A Gaussian Process Framework for Structured, Flexible, and Interpretable Machine-Learning Models

Annamaria Prati, Yehu Chen, Jacob Montgomery, and Roman Garnett

October 7, 2023

Abstract

Standard models in political science require scholars to make strong assumptions about the data generating process, often with limited guidance from theory or background knowledge. Modern machine learning approaches allow researchers to take a more agnostic approach to model building, but often fail to provide natural measures of uncertainty for key quantities and make it difficult to encode *a priori* knowledge. In this paper, we introduce the powerful Gaussian process (GP) framework, which offers a satisfying compromise between the restrictive assumptions of traditional linear models and the highly agnostic assumptions embedded within many machine learning methods. We begin by introducing Gaussian process regression (GPR), noting that the vast majority of linear models already in the literature are actually special (and restrictive) cases of GPR. We then illustrate how to leverage the power of GPs to build more flexible – but structured – models in the presence of clustering and spatiotemporal autocorrelation.

1 Flexibility versus structure

Social scientists are typically interested in learning substantively-motivated quantities such as marginal effects, first differences, or treatment effects in settings where they are hard to isolate accurately from background factors such as clustering, temporal autocorrelation, or spatial confounding. For example, Blair, Di Salvatore and Smidt (2023) asks whether a UN peacekeeping mission can promote democracy in countries experiencing a civil war. In this case, prior literature indicates that there are multiple possible confounding variables to take into account (e.g., poverty, reliance on natural resources, regional instability, etc.). Furthermore, there are additional pitfalls requiring researchers to account for spatiotemporal autocorrelation as well as clustering at the country level.

In these settings, scholars face a dilemma. Relying on traditional methods, researchers can build some variant of a linear model that incorporates specific structures they believe are relevant to the question at hand. Common solutions might be including fixed or random effects, controlling for confounding variables such as time (e.g. Paglayan, 2021) or latitude (e.g. Ahmed and Stasavage, 2020; de Kadt and Larreguy, 2018), or including spatial or temporal lags (e.g. Acemoglu et al., 2008; Di Salvatore, 2019).

The advantage of these models is that they are readily available, simple to estimate, easy to interpret, and allow the researcher to encode strong *a priori* knowledge about the data generating process. The disadvantage, however, is that they require researchers to make firm commitments to specifications even where there is little background knowledge to guide such decisions. What variables should be included? Should they be entered in linearly, with polynomial terms, or interactions? Should we control for lagged values of the outcome, and, if so, for how many previous periods? These and other questions must often be answered with limited guidance from theory and are often consequential to the results.

An attractive option in these settings is to rely on more flexible machine learning method such as a generalized additive models (Beck and Jackman, 1998), Bayesian model averaging

(Bartels, 1997; Montgomery and Nyhan, 2010), neural network (Beck, King and Zeng, 2000), random forest (Athey, Tibshirani and Wager, 2019; Hill and Jones, 2014; Montgomery and Olivella, 2018), kernel regression (Hainmueller and Hazlett, 2014), and more (e.g. Argyle et al., 2023; Grimmer, Messing and Westwood, 2017; Kleinberg et al., 2018; Torres and Cantú, 2022). Machine learning offer flexibility, an especially appealing feature when the researcher is uncertain about model specification. The disadvantages are that these models are difficult to interpret and often provide no natural estimates of uncertainty for key quantities of interest. In addition, situations where researchers have strong beliefs they wish to encode into the model, it can be difficult or even impossible to do so. Even tasks such as including fixed effects or interaction terms can require customized software, since existing implementations will not, for instance, include lower-order terms or all fixed-effects dummies (Montgomery and Nyhan, 2010). In short, these models can be *too agnostic* when we are already aware of some (but not all) potential issues.

To address this dilemma, this paper introduces Gaussian Process Regression (GPR) (Rasmussen and Williams, 2006) for Political Science research. While these models have been widely studied in computer science (Aglietti et al., 2020; Alaa and van der Schaar, 2017; Arbour et al., 2021; Flaxman, Neill and Smola, 2015; Hensman, Fusi and Lawrence, 2013; Rasmussen and Williams, 2006; Reynolds, 2009; Witty et al., 2020), they remain remarkably rare in Political Science (but see Chen, Garnett and Montgomery, 2023; Gill, 2021*a*,*b*). We argue that GPR is a machine learning framework that offers a satisfying compromise; researchers can encode domain knowledge when available but allow for high levels of flexibility where appropriate. Moreover, as a Bayesian model, GPR lends itself naturally to producing meaningful measures of uncertainty for key quantities of interest such as marginal effects. Indeed, as we demonstrate below, many of the linear models already favored by quantitative scholars are themselves special (restricted) examples of a GPR model. Thus, adopting the GP framework may represent a win-win, simply adding flexibility to existing practice.

To motivate GPR, we begin by briefly discussing alternative approaches to model building in the social sciences. We then provide an overview of the GR framework and approach to inference, giving special attention to how GPR is closely related to standard methods in the field. We then report a simulation study to compare the method to standard practices in the field, before reporting to applications that illustrate how to leverage the power of GPR to build flexible, but structured, models in the presence of clustering and spatio-temporal autocorrelation. These examples are designed to illustrate how we can begin with a simple linear model, and extend it naturally to account for increasingly complex problems, all within the same framework. All estimation is done in the `gpytorch` framework, which allows for modular construction and fast estimation.[1] We conclude with a discussion of avenues for future work.

Throughout, our aim is to provide an approachable overview of GPR for a social science audience, something that is currently missing from the literature. This is especially important because the GP literature is primarily aimed at building *predictive* models; standard quantities of interest to social scientists are not discussed (Rasmussen and Williams, 2006).

2 Flexible but structured models

Generically, model building based on observed covariates can be denoted,

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \tag{1}$$

where y_i is the outcome of interest for observation i, $\mathbf{x}_i$ is a vector of predictive features, and ε_i represents the error term. To begin, we make no assumptions about the structure of the error term, leaving that to our later discussion of inference. Instead, our primary focus is on the prior structure and assumptions we place on $f(\mathbf{x}_i)$.[2]

[1] Future versions of this paper will included detailed appendices with code and examples for applied scholars to use.

[2] We can generalize the presentation by removing the error term and specifying the problem as $y_i = g\big(f(\mathbf{x}_i); \theta\big)$, where $g(\cdot)$ is an inverse link function with ancillary parameters θ. However, this more generalized presentation makes the exposition more cumbersome and adds little to our main argument.

2.1 Approaches to model building

In the classic linear model, Equation 1 becomes

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i. \tag{2}$$

Here, f is just a weighted sum (linear combination) of the input vector $\mathbf{x}_i$, where the "weights" are represented by the vector β. In this formulation, f is assumed to be a function within the function space spanned by $\{\mathbf{x}_i\}$. That is, we are assuming that f can be approximated accurately by a linear combination of the observed covariates.

Where we need more flexibility in the functional form, we can also create a basis expansion, $\phi(\mathbf{x}_i)$, including such things as squared terms, interactions, or even splines. In this case, Equation 1 becomes,

$$y_i = \phi(\mathbf{x}_i)^T \beta + \varepsilon_i. \tag{3}$$

While more complex, this formulations still assumes that f exists in the function space spanned by the explicit inputs $\{\phi(\mathbf{x}_i)\}$. That is, we are assuming that f can be approximated by a linear additive combination of the transformed inputs.

A key advantage of this approach is that downstream inference is generally focused on the regression parameters β. These are (to varying degrees) interpretable and also come with well-established measures of uncertainty across the frequentist, Bayesian, and maximum likelihood paradigms. The downside is that this model can introduce significant error when f cannot be well approximated by $\phi(\mathbf{x})^T \beta$. In most settings, there are an *extremely* large number of potential basis expansions to consider, and it is difficult for scholars to know when they have identified a "good" approximation of f.

Automatic basis expansion. To address this challenge, scholars have increasingly relied on flexible machine learning models that seek to automatically generate appropriate basis expansions for Equation 3. Examples generalized additive models (GAM) (Beck and

Jackman, 1998), random forests (Athey, Tibshirani and Wager, 2019; Hill and Jones, 2014; Montgomery and Olivella, 2018), neural networks (Beck, King and Zeng, 2000), and double or triple machine learning (Chernozhukov et al., 2018; Ratkovic and Tingley, 2023). Although different in their details, these methods share a general structure where an algorithm searches through a high-dimensional space of possible choices to identify a basis expansion $\phi(\mathbf{x}_i)$ and weights vector β such that f is well-approximated by $\phi(\mathbf{x}_i)^T\beta$. The curse of dimensionality ensures that not all expansions can be considered, so each method follows its own algorithmic heuristic to explore this space. Past research shows that, with appropriate regularization and sufficient data, these models often faithfully approximate f and make accurate predictions of $\{y_i\}$.

While attractive in principal, there are at least three drawbacks to this approach in practice. First, model outputs from these machine learning models are very difficult to interpret. While the final model structure $\phi(\mathbf{x}_i)^T\beta$ is still a linear model, the automatic approach to basis expansion almost ensures that the β parameters themselves will be at best distantly related to the substantive quantities of interest. So, for instance, random forests offer nothing like a regression coefficient that captures the average marginal relationship between a raw input and the outcome even though it is simply a weighted summation of basis expansions.

Second, even where such quantities can be calculated (e.g., via simulation), they often do not come with appropriate measures of uncertainty. For example, even simple models such as the LASSO models are unable to provide valid standard errors (Casella et al., 2010). In some cases, it is possible to approximate standard errors using jackknife or bootstrapping algorithms (e.g., Sexton and Laake, 2009). However, even where this is not computationally prohibitive, methods such as bootstrapping can become immensely complex in applied settings. For instance, when data exhibits clustering or spatiotemporal autocorrelation, bootstrapping methods must be adjusted in *ad hoc* ways to account for the non-independence between observations (Cameron, Gelbach and Miller, 2008; Esarey and Menger, 2019; Jackson, 2020).

Third, as they are implemented, these algorithms are so agnostic that they typically do not allow researchers to encode *a priori* knowledge into their estimation procedure. With panel data, for instance, we can account for unit-level and time-level shocks by adding random effects to the standard linear model. Now also indexing by t, Equation 1 becomes,

$$y_{it} = \phi(\mathbf{x}_{it})'\beta + u_i + v_t + \varepsilon_{it}, \tag{4}$$

where u_i and v_t are assigned some prior structure.[3] However, models such as random forests are extremely difficult to adjust for these settings, offering no way to ensure the model is accounting for spatial or temporal dynamics.

