Bayesian Density Regression with Discontinuity

Haoliang Zheng *
Department of Statistics & Data Science, National University of Singapore

Shuangjie Zhang
Department of Statistics, University of California Santa Cruz

Rik S. Sen
Department of Finance, University of Georgia

Surya T. Tokdar
Department of Statistical Science, Duke University

Abstract

Many real-world variables exhibit a phenomenon of bunching right above a predefined threshold. While smoothing techniques could be used to detect such bunching and measure its magnitude, relating these measures to important covariates requires a model-based statistical analysis. We present here a Bayesian method that models the density function through a smooth polynomial basis expansion augmented with a half-kernel to introduce a hard discontinuity at the given threshold. Unlike existing density discontinuity methods, our model incorporates covariates to influence the nonnegative jump size. Posterior inference is carried out by a data augmented Gibbs sampler that can scale to large data sets. We examine whether sharper inference could be obtained by truncating the response data to a shorter interval. Our findings indicate modest truncation may be beneficial to avoid the need of modeling irrelevant features away from the threshold. We study a WAIC criterion for determining a reasonable truncation amount and other hyper-parameters. We illustrate the performance of our model on simulated data and use it to study a corporate proposal voting dataset with a known pass/fail threshold.

Keywords: Bayesian Density Regression; Discontinuity; Regression Discontinuity Designs; Rejection History Algorithm.

*Address for Correspondence: XXXX. E-mail: xxxx.

1 Introduction

From corporate tax laws to insurance contracts, many institutional policies subject an underlying continuous variable to an artificial cutoff, resulting in a large change of the density even for small changes in the endogenous variable (also called running variable or forcing variable in other contexts). When feasible, agents may regulate the magnitude of the endogenous variable to stay on the preferred side of a known cutoff (Lee, 2008). Such regulations may manifest at the population level in the form of a jump discontinuity of the distribution of the endogenous variable at the cutoff value. For example, the presence of a pass/fail threshold may generate discontinuities in grade distributions, which is found to be a consequence of teacher discretion(Diamond and Persson, 2016). Studying this phenomenon of bunching and measuring its magnitude helps to uncover the behaviors of responses to a policy at a given threshold. Noticeably, different from the regression discontinuity designs (RDDs) in causal inference focusing on the treatment status changing discontinuously according to the endogenous variable, the main focus in this paper is the distribution of the endogenous variable itself.

Statistical approaches such as density discontinuity estimation have been vastly developed to detect or estimate the discontinuity in the density of a variable. Local polynomial density estimator on data binning was first introduced in Cheng (1997) and Cheng et al. (1997) to reduce the boundary bias. Besides the local linear estimator, other related work, such as the boundary kernel method (Müller, 1991; Zhang and Karunamuni, 1998), could be both applied to the density discontinuity estimation. More recently, McCrary (2008) developed a pre-binning approach to test discontinuity based on the logarithmic difference in the density function around the threshold. An empirical likelihood estimation procedure was used by Otsu et al. (2013) based on boundary-corrected kernels. Cattaneo et al. (2020) employed local polynomial density estimators to fit the empirical cumulative density function. A good review paper of the density discontinuity approach is by Jales and Yu

(2017). However, there appear to be very few methodologies relating such discontinuity magnitude to important covariates, calling for a model-based statistical analysis. We take a step towards resolving this gap by providing a Bayesian density regression framework that relates the discontinuity size with covariates.

The heuristic idea underlying our approach, and differentiating it from other existing ones, is that we treat the jump as a manifestation of the depletion of observations right below the discontinuity point(threshold). This could happen if either (1) responses that are likely to be below the threshold by a small margin are not recorded or (2) these types of responses were re-sampled, and only the result of the re-sampling is reported. In other words, we are assuming that the dataset contains biased sampling, where the sampling bias negatively affects the probability of inclusion of responses whose results were marginally below the threshold. Thus, the density regression framework is well-suited to embed the discontinuity into the link function, which is a set of polynomial basis functions under our choice among others.

A second novelty of our approach is that our method is compatible with the use of data trimming to promote the estimation of discontinuity around the threshold and reduce computation complexity. We make use of a truncated baseline density before bias sampling to focus on the primary goal of interest: density shape around the given threshold and discontinuity magnitude around the threshold. Discarding the tails of the density follows the assumption that the influence of discontinuity magnitude should fade away when the response is away from the threshold. This trimming approach refrains from the overfitting problem in density regression when we use a set of polynomial basis and saves the computation resources for a big dataset. As would be demonstrated later, the trimming model achieves even more accurate estimations regarding discontinuity with fewer basis functions when we have a complicated design shape. Furthermore, we treat the order of polynomial basis functions and the trimming level as data-driven parameters. We would like to

strike a balance between missing important density features around the given threshold by choosing the order too small and wasting computation on an overly conservative polynomial level. We evaluate the widely applicable information criterion (WAIC, Watanabe (2013)) of models with different orders and trimming choices to choose the optimal combination. We illustrate in simulations that the WAIC criteria chosen model performs better in discontinuity estimation and posterior predictive densities.

The rest of the article is structured as follows. In section 2, we start with the density regression and introduce our Bayesian density regression with discontinuity model. Hereafter, we shall call this model BDRD. In section 3, we introduce an efficient algorithm for fast implementation on big datasets and WAIC for selecting the polynomial order K. In section 4, we provide extensive simulation studies and build a truncated version of BDRD in section 5. In section 6, we apply our models to analyze a real corporate proposal voting dataset. The voting is subjected to a majority cutoff for a pass or fail. In section 7, we discuss the future direction in the field of Bayesian density regression with discontinuity.

2 From density regression to BDRD

Density regression refers to regression techniques where one models and estimates the entire conditional density $p(y \mid \mathbf{x})$ of a response y given predictors $\mathbf{x} = (x_1, \cdots, x_p)$ with fewer structural assumptions than those underlying the routine multilinear regression. Typically density regression is performed under smoothness assumptions of the conditional density function. The focus of this section is thus an important extension where the conditional density potentially contains a sharp jump discontinuity at a given threshold.

