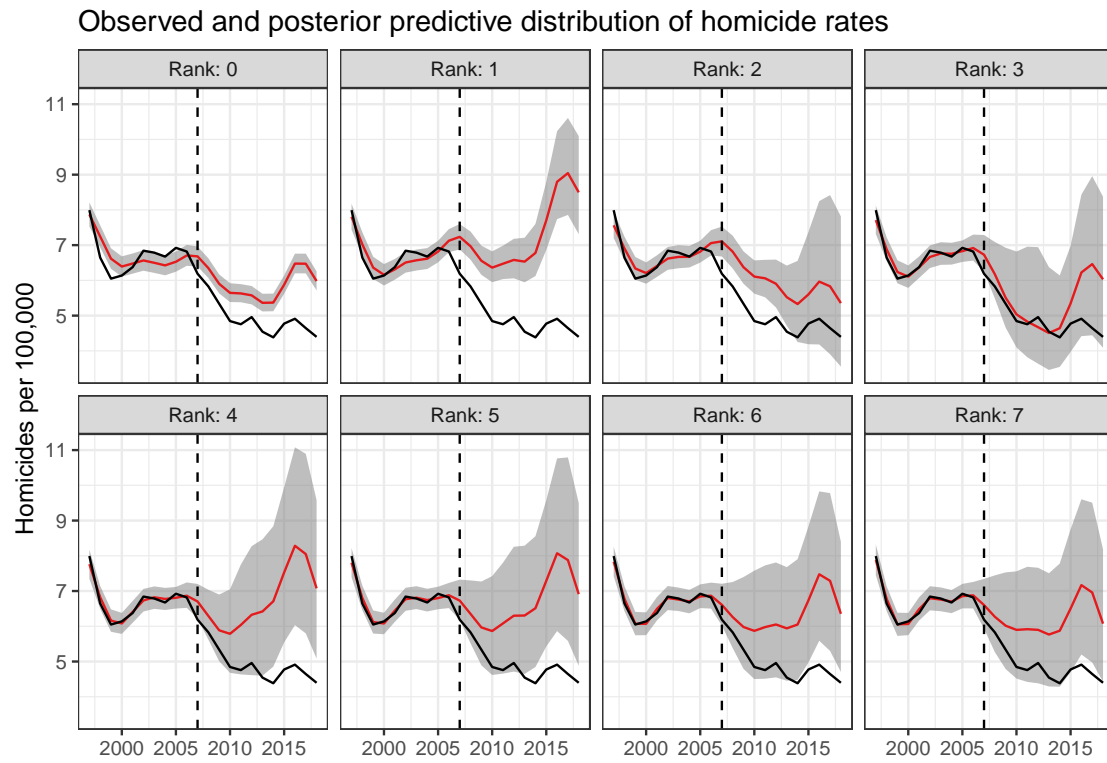
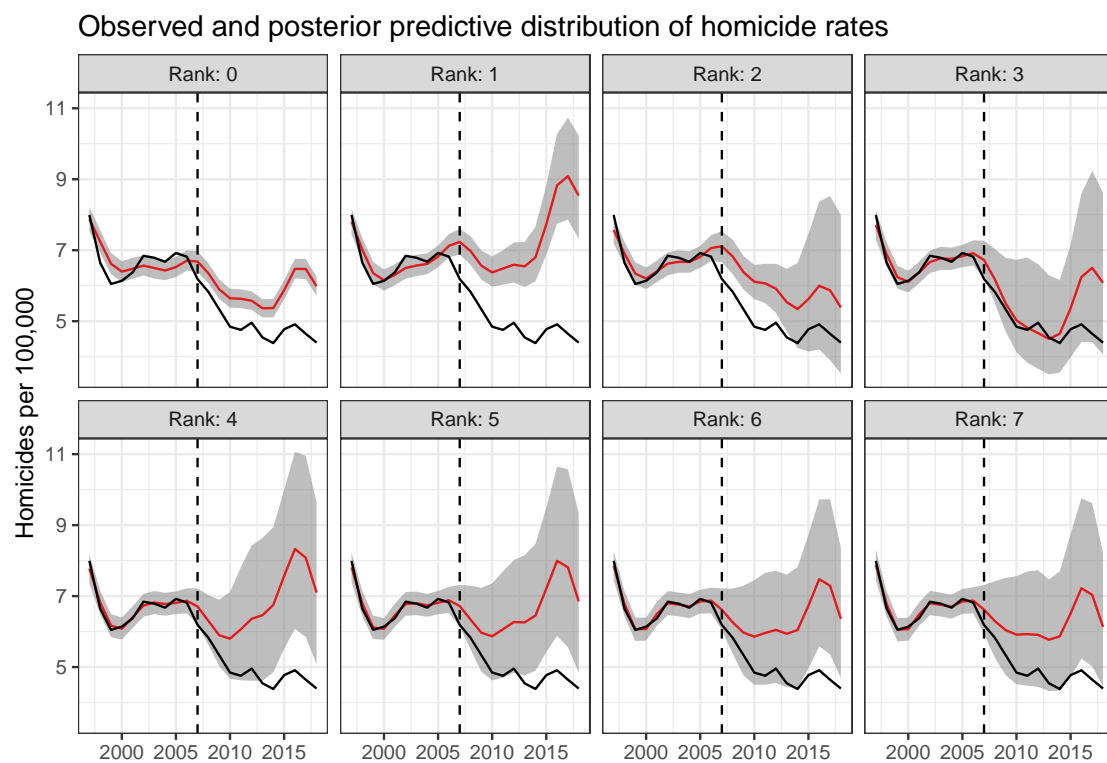Figure C.7: Estimates from MTGP for Normal model