2.2 Related work

To sure, GPR is not the only machine learning model that of interest to Political Science. Indeed, these methods have become increasingly common across Political Science to analyze datasets large (e.g. Imai, Lo and Olmsted, 2016) and small (e.g. Broniecki, Leemann and Wüest, 2022). It has been intuitively used for prediction (e.g. Cranmer and Desmarais, 2017; Muchlinski et al., 2016; Streeter, 2019), but those predictive powers have also been extended as proposed solutions for when data is sparse or missing (Chen, Garnett and Montgomery, 2023; Ratkovic and Tingley, 2017), to construct variables (e.g. Carroll and Kenkel, 2019; Chiu and Xu, 2023; Fong and Tyler, 2021; King, Pan and Roberts, 2017; Knox and Lucas, 2021; Mitts, Phillips and Walter, 2022), and to build datasets (e.g. Barari, Lucas and Munger, 2021; Gohdes, 2020). Less common, but of great interest to social science scholars, is the use of machine learning for estimating treatment effects (e.g. Fong and Grimmer, 2023; Grimmer, Messing and Westwood, 2017; Knox, Lucas and Cho, 2022; Ratkovic and Tingley, 2017).

As noted above, a particular body of work has specifically focused on on using machine

[3] In this setting, fixed effects models are simply where we place a white noise prior on these parameters.

learning to building more robust and flexible models to test substantive claims. This literature focuses on three key and inter-related concerns: how to choose covariates, choosing optimal specifications, and how to compare competing models with different specifications. For example, in regards to the first concern, neural networks, random forests, and similar methods assign weights to different covariates to maximize an objective function, essentially "pruning" the set of covariates to only include the best predictors of the outcome (Athey, Tibshirani and Wager, 2019; Beck, King and Zeng, 2000; Hill and Jones, 2014; Montgomery and Olivella, 2018; Torres and Cantú, 2022). Alternatively, other methods, such as gradient boosting, weight the individual data points (Kleinberg et al., 2018). Beyond covariate choices, researchers also face the question of how to optimally fit a model, especially when the data exhibits non-linear relationships. For example, GAMs allow researchers to choose any non-parametric function for each independent variable, accommodating non-linearity and other arbitrary relationships (Beck and Jackman, 1998). Model averaging methods are designed to allow researchers to combine insight from multiple model configurations without having to choose "the best" (Bartels, 1997; Grimmer, Messing and Westwood, 2017; Montgomery and Nyhan, 2010). GPR, however, is particularly attractive because it addresses all three of these concerns in a single framework by finding the optimal distribution over an infinite number of functions of any specification and any combination of covariate weights. Indeed, GPs are in the class of models called "universal approximators," meaning (roughly) that given enough data they can always accurately learn f even if the kernel is incorrectly specified.[4]

The work most similar to our own is kernel regularized least squares (KRLS) (Hainmueller and Hazlett, 2014; Mohanty and Shaffer, 2019) and the kernel smoothing estimator (Hainmueller, Mummolo and Xu, 2019; Li and Racine, 2010). KRLS in particular is mathematically very similar to basic GPR, although it comes from the frequentist tradition. A related literature uses kernels within a matching framework for causal estimation (Fong, Hazlett and Imai,

[4] The most commonly discussed universal approximator models are neural networks. Remarkably, Lee et al. (2018) shows that a single-layer neural network with an infinite number of nodes is itself a GP.

2018; Hazlett, 2020; Hazlett and Xu, 2018). This ties closely to GPR since, as we show below, it is possible to characterize GPR predictions for counter-factual values as a weighted sum of observations, where weights are assigned via "closeness" in a kernel. Another related line of work is triple machine learning (Ratkovic and Tingley, 2023), which also seeks to leverage the flexibility of machine learning methods while maintaining many of the advantages of the standard linear models.

Relative to existing kernel methods, the advantage of GPR is that it offers us a way to systematically build models that include structure when it is needed, but can be highly flexible where theory and prior beliefs provide limited guidance. We can assume functional forms are linear and additive, non-linear and smooth, or anything in between. We can impose sharp restrictions on some parameters, assume the function is strictly additive in covariates, or we can allow for high-level interactions and only loose expectations about how functional forms will be shaped. Moreover, as we illustrate below, GPR is modular, allowing us to start simple and add complexity, all while providing a coherent method for comparing alternative model specifications via Bayesian model selection methodologies. In contrast, KRLS and related methods have primarily been implemented for experimental or cross-sectional settings. For instance, we are aware of no extensions that would make it appropriate for building a model for panel data.

In addition, since GPR is a Bayesian model, it is relatively straightforward to either derive or construct common quantities of interest from the posterior, which come with natural measures of uncertainty. In contrast, implementations for KRLS and related kernel models rely on bootstrapping methods for standard errors, which can become highly inaccurate in settings with complex error structures.

The goals of triple machine learning are similar to our own, in that we can leverage the power of flexible models but retain the simplicity of the classic linear modeol. However, the method itself is very different requiring an extensive procedure (including multiple splits in the data) with the goal of identifying an optimal basis expansion $\{\phi(\mathbf{x_i})\}$ that allows us to approximate f. GPR, however, follows a completely different approach that allows us to avoid the problem of constructing a covariate set altogether.

3 Gaussian process regression

GPR is a Bayesian nonparametric approach that seeks to sidestep the issue of basis exploration and selection by placing a prior on the function f itself, not on specific regression parameters (e.g., β).[5] The goal is choose a prior structure that is flexible, but still encodes prior knowledge about the DGP. An additional concern is to do so while remaining mathematically (or at least numerically) tractable. This is achieved by (i) marginalizing[6] out β and (ii) re-representing the problem so inference is conducted not using the (possibly infinitely large) basis expansion space $\{\phi(\mathbf{x}_i)\}$, but rather the much more tractable input feature space $\{\mathbf{x}_i\}$.

In this section, we provide a high-level overview of the method, beginning with model specification and moving to inference and interpretation. We then provide a short simulation study before moving to our applications.

3.1 Basic model specification

Letting $f_i = f(\mathbf{x}_i)$ and assuming we have n observations, we let

$$\{f_i\} \sim \mathcal{GP}(\mathbf{0}; \mathbf{K}). \tag{5}$$

Here $\mathcal{GP}(\cdot)$ denotes a Gaussian process prior, which is equivalent to an n-dimensional multivariate normal prior. Thus, $\mathbf{0}$ is a vector of length n filled with zero and $\mathbf{K}$ is an $n \times n$

[5] Bayesian nonparametric models have become increasingly common in the social sciences, but primarily based on Dirichlet process priors rather than GPs (e.g., Bisbee, 2019; Moser, Rodríguez and Lofland, 2021; Shiraito, Lo and Olivella, 2023).

[6] You can, if you wish, still include such parameters but they are unnecessary and will be problematic with high-dimensional datasets.

positive-definite covariance matrix. For reasons that become clear below, we refer to this is the kernel matrix or kernel covariance matrix.

At a very high level, there are three properties of GP models that make them especially useful.[7] First, *any linear model with an explicit basis function can be represented as a* GPR *with a specific kernel* (Rasmussen and Williams, 2006). Consider the linear formulation $\phi(\mathbf{x}_i)^T\beta$ with prior $\beta \sim N(\mathbf{0}, \Sigma_\beta)$ on our regression coefficients. For any two observations i and j, we can show that (f_i, f_j) are jointly normal with mean $\mathbf{0}$ and covariance $\phi(\mathbf{x_i})^T\Sigma_\beta\phi(\mathbf{x_j})$ (Rasmussen and Williams, 2006, p. 14). By extension, the entire regression function model can then be re-represented as

$$\{f_i\} \sim \mathcal{GP}\big(\mathbf{0}, \mathbf{K}_{\text{lm}}\big), \tag{6}$$

where the ij^{th} element of K_{lm} is $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\Sigma_\beta\phi(\mathbf{x}_j)$. Rasmussen and Williams (2006) show that this result generalizes such that any linear model[8] based on explicit set of inputs can be re-represented as a GP. Intuitively, what this means is that model building is no longer about specifying the correct linear basis expansion (viz. choosing covariates, interactions, or polynomial terms), but rather about setting up the correct covariance kernel $\mathbf{K}$.

Second, using the "kernel trick", any GP model can be specified as a function of the *untransformed* input vectors $\{\mathbf{x}\}$ with no need for any regression parameters β. This feature results from the fact that any model that is defined entirely based on the inner product of a transformation of raw features can be re-represented as a function of the original unmodified features using a kernel function $k(\cdot, \cdot)$. Generically, $\psi(\mathbf{x}_i) \cdot \psi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$, where $\psi(\cdot)$ is a re-representation of $\mathbf{x}$. For example, in the case of the basic GPR shown in Equation 6, we see that the ij^{th} element of the covariance matrix is $\phi(\mathbf{x}_i)\Sigma_\beta\phi(\mathbf{x}_j)$. Letting $\psi(\mathbf{x}) = \Sigma_p^{1/2}\phi(\mathbf{x})$, we can see that the appropriate kernel function that corresponds exactly to this model is $k(\mathbf{x}_i, \mathbf{x}_j) = \psi(\mathbf{x}_i) \cdot \psi(\mathbf{x}_j)$. This again generalizes such that any linear model with an explicit basis function can be re-represented purely as a function of the input features through an

[7] This section relies heavily on results in Rasmussen and Williams (2006) Chapters 2 and 4.

[8] This result depends on placing Gaussian priors on the model parameters β.

appropriate kernel function $k(\cdot,\cdot)$.

This second result is primarily useful in that we have removed consideration (and estimation) of the β parameters. In practice, using this result directly would still require us to commit to some model specification via an explicit basis expansion. From this, we could derive the corresponding kernel. This mathematical feature can sometimes be useful by itself, since the kernel representation may be more computationally tractable in high dimensional settings.

However, the third and most remarkable feature of GP models is that we can flip this logic on its head by choosing a kernel function that encompasses a space of potential basis expansions. That is, we can move away from specifying $\phi(\cdot)$ entirely, instead choosing a kernel function $k(\cdot,\cdot)$ that represents the *class* of basis expansion we wish to consider. Amazingly, there are many well-studied kernels that correspond to infinitely large basis expansions – something that would be impossible to specify or even consider in a traditional framework. Moreover, the kernel requires only the raw input vectors $\{\mathbf{x}_i\}$, which typically have a much lower dimensionality than the basis expansion.

