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

## 1 POSTERIOR ANALYSIS: MATHEMATICAL DETAILS

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

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

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

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

$$\text{var}_{\text{post}}\big(\gamma_1(t)\big) \leq \frac{n(1-\rho^2)+s^2}{n+s^2} \tag{3}$$

$$\text{var}_{\text{post}}\big(\delta(t)\big) \leq \frac{(s^2)(1+\sigma^2)}{n+n\sigma^2+s^2} + \frac{n(1-\rho^2)+s^2}{n+s^2} \tag{4}$$

Note that $t$ is arbitrary post-treatment time period so the above results hold for the entire post-treatment time series. Assume we can achieve the large-sample limit. We first observe that the treatment factual outcome can be *exactly* identified, as the posterior variance of $\gamma_1(t) + \delta(t)$ (14) will shrink to zero in the large-sample limit ($n \to \infty$). The joint posterior over treated counterfactual outcomes and treatment effects also becomes increasingly negatively correlated to $-1$ as $n \to \infty$. Hence, as long as one hypothesizes a counterfactual or a treatment effect with no uncertainty, the posterior on the other will collapse to a Dirac delta function. Moreover, (22) shows that the treated counterfactual outcome is *partially* identified up to an upper bound depending on the inter-group correlation parameter $\rho$, and can be *exactly* identified if trends were perfectly correlated (parallel) because this upper bound will shrink to zero. Finally, the treatment effect is also *partially* identified up to the same upper bound and can be *exactly* identified if trends were perfectly correlated (parallel).

### 1.1 Joint Posterior of Treatment Effect and Post-treatment Treated Counterfactual

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

$$\begin{bmatrix} \delta \\ \gamma_1 \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \sigma^2 \mathbf{1}_n^T \\ 0 & 1 & \mathbf{1}_n^T \\ \sigma^2 \mathbf{1}_n & \mathbf{1}_n & (1+\sigma^2)\mathbf{1}_n \mathbf{1}_n^T + s^2 \mathbb{I}_n \end{bmatrix} \right), \tag{5}$$

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

$$\begin{aligned}
\text{var}[\delta, \gamma_1 \mid \mathbf{y}] &= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \left( \mathbf{1}_n^T \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix} \right) \left( (1+\sigma^2)\mathbf{1}_n \mathbf{1}_n^T + s^2 \mathbb{I}_n \right)^{-1} \left( \mathbf{1}_n \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}^T \right) && (6) \\
&= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{s^2} \left( \mathbf{1}_n^T \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix} \right) \left( \frac{1+\sigma^2}{s^2} \mathbf{1}_n \mathbf{1}_n^T + \mathbb{I}_n \right)^{-1} \left( \mathbf{1}_n \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}^T \right) && (7) \\
&= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{s^2} \left( \mathbf{1}_n^T \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix} \right) \left( \mathbb{I}_n - \frac{1+\sigma^2}{A} \mathbf{1}_n \mathbf{1}_n^T \right) \left( \mathbf{1}_n \otimes \begin{bmatrix} \sigma^2 \\ 1 \end{bmatrix}^T \right) && (8) \\
&= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} - \frac{n}{A} \begin{bmatrix} \sigma^4 & \sigma^2 \\ \sigma^2 & 1 \end{bmatrix} && (9) \\
&= \frac{1}{A} \begin{bmatrix} \sigma^2(n+s^2) & -n\sigma^2 \\ -n\sigma^2 & n\sigma^2 + s^2 \end{bmatrix} && (10)
\end{aligned}$$

where $A = n + n\sigma^2 + s^2$. Hence we can compute the posterior correlation between $\delta$ and $\gamma_1$ as

$$\begin{aligned}
\text{cor}(\delta, \gamma_1) &= -\frac{n\sigma^2}{\sqrt{\sigma^2(n+s^2)(n\sigma^2+s^2)}} && (11) \\
&= -\frac{1}{\sqrt{(1+s^2/n)(1+s^2/(n\sigma^2))}} \to -1 \text{ if } n \to \infty && (12)
\end{aligned}$$

