Bayesian Semiparametric 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 Sen
Department of Finance, University of Georgia

Surya T. Tokdar
Department of Statistical Science, Duke University

Abstract

In many real-world scenarios, variables show a distinct bunching pattern just above a known threshold. Detecting and quantifying the extent of such bunching behavior is achieved using density smoothing methods. However, a researcher may also want to correlate the magnitude of bunching with pertinent covariates. To this end, we introduce a Bayesian approach that models the density function using a smooth polynomial basis expansion, supplemented with a half-kernel to introduce a discontinuity at the specified threshold. Our model allows the covariates to impact the size of the jump at the threshold. We employ a data-augmented Gibbs sampler for posterior inference, that can also be used for large datasets. We also introduce a model selection criterion to formally compare and select between different versions of the model one could fit to the data. The efficacy of our model is demonstrated through simulated data and its application to a corporate proposal voting dataset with a known pass/fail threshold.

Keywords: Density discontinuity; Bunching estimation; Bayesian density regression; Double intractability; Data augmentation

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

1 Introduction

Many institutional policies attach a dichotomous outcome to a continuous endogenous variable. In response, agents may manipulate choices to stay on the preferred side of a threshold value, causing the empirical distribution of the variable to manifest a jump discontinuity at the cutoff. The existence of such density discontinuity has been noted for taxable earning (where the marginal tax rate increases on crossing certain thresholds of taxable income), firm size (where certain regulations apply only for firms that are larger than a certain value), student grade (where scoring below the passing threshold implies failing the course), etc.; see Jales and Yu (2017) and the references therein. Detecting and measuring a density discontinuity offer useful insight into agents' incentives and sensitivity to a policy variable that changes discontinuously around the threshold. Many statistical methods to measure the size of the discontinuity have been proposed for this task (e.g., McCrary, 2008; Otsu et al., 2013; Cattaneo et al., 2020). However, these methods do not address the natural follow-on question of relating density discontinuity to measured covariates, i.e., is the density discontinuity larger for certain types of agents (e.g., what kind of people report a taxable income that avoided crossing a tax threshold?). Our paper presents a method to answer such questions.

Our motivating application is a study of corporate proposal voting (Section 4). For certain significant corporate decisions, the management of the company is required to put up a proposal that is voted on by the shareholders and "passes" before it can be implemented. If the fraction of "yes" votes is more than a predefined threshold (typically 50%) then the proposal passes, otherwise it fails. The management would prefer that their proposal not fail, and they have various tools at their disposal to try and achieve this. The histogram of the fraction votes supporting a proposal ($n \approx 20,000$) clearly shows density discontinuity at the cutoff – there are many more proposals that pass narrowly than those that fail narrowly – and the discontinuity is statistically significant (Figure 1). These conclusions do not hold with equal force when the data is segmented along measured attributes. The statistical evidence of a jump is more ambiguous for proposals with a positive recommendation from the Institutional Shareholder Services (a voting advisory entity that recommends how shareholders should vote on each proposal) than those with a negative recommendation, even though the former group is about 6 times larger in size. The jump size itself seems to diminish if a higher number of financial analysts cover the firm. A discontinuity is barely detectable except in samples with the lowest 33% of analyst coverage. A similar pattern can be noticed for firm size but not for the firm's Q ratio (the ratio of the market value and the book value of the firm – a measure of future growth opportunities of the firm that is often found to be correlated with outcomes examined in corporate governance studies). In other words, the existence of a density discontinuity and its magnitude both appear to be associated with some covariates but not others.

While segmentation analyses are useful in understanding how agents' response to policy is influenced by covariates, the qualitative nature of these analyses is unsatisfactory for several reasons. Lacking a framework to relate results from different segments to one another, such analyses fail to establish a clear pattern of association between density discontinuity and the variable, especially when the variable is continuous or ordinal. For example, results presented in Figure 1 cannot be used to quantify if greater analyst coverage is associated with lower agent manipulation, or whether Q ratio has no association whatsoever.

