



A nonparametric Bayesian methodology for regression discontinuity designs

Zach Branson^{a,*}, Maxime Rischard^a, Luke Bornn^b, Luke W. Miratrix^c

^a Department of Statistics, Harvard University, United States

^b Department of Statistics and Actuarial Science, Simon Fraser University, Canada

^c Graduate School of Education, Harvard University, United States

ARTICLE INFO

Article history:

Received 14 March 2018

Received in revised form 30 September 2018

Accepted 18 January 2019

Available online 28 January 2019

Keywords:

Regression discontinuity designs

Gaussian process regression

Bayesian nonparametrics

Coverage

Posterior consistency

ABSTRACT

One of the most popular methodologies for estimating the average treatment effect at the threshold in a regression discontinuity design is local linear regression (LLR), which places larger weight on units closer to the threshold. We propose a Gaussian process regression methodology that acts as a Bayesian analog to LLR for regression discontinuity designs. Our methodology provides a flexible fit for treatment and control responses by placing a general prior on the mean response functions. Furthermore, unlike LLR, our methodology can incorporate uncertainty in how units are weighted when estimating the treatment effect. We prove our method is consistent in estimating the average treatment effect at the threshold. Furthermore, we find via simulation that our method exhibits promising coverage, interval length, and mean squared error properties compared to standard LLR and state-of-the-art LLR methodologies. Finally, we explore the performance of our method on a real-world example by studying the impact of being a first-round draft pick on the performance and playing time of basketball players in the National Basketball Association.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Recently there has been a renewed interest in regression discontinuity designs (RDDs), which originated with [Thistlethwaite and Campbell \(1960\)](#). In an RDD, the treatment assignment is discontinuous at a certain covariate value, or “threshold”, such that only units whose covariate is above the threshold will receive treatment. There are many examples of RDDs in the applied econometrics literature: the United States providing additional funding to only the 300 poorest counties for the Head Start education program ([Ludwig and Miller, 2007](#)); schools mandating students to attend summer school if their exam scores are below a threshold ([Matsudaira, 2008](#)); colleges offering financial aid to students whose academic ability is above a cutoff ([Van der Klaauw, 2002](#)); and Medicare increasing insurance coverage after age 65 ([Card et al., 2004](#)). The main goal of an RDD is to estimate a treatment effect while addressing likely confounding by the covariate that determines treatment assignment.

One of the most popular methodologies for estimating the average treatment effect at the threshold in an RDD is local linear regression (LLR), which places larger weight on units closer to the threshold. Implementation of LLR is straightforward

* Corresponding author.

E-mail address: zbranson@g.harvard.edu (Z. Branson).

¹ This research was supported by the National Science Foundation, USA Graduate Research Fellowship Program under Grant No. 1144152, by the National Science Foundation, USA under Grant No. 1461435, by DARPA, USA under Grant No. FA8750-14-2-0117, by ARO, USA under Grant No. W911NF-15-1-0172, and by NSERC, Canada. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, DARPA, ARO, or NSERC.

and there is a wide literature on its theoretical properties. However, recent works have found that LLR can exhibit poor inferential properties – such as confidence intervals that tend to undercover – which has motivated a strand of literature started by [Calonico et al. \(2014\)](#) that modifies LLR to improve coverage and other inferential properties.

Adding to this literature, we propose a nonparametric regression approach that acts as a Bayesian analog to LLR for sharp regression discontinuity designs. Our approach utilizes Gaussian process regression (GPR) to provide a flexible fit for treatment and control responses by placing a general prior on the mean response functions. While GPR has been widely used in the machine learning and statistics literature, it has not previously been proposed for estimating treatment effects in RDDs. Thus, our main contribution is outlining how to use Gaussian processes to make causal inferences in RDDs and assess how such a methodology compares to current LLR methodologies.

In the remainder of this section, we review RDDs and LLR methodologies for estimating the average treatment effect at the threshold. In Section 2, we outline GPR for sharp RDDs and note various analogies to LLR, which builds intuition for implementing our method. In Section 3, we establish that our method is consistent in estimating the average treatment effect at the boundary. In Section 4, we show via simulation that our method exhibits promising coverage, interval length, and mean squared error properties compared to standard LLR and state-of-the-art LLR methodologies. In Section 5, we use GPR on data from the National Basketball Association (NBA) to estimate the effect of being a first-round versus a second-round pick on basketball player performance and playing time, and we find that GPR detects treatment effects that are more in line with previous results in the sports literature than do LLR methodologies. In Section 6, we conclude by discussing extensions to our methodology to tackle problems beyond sharp RDDs.

1.1. Overview of regression discontinuity designs

We follow the notation of [Imbens and Lemieux \(2008\)](#) and discuss RDDs within the potential outcomes framework: For each unit i , there are two potential outcomes, $Y_i(1)$ and $Y_i(0)$, corresponding to treatment and control, respectively, and a covariate X_i . Only one of these two potential outcomes is observed, but X_i is always observed. Let W_i denote the assignment for unit i , where $W_i = 1$ if i is assigned to treatment and $W_i = 0$ if unit i is assigned to control. The observed outcome for unit i is then

$$y_i = W_i Y_i(1) + (1 - W_i) Y_i(0) \quad (1)$$

Often, one wants to estimate the average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)]$, but usually this treatment effect is confounded by X (and possibly other unobserved covariates), such that examining the difference in mean response between treatment and control is not appropriate. In these cases, methods such as stratification, matching, and regression are often employed to address covariate confounding when estimating the average treatment effect. However, such methods are only appropriate when there is sufficient overlap in the treatment and control covariate distributions, i.e.,

$$0 < P(W_i = 1 | X_i = x) < 1 \quad (2)$$

For example, this overlap assumption is essential for propensity score methodologies, where the relationship between W_i and X_i is estimated and then accounted for during the analysis.

In an RDD, the relationship between the treatment assignment W_i and the covariates is known. More specifically, it is known that the function $P(W = 1 | X = x)$ is discontinuous at some threshold or boundary $X = b$. In this paper we focus on a special case, *sharp RDDs*, where the treatment assignment for a unit i is

$$W_i = \begin{cases} 1 & \text{if } X_i \geq b \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The covariate X that determines treatment assignment in an RDD is often called the “running variable” in order to not confuse it with other background covariates that do not necessarily determine treatment assignment. The more general case, where $P(W = 1 | X = x)$ is discontinuous at $X = b$ but is not necessarily 0 or 1, is known as a *fuzzy RDD*.

[Rubin \(1977\)](#) states that when treatment assignment depends on one covariate, estimating $\mathbb{E}[Y_i(1) | X_i]$ and $\mathbb{E}[Y_i(0) | X_i]$ is essential for estimating the average treatment effect. Furthermore, treatment effect estimates are particularly sensitive to model specification of $\mathbb{E}[Y_i(1) | X_i]$ and $\mathbb{E}[Y_i(0) | X_i]$ when there is not substantial covariate overlap, as in a sharp RDD. Aware of this sensitivity, RDD analyses typically do not attempt to estimate the average treatment effect. Instead, they focus on the estimand that requires the least amount of extrapolation to overcome this lack of covariate overlap: the average treatment effect *at the boundary* b . Defining the conditional expectations for treatment and control as

$$\mu_T(x) \equiv \mathbb{E}[Y_i(1) | X_i = x], \quad \mu_C(x) \equiv \mathbb{E}[Y_i(0) | X_i = x], \quad (4)$$

the treatment effect at the boundary b is ([Imbens and Lemieux, 2008](#))

$$\tau = \lim_{x \downarrow b} \mathbb{E}[y_i | X_i = x] - \lim_{x \uparrow b} \mathbb{E}[y_i | X_i = x] = \mu_T(b) - \mu_C(b). \quad (5)$$

The notation $\mu_T(x)$ and $\mu_C(x)$ emphasizes that the goal of an RDD requires estimating two unknown mean functions. [Hahn et al. \(2001\)](#) showed that sufficient conditions for τ to be identifiable in a sharp RDD are that $\mu_C(x)$ and $\mu_T(x) - \mu_C(x)$ are continuous at b . They further state that “we can use any nonparametric estimator to estimate” $\mu_T(x)$ and $\mu_C(x)$, and recommended local linear regression (hereafter called LLR), which is currently the most popular methodology for estimating the average treatment effect at the boundary in an RDD.

1.2. Review of local linear regression

The goal of an RDD is to estimate τ defined in (5), i.e., to estimate $\mu_T(b)$ and $\mu_C(b)$. LLR estimates $\mu_T(b)$ as

$$\hat{\mu}_T(b) = X_b(X_T^T W_T X_T)^{-1} X_T^T W_T Y_T \quad (6)$$

where $X_b \equiv (1 \quad b)$, X_T is the $n_T \times 2$ design matrix corresponding to the intercept and running variable X for treated units, Y_T is the n_T -dimensional column vector of treated units' responses, and W_T is a $n_T \times n_T$ diagonal weight matrix whose entries are

$$(W_T)_{ii} \equiv K\left(\frac{x_i - b}{h}\right), \quad i = 1, \dots, n_T \quad (7)$$

for some kernel $K(\cdot)$ and bandwidth h . The estimator $\hat{\mu}_C(b)$ is analogously defined for the control. Then, the estimator for the treatment effect is $\hat{\tau} = \hat{\mu}_T(b) - \hat{\mu}_C(b)$. Here, $\hat{\mu}_T(b)$ and $\hat{\mu}_C(b)$ are weighted least squares estimators, where the weights depend on units' closeness to the boundary b . To perform local polynomial regression, one appends higher orders of X to the design matrices X_T and X_C .

The RDD literature has focused on LLR largely because of its boundary bias properties. For example, [Hahn et al. \(2001\)](#) recommend LLR over alternatives like kernel regression because [Fan, 1992](#) showed that LLR exhibits better bias properties for boundary points than kernel regression. For more details on the bias comparison between kernel regression and LLR, see [Imbens and Lemieux \(2008\)](#) (p. 624–625) as well as [Fan and Gijbels \(1992\)](#) and [Porter \(2003\)](#).

Furthermore, LLR's implementation is straightforward once a kernel $K(\cdot)$ and bandwidth h are chosen. The most common choice of $K(\cdot)$ is the rectangular or triangular kernel; [Imbens and Lemieux \(2008\)](#) argue that more complicated kernels rarely make a difference in estimation. Much more attention has been given to the bandwidth choice h , largely because the bias is characterized by h . In the 2000s, choosing an appropriate h for LLR in RDDs was an open problem: For example, [Ludwig and Miller \(2007\)](#) stated that “there is currently no widely-agreed-upon method for selection of optimal bandwidths...so our strategy is to present results for a broad range of candidate bandwidths”. One widely-used bandwidth selection method is that of [Imbens and Kalyanaraman \(2012\)](#), who derived a data-driven, MSE-optimal bandwidth for LLR estimators. This provided practitioners with clear guidelines for implementing LLR for RDDs, which made its use very popular.

The bandwidth is arguably the most important choice to be made in the LLR methodology for RDDs, because the treatment effect is often sensitive to the bandwidth choice. This motivates sensitivity checks such as that in [Ludwig and Miller \(2007\)](#), where the treatment effect is estimated several times with different bandwidths to ensure that estimates do not vary too greatly. Some have noted that confidence intervals from LLR have a tendency to undercover when a single bandwidth is chosen for inference when the treatment effect is sensitive to the bandwidth choice ([Armstrong and Kolesár, 2017](#); [Gelman and Imbens, 2018](#)). A recent extension to the LLR methodology – that of [Calonico et al. \(2014\)](#), hereafter called “robust LLR” – was one of the first methods to address the undercoverage issue of LLR by incorporating a bias correction and inflated confidence intervals corresponding to the uncertainty in estimating the bias correction. Because of its promising inference properties, [Calonico et al. \(2014\)](#) has arguably become the state-of-the-art for conducting inference for the average treatment effect at the boundary in an RDD.

