Political Analysis (2026), **34**, 569–588
doi:10.1017/pan.2026.10032

PA

ARTICLE

Inference at the Data's Edge: Gaussian Processes for Estimation and Inference in the Face of Extrapolation Uncertainty

Soonhong Cho[1], Doeun Kim[1] and Chad Hazlett[1,2]

[1]Political Science, UCLA, USA; [2]Political Science, Statistics & Data Science, UCLA, USA

Corresponding author: Chad Hazlett; Email: chazlett@ucla.edu

(Received 12 November 2024; revised 25 October 2025; accepted 25 October 2025; published online 10 March 2026)

Abstract

Many inferential tasks involve fitting models to observed data and predicting outcomes at new covariate values, requiring interpolation or extrapolation. Conventional methods select a single best-fitting model, discarding fits that were similarly plausible in-sample but would yield sharply different predictions out-of-sample. Gaussian processes (GPs) offer a principled alternative. Rather than committing to one conditional expectation function, GPs deliver a posterior distribution over outcomes at any covariate value. This posterior effectively retains the range of models consistent with the data, widening uncertainty intervals where extrapolation magnifies divergence. In this way, the GP's uncertainty estimates reflect the implications of extrapolation on our predictions, helping to tame the "dangers of extreme counterfactuals" (King and Zeng, 2006). The approach requires (i) specifying a covariance function linking outcome similarity to covariate similarity and (ii) assuming Gaussian noise around the conditional expectation. We provide an accessible introduction to GPs with emphasis on this property, along with a simple, automated procedure for hyperparameter selection implemented in the R package `gpss`. We illustrate the value of GPs for capturing counterfactual uncertainty in three settings: (i) treatment effect estimation with poor overlap, (ii) interrupted time series requiring extrapolation beyond pre-intervention data, and (iii) regression discontinuity designs where estimates hinge on boundary behavior.

Keywords: causal inference; Gaussian process regression; machine learning; regression discontinuity; interrupted time series; positivity

Edited by: Dorothy Kronick

1. Introduction

Many inferential tasks rely on training models to predict outcomes from observed covariates, then extending these predictions to new covariate locations through interpolation or extrapolation. During model fitting, different parameter choices can often produce models that fit similarly well in-sample but diverge sharply in data-sparse regions or beyond the support of the data. Nevertheless, standard practice selects only a single best-fitting conditional expectation function (CEF). When extrapolating, uncertainty intervals from this single model cannot account for the different predictions that might have been produced by other plausible fits. This approach thus omits a source of uncertainty we refer to as "extrapolation uncertainty." The problem is especially acute in settings requiring predictions in data-sparse regions, such as covariate-adjusted comparisons lacking common support, interrupted time series (ITS), and regression discontinuity (RD).

We recommend Gaussian process (GP) regression as one principled approach to address this concern in such settings. GPs yield a posterior distribution for outcomes at each covariate location,

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Political Methodology.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2026.10032

not just a single best-fitting CEF estimate. In data-sparse regions, this posterior effectively retains and encompasses the range of plausible fits, widening uncertainty intervals to reflect their divergence.

Despite their appeal, GPs remain relatively unfamiliar in the social sciences, with only a few recent exceptions (e.g., Ben-Michael *et al.* 2023; Branson *et al.* 2019; Hinne *et al.* 2022; Ornstein and Duck-Mayr 2022; Prati *et al.* 2023). We therefore begin with an accessible introduction. One barrier to wider adoption has been the practical challenge of tuning three highly interdependent hyperparameters. We simplify this by fixing one parameter, selecting a second based only on the covariates, and estimating the third with an automated line search. This fully automated procedure, implemented in the R package `gpss`, has proven stable in our simulations and applications.

We illustrate the practical value of GPs in three settings. First, in treatment–control comparisons, we illustrate how GPs handle failures of overlap or common support in terms of their uncertainty intervals for both average and conditional effect estimates. Second, in ITS designs, GPs provide a principled framework for extrapolating beyond pre-intervention data while making specification choices explicit. Third, in RD designs, GPs approach edge estimation differently from standard optimal-bandwidth local polynomial methods (e.g., Calonico, Cattaneo, and Titiunik 2014), relying on the GP's posterior distribution for the outcome at the cutoff from models trained on either side. This proves especially valuable in samples smaller than those suitable for the local polynomial approach. In particular, the GP approach reduces the influence of noisy observations near the cutoff that can drive extreme polynomial fits, improving bias and RMSE while maintaining good coverage, often with shorter intervals.

2. The GP Framework

We briefly outline the GP framework, drawing on concepts familiar to most quantitative social scientists. For a classical treatment, readers may also wish to reference Rasmussen and Williams (2006). Our discussion is motivated by—and emphasizes the implications for—how we think about uncertainty estimation for new observations that may lie at varying distances from the observed data. Throughout, we assume (training) data consisting of n independent tuples $\{Y_i, X_i\}_{i=1}^{n}$ drawn from a common joint distribution or data-generating process (DGP). Here, $Y_i \in \mathcal{Y}$ denotes a scalar outcome, and $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes a vector of covariates.

A Distributional Outcome: Consider first the outcome Y, a vector of n observations ($i \in \{1, \ldots, n\}$) drawn from a multivariate normal distribution $Y \sim \mathcal{N}(\mu, \Sigma)$. We defer momentarily how μ and Σ will be determined.

Smoothness, Covariance, and Functional Form: Second, suppose that Y_i and Y_j should be more similar in value when X_i and X_j are similar in value. This can be restated as requiring that the covariance of Y_i with Y_j (across varying DGP) must be higher when X_i and X_j are more similar. More formally, choose a kernel function $k(\cdot,\cdot)$ governing the relationship between the covariance of Y_i with Y_j and the distance between X_i and X_j according to $cov(Y_i, Y_j) = \sigma_f k(X_i, X_j)$. This can take a particular form, such as $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$ (the Gaussian kernel) whereby observations i and j will have maximal covariance when $X_i = X_j$, and covariance decreases toward zero as X_i and X_j become distant. Other kernel functions can be applied to various purposes we discuss below.

Consider the kernel matrix $\mathbf{K}$, containing all the pairwise kernel evaluations, i.e., $\mathbf{K}_{i,j} = k(X_i, X_j)$. As this describes the full variance–covariance matrix of the vector Y, we now have the prior model

$$Y \mid X \sim \mathcal{N}(\mu, \sigma_f \mathbf{K}), \tag{1}$$

where μ is a length-n vector giving the prior mean at each point, and σ_f is a scaling parameter governing the prior variance. Further, the CEF is not expected to go through each observation; rather, the observations are spread around their conditional expectations with some irreducible error, σ^2. This revises our prior model to

$$Y \mid X \sim \mathcal{N}(\mu, \sigma_f \mathbf{K} + \sigma^2 I). \tag{2}$$

https://doi.org/10.1017/pan.2026.10032

Simplification: As a prior model, we emphasize that this represents only possible draws of the function before observing any data. We make two simplifications. First, we set $\mu = 0$ at all points. This is because (i) having not yet seen the data, we have no preference for any particular choice of μ and (ii) we globally demean Y for modeling purposes (adding the mean back to predictions afterward). As we elaborate in Section 2.1, this does not effectively constrain the shape of the posterior CEF, which will shape itself around the data as we describe below.

Second, we set $\sigma_f = 1$. Recall that σ_f governs the prior variance of the underlying function at any point $X = x$, before observing data. With this choice, the solution for the posterior CEF at the training points becomes $\mathbf{K}(\mathbf{K}+\sigma^2 I)^{-1} Y$ (see Expression (4) below). This exactly reproduces the solution for the CEF obtained through kernel ridge regression (e.g., Hainmueller and Hazlett 2014), which rests only on regularized loss-minimization in a function space without invoking any notion of prior variance, except implicitly through the chosen degree of regularization. Fixing $\sigma_f = 1$ can thus be understood not as a restriction but as parameterization that will rescale and alter the interpretation of σ^2. Practically speaking, this greatly improves identifiability of σ^2, allowing it to be tuned efficiently via a simple line search (see Section 2.1). Together, these adjustments leave us with a prior model of

$$Y \mid X \sim \mathcal{N}(0, \mathbf{K} + \sigma^2 I). \tag{3}$$

Conditioning on the Data: If we are given X_j and asked to guess Y_j, we can only guess Y_j is distributed $N(0,\sigma^2)$; i.e., we are maximally ignorant (Figure 1, blue). However, suppose we observe another unit, $\{X_i, Y_i\}$. The distance between X_i and X_j (or more properly, $k(X_i,X_j)$) tells us how Y_i and Y_j will covary. If X_j is close to X_i, we can guess that Y_j is more likely to be close in value to Y_i. If Y_i takes a large value, for example, we can expect Y_j to be larger, and this information somewhat reduces our uncertainty. Conditioning on the data in this way leads to a (posterior) distribution for the unobserved Y_j (Figure 1, red).

Scaling this logic, consider n (training) observations $\{X_i, Y_i\}_{i=1}^n$ and n^* test observations with observed covariates $\{X^*\}$ and unknown outcome Y^* (an n^*-length vector). Our belief about Y^* is then given by

$$Y^* \mid X^*, Y, X \sim \mathcal{N}(\mathbf{K}_*(\mathbf{K}+\sigma^2 I_n)^{-1} Y,\ \mathbf{K}_{*,*} + \sigma^2 I_{n^*} - \mathbf{K}_*(\mathbf{K}+\sigma^2 I_n)^{-1}\mathbf{K}_*^\top), \tag{4}$$

where $\mathbf{K}_{*,*}$ denotes the $(n^* \times n^*)$ kernel matrix between test points, and $\mathbf{K}_*$ the $(n^* \times n)$ kernel matrix between test and training points. We defer derivations to Rasmussen and Williams (2006) or other sources. This is a posterior distribution for Y^* given the data (and the choice of kernel). Further, since the normal distribution has identical mean and mode, the mean argument in posterior $(\mathbf{K}_*(\mathbf{K}+\sigma^2 I_n)^{-1} Y)$ is both the CEF and the *maximum a posteriori* (MAP) estimator.

The consequences of this expression are illustrated in Figure 2. For values of X closer to observations in the data, the covariance of Y^* with the corresponding observed Y will be higher, providing more information to update our guess of Y^* and reduce our uncertainty, whereas uncertainty balloons for predictions farther from observed data. This entire posterior distribution for any point is available in closed form, making Markov chain sampling unnecessary.