Existing Bayesian density regression methods almost exclusively have the mixture form $p(y \mid \mathbf{x}) = \sum_{k=1}^{\infty} \pi_h(\mathbf{x}) g(y \mid \theta_h(\mathbf{x}))$ which provides flexible modelings of densities (Griffin and Steel, 2006; Dunson et al., 2007; Chung and Dunson, 2009; Orlandi et al., 2021). A simpler representing formula we consider is $p(y \mid \mathbf{x}) \propto g(y \mid \cdot)\Phi(r(y, \mathbf{x}))$, where $g(y \mid \cdot)$ is a

smooth baseline density and $\Phi(\cdot)$ is a link function. We refer to $\Phi(r(y, \mathbf{x}))$ as a bias function since we treat the discontinuity as a sequence of bias sampling, and we model it through a discontinuous bias function $\Phi(r(y, \mathbf{x}))$. Many forms of $r(y, \mathbf{x})$ have been studied extensively. See van der Vaart and van Zanten (2008), Tokdar et al. (2010), and Riihimäki and Vehtari (2014), among many others, for examples of using a Gaussian process. Most recently, Li et al. (2022) proposed to use soft Bayesian additive regression trees (SBART), which allows for developing default priors with decent properties. Different from soft decision trees based on covariates $\mathbf{x}$, we introduce a hard discontinuity by extracting a positive half-kernel at the given threshold from $r(y, \mathbf{x})$ which are a set of polynomial basis functions. We then multiply the half kernel with a covariate-dependent regression to quantify the discontinuity magnitude and assess how it is related to measured covariates $\mathbf{x}$. By doing this, the discontinuity is easily encoded in the condition density $p(y \mid \mathbf{x})$ through the discontinuity in the bias function $\Phi(r(y, \mathbf{x}))$.

Below we describe our BDRD model specific to a density within $(-1, 1)$ with a positive jump at $y = 0$ with a depression to the left. For any given variable, one can shift the threshold to $y = 0$ and scale the measured response to $(-1, 1)$. A negative drop can also be achieved by replacing the positive half-kernel with a negative half-kernel. A more ambitious approach is to assume the entire real line for the response and multiple discontinuity points. However, this is not pursued in the current paper due to the additional computing challenges of fitting such models.

We start with the density regression model

$$p(y \mid \mathbf{x}) = \frac{g(y \mid \cdot)\Phi(r(y, \mathbf{x}))}{\int_{-1}^{1} g(t \mid \cdot)\Phi(r(t, \mathbf{x}))\mathrm{dt}}, \tag{1}$$

where $\mathbf{x} = (x_1, \cdots, x_p) \in \mathbb{R}^p$ and $y \in (-1, 1)$. We refer $g(y \mid \cdot)$ to the baseline density before sampling bias, $\Phi(\cdot)$ to the link function and $\Phi(r(y, \mathbf{x}))$ to the bias function. To bring in a positive discontinuity to the bias function, we augment a set of basis functions with a

half-kernel

$$r(y, \mathbf{x} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda) = \sum_{k=1}^{K} (\mathbf{x}'\boldsymbol{\beta}_k) P_k(y) - (\mathbf{x}'\boldsymbol{\alpha})_+ Q(y \mid \lambda) \mathrm{I}(y < 0), \tag{2}$$

where $P_1(y), \cdots, P_K(y)$ are a set of basis functions over $(-1, 1)$ and corresponding coefficients $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_k^{(1)}, \boldsymbol{\beta}_k^{(2)}, \cdots, \boldsymbol{\beta}_k^{(p)})$ for each set. In our setting, we take the basis functions to be normalized Legendre polynomial basis functions. Legendre polynomial basis functions are normalized by making $\int_{-1}^{1} P_k(y)\mathrm{d}y = 0$ and $\int_{-1}^{1} P_k^2(y)\mathrm{d}y = 1$. Orthogonal and normalized basis functions are beneficial for the estimation of $\boldsymbol{\beta}$. The first few orders of polynomial basis functions are $P_1(y) = \sqrt{\frac{3}{2}}y$, $P_2(y) = \sqrt{\frac{5}{8}}(3y^2 - 1)$, $P_3(y) = \sqrt{\frac{7}{8}}(5y^3 - 3y)$ and $P_4(y) = \sqrt{\frac{9}{128}}(35y^4 - 30y^2 + 3)$. We will illustrate in Section 3 a way to select the order K. In (2), $Q(y \mid \lambda)$ is a positive function which equals to 1 at the threshold $y = 0$ and monotonically decreases in $|y|$, with a rate of decay controlled by the persistence parameter λ. The range of λ reflects our assumption that the influence of diffused erosion should fade away when y is away from the threshold 0. In our context, we choose $Q(y \mid \lambda) = e^{-\frac{y^2}{2\lambda^2}}$ with $\lambda \in (0, 0.32)$. The jump size at $y = 0$ is parameterized as $(\mathbf{x}'\boldsymbol{\alpha})_+$ where v_+ denotes the positive part of a real number v. Clearly, the jump is always non-negative, and the estimation of coefficients $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$ is our primary goal, quantifying the relationship between the discontinuity and covariates. A drop discontinuity can be achieved easily by taking the negative part. To complete the specification of the bias function, we use the cumulative distribution function of the standard logistic distribution $\Phi(z) = (1 + e^{-z})^{-1}$ for the link function $\Phi(\cdot)$. The positive valued, bounded and smooth link function ensures that $p(y \mid \mathbf{x})$ is a well-defined probability density function over $(-1, 1)$. To facilitate efficient statistics inferences, we choose the CDF of a standard logistic distribution for $\Phi(\cdot)$ and perform the data augmentation scheme similar to the rejection history algorithm in Rao et al. (2016) and Li et al. (2022). After adopting the data augmentation method, we update all parameters in an efficient Markov chain Monte Carlo (MCMC) for a big dataset which is a technical problem in Bayesian density regression, especially with discontinuity.