A concrete example is helpful. The most common kernel is the squared exponential (SE) function[9],

$$k(x_i, x_j) = \sigma^2 \exp\big(-\frac{(x_i - x_j)^2}{2l^2}\big),$$

where σ and l are hyperparameters discussed below. Choosing this kernel is equivalent *infinite-dimensional basis expansion* of $\mathbf{x}_i$.[10] Substantively, it encodes the assumption that we expect f to be very *smooth*, but can otherwise take on any shape. With this, we can build models that are highly flexible; far more flexible than any standard linear model will allow.

In total, GPR allows us take a very different approach to model building. Instead of trying to find the "correct" model specification or allowing an algorithm to search through a

[9] In other literatures this is termed a radial basis function (RBF) or a Gaussian kernel.

[10] In future versions of this paper, we will provide a short appendix summarizing common kernels in the literature. See Rasmussen and Williams (2006) Chapter 4 for an introduction.

space of possible feature representations, we instead focus on specifying a covariance kernel $\mathbf{K}$ that encodes our beliefs about the data generating process. Kernels can correspond to explicit assumptions, such as the basic linear regression example above. Alternatively, they far more agnostic and flexible, such as the SE that corresponds to an infinitely large basis expansion. And, as we show below, we can construct kernels that have both structure and flexibility as suits the specific application.

3.2 Encoding *a priori* knowledge

An additional feature of GPR is that it is relatively straightforward to customize for specific settings, largely due to the fact that affine transformations of Gaussians are also Gaussian. So, for instance, we might begin with *a priori* expectations that f will take on a specific shape (e.g., linear) but wish to *allow* it to deviate from linearity when appropriate.

To do this, we can specify a mean function $\mathbb{E}(f(\mathbf{x})) = m(\mathbf{x}_i)$, that encodes our expectations. Our GPR then becomes

$$\{f_i\} \sim GP(\{m(\mathbf{x_i})\}, \mathbf{K}). \tag{7}$$

If $m(\cdot)$ includes additional parameters (e.g., $m(\mathbf{x}) = \mathbf{x}^T\beta$), these will also need to be estimated. While this is feasible, we can also marginalize out the mean function into a new GPR with a different kernel, $\{f_i\} \sim GP(\mathbf{0}, \mathbf{K}_{\text{alt}})$

Further, we can build more complex models when needed by combining GPs. So, for instance, we may wish to include unit and time-level random effects as in Equation 4. Assuming Gaussian hierarchical priors, this can be written as

$$\{f_i\} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}) + \mathcal{GP}(\mathbf{0}, \mathbf{K}_u) + \mathcal{GP}(\mathbf{0}, \mathbf{K}_v),$$

which becomes

$$\{f_i\} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K} + \mathbf{K}_u + \mathbf{K}_v),$$

The kernels $\mathbf{K}_u$ and $\mathbf{K}_v$ could be chosen to reflect assumed independence, or to encode additional structure (e.g., smoothing over geography or some other covariate).

More interestingly, we could place independent GP priors on each unit, reflecting the assumption that $\mathbb{E}(\mathbf{f}_i)$ should move smoothly over time. As we illustrate below, this is (yet again) another GP with a slightly different kernel specification. This would be analogous to a hierarchical trajectory model, but without any assumed functional form for the unit-level movement through time. This is just one configuration of a GP kernel that would be very difficult to specify and implement in a traditional framework, but is easily incorporated into a GPR.

3.3 Inference, errors, and interpretation

Thus far we have focused entirely on specifying the GP prior on f, ignoring the error term ϵ_i and posterior inference. To address these issues, we begin by assuming Gaussian error. We return to the issue of non-Gaussian likelihoods below.

To calculate our posterior, we return to the basic formulation of $\{y_i\} \sim f(\mathbf{x}_i) + \epsilon_i$, but now we add the assumption that $\epsilon_i \sim N(0, \sigma^2_{\text{noise}})$. Since Gaussians are additive, we can see that $cov(y_i, y_j) = k(x_i, x_j) + \sigma^2_{noise} I(i = j)$, where $I(\cdot)$ is the usual indicator function, which is one when $i = j$ and zero otherwise. In matrix notation this is

$$cov(\mathbf{y}) = \mathbf{K} + \mathbf{I}\sigma^2_{\text{noise}}.$$

Thus, the complete likelihood is

$$\mathbf{y} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K} + \mathbf{I}\sigma^2_{\text{noise}}), \tag{8}$$

where σ^2_{noise} is a hyperparameter. We can write the likelihood function more generally as, $p(\mathbf{y}|\mathbf{f}, \mathbf{X}, \theta)$, where θ is the collection of σ^2_{noise} and any hyperparameters in $\mathbf{K}$.

In standard Bayesian models, the goal would then be to reverse this to find the posterior distribution, $p(\mathbf{f}|\mathbf{X}, \mathbf{y}, \theta)$. For social science purposes, however, this specification is not directly useful as it would characterize f only at the *observed* values. For many quantities such as first-differences or causal effects, we will want to understand f at a broader set of values including those not yet observed. That is, we need to extract posterior estimates for f both at observed locations in our data but also at hypothetical locations where are covariates take on a different (hypothetical) value. For instance, we might want to know how the expected outcome, $\mathbb{E}(f_i|\mathcal{D}, \theta)$, will change in a response to a one unit increase in some predictor for a specific unit.

To do this, we will introduce the notation that $\mathbf{X}$ represents the matrix of observed covariates and $\mathbf{X}^*$ represents some set of observed/unobserved values we want to consider. We can then let $\mathbf{K}(\mathbf{X}, \mathbf{X})$ be the kernel matrix where the ij^{th} element is $k(\mathbf{x}_i, \mathbf{x}_j)$. We can also specify $\mathbf{K}(\mathbf{X}^*, \mathbf{X})$, where the ij^{th} element represents $k(\mathbf{x}_i^*, \mathbf{x}_j)$. Our goal is to estimate the posterior for $\{\mathbf{f}^*\} = \mathbb{E}(\mathbf{y}^*|X, \mathbf{y}, \mathbf{X}^*, \theta)$. With this notation, we can then specify the joint distribution of the data and f^* as:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{GP}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \mathbf{I}\sigma^2_{\text{noise}} & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*) \end{bmatrix}\right). \tag{9}$$

Using standard Gaussian identities, and letting $\mathcal{D} = (\mathbf{y}, \{\mathbf{x}_i\}, \mathbf{X}^*)$, we can then derive the conditional distribution

$$\mathbf{f}^*|\mathcal{D}, \theta \sim \mathcal{GP}\big(\mu_{f^*|\mathcal{D},\theta}, K_{f^*|\mathcal{D},\theta}\big), \tag{10}$$

where

$$\mu_{f^*|\mathcal{D},\theta} = K(\mathbf{X}^*, \mathbf{X})\big[K(\mathbf{X}, \mathbf{X}) + \mathbf{I}\sigma^2_{\text{noise}}\big]^{-1}\mathbf{y} \tag{11}$$

$$K_{f^*|\mathcal{D},\theta} = K(\mathbf{X}^*, \mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X})\big[K(\mathbf{X}, \mathbf{X}] + \mathbf{I}\sigma^2_{\text{noise}}\big]^{-1}K(\mathbf{X}, \mathbf{X}^*) \tag{12}$$

and $K(\mathbf{X}, \mathbf{X}) + \mathbf{I}\sigma^2_{\text{noise}}$ is an $n \times n$ matrix. This inversion is the most computationally expensive part of the algorithm.

Note that in this construction $\mathbb{E}(\mathbf{f}_i^*|\mathcal{D},\theta)$ in Equation 11 is a function only of the kernel outputs and the vector observed outcomes $\mathbf{y}$. Let w_i be the i^{th} element of $K(\mathbf{X}^*,\mathbf{X})\big[K(\mathbf{X},\mathbf{X})+\mathbf{I}\sigma^2_{\text{noise}}\big]^{-1}$. Then, $\mathbb{E}(f_i|\mathcal{D},\theta)$ is,

$$\sum_{i=1}^{n} w_i y_i.$$

Intuitively, this means that any predicted hypothetical (or counterfactual) value, is estimated as the weighted sum of the outcomes where the weights reflect proximity in the learned kernel space. This relates GPR back to the kernel-weighted matching methods discussed above, since all counterfactual values are implicitly weighted sums of the observed outcome values.

With these quantities in hand, we can easily calculate quantities of interest such as $f_i - f_i^*$ for a single individual or averaging these values over the entire observed population to get first differences. Indeed, since these are again affine transformations of a GP, these quantities are themselves GPs with analytical solutions.[11]

Amazingly, we can use this same approach to calculate the average marginal effects since differentiation is itself a linear operator (See Rasmussen and Williams, 2006, Section 9.4). Let $\frac{\partial f_i}{\partial \mathbf{x}_{dj}}$ be the partial derivative of a function along dimension d. For any observation pair ij, we have

$$cov\left(f_i, \frac{\partial f_j}{\partial x_{dj}}\right) = \frac{\partial k(\mathbf{x}_i,\mathbf{x}_j)}{\partial \mathbf{x}_{dj}}, cov\left(\frac{\partial f_i}{\partial x_{di}}, \frac{\partial f_j}{\partial x_{dj}}\right) = \frac{\partial^2 k(\mathbf{x}_i,\mathbf{x}_j)}{\partial x_{di} x_{dj}}$$

. Which means we can derive the full posterior as,

$$\begin{bmatrix} \mathbf{f}|\mathcal{D} \\ \nabla \mathbf{f}|\mathcal{D} \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} \mu_{\mathbf{f}|\mathcal{D}} \\ \nabla \mu_{\mathbf{f}|\mathcal{D}} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mu|\mathcal{D}} & \nabla \mathbf{K}^T_{\mu|\mathcal{D}} \\ \nabla \mathbf{K}_{\mu} & \nabla^2 \mathbf{K}_{\mu} \end{bmatrix}\right), \tag{13}$$

With this result, we can again easily calculate quantities such as the marginal effect of a covariate at a specific location or average this quantity across observed and/or counterfactual

[11] In future versions of this paper we will provide an appendix with the full posteriors for these and other quantities of interest.

values.

While GP models have closed-form solutions with Gaussian errors, the posteriors must be approximated for other error structures (Rasmussen and Williams, 2006). However, here we can lean on nearly two decades of work that has provided a wide array of methods (e.g., Laplace approximation, expectation propagation, variational inference, Markov chain Monte Carlo sampling) to approximate this posterior with varying levels of accuracy and computational complexity (Brooks et al., 2011; Girolami and Rogers, 2006; Hensman, Matthews and Ghahramani, 2015; Rasmussen and Williams, 2006; Titsias, 2009). Perhaps the simplest approach is Laplace approximation, which can provide good (and fast) approximations when the posterior is well behaved and unimodal.