Note also that the variance in Expression (4), $\mathbf{K}_{*,*} + \sigma^2 I_{n^*} - \mathbf{K}_*(\mathbf{K}+\sigma^2 I_n)^{-1}\mathbf{K}_*^\top$, describes our (un)certainty about points Y^*. In contrast, when we are interested in uncertainty about the CEF itself, the relevant variance is $\mathbf{K}_{*,*} - \mathbf{K}_*(\mathbf{K}+\sigma^2 I_n)^{-1}\mathbf{K}_*^\top$, which excludes the irreducible noise term $(\sigma^2 I_{n^*})$ that reflects how individual Y values scatter around their conditional mean.

Homoskedasticity: A key assumption employed by this model choice is that Y is distributed normally around its CEF/MAP, with constant variance. Violating this assumption can influence interval size in particular, as we illustrate in Section 3.3 and Section A.5.1 of the Supplementary Material. Note however that this does not imply constant variance in the posterior uncertainty, which will widen or narrow adaptively, reflecting local data sparsity.

Interpreting the Function Space and Comparison to Kernel Regularized Least Squares (KRLS): For additional context, in Section A.1 of the Supplementary Material, we remark on the relationship between

https://doi.org/10.1017/pan.2026.10032

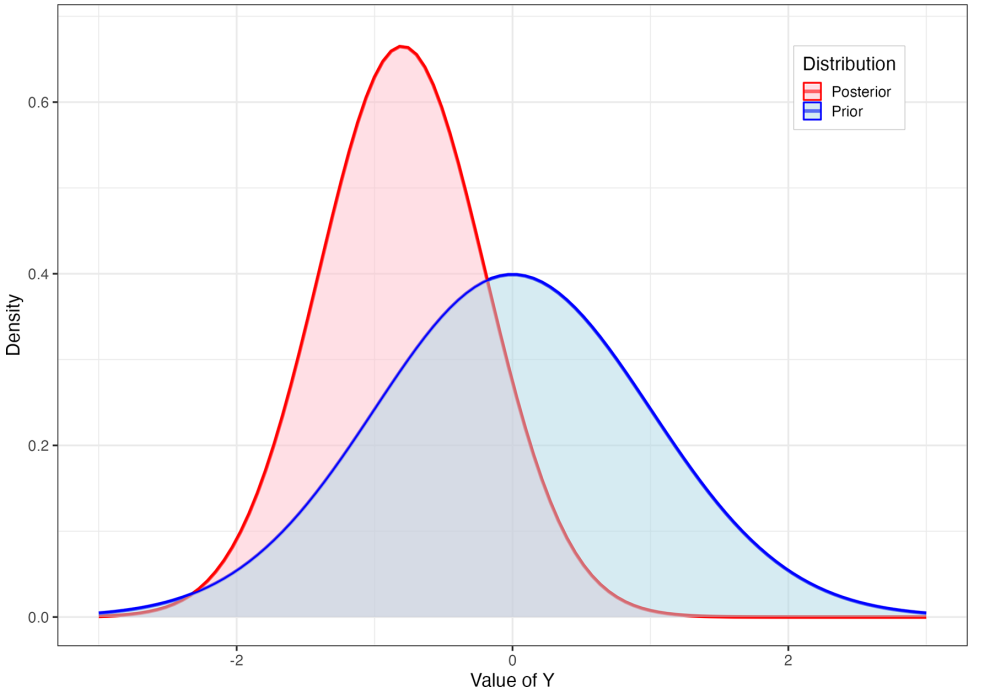


Figure 1. *Blue* (prior): Distributional belief for an unseen Y^* knowing only that it will come from a normal distribution with mean 0 and variance σ^2. *Red* (posterior): Our revised belief regarding Y^* assuming $cor(Y^*, Y_{obs}) = 0.8$, and having observed $Y_{obs} = -1$.

GP and KRLS, which have identical function spaces for the CEF, but handle uncertainty very differently (discussed in Section 2.2). Both also involve regularization, but motivated and tuned differently.

2.1. Additional Details

We now consider a set of interrelated scaling and hyperparameter selection choices, including the kernel choice. It is not clear whether an optimal solution to these questions exists in a general sense, yet users must make these choices, and some choices can lead to poor performance. We offer guidance that we believe is at least reasoned, practical, and shows good performance to the limits of our testing.

Demeaning and Rescaling: We demean and scale the covariates to have unit variance. This is common for kernel-based machine learning methods, and avoids dependency on unit-of-measure decisions. As noted, we also (globally) demean and scale Y to have unit variance, as in some other kernelized approaches (see, e.g., Hainmueller and Hazlett 2014). This is also required for the interpretation of σ^2 as 1-R^2 noted below.

Bandwidth Selection: For the Gaussian kernel $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$, we must choose the bandwidth or "length-scale," b. There are numerous strategies for doing so. One reasonable approach is to set b equal to (or proportional to) the number of dimensions of X, so that the result does not explode or go to zero as the number of dimensions changes (as in Hainmueller and Hazlett 2014). Another recent proposal, offered in Hartman, Hazlett, and Sterbenz (2025), chooses the value of b that leads to the highest estimated variance across the (off-diagonal) elements of $\mathbf{K}$. This simply ensures that the columns of $\mathbf{K}$ stand to be highly informative rather than having $\mathbf{K}$ approach the identity matrix (if b is too small) or a block of ones (if b is too large). This is the approach we take.

Choosing σ^2: Having fixed $\sigma_f = 1$ and selected the kernel bandwidth based directly on the distribution of $\mathbf{K}$, we are left with only one hyperparameter to tune. As advocated by Rasmussen and Williams

https://doi.org/10.1017/pan.2026.10032

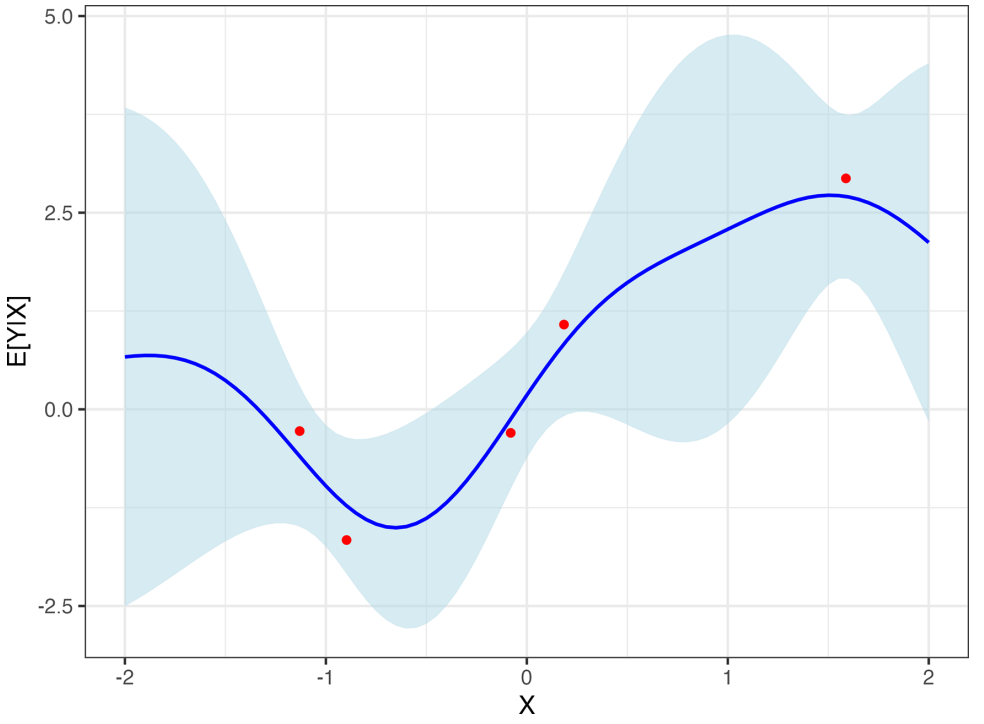


Figure 2. Posterior CEF with its 95% CI. Inferred distribution (spread vertically) about Y^* as a function of covariate X, after seeing five observations (red dots). The key assumption is on the covariance between points as a function of their X values, here given by $cov(Y^*, Y_i) = exp(-||X_i - X^*||^2/b)$.

(2006), we tune the residual variance term, σ^2, by maximizing the log marginal likelihood,

$$\log\, p(Y \mid X) = -\frac{1}{2} Y^\top (\mathbf{K} + \sigma^2 I)^{-1} Y - \frac{1}{2}(\log|\mathbf{K} + \sigma^2 I|) - \frac{n}{2}\log 2\pi. \tag{5}$$

In principle, since one could always choose a "wigglier CEF" (even for a fixed kernel bandwidth) together with a smaller σ^2, this quantity might not always be well-identified on the data. While we remain concerned about this in the general case, empirically, this approach appears to perform very well at least over the cases we examine. In particular, we see that for each DGP we attempt, the estimated σ^2 values are very close to 1-R^2, as our scaling choice should lead to. This may fail when the DGP does not have the Gaussian, homoskedastic noise on which this likelihood estimate is based.

Combining Kernels: A useful fact about kernels used for such kernel regression tasks is that they can be added together, with the resulting kernel remaining valid (positive semi-definite), and providing access to a space of functions that combines the underlying features of the constituent kernels (see, e.g., Schulz *et al.* 2017). For example, consider a kernel $\mathbf{K}^{poly}$ that provides access to polynomial models, i.e., $\mathbf{K}^{poly}_{i,j} = 1 + \langle x_i, x_j \rangle^p$, where p is the desired degree of the polynomial. Such a kernel matrix can be added pointwise to the Gaussian kernel matrix. The resulting kernel matrix allows the GP to represent functions that are both locally smooth (from the Gaussian kernel), and exhibit polynomial growth rather than translation invariance (from the polynomial). Similarly, including a periodic kernel enables the model to capture cyclic structures such as seasonality. We demonstrate this below to explore models for ITS analysis.

Stationary vs. Non-Stationary Kernels for Mild vs. Extreme Extrapolation. The next question is whether one is interested in making inferences nearer to observed data (e.g., interpolation, edge estimation, or very mild extrapolation), where we can rely on the covariance between observed points

https://doi.org/10.1017/pan.2026.10032

and target points directly, or if one wishes to make inferences well beyond the data's edge. In the former case, we can employ a "stationary kernel" that operates only on the relative distances between observations. In these settings, we rely on the Gaussian kernel because (i) it directly works with the logic that observations with more similar values should have greater covariance while covariance should drop to zero as observations grow far apart in the covariate space; (ii) it is commonly taken to be the workhorse kernel for many machine learning approaches involving kernels; and (iii) it is an example of a "universal kernel" for the continuous functions, meaning that it can represent any continuous function given sufficient data (Micchelli, Xu, and Zhang 2006).