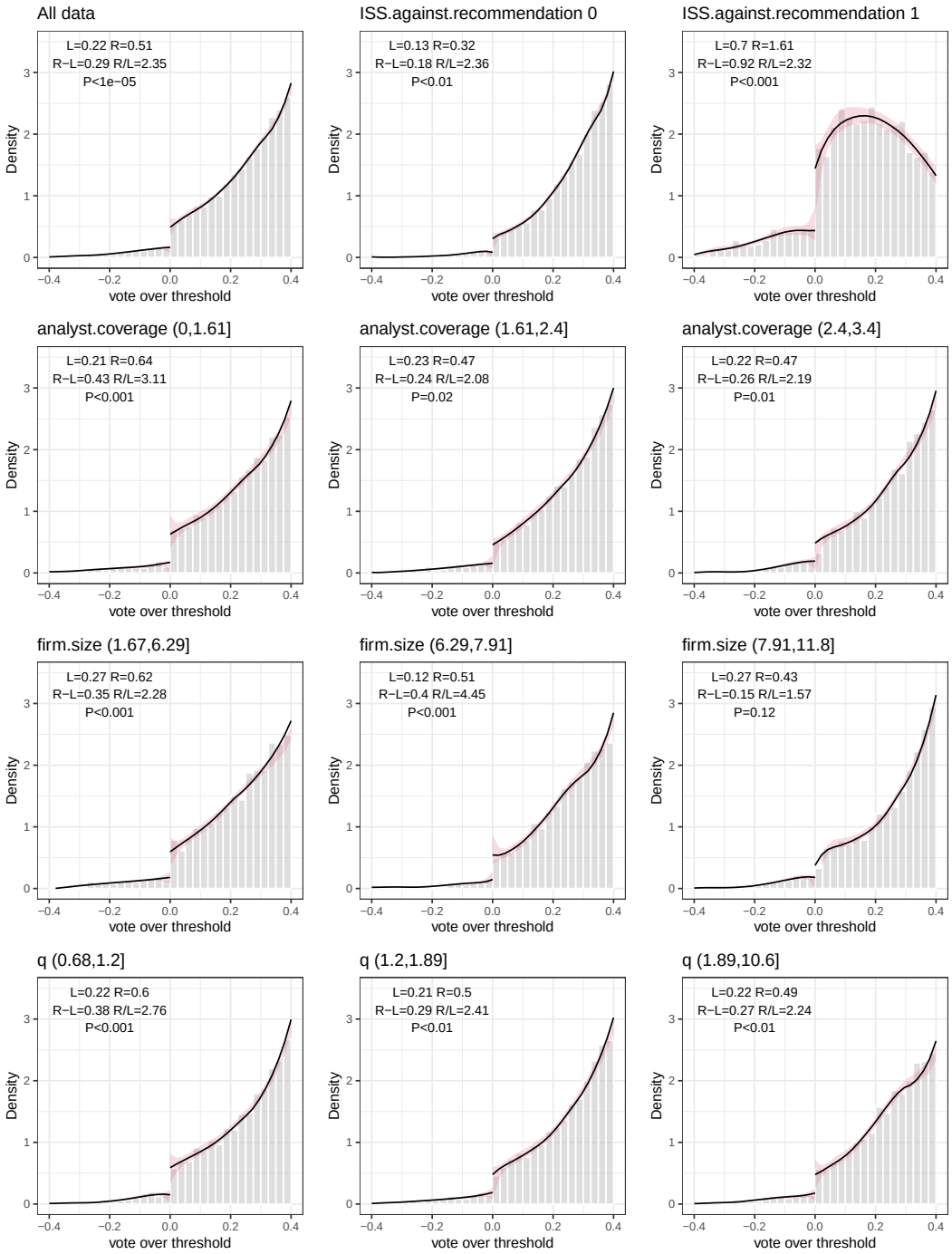


Figure 1: Density discontinuity testing for corporate proposal voting. Estimates and discontinuity tests were done using rddensity R-package by Cattaneo, with cubic local polynomial fit used for density estimation. For all continuous covariates, data was segmented into equal thirds according the sorted values of the covariate in concern.

Moreover, absent a joint analysis of association, it is impossible to answer whether density discontinuity exhibits association with a particular covariate of interest, say, analyst coverage, when adjusted for all remaining variables, including firm size. Although this problem could be partially alleviated by jointly segmenting over two or more covariates, such an

extension is difficult to put into practice. A joint segmentation analysis exacerbates the interpretability challenges mentioned above. It may also lead to statistical inefficiency and issues of multiple testing by partitioning data into many segments; one or more segments may end up with very little data to reliably analyze density discontinuity.