3.4 Hyperparameters, model selection, and regularization

As discussed above, the GPR framework allows us to re-represent models such that we can marginalize out standard regression parameters. Instead, the modeling problem becomes selecting the correct kernel. Obviously this does not come without cost, as the kernel(s) have hyperparameters that must somehow be specified. For instance, if using the SE kernel for noisy observations we have $\theta = (\sigma, l, \sigma^2_{\text{noise}})$.[12]

There are a number of ways to select these parameters. Most simply, we could choose them based on either *a priori* knowledge or through some form of cross-validation. However, this first option is likely to be difficult to justify, and cross-validation can be extremely computationally expensive for all but the smallest datasets.

A more principled Bayesian approach would be to place a prior on θ, $p(\theta)$, and then use standard Bayesian methodologies to marginalize these parameters away. This can be done using some form of Markov chain Monte Carlo (MCMC) sampling or Laplace approximation. This would allow us to leverage the posterior on f while fully incorporating our uncertainty

[12] Note that in this presentation l is a scalar. However, in many cases it may be desirable to estimate a separate l for each input dimension or to use an automatic relevance determination kernel.

about θ.[13]

In practice, however, these approaches can be very expensive computationally, especially with larger datasets. We have found that our models perform adequately with the simpler approach of using θ^{MAP}, which are the *maximum a posteriori* estimates. Intuitively, this can be considered an "empirical Bayes" approach where these hyperparameters are selected to maximize the marginal loglikelihood,

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi. \tag{14}$$

Note that only the first component of this equation includes the outcome y. The latter components can be interpreted as a complexity penalty that can serve to create natural Bayesian regularization to avoid overfitting. Thus, Rasmussen and Williams (2006, p. 111) state:

> Notice that the trade-off between data-fit and model complexity is automatic; there is no need to set a parameter externally to fix the trade-off. ... Thus, a model complexity which is well suited to the data can be selected using the marginal likelihood.

Finally, note that Equation 14 (the marginal log-likelihood) can be easily used to summarize the model, which facilitates direct calculations of Bayes factors for model testing. Likewise, we can also calculate common fit statistics such as the Bayesian information criterion (BIC).

4 Applications

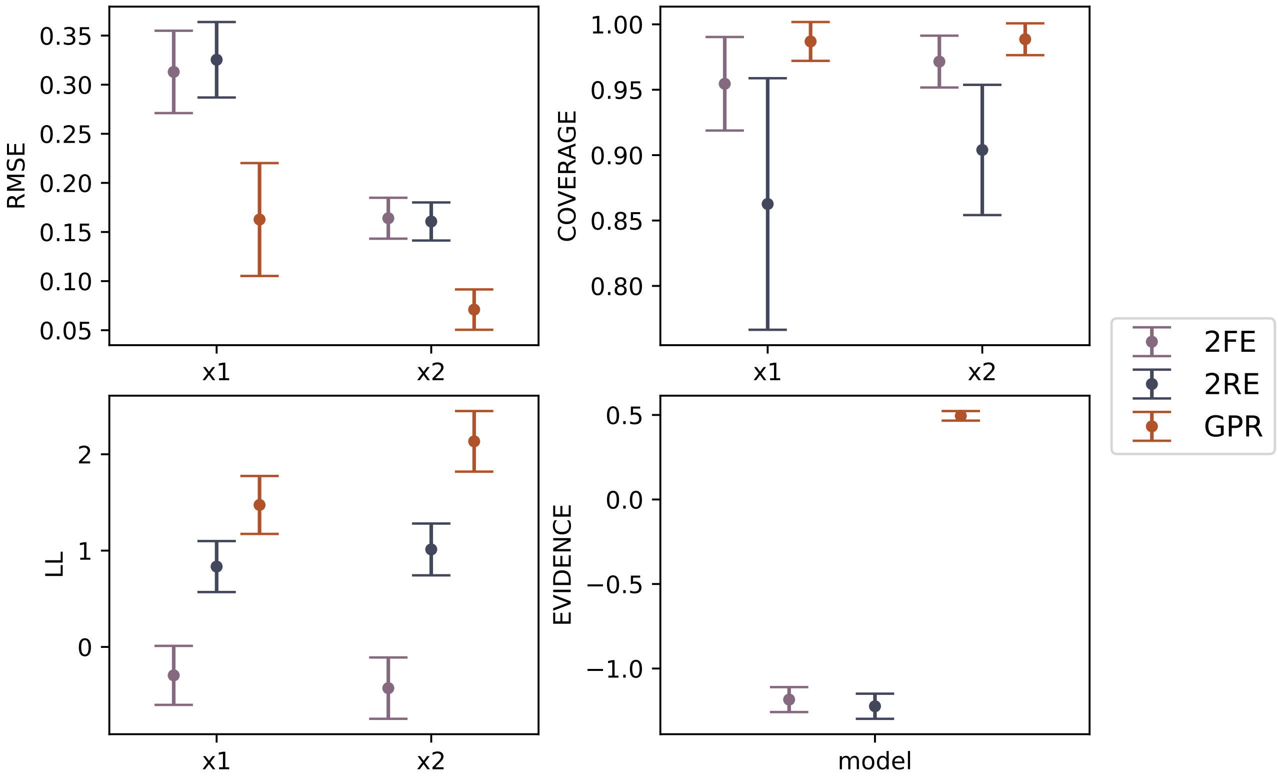
In this section, we show a simulation and two applications of GPR. First we present the findings from a simulation study, where we show that across several model diagnostic statis-

[13] In future versions of this paper, we hope to compare alternative methods of handling hyperparameters in GPYTORCH.

tics GPR performs no worse and often better than fixed effects and random effects for panel data. Moving on to the applications, we first show how GPR can be used with panel data to challenge recent findings in a debate on international organization and human rights. In the second application, we add a spatial component to GPR with an example on accountability and infant mortality.

4.1 Simulation

Figure 1: Model fit statistic comparisons with 95% confidence interval for RMSE (upper left), coverage (upper right), log likelihood (lower left), and evidence (lower right), between two-way fixed effects (purple), two-way random effects (blue), and GPR (orange) for simulated times series data. Across all statistics, GPR does no worse and often better than two-way fixed effects and two-way random effects.



To show the benefits of GPR, we ran a times series simulation study with two covariates ($D = 2$), 50 observations ($N = 50$), and 20 time periods ($T = 20$). The data generating process for the outcome y_{it} is a non-linear covariate effect and an individualized time effect.

The general data and model set-up can be summarized as follows:

$$x_{id} \sim N(0,1), d = 1, 2 \tag{15}$$

$$y_{it} = f(x_i) + g_i(t) + \varepsilon \tag{16}$$

$$f(x_i) = x_{i1}^2 - x_{i1} \times x_{i2} \tag{17}$$

$$g_i(t) \sim GP(0, K_t), i = 1, ..., N \tag{18}$$

$$K_t(t_1, t_2) \sim \exp\left(-\frac{1}{2}(t_1 - t_2)^2 \right) \tag{19}$$

$$\varepsilon \sim N(0, 0.1) \tag{20}$$

Note that $f(x_i)$ is non-linear but static across time. The time effect, $g_i(t)$, is unit specific and generated from a Gaussian Process.

We build a simple Gaussian process model that takes covariate , time and unit-index as inputs with the following kernel structure:

$$f(x, t, i) \sim \mathcal{GP}(0, K_x + K_t * K_i) \tag{21}$$

$$K_x(x, x') = \rho_x^2 \exp(-\frac{1}{2\ell_1^2}(x_1 - x_1')^2 - \frac{1}{2\ell_2^2}(x_2 - x_2')^2) \tag{22}$$

$$K_t(t_1, t_2) = \rho_t^2 \exp(-\frac{1}{2\ell_t^2}(t_1 - t_2)^2), \quad K_i(i, i') = 1 \text{ if } i == i' \text{ else } 0 \tag{23}$$

In this simulation, we focus on the static marginal effect and hence decompose the kernel structure to the sum of the covariate kernel and the time kernel. The model could also implement a joint kernel of the covariates and time. We use a zero mean function for simplicity.

We simulated three models: a two-way fixed effects model, a two-way random effects model, and a GPR model. More precisely, we define the two-way fixed effects model and the two-way random effects model, respectively, as

$$y_{it} = \beta_1 x_{i1} + \beta_2 x_{i2} + a_i + b_t + \varepsilon \tag{24}$$

$$y_{it} = \beta_1 x_{i1} + \beta_2 x_{i2} + a_i + b_t + \varepsilon \tag{25}$$

$$a_i \sim \mathcal{N}(0, \sigma_a^2) \tag{26}$$

$$b_t \sim \mathcal{N}(0, \sigma_b^2) \tag{27}$$

We then compared key statistics for model fit: RMSE, coverage, log likelihood, and evidence (marginal likelihood). Theses comparisons, including the 95% confidence intervals, are visually shown in Figure 1, where two-way fixed effects is shown in purple, two-way random effects is shown in blue, and GPR is shown in orange. In all cases, GPR does no worse and often better than two-way fixed effects and two-way random effects. More specifically, GPR significantly outperforms two-way fixed effects and two-way random effects in terms of log likelihood for both covariates and evidence.

Additional statistics from the simulation can be found in the Supplementary Material.

4.2 International shaming and human rights

Scholars debate whether the international community "naming and shaming" a country about its human rights practices might be counterproductive, potentially leading to a backlash and worsening human rights conditions. In one of the most cited papers from this debate, Hafner-Burton (2008) finds that when Amnesty International, an international human rights organization, publicly shames a country about their physical integrity rights, then practices around physical integrity rights will improve while practices around political rights will deteriorate. However, Strezhnev, Kelley and Simmons (2021) fail to replicate her results. In a modified specification, they find that Amnesty International's public shaming on physical integrity rights leads to an increase on the physical integrity rights index the following year, where a higher score indicates worse practices.

In this example, we demonstrate the GP framework for multiple time series using the dataset from Strezhnev, Kelley and Simmons (2021), which corrects some of the original dataset from Hafner-Burton (2008). This dataset spans from 1984 to 2001 and covers 140 countries.[14]. The independent variable of interest is $\mathit{AIShame}_t$, which is binary and indicates whether Amnesty International shamed a country for its physical integrity rights practices in a given year t. The dependent variable of interest is PIRI_{t+1}, the physical integrity rights index in the following year as constructed by Hafner-Burton (2008). PIRI is a composite index of repression indices obtained by adding together the scores for the four physical integrity measures: killing, torture, imprisonment, and disappearances. This yields a variable that ranges from 0 (no violations on any of the four measures) to 8 (worst scores on all four measures), though we follow previous work in treating it as continuous.