By contrast, for some applications, inferences are attempted farther from the edge of the data. Our ITS application below is a leading example, though our poor-overlap case also requires moderate to severe extrapolation. Under stationary kernels like the Gaussian, the predicted CEF will "return to the mean" far from the edge of the data. While this may be reasonable, the uncertainty can reach a maximum (determined by $\hat{\sigma}^2$) and stop growing. Without addressing this issue, it would appear that the GP does not extrapolate well (see, e.g., Wilson and Adams 2013). However, when investigators need to extrapolate farther (relative to the scale of the kernel bandwidth), they may need to entertain functions that show (i) periodic behavior (in the case of time series data) and/or (ii) continued growth, e.g., with some continuity in the first or higher derivative, as in linear or polynomial prediction. The GP can accommodate this by combining kernels, for example, adding a Gaussian kernel with a polynomial kernel of the desired order. In these settings, the GP becomes a device for illustrating the wide range of possible functions (and our uncertainty over them). We illustrate this use in Section 3.2 below.

Prior Mean Functions: As noted above, we set $\mu = 0$. This is because prior to seeing any data, we would have no particular reason to prefer a given functional form, direction of change, or any other feature of the CEF not already encoded by choosing the multivariate normal distribution with covariance given by the kernel function. After conditioning, the posterior CEF can approximate whatever smooth functional form the data support.[1] This choice does not restrict the posterior fit, which adapts around the observed data as in Expression (4) (see, e.g., Figure 2). Regularization nevertheless occurs through σ^2, which governs the trade-off between fit and smoothness and plays the same role as the regularization parameter in kernel regression (e.g., λ in KRLS; Hainmueller and Hazlett 2014).

Taking RD as an example, the GP can represent CEFs that are approximately linear, polynomial, or otherwise smooth on each side of the cutoff. If extrapolation far beyond the observed data is needed, prior expectations about functional form can be encoded through a (non-stationary) kernel rather than a mean function. This distinguishes our strategy from approaches in Branson *et al.* (2019), Rischard *et al.* (2021), and Ornstein and Duck-Mayr (2022).[2]

2.2. Comparison to Conventional Uncertainty Estimation under Extrapolation

For tasks involving extrapolation, a key advantage of the GP is its ability to incorporate what we refer to as "extrapolation uncertainty" in its reported posterior uncertainty intervals. Conventional approaches to uncertainty estimation in parametric models involve selecting a single best-fitting model and using that fitted model alone for extrapolation (as well as to compute residuals that inform uncertainty intervals). This is problematic because, within a sufficiently flexible model space, many model fits can explain the observed data similarly well but will make radically different predictions outside of the observed data. Our uncertainty over which of those fits to choose must be maintained if we wish to sustain appropriate uncertainty over the predictions they make when extrapolated. This "extrapolation

[1] These CEFs are linear in the kernel matrix, corresponding to the empirical subspace of the reproducing kernel Hilbert space (RKHS) associated with the kernel.

[2] We expect that adding a non-zero prior mean function will have little impact on estimation, particularly when the mean function lies within the function space implied by the kernel and inference is restricted to regions near the observed data. Branson *et al.* (2019) likewise report that using a zero mean instead of a linear mean function produced "largely the same" results in their simulations.

https://doi.org/10.1017/pan.2026.10032

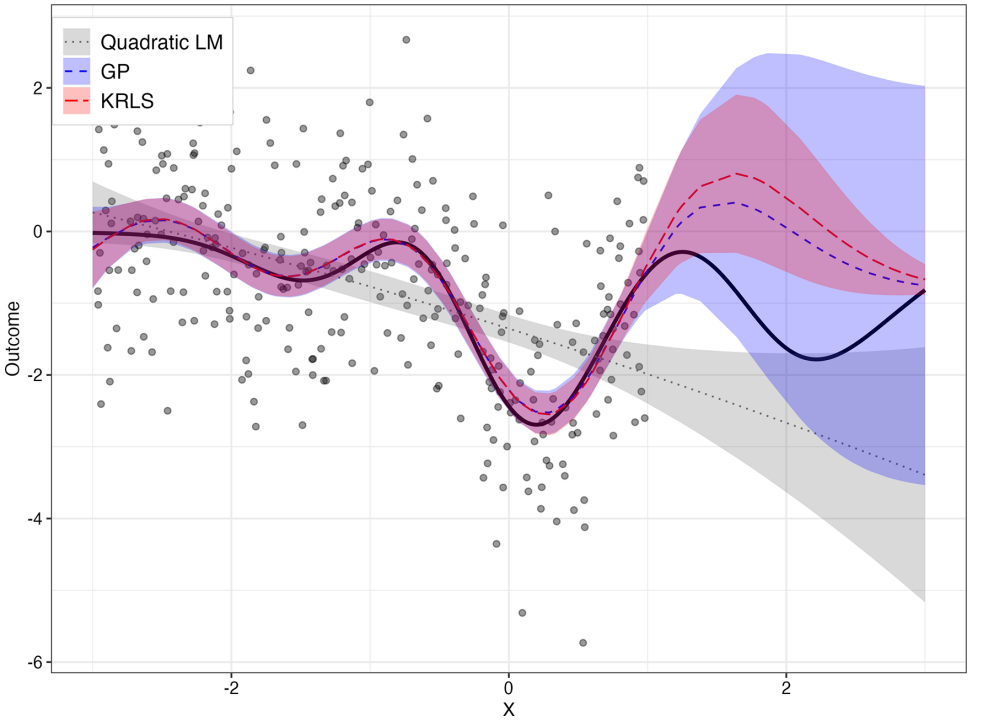


Figure 3. Comparison of uncertainty for conditional expectation functions given by GP (blue), KRLS (red), and a quadratic polynomial (gray). Models are fitted on X data between −3 and 1, then extrapolated over X between 1 and 3. The solid line shows the (randomly drawn) true function, while dashed/dotted lines indicate each model's estimated CEF. Bands indicate 95% CIs for the CEFs.

uncertainty," however, is lost once we choose a single fitted model, discarding all others. Contrast this with the GP. While the GP does commit to a specific model space and set of assumptions described above, the process of conditioning on the data and inspecting the posterior is different from that of choosing and extrapolating only a single best-fitting model. The GP's posterior uncertainty band instead reflects the range of CEFs that remain plausible, to varying degrees, given the observed data. The posterior uncertainty at new points effectively extends these functions beyond the data, encompassing their divergent predictions and thereby incorporating increased uncertainty as we move farther from the conditioning data.

We illustrate this in Figure 3. First, consider a linear model (LM) fitted by OLS. The classical variance estimate for the coefficient under homoskedastic and independent (spherical) errors will be $\widehat{\mathbb{V}}(\hat{\beta} \mid \mathbf{X}, Y) = (\mathbf{X}^\top \mathbf{X})^{-1} \hat{\sigma}^2 I$. The value of $\hat{\sigma}^2$ is proportional to $\sum_i (Y_i - \mathbf{X}_i^\top \hat{\beta})^2$, that is, the sum of the squared *fitted* residuals.[3] Consequently, the estimated variance of the predicted value of the CEF at some point $\mathbf{X}_j$ will be $\widehat{\mathbb{V}}(\hat{Y}(\mathbf{X}_j)) = \hat{\mathbb{V}}(\mathbf{X}_j^\top \hat{\beta}) = \mathbf{X}_j^\top \hat{\mathbb{V}}(\hat{\beta}) \mathbf{X}_j$. While this quantity increases for predictions at $\mathbf{X}_j$ points farther from the mean of the data, this is blind to a point's distance from supporting data. The `quadratic LM` line and uncertainty envelope in Figure 3 illustrate this in the case of a fitted quadratic model.

This problem is not one of relying on a restrictive model space. We can demonstrate this by comparing GP and KRLS, employing the same kernel (Figure 3). Since these models have identical model spaces for the CEF, their CEFs match very closely, differing only due to hyperparameter tuning.

[3] Alternative variance estimators, such as heteroskedasticity robust, cluster-robust, etc., will have a more complex form rooted in a different choice of variance–covariance matrix for the residuals, Σ. However, the problem we illustrate here is not due to misspecification of the Σ and persists even when Σ is correctly chosen.

https://doi.org/10.1017/pan.2026.10032

The concern, however, is the uncertainty intervals. KRLS employs the conventional fitted-function variance estimation. This results in a collapse at points farther from the observed data, because the model is structured such that for any one model, we can be sure the CEF it predicts returns to the mean of Y as we leave the support of the data. However, this is the opposite of what we would want, which is to see increased uncertainty about the value of Y as we move further from the support of the data. As Figure 3 shows, the predictive uncertainty from the GP model provides this, widening as we move toward and then beyond the edge of the data.

3. Use Cases for GP

In this section, we illustrate the useful properties of GPs in three settings where inferences depend on understanding uncertainty that arises from model dependency as we move toward or beyond the edge of the observed data.[4]

3.1. Comparing Groups with Poor Overlap

In cross-sectional comparisons, treated and control groups must often be compared as if they had the same covariate distributions. For example, we have an observed confounder X, and assuming the absence of unobserved confounders (i.e., under selection on observables or conditional ignorability), we wish to make comparisons between treated and control units adjusting for X.

Adjustment strategies of all kinds are vulnerable to extrapolation bias when there are regions of X containing only treated or only control units. Here, treated units have values of X between −3 and 1, whereas control units have values between −2 and 3. Thus, a model for the treatment outcome must be extrapolated in areas with few or no treated units ($X > 1$) and a model for the control outcome must be extrapolated in areas with few or no control units ($X < -2$). Treatment effect estimates—computed by comparing these two models for each unit (i.e., g-computation, regression imputation, etc.)—consequently suffer in the regions of poor overlap insofar as these model fits fail to predict well in those regions.

For each simulation iteration, we draw a random function to serve as the CEF of Y given X, with a constant treatment effect of 3. To favor non-GP methods, we draw the CEF as a quadratic function $f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ with the coefficients independently drawn from a standard normal distribution. Following this, "observed" samples are drawn from each of these CEFs ($N = 500$) with the addition of noise, distributed normally and with variance calibrated to ensure an overall true R^2 of 0.5.[5]

Figure 4 presents a single example of a random function drawn from this space (dotted lines), along with the simulated data under treatment (red) and control (blue). The three models tested fit the data similarly well within the region of common support ($-2 < X < 1$). In the regions with poor overlap ($X < -2$ and $X > 1$), however, the result is highly sensitive to the choice of modeling approach. Notably, the uncertainty estimates for the LM and BART show insufficient uncertainty to accommodate this model dependency under poor overlap.