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

$$\begin{aligned}
\text{var}(\delta + \gamma_1) &= \frac{\sigma^2(n+s^2) - 2n\sigma^2 + (n\sigma^2 + s^2)}{n + n\sigma^2 + s^2} && (13) \\
&= \frac{s^2(1+\sigma^2)}{n + n\sigma^2 + s^2} \to 0 \text{ if } n \to \infty && (14)
\end{aligned}$$

We can see that although $\delta$ and $\gamma_1$ are uncorrelated a priori, the posterior correlation will approach to negative one in the limit of infinitely many data ($n \to 0$). In addition, the posterior variance of their sum will also approach to zero in the limit of infinitely many data ($n \to 0$). Hence, if we hypothesize a counterfactual or a treatment effect, then the posterior on the other will collapse.

### 1.2 Posterior of Post-treatment Control Factual

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

$$\begin{aligned}
\text{var}[\gamma_0 \mid \mathbb{Y}_{\text{obs}}] &\leq \text{var}[\gamma_0 \mid y_1, \ldots, y_n] && (15) \\
&= 1 - \mathbf{1}_n^T \Big( \mathbf{1}_n \mathbf{1}_n^T + s^2 \mathbb{I}_n \Big)^{-1} \mathbf{1}_n && (16) \\
&= \frac{s^2}{n + s^2} \to 0 \text{ if } n \to \infty && (17)
\end{aligned}$$

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

### 1.3 Posterior of Post-treatment Treated Counterfactual

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

$$\text{cov}\begin{bmatrix}\gamma_1 \\ \mathbf{y}_0\end{bmatrix} = \begin{bmatrix}1 & \rho\mathbf{1}_n^T \\ \rho\mathbf{1}_n & \mathbf{1}_n\mathbf{1}_n^T + s^2\mathbb{I}_n\end{bmatrix} \tag{18}$$

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

$$\begin{aligned}
\text{var}\big[\gamma_1 \mid \mathbb{Y}_{\text{obs}}\big] &\leq \text{var}\big[\gamma_1 \mid \mathbf{y}_0] && (19)\\
&= 1 - \rho\mathbf{1}_n^T\Big(\mathbf{1}_n\mathbf{1}_n^T + s^2\mathbb{I}_n\Big)^{-1}\rho\mathbf{1}_n^T && (20)\\
&= 1 - \rho^2\frac{n}{n+s^2} && (21)\\
&= \frac{n(1-\rho^2)+s^2}{n+s^2} \to 1-\rho^2 \text{ if } n \to \infty && (22)
\end{aligned}$$

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

### 1.4 Posterior of Treatment Effect

Suppose we have $n$ noisy post-treatment control observations $y_0^{(i)} = \gamma_0 + u_i + \varepsilon_i$ and $n$ noisy post-treatment control observations $y_1^{(i)} = \gamma_1 + v_i + \delta + \varepsilon_i$ with white noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$, where $u_i$ and $v_i$ are unit deviations at time $t$ and $s^2 = \sigma_u^2 + \sigma^2_{\text{noise}}$, respectively. Suppose the group correlation parameter is $\rho = \text{cov}(\gamma_1, \gamma_0)$. Using Eq. 14 and 22, we can derive an upper bound of posterior variance of $\delta$ as

$$\begin{aligned}
\text{var}[\delta \mid \mathbb{Y}_{\text{obs}}] &= \text{var}[\delta + \gamma_1 \mid \mathbb{Y}_{\text{obs}}] + \text{var}[\gamma_1 \mid \mathbb{Y}_{\text{obs}}] && (23)\\
&\leq \frac{s^2(1+\sigma^2)}{n + n\sigma^2 + s^2} + \frac{n(1-\rho^2)+s^2}{n+s^2} \to 1-\rho^2 \text{ if } n \to \infty && (24)
\end{aligned}$$