The original model from Hafner-Burton (2008) includes up to three lags of the dependent variable that Strezhnev, Kelley and Simmons (2021) found to be insignificant with the updated data, so we follow the lead of Strezhnev, Kelley and Simmons (2021) to only include one lag. Therefore the baseline linear model can be specified as

$$\begin{aligned}\text{PIRI}_{i,t+1} =& \text{AIShame}_{i,t} + \text{PIRI}_{i,t} + \text{CAT}_{i,t} + \text{ICCPR}_{i,t} + \text{Democracy}_{i,t} \\ & + \log(\text{GDPperCapita})_{i,t} + \log(\text{Pop})_{i,t} + \text{CivilWar}_{i,t} + \text{War}_{i,t} + u_t + \varepsilon_{it} \qquad (28)\end{aligned}$$

where *AIShame* indicates whether Amnesty International shamed country i in year t; *PIRI* is the physical integrity rights index for country i in year t; *CAT* and *CCPR* are indicators for whether country i in year t has signed the Convention Against Torture or the International Covenant on Civil and Political Rights, respectively; *Democracy* is an indicator for whether country i has a Polity IV score above 6 in year t; *log(GDPperCapita)* is the logged GDP in country i in year t; *log(Pop)* is the logged population in country i in year t; *CivilWar* and *War* are indicators for whether country i in year t is experiencing a civil war or interstate war, respectively; u_t represents year fixed effects; and ε_{it} is an error term.

In this case, we show the additive properties of GP models by constructing a GP for units

[14] We remove New Zealand and the Netherlands for our analyses since they have a perfect score for the full time period, which slightly changes the baseline results

over time, covariates, and our main variable of interest:

$$\text{PIRI}_{i,t+1} = f_i(t) + f(\mathbf{x}_{i,t}) + f(a_{i,t}) + \varepsilon_{i,t} \tag{29}$$

where $f_i(t)$ is nonlinear unit trends, $f(\mathbf{x}_i)$ is the GP for the covariates, $f(a_i)$ is the GP for the treatment effect of Amnesty International's shaming, and ε_{it} is independent Gaussian error term. We place an independent GP prior on the unit-level time trends:

$$\mathbf{f}_i \sim GP\big(0, \mathbf{K}_u\big) \tag{30}$$

where $\mathbf{K}_u$ follows the squared exponential kernel with time as the only input feature (year). This allows us to exclude both the lagged value and the year fixed effects. The hyperparameters are set to standard starting values and then updated via MAP.[15]

Turning now to the covariates, we specified a zero mean function and then created a kernel by adding two sets of squared exponential kernels together – one set for continuous covariates, demarcated with the subscript $c \in \{\text{GDP}, \text{Pop}, \text{PIRI}\}$, and one set for binary covariates, demarcated with the subscript $b \in \{\text{CAT}, \text{CCPR}, \text{Dem}, \text{Civil}, \text{War}\}$. Like before, we set the to standard starting values, which were updated during the modelling process using MAP.[16]

Finally, we also modeled the "effect" of Amnesty International's shaming practices as:

$$f(\mathbf{a}) \sim GP(\mathbf{0}, \mathbf{K}_a) \tag{31}$$

where we again use the squared exponential kernel as our covariance function, and set the parameters σ^2 and ρ to standard starting values which were updated during the modeling

[15] This gives us the hyperparameters $\{\sigma_i^2, \sigma_t^2, \rho_i, \rho_t\}$ set to $\{0.25, 0.25, 4, 0.01\}$. The MAP values are $\{0.72, 0.72, 4, 0.01\}$.

[16] For the continuous covariates, we have the hyperparameters $\{\sigma^2_{GDP}, \rho_{GDP}, \sigma^2_{Pop}, \rho_{Pop}, \sigma^2_{PIRI}, \rho_{PIRI}\}$ set to $\{4, 0.25, 4, 0.25, 4, 0.25\}$. The MAP values are $\{2.75, 0.012, 2.84, 0.012, 3.08, 0.011\}$. All binary covariate hyperparameters $\{\sigma^2_b, \rho_b\}$ were set to $\{0.01, \text{None}\}$. The MAP values are $\{0.01, 0.69\}$.

Table 1: Comparing results from linear model and GPR on the effect of shaming by Amnesty International on a country's physical integrity rights index.

	Linear Model Results				GPR Results			
	Effect	SE	t-value	p-value	Effect	SE	t-value	p-value
AIShame	0.181	0.076	2.398	0.017	0.038	0.214	0.178	0.57
Model Evidence	-3561.4				-3362.2			
BIC	7360.5				6824.0			

process using MAP.[17]

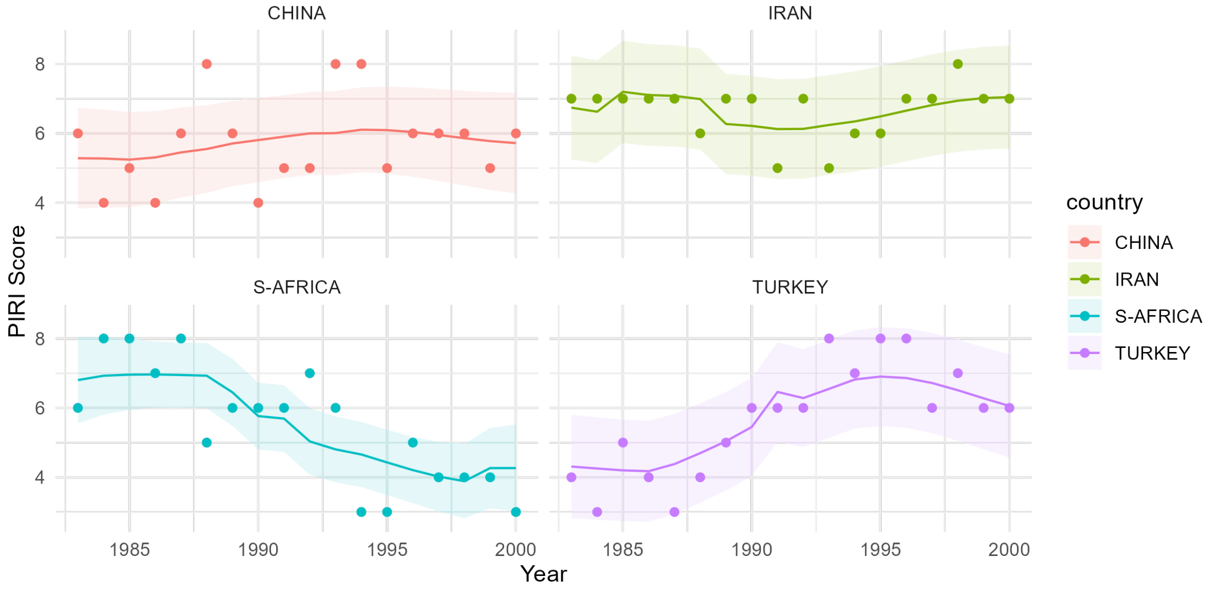
The results of our model and the baseline are shown in Table 1. Using GPR, we find that shaming by Amnesty International leads to a statistically insignificant 0.0038 (se$= 0.214, p > 0.1$) point increase in the physical integrity rights scale. The effect size is almost one fifth the size of the effect found in Strezhnev, Kelley and Simmons (2021), which found that shaming lead to a 0.181 (0.076) point increase on the index. Regarding model fit, our GPR model has slightly higher model evidence of -3362.2 than the original model (-3561.4) and a lower BIC (6824 and 7360.5, respectively).

Beyond the different magnitude of the effect, GPR can be used to model individual country trajectories, in essence tracking the trajectory of individual countries over time. Sample country trajectories with the model are shown in Figure 2, where, in the upper left and going clockwise, we show China, Turkey, South Africa, and Iran. In particular, these trends show the relatively flat indices of China and Iran during this time, while Turkey's practices worsen and South Africa's practices improve.

One concern with GPR is that it is so flexible that it may result in overfitting, leading to poor predictive performance out of sample. To test this, we held out the last five years (1996-2000) of data for 30 randomly selected countries and refitted both the GPR and the linear regression models. We then used our revised models to predict out of sample the last five years for each of the 30 countries. 3 shows six of of these countries, where the points are the observed outcome values and the solid lines show the respective models predicted

[17] This gives us the hyperparamters $\{\sigma^2, \rho$ set to $\{0.01, 0.25\}$. The MAP values are $\{0.01, 0.25\}$.

Figure 2: Individual country trends for PIRI Scores as modeled by GPR for four random countries. The solid lines show the estimated trend from GPR, the points show the observed PIRI score, and the shaded region gives the 95% confidence region. Note that the GPR nicely captures the non-linearity of the PIRI scores over time for countries that remain relatively constant over the study window (Iran), decline or increase (South Africa and Turkey), or are relatively volatile (China).



trend and the shaded regions show the prediction intervals. The top panel shows prediction data from the GPR model, and the bottom panel shows the prediction data from the linear model. While the predicted values are similar across both models, the GPR predictions are smoother with similar coverage – helpful properties when forecasting noisy data.

4.3 Accountability and infant mortality

In our second applicaiton, we show how the kernel structure of GPR allows us to easily add spatial components to a times series analysis. The use of GPR for spatial analysis is not new in Political Science (e.g. Monogan and Gill, 2016), but it is not widely used either. This example further extends this nascent work to show how GPR can simultaneously model both temporal and spatial effects. Specifically, we replicate a model from Lührmann, Marquardt and Mechkova (2020), which uses infant mortality to validate a new measure of government accountability.