The bottom row of Figure 4 shows what these estimated CEFs and uncertainty intervals imply for conditional average treatment effects (CATEs) and the average treatment effect (ATE). Since the true CATE is constant, the apparently linear change in the treatment effect as a function of X for the LM (left) is erroneous. This is a byproduct of the poor overlap, which led to fitted models with very different slopes for treated versus control groups. While the problem is less pronounced for BART, its uncertainty estimates still fail to reflect the distance from relevant observed data. Its uncertainty over the CATE remains nearly constant across areas of good and poor overlap, producing widely insufficient coverage in

[4] Replication code and data are available at Cho, Kim, and Hazlett (2025).

[5] Throughout our analyses, for realism, we calibrate R^2 values in simulations to 0.3–0.5, except where we vary them more widely to demonstrate how performance depends on it. In this simulation, we set this R^2 to 0.5, though in Figure 4 $R^2 = 0.8$ for improved visualization.

https://doi.org/10.1017/pan.2026.10032

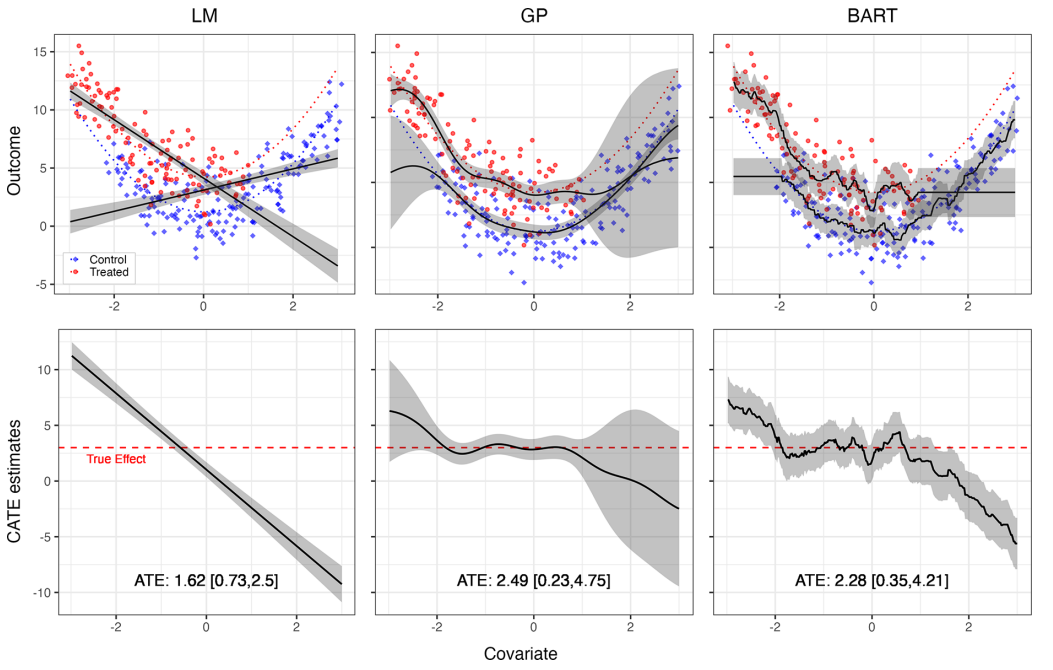


Figure 4. Uncertainty quantification of CATE under varying overlap. *Top row*: Illustration from one draw of the simulation setting, with a good overlap region ($-2 < X < 1$) and no overlap elsewhere. For treated units (red) and control units (blue), the true CEF (dotted lines) and estimated CEF (black solid lines) with corresponding uncertainty bands are drawn. As shown at both ends of the covariate value, GP allows for growing uncertainty bands adaptive to the degree of overlap. The upper bound of the confidence interval in extrapolation depends on the variability in the fitted data. *Bottom row*: The CATE estimates with 95% confidence bands. The wider bands of GP in poor-overlap regions propagate to higher uncertainty in CATE estimation. The red dashed line represents the true effect size.

the areas of weaker overlap. In contrast, GP's uncertainty estimates adaptively reflect model dependence due to data sparsity—narrower where common support is better, wider where it is weaker.

We repeat this process across 500 iterations, drawing different random functions each time with sample sizes of 500 as described above. Figure 5 illustrates the behaviors of each method across these iterations. The coverage rate for the CATE at each possible value of X (top left) is problematic for the LM at all values of X. BART generally overcovers in areas of common support but radically undercovers in non-overlap regions, dropping below 50% on one side and 25% on the other. The GP achieves approximately nominal coverage where common support is good. In non-overlap areas, it sustains good coverage initially on one side, but still transitions to undercoverage, dropping as low as roughly 75%. While this undercoverage is much less severe than the others, it remains imperfect. This imperfection serves as a warning of the limitations of the GP approach. Specifically, the example is an adversarial one with common support failing for half of the range of X, while the underlying DGPs are quadratic functions that can diverge radically beyond the edges of their training data. In practice, we would recommend GP with a non-stationary kernel when this degree of extrapolation is required (see Section 3.2).

The interval lengths required to achieve this (top right) are also of interest. The drastic undercoverage of LM results from inappropriately short intervals. BART maintains almost constant interval lengths throughout, which are excessively long for the region of good common support. GP exhibits adaptive behavior, maintaining narrow intervals in regions of good overlap and wider intervals in areas of poor overlap.

Finally, while conditional uncertainty and coverage are of potential interest for many queries and demonstrate the behavior of interest here, investigators may be interested primarily in average effects.

https://doi.org/10.1017/pan.2026.10032

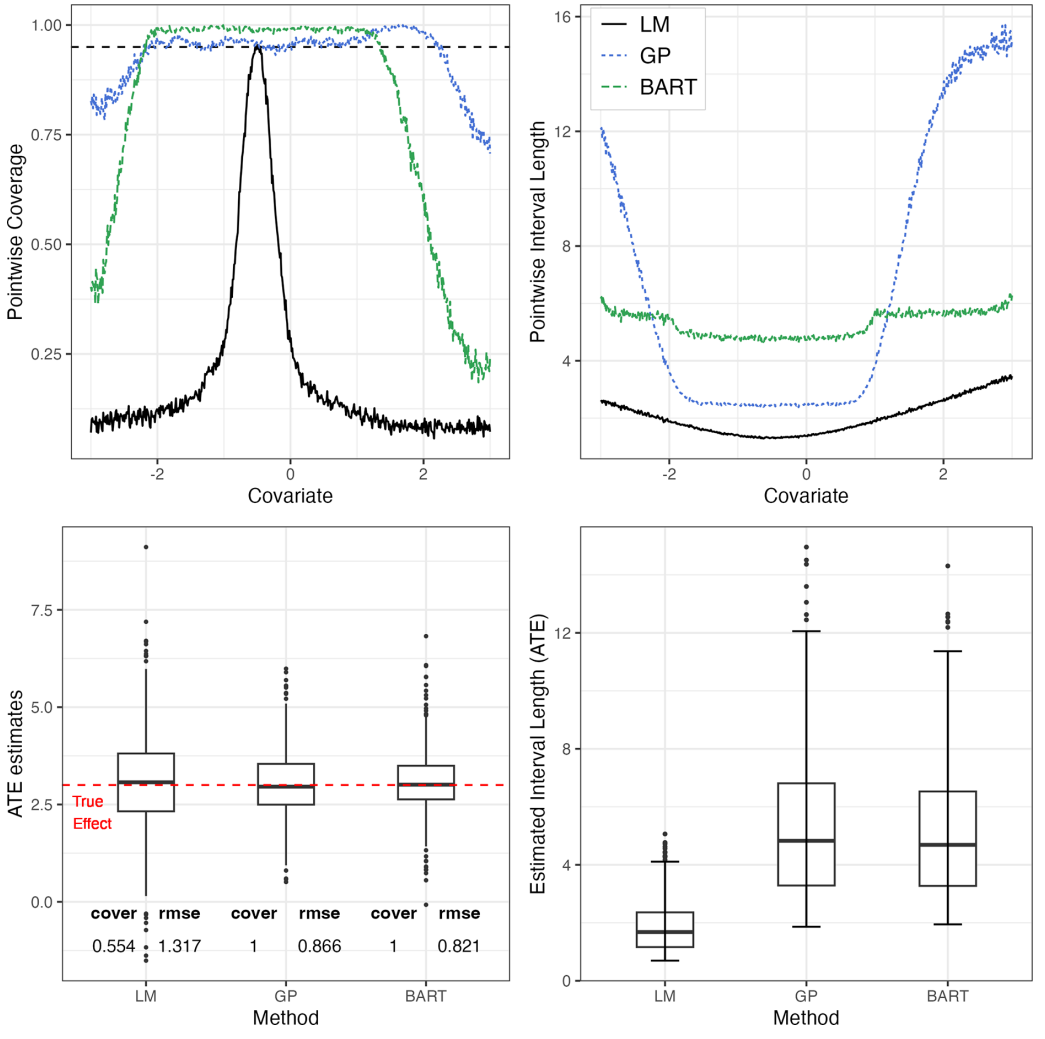


Figure 5. Pointwise (top) and average (bottom) performance of GP model compared to LM and BART. *Top row*: The graph on the left displays the pointwise coverage rate of the ATE estimators by the three models. On the right, the lengths of corresponding 95% confidence intervals are shown over 500 simulations ($n = 500$). *Bottom row*: Boxplots represent the distributions of the ATE estimates (*left*) and average interval lengths (*right*). The true treatment effect is denoted by the red dashed line. The coverage rates and RMSEs (of ATE) for each method are also shown.

Even here, however, the poor behavior of LM under weak overlap leads to an RMSE 1.52 times larger than that of GP, together with severe undercoverage (55%).

In Section A.2 of the Supplementary Material, we expand these findings with additional simulations across various function spaces and signal-to-noise ratios (R^2).

3.1.1. *Illustration with a Multivariate Benchmark under Varying Overlap*

In Section A.3 of the Supplementary Material, we use the well-known National Supported Work Demonstration (NSW) example (Dehejia and Wahba 1999; LaLonde 1986) to demonstrate how to apply GP for covariate-adjusted comparison with multiple covariates, with varying overlap quality. GP produces a point estimate very close to the experimental benchmark. Further, the counterfactual uncertainty estimates from GP demonstrate their expected relationship to local data density.

https://doi.org/10.1017/pan.2026.10032

3.2. Interrupted Time Series with GP

3.2.1. Background

The ITS design is used to assess the effect of events or shocks experienced universally after a specific time point (Bernal, Cummins, and Gasparrini 2017; Box and Jenkins 1976; Box and Tiao 1975. We observe a time series of non-treatment outcomes, $Y_t(0)$ prior to the event of interest. The event occurs at time $t = T$ and all Y_t outcomes measured after that time are treatment outcomes, $Y_t(1)$. We train a model on pre-treatment data then must extrapolate it to the post-treatment era. These extrapolated predictions of $Y_t(0)$ are then compared to actual post-treatment outcomes ($Y_t(1)$) at any $t \geq T$.