1.3. Other methods besides LLR

Other methodologies besides LLR have been proposed for estimating the average treatment effect at the boundary in a sharp RDD. For example, many practitioners have used high-order global polynomials to estimate $\mu_T(x)$ and $\mu_C(x)$: [Matsuda-daira \(2008\)](#) argued for a global third-order polynomial regression, and also considered fourth- and fifth-order polynomials as a sensitivity check; similarly, [Van der Klaauw \(2002\)](#) used a global third-order polynomial and noted that LLR could have been an alternative; finally, [Card et al. \(2004\)](#) argued for using a global third-order polynomial regression instead of LLR because the running variable, age, was discrete. However, in recent years many have argued against the use of high-order polynomials in RDDs because of their tendency to yield point estimates and confidence intervals that are highly sensitive to the order of the polynomial ([Calonico et al., 2015](#); [Gelman and Imbens, 2018](#)).

Others have focused on local randomization methodologies, where units within a window around the boundary are viewed as-if randomized to treatment and control. For example, [Cattaneo et al. \(2015\)](#) recommend a series of covariate balance tests to decide the window around the boundary such that the as-if randomized assumption is most plausible. [Li et al. \(2015\)](#) extended these ideas to develop a notion of a local overlap assumption and used a Bayesian hierarchical modeling approach for deciding the window around the boundary where this assumption is most plausible. [Cattaneo et al. \(2017\)](#) compared the local randomization approach to local polynomial estimators, and they extended the local randomization approach to incorporate adjustments via parametric models as well.

Finally, others have developed Bayesian methodologies for RDDs. [Li et al. \(2015\)](#) propose a principal stratification approach that provides alternative identification assumptions based on a formal definition of local randomization. [Geneletti et al. \(2015\)](#) propose a Bayesian methodology that incorporates prior information in the treatment effect. [Chib and Greenberg \(2014\)](#) use Bayesian splines to estimate treatment effects in RDDs. [Chib and Jacobi \(2015\)](#) propose a Bayesian methodology specific to fuzzy RDDs.

1.4. Our proposal: Gaussian process regression for sharp RDDs

We propose a methodology that utilizes Gaussian process regression (GPR), which is one of the most popular nonparametric methodologies in the machine learning and Bayesian modeling literature for estimating unknown functions (Rasmussen and Williams, 2006). The notion of using GPR for RDDs is very much in line with the claim in Hahn et al. (2001) that any nonparametric estimator can be used to estimate the treatment and control response in an RDD. However, to our knowledge, GPR has not been previously proposed for RDDs.

Similar to LLR, our GPR methodology provides a flexible fit to the mean functions $\mu_T(x)$ and $\mu_C(x)$. Furthermore, our methodology can incorporate both prior knowledge and uncertainty in various parameters in the RDD problem – such as how units are weighted near the boundary – which is not necessarily as straightforward with current LLR methodologies. Finally, our GPR methodology can be used in conjunction with a local randomization perspective. Our methodology adds to the strand of literature started by Calonico et al. (2014) that addresses the undercoverage of standard LLR, as well as the strand of literature on Bayesian methodologies for RDDs.

2. GPR models to estimate the average treatment effect at the boundary

First we review notation for GPR and how GPR is used to estimate a single unknown function. We then discuss GPR models that estimate the two unknown mean functions in sharp RDDs.

2.1. Notation for estimating one unknown function

Define a dataset $\{x_i, y_i\}_{i=1}^n$ of responses $\mathbf{y} = (y_1, \dots, y_n)$ that varies around some unknown function of the covariate $\mathbf{x} = (x_1, \dots, x_n)$:

$$\mathbf{y} = f(\mathbf{x}) + \epsilon \quad (8)$$

where $\epsilon \sim \mathcal{N}_n(0, \sigma_y^2 \mathbf{I}_n)$ and $f(\mathbf{x}) \equiv (f(x_1), \dots, f(x_n))$. If the goal is to well-estimate $\mathbb{E}[f(x^*)]$ for a particular covariate value x^* , one option is to specify a functional form for $f(\mathbf{x})$ and then predict $\mathbb{E}[f(x^*)]$ from this specified model, such as local linear regression, as discussed in Section 1. Instead of specifying a functional form for $f(\mathbf{x})$, we consider nonparametrically inferring $f(\mathbf{x})$ by placing a prior on $f(\mathbf{x})$. A Gaussian process is one such prior:

$$f(\mathbf{x}) \sim \text{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')) \quad (9)$$

for some unknown mean function $m(\mathbf{x})$ and covariance function $K(\mathbf{x}, \mathbf{x}')$. The notation $f(\mathbf{x}) \sim \text{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$ denotes a Gaussian process prior on the unknown function $f(\mathbf{x})$, which states that, for any (x_1, \dots, x_n) , the joint distribution $(f(x_1), \dots, f(x_n))$ is an n -dimensional multivariate normal distribution with mean vector $(m(x_1), \dots, m(x_n))$ and covariance matrix $K(\mathbf{x}, \mathbf{x}')$ whose (i, j) entries are $K(x_i, x_j)$.

There are many choices one could make for the mean and covariance functions. A common choice for the mean function is $m(\mathbf{x}) = \mathbf{0}$; a common choice for the covariance function is the squared exponential covariance function, whose entries are

$$K(x_i, x_j) \equiv \sigma_{GP}^2 \exp\left(-\frac{1}{2\ell^2}(x_i - x_j)^2\right). \quad (10)$$

Placing a Gaussian process prior with a squared exponential covariance function on $f(\mathbf{x})$ assumes that $f(\mathbf{x})$ is infinitely differentiable, which is similar to other assumptions in the RDD literature (e.g., Assumption 3.3 of Imbens and Kalyanaraman, 2012 and Assumption 1 of Calonico et al., 2014). The covariance parameters σ_{GP}^2 and ℓ are called the variance and lengthscale, respectively. The variance determines the amplitude of $f(\mathbf{x})$, i.e., how much the function varies from its mean. The lengthscale determines the smoothness of the function: Small lengthscales correspond to $f(\mathbf{x})$ changing rapidly. Most importantly, the covariance function assumes that the response at a particular covariate value $f(x^*)$ will be similar to the response at covariate values close to x^* .

Given the Gaussian process prior with mean function $m(\mathbf{x})$ and covariance function $K(\mathbf{x}, \mathbf{x}')$, as well as their parameters, the posterior for $f(x^*)$ at any particular covariate value x^* can be obtained via standard conditional multivariate normal theory (for an exposition, see Rasmussen and Williams, 2006, Pages 16–17):

$$f(x^*)|\mathbf{x}, \mathbf{y} \sim N(\mu^*, \sigma_{GP}^2 - \Sigma^*), \quad \text{where} \quad (11)$$

$$\begin{aligned} \mu^* &\equiv m(x^*) + K(x^*, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \mathbf{I}_n]^{-1}(\mathbf{y} - m(\mathbf{x})) \\ \Sigma^* &\equiv K(x^*, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \mathbf{I}_n]^{-1}K(\mathbf{x}, x^*) \end{aligned} \quad (12)$$

The above posterior can thus be used to obtain point estimates and credible intervals for the value of an unknown function at a particular covariate value x^* . In practice, the covariance parameters are estimated from the data, such as through maximum likelihood or cross-validation (Rasmussen and Williams, 2006, Chapter 5). In Section 2.2 we assume that the covariance parameters are fixed, and in Section 2.3 we extend to a full Bayesian approach that places priors on ℓ , σ_{GP}^2 , and σ_y^2 .

2.2. GPR models for sharp RDDs

The notion of using GPR to estimate the average treatment effect at the boundary in an RDD suggests a class of models that has not previously been considered in the RDD literature. We focus on two GPR models, which correspond to different assumptions placed on the unknown response functions $\mu_T(x)$ and $\mu_C(x)$. For each model we show the resulting posterior for the average treatment effect at the boundary and compare it to its analogous LLR model. In Section 2.3 we discuss how – unlike LLR methodologies – the uncertainty in how units are weighted can be incorporated into these GPR models.

For both GPR models, we assume that the treatment response $Y_i(1)$ and the control response $Y_i(0)$ have the following relationship with the running variable x_i

$$Y_i(1) = \mu_T(x_i) + \epsilon_{i1}, \quad \text{and} \quad Y_i(0) = \mu_C(x_i) + \epsilon_{i0}, \quad \text{where} \quad (13)$$

$$\mu_T(x_i) \perp \mu_C(x_i), \quad \epsilon_{i1} \stackrel{iid}{\sim} N(0, \sigma_{y1}^2), \quad \text{and} \quad \epsilon_{i0} \stackrel{iid}{\sim} N(0, \sigma_{y0}^2)$$

Thus, intuitively, the procedure outlined in Section 2.1 can simply be performed twice – once for $\mu_T(x)$ and once for $\mu_C(x)$. However, there are assumptions on $\mu_T(x)$ and $\mu_C(x)$ that, if true, can simplify our GPR models and make inference more precise.

In particular, assumptions can be placed on the covariance structure of $\mu_T(x)$ and $\mu_C(x)$. The two models we present correspond to two different sets of assumptions – the first assumes that the covariance structure of $\mu_T(x)$ and $\mu_C(x)$ are the same, while the second allows them to be different. In both models, we assume that $\mu_T(x)$ and $\mu_C(x)$ are stationary processes, i.e., the covariance parameters of $\mu_T(x)$ and $\mu_C(x)$ do not vary with the running variable X . We discuss cases when this stationarity assumption is inappropriate in Sections 4 and 6.

Same Covariance Assumption: $\text{Cov}(\mu_T(x)) = \text{Cov}(\mu_C(x))$, and $\mu_T(x)$ and $\mu_C(x)$ are stationary processes.

If the Same Covariance Assumption holds, a natural LLR procedure is to fit local linear regressions on both sides of the boundary with the same bandwidth but different intercepts and slopes. This is largely the standard practice in the RDD literature (Imbens and Lemieux, 2008). Analogously, we place Gaussian process priors on $\mu_T(x)$ and $\mu_C(x)$ for given mean functions $m_T(\mathbf{x})$ and $m_C(\mathbf{x})$ and covariance function $K(\mathbf{x}, \mathbf{x}')$:

$$\begin{aligned} \mu_T(x) &\sim \text{GP}(m_T(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')) \\ \mu_C(x) &\sim \text{GP}(m_C(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')) \end{aligned} \quad (14)$$

Then, estimates $\hat{\mu}_T(b)$ and $\hat{\mu}_C(b)$ are obtained, which results in a treatment effect estimate $\hat{\tau} = \hat{\mu}_T(b) - \hat{\mu}_C(b)$. We now outline how such estimates $\hat{\mu}_T(b)$ and $\hat{\mu}_C(b)$ are obtained. Using standard conditional multivariate normal theory as in Section 2.1, we have the following posteriors for $\mu_T(b)$ and $\mu_C(b)$:

$$\begin{aligned} \mu_T(b) | \mathbf{x}, \mathbf{y} &\sim N(\mu_{b|T}, \sigma_{GP}^2 - \Sigma_{b|T}) \\ \mu_C(b) | \mathbf{x}, \mathbf{y} &\sim N(\mu_{b|C}, \sigma_{GP}^2 - \Sigma_{b|C}), \quad \text{where} \end{aligned} \quad (15)$$

$$\begin{aligned} \mu_{b|T} &\equiv m_T(b) + K(b, \mathbf{x}_T)[K(\mathbf{x}_T, \mathbf{x}_T) + \sigma_y^2 \mathbf{I}]^{-1}(\mathbf{y}_T - m_T(\mathbf{x}_T)) \\ \mu_{b|C} &\equiv m_C(b) + K(b, \mathbf{x}_C)[K(\mathbf{x}_C, \mathbf{x}_C) + \sigma_y^2 \mathbf{I}]^{-1}(\mathbf{y}_C - m_C(\mathbf{x}_C)) \\ \Sigma_{b|T} &\equiv K(b, \mathbf{x}_T)[K(\mathbf{x}_T, \mathbf{x}_T) + \sigma_y^2 \mathbf{I}]^{-1}K(\mathbf{x}_T, b) \\ \Sigma_{b|C} &\equiv K(b, \mathbf{x}_C)[K(\mathbf{x}_C, \mathbf{x}_C) + \sigma_y^2 \mathbf{I}]^{-1}K(\mathbf{x}_C, b) \end{aligned} \quad (16)$$

Here, $\mu_{b|T}$ and $\mu_{b|C}$ denote the posterior mean for $\mu_T(b)$ and $\mu_C(b)$, respectively, which are in the definition of the treatment effect τ defined in (5). Note that $\mu_{b|T}$ and $\mu_{b|C}$ are weighted averages of the observed response, where the weights $K(b, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \mathbf{I}]^{-1}$ depend on the covariance parameters ℓ , σ_{GP}^2 , and σ_y^2 , as well as \mathbf{x} . For more discussion on the behavior of the weights $K(b, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \mathbf{I}]^{-1}$, see Rasmussen and Williams (2006, Section 2.6).