For the baseline density before sampling bias $g(y \mid \cdot)$, we set it to be transformed from a $\text{Be}(\gamma_1, \gamma_2)$ random variable by 2 then minus 1. Parameter $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ controls the shape of baseline density. There are several reasons why we consider a parametric model in the whole framework. First, the computation of a parametric model with discontinuity will be easier than a nonparametric one. Especially under the discontinuity condition, the calculation of the normalizing constant becomes complicated under a nonparametric model. With a parametric model, we are able to propose a data-augmented algorithm in Section 3 to obtain posterior samples without much cost. Second, we can make direct statistical inferences through the parameter $\boldsymbol{\alpha}$, which is the essence of our analysis regarding jump discontinuity. Third, our proposed model frame accommodates more customized versions using different parametric models. For example, one may consider using Gaussian basis functions equally spaced over $(-1, 1)$.

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \boldsymbol{\gamma})$ denote all parameters included. With the aforementioned choices, the model for the conditional density of $y \mid \mathbf{x}$ is written as

$$p(y \mid \mathbf{x}, \boldsymbol{\theta}) = \frac{g(y \mid \boldsymbol{\gamma})\Phi(\sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(y) - (\mathbf{x}'\boldsymbol{\alpha})_+ Q(y \mid \lambda)\mathrm{I}(y < 0))}{\int_{-1}^{1} g(t \mid \boldsymbol{\gamma})\Phi(\sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(t) - (\mathbf{x}'\boldsymbol{\alpha})_+ Q(t \mid \lambda)\mathrm{I}(t < 0))\mathrm{dt}}. \tag{3}$$

The likelihood given observations $(\mathbf{x}_i, y_i)$ for $1 \leq i \leq n$ is $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$. It remains to specify prior distributions on $\boldsymbol{\theta}$. We assume $\gamma_{1,2} \overset{iid}{\sim} \text{Ga}(a_\gamma, b_\gamma)$, $\boldsymbol{\beta}^{(j)} = (\boldsymbol{\beta}_1^{(j)}, \cdots, \boldsymbol{\beta}_K^{(j)}) \overset{iid}{\sim} \text{N}(0, \text{I}_K)$, $\boldsymbol{\alpha} \sim \text{N}(0, \text{I}_p)$ and $\lambda \sim \text{Be}(a_\lambda, b_\lambda) \times 0.32$.

3 Model Fitting and Estimation

In this section, we describe how we conduct inference for model (3) and address the computational challenges associated with discontinuity in the Bayesian density regression framework. Then we summarize a comprehensive MCMC algorithm at the end of the section and discuss how to determine the order of polynomials K.

In our model, the normalization term is parameter-dependent, which makes it doubly-

intractable (Murray et al., 2012). One may use standard approximations for the integral or variational methods for the posterior distributions. These operations could be quite expensive when we have a big dataset and more covariates. However, thanks to the rejection history data augmentation by Rao et al. (2016) and proposition 2 in Li et al. (2022), we are able to avoid the intractable integral in the denominator. The idea is to impute the rejected observations $\tilde{\boldsymbol{y}}_i$ until accepting each observation y_i, which results in conditional independence among all joint samples $\boldsymbol{y}_i^\star = (\tilde{\boldsymbol{y}}_i, y_i)$ including both accepted and rejected. The joint likelihood function of $\boldsymbol{y}_i^\star$ no longer contains intractable terms. See derivations of the augmented state-space in Rao et al. (2016) for more details.

From here, one can follow standard techniques to update $\boldsymbol{\theta}$. For example, under our choice of link function, CDF of standard logistic distribution, we introduce Pólya-Gamma latent variables $\omega_i \mid \boldsymbol{y}_i^\star, x_i, \boldsymbol{\theta} \sim \text{PG}(1, r(\boldsymbol{y}_i^\star, x_i, \boldsymbol{\theta}))$ and forge a fully Gibbs update for $\boldsymbol{\beta}$ (Polson et al., 2013). Details of posterior derivations are in the supplementary file. Adaptive metropolis algorithm (Andrieu and Thoms, 2008) can be used for updating $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Especially when $h = 1$, independent metropolis algorithm is available for $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ after using the reparameterization $\boldsymbol{\nu} = (\nu_1, \nu_2)$ with $\nu_1 = \log(\frac{\gamma_1}{\gamma_2})$ and $\nu_2 = \gamma_1 + \gamma_2$.

For parameter λ, we consider ensemble MCMC algorithm in Neal (2011) to produce computational advantages. There are two facts being considered when we choose the ensemble MCMC algorithm. First, generating rejection history and Pólya-Gamma random variables step takes most part of the computing time. After the data augmentation, however, the computation of likelihood becomes much faster. This situation is called "fast and slow variables" by Neal (2011), where ensemble MCMC is specifically designed for. Second, because we restrict λ to a certain range (0, 0.32), it is easy to come up with a proper proposal distribution to sample from, e.g. a Beta distribution multiplied by 0.32. In practice, we use $0.32 \times \text{Be}(1, 2)$ as the proposal distribution. A sketch of the MCMC sampler is summarized in Algorithm 1. More details are in the Supplementary file.

Algorithm 1 MCMC algorithm for sampling the parameters of the model

Step 1 Sample rejected observations $\tilde{\boldsymbol{y}}_i$ for each observation i and combine both rejected and accepted observations $\boldsymbol{y}_i^\star = (\tilde{\boldsymbol{y}}_i, y_i)$ as the total observations for each observation i

Step 2 Sample Pólya-Gamma latent variables $\omega_i \mid \boldsymbol{y}_i^\star, x_i, \boldsymbol{\theta} \sim \text{PG}(1, r(\boldsymbol{y}_i^\star, x_i, \boldsymbol{\theta}))$

Step 3 Update $\boldsymbol{\beta}^{(j)} \mid \boldsymbol{\beta}^{(-j)}$ through conditional normal updates