The GP approach is potentially valuable here for several reasons. First, it handles autocorrelation in the outcomes naturally. Second, the covariance function (kernel) can be designed to accommodate not only the "smoothness" we expect but also secular trends that will continue on over time, and periodic trends such as seasonality. Third and most relevant to our discussion, its estimates of uncertainty take into account the distance (in time) to the observed data and the consequent model dependency. We note that regardless of estimation approach, making causal claims from ITS requires additional demanding identification assumptions (see Section A.4 of the Supplementary Material).

3.2.2. Illustrative Application

In June 2008, the Supreme Court ruling in *District of Columbia v. Heller*, 554 U.S. 570 (2008), affirmed the individual's right to keep and bear arms for self-defense and other purposes, striking down DC's prior ban. We examine how the number of handgun background checks per 100,000 population (as a proxy for handgun sales) responded to this decision.[6] Figure 6 shows results for DC and Vermont, with seven years of pre-treatment and one year of post-treatment data. In DC, the increase in background checks immediately after *Heller* is substantial compared to the expectations suggested by the GP model, even given additional uncertainty due to extrapolation. Results are less clear in Vermont, with many of the post-treatment observations falling within the interval expected post-treatment (Figure 6, *bottom*).

Further, the choice of kernel encodes the set of extrapolating functions the investigator deems plausible. Figure 7 demonstrates the impact that these choices can have. The left panel shows results with only the Gaussian kernel. As a stationary kernel, this is not suitable for conservative extrapolation as noted above.

We then consider two additions to the kernel. First, we add a periodic kernel (with a period of one year) to capture possible annual cycles. Second, we consider polynomial functions that describe how the CEF is allowed to extend beyond the edge of the data. The *center* panel of Figure 7 employs a Gaussian + periodic + linear kernel. A sizable fraction of post-treatment observations still fall above the predicted counterfactual level. Adding a quadratic growth component rather than linear (*right* panel), however, the observed post-treatment outcomes fall almost entirely within the predicted envelope, implying no estimated effect.

We do not expect the user will typically know which choice best describes how $Y(0)$ would evolve once it becomes unobservable. Further, the observed pre-treatment data are not well-equipped to choose among these options without additional strong assumptions. The benefit of GP is not that it can somehow "know" the right extrapolation model, but that it provides a tool to consider capacious spaces of functions while fully characterizing uncertainty over predictions given this. The price of our inability to rule out more extreme extrapolating models (e.g., higher-order polynomials) is seen in greater uncertainty over the post-treatment counterfactual, and weakened ability to make inferences. The GP makes this tradeoff explicit and clear. This approach can also be used in the inverse, revealing what assumption (on the extrapolation function) a user would have to defend in order to support a given conclusion. Here, for instance, to claim there is a putative effect requires arguing that the extrapolating function is linear, but not quadratic.

[6] Thanks to Jack Kappelman for these data.

https://doi.org/10.1017/pan.2026.10032

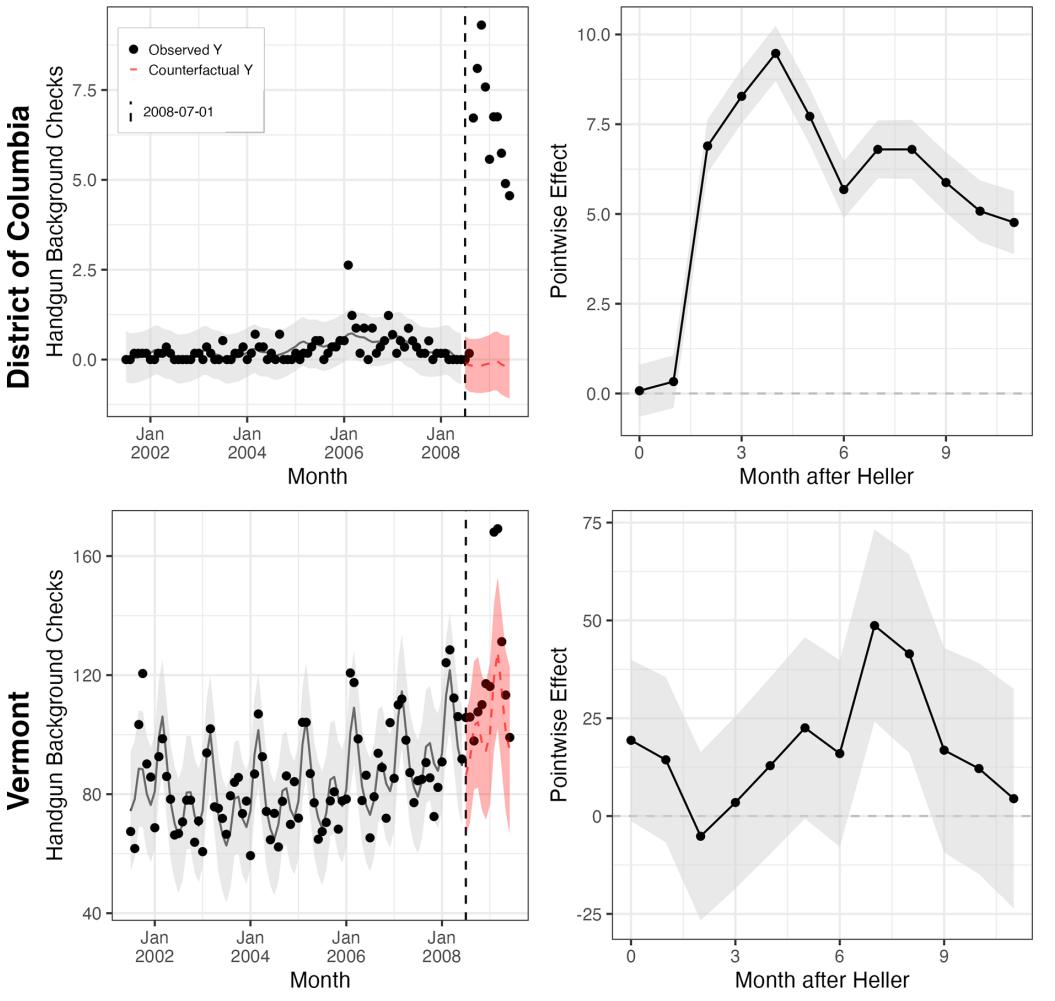


Figure 6. *Left column:* Monthly rate of handgun background checks per 100,000 population in D.C. (top) and Vermont (bottom), seven years before and one year after the *Heller* ruling. The red envelope shows the 95% predictive interval for the GP-estimated non-treatment outcome, post-treatment. *Right column:* Point estimates and confidence band for the differences between the observed outcome (treated) and the predicted (non-treatment) counterfactual at each month. The kernel choice combines (adds) the linear, periodic, and Gaussian kernels.

3.3. Regression Discontinuity

The RD design is a widely used tool in the social sciences for cases when decisions about treatment turn on a sharp cutoff in some variable. While this approach rests on relatively credible identification assumptions, it poses estimation challenges as it relies on point estimates with uncertainty intervals precisely where the underlying model(s) are at the very edge of their support.

When fitting models below/above the cutoff in RD, decisions must be made regarding the bias-variance tradeoff: how much should we seek to reduce variance by using observations farther from the cutoff, given that including such data can bias our estimate? Even when optimizing this choice by some criterion, the consequent bias must be dealt with when estimating uncertainty. Stommes, Aronow, and Sävje (2023) review varied approaches to dealing with this bias and its implications for inference. Among such approaches, Calonico *et al.* (2014) provide a now standard optimal bandwidth selection technique, bias-adjustment of the point estimate, and bias-adjusted inference, which we refer to by the

https://doi.org/10.1017/pan.2026.10032

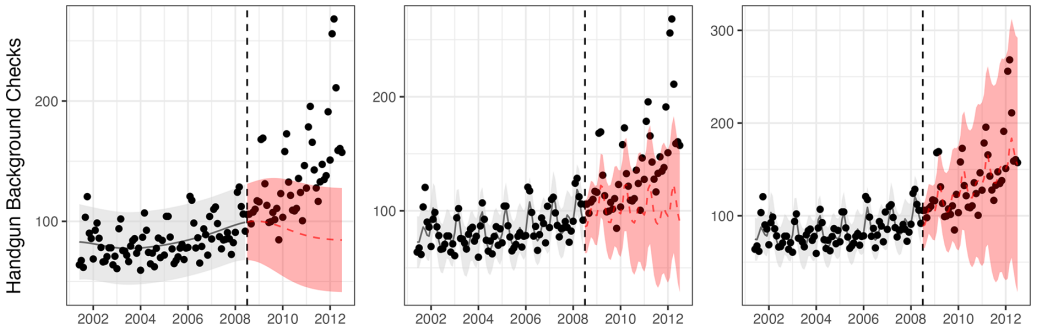


Figure 7. GP results with three different kernels for Vermont. The post-treatment period is extended to four years to better illustrate the behavior under extreme extrapolation. *Left:* Gaussian kernel; *Center:* "Gaussian + periodic + linear" kernel; *Right:* "Gaussian + periodic + quadratic" kernel.

name of the widely used software implementation, `rdrobust` (Calonico, Cattaneo, and Titiunik 2015; Calonico *et al.* 2017).

The GP approach is attractive for RD, first, because it offers a flexible function space that makes only weak assumptions about smoothness (when using the Gaussian kernel, for example). Second, and more importantly, it provides a principled approach to uncertainty quantification for edge prediction or mild extrapolation. In cases with considerable data near the threshold, an optimally chosen parametric or semi-parametric approach, like those employed by `rdrobust`, can establish reliable point estimates and uncertainty intervals at the edge of the data, as demonstrated by Calonico *et al.* (2014). However, with smaller samples or sparse data near the cutoff, multiple models may fit the observed data similarly well while yielding different predictions at the boundary. The GP approach provides uncertainty estimates that retain this uncertainty over the CEF, allowing inference without relying on large-sample asymptotics.