The posterior for the treatment effect under the Same Covariance Assumption is then

$$\tau \equiv \mu_T(b) - \mu_C(b) | \mathbf{x}, \mathbf{y} \sim N(\mu_{b|T} - \mu_{b|C}, 2\sigma_{GP}^2 - \Sigma_{b|T} - \Sigma_{b|C}). \quad (17)$$

where we have also used the independence of $\mu_T(x)$ and $\mu_C(x)$ stated in (13). If the Same Covariance Assumption does not hold, one can still assume that the mean treatment and control response processes are stationary, but allow both the mean and covariance to vary on either side of the boundary.

Stationary Assumption: $\mu_T(x)$ and $\mu_C(x)$ are stationary processes.

The posterior in this case would be identical to (17), except the means $\mu_{b|T}$ and $\mu_{b|C}$ and covariances $\Sigma_{b|T}$ and $\Sigma_{b|C}$ are instead defined as

$$\begin{aligned} \mu_{b|T} &\equiv m_T(b) + K_T(b, \mathbf{x}_T)[K_T(\mathbf{x}_T, \mathbf{x}_T) + \sigma_{y1}^2 \mathbf{I}]^{-1}(\mathbf{y}_T - m_T(\mathbf{x}_T)) \\ \mu_{b|C} &\equiv m_C(b) + K_C(b, \mathbf{x}_C)[K_C(\mathbf{x}_C, \mathbf{x}_C) + \sigma_{y0}^2 \mathbf{I}]^{-1}(\mathbf{y}_C - m_C(\mathbf{x}_C)), \\ \Sigma_{b|T} &\equiv K_T(b, \mathbf{x}_T)[K_T(\mathbf{x}_T, \mathbf{x}_T) + \sigma_{y1}^2 \mathbf{I}]^{-1}K_T(\mathbf{x}_T, b) \\ \Sigma_{b|C} &\equiv K_C(b, \mathbf{x}_C)[K_C(\mathbf{x}_C, \mathbf{x}_C) + \sigma_{y0}^2 \mathbf{I}]^{-1}K_C(\mathbf{x}_C, b) \end{aligned} \quad (18)$$

i.e., the shared covariance $K(\cdot, \cdot)$ is replaced with $K_T(\cdot, \cdot)$ for units receiving treatment and $K_C(\cdot, \cdot)$ for units receiving control. The analogous LLR methodology would be to allow different intercepts, slopes, and bandwidths on either side of the boundary. However, using different bandwidths on either side of the boundary is rarely done in practice. For example, [Imbens and Lemieux \(2008\)](#) argue that if the curvature of $\mu_T(x)$ and $\mu_C(x)$ is the same, then the large-sample optimal bandwidths should be the same; and, furthermore, there is additional variance in estimating two optimal bandwidths rather than one, due to the smaller sample used to estimate each bandwidth. Thus, a benefit of the Same Covariance Assumption is that it allows researchers to use the entire data to estimate one set of covariance parameters, instead of estimating two separate sets of covariance parameters for treatment and control. However, when fitting our GPR model, we do not recommend sharing information between $\mu_T(x)$ and $\mu_C(x)$ beyond estimating their covariance structure — this follows the general practice in the RDD literature to fit separate regression functions (that may nonetheless share the same bandwidth) for the treatment and control groups.

The above posteriors for these two GPR models assume fixed mean and covariance parameters. In practice, maximum-likelihood or cross-validation can be used for estimating these parameters. In Section 2.3, we extend the above GPR models to a full-Bayesian approach that incorporates uncertainty in the mean and covariance parameters.

2.3. Accounting for mean and covariance function uncertainty

The GPR models in Section 2.2 assume that $\mu_T(b)$ and $\mu_C(b)$ will be similar to $\mu_T(x)$ and $\mu_C(x)$, respectively, for x near b . The extent of this similarity is determined by the covariance function $K(\mathbf{x}, \mathbf{x}')$ and its parameters. In particular, recall that in Section 2.2 we showed that the posterior mean of the average treatment effect for GPR is characterized by a difference of two weighted averages, where the weights are of the form $K(b, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \mathbf{I}]^{-1}$. Thus, incorporating uncertainty in the covariance parameters in turn incorporates uncertainty in how units are weighted when estimating the average treatment effect.

Denote the mean function parameters by θ_m and the covariance function parameters by θ_K . For example, consider the mean functions

$$\begin{aligned} m_T(\mathbf{x}) &= \mathbf{h}(\mathbf{x})^T \boldsymbol{\beta}_T \\ m_C(\mathbf{x}) &= \mathbf{h}(\mathbf{x})^T \boldsymbol{\beta}_C \end{aligned} \quad (19)$$

where $\mathbf{h}(\mathbf{x}) = (1, x, \dots, x^{p-1})$, and $\boldsymbol{\beta}_T$ and $\boldsymbol{\beta}_C$ are the corresponding p -dimensional column vectors. In this case, $\theta_m = (\boldsymbol{\beta}_T, \boldsymbol{\beta}_C)$. For the squared exponential covariance function defined in (10), $\theta_K = (\ell, \sigma_{GP}^2, \sigma_y^2)$.

In order to incorporate uncertainty in θ_m and θ_K , one can first obtain draws $1, \dots, D$ from the joint posterior of (θ_m, θ_K) , rather than obtaining point-estimates $\hat{\theta}_m$ and $\hat{\theta}_K$. Then, for each draw $(\theta_m, \theta_K)_1, \dots, (\theta_m, \theta_K)_D$, one draws from the posterior for τ , defined in (17).

Section 2.2 already defines the likelihood for the GPR models, so all that remains is to specify priors for (θ_m, θ_K) in order to obtain draws from the joint posterior of (θ_m, θ_K) . Priors for (θ_m, θ_K) will depend on the choice of mean and covariance functions. For example, for the mean functions defined in (19), we recommend $\mathcal{N}(\mathbf{0}, B)$ as a prior for $\boldsymbol{\beta}_T$ and $\boldsymbol{\beta}_C$, where B is a $p \times p$ diagonal matrix with reasonably large entries ([Rasmussen and Williams, 2006](#), Pages 28–29). For the squared exponential covariance function given by (10), we recommend half-Cauchy priors for the covariance parameters $\frac{1}{\ell^2}$ and σ_{GP}^2 and noise σ_y^2 , following advice from [Gelman \(2006\)](#) about stable priors for variance parameters.

Now we prove that our GPR methodology is consistent in estimating the average treatment effect at the boundary. First we establish consistency when the covariance parameters are fixed, and then we consider the case where priors are placed on the covariance parameters.

3. Posterior consistency of our GPR models

Gaussian processes are known to exhibit posterior consistency under minimal assumptions. [Ghosal and Roy \(2006\)](#) proved posterior consistency of binary GPR for fixed covariance parameters, and [Choi and Schervish \(2007\)](#) proved posterior consistency of GPR when the response is continuous. More generally, [van der Vaart and van Zanten \(2008\)](#) studied the contraction rate for Gaussian process priors for density estimation and regression problems, and [van der Vaart and van Zanten \(2009\)](#) extended these results to when a prior is placed on the lengthscale of a Gaussian process.

Here we evaluate our Bayesian methodology from a frequentist point-of-view, which assumes a fixed treatment effect at the boundary. The GPR models in Section 2.2 estimate the treatment effect as the difference between two Gaussian process regressions; thus, our posterior of the treatment effect is consistent if the separate GPRs on either side of the discontinuity are consistent. First we establish posterior consistency assuming the covariance parameters are fixed, as in Section 2.2, and then we extend these results to when a prior is placed on the covariance parameters, as in Section 2.3.

We prove posterior consistency assuming the Stationary Assumption in Section 2.2, but the results also hold for the Same Covariance Assumption. Furthermore, we assume the mean functions $m_T(x) = 0$, $m_C(x) = 0$ and the squared exponential covariance function defined in (10). Discussion about the extent to which these results extend to other choices for the mean and covariance functions is in the [Appendix](#). The other assumptions necessary for [Theorems 1 and 2](#) follow [van der Vaart and van Zanten \(2009\)](#) and are also given in the [Appendix](#).

Theorem 1. Assume that the Stationary Assumption holds, the covariance functions $K_T(x, x)$ and $K_C(x, x)$ are fixed, and Assumptions A1, A2, and A3 given in the [Appendix](#) hold. Denote the true average treatment effect at the boundary as $\tau^* = \mu_T^*(b) - \mu_C^*(b)$, where $\mu_T^*(x)$ and $\mu_C^*(x)$ are the true mean treatment and control response functions in the model (13). Let $\prod(\tau | x_1, \dots, x_n)$ denote the posterior distribution of the average treatment effect at the boundary, defined in (17). Then, this posterior is consistent, in the sense that

$$\prod(\tau : h(\tau, \tau^*) \geq M\epsilon_n | x_1, \dots, x_n) \xrightarrow{P_{\tau^*}} 0 \quad (20)$$

for sufficiently large M , where h is the Hellinger distance, and ϵ_n is the rate at which the posterior of τ contracts to the true τ^* .

The proof of [Theorem 1](#), as well as a discussion about the nature of the contraction rate, is given in the [Appendix](#). [Theorem 2](#) establishes posterior consistency when a prior is placed on the lengthscale parameter ℓ , instead of being held fixed (see [Section 2.3](#)). A discussion about posterior consistency when an additional prior is placed on σ_{GP}^2 is in the [Appendix](#).

Theorem 2. Assume that the Stationary Assumption holds, the σ_{GP}^2 parameters in $K_T(x, x)$ and $K_C(x, x)$ are fixed, and Assumptions A1, A2, A3, and A4 given in the [Appendix](#) hold. Then, [Theorem 1](#) holds.

The proof of [Theorem 2](#) is given in the [Appendix](#). A corollary follows from the proofs of [Theorems 1](#) and [2](#).

Corollary 1. [Theorems 1](#) and [2](#) hold if the Same Covariance Assumption holds instead of the Stationary Assumption.

4. Simulation results

Here we investigate how our Gaussian Process model compares to standard LLR and the robust LLR method introduced in [Calonico et al. \(2014\)](#). We choose these two methods because the former is the standard in both applied work and the RDD literature at large, and the latter is a recent method that attempts to solve the undercoverage issue of standard LLR. We focus on the GPR model assuming the Same Covariance Assumption in [Section 2.2](#) and use the mean functions

$$\begin{aligned} m_T(\mathbf{x}) &= \beta_{0T} + \beta_{1T}\mathbf{x} \\ m_C(\mathbf{x}) &= \beta_{0C} + \beta_{1C}\mathbf{x} \end{aligned} \quad (21)$$

and the squared exponential covariance function given by (10). These assumptions are analogous to the LLR procedure of fitting separate local linear regressions in treatment and control with differing slopes but the same bandwidth. Specification of the mean function in the Gaussian process prior is typically not consequential for estimation; however, as discussed in [Rasmussen and Williams \(2006, Section 2.7\)](#), there can be some benefits to specifying a mean function, as we do here. In particular, the above specification allows GPR predictions to pull towards a global linear trend instead of a global mean (which would be the case if we used a zero mean function – a common choice in the literature – for the Gaussian process prior). This can be useful within the context of extrapolation towards the boundary, as in an RDD. In the [Appendix](#) in [Table 3](#), we present simulation results for our GPR methodology using a zero mean function instead of the above linear mean function for the Gaussian process prior. The results for that case are largely the same as the results presented here, which suggests that our results are insensitive to specification of the mean function in the Gaussian process prior.