Arguably, these shortcomings could be overcome by using regression tools to jointly analyze the association between covariates and (conditional) density discontinuity. To the best of our knowledge, no such statistical tool currently exists in the literature. Density estimation, with or without discontinuity, is mostly taken to be a data smoothing exercise whose statistical validity rests on large sample estimation theory (Silverman, 2018; Cattaneo et al., 2020). Although seemingly model-free, many such smoothing techniques closely relate to (regularized) likelihood-based estimators in suitably defined high- or infinite-dimensional model settings involving mixture decompositions or basis expansions (Kiefer and Wolfowitz, 1956; Good and Gaskins, 1971; Leonard, 1978; Silverman, 1982; Ferguson, 1983), with equally compelling asymptotic guarantees (van der Vaart and van Zanten, 2009; Shen et al., 2013; Chen, 2017). Such model-based reformulations and the associated likelihood-based estimation are particularly useful in settings where the estimation of the density function is one piece within a larger estimation task, as is the case with the density discontinuity regression problem pursued here. Indeed, our primary goal is to estimate a parametric relation between covariates and the density discontinuity, treating the rest of the density function as a nuisance parameter. In general, likelihood-based methods, especially the integrated likelihood approach of Bayesian methods are well-suited for handling high-dimensional nuisance parameters (Kalbfleisch and Sprott, 1973; Berger et al., 1999).

We pursue such a likelihood-based inference method within a generalized linear model setting. The jump size is modeled as a linear function of the covariates so that the coefficients determining this linear function can be interpreted as each measuring the association of the corresponding attribute, adjusted for other variables in the model. Our formulation builds upon the mostly Bayesian literature on density regression, where one attempts to estimate the conditional density functions of a response given predictors (Dunson et al., 2007; Tokdar et al., 2010). Density discontinuity is incorporated into this framework as a discontinuous biasing function with a parametric form involving the linear function of covariates mentioned above (Section 2.1). The biasing function modifies an otherwise smooth conditional density function which also needs to be estimated from data. A Bayesian formulation proves particularly helpful in parameter estimation to sidestep some intractable integrals associated with the likelihood function that arise from the biasing operation (Section 2.2). The resulting estimation method is complemented by a model selection technique based on an information criterion due to Watanabe (2013), and its practical utility is discussed in Section 2.3.

Understanding the behavior of agents in response to policies is not merely an academic pursuit; it has profound implications for economics, corporate governance, and regulatory effectiveness. For instance, when taxpayers manipulate earnings to avoid higher tax brackets, it can lead to significant revenue losses for governments and distort economic indicators. Similarly, if companies maneuver to receive favorable votes on proposals, it could reflect issues in shareholder democracy and corporate transparency. The methodology in this paper for detecting and interpreting heterogeneity in density discontinuities would help get a more nuanced understanding of how different agents alter their behavior differently in response to policies, thereby informing more effective policy designs.

2 Method

2.1 Modeling conditional density with discontinuity

Let Y denote the endogenous variable of interest. Without loss of generality, $Y \in (-1, 1)$ and the cutoff point is zero. Given a vector of covariates $\mathbf{x} = (x_1, \ldots, x_p)$, the conditional density of Y could be expressed as $p(y|\mathbf{x}) \propto g(y)\Phi(r(\mathbf{x}, y))$, where $g(y)$ is a baseline density, $r(\mathbf{x}, y)$ is a biasing function influenced by the covariates and filtered through a nonlinear transformation $\Phi(\cdot)$. Borrowing terminology from generalized linear models, one may call Φ^{-1} a link function. To embed a parametric form of jump discontinuity, we further decompose the biasing function as $r(\mathbf{x}, y) = r_c(\mathbf{x}, y) + (\mathbf{x}'\boldsymbol{\alpha})_+ j(y)$, where $r_c(\mathbf{x}, y)$ is continuous in y, and $j(y)$ has a jump discontinuity of unit length at $y = 0$ but is otherwise continuous. Consequently, $r(\mathbf{x}, y)$ inherits a jump discontinuity at the cutoff with jump length $(\mathbf{x}'\boldsymbol{\alpha})_+ := \max(\mathbf{x}'\boldsymbol{\alpha}, 0)$, which depends on the covariates through a linear combination determined by the coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$. Estimating $\boldsymbol{\alpha}$ from data is the primary goal.