3.3.1. RD Case 1: Total Random Simulation

We first consider the simple "total random case," in which the running variable (X) and the outcome (Y) are both drawn independently at random, with zero treatment effect.[7] For each of 500 simulations, we draw a sample of size 500, with $X_i \sim \mathcal{N}(0,1)$, and $Y_i \sim \mathcal{N}(0,\sigma^2)$. The value of σ is varied from near zero up to 3, reflecting various possible scales of the outcome variable relative to the forcing variable. We use a cutoff of 0 for X to determine treatment eligibility status. We employ `rdrobust` with software defaults and utilize the bias-adjusted point estimates with the "robust" confidence intervals. To compute the GP estimate (`GPrd`), we first use the GP function from `gpss` to estimate a model for Y given X for the treated units ($X > 0$), and another for the control units ($X \leq 0$). From this, we obtain the point estimate and standard error for the prediction of the expected Y at precisely $X = 0$ from each model. The treatment effect is estimated as the difference between these point estimates, with standard error given by the square root of the sum of the variances of the two predicted outcomes.

Figure 8 shows simulation results in this setting. Looking first at the top row, we see that `GPrd` has a higher coverage rate across all levels of σ, at 97%–98%, while `rdrobust` has a coverage rate between

[7]Here, we examine RD designs with a single running variable and no covariates. As shown by Calonico *et al.* (2019), including covariates may be desirable in some cases, if the intention is to improve precision. To include covariates in GP, they need only be included as additional columns in the data alongside the running variable. The kernel matrix entries are computed over the running variable and covariates, just as it is in the multivariate examples above. The RD estimator itself is still taken with respect to the running variable: for each individual i, the predicted Y_i is obtained from the model below the cutoff and from the model above the cutoff, setting X to the cutoff value but leaving the covariates at observed values. Consequently, the models account for any differences between groups on covariates above and below the cutoff, through their impact on the predictions of Y.

https://doi.org/10.1017/pan.2026.10032

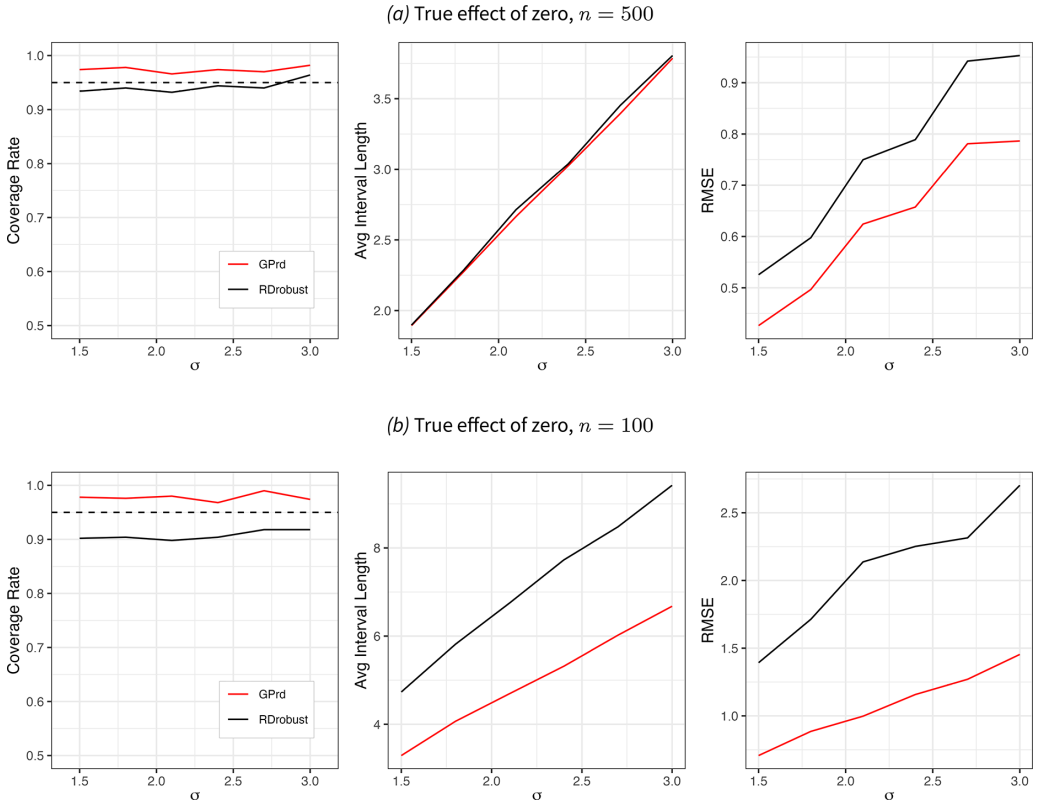


Figure 8. The coverage rate, average length of 95% confidence interval, and RMSE of `GPrd` (red) and `rdrobust` with "robust" options (black) in the total random setting. The ratio of the variance of *Y* to *X* is given by σ on the horizontal axis. *Top row:* True effect size is zero, 500 observations. *Bottom row:* Effect size is zero, sample size is reduced to 100. Results with a true effect size of three look identical.

93% and 96%. The two approaches have essentially identical interval sizes, which appropriately scale linearly with σ (middle column). `GPrd` also shows much smaller RMSE across the 500 simulations. This RMSE benefit is amplified in smaller samples (bottom row). We examine and offer an explanation for this behavior in Section 3.3.3.

3.3.2. RD Case 2: Latent Variable Confounding Simulation

The total random case is a useful starting point, but may not produce sufficient risk of bias because there is no relationship between X and $Y(d)$ on either side of the cutoff such that incorporating information farther from the cutoff risks adding bias to the estimate. To better simulate this concern, we consider a latent confounding formulation. In each of 500 replications, we generate a random sample of size 200 according to the following:

1. Draw a latent variable: $\mu \sim \mathcal{U}(-0.25, 0.25)$.
2. Create our running variable, X, loosely resembling vote share in an electoral RD as a sigmoidal function of μ, with $x_i = 1/(1 + exp(-s * \mu_i))$. The parameter s represents a varying degree of steepness of this sigmoid, such that larger values produce a steeper sigmoid, increasing the risk of bias due to misspecification (see Figure 9).
3. The treatment D represents "winning the election," coded as 1 if the vote share is greater than 0.5, and 0 otherwise.

https://doi.org/10.1017/pan.2026.10032

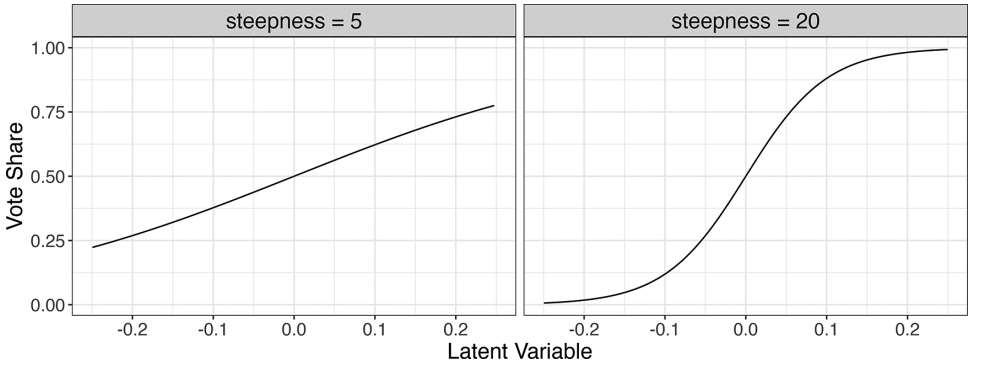


Figure 9. Relationship between simulated latent variable μ (horizontal axis) and vote share X (vertical axis) at two steepness (s) parameters used in the simulation.

4. Generate the outcome $Y_i = 1.5\mu_i + \tau D + \varepsilon$, where $\tau = 3$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The σ will be varied by simulation setting to explore a range of realistic signal-to-noise ratios (R^2).

Next, purely data-driven RD specification risks using distant observations to estimate $\mathbb{E}[Y|X]$ at the cutoff, compromising the motivating logic for using RD. To address this concern, we consider options that set the bandwidth or trim the sample (or both) based on outside causal assumptions:

- `gp_causal`: The investigator proposes a kernel bandwidth based on what they are willing to believe about how close units may covary in their outcomes. For example, it might be reasonable to propose that two observations whose vote shares are 2% apart may be highly correlated, e.g., at 0.9. This implies a surprisingly small value for b at 0.005.[8]
- `gp_causaltrim`: After choosing the `causaltrim` bandwidth, regions that are effectively irrelevant to estimation at the cutoff can be removed. This aids in transparency, but is also helpful because the value of σ^2 is a single global parameter, so trimming the data first ensures it is determined over only the range of the relevant range of the sample. In our simulation, we remove observations with vote shares below 0.4 or above 0.6, on the premise that users would likely argue that data outside this range are of no value to modeling what happens at the cutoff.

Figure 10 summarizes the results over 500 iterations at each choice for the steepness (s) of the latent relationship between μ and Y and at four different levels of R^2. All the approaches show similar interval sizes, with the exception of `rdrobust`, which yields wider intervals when steepness equals 20. While `rdrobust` tends to slightly undercover the true effect, the GP-based approaches exhibit slight overcoverage, except for `gp_causaltrim` at a steepness of 20.

3.3.3. RD Case 3: Empirical Application with Benchmarking by Placebo Cutoffs

Finally, we study the performance of `GPrd` on real data, in a case where the "true" effect is knowable for benchmarking purposes. To do so, we consider the data from Lee, Moretti, and Butler (2004), in which the forcing variable is Democrats' (two-party) vote share in U.S. House Elections (1948–1990), and the outcome of interest is a measure of how liberal each representative's vote record is assessed to be, called the ADA score. While we cannot know the true effect size, it would be arguably zero when using any value other than 0.5 as a (placebo) cutoff.[9]

[8] $k(X_i, X_j) = exp\left(-\frac{(X_i - X_j)^2}{b}\right) \rightarrow 0.9 \approx exp\left(-\frac{0.02^2}{0.005}\right)$. Correlations equal covariances here due to the rescaling of the outcome. The running variable is not rescaled, so that the 0.02 in the numerator remains correct.