As discussed in [Section 2.3](#), we took a full-Bayesian approach to our GPR methodology and placed independent $\mathcal{N}(0, 100^2)$ priors on the mean function parameters in (21) and independent half-Cauchy(0, 5) priors on the covariance parameters. These choices for the prior distributions are in line with common recommendations in the Bayesian data analysis literature: The choice of Normal priors on the mean function parameters follows recommendations from [Rasmussen and Williams \(2006\)](#), and the choice of half-Cauchy priors on the covariance parameters follows recommendations from [Gelman et al. \(2006\)](#), [Polson and Scott \(2012\)](#), and [Gelman et al. \(2013, Chapter 5\)](#). We used the R package `rstan` ([Carpenter et al., 2016](#)) to sample from the posterior of these parameters. In the [Appendix](#) in [Table 3](#), we discuss simulation results for GPR when we instead plug in the MLE for the covariance parameters; however, we found that the full-Bayesian approach is preferable in terms of inferential properties, which suggests that it is beneficial to incorporate uncertainty in the covariance parameters for our GPR method.

We conduct a simulation study based on simulations from [Imbens and Kalyanaraman \(2012\)](#) and [Calonico et al. \(2014\)](#). In all simulations, we generated 1000 datasets of 500 observations $\{(x_i, y_i, \epsilon_i) : i = 1, \dots, 500\}$, where $x_i \sim 2\text{Beta}(2, 4) - 1$, $\epsilon_i \sim \mathcal{N}(0, 0.1295^2)$, and $y_i = \mu_j(x_i) + \epsilon_i$ for different mean functions $\mu_j(x_i)$.

We consider seven different mean functions (see [Fig. 1](#)), which we call Lee, Quad, Cate 1, Cate 2, Ludwig, Curvature, and Cubic. Lee, Quad, Cate 1, and Cate 2 were used in [Imbens and Kalyanaraman \(2012\)](#), and Lee, Ludwig, and Curvature were used in [Calonico et al. \(2014\)](#); details about these datasets can be found in [Imbens and Kalyanaraman \(2012\)](#) (Page 18) and [Calonico et al. \(2014\)](#) (Page 20). We also introduce the Cubic mean function as a comparison to the Quad mean function, because in the Quad mean function the linear trends on either side of b are the opposite sign, whereas those for the Cubic mean function are the same sign.

The boundary for each dataset is $b = 0$. The treatment effect is $\tau = 0.04$ for Lee and Curvature, $\tau = 0$ for Quad and Cubic, $\tau = 0.1$ for Cate 1 and 2, and $\tau = -3.35$ for Ludwig. Also displayed in [Fig. 1](#) are a set of sample points $\{(x_i, y_i), i = 1, \dots, 500\}$ for each mean function, which shows what one dataset looks like for each mean function. One can see that – although $\epsilon_i \sim \mathcal{N}(0, 0.1295^2)$ for all mean functions – the relative noise varies across mean functions.

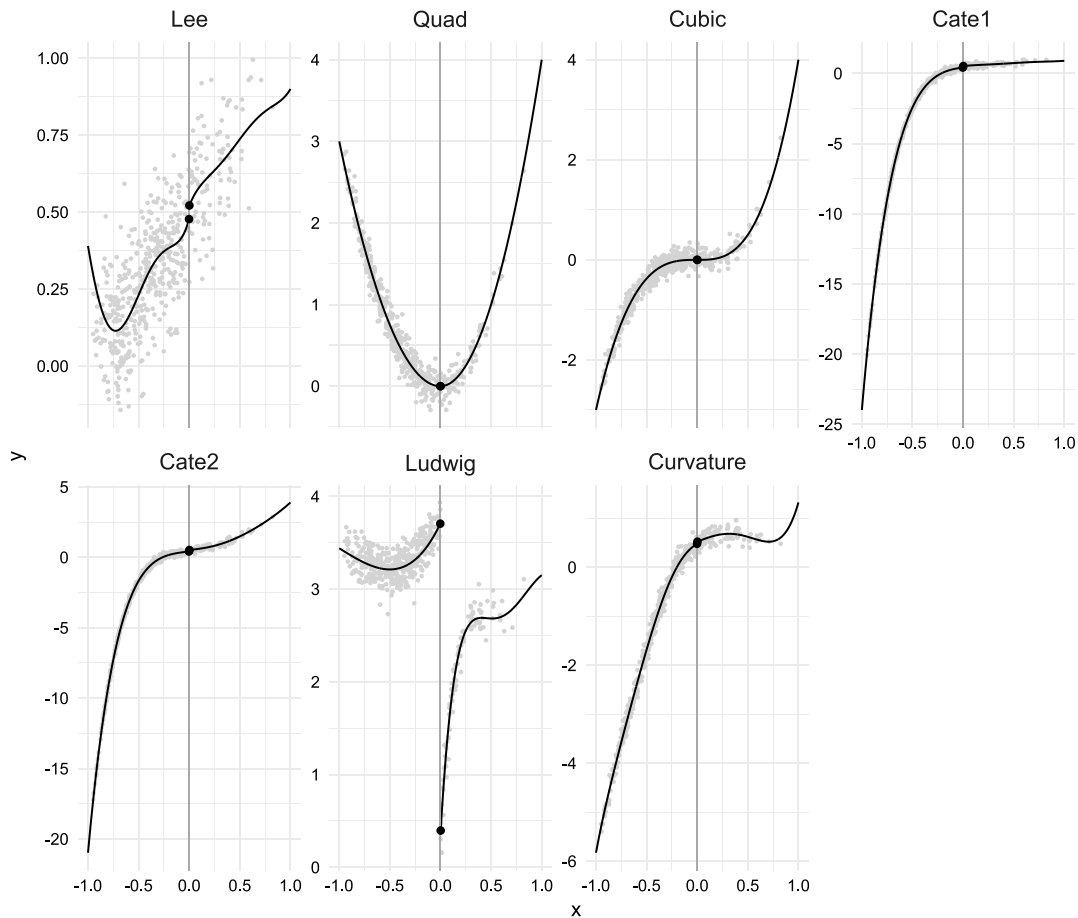


Fig. 1. Mean function for the datasets used in our simulation study. Lee, Quad, Cate 1, and Cate 2 were used in [Imbens and Kalyanaraman \(2012\)](#), and Lee, Ludwig, and Curvature were used in [Calonico et al. \(2014\)](#). Displayed in gray are a set of sample points $\{(x_i, y_i), i = 1, \dots, 500\}$ for each mean function.

For standard LLR and robust LLR we used the `rdrobust` R package ([Calonico et al., 2017](#)). For both methods, we used an MSE-optimal bandwidth that is the default in `rdrobust`. Simulation results using the bandwidth introduced in [Imbens and Kalyanaraman \(2012\)](#) – also known as the IK bandwidth, which has been widely used in practice – are provided in the [Appendix](#) in [Table 4](#), and results using the coverage error rate optimal bandwidth – an alternative bandwidth choice within the `rdrobust` R package that is also discussed in [Calonico et al. \(2018a\)](#) – are provided in the [Appendix](#) in [Table 5](#). The results using those bandwidths are largely the same as the results presented here. Simulation results for other bandwidth choices appear in [Calonico et al. \(2014\)](#).

[Imbens and Kalyanaraman \(2012\)](#) ran a simulation study that focused on the Lee, Quad, Cate 1, and Cate 2 mean functions, and they compared different bandwidth selectors for LLR in terms of bias and root mean squared error (RMSE). Similarly, [Calonico et al. \(2014\)](#) ran a simulation study that focused on the Lee, Ludwig, and Curvature mean functions, and they compared different bandwidth selectors for LLR and their methodology in terms of coverage and mean interval length (IL). We synthesize these simulation studies and report in [Fig. 2](#) how LLR, robust LLR, and two versions of our GPR methodology perform on the seven mean functions in [Fig. 1](#) in terms of coverage, IL, absolute bias, and RMSE. The numbers plotted in [Fig. 2](#) are in [Tables 2 and 3](#) in the [Appendix](#). Point estimates and 95% confidence intervals for LLR and robust LLR were obtained from `rdrobust`. Point estimates and 95% credible intervals for our methodology corresponded to the mean and 2.5% and 97.5% quantiles, respectively, of the posterior of the average treatment effect, shown in [\(17\)](#).

Robust LLR is meant to improve the coverage of LLR, and indeed it does for all datasets. The better coverage is in part due to wider confidence intervals (see the systematically higher mean interval length at the top right of [Fig. 2](#)) and in part due to better bias properties (see the bottom left of [Fig. 2](#)). Robust LLR also tends to exhibit worse RMSE than LLR (see the bottom right of [Fig. 2](#)).

Our primary method (“GPR”) tends to exhibit narrower intervals than both LLR and robust LLR. GPR also exhibits better coverage than both methods, except for the Lee and Ludwig datasets. Furthermore, our method tends to exhibit lower RMSE than both LLR and robust LLR. However, our method always exhibits more bias than robust LLR, which explicitly uses a bias correction.

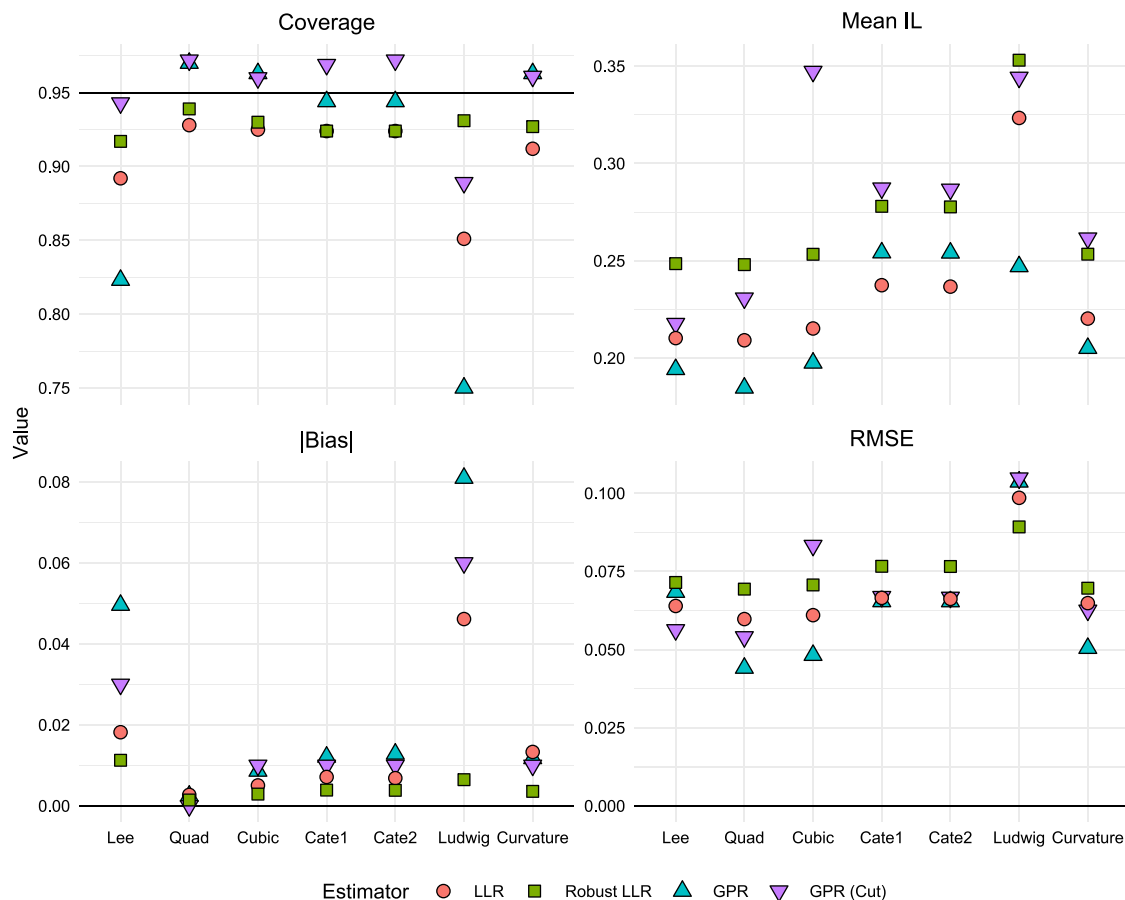


Fig. 2. The coverage, mean interval length (IL), absolute bias, and root mean squared error (RMSE) for LLR, robust LLR, and our GPR method.