Step 4 Update $\boldsymbol{\alpha}$ in adaptive MCMC algorithm

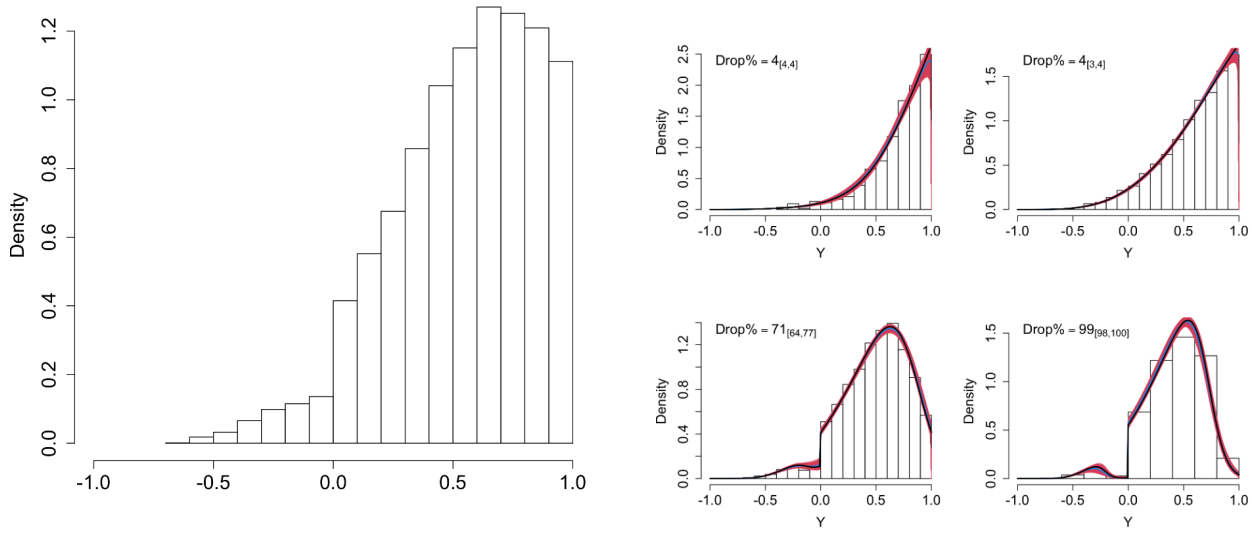
Step 5 Update $\boldsymbol{\gamma}$ in adaptive MCMC algorithm. If $h = 1$, transform $\boldsymbol{\nu} = (\nu_1, \nu_2)$ with $\nu_1 = \log(\frac{\gamma_1}{\gamma_2})$ and $\nu_2 = \gamma_1 + \gamma_2$.

Step 6 Update λ in ensemble MCMC: proposing $M-1$ many alternative $\tilde{\lambda}_{1:(M-1)}$, computing the transition probability, accept one proposal

In the previous sections, our models are defined and estimated by assuming a fixed number of basis functions. Next, we illustrate how we find the best choice of the polynomial order K among others. Various Bayesian and non-Bayesian model selection criteria can be applied, such as AIC(Akaike, 1998), BIC(Schwarz, 1978) and DIC (Spiegelhalter et al., 2002). We adopt a more fully Bayesian approach, namely widely available information criterion (WAIC, Watanabe and Opper (2010)), computing the log pointwise posterior predictive density with an adjustment for the effective number of parameters. After we get the posterior iteration of parameters, we have to calculate the normalizing constant for each observation in (3). We use the trapezoidal rule to obtain numerical integration in the denominator of the conditional density.

4 Simulation 1

In this section, we use simulations to assess BDRD's performance to recover the discontinuity estimation and further provide comparisons of different models under different orders K. To construct synthetic conditional densities, we first include an intercept x_{i1} and sample a continuous covariate x_{i2} from $\text{N}(0, 1)$, so we have $\mathbf{x}_i = (x_{i1}, x_{i2})'$ with $p = 2$. We consider the



(a) Density histogram of simulated y_i (b) Posterior Predictive Density of $y_i^\star$

Figure 1: [Simulation 1] Density histogram of simulated y_i and posterior predictive density of $y_i^\star$ under $K = 2$ is plotted in (a) and (b), respectively. In (b), the histogram consists of simulated y_i whose $|x_{i2} - x_2^\star| < 0.25$. The red lines are the posterior predictive density in each posterior iteration, the blue lines are the posterior predictive mean density, and the black lines are the truth density for new $y_i^\star$.

scenario with 20k samples and set the coefficients $\boldsymbol{\beta}^{\text{tr}}_{K^{\text{tr}}\times p}$ as $\boldsymbol{\beta}^{\text{tr}}_{2\times 2} = \begin{pmatrix} 1.2598 & 0 \\ -1.6498 & -1.3296 \end{pmatrix}$, where the values of $\boldsymbol{\beta}^{\text{tr}}$ are obtained by fitting a mixture of Beta distributions by the quasi-Newton optimization method. The mixture of Beta distributions is $0.5\text{Be}(10, 10)+ 0.4\text{Be}(2, 1)+0.1\text{Be}(1, 1)$. For other parameters, we set $\boldsymbol{\alpha}^{\text{tr}} = (1, 4)$ implying a jump discontinuity, $\boldsymbol{\gamma}^{\text{tr}} = (4, 1)$ and $\lambda^{\text{tr}} = 0.16$. We again utilize the rejection history algorithm to sample the discontinuous variable y_i. The simulated densities are shown in Fig 1 (a).

We run BDRD model under $K = 2, 3, 4$ under this simulation setting with 100 replicates of simulated data. The hyper-parameters we use follows Section 2, namely, $a_\gamma = 3, b_\gamma = 3, a_\lambda = 1, b_\lambda = 2$. We summarise the average length and coverage of posterior credible

intervals, mean squared error (MSE) of posterior mean estimates, and percentage of time having the smallest WAIC in Table 1. The result shows that WAIC chooses the correct order $K = 2$, which enjoys the smallest MSE for all parameter estimates and the shortest length of credible intervals. Although $K = 3$ and 4 shares some probability of being chosen as the best model, the 95% credible intervals of redundant coefficients β_{kp} include 0 most of the time. Furthermore, we plot the posterior predictive density of $y_i^\star$ with new covariate $x_2^\star = (-2, -0.75, 0.5, 1.75)$ from one replicate in Fig 1 (b). The good posterior predictive density indicates a good fit of the model.