[9] In each placebo analysis, data are included only from below 0.5 or above it, but not from both, to avoid potential contamination due to a real treatment effect.

https://doi.org/10.1017/pan.2026.10032

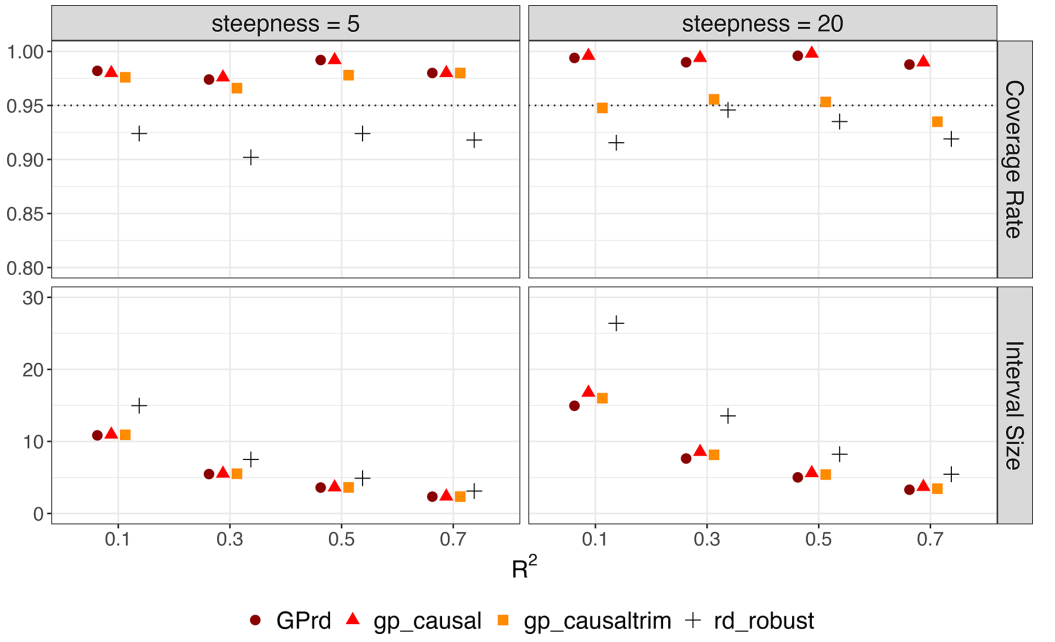


Figure 10. RD simulation results with latent variable (effect size = 3). Each point represents the coverage rate, or average length of 95% confidence interval of `GPrd`, `gp_causal`, `gp_causaltrim`, and `rdrobust` across 500 iterations of the latent variable simulations ($n = 200$). The horizontal axis represents the R^2.

We consider three estimators: `GPrd`, `gp_causaltrim`, and `rdrobust`. Figure 11 shows results with the full data. The choice of estimator matters little to the point estimates at the true cutoff (0.5), though `GPrd` and `gp_causaltrim` are more conservative in their uncertainty estimates. Looking at the eight placebo cutoffs, the differences are relatively small and while `rdrobust` suggests a statistically significant estimate at the cutoff of 0.35, `GPrd` and `gp_causaltrim` very nearly do as well. At the cutoff of 0.65, the `rdrobust` interval again excludes zero, while the intervals for the GP approaches do not, but the actual difference is again relatively small.

However, the Lee *et al.* (2004) data initially contain 13,577 observations without missingness on the running variable and outcome. While each of the placebo cutoff analyses uses only a subset of these points, there are still thousands of observations available to each of these estimates. The similarity of the GP and `rdrobust` approaches in this context fits the prediction that they should be similar with ample data. We also wish to know whether GP provides a suitable approach to RD for smaller sample sizes. We thus use the same data as the basis for a second analysis in which we examine much smaller sub-samples. Specifically, we:

- fix a (placebo) cutoff point: 0.3, 0.4, 0.5, 0.6, 0.7 (0.5 is the true cutoff);
- limit the data to just 0.1 above and below the cutoff in question (e.g., for cut = 0.4, data with $0.3 < X < 0.5$ are used) to maintain symmetry and comparability across cutoff estimates;
- sub-sample 200 observations (without replacement) from the remaining eligible observations;
- estimate the treatment effect and its uncertainty using various models.

We repeat this 500 times. Figure 12 shows the results. For coverage and RMSE calculations, we use the "full sample" estimates above as if it were the long-run target to reveal the change in behavior due solely to sample size. In general, the GP approaches are similar and perform well. Coverage rates vary from 96% to 99% across the GP methods, while the `gp_causaltrim` approach shows greater efficiency. As this approach makes stricter (and arguably transparency- and credibility-improving) choices, it is fortunate to also find it has desirable behavior. The `rdrobust` approach is not designed for the small

https://doi.org/10.1017/pan.2026.10032

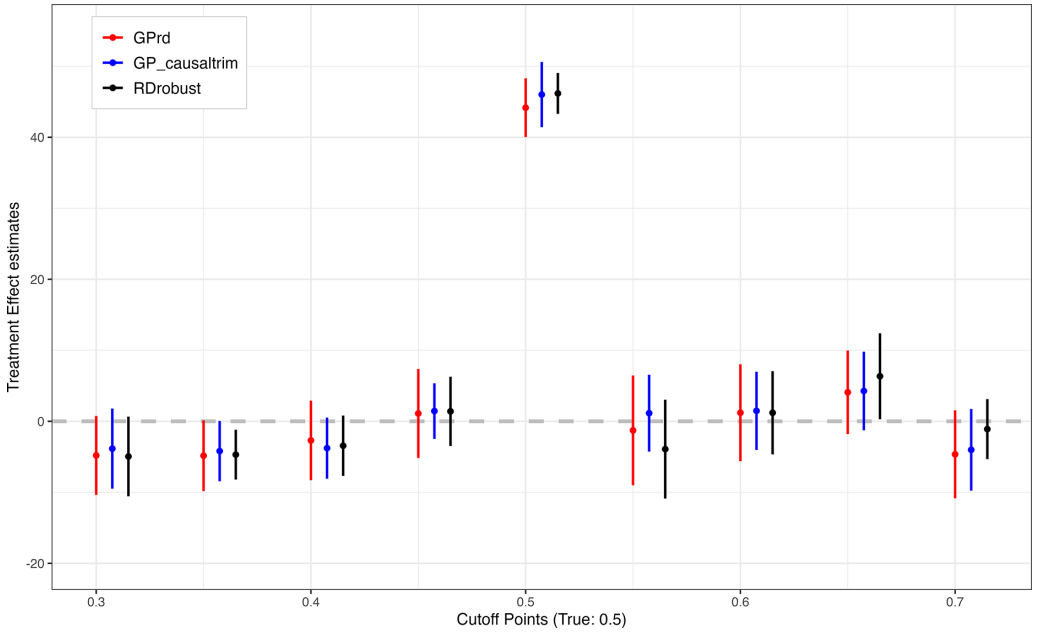


Figure 11. RDD estimates using close election data with placebo cutoff points and different GP RD Specifications. The 95% confidence interval of `GPrd` (*red*), `gp_causaltrim` (*blue*), and `rdrobust` (*black*) across eight different placebo cutoffs and the true cutoff point at 0.5 are presented.

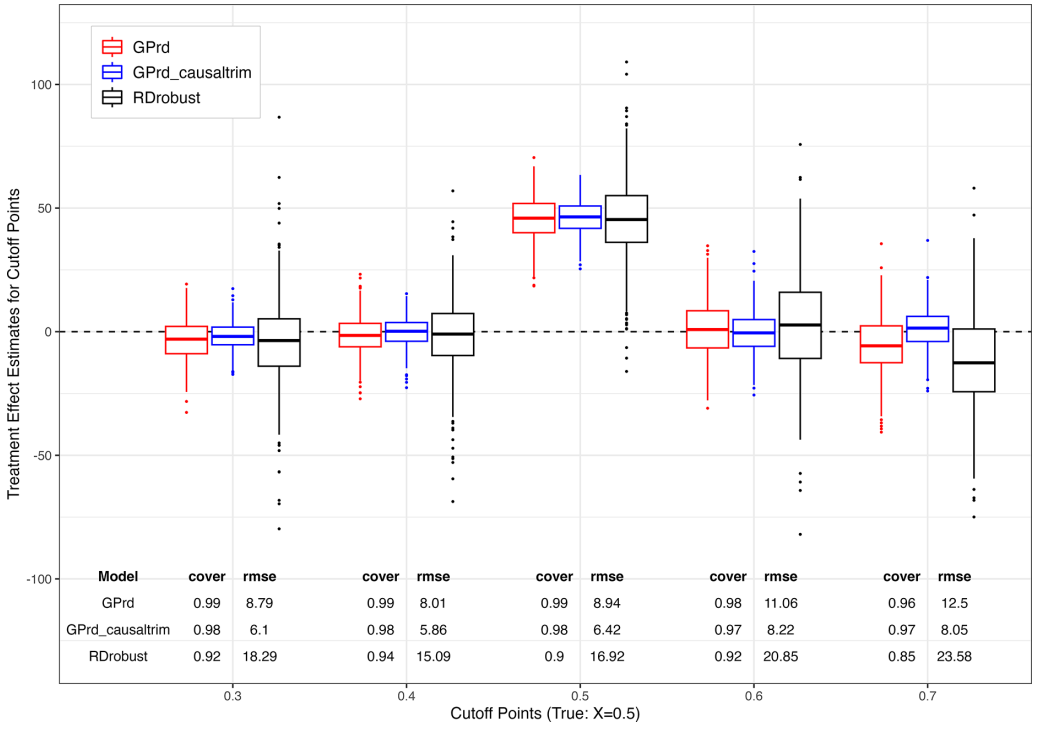


Figure 12. RDD placebo cutoff subsampling simulation. Boxplots show the distribution of treatment effect estimates from `GPrd` (*red*), `gp_causaltrim` (*blue*), and `rdrobust` (*black*) across four different placebos and true cutoff point at 0.5.

https://doi.org/10.1017/pan.2026.10032

sample setting, but we include it here to illustrate its sensitivity to sample size for those who might use it in smaller sample settings regardless. It shows occasional erratic estimates (omitted from Figure 12 with RMSE values roughly twice those of `GPrd`.[10]

3.3.4. Heteroskedasticity

As noted, the GP prior and the likelihood procedure for choosing σ^2 assume the DGP has normal errors with constant variance. In Section A.5.1 of the Supplementary Material, we study three types of heteroskedastic noise, varying by degrees up to an extreme where the variance at some locations in $\mathcal{X}$ is four times that at others. Using the "fully random" simulation setting, across types, intensities of heteroskedasticity, and sample sizes (ranging from 100 to 500), `GPrd` performs very well on coverage rates, interval lengths, and RMSE in most circumstances. The exception occurs with one heteroskedasticity pattern (decreasing variance toward the cutoff, from both sides), with larger sample sizes and greater heteroskedasticity, where GPrd still has conservative coverage and preferable RMSE, but larger interval sizes.