For the Lee and Ludwig datasets, our method did worse than robust LLR in terms of coverage and bias, and this may be because it is inappropriate to assume that $\mu_T(x)$ and $\mu_C(x)$ are stationary processes – i.e., that the covariance parameters do not vary across the running variable X – in these cases. By assuming stationarity, our GPR model uses data both close to and far from the boundary to estimate the single variance σ_{GP}^2 and lengthscale ℓ . This assumption is related to the stability of the second derivative of $\mu_T(x)$ and $\mu_C(x)$, because the covariance parameters of a Gaussian process dictate their derivative processes; see Wang (2012) for a further discussion of this relationship. Fig. 3 displays the absolute second derivative of the seven mean functions (in blue). The Lee and Ludwig datasets are characterized by the absolute second derivative rapidly increasing near the boundary. Our GPR methodology likely does not do well for these datasets because we are using data far from the boundary to estimate the overall curvature of the mean function, which leads us to underestimating the curvature of the Lee and Ludwig mean functions at the boundary.

Fig. 3 also displays $\frac{\hat{\sigma}_{GP}}{\hat{\ell}}$, the ratio of the maximum-likelihood estimates of the covariance parameters, which was computed within a sliding window of a noiseless version of the seven mean functions. Although there is not a one-to-one correspondence between the absolute second derivative and $\frac{\hat{\sigma}_{GP}}{\hat{\ell}}$, their behavior is notably similar, which reinforces the idea that both the variance σ_{GP}^2 and lengthscale ℓ play a role in capturing the curvature of the mean function. Furthermore, this connection between the second derivative and the covariance parameters further suggests a similarity between the covariance parameters in our GPR methodology and the bandwidth in the LLR methodology, because the IK bandwidth is estimated as a nonlinear function of the estimated second derivative at the boundary b (Imbens and Kalyanaraman, 2012).

If one did not believe the Stationary Assumption held, one alternative would be to only use data close to the boundary when fitting our GPR model. This is our second method, “GPR (Cut)”, whose coverage, mean IL, absolute bias, and RMSE is also displayed in Fig. 2. For each of the 1000 replications of the seven mean functions, we first estimated the IK bandwidth with a rectangular kernel; then, we fit our GPR model within this estimated bandwidth. This procedure improves upon our GPR model for and only for the Lee and Ludwig datasets – the coverage increased to 94.3% and 88.9%, respectively, and the bias improved for the Lee and Ludwig datasets while staying the same for the other datasets. While the coverage also improved for the Cate 1 and Cate 2 datasets, this is likely due to the increase in the interval length. Because results improved only for

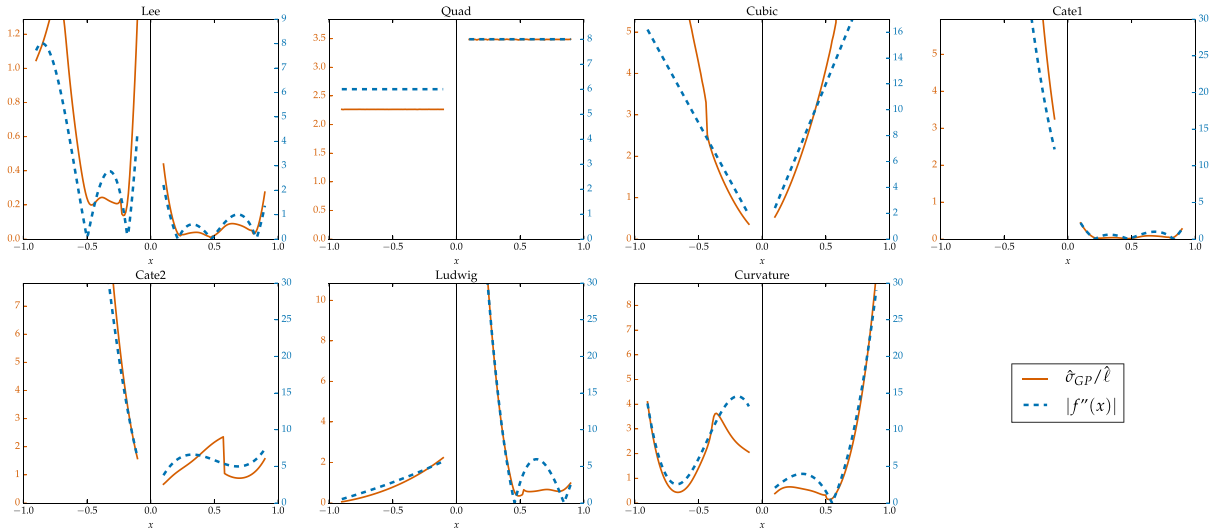


Fig. 3. The absolute second derivative (blue dotted line) of the seven mean functions shown in Fig. 1, and the ratio of the maximum-likelihood estimates of the variance and lengthscale (orange solid line), computed within a sliding window of a noiseless version of the seven mean functions. The sliding window was the range $[x - 0.1, x + 1]$ for $x \in [-0.9, -0.1]$ in the control group (left-hand side) and $x \in [0.1, 0.9]$ in the treatment group (right-hand side).

these two datasets, this further demonstrates that using our GPR model on the whole dataset can be beneficial when $\mu_T(x)$ and $\mu_C(x)$ are stationary processes; otherwise, it may be preferable to only fit our GPR model to data close to the boundary.

Furthermore, this suggests that our method can be combined with a local randomization perspective for RDDs (e.g., Li et al., 2015): One can first determine the window around the boundary where units are “as-if randomized” by using covariate balance tests such as Cattaneo et al. (2015) and Li et al. (2015) and then use our GPR methodology within this window around the boundary. This approach is similar to Cattaneo et al. (2017), who combined the local randomization perspective with parametric models for estimating the average treatment effect at the boundary.

Overall, GPR performs well compared to LLR and robust LLR. In particular, our GPR method tends to yield better interval length and RMSE properties than LLR and robust LLR, and it also yields better coverage when the underlying mean functions are stationary across the running variable X . The issue of undercoverage in LLR methodologies has been relatively unaddressed in the RDD literature, except for robust LLR (Calonico et al., 2014), and so our GPR method can be viewed as the second method to yield promising coverage properties for RDD analyses while also providing a flexible fit to the underlying mean functions $\mu_T(x)$ and $\mu_C(x)$. When $\mu_T(x)$ and $\mu_C(x)$ are nonstationary across X , our GPR methodology could be extended to incorporate a lengthscale function $\ell(x)$ instead of a single lengthscale ℓ . However, such a lengthscale function would either need to be specified (see Rasmussen and Williams 2006, Chapter 4, for an example), or estimated via another Gaussian process prior (Plagemann et al., 2008). In a similar vein, there has also been work on using dimension expansion to model nonstationary processes (Bornn et al., 2012). However, in an RDD, we only need a good estimate of the covariance parameters near the boundary, rather than across the entire mean functions $\mu_T(x)$ and $\mu_C(x)$. More work needs to be done to determine the optimal amount of data to include in our GPR model for estimating these parameters and the average treatment effect at the boundary in the case of nonstationary processes.

Now we compare how LLR, robust LLR, and GPR perform on a real dataset from the National Basketball Association.

5. Empirical example: The NBA draft

The National Basketball Association (NBA) draft, held annually, is divided into 2 rounds, where each NBA teams gets one selection per round to draft a player of their choice. Because players are picked sequentially, there is no reason to believe there is a marked skill difference between the last pick of the first round and first pick of the second round. However, because of the difference in the perceived value of first-round versus second-round picks, as well as differing contract structures between the two rounds, we suspect that first-round picks are treated more favorably and given more playing time than their second-round colleagues, above and beyond what can be explained by differences in skill. As such, we seek to explore if there is a difference between first- and second-round picks in both skill and playing time.

We want to estimate the treatment effect of having a second-round contract instead of a first-round contract on four basketball player outcomes: box plus-minus, win shares played, number of minutes played, and number of games played. The first two are overall measures of player performance (Kubatko, 2009; Myers, 2015), while the latter two are measures of playing time. Our data include the pick number and the four aforementioned outcomes for 1238 NBA basketball players drafted between 1995 and 2016. Due to anomalies created by the NBA expanding from 29 teams to 30 teams in 2004, as well as some years with teams forfeiting picks, we shifted the pick numbers to ensure that $b = 30.5$ marked the

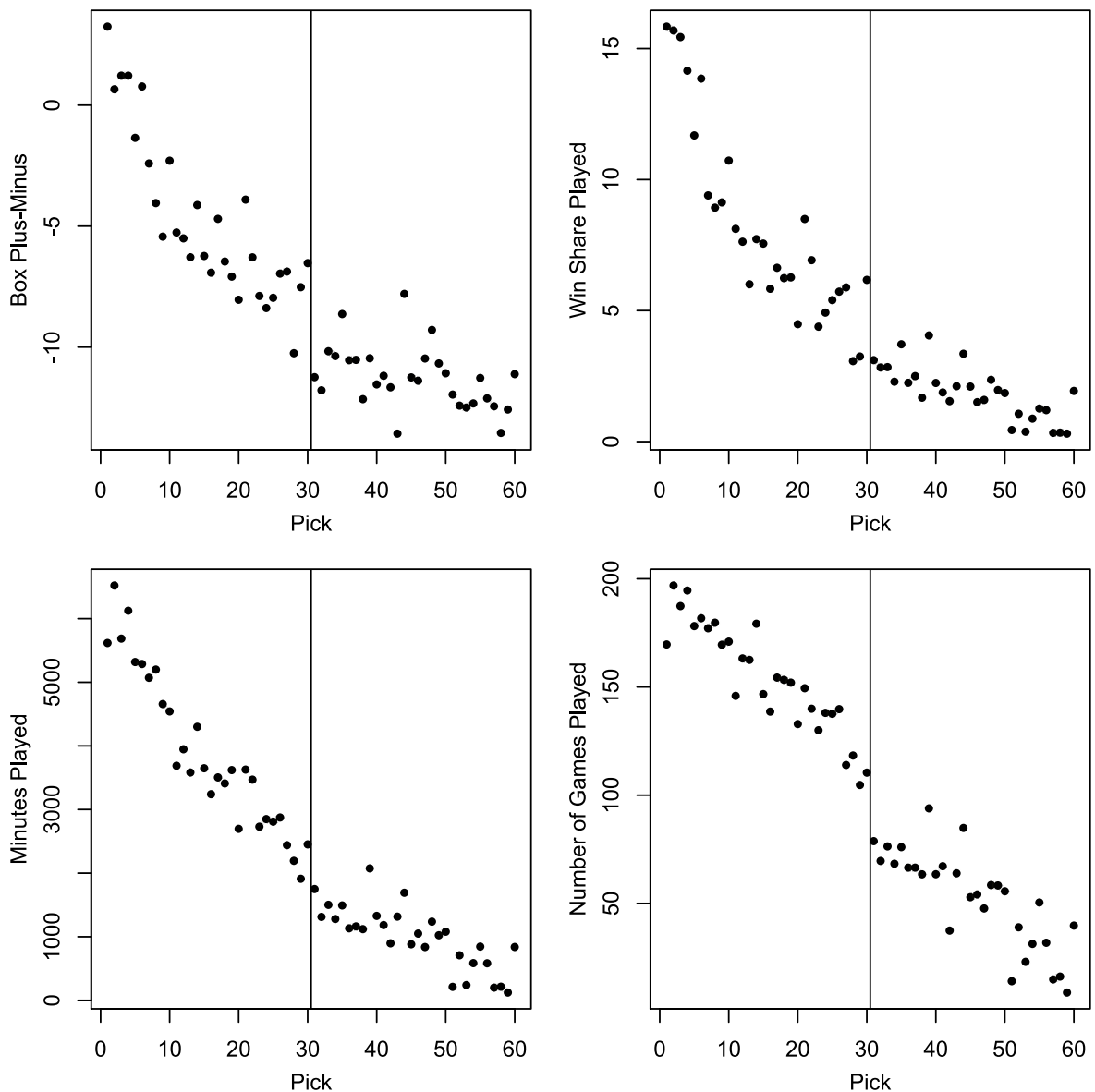


Fig. 4. Four basketball-player outcomes – number of minutes played, box plus-minus, win shares played, and number of games played – across picks.

discontinuity between the first- and second-round picks in each year. [Fig. 4](#) displays the NBA player data grouped by pick number for each of the four outcomes (after aligning pick numbers). Grouping by pick number allows us to understand the average performance and playing time of players drafted at each pick number, which is a standard approach in draft evaluation ([Silver, 2014](#)).