5 Truncated BDRD

In real applications, we may encounter a big dataset which has a complicated density shape away from the discontinuity at the threshold, which requires a higher order of basis function to catch those patterns at the tail. We think that the discontinuity's impact decreases with the distance from the response to the threshold. Thus, we are able to use part of the data around the threshold to fit a simpler model and ignore the tail points. Following this vein, we introduce a truncated version of model (3) which can be seen as a "data trimming" approach. Suppose we only use the responses between $(-h, h)$ with $0 < h \leq 1$, then a truncated BDRD model can be built as

$$p(y \mid \mathbf{x}, \boldsymbol{\theta}) = \frac{g(y \mid \boldsymbol{\gamma})\Phi(\sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(y) - (\mathbf{x}'\boldsymbol{\alpha})_+ Q(y \mid \lambda)\mathrm{I}(y < 0))}{\int_{-h}^{h} g(t \mid \boldsymbol{\gamma})\Phi(\sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(t) - (\mathbf{x}'\boldsymbol{\alpha})_+ Q(t \mid \lambda)\mathrm{I}(t < 0))\mathrm{dt}}, \tag{4}$$

where $y \in (-h, h)$. When $h = 1$, model (4) returns to the standard version. Under the same conditions, we can view model (4) using a truncated baseline sampling distribution $g(y \mid \cdot)$ on $(-h, h)$ instead of $(-1, 1)$. From this point of view, the data-augmented method can also be adapted to a truncated version by augmenting variables from a truncated distribution.

When we apply the truncated model, we benefit from fewer observations and tend to select fewer basis functions. It reduces the computational cost and puts more emphasis on

Table 1: [Simulation 1] Average length of 95% credible interval, mean square error (MSE) of posterior mean estimates and coverage rate of 95% credible interval are computed based on 100 replicates

		$K=2$			$K=3$			$K=4$		
$\boldsymbol{\theta}$	Tr	Length	MSE	Coverage	Length	MSE	Coverage	Length	MSE	Coverage
α_1	1	0.748	**0.045**	0.94	0.839	0.048	0.95	0.871	0.051	0.95
α_2	4	1.166	**0.150**	0.83	1.227	0.231	0.77	1.244	0.223	0.78
β_{11}	1.2598	0.655	**0.033**	0.90	0.805	0.082	0.83	0.884	0.102	0.81
β_{12}	0	0.267	**0.004**	0.97	0.398	0.016	0.86	0.412	0.014	0.92
β_{21}	-1.6498	0.416	**0.016**	0.90	0.774	0.107	0.72	1.020	0.167	0.78
β_{22}	-1.3296	0.326	**0.007**	0.92	0.573	0.042	0.81	0.609	0.039	0.86
β_{31}	0				0.383	0.020	0.79	0.699	0.051	0.83
β_{32}	0				0.291	0.010	0.83	0.470	0.017	0.93
β_{41}	0							0.268	0.006	0.91
β_{42}	0							0.275	0.004	0.98
γ_1	4	0.363	**0.008**	0.96	0.479	0.014	0.95	0.570	0.018	0.96
γ_2	1	0.071	**0.000**	0.96	0.085	0.000	0.98	0.096	0.000	0.99
λ	0.16	0.038	**0.000**	0.95	0.042	0.000	0.96	0.050	0.000	0.94
WAIC			**63%**			24%			13%	
α_1^{adp}	1	0.783	**0.04356**	**0.96**						
α_2^{adp}	4	1.189	0.175	0.81						

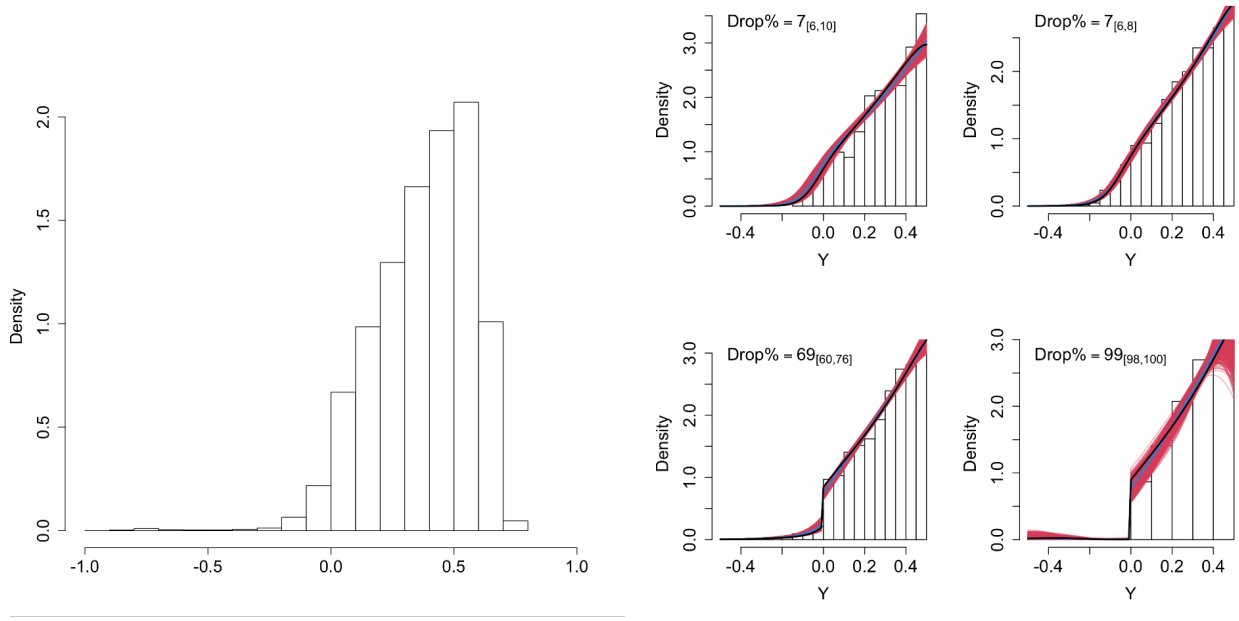
the discontinuity analysis. But this sacrifices observations distant from the threshold and property of orthogonality of basis functions during trimming. There's a determination of trimming versus no trimming and the value of h, then we also follow the aforesaid WAIC rules to help users make decisions. The truncated model can be seen as a series of nested models using a subset of the entire data to fit the BDRD models. To make a reasonable comparison, the calculation of WAIC needs to be based on the smallest subset of all trimmed

datasets under all models. One may also use expert information or empirical density to determine the trimming level h. For example, in simulation 2 in the next section, we observe that the response 0.5 away from the threshold contributes very little to the estimation of discontinuity, then we let $h = 0.5$ in simulation 2. More generally, one may also choose an optimal $h^\star$ among candidates through other criteria, or even an asymmetric trimming strategy.