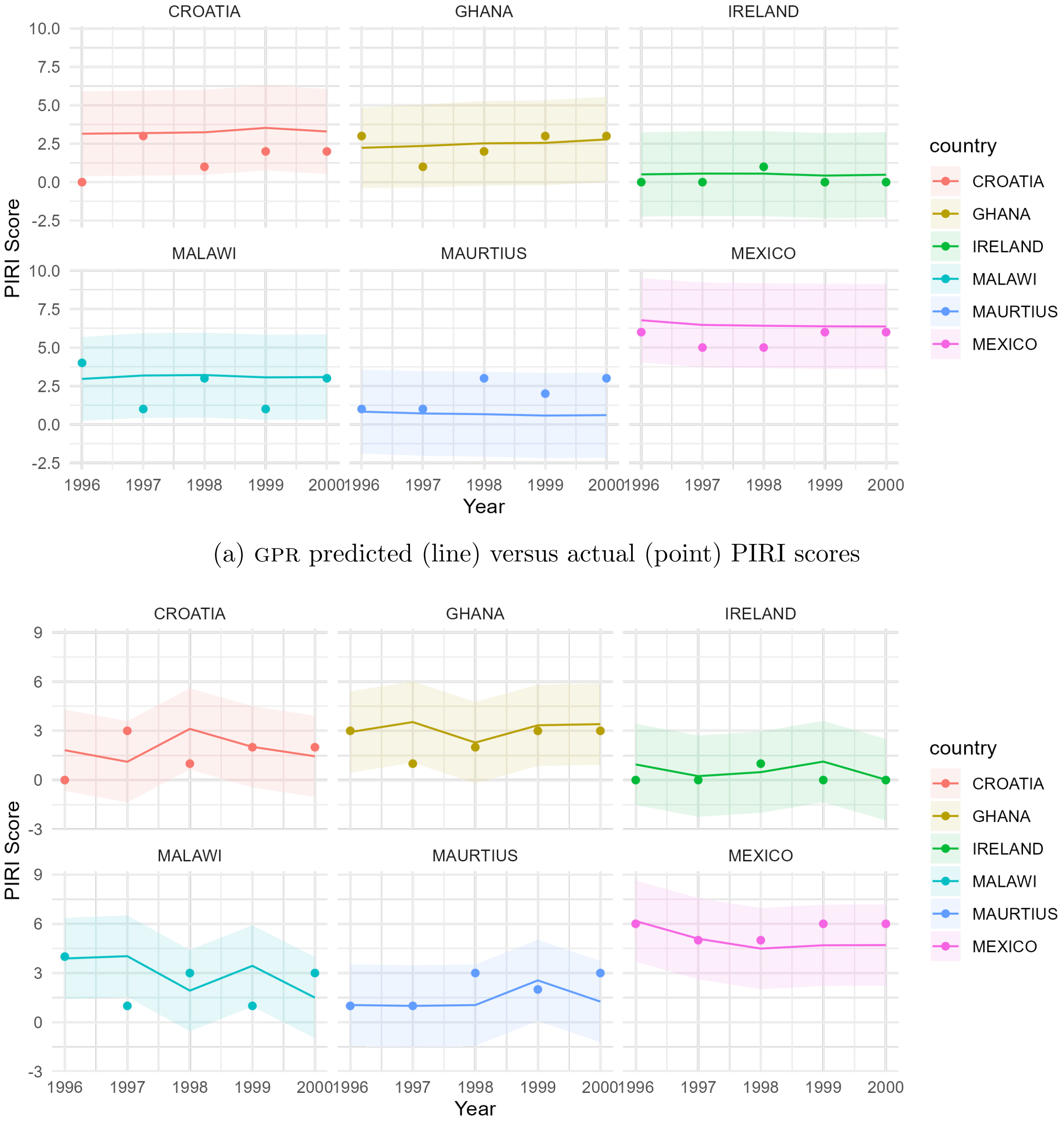
The original model by Lührmann, Marquardt and Mechkova (2020) includes country and year fixed effects to account for spatial and temporal dependence, as well as a variable measuring the regional average of infant mortality. The original model can be mathematically defined as

$$\begin{aligned}
\text{Infant Mortality} = {} & \text{Accountability} + \text{Foreign Aid} + \ln(\text{GDP/capita}) \\
& + \text{Economic Growth} + \text{Resource Dependence} \\
& + \text{Economic Inequality} + \ln(\text{Population}) + \text{Urbanization} \\
& + \text{Political Violence} + \text{Communist} \\
& + \text{Regional Infant Mortality Average } + \text{Political Corruption Index} \\
& + u_i + v_t + \varepsilon_{it} \qquad (32)
\end{aligned}$$

where u_i is the country fixed effects and v_t is the series of year fixed effects.

When replicating these spatial and temporal dependencies using GPR, we leverage the

Figure 3: Prediction intervals using GPR (Panel A) and OLS regression (Panel B) for six randomly selected countries: Croatia, Ghana, Ireland, Malawi, Mauritius, and Mexico. For each plot, the y axis indicates the PIRI score and the x axis indicates the year. The points show the observed data point, the solid line shows the predicted value, and the prediction interval is shaded. While the predicted values are similar across both models, the GPR predictions are smoother with similar coverage.



(a) GPR predicted (line) versus actual (point) PIRI scores

(b) OLS regression predicted (line) versus actual (point) PIRI scores

Table 2: Comparing results from Lührmann, Marquardt and Mechkova (2020) and GPR on the effect of the accountability on infant mortality

	Linear Model Results				GPR Results			
	Effect	SE	t-value	p-value	Effect	SE	t-value	p-value
Accountability	−4.34	0.35	−12.38	1.7e-24	−4.29	0.38	−10.87	2.2e-20
Model Evidence	−15435.67				−10117.09			
BIC	32664.42				1537.29			

flexibility of the kernel to learn the optimal fit, doing away with the fixed effects and spatial averages. In this case, the GPR model can be defined as

$$\begin{aligned}
\text{Infant Mortality} &= \text{Accountability} + \text{Foreign Aid} + \ln(\text{GDP/capita}) \\
&\quad + \text{Economic Growth} + \text{Resource Dependence} \\
&\quad + \text{Economic Inequality} + \ln(\text{Population}) + \text{Urbanization} \\
&\quad + \text{Political Violence} + \text{Communist} \\
&\quad + \text{Regional Infant Mortality Average } + \text{Political Corruption Index} \\
&\quad + u_i(t) \qquad (33) \\
u_i(t) &\sim GP\big(b_i, K(x) + K(t) * K(g)\big) \qquad (34)
\end{aligned}$$

where the GP is defined to have a zero mean, and each K follows the squared exponential kernel, taking as inputs the model's covariates, time, and longitude and latitude, respectively. We again leverage the additive properties of Gaussian processes to combine the three kernels together, multiplying the kernels for space and time to allow for maximum dependencies. The hyperparameters are set to standard starting values and updated using MAP. The starting values and MAP values are shown in the Supplementary Material.

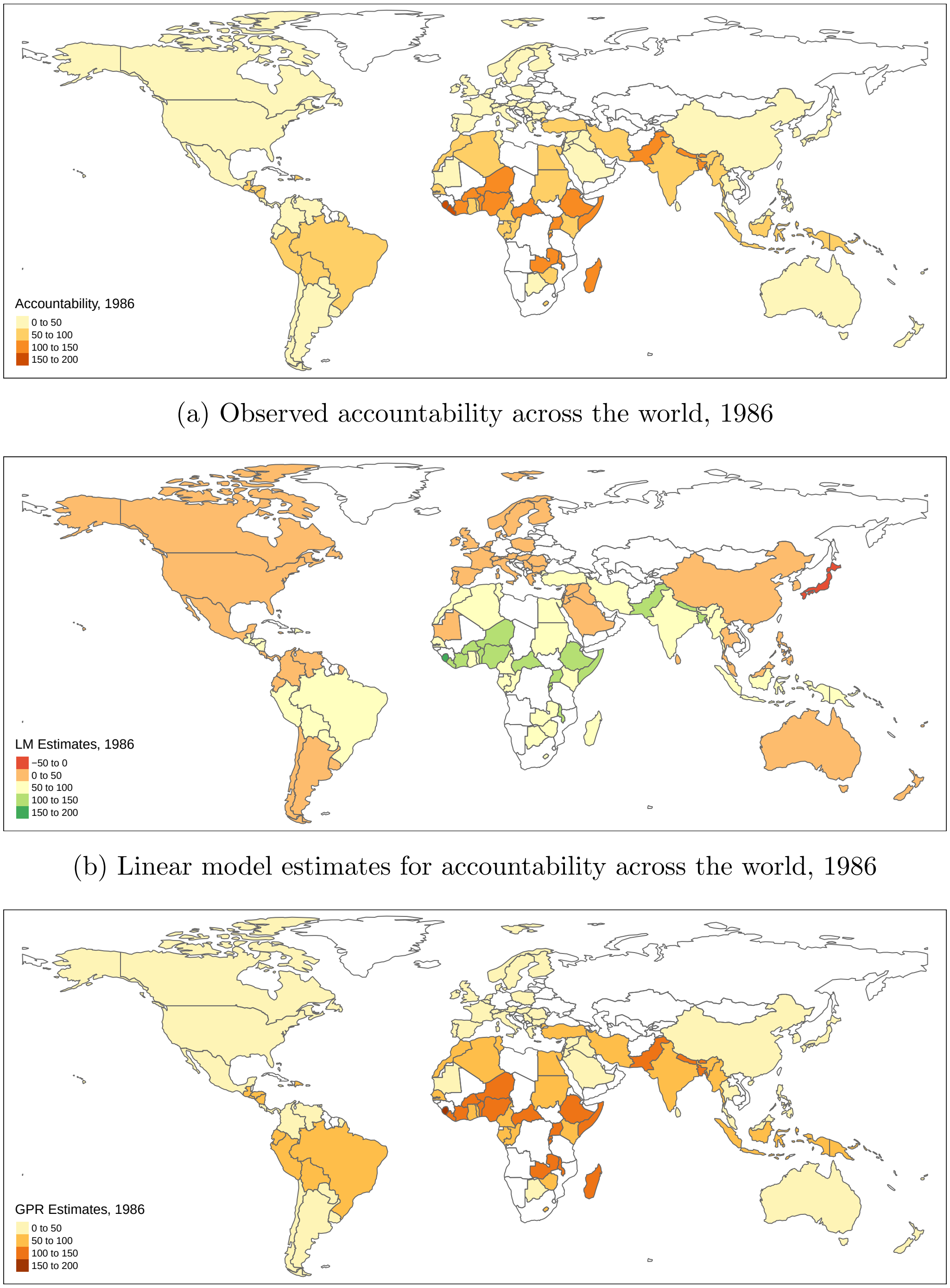
For our analyses, we use the data from Cook, Hays and Franzese (2023), who also replicate Lührmann, Marquardt and Mechkova (2020), since their data includes the latitude and longitude necessary for the geospatial component of the kernel. The GPR gradient and linear regression results are shown in Table 2. Using the GP model, we find that a one unit increase on the accountability index is associated with a 4.29 percentage point decrease in the infant mortality rate, on average and all else constant. This is comparable to the effect found in Lührmann, Marquardt and Mechkova (2020), which found that a one unit increase on the accountability index was associated with a 4.34 percentage point decrease in the infant mortality index. Our standard error is also comparable to the original finding. Regarding model fit, our GPR model has higher model evidence and a lower BIC score.