Density regression models of this type, without discontinuity, have been explored earlier (Tokdar et al., 2010; Li et al., 2022), focusing mainly on nonparametric Bayesian estimation of $r_c(\mathbf{x}, y)$ under prior distributions that allow great flexibility of shape and smoothness. Here, to keep the focus on estimation of $\boldsymbol{\alpha}$, we consider a simpler formulation $r_c(\mathbf{x}, y) = \sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(y)$, determined by unknown coefficient vectors $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^\top$, $1 \le k \le K$, where $P_1(y), \cdots, P_K(y)$ are normalized Legendre polynomial basis of maximum degree K, which we call the *order* of the model. The Legendre polynomials satisfy $\int_{-1}^{1} P_k(y)\mathrm{d}y = 0$, $\int_{-1}^{1} P_k^2(y)\mathrm{d}y = 1$, and $\int_{-1}^{1} P_k(y)P_l(y)\mathrm{d}y = 0$ if $k \neq l$. The first few polynomials are

$$P_1(y) = a_1 y, \quad P_2(y) = a_2(3y^2 - 1), \quad P_3(y) = a_3(5y^3 - 3y), \quad P_4(y) = a_4(35y^4 - 30y^2 + 3),$$

with $a_1 = \sqrt{3/2}$, $a_2 = \sqrt{5/8}$, $a_3 = \sqrt{7/8}$, and $a_4 = \sqrt{9/128}$. Taking $K = \infty$ produces an orthonormal basis of the L_2 space on $(-1, 1)$, but we restrict to a finite K. Section 2.2 introduces a data-driven choice of K.

One could consider various shapes for the jump function $j(y)$ to capture different types of agent manipulation. For illustrative purposes, we shall focus on a parametric half-kernel $j(y) = j_\lambda(y) = -\exp\{-\frac{y^2}{2\lambda^2}\} \cdot \mathrm{I}(y \le 0)$, with $j_\lambda(y) = 0$ for $y > 0$, and $j_\lambda(y) < 0$ and monotonically decreasing for $y \le 0$ with $j_\lambda(0) = -1$ and $\lim_{y\to-\infty} j_\lambda(y) = 0$. Such a jump function could be justified for corporate proposal voting if it is believed that management has some ability to detect, and withhold from voting, proposals that lack majority support by a small margin; the propensity of withholding decaying as the margin grows. The rate of this decay is controlled by the persistence parameter $\lambda > 0$. See Section 5 for further comments on $j(y)$.

We complete the model formulation by taking $g(y) = g_{\boldsymbol{\gamma}}(y) \propto (\frac{1+y}{2})^{\gamma_1 - 1}(\frac{1-y}{2})^{\gamma_2 - 1}$, with $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ to be estimated from data. A parametric form of the baseline density adds extra shape flexibility, especially at the boundaries of the support $(-1, 1)$. Notice that $Y \sim g_{\boldsymbol{\gamma}}(y)$ if and only if $\frac{1+Y}{2} \sim \mathrm{Be}(\gamma_1, \gamma_2)$.

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \boldsymbol{\gamma})$ denote all parameters of the model. Given observations $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$, the likelihood function is $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i)$, where

$$p_{\boldsymbol{\theta}}(y|\mathbf{x}) = \frac{g_{\boldsymbol{\gamma}}(y)\Phi(\sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(y) + (\mathbf{x}'\boldsymbol{\alpha})_+ j_\lambda(y))}{\int_{-1}^{1} g_{\boldsymbol{\gamma}}(t)\Phi(\sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(t) + (\mathbf{x}'\boldsymbol{\alpha})_+ j_\lambda(t))\mathrm{d}t}. \tag{1}$$

The denominator in (1) cannot be evaluated in closed form, making the likelihood function intractable and unfit for maximum likelihood estimation. For Bayesian computation, an intractable likelihood function usually leads to a doubly intractable posterior distribution that is difficult to compute or even approximate with Markov chain Monte Carlo (Murray et al., 2006; Caimo and Mira, 2015). Fortunately, for the present model, double intractability could be resolved with a clever data augmentation strategy when Φ is a bounded transformation (Adams et al., 2009; Rao et al., 2016). In the following, we consider $\Phi(r) = (1 + e^{-r})^{-1}$, which is associated with the logit link $\Phi^{-1}(u) = \log \frac{u}{1-u}$, which offers additional computational advantages when the coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are assigned Gaussian prior distribution (Polson et al., 2013).