Even though the running variable in this case is discrete – which can cause complications in some regression discontinuity analyses ([Lee and Card, 2008](#); [Kolesár and Rothe, 2018](#)) – we nonetheless apply LLR, robust LLR, and our GPR method to these data to compare how they perform in practice. As in [Section 4](#), we used the `rdrobust` R package to implement LLR and robust LLR using the default MSE-optimal bandwidth. For GPR, we used the squared exponential covariance function assuming the Same Covariance Assumption; furthermore, we took the full-Bayesian approach to our GPR methodology and – similar to [Section 4](#) – placed independent $\mathcal{N}(0, 100^2)$ priors on the mean function parameters and half-Cauchy(0, 5) priors on the covariance parameters.

Table 1

Treatment effect estimation for LLR, robust LLR, and GPR on NBA data.

Outcome	LLR		Robust LLR		GPR	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Box plus-minus	−3.06	[−5.20, −0.92]	−3.37	[−5.86, −0.88]	−2.42	[−5.02, 0.01]
Win shares	−1.58	[−4.09, 0.93]	−1.73	[−4.76, 1.29]	−1.08	[−3.04, 0.88]
Minutes played	−446.29	[−1042.81, 150.23]	−406.63	[−1123.99, 310.72]	−640.75	[−1259.65, 5.51]
Games played	−29.71	[−40.20, −19.22]	−28.16	[−40.81, −15.51]	−32.00	[−45.69, −18.14]

Point estimates and 95% confidence intervals for the treatment effect on each of the four outcomes: box plus-minus, win shares played, number of minutes played, and number of games played. Statistically significant point estimates are in bold.

Fig. 5 shows the estimated mean functions and corresponding confidence intervals for LLR and GPR, and Table 1 shows the treatment effect point estimates and corresponding confidence intervals for LLR, robust LLR, and GPR.² The estimated mean functions for LLR and GPR are quite similar to each other for all outcomes, with GPR yielding slightly wider confidence intervals, as expected. All three methods find the number of games played for second-round picks to be significantly lower than that of first-round picks. Furthermore, LLR and robust LLR find the box plus-minus for second-round picks to be significantly lower than that of first-round picks; meanwhile, GPR finds this difference to be borderline insignificant. Out of these three methods, the results from our GPR method are most in line with previous reports – both qualitative (Barber, 2016) and quantitative (Koenig, 2012) – on the difference between first- and second-round NBA basketball players, which have claimed that there is a difference in attention given to first-round picks (e.g., games played) but not a difference in player ability (e.g., box plus-minus and win shares).

In summary, using our GPR methodology, we find that the treatment effect of being a second-round pick significantly reduces the number of games played and marginally reduces the number of minutes played. This suggests that there is a drop-off in playing time for second-round players relative to their first-round counterparts beyond that explainable by the natural drop-off in playing time between successive picks. Furthermore, we find that there is not a significant difference in player ability between first- and second-round basketball players at the boundary between the 30th and 31st picks that divides the first and second rounds of the NBA draft.

6. Discussion and conclusion

Local linear regression (LLR) and its variants are currently the most common methodologies for estimating the average treatment effect at the boundary in RDDs. These methods are popular because they are easy to implement and there is a large literature on their theoretical properties. However, recent works have noted that LLR tends to yield confidence intervals that undercover, and new methodologies – namely that of Calonico et al. (2014) – have tried to address this issue. As an alternative to LLR, we proposed a Gaussian Process regression (GPR) methodology that flexibly fits the treatment and control response functions by placing a general prior on the mean response functions. We showed via simulation that our GPR methodology tends to outperform standard LLR and the state-of-the-art methodology of Calonico et al. (2014) in terms of coverage, interval length, and RMSE. Furthermore, we used our GPR methodology on a real-world sharp RDD in the National Basketball Association (NBA) draft and found that GPR yielded results that were more in line with previous reports on the NBA draft than were results from LLR methods. Overall, our methodology addresses the undercoverage issue commonly seen in RDDs without sacrificing too much power to detect treatment effects, thereby adding to the growing literatures on improving inference for RDDs (Calonico et al., 2014, 2018b) and on Bayesian methods for RDDs (Chib and Greenberg, 2014; Chib and Jacobi, 2015; Geneletti et al., 2015; Li et al., 2015).

Our methodology focuses on flexibly fitting the mean treatment and control responses while also improving coverage properties; however, there are many other issues of interest in the RDD literature. For example, we only consider sharp RDDs; Li et al. (2015) and Chib and Jacobi (2015) provide Bayesian methodologies for fuzzy RDDs which could likely be combined with our Gaussian process approach. Furthermore, we focused on RDDs that only have one background covariate – the running variable – but other covariates could be included in our GPR methodology to improve the precision of our treatment effect estimator (e.g., Calonico et al., 2016). See Imbens and Lemieux (2008) for a review of these other RDD concerns.

Furthermore, while our method can incorporate prior knowledge and uncertainty in various parameters that are typically discussed in the RDD literature, it is not necessarily clear when this should be done for our GPR model. For example, Hall and Kang (2001) found that incorporating the uncertainty in the bandwidth – which is somewhat analogous to the covariance parameters in our GPR model – for density estimators via bootstrapping can be inconsequential or even detrimental in some cases. Although our simulation study suggests that it is beneficial to take a full-Bayesian approach and propagate uncertainty

² Robust LLR yields an inflated confidence interval for the treatment effect specifically, which depends on a bias correction at the boundary. Thus, robust LLR does not yield confidence intervals for the entire mean functions, because the bias correction is boundary-specific. Thus, only LLR and GPR are displayed in Fig. 5, while LLR, robust LLR, and GPR are all discussed in Table 1. Furthermore, for robust LLR in Table 1, we report the bias-corrected point estimate given by the `rdrobust` package.

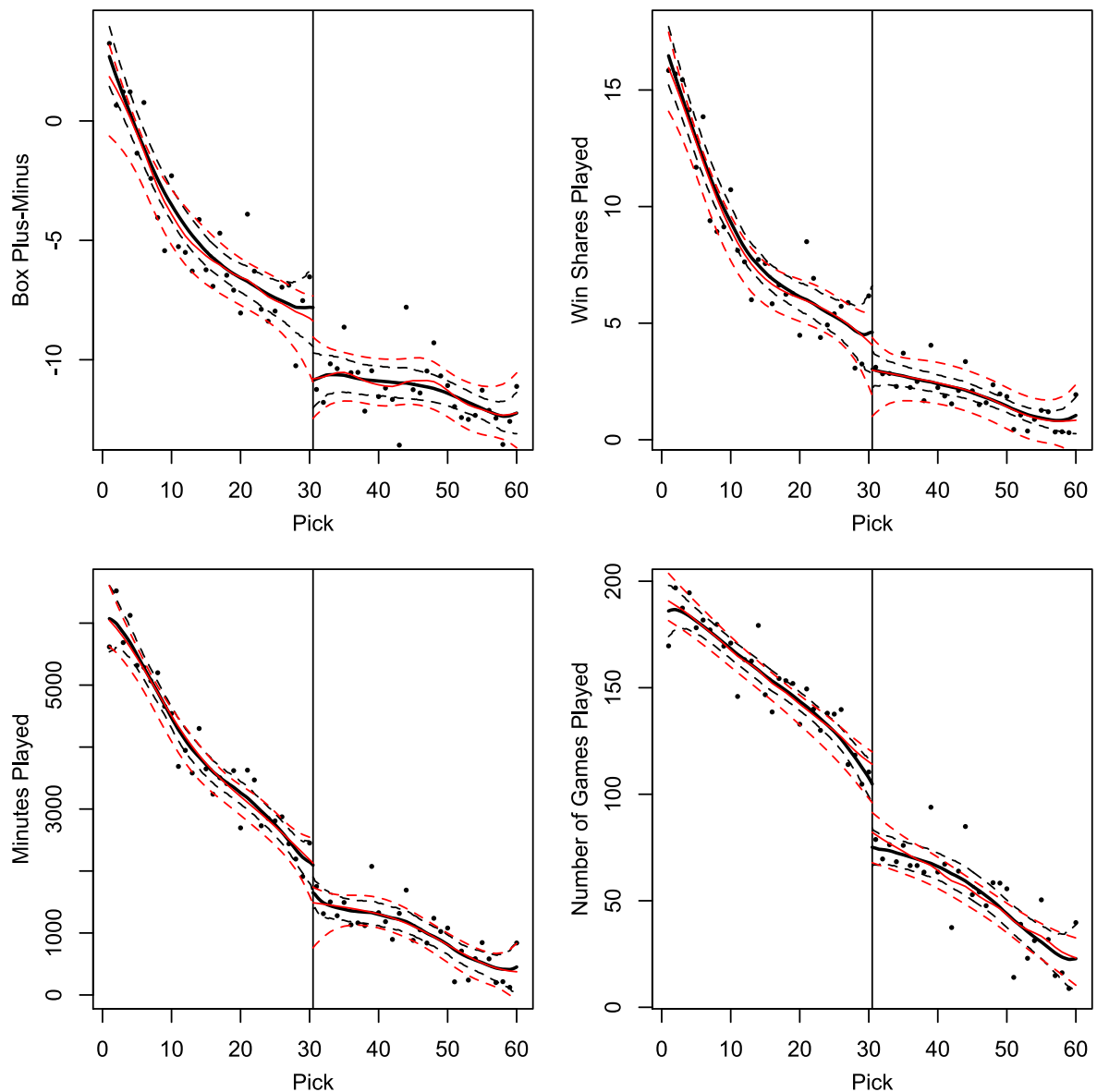


Fig. 5. The estimated mean functions (solid lines) and corresponding confidence intervals (dashed lines) for LLR (black lines) and GPR (red lines). The lines for LLR were produced by the `rdd` R package (Dimmery, 2013), but using the bandwidth estimated by the `rddrobust` R package. The two treatment groups are the first round of picks (picks 30 and below) and second round of picks (picks 31 and above). We set the boundary to be $b = 30.5$ to minimize the amount of extrapolation that needs to be conducted on both sides of the boundary to estimate the treatment effect. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in the GPR model parameters when estimating treatment effects in RDDs, more work needs to be done to determine when it is most appropriate to incorporate uncertainty in these parameters. Furthermore, we only focused on the squared exponential covariance function for Gaussian processes, but other covariance functions should be considered. Because estimating the average treatment effect at the boundary in an RDD is fundamentally an extrapolation issue, covariance functions whose purpose is to extrapolate well (e.g., Wilson and Adams, 2013) may be particularly suitable for RDDs.

Additionally, although our GPR methodology exhibits arguably better coverage, interval length, and RMSE properties than standard methodologies in the literature, there are ways our methodology could be improved, even for the case of a sharp RDD with only one covariate. Our methodology does not perform well when the mean response functions $\mu_T(x)$ and $\mu_C(x)$ are nonstationary across the running variable X . More work needs to be done to model nonstationary processes within the context of RDDs, such as by estimating a lengthscale function $\ell(x)$ or by determining the optimal amount of data to use when estimating the covariance parameters for our GPR methodology. Relatedly, our simulation results suggest that our

GPR methodology can be used in combination with a local randomization framework for RDDs (such as that seen in [Li et al., 2015](#)).