5.1 Simulation 2

We provide another simulation here to compare the performance of the truncated model and the normal model. In simulation 2, we also consider the same two covariates $\mathbf{x}_i = (x_{i1}, x_{i2})'$ as simulation 1. But this time, we consider a higher true order of polynomials densities with $K^{\text{tr}} = 4$. After fitting another mixture of Beta distributions $0.4\text{Be}(10, 10)+0.5\text{Be}(0.1, 100)+0.1\text{Be}(0.1, 10)$, we manually set the coefficients $\boldsymbol{\beta}^{\text{tr}}_{K^{\text{tr}}\times p}$ as $\boldsymbol{\beta}^{\text{tr}}_{4\times 2} = \begin{pmatrix} -5 & 0 \\ -1 & -1.5 \\ -8 & 0.5 \\ 1 & -1 \end{pmatrix}$. For other parameters, we use the same setting as simulation 1 by letting $\boldsymbol{\alpha}^{\text{tr}} = (1, 4), \boldsymbol{\gamma}^{\text{tr}} = (4, 1)$ and $\lambda^{\text{tr}} = 0.16$. The simulated densities are shown in Fig 2 (a), where there is a quickly decreasing tail after 0.5 as we design.

We run the BDRD model and truncated BDRD model under $K = 3, 4$ setting with 100 replicates. The hyper-parameters we use follows Section 2, namely, $a_\gamma = 3, b_\gamma = 3, a_\lambda = 1, b_\lambda = 2$. For the truncation level h, we fix it at $h = 0.5$ here due to the clear drop after 0.5 in Fig 2 (a). We summarise the same measurement quantities: average length and coverage of posterior credible intervals, mean squared error (MSE) of posterior mean estimates, and percentage of time having the smallest WAIC in Table 2. WAIC chooses the trimming model with $K = 3$ at 48% of the time, and the no-trimming model with $K = K^{\text{tr}} = 4$ comes second. WAIC switching between the full model and the



(a) Density histogram of simulated y_i (b) Posterior Predictive Density of $y_i^\star$

Figure 2: [Simulation 2] Density histogram of simulated y_i and posterior predictive density of $y_i^\star$ model using $K = 3, h = 0.5$ is plotted in (a) and (b), respectively. In (b), the histogram consists of simulated y_i whose $|x_{i2} - x_2^\star| < 0.25$. The red lines are the posterior predictive density in each posterior iteration, the blue lines are the posterior predictive mean density, and the black lines are the truth density for new $y_i^\star$.

trimmed lower-ordered model demonstrates the robustness of WAIC and the benefits of trimming. The trimming model bears larger credible intervals due to fewer observations in trimmed data. For the discontinuity magnitude regression intercept α_1, trimming models have smaller MSE. In terms of another parameter of interest α_2, the no-trimming model with $K = K^{\text{tr}} = 4$ has the smallest MSE. The posterior predictive density distribution $y_i^\star$ under $x_2^\star = (-2, -0.75, 0.5, 1.75)$ is plotted in Fig 2 (b).

Table 2: [Simulation 2]Average length of 95% credible interval, mean square error (MSE) of posterior mean estimates and coverage rate of 95% credible interval are computed based on 100 replicates

		No-trimming						Trimming								
		$K=3$			$K=4$			$K=2$			$K=3$			$K=4$		
$\boldsymbol{\theta}$	Tr	Length	MSE	Coverage	Length	MSE	Coverage	Length	MSE	Coverage	Length	MSE	Coverage	Length	MSE	Coverage
α_1	1	0.820	0.178	0.88	0.909	0.106	0.83	0.696	**0.057**	**0.89**	0.778	0.086	0.79	0.791	0.089	0.81
α_2	4	1.357	1.044	0.73	1.279	**0.231**	0.76	1.173	0.789	0.25	1.292	0.582	0.46	1.288	0.577	0.41
β_{11}	-5	0.583	0.654	0.22	0.710	**0.105**	0.68	1.830	184.595	0.00	3.367	52.185	0.00	3.433	47.547	0.00
β_{12}	0	0.536	0.361	0.02	0.614	**0.029**	0.94	1.298	6.765	0.00	3.102	0.441	0.99	3.122	0.204	1.00
β_{21}	-1	0.660	0.599	0.03	0.925	**0.062**	0.94	0.927	0.085	0.88	1.016	0.071	0.94	2.501	1.407	0.65
β_{22}	1.5	0.553	1.058	0.00	0.835	**0.046**	0.97	0.468	1.652	0.00	0.581	1.928	0.00	1.708	0.377	0.87
β_{31}	-8	0.766	1.008	0.36	0.878	**0.200**	0.53				2.004	14.459	0.00	2.154	17.302	0.00
β_{32}	0.5	0.715	0.129	0.71	0.832	**0.073**	0.87				2.013	0.333	0.95	2.064	0.178	0.99
β_{41}	1				0.638	**0.024**	0.96							2.046	0.132	1.00
β_{42}	-1				0.573	**0.014**	0.98							1.453	0.094	0.99
γ_1	4	0.527	1.408	0.04	0.512	**0.033**	0.85	2.164	0.558	0.89	1.839	0.895	0.57	1.788	1.216	0.42
γ_2	1	0.309	0.943	0.03	0.311	**0.016**	0.77	1.101	0.044	0.99	0.940	0.142	0.77	0.889	0.207	0.61
λ	0.16	0.065	0.001	0.96	0.101	**0.001**	0.81	0.100	0.006	0.05	0.108	0.002	0.76	0.121	0.002	0.73
WAIC			0%			42%			0%			**48%**			10%	
α_1^{adp}	1	0.826	**0.0797**	0.82												
α_2^{adp}	4	1.290	0.398	0.64												

6 Corporate voting dataset

6.1 Data Insight

The original corporate voting dataset contains 52 variables and 30,566 data points. For each corporate proposal, the response value is the distance from the threshold, which is derived by subtracting pass threshold from proposal pass rate. After cleaning data and establishing some covariates, finally the data used for model is left 5 predictors and 19,775 data points. The explanation of predictors are as follows in Table 3. The density of response variable is in Fig 3.