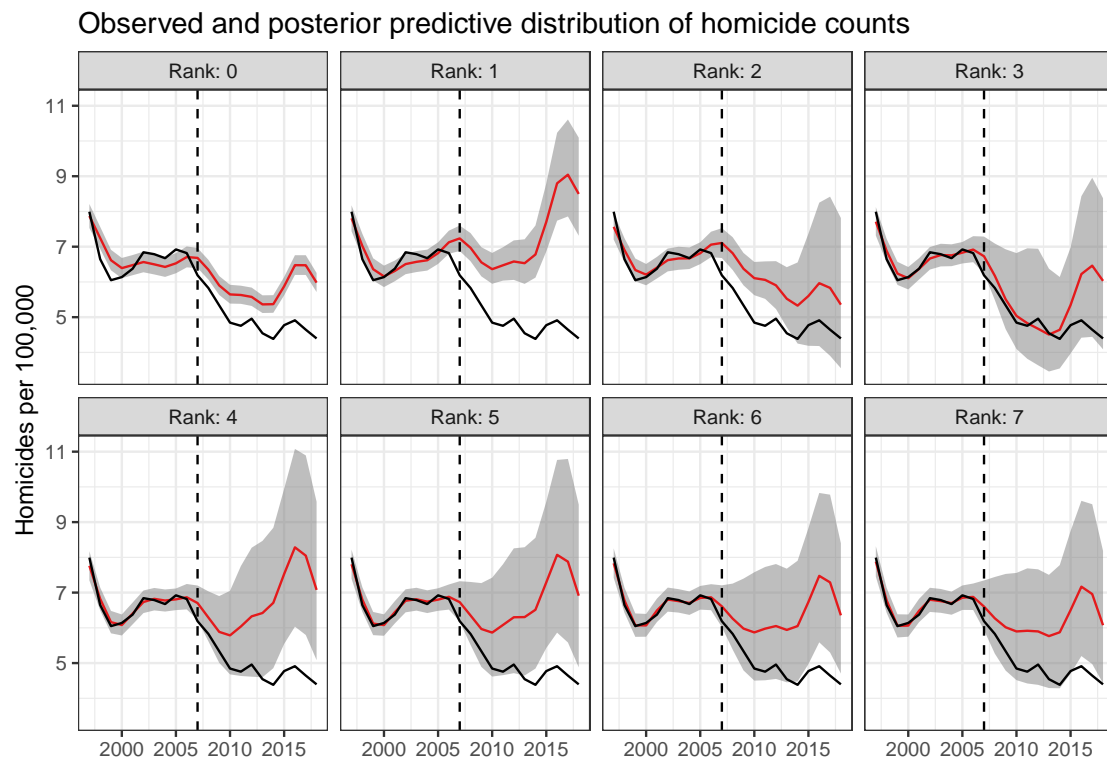Figure C.8: Estimates from MTGP for Normal model with a linear adjustment for background covariates.