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

## 2 ADDITIONAL MODEL DETAILS

We briefly provide additional details about the individual components of our model and the hyperpriors used in Eq. (9). We put a multi-task GP prior with constant mean and SE kernel for the group trends $\{\gamma_g\}$. We place a shared GP prior with zero mean and SE kernel on all the unit deviations $\{u_i\}$. We put a GP prior with linear mean and SE kernel on the effects from covariates $h$. We place a GP prior with zero mean and SE kernel but scaled smoothly from zero at the time of intervention $T_0$ to a fixed output scale at some later time $T_1 = T_0 + \Delta T$ for the treatment effect process $\delta$. To enforce the full effect time $T_1$ is always no earlier than intervention time $T_0$, we require $\Delta T \geq 0$. Observation noise is assumed to be i.i.d. Gaussian. Table 1 summarizes all the model hyperparameters.

Table 1: Table of notations.

| Component | Prior | Hyperparameter |
|---|---|---|
| Group trends $\gamma_g(t)$ | constant $\mu_g(t)$<br>$K_\gamma(t,t') \cdot K_{\text{task}}(m,m')$ | $c_\gamma$<br>$\ell_\gamma$, $\lambda_\gamma$, $\rho$ |
| Unit deviation $u_i(t)$ | zero $\mu_u(t)=0$<br>$K_u(t,t')$ | N/A<br>$\ell_u$, $\lambda_u$ |
| Covariate $h(x)$ | linear $\mu_x(x)$<br>$K_x(x,x')$ | slope $a_x$<br>$\ell_x$, $\lambda_x$ |
| Treatment effect $\delta(t)$ | zero $\mu_\delta(t)=0$<br>$K_\delta(t,t')$ | N/A<br>$\ell_\delta$, $\lambda_\delta$, $\Delta T$ |
| General noise $\varepsilon$ | $\mathcal{N}(0,\sigma^2_{\text{noise}})$ | $\sigma_{\text{noise}}$ |

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

## 3 INFERENCE

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

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

$$p\big(\delta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}, \theta\big) \propto p\big(\mathbb{Y}^{(1)}_{\text{obs}} \mid \mathbb{Y}^{(0)}, \delta, \theta\big)\, p\big(\mathbb{Y}^{(0)}_{\text{obs}} \mid \mathbb{Y}^{(0)}, \theta\big)\, p(\delta \mid \theta). \tag{25}$$

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

$$p\big(\delta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}\big) = \int p\big(\delta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}, \theta\big)\, p\big(\theta \mid \mathbb{Y}_{\text{obs}}, \mathbb{Y}^{(0)}, \delta\big)\, \mathrm{d}\theta. \tag{26}$$

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

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

# 4 HYPERPRIORS

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

The sampling is done via Hamiltonian Markov chain Monte Carlo.[1] We sampled five chains using a random restart around the MAP estimator for initialization. Specifically, we initialized each chain by perturbing the MAP hyperparamters (in the transformed space) with an additive Gaussian jitter with standard deviation equal to 0.1. For each chain, 3000 samples were collected after a burn-in of 1000 samples, which we found to be typically sufficient.

## 4.1 Simulation Studies

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

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

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

## 4.2 LocalNews

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

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

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

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

The hyperpriors for LocalNews model are summarized below:

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

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

# 5 BASELINE IMPLEMENTATION DETAILS

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

1. The 2FE model was implemented using the standard approach by introducing time/unit indicators into the regression model as fixed effects, but used the ordinary least square estimator with robust standard errors (Bell and McCaffrey, 2002) to account for noise heteroscedasticity among units. OLS with the robust standard error is similar to standard OLS, but assumes the noise may be heteroscedastic, such that the noise matrix is clustered as a blocked matrix, in our case, by unit. The implementation relied on the `lm_robust` function from `estimatr` library in R. We used the default setting of HC2 type.