3.3.5. Improved RMSE: Why and When?

Though our initial motivation for employing GPs for RD was their inferential approach given the edge estimation problem, their most notable advantage is on RMSE. This improvement stems neither from employing more flexible functions, nor from the occasional extreme estimates produced by `rdrobust` in small samples. Even within the interquartile range, `rdrobust` exhibits considerably more variability than `GPrd` (see Figure 12 and Section A.5.3 of the Supplementary Material). As Ornstein and Duck-Mayr (2022) emphasize, local polynomial models may be especially susceptible to extreme behaviors at the edges. Complementing this, we argue that `GPrd` has especially low RMSE due to the role of σ^2 in regularizing the estimated CEFs on each side of the cutoff. When a small number of observations deviate substantially from others on a given side of the cutoff, GP treats them as noise rather than substantially deflecting the CEF to better fit them.[11]

To investigate this, in Section A.5.4 of the Supplementary Material, we add seven simulation settings from prior RD literature, modified to control the signal-to-noise ratio ("true R^2"). In five of the cases, `GPrd` performs well, with nominal or better coverage and lower RMSE than `rdrobust`. One simulation (from Ludwig & Miller 2007) includes a very severe deflection in the CEF at the cutoff, and on this case `GPrd` underperforms `rdrobust`. The "Lee" simulation also shows a substantial deflection. Here, `GPrd` outperforms `rdrobust` on RMSE when there is sufficient noise ($R^2 \leq 0.55$) but has higher RMSE at lower noise levels. These results fit with the reasoning that `GPrd` should excel where regularization of the CEF (especially near the cutoff) is a "good bet." We note, however, that GP can still handle relatively steep CEF changes—as demonstrated by the performance of all GP methods with the steep sigmoidal CEF in Figure 10, across all noise levels.

4. Discussion

Our analysis shows that GPs provide a practical and principled tool for causal inference problems that hinge on predictions at the edge of the data. Their central advantage is the ability to incorporate extrapolation uncertainty, widening intervals as predictions rely more heavily on assumptions beyond

[10] Some extreme `rdrobust` estimates are truncated for graphical legibility. We set a truncation threshold (drop if $|\hat{\tau}| \geq 150$) which retains over 99% of the estimates while maintaining the visual clarity of the figure. For more discussion on erratic estimates of `rdrobust` with small sample sizes in simulation settings (see Section A.5.3 of the Supplementary Material).

[11] This regularization of CEFs on either side of the cutoff mutes only the influence of abrupt, localized changes near the boundary. Recall also that the mean/intercept of each CEF is not regularized. Consequently, this approach does not shrink the RD estimate itself toward zero. This is evidenced in our simulations, as they show identical RMSE regardless of the true effect size used.

https://doi.org/10.1017/pan.2026.10032

the observed support. We demonstrate both the benefits and limitations of this approach in settings where poor overlap or extrapolation pose the greatest risks, including treatment–control comparisons with limited covariate support, ITS, and RD, particularly in small to moderate samples.

A second contribution is to make GP methods more accessible for applied researchers. We (i) explain the approach in terms familiar to social scientists, with emphasis on uncertainty quantification; (ii) simplify and automate the fitting procedure, including hyperparameter management; and (iii) provide an implementation in the R package `gpss`, available at https://doeunkim.org/gpss.

4.1. Limitations and Future Research

Several limitations of the GP framework deserve emphasis. First, GP performance depends on its assumptions. The kernel function $k(X_i, X_j)$ is taken to approximate $cov(Y_i, Y_j)$. In practice, universal kernels such as the Gaussian perform well for predictions near observed data (within the kernel width), even if they do not match the true covariance structure exactly. For more extreme extrapolation, however, non-stationary kernels may be needed to encode plausible ways the CEF could evolve beyond the data, as we illustrated in the ITS setting. In addition, the GP prior assumes that Y follows a multivariate normal distribution, and our procedure for choosing σ^2 relies on maximum likelihood under this assumption. Heteroskedastic or non-Gaussian residuals can violate these conditions. We briefly explored heteroskedasticity in the RD setting (Section A.5.1 of the Supplementary Material), but more work is needed to understand when such violations materially affect inference, and how to relax the assumptions. Finally, σ^2 acts as a regularization parameter, smoothing the fitted CEF. This shrinkage is often beneficial but can be costly when the true CEF features sharp deflections or discontinuities.

A second set of limitations concerns computation. Many GP implementations, including `gpss`, scale poorly beyond a few thousand observations because of the cost of constructing and inverting the kernel matrix. Promising approaches, such as Nyström approximations, random feature methods, and kernel sketching (e.g., Chang and Goplerud 2024), may help extend GPs to larger datasets.

Third, our empirical investigations are still limited. We have demonstrated performance in settings with one or a modest number of covariates (up to 10 in Section A.3 of the Supplementary Material), but not yet in higher-dimensional problems. Likewise, while we have begun to examine how GP compares to other approaches in the various settings above, future work should further compare GP with other methods across varied tasks and DGPs.

Finally, the logic developed here suggests potential applications beyond the designs we studied. In problems of generalizability and transportability, for instance, GP provides a natural way to construct treatment effect estimates for populations with different covariate distributions, while reflecting the added uncertainty when such estimates rely on extrapolation. We leave this and other promising application areas to future research.

Acknowledgments. We thank two anonymous reviewers, Max Goplerud, Ian Lundberg, attendees of the 2024 Summer Polmeth meeting, and members of the UCLA Practical Causal Inference Lab learning group for their comments and suggestions.

Funding Statement. This research was not funded by any specific grant from any funding agency.

Data Availability Statement. Replication materials are available at https://doi.org/10.7910/DVN/7G092W (Cho *et al.* 2025).

Author Contributions. Authorship order is alphabetical for the first two authors. C.H. is the senior co-author. All authors contributed equally.

Competing Interests. The authors declare no competing interests.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2026.10032.

https://doi.org/10.1017/pan.2026.10032

References

Ben-Michael, E., D. Arbour, A. Feller, A. Franks, and S. Raphael. 2023. "Estimating the Effects of a California Gun Control Program with Multitask Gaussian Processes." *The Annals of Applied Statistics* 17 (2): 985–1016.

Bernal, J. L., S. Cummins, and A. Gasparrini. 2017. "Interrupted Time Series Regression for the Evaluation of Public Health Interventions: A Tutorial." *International Journal of Epidemiology* 46 (1): 348–355.

Box, G. E. P., and G. M. Jenkins. 1976. *Time Series Analysis Forecasting and Control*. Hoboken, New Jersey: Wiley Series in Probability and Statistics.

Box, G. E. P., and G. C. Tiao. 1975. "Intervention Analysis with Applications to Economic and Environmental Problems." *Journal of the American Statistical Association* 70 (349): 70–79.

Branson, Z., M. Rischard, L. Bornn, and L. W. Miratrix. 2019. "A Nonparametric Bayesian Methodology for Regression Discontinuity Designs." *Journal of Statistical Planning and Inference* 202: 14–30.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik. 2017. "Rdrobust: Software for Regression-Discontinuity Designs." *The Stata Journal* 17 (2): 372–404.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik. "Regression Discontinuity Designs Using Covariates." *Review of Economics and Statistics* 101 (3): 442–451.

Calonico, S., M. D. Cattaneo, and R. Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6): 2295–2326.

Calonico, S., M. D. Cattaneo, and R. Titiunik. 2015. "Rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs." *The R Journal* 7 (1): 38.

Chang, Q., and M. Goplerud. 2024. "Generalized Kernel Regularized Least Squares." *Political Analysis* 32 (2): 157–171.

Cho, S., D. Kim, and C. Hazlett. 2025. "Replication Data for: Inference at the Data's Edge: Gaussian Processes for Estimation and Inference in the Face of Extrapolation Uncertainty." Harvard Dataverse. https://doi.org/10.7910/DVN/7G092W

Dehejia, R. H., and S. Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–1062.

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

Hartman, E., C. Hazlett, and C. Sterbenz. 2025. "Kpop: A Kernel Balancing Approach for Reducing Specification Assumptions in Survey Weighting." *Journal of the Royal Statistical Society Series A: Statistics in Society* 188 (3): 875–895.

Hinne, M., D. Leeftink, M. A. J. van Gerven, and L. Ambrogioni. 2022. "Bayesian Model Averaging for Nonparametric Discontinuity Design." *PLoS One* 17 (6): e0270310.

King, G., and L. Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14 (2): 131–159.

LaLonde, R. J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review* 76 (4): 604–620.

Lee, D. S. 2008. "Randomized Experiments from Non-random Selection in US House Elections." *Journal of Econometrics* 142 (2): 675–697.

Lee, D. S., E. Moretti, and M. J. Butler. 2004. "Do Voters Affect or Elect Policies? Evidence from the US House." *The Quarterly Journal of Economics* 119 (3): 807–859.

Ludwig, J., and D. L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *The Quarterly Journal of Economics* 122 (1): 159–208.

Micchelli, C. A., Y. Xu, and H. Zhang. 2006. "Universal Kernels." *Journal of Machine Learning Research* 7 (12): 2651–2667.

Ornstein, J., and J. B. Duck-Mayr. 2022. "*Gaussian Process Regression Discontinuity*." Technical report, Working Paper.

Prati, A., Y. Chen, J. Montgomery, and R. Garnett. 2023. "*A Gaussian Process Framework for Structured, Flexible, and Interpretable Machine-Learning Models*." *Working Paper*.

Rasmussen, C. E., and C. K. Williams. 2006. *Gaussian Processes for Machine Learning*, Vol. 1. Cambridge, MA: MIT Press.

Rischard, M., Z. Branson, L. Miratrix, and L. Bornn. 2021. "Do School Districts Affect NYC House Prices? Identifying Border Differences Using a Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs." *Journal of the American Statistical Association* 116 (534): 619–631.

Schulz, E., J. B. Tenenbaum, D. Duvenaud, M. Speekenbrink, and S. J. Gershman. 2017. "Compositional Inductive Biases in Function Learning." *Cognitive Psychology* 99: 44–79.

Stommes, D., P. M. Aronow, and F. Sävje. 2023. "On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science." *Research & Politics* 10 (2): 20531680231166457.

Wilson, A., and R. Adams. 2013. "Gaussian Process Kernels for Pattern Discovery and Extrapolation." In *International Conference on Machine Learning*, 1067–1075. PMLR.

Cite this article: Cho, S., Kim, D. and Hazlett, C. (2026). Inference at the Data's Edge: Gaussian Processes for Estimation and Inference in the Face of Extrapolation Uncertainty. *Political Analysis*, 569–588. https://doi.org/10.1017/pan.2026.10032

https://doi.org/10.1017/pan.2026.10032