Arguably, from the point of view of interpretability, a more compelling choice would be $\Phi(r) = e^r$, for which $\log \frac{p(0+|\mathbf{x})}{p(0-|\mathbf{x})} = (\mathbf{x}'\boldsymbol{\alpha})_+$. This "log link" directly relates the coefficient vector $\boldsymbol{\alpha}$ to the logarithm of the ratio of the density values just above and below the cutoff, which is often taken as a key measure of the discontinuity size (Cattaneo et al., 2020). However, the fact that the exponential function is unbounded makes the double intractability problem mentioned above quite challenging (Tokdar et al., 2010). The logit link formulation adopted here lacks a similarly direct quantitative interpretation of $\boldsymbol{\alpha}$. Qualitatively speaking, $\log \frac{p(0+|\mathbf{x})}{p(0-|\mathbf{x})}$ is monotonically increasing in $x_j\alpha_j$, when all other covariates are held fixed. So the sign of each α_j can be interpreted as a measure of positive versus negative association of the j-th covariate with the jump size. Toward a more quantitative interpretation, one could consider the shifted logistic link $\Phi(r) = (1 + e^{a-r})^{-1}$, for a fixed constant shift a. For this choice,

$$\log \frac{p(0+|\mathbf{x})}{p(0-|\mathbf{x})} = (\mathbf{x}'\boldsymbol{\alpha})_+ + \log \frac{1+e^{r_c(\mathbf{x},0)-(\mathbf{x}'\boldsymbol{\alpha})_+-a}}{1+e^{r_c(\mathbf{x},0)-a}} \to (\mathbf{x}'\boldsymbol{\alpha})_+$$

as $a \to \infty$. Therefore the shifted logit link, with a large enough shift, mimics the model formulation and interpretation one would have with the log link. The computational method we introduce below works with any shift amount. But there is a price to pay in terms of increased computing time for larger values of a.

2.2 Parameter estimation

The data augmentation strategy of Adams et al. (2009), later refined by Rao et al. (2016) and Li et al. (2022), is based upon the concept of rejection sampling. Given a transformation $\Phi : (-\infty, \infty) \to (0, 1)$, one can simulate random numbers from a density $p(y) \propto g(y)\Phi(r(y))$ as follows:

Step 1. Draw a random number y from the density $g(y)$

Step 2. Draw a random number u from the unit interval

Step 3. Terminate and return y if $u < \Phi(r(y))$; otherwise, go back to Step 1.

In the context of our model, each observation y_i could be seen as the random number returned by this rejection sampling algorithm, applied with $r(y) = r(\mathbf{x}_i, y)$. If we had access to the rejected intermediate draws $\tilde{y}_{ij}$, $j = 1, \ldots, J_i$, a complete data likelihood could be written as

$$\tilde{L}(\boldsymbol{\theta}; \tilde{\mathbf{y}}) = p(\mathbf{y}, \tilde{\mathbf{y}}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[g_{\boldsymbol{\gamma}}(y_i)\Phi(r(\mathbf{x}_i, y_i)) \times \prod_{j=1}^{J_i} g_{\boldsymbol{\gamma}}(\tilde{y}_{ij})\{1 - \Phi(r(\mathbf{x}_i, \tilde{y}_{ij}))\} \right], \quad (2)$$

where $r(\mathbf{x}, y) = r_{\boldsymbol{\beta},\boldsymbol{\alpha},\lambda}(\mathbf{x}, y) = \sum_{k=1}^{K}(\mathbf{x}'\boldsymbol{\beta}_k)P_k(y) + (\mathbf{x}'\boldsymbol{\alpha})_+ j_\lambda(y)$. The complete data likelihood does not involve any intractable integrals and can be easily computed.

Of course, the rejection history $\tilde{\mathbf{y}}$ is not available to us but – treating it as missing data – we could attempt to approximate the joint posterior distribution $p(\boldsymbol{\theta}, \tilde{\mathbf{y}}|\mathbf{y}) \propto \tilde{L}(\boldsymbol{\theta}, \tilde{\mathbf{y}})\pi(\boldsymbol{\theta})$ by Markov chain Monte Carlo, under a suitable prior density $\pi(\boldsymbol{\theta})$. In this work, we use independent prior specifications on the individual scalar elements of $\boldsymbol{\theta}$: $\gamma_l \sim \text{Ga}(a_\gamma, b_\gamma)$, $\beta_{kj} \sim \text{N}(0, h_\beta^2)$, $\alpha_j \sim \text{N}(0, h_\alpha^2)$, $\frac{\lambda}{0.32} \sim \text{Be}(a_\lambda, b_\lambda)$. This is a fairly generic specification with the exception that (jointly) normal priors on $(\beta_{1j}, \ldots, \beta_{Kj})$, with independence across $j = 1, \ldots, p$, are convenient due to their partial conjugacy properties as discussed below.