2. The GSC model was implemented using software[2] provided by Xu (2017). We set the number of maximal factors to be 10 and allow the built-in cross validation procedure to select the optimal number of factors. We impose unit and day fixed effects besides interactive effects.

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

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

5. The ICM model was implemented using the Matlab gpml software [4] following the setup in Arbour et al. (2021). We used independent standard normal priors for the scalar coefficients, and set the number of latent processes to be 5. The details for MCMC sampling are the same with MGP-PANEL to ensure comparability.

# 6 ABLATION STUDY

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

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

The **naïve** estimator shows the value of including dynamics in treatment effect structures into the model. Specifically, naïve only extrapolates the post-treatment periods for the treated group after conditioning on the pre-treatment observations for the treated group and the pre- and post-treatment observations for the control group, and then subtracts the observed outputs from the posterior predictions in the post-treatment periods for the treated group as estimations of treatment effects. This estimator is equivalent to ours except that the GP prior on ATT is assumed to be infinitely wide (Arbour et al., 2021). Hence, naïve estimator allows ATT to be arbitrary but incorporates no temporal correlation.

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

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

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

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

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

Table 2: Performance measures of MGP-PANEL compared to ablated models across selected settings of correlation parameters (standard errors are shown in parentheses). Among all models, MGP-PANEL has the highest coverage and LL scores in all settings. Although the MAP estimator has lower RMSE than MGP-PANEL, the differences are not significant in a paired $t$-test.

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

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

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

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

Table 2 shows the averaged RMSE, $95\%$ confidence interval coverage rate, and log likelihood scores of MGP-PANEL and ablated models for different levels of correlation. Among all models, MGP-PANEL has the highest COVERAGE and LL scores in all settings, but the MAP estimator has the lowest RMSE. A reason for this may be that MGP-PANEL is designed for a better coverage rate and log likelihood due to the fully Bayesian inference that absorbs the model and hyperparameter uncertainty. Note that although the MAP estimator has a better RMSE than the MGP-PANEL estimator, the differences are not significant in a paired t-test.

Several implications could be drawn from Table 2. First, the **BLR** and **perfect corr** estimators have the worst performance among all models, emphasizing that inferring time trends correctly is the most critical aspect of accurately estimating treatment effects using time series. Second, the **uncorr trend** model has better performance than the **BLR** and **perfect corr** estimators but is still much worse than the other models, indicating the vital role of regularization in modeling time effects when temporal structure exists. Finally, the **naïve**, **uncorr trend** and **no unit trends** estimators perform are just slightly worse than MGP-PANEL and MAP models, encouraging practitioners to further regularize treatment effects, fully make use

of post-treatment data and take into consideration unit-level heterogeneity in groups whenever possible.

# 7 ADDITIONAL SIMULATION

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

Table 3: Performance scores of MGP-PANEL compared to baseline estimators under different data generating process (DGP) averaged across different random seeds, including using a student-t noise model with degree of freedom equal 4 (non normal error), modeling non smooth time trends using GP with Matérn kernel (non smooth trends), observing one unit per group (few units) and modeling group trends using independent GP (independent GP trends). Performance scores of GSC and 2FE under few units setting are not available due to bugs in released code from original authors.