Finally, a promising avenue for future research is extending our GPR methodology beyond one-dimensional sharp RDDs. In particular, recent work has explored geographic RDDs, where spatial variation in outcomes must be accounted for when estimating the treatment effect along a geographic boundary ([Keele and Titiunik, 2014](#); [Keele et al., 2015](#)). Because GPR has been widely used in spatial statistics ([Banerjee et al., 2014](#); [Cressie, 2015](#)), it may be particularly suitable for geographic RDDs. We explore the use of our GPR methodology for geographic RDDs in [Rischard et al. \(2018\)](#).

Appendix

A.1. Complete simulation results from Section 4

See [Tables 2–5](#).

A.2. Assumptions for posterior consistency proofs

Assumptions on the Running Variable X (Assumptions A1)

1. Assume the control running variable values $\{x_i\}_{i=1}^{n_C}$ are known elements of $[b_C, b]$ and the treatment running variable values $\{x_i\}_{i=1}^{n_T}$ are known elements of $[b, b_T]$, for some b_C, b_T , and boundary b .

Assumptions on the Response Function (Assumptions A2)

1. Let $C^\alpha[b, b_T]$ and $C^\alpha[b_C, b]$ be Hölder spaces of α -smooth functions $f : [b, b_T] \rightarrow \mathbb{R}$ and $f : [b_C, b] \rightarrow \mathbb{R}$, respectively. Assume $\mu_T(x) \in C^\alpha[b, b_T]$ and $\mu_C(x) \in C^\alpha[b_C, b]$, where $\mu_T(x)$ and $\mu_C(x)$ are the treatment and control response functions.
2. The treatment and control responses $\{y_i\}_{i=1}^{n_T}$ and $\{y_i\}_{i=1}^{n_C}$ have the following relationships with the running variable:

$$y_i = \mu_T(x_i) + \epsilon_i, \quad i = 1, \dots, n_T \quad (22)$$

$$y_i = \mu_C(x_i) + \epsilon_i, \quad i = 1, \dots, n_C \quad (23)$$

for mean response functions $\mu_T(x)$ and $\mu_C(x)$ and independent errors $\{\epsilon_i\}_{i=1}^{n_C} \sim N(0, \sigma_{y0}^2)$ and $\{\epsilon_i\}_{i=1}^{n_T} \sim N(0, \sigma_{y1}^2)$.

Assumptions on the Noise (Assumptions A3)

1. The priors on σ_{y0} and σ_{y1} have support on compact intervals that are subsets of $(0, \infty)$ which contain the true errors σ_{y0} and σ_{y1} , respectively.

Assumptions on the Prior for ℓ (Assumptions A4)

1. The lengthscale ℓ has a prior distribution κ such that, for positive constants C_1, D_1, C_2, D_2 , nonnegative constants p, q , and every sufficiently large $a > 0$,

$$C_1 a^p \exp(-D_1 a \log^q a) \leq \kappa(a) \leq C_2 a^p \exp(-D_2 a \log^q a) \quad (24)$$

Table 2
Simulation for $n = 500$, shown in [Fig. 2](#).

Dataset	Method				Robust LLR				GPR			
	EC	$\overline{\text{IL}}$	Bias	RMSE	EC	$\overline{\text{IL}}$	Bias	RMSE	EC	$\overline{\text{IL}}$	Bias	RMSE
Lee	89.2%	0.21	0.02	0.06	91.7%	0.25	0.01	0.07	82.3%	0.19	0.05	0.07
Quad	92.8%	0.21	0.00	0.06	93.9%	0.25	0.00	0.07	97.0%	0.18	0.00	0.04
Cubic	92.5%	0.22	−0.01	0.06	93.0%	0.25	0.00	0.07	96.3%	0.20	−0.01	0.05
Cate 1	92.4%	0.24	−0.01	0.07	92.4%	0.28	0.00	0.08	94.4%	0.25	−0.01	0.07
Cate 2	92.4%	0.24	−0.01	0.07	92.4%	0.28	0.00	0.08	94.4%	0.25	−0.01	0.07
Ludwig	85.1%	0.32	0.05	0.10	93.1%	0.35	0.01	0.09	75.0%	0.25	0.08	0.10
Curvature	91.2%	0.22	−0.01	0.06	92.7%	0.25	0.00	0.07	96.3%	0.21	−0.01	0.05

Simulations assessing the empirical coverage (EC), mean interval length ($\overline{\text{IL}}$), bias, and root mean squared error (RMSE) for local linear regression (LLR), robust LLR, and our Gaussian Process Regression (GPR) method, where we used the MSE-optimal default bandwidth in `rdrobust` when implementing LLR and robust LLR. These methods were performed on 1000 replications of seven different datasets, which were also used in [Imbens and Kalyanaraman \(2012\)](#) and [Calonico et al. \(2014\)](#). A plot of these numbers is shown in [Fig. 2](#).

Table 3

Simulation results for GPR, GPR (Cut), GPR (Zero Mean), and GPR (MLE).

Data	Method				GPR (Cut)				GPR (Zero Mean)				GPR (MLE)			
	GPR				GPR (Cut)				GPR (Zero Mean)				GPR (MLE)			
	EC	\bar{IL}	Bias	RMSE	EC	\bar{IL}	Bias	RMSE	EC	\bar{IL}	Bias	RMSE	EC	\bar{IL}	Bias	RMSE
Lee	82.3%	0.19	0.05	0.07	94.3%	0.22	0.03	0.06	81.6%	0.18	0.05	0.07	76.4%	0.15	0.04	0.06
Quad	97.0%	0.18	0.00	0.04	97.2%	0.23	−0.00	0.05	97.0%	0.19	0.01	0.04	95.9%	0.18	0.00	0.05
Cubic	96.3%	0.20	−0.01	0.05	96.0%	0.35	−0.01	0.08	96.7%	0.20	0.00	0.05	91.1%	0.19	−0.03	0.06
Cate 1	94.4%	0.25	−0.01	0.07	96.9%	0.29	−0.01	0.07	95.2%	0.26	0.00	0.07	93.8%	0.25	0.00	0.07
Cate 2	94.4%	0.25	−0.01	0.07	97.2%	0.29	−0.01	0.07	95.4%	0.26	0.00	0.07	94.2%	0.25	0.00	0.07
Ludwig	75.0%	0.25	0.08	0.10	88.9%	0.34	0.06	0.10	64.5%	0.24	0.10	0.12	52.0%	0.24	0.11	0.13
Curvature	96.3%	0.21	−0.01	0.05	96.1%	0.26	−0.01	0.06	96.9%	0.20	0.00	0.05	89.6%	0.20	−0.03	0.06

Simulation results assessing the empirical coverage (EC), mean interval length (\bar{IL}), bias, and root mean squared error (RMSE) for GPR (which uses the full Bayesian approach discussed in Section 2.3), GPR using only data within the IK bandwidth with a rectangular kernel (called GPR (Cut)), GPR using a zero mean function in the prior (19) instead of a linear trend (called GPR (Zero Mean)), and GPR plugging in the MLE for the covariance parameters (called GPR (MLE)). Note that the GPR columns are the same as those in Table 2. GPR (Cut) performs much better than GPR for the Lee and Ludwig datasets, and GPR (Cut) performs marginally better than GPR for the Cate 1 and Cate 2 datasets. Otherwise, GPR (Cut) is equal to or worse than GPR. This demonstrates that GPR on the whole dataset can be beneficial when $\mu_T(x)$ and $\mu_C(x)$ are stationary processes; otherwise, it may be preferable to only fit our GPR model to data close to the boundary. Furthermore, GPR (Zero Mean) performs similarly to GPR for most of the datasets; this suggests that our results are generally insensitive to specification of the mean function in the Gaussian process prior. However, GPR does perform better than GPR (Zero Mean) for the Ludwig dataset. This is likely because, as discussed in Section 4, when GPR extrapolates to the boundary, its predictions will be pulled towards the global linear trend, while predictions from GPR (Zero Mean) will be pulled towards the global mean. This also suggests why, for the Ludwig dataset, the bias for GPR (Zero Mean) is higher than the bias for GPR. Finally, GPR that simply plugs in the MLE of the covariance parameters tends to perform worse than the full-Bayesian GPR approach, especially in terms of coverage. This further suggests that incorporating uncertainty in the covariance parameters for our GPR method can lead to promising inferential properties.

Table 4Simulation for $n = 500$, using the IK bandwidth.

Dataset	Method				Robust LLR				GPR			
	LLR				Robust LLR				GPR			
	EC	\bar{IL}	Bias	RMSE	EC	\bar{IL}	Bias	RMSE	EC	\bar{IL}	Bias	RMSE
Lee	82.7%	0.15	0.04	0.05	91.2%	0.27	0.01	0.08	82.3%	0.19	0.05	0.07
Quad	94.6%	0.15	0.00	0.04	93.2%	0.24	0.00	0.07	97.0%	0.18	0.00	0.04
Cubic	93.7%	0.19	−0.01	0.05	93.8%	0.24	0.01	0.06	96.3%	0.20	−0.01	0.05
Cate 1	91.2%	0.21	−0.01	0.06	92.9%	0.26	0.01	0.07	94.4%	0.25	−0.01	0.07
Cate 2	92.1%	0.22	−0.01	0.06	92.8%	0.26	0.01	0.07	94.4%	0.25	−0.01	0.07
Ludwig	31.5%	0.23	0.15	0.16	89.8%	0.27	0.04	0.08	75.0%	0.25	0.08	0.10
Curvature	84.4%	0.19	−0.03	0.06	94.8%	0.23	0.00	0.06	96.3%	0.21	−0.01	0.05

Simulations assessing the empirical coverage (EC), mean interval length (\bar{IL}), bias, and root mean squared error (RMSE) for local linear regression (LLR), robust LLR, and our Gaussian Process Regression (GPR) method, where we used the bandwidth of [Imbens and Kalyanaraman \(2012\)](#) when implementing LLR and robust LLR. These results are largely the same as the results presented in Table 2: Our method performed best in terms of coverage, except for the Lee and Ludwig datasets. Furthermore, our method tended to yield narrower intervals and lower RMSE than robust LLR. Our method also tended to yield more bias than robust LLR.

Table 5Simulation for $n = 500$, using the CER bandwidth.

Dataset	Method				Robust LLR				GPR			
	LLR				Robust LLR				GPR			
	EC	\bar{IL}	Bias	RMSE	EC	\bar{IL}	Bias	RMSE	EC	\bar{IL}	Bias	RMSE
Lee	91.1%	0.25	0.01	0.07	91.8%	0.27	0.01	0.08	82.3%	0.19	0.05	0.07
Quad	92.3%	0.24	−0.00	0.07	92.7%	0.27	−0.00	0.08	97.0%	0.18	0.00	0.04
Cubic	92.2%	0.25	−0.00	0.07	92.1%	0.27	0.00	0.08	96.3%	0.20	−0.01	0.05
Cate 1	91.9%	0.28	−0.00	0.08	92.3%	0.30	0.00	0.08	94.4%	0.25	−0.01	0.07
Cate 2	92.0%	0.28	−0.00	0.08	92.3%	0.30	0.00	0.08	94.4%	0.25	−0.01	0.07
Ludwig	91.5%	0.39	0.02	0.10	92.9%	0.41	0.00	0.10	75.0%	0.25	0.08	0.10
Curvature	91.4%	0.26	−0.00	0.07	93.3%	0.28	0.00	0.08	96.3%	0.21	−0.01	0.05

Simulations assessing the empirical coverage (EC), mean interval length (\bar{IL}), bias, and root mean squared error (RMSE) for local linear regression (LLR), robust LLR, and our Gaussian Process Regression (GPR) method, where we used the coverage error rate (CER) optimal bandwidth – an alternative bandwidth choice within the `rdrobust` R package that is also discussed in [Calonico et al. \(2018a\)](#). These results are largely the same as the results presented in Tables 2 and 4: Our method performed best in terms of coverage, except for the Lee and Ludwig datasets. Furthermore, our method tended to yield narrower intervals and lower RMSE than robust LLR. Our method also tended to yield more bias than robust LLR.