A customized Gibbs sampling algorithm seems appealing to make draws of $(\boldsymbol{\theta}, \tilde{\mathbf{y}})$ from their joint posterior distribution. The algorithm can be initialized at an arbitrary $\boldsymbol{\theta}^{(0)}$. At iteration $t = 1, 2, \ldots$, one generates $\tilde{\mathbf{y}}^{(t)}$ from the conditional posterior $p(\tilde{\mathbf{y}}|\boldsymbol{\theta} = \boldsymbol{\theta}^{(t-1)}, \mathbf{y})$ by running the rejection sampling algorithm once for each data unit i and storing the intermediate rejected draws $\tilde{y}_{ij}$, and discarding the final accepted point; see Rao et al. (2016) for a theoretical validation. Once $\tilde{\mathbf{y}}^{(t)}$ is imputed, the complete data likelihood is used to simulate $\boldsymbol{\theta}^{(t)}$ in accordance to the conditional posterior

$$p(\boldsymbol{\theta}|\tilde{\mathbf{y}} = \tilde{\mathbf{y}}^{(t)}, \mathbf{y}) = p(\boldsymbol{\gamma}|\tilde{\mathbf{y}} = \tilde{\mathbf{y}}^{(t)}, \mathbf{y}) \times p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \lambda|\tilde{\mathbf{y}} = \tilde{\mathbf{y}}^{(t)}, \mathbf{y}).$$

The first factor in this product is the ordinary posterior density of $\boldsymbol{\gamma}$ under a beta observation model with data $(y_{ij}^\star : 1 \le j \le J_i + 1; 1 \le i \le n)$, where $y_{ij}^\star = \tilde{y}_{ij}^{(t)}$, $1 \le j \le J_i$ are the imputed points and $y_{J_i+1}^\star = y_i$ are the observed data points. One could use any number of Metropolis type algorithms to draw $\boldsymbol{\gamma}^{(t)}$ from this posterior. We adopt an independence Metropolis sampler which uses an easily computed normal approximation as the proposal distribution; see Appendix A for more details.

The second factor equals the posterior density of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \lambda)$ under a binary regression model: $z_{ij}^\star \sim \text{Bern}(\Phi(r(\mathbf{x}_i, y_{ij}^\star)))$, with observed binary response values $z_{ij}^\star = I(j = J_{i+1})$ marking whether $y_{ij}^\star$ is observed $(z_{ij}^\star = 1)$ or imputed $(z_{ij}^\star = 0)$. At this point, exploiting the fact that Φ^{-1} is the logistic link, one can introduce Pólya-Gamma latent variables $\omega_{ij} \sim \text{PG}(1, r(y_{ij}^\star, \mathbf{x}_i))$ and reinterpret the conditional posterior of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \lambda)$ as one arising from the Gaussian regression model

$$u_{ij} = \sum_{k=1}^{K}(\mathbf{x}_i'\boldsymbol{\beta}_k)P_k(y_{ij}^\star) + (\mathbf{x}_i'\boldsymbol{\alpha})_+ j_\lambda(y_{ij}^\star) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \text{N}(0, w_{ij}^{-1}), \tag{3}$$

with observed response values $u_{ij} = (z_{ij}^\star - 0.5)/w_{ij}$; see Polson et al. (2013) for more details. Next, one cycles through $j = 1, \ldots, p$, and draws $(\beta_{1j}^{(t)}, \ldots, \beta_{Kj}^{(t)})$ from their joint conditional posteriors which are Gaussian due to conjugacy. A similar conjugate update does not work for $\boldsymbol{\alpha}$ because we restrict the jump size to be non-negative. Instead, $\boldsymbol{\alpha}^{(t)}$ can be drawn (from the likelihood under (3)) by an adaptive Metropolis algorithm which can automatically tune the step size of the underlying random walk to attain an overall acceptance rate of 23%. We use the AM Algorithm of Andrieu and Thoms (2008) to carry out this task. Finally, λ is updated according to an ensemble Markov chain sampling algorithm due to Neal (2011); see Appendix A.