Despite these seemingly similar coefficients and standard errors, these improved fit statistics (model evidence and BIC) are important. In practice, they allow the GPR model to offer individual country estimates that are much more similar to the observed values than the linear model would allow. This similarity is demonstrated in Figure 4, which compares the observed *Accountability* measure (top panel) to the estimates from the linear regression (middle panel) and the GPR (bottom panel). Since the measure varies across time and space, we are using 1986 as an example snapshot in time since it is the mean year in the dataset. Note that *Accountability* is an aggregation of several positive value indices and therefore can only take on positive values, however the linear model estimates some countries to have negative values. In contrast, the map showing the estimates from the GPR is very similar to the map showing the observed measures of accountability. This demonstrates the utility of the country trend $u_i(t)$ that incorporates space and time through its kernel function over linear two-way fixed effects when examining estimates at the country level.

5 Conclusion

In this paper, we have shown how GPR allows social scientists to model quantities of interest in a structured yet flexible framework. By seamlessly integrating prior knowledge and observations, Gaussian processes empower researchers to make informed predictions and insights while maintaining appropriate bounds for underlying uncertainty. Since our many models

Figure 4: Comparison of the observed accountability measure (Panel A), the estimated accountability measure from a linear model (Panel B), and the estimated accountability measure from a GPR model (Panel C), using 1986 as an example snapshot in time. Note that the linear model estimates negative values, which are invalid for the accountability measure, while the GPR estimates are similar to the observed accountability measure.



(a) Observed accountability across the world, 1986

(b) Linear model estimates for accountability across the world, 1986

(c) GPR estimates for accountability across the world, 1986 already widely used in current literature are actually special cases of GPR, it offers a powerful framework for extending basic practices to allow for far more flexibility than traditional le. This underscores the universality and broad applicability of the method, and potentially moves the discipline toward a more cohesive understanding of the varied models in our arsenal. Thus, a primary goal for this paper is to provide an approachable explanation of this framework.

As we continue to develop this paper, we hope to add several new components to help make GPR more accessible to social scientists. First and foremost is to add more examples, including a cross sectional case, a single stream time series, and a model with binary dependent variables. Key to these additional examples would be expanding the discussion on kernel functions: there are many alternatives to the squared exponential kernel, and we hope to help guide scholars through choosing the best kernel for their particular choices. Furthermore, we hope to add an appendix explaining how to use the software `gpytorch`, since admittedly it can have a steep learning curve.

We believe that GPR could also be integral in the development of new methods in Political Science. For example, GPR has been used to relax the parallel trends assumption for difference-in-differences analysis with one treatment period (Chen et al., 2023), but this could also be extended to differences-in-differences with staggered adoption or when units enter and exit treatment status. We also believe that our current example demonstrating GPR's use for geospatial analysis is the most basic form, and that the kernel structure could be refined to better leverage spatial effects. Given that GPR was initially designed for spatial predictions in the mining industry (Cressie, 2015), we are optimistic about the potential applications to Political Science. More fundamentally, GPR could also provide an elegant way to address the varied needs for interactions in Political Science through the kernel function. Given the importance and widespread usage of interactions in Political Science, we believe this deserves a stand alone paper.

In conclusion, we believe that Gaussian processes are a powerful analytical tool that embraces the complexities of Political Science and enriches social science's methodological arsenal. We believe that the relative complexity of these models is their key barrier to entry, as this paper is an initial step in breaking down that barrier. With this greater understanding and hopefully wider usage, the inherent advantages of flexibility coupled with interpretability will contribute to more nuanced understandings of our discipline's quantitative pursuits.

References

Acemoglu, Daron, Simon Johnson, James A Robinson and Pierre Yared. 2008. "Income and Democracy." *American Economic Review* 98(3):808–842.

Aglietti, Virginia, Theodoros Damoulas, Mauricio Álvarez and Javier González. 2020. "Multi-task Causal Learning with Gaussian Processes." *Advances in Neural Information Processing Systems* 33:6293–6304.

Ahmed, Ali T. and David Stasavage. 2020. "Origins of Early Democracy." *American Political Science Review* 114(2):502–518.

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

Arbour, David, Eli Ben-Michael, Avi Feller, Alex Franks and Steven Raphael. 2021. "Using Multitask Gaussian Processes to Estimate the Effect of a Targeted Effort to Remove Firearms." *arXiv preprint arXiv:2110.07006* .

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31(3):337–351.

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized random forests." *The Annals of Statistics* 47(2):1148 – 1178.

Barari, Soubhik, Christopher Lucas and Kevin Munger. 2021. "Political Deepfakes Are As Credible As Other Fake Media And (Sometimes) Real Media." *OSF preprints* 13.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41(2):641–674.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):21–35.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2):596–627.

Bisbee, James. 2019. "BARP: Improving Mister P Using Bayesian Additive Regression Trees." *American Political Science Review* 113(4):1060–1065.

Blair, Robert A., Jessica Di Salvatore and Hannah M. Smidt. 2023. "UN Peacekeeping and Democratization in Conflict-Affected Countries." *American Political Science Review* p. 1–19.

Broniecki, Philipp, Lucas Leemann and Reto Wüest. 2022. "Improved Multilevel Regression with Poststratification through Machine Learning (autoMrP)." *The Journal of Politics* 84(1):597–601.

Brooks, Steve, Andrew Gelman, Galin Jones and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. CRC Press.

Cameron, A Colin, Jonah B Gelbach and Douglas L Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90(3):414–427.

Carroll, Robert J. and Brenton Kenkel. 2019. "Prediction, Proxies, and Power." *American Journal of Political Science* 63(3):577–593.

Casella, George, Malay Ghosh, Jeff Gill and Minjung Kyung. 2010. "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis* 5(2):369 – 411.

Chen, Yehu, Annamaria Prati, Jacob Montgomery and Roman Garnett. 2023. A Multi-Task Gaussian Process Model for Inferring Time-Varying Treatment Effects in Panel Data. In *International Conference on Artificial Intelligence and Statistics*. PMLR pp. 4068–4088.

Chen, Yehu, Roman Garnett and Jacob M. Montgomery. 2023. "Polls, Context, and Time: A Dynamic Hierarchical Bayesian Forecasting Model for US Senate Elections." *Political Analysis* 31(1):113–133.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21(1):C1–C68.

Chiu, Albert and Yiqing Xu. 2023. "Bayesian Rule Set: A Quantitative Alternative to Qualitative Comparative Analysis." *The Journal of Politics* 85(1):280–295.

Cook, Scott J., Jude C. Hays and Robert J. Franzese. 2023. "STADL Up! The Spatiotemporal Autoregressive Distributed Lag Model for TSCS Data Analysis." *American Political Science Review* 117(1):59–79.

Cranmer, Skyler J. and Bruce A. Desmarais. 2017. "What Can We Learn from Predictive Modeling?" *Political Analysis* 25(2):145–166.

Cressie, Noel. 2015. *Statistics for Spatial Data*. John Wiley & Sons.

de Kadt, Daniel and Horacio A. Larreguy. 2018. "Agents of the Regime? Traditional Leaders and Electoral Politics in South Africa." *The Journal of Politics* 80(2):382–399.

Di Salvatore, Jessica. 2019. "Peacekeepers against Criminal Violence—Unintended Effects of Peacekeeping Operations?" *American Journal of Political Science* 63(4):840–858.

Esarey, Justin and Andrew Menger. 2019. "Practical and Effective Approaches to Dealing With Clustered Data." *Political Science Research and Methods* 7(3):541–559.

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

Fong, Christian, Chad Hazlett and Kosuke Imai. 2018. "Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements." *The Annals of Applied Statistics* 12(1):156–177.

Fong, Christian and Justin Grimmer. 2023. "Causal Inference with Latent Treatments." *American Journal of Political Science* 67(2):374–389.

Fong, Christian and Matthew Tyler. 2021. "Machine Learning Predictions as Regression Covariates." *Political Analysis* 29(4):467–484.

Gill, Jeff. 2021*a*. "Measuring Constituency Ideology Using Bayesian Universal Kriging." *State Politics & Policy Quarterly* 21(1):80–107.

Gill, Jeff. 2021*b*. "Political Science Is a Data Science." *The Journal of Politics* 83(1):1–7.

Girolami, Mark and Simon Rogers. 2006. "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors." *Neural Computation* 18(8):1790–1817.

Gohdes, Anita R. 2020. "Repression Technology: Internet Accessibility and State Violence." *American Journal of Political Science* 64(3):488–503.

Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413–434.

Hafner-Burton, Emilie M. 2008. "Sticks and Stones: Naming and Shaming the Human Rights Enforcement Problem." *International Organization* 62(4):689–716.

Hainmueller, Jens and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2):143–168.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27(2):163–192.

Hazlett, Chad. 2020. "Kernel Balancing: A Flexible Non-Parametric Weighting Procedure for Estimating Causal Effects." *Statistica Sinica* 30(3):1155–1189.

Hazlett, Chad and Yiqing Xu. 2018. "Trajectory Balancing: A General Reweighting Approach to Causal Inference With Time-Series Cross-Sectional Data." *Available at SSRN 3214231* .

Hensman, James, Alexander Matthews and Zoubin Ghahramani. 2015. Scalable Variational Gaussian Process Classification. In *Artificial Intelligence and Statistics*. PMLR pp. 351–360.

Hensman, James, Nicolo Fusi and Neil D Lawrence. 2013. "Gaussian Processes for Big Data." *arXiv preprint arXiv:1309.6835* .

Hill, Daniel W. and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3):661–687.

Imai, Kosuke, James Lo and Jonathan Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review* 110(4):631–656.

Jackson, John E. 2020. "Corrected Standard Errors with Clustered Data." *Political Analysis* 28(3):318–339.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2017. “How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument.” *American Political Science Review* 111(3):484–501.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133(1):237–293.

Knox, Dean and Christopher Lucas. 2021. “A Dynamic Model of Speech for the Social Sciences.” *American Political Science Review* 115(2):649–666.

Knox, Dean, Christopher Lucas and Wendy K Tam Cho. 2022. “Testing Causal Theories with Learned Proxies.” *Annual Review of Political Science* 25:419–441.

Lee, Jaehoon, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington and Jascha Sohl-Dickstein. 2018. “Deep Neural Networks as Gaussian Processes.” *arXiv preprint arXiv:1711.00165* .

Li, Qi and Jeffrey S Racine. 2010. “Smooth Varying-Coefficient Estimation and Inference for Qualitative and Quantitative Data.” *Econometric Theory* 26(6):1607–1637.

Lührmann, Anna, Kyle L. Marquardt and Valeriya Mechkova. 2020. “Constraining Governments: New Indices of Vertical, Horizontal, and Diagonal Accountability.” *American Political Science Review* 114(3):811–820.