Table 3: Explanation of all variables in the corporate proposal voting data

Variable Name	Explanation
from.requirement.threshold	Response Variable, distance from the threshold
ISS.recommendation	ISS is a company providing opinion or strategy for the companies. Most is 1, which means the vote is supported by the ISS company.
analyst.coverage	How many analyst actively tracking and publishing opinions on a company and its stock.
past.stock.return	Reutrn of each stock
q	The Q ratio, also known as Tobin's Q, equals the market value of a company divided by its assets' replacement cost.
firm.size	Firm size

From the density plot, we can convince that there is indeed some discontinuity at threshold 0. Therefore, our BDRD can be implemented for estimating the cause of discontinuity and measure the discontinuity magnitude.

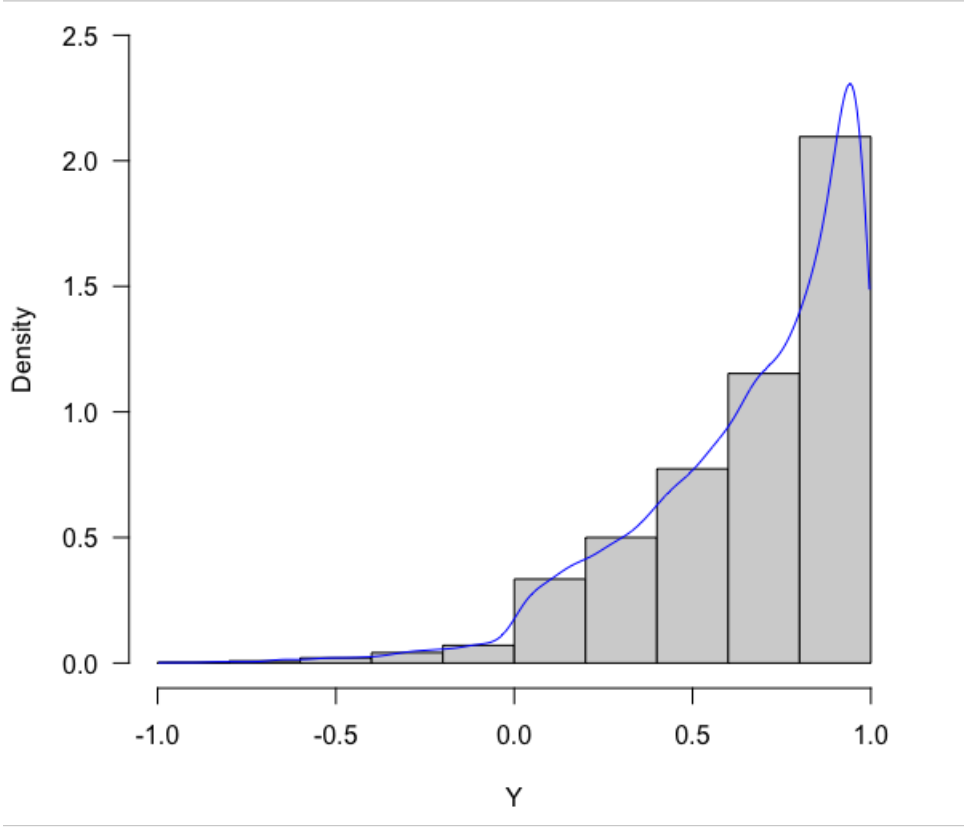


Figure 3: Histogram of response in corporate voting data with kernel density estimation

6.2 BDRD Results

We consider truncated BDRD model with $h = 0.5, 0.6, \ldots, 1$ and $K = 2, 3$. For each model, we run 10 parallel chains from different initializing points, with the burn-in size of $15,000$ and posterior samples of $25,000$. The prior we use here is independent standard normal.

Below we use $\alpha_1, \alpha_2, \alpha_3, \ldots$ to denote the parameter for intercept, ISS. recommendation, analyst.coverage and so on. The 95% credible intervals of $\alpha_1, \ldots, \alpha_6$ are as follows in Fig 4. We can tell that results based on models with $K = 2$ and $h = 1, \ldots, 0.7$ will result in same conclusion, while results based on models with $K = 2$ and $h = 0.6, 0.5$ and models with $K = 3$ and $h = 0.8, \ldots, 0.5$ will result in another conclusion. The reason is that flexibility

of model with order 2 is not enough, which requires more trimming level.