A.3. Proof of Theorems 1 and 2

Proof of Theorem 1. Assume that the Stationary Assumption holds, the covariance functions $K_T(x, x)$ and $K_C(x, x)$ are fixed, and Assumptions A1, A2, and A3 hold. Then, according to Theorem 2.2 in van der Vaart and van Zanten (2009), for sufficiently large M_T and M_C ,

$$\prod (\mu_T : h(\mu_T, \mu_T^*) \geq M_T \epsilon_{n_T} | x_1, \dots, x_{n_T}) \xrightarrow{P_{\mu_T^*}} 0, \quad \text{and} \quad (25)$$

$$\prod (\mu_C : h(\mu_C, \mu_C^*) \geq M_C \epsilon_{n_C} | x_1, \dots, x_{n_C}) \xrightarrow{P_{\mu_C^*}} 0$$

for some contraction rates ϵ_{n_T} and ϵ_{n_C} , where h is the Hellinger distance. The nature of the contraction rates ϵ_{n_T} and ϵ_{n_C} are discussed in van der Vaart and van Zanten (2009), and depend on the differentiability and smoothness of the true $\mu_T^*(x)$ and $\mu_C^*(x)$. Note that the only covariate value for which both of these hold is $x = b$, because by Assumption A1, the intersection of the supports of $\{x_i\}_{i=1}^{n_C}$ and $\{x_i\}_{i=1}^{n_T}$ is only the boundary b .

Because the Hellinger distance is symmetric and satisfies the triangle inequality,

$$h(\tau, \tau^*) = h(\mu_T(b) - \mu_C(b), \mu_T^*(b) - \mu_C^*(b)) \quad (26)$$

$$\leq h(\mu_T(b), \mu_T^*(b)) + h(\mu_C(b), \mu_C^*(b)) \quad (27)$$

$$\leq M_T \epsilon_{n_T} + M_C \epsilon_{n_C} \quad (28)$$

where the last inequality holds with posterior probability 1, by (25). Therefore,

$$\prod (\tau : h(\tau, \tau^*) \geq M \epsilon_n | x_1, \dots, x_n) \xrightarrow{P_{\tau^*}} 0 \quad (29)$$

where $M \equiv \sqrt{2} \cdot \max(M_T, M_C)$ and $\epsilon_n \equiv \sqrt{2} \cdot \max(\epsilon_{n_T}, \epsilon_{n_C})$, so that $M \epsilon_n \geq M_T \epsilon_{n_T} + M_C \epsilon_{n_C}$. ■

Proof of Theorem 2. Assume that the Stationary Assumption holds, the σ_{GP}^2 parameters in $K_T(x, x)$ and $K_C(x, x)$ are fixed, and Assumptions A1, A2, A3, and A4 hold. By Theorem 3.1 of van der Vaart and van Zanten (2009), (25) holds, and then the proof of Theorem 2 is identical to that of Theorem 1. ■

A.4. Extending Theorems 1 and 2 to other mean and covariance functions and random variance

Rasmussen and Williams (2006) (Section 2.7) discuss other choices of mean functions besides $m(x) = 0$, particularly mean functions of the form $m(x) = \mathbf{h}(x)^T \boldsymbol{\beta}$ for some fixed basis functions $\mathbf{h}(x)$. However, Rasmussen and Williams (2006) argue that different choices of $m(x)$ are more for interpretability than predictive accuracy, because $m(x) = 0$ does not constrain the posterior mean to be zero, and so different choices of $m(x)$ likely do not affect the consistency results in Section 3. To the best of our knowledge, the literature has focused primarily on the choice $m(x) = 0$ for posterior consistency of GPR (e.g., Choi and Schervish, 2007 and van der Vaart and van Zanten, 2009).

van der Vaart and van Zanten (2009) discuss how their Theorems 2.2 and 3.1 (which correspond to our Theorems 1 and 2, respectively) extend to covariance functions besides the squared exponential. Specifically, their results hold for processes whose spectral measure has subexponential tails; the squared exponential process falls under this class, but van der Vaart and van Zanten (2009) discuss other processes that fall under this class as well, and our Theorems 1 and 2 also hold for those processes.

Finally, the literature has focused on posterior consistency for the case when a prior is placed on the lengthscale ℓ but not on the variance σ_{GP}^2 , as in van der Vaart and van Zanten (2009). To the best of our knowledge, Choi (2007) is the only work to consider posterior consistency when priors are placed on both ℓ and σ_{GP}^2 ; however, these results only hold for binary Gaussian process regression, and their necessary assumptions are more restrictive than those presented in this paper. We leave posterior consistency of our GPR method when priors are placed on both ℓ and σ_{GP}^2 as future work.

References

- Armstrong, T.B., Kolesár, M., 2017. A simple adjustment for bandwidth snooping. *Rev. Econom. Stud.* 85 (2), 732–765.
- Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. *Hierarchical Modeling and Analysis for Spatial Data*. Crc Press.
- Barber, H., 2016. Where are all the second round picks? nobody knows and the nba doesn't seem to care. https://www.huffingtonpost.com/houston-barber/where-are-all-the-second-_b_9722192.html. Accessed: 2017-11-21.
- Bornn, L., Shaddick, G., Zidek, J.V., 2012. Modeling nonstationary processes through dimension expansion. *J. Amer. Statist. Assoc.* 107 (497), 281–289.
- Calonico, S., Cattaneo, M.D., Farrell, M.H., 2018a. On the effect of bias estimation on coverage accuracy in nonparametric inference. *J. Amer. Statist. Assoc.* 1–13.
- Calonico, S., Cattaneo, M.D., Farrell, M.H., 2018b. Optimal bandwidth choice for robust bias corrected inference in regression discontinuity designs. *arXiv preprint arXiv:1809.00236*.
- Calonico, S., Cattaneo, M.D., Farrell, M.H., Titiunik, R., 2016. Regression discontinuity designs using covariates. *Rev. Econ. Stat.* (0).
- Calonico, S., Cattaneo, M.D., Farrell, M.H., Titiunik, R., 2017. *rdrobust: Software for regression discontinuity designs*. *Stata J.* 17 (2), 372–404.

- Calonico, S., Cattaneo, M.D., Titiunik, R., 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82 (6), 2295–2326.
- Calonico, S., Cattaneo, M.D., Titiunik, R., 2015. Optimal data-driven regression discontinuity plots. *J. Amer. Statist. Assoc.* 110 (512), 1753–1769.
- Card, D., Dobkin, C., Maestas, N., 2004. The impact of nearly universal insurance coverage on health care utilization and health: evidence from medicare. Tech. rep., National Bureau of Economic Research.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A., 2016. Stan: A probabilistic programming language. *J. Stat. Softw.* 20.
- Cattaneo, M.D., Frandsen, B.R., Titiunik, R., 2015. Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *J. Causal Inference* 3 (1), 1–24.
- Cattaneo, M.D., Titiunik, R., Vazquez-Bare, G., 2017. Comparing inference approaches for rd designs: A reexamination of the effect of head start on child mortality. *J. Policy Anal. Manage.*
- Chib, S., Greenberg, E., 2014. Nonparametric bayes analysis of the sharp and fuzzy regression discontinuity designs. Technical Report, Washington University in St. Louis.
- Chib, S., Jacobi, L., 2015. Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling. *J. Appl. Econometrics*.
- Choi, T., 2007. Alternative posterior consistency results in nonparametric binary regression using gaussian process priors. *J. Statist. Plann. Inference* 137 (9), 2975–2983.
- Choi, T., Schervish, M.J., 2007. On posterior consistency in nonparametric regression problems. *J. Multivariate Anal.* 98 (10), 1969–1987.
- Cressie, N., 2015. *Statistics for Spatial Data*. John Wiley & Sons.
- Dimmery, D., 2013. rdd: Regression discontinuity estimation. R package version 0.54.
- Fan, J., 1992. Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* 87 (420), 998–1004.
- Fan, J., Gijbels, I., 1992. Variable bandwidth and local linear regression smoothers. *Ann. Statist.* 2008–2036.
- Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* 1 (3), 515–534.
- Gelman, A., Imbens, G., 2018. Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econom. Statist.* 1–10.
- Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. *Bayesian Data Analysis*. Chapman and Hall/CRC.
- Geneletti, S., O’Keeffe, A.G., Sharples, L.D., Richardson, S., Baio, G., 2015. Bayesian regression discontinuity designs: Incorporating clinical knowledge in the causal analysis of primary care data. *Stat. Med.* 34 (15), 2334–2352.
- Ghosal, S., Roy, A., 2006. Posterior consistency of gaussian process prior for nonparametric binary regression. *Ann. Statist.* 2413–2429.
- Hahn, J., Todd, P., Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69 (1), 201–209.
- Hall, P., Kang, K.-H., 2001. Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *Ann. Statist.* 1443–1468.
- Imbens, G., Kalyanaraman, K., 2012. Optimal bandwidth choice for the regression discontinuity estimator. *Rev. Econom. Stud.* 79 (3), 933–959.
- Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. *J. Econometrics* 142 (2), 615–635.
- Keele, L.J., Titiunik, R., 2014. Geographic boundaries as regression discontinuities. *Polit. Anal.* 23 (1), 127–155.
- Keele, L., Titiunik, R., Zubizarreta, J.R., 2015. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. Roy. Stat. Soc. Ser. A (Stat. Soc.)* 178 (1), 223–239.
- Van der Klaauw, W., 2002. Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *Internat. Econom. Rev.* 43 (4), 1249–1287.
- Koenig, A., 2012. Coming up just short: The marginal effect of being a first round pick in the nba draft.
- Kolesár, M., Rothe, C., 2018. Inference in regression discontinuity designs with a discrete running variable. *Amer. Econ. Rev.* 108 (8), 2277–2304.
- Kubatko, J., 2009. Calculating win shares. *Basketball-Reference. com*.
- Lee, D.S., Card, D., 2008. Regression discontinuity inference with specification error. *J. Econometrics* 142 (2), 655–674.
- Li, F., Mattei, A., Mealli, F., et al., 2015. Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *Ann. Appl. Stat.* 9 (4), 1906–1931.
- Ludwig, J., Miller, D.L., 2007. Does head start improve children’s life chances? evidence from a regression discontinuity design. *Q. J. Econ.* 122 (1), 159–208.
- Matsudaira, J.D., 2008. Mandatory summer school and student achievement. *J. Econometrics* 142 (2), 829–850.
- Myers, D., 2015. About box plus/minus (bpm). *Basketball-Reference. com*, 15.
- Plagemann, C., Kersting, K., Burgard, W., 2008. Nonstationary gaussian process regression using point estimates of local smoothness. *Mach. Learn. Knowl. Discov. Databases* 204–219.
- Polson, N.G., Scott, J.G., 2012. On the half-cauchy prior for a global scale parameter. *Bayesian Anal.* 7 (4), 887–902.
- Porter, J., 2003. Estimation in the Regression Discontinuity Model. Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison.
- Rasmussen, C.E., Williams, C.K.I., 2006. *Gaussian Processes for Machine Learning*. MIT Press.
- Rischard, M., Branson, Z., Miratrix, L., Bornn, L., 2018. A bayesian nonparametric approach to geographic regression discontinuity designs: Do school districts affect nyc house prices? arXiv preprint arXiv:1807.04516.
- Rubin, D.B., 1977. Assignment to treatment group on the basis of a covariate. *J. Educ. Stat.* 2 (1), 1–26.
- Silver, N., 2014. How much is winning the nba draft lottery really worth? <https://fivethirtyeight.com/features/how-much-is-winning-the-nba-draft-lottery-really-worth/>. Accessed: 2017-11-21.
- Thistlethwaite, D.L., Campbell, D.T., 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *J. Educ. Psychol.* 51 (6), 309–317.
- van der Vaart, A.W., van Zanten, J.H., 2008. Rates of contraction of posterior distributions based on gaussian process priors. *Ann. Statist.* 1435–1463.
- van der Vaart, A.W., van Zanten, J.H., 2009. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *Ann. Statist.* 2655–2675.
- Wang, X., 2012. *Bayesian Modeling using Latent Structures* (Ph.D. thesis), Citeseer.
- Wilson, A.G., Adams, R.P., 2013. Gaussian process kernels for pattern discovery and extrapolation. In: *ICML* (3). pp. 1067–1075.