Mitts, Tamar, Gregoire Phillips and Barbara F. Walter. 2022. “Studying the Impact of ISIS Propaganda Campaigns.” *The Journal of Politics* 84(2):1220–1225.

Mohanty, Pete and Robert Shaffer. 2019. “Messy Data, Robust Inference? Navigating Obstacles to Inference with bigKRLS.” *Political Analysis* 27(2):127–144.

Monogan, James E and Jeff Gill. 2016. “Measuring State and District Ideology with Spatial Realignment.” *Political Science Research and Methods* 4(1):97–121.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18(2):245–270.

Montgomery, Jacob M. and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Moser, Scott, Abel Rodríguez and Chelsea L. Lofland. 2021. "Multiple Ideal Points: Revealed Preferences in Different Domains." *Political Analysis* 29(2):139–166.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24(1):87–103.

Paglayan, Agustina S. 2021. "The Non-Democratic Roots of Mass Education: Evidence from 200 Years." *American Political Science Review* 115(1):179–198.

Rasmussen, Carl Edward and Christopher KI Williams. 2006. *Gaussian Processes for Machine Learning*. Vol. 2 MIT press Cambridge, MA.

Ratkovic, Marc and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 25(1):1–40.

Ratkovic, Marc and Dustin Tingley. 2023. "Estimation and Inference on Nonlinear and Heterogeneous Effects." *The Journal of Politics* 85(2):421–435.

Reynolds, Douglas. 2009. Gaussian Mixture Models. In *Encyclopedia of Biometrics*, ed. Stan Z. Li and Anil Jain. Boston, MA: Springer pp. 659–663.

Sexton, Joseph and Petter Laake. 2009. "Standard errors for bagged and random forest estimators." *Computational Statistics & Data Analysis* 53(3):801–811. Computational Statistics within Clinical Research.

Shiraito, Yuki, James Lo and Santiago Olivella. 2023. "A Nonparametric Bayesian Model for Detecting Differential Item Functioning: An Application to Political Representation in the US." *Political Analysis* 31(3):430–447.

Streeter, Shea. 2019. "Lethal Force in Black and White: Assessing Racial Disparities in the Circumstances of Police Killings." *The Journal of Politics* 81(3):1124–1132.

Strezhnev, Anton, Judith G Kelley and Beth A Simmons. 2021. "Testing for Negative Spillovers: Is Promoting Human Rights Really Part of the "Problem"?" *International Organization* 75(1):71–102.

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

Torres, Michelle and Francisco Cantú. 2022. "Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data." *Political Analysis* 30(1):113–131.

Witty, Sam, Kenta Takatsu, David Jensen and Vikash Mansinghka. 2020. Causal Inference Using Gaussian Processes with Structured Latent Confounders. In *International Conference on Machine Learning*. PMLR pp. 10313–10323.

A Supplementary Material

A.1 Additional Simulation Statistics

mean	std	model	measure
0.313007	0.041923	2FE	RMSE_x1
0.95456	0.035763	2FE	COVERAGE_x1
-0.29531	0.306315	2FE	LL_x1
0.164036	0.020831	2FE	RMSE_x2
0.97152	0.019817	2FE	COVERAGE_x2
-0.42775	0.318487	2FE	LL_x2
0.325387	0.038425	2RE	RMSE_x1
0.86264	0.096159	2RE	COVERAGE_x1
0.834301	0.264644	2RE	LL_x1
0.160712	0.019385	2RE	RMSE_x2
0.90396	0.049791	2RE	COVERAGE_x2
1.012222	0.268905	2RE	LL_x2
0.162702	0.057474	GPR	RMSE_x1
0.98692	0.014851	GPR	COVERAGE_x1
1.474003	0.300649	GPR	LL_x1
0.070989	0.020534	GPR	RMSE_x2
0.9886	0.012162	GPR	COVERAGE_x2
2.134344	0.31437	GPR	LL_x2
-1.18376	0.073836	2FE	EVIDENCE
-1.22316	0.074353	2RE	EVIDENCE
0.49476	0.028307	GPR	EVIDENCE

A.2 Accountability and infant mortality hyperparameters

Hyperparameter	Starting Value(s)	MAP Value
σ_t^2	6.4	6.4
σ_g^2	1	1
$\rho_{t,g}$	10	41.7
$m_{\text{covariates}}$	$\{-4.34, -0.05, -10.14, 0.03, 0.04, -0.08, -18.24, -0.11, 0.64, -3.71\}$	$\{-4.3, -0.03, -0.01, 0.007, 0.01, 0.008, -0.18, -0.008, 0.61, -3.36\}$
$\sigma^2_{\text{covariates}}$	$\{3.61, 3.61, 3.61, 3.61, 3.61, 3.61, 3.61, 3.61, 3.61, 3.61\}$	$\{3.99, 5.03, 0.43, 5.0, 4.95, 4.8621, 0.37, 4.8, 4.8, 0.26\}$
$\rho_{\text{covariates}}$	None	7.27

A.3 First difference and AMCE under GP priors

Priors are essential in Bayesian methods as they encode either the initial belief of the parameters of interests or serve as regularization that penalizes overfitting. Here we derive the analytical forms of our first-difference and derivative estimators for AMCE under our GP framework. For the purpose of demonstration, we first consider a general GP prior $\mathbf{f} \sim \mathcal{GP}(\boldsymbol{\mu}, \mathbf{K})$ on a d-dim input space $\mathcal{X}$. WOLG, suppose we are interested in finding the marginal effect of a binary variable $\boldsymbol{x}_1 = 0, 1$. As GP is closed under linear transformation, the induced prior on first-difference estimator is

$$\begin{aligned}
\hat{\pi}(\boldsymbol{x}_1 = 1, \boldsymbol{x}_1 = 0; \boldsymbol{x}_{-1}) &= \mathbf{f}(\boldsymbol{x}_1 = 1, \boldsymbol{x}_{-1}) - \mathbf{f}(\boldsymbol{x}_1 = 0, \boldsymbol{x}_{-1}) && (35)\\
&\sim \mathcal{GP}\big(\boldsymbol{\mu}(\boldsymbol{x}_1 = 1, \boldsymbol{x}_{-1}) - \boldsymbol{\mu}(\boldsymbol{x}_1 = 0, \boldsymbol{x}_{-1}), && (36)\\
&\qquad \mathbf{K}(\boldsymbol{x}_1 = 1, \boldsymbol{x}_{-1}; \boldsymbol{x}_1 = 1, \boldsymbol{x}_{-1}) + \mathbf{K}(\boldsymbol{x}_1 = 0, \boldsymbol{x}_{-1}; \boldsymbol{x}_1 = 0, \boldsymbol{x}_{-1})\big) && (37)
\end{aligned}$$

For zero mean $\boldsymbol{\mu}(\boldsymbol{x}) = 0$ and RBF kernel $\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}') = \rho^2 \exp\big(-(\boldsymbol{x} - \boldsymbol{x}')^2/2/\ell^2\big)$, the induced prior reduces to a zero-meaned Gaussian $\hat{\pi}(\boldsymbol{x}_1 = 1, \boldsymbol{x}_1 = 0; \boldsymbol{x}_{-1}) \sim \mathcal{N}(0, 2\rho^2)$. Therefore, the prior variance on the first-difference estimator is proportional to the output scales: the larger ρ^2 is, the flatter or more uninformative prior we have on the first difference.

Secondly, the AMCE under GP priors also reduces a Gaussian. Now assume $\boldsymbol{x}_1$ is continuous, so AMCE at $\boldsymbol{x}_1 = x, \boldsymbol{x}_{-1}$ can be written as the derivative of $\mathbf{f}$:

$$\begin{aligned}
\hat{\pi}(\boldsymbol{x}_1 = x, \boldsymbol{x}_{-1}) &= \frac{\partial}{\partial \boldsymbol{x}_1}\mathbf{f}(\boldsymbol{x}_1 = x, \boldsymbol{x}_{-1}) && (38)\\
&\sim \mathcal{GP}\big(\frac{\partial}{\partial \boldsymbol{x}_1}\boldsymbol{\mu}(\boldsymbol{x}_1 = x, \boldsymbol{x}_{-1}), \frac{\partial^2}{\partial \boldsymbol{x}_1 \partial \boldsymbol{x}_1'}\mathbf{K}(\boldsymbol{x}_1 = x, \boldsymbol{x}_{-1}; \boldsymbol{x}_1' = x, \boldsymbol{x}_{-1}')\big) && (39)
\end{aligned}$$

For a linear mean function $\boldsymbol{\mu}(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$ and RBF kernel $\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}') = \rho^2 \exp\big(-(\boldsymbol{x}_{-1} - \boldsymbol{x}_{-1}')^2/2/\ell^2\big) \exp\big(-(\boldsymbol{x}_1 - \boldsymbol{x}_1')^2/2/\ell^2\big)$, we have again an induced Gaussian prior $\hat{\pi}(\boldsymbol{x}_1 = x, \boldsymbol{x}_{-1}) \sim \mathcal{N}(\boldsymbol{\beta}_1, \rho^2/\ell^2)$. Again we can see that the larger ρ^2 is, the flatter or more uninformative prior we have on the derivative of $\mathbf{f}$. In addition, the length sfcale ℓ inversely affects the prior variance ρ^2/ℓ^2 on first derivative, meaning that the larger ℓ is, the smoother the response surface becomes and the tighter the prior variance is.

A.4 GP kernels for common panel analysis models

Common panel analysis models can be written under GP framework. Suppose we have panel data indexed by unit i and time t.

Panel model	Original model	GP model
Fixed effect (ungrouped)	α_i	$\mu(i) = \alpha_i$
Fixed effect (grouped)	$\alpha_g(i \in g)$	$\mu(i) = \alpha_i(i \in g)$
Random effect (ungrouped)	$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$	$K(i; i') = \sigma_\alpha^2 \mathbb{I}[i = i']$
Random effect (grouped)	$\alpha_g(i \in g) \sim \mathcal{N}(0, \sigma_g^2)$	$K(g; g') = \sigma_g^2 \mathbb{I}[g = g']$
Temporal effect	γ_t	$\mu(t) = \gamma_t$
Clustered standard errors	$\varepsilon_g \sim \mathcal{N}(0, \sigma_g^2)$	$K(g, g') = \sigma_g^2 \mathbb{I}[g = g']$
Linear covariates	$\boldsymbol{\beta}^T \boldsymbol{x}, \boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$	$K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{\Sigma} \boldsymbol{x}'$