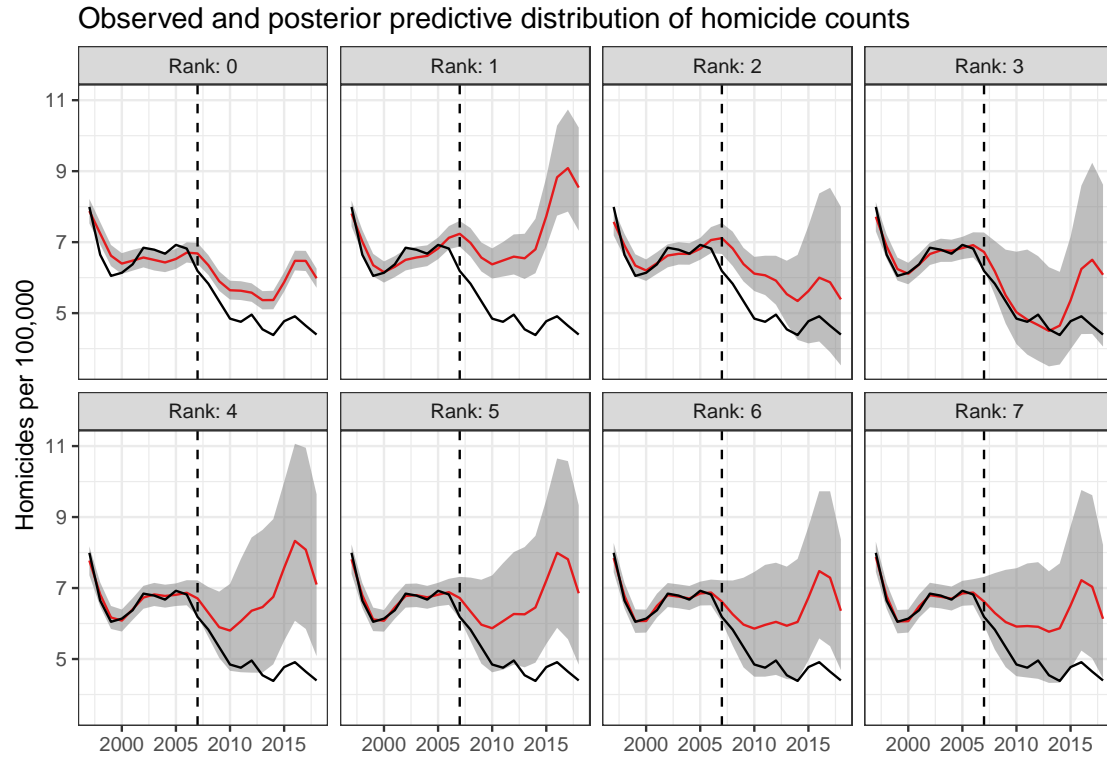
Figure C.9: Estimates from MTGP with Poisson link

Figure C.10: Estimates from MTGP with Poisson link and linear covariate adjustment.
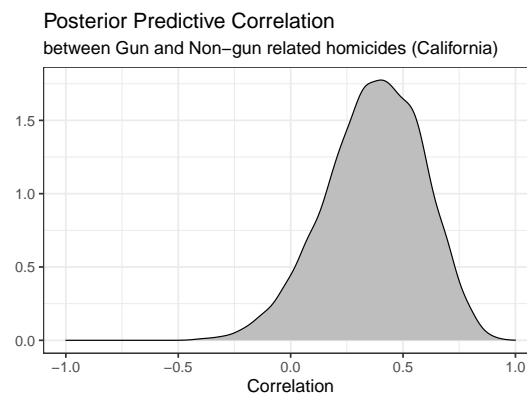

Figure C.11: Posterior predictive correlation between gun and non-gun related homicides in California, pre-treatment.