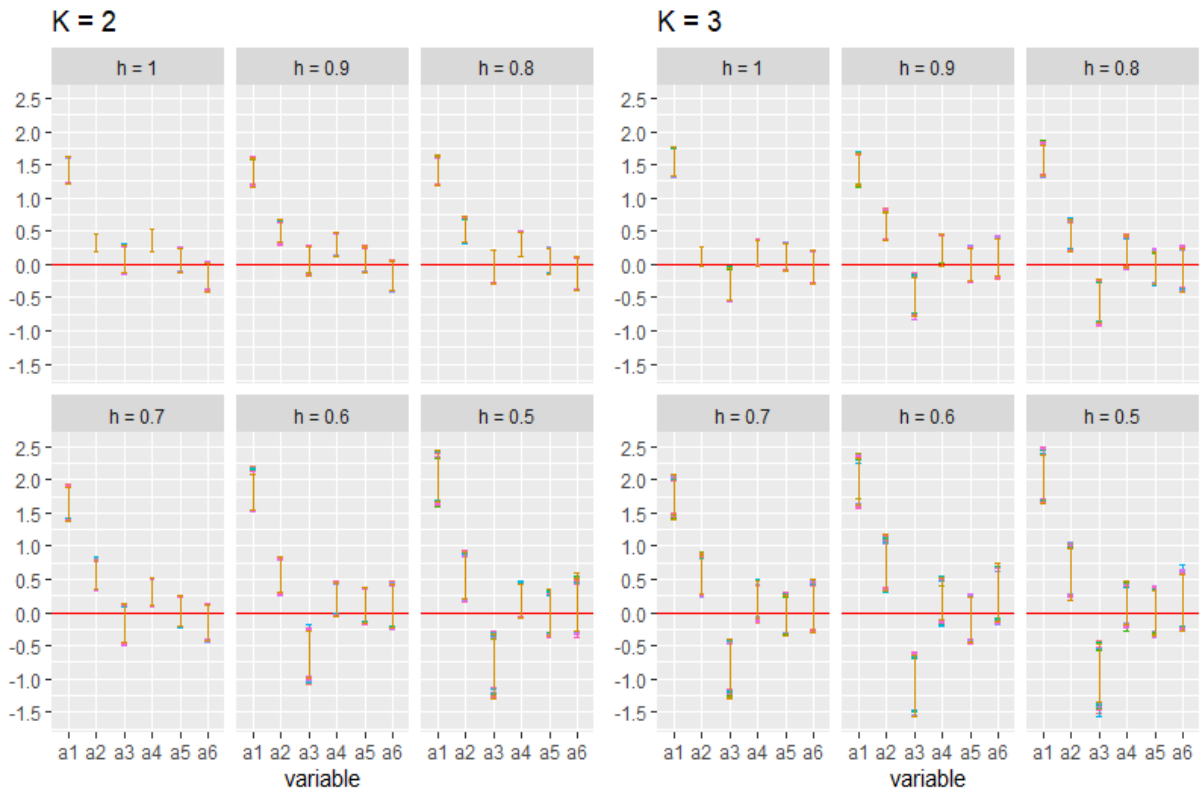


Figure 4: 95% credible intervals of $\alpha_1, \ldots, \alpha_6$ from models with different trimming and order, with different color denoting different chains

The WAIC of all models with different trimming and order based on 10 chains is in Table 4. The pattern shows that WAIC decreases as order and trimming level increases. When $K = 4$, we find multi-modality issue in the posterior samples, which indicates too much flexibility. Therefore, we may take truncated RDRD with $K = 3$ and $h = 0.5$ as the final model. The conclusion is: ISS.recommendation has significantly positive effect on the discontinuity, analyst.coverage has significantly negative effect on the discontinuity, and other predictors has no significant effect.

Table 4: Average WAIC of all models with different trimming and order based on 10 chains

	$h = 1$	$h = 0.9$	$h = 0.8$	$h = 0.7$	$h = 0.6$	$h = 0.5$
$K = 2$	-4582.394	-4620.850	-4658.979	-4697.427	-4730.162	-4752.063
$K = 3$	-4611.965	-4650.481	-4692.916	-4721.809	-4746.226	-4756.414

7 Limitation and Discussion

In this article, we proposed a new method for density regression with discontinuity. It uses a Bayesian density regression framework incorporating covariates to estimate the discontinuity magnitude and can be adapted to a truncated version. On simulated data, we illustrated how BDRD captures the relationship between the covariate and the discontinuity. In addition, we provide a WAIC method to choose the order and the trimming level. Using corporate voting data, we showed how BDRD output the effect of XXX on the conditional distribution of corporate voting rate. XXX

One can extend the model using some shrinkage priors on $\beta_{K\times p}$ to stronger identifiability of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For example, under our choices of a polynomial function, shrinkage priors on the odds orders lead to better identifiability of discontinuity magnitude estimation. Furthermore, a non-parametric structure can be accommodated into the baseline density, while it requires careful design of discontinuity magnitude to achieve reasonable interpretation. Lastly, our model needs to be modified to consider the estimation of multiple discontinuity points in the future.

SUPPLEMENTARY MATERIAL

Title: Brief description. (file type)

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer.

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive mcmc. *Statistics and computing 18*(4), 343–373.

Cattaneo, M. D., M. Jansson, and X. Ma (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association 115*(531), 1449–1455.

Cheng, M.-Y. (1997). Boundary aware estimators of integrated density derivative products. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(1), 191–203.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997). On automatic boundary corrections. *The Annals of Statistics 25*(4), 1691–1708.

Chung, Y. and D. B. Dunson (2009). Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association 104*(488), 1646–1660.

Diamond, R. and P. Persson (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Technical report, National Bureau of Economic Research.

Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 163–183.

Griffin, J. E. and M. J. Steel (2006). Order-based dependent dirichlet processes. *Journal of the American statistical Association 101*(473), 179–194.

Jales, H. and Z. Yu (2017). Identification and estimation using a density discontinuity approach. *Regression Discontinuity Designs*.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics 142*(2), 675–697.

Li, Y., A. R. Linero, and J. Murray (2022). Adaptive conditional distribution estimation with bayesian decision tree ensembles. *Journal of the American Statistical Association*, 1–14.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics 142*(2), 698–714.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika 78*(3), 521–530.

Murray, I., Z. Ghahramani, and D. MacKay (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

Neal, R. M. (2011). Mcmc using ensembles of states for problems with fast and slow variables such as gaussian process regression. *arXiv preprint arXiv:1101.0387*.

Orlandi, V., J. Murray, A. Linero, and A. Volfovsky (2021). Density regression with bayesian additive regression trees. *arXiv preprint arXiv:2112.12259*.

Otsu, T., K.-L. Xu, and Y. Matsushita (2013). Estimation and inference of discontinuity in density. *Journal of Business & Economic Statistics 31*(4), 507–524.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association 108*(504), 1339–1349.

Rao, V., L. Lin, and D. B. Dunson (2016). Data augmentation for models based on rejection sampling. *Biometrika 103*(2), 319–335.

Riihimäki, J. and A. Vehtari (2014). Laplace approximation for logistic gaussian process density estimation and regression. *Bayesian analysis 9*(2), 425–448.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology) 64*(4), 583–639.

Tokdar, S. T., Y. M. Zhu, and J. K. Ghosh (2010). Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian analysis 5*(2), 319–344.

van der Vaart, A. W. and J. H. van Zanten (2008). Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics 36*(3), 1435–1463.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research 14*(27), 867–897.

Watanabe, S. and M. Opper (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research 11*(12).

Zhang, S. and R. J. Karunamuni (1998). On kernel density estimation near endpoints. *Journal of statistical Planning and inference 70*(2), 301–316.