| | MGP-PANEL | GSC | 2FE | CMGP | DM-LFM | ICM | LTR |
|---|---|---|---|---|---|---|---|
| DGP | | | | RMSE | | | |
| Non normal error | *0.0215(22)* | 0.0536(20) | 0.0461(14) | **0.0181(14)** | 0.0497(9) | 0.0454(14) | 0.0228(47) |
| Non smooth trends | **0.0102(23)** | 0.0389(10) | 0.0381(15) | 0.0210(9) | 0.0351(11) | 0.0312(17) | 0.0342(37) |
| Few units | **0.0224(13)** | N/A | N/A | *0.0337(105)* | 0.0923(16) | 0.0918(35) | *0.0239(37)* |
| Independent GP trends | **0.0190(18)** | 0.0933(58) | 0.0613(96) | 0.0298(27) | 0.0533(21) | 0.0434(18) | 0.0299(49) |
| DGP | | | | COVERAGE | | | |
| Non normal error | **1.000**(0) | **1.000**(0) | *0.982*(12) | **1.000**(0) | **1.000**(0) | 0.758(15) | 0.665(91) |
| Non smooth trends | **1.000**(0) | 0.960(10) | 0.950(10) | **1.000**(0) | **1.000**(0) | 0.860(18) | 0.470(46) |
| Few units | **1.000**(0) | N/A | N/A | **1.000**(0) | **1.000**(0) | 0.960(18) | **1.000**(0) |
| Independent GP trends | 0.923(25) | *0.990(10)* | 0.780(56) | *0.923(37)* | **1.000**(0) | 0.842(24) | 0.773(204) |
| DGP | | | | LL | | | |
| Non normal error | **2.304(54)** | 1.474(32) | 1.657(12) | *2.273(23)* | 1.572(21) | 1.092(132) | 0.867(740) |
| Non smooth trends | **2.736**(193) | 1.753(11) | 1.815(35) | *2.396*(31) | 1.900(27) | 1.778(65) | −1.999(1.029) |
| Few units | **1.981(75)** | N/A | N/A | 1.279(43) | 0.881(41) | 0.900(26) | *1.846(76)* |
| Independent GP trends | *1.912(127)* | 0.905(62) | 1.033(326) | **2.079(106)** | 1.451(43) | 1.366(15) | 1.324(64) |

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

Table 4: Demonstration of model selection when the true noise model is student-t distributed or the group trends are generated from Matérn kernels. Lower Bayesian information criterion (BIC) is preferred.

| DGP | Student-t distributed noise | | group trends with Matérn kernels | |
|---|---|---|---|---|
| model | Student-t noise | Gaussian noise | Matérn kernel | Gaussian kernel |
| BIC | $-\mathbf{1014}$ | $-863$ | $-\mathbf{847}$ | $-832$ |

## 8 CASE STUDY: COMPARISON WITH BASELINES

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

Table 5: Comparison of estimated averaged treatment effects and uncertainties between MGP-PANEL and other baselines at various post-treatment time periods in the applied study. Amongst all models, MGP-PANEL has the lowest uncertainty on ATE and learned a minor instantaneous treatment effect.

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

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

## 9 CASE STUDY: NON-GAUSSIAN LIKELIHOODS

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

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

We specifically examine an increase in conflict along the Afghanistan–Pakistan border in 2008. In 2008, President Pervez Musharraf of Pakistan, a U.S. ally, became embroiled in a corruption scandal, culminating in a coalition government agreeing to impeach him on August 11, 2008. Musharraf resigned on August 18 Perlez (2008). We assume that the

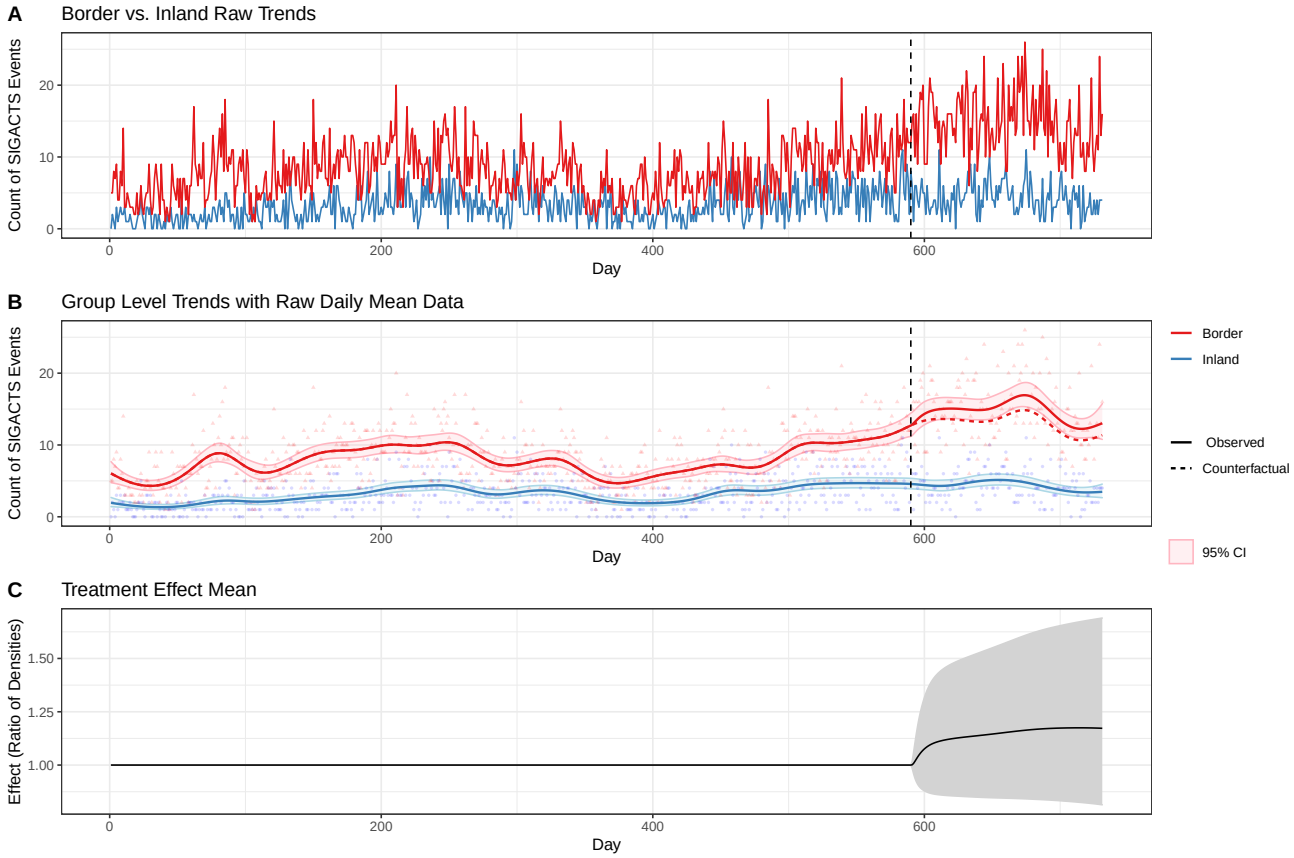


Figure 1: Panel A shows the trends of the raw group counts over time, where the border province count (treatment) is in red and the inland province count (control) is in blue. Panel B shows the fitted group level trends with 95% credible intervals (solid lines and shaded regions), the modeled counterfactual trend of the treated group (dashed red line), and the raw group counts as points. Panel C shows the posterior of the average treatment effect on the treated interpreted as a ratio of the density of direct fire on the border to the density of direct fire inland with a 95% credible interval.

increased violence started on Aug 12, 2008 (the day after the government agreed to impeach Musharraf) and compared the aggregated number of reported direct fire events along the border to the number of events inland. We hypothesize that, after announcing plans to impeach the president, the coalition sought to establish their independence from the West by violently confronting the U.S. and ISAF forces along the Afghanistan–Pakistan border, so we should observe more direct fire events in the border region.

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

$$y(t) \sim \text{Poisson}\big(\lambda(t)\big) \tag{27}$$

$$\log\big(\lambda(t)\big) \sim \mathcal{GP}(\mu_y, K_y) \tag{28}$$

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

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

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

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

**References**


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

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

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

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

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

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

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

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

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

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

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

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