2.3 Model selection

A critical issue in model fitting is the choice of the polynomial basis order K. For the corporate proposal voting data, substantive differences can be seen in the estimate of $\boldsymbol{\alpha}$ when using $K = 2$ versus $K = 3$, including a potential sign change of the coefficient for analyst coverage; see Appendix/Supplementary Figure FF. Intuitively, if K was chosen too large or too small, the model could overfit or underfit the continuous piece of the density, and potentially distort the estimate of $\boldsymbol{\alpha}$ as an artifact. A data-driven choice of K could be derived with formal model selection criteria such as the well known AIC or BIC which assign a numeric score to a fitted a model by a combining of an assessment of its goodness of fit with a measure of the underlying model complexity. For Bayesian models fitted with Markov chain Monte Carlo approximation, a particularly attractive model selection criterion is the Watanabe-Akaike information criterion (WAIC; Watanabe, 2013), given by

$$\text{WAIC} \;=\; -2\sum_{i=1}^{n} \log \mathrm{E}_{\text{post}}\{p_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i)\} + 2\sum_{i=1}^{n} \mathrm{var}_{\text{post}}\{\log p_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i)\} \tag{4}$$

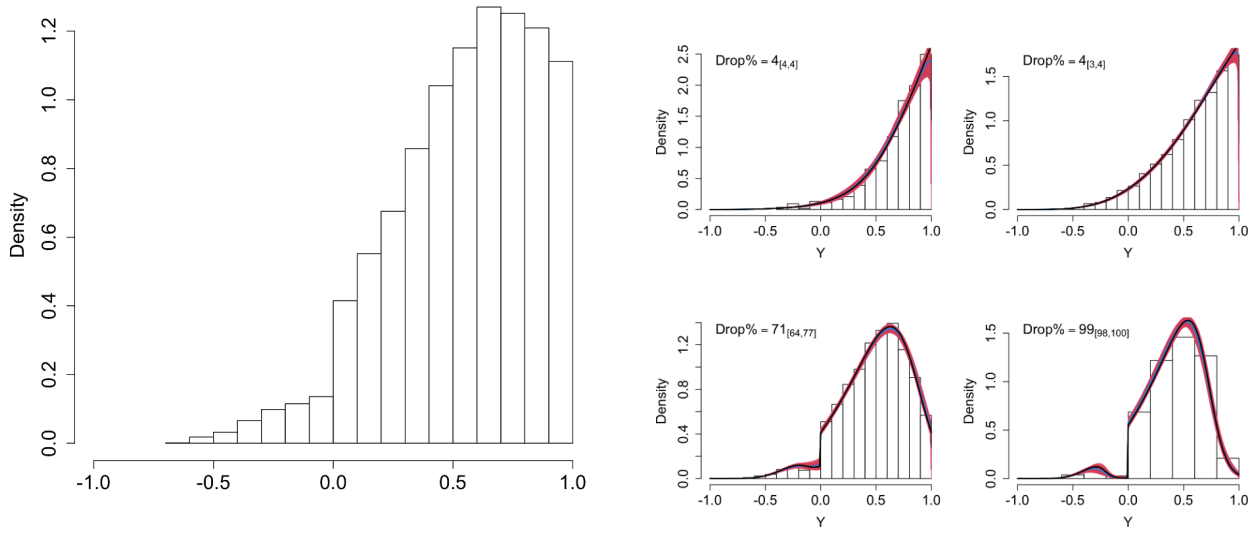
where the first term provides an assessment of model fit while the second term measures model complexity; for either term, smaller values are preferable. The WAIC can also be seen as an asymptotic limit of a leave-one-out cross-validation based scoring of model fit; see Gelman et al. (2014) for more details.

In (4), E_{post} and $\mathrm{var}_{\text{post}}$ denote expectation and variance under posterior $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. These operations can be carried out via Monte Carlo approximation once a sample of posterior draws of $\boldsymbol{\theta}$ is available. However, calculation of WAIC in doubly intractable problems poses additional challenges because (4) also depends on $p_{\boldsymbol{\theta}}(y|\mathbf{x})$ being easily computable. The data augmentation strategy we used for posterior computation appears less helpful in avoiding the normalization operation in (1); see Li et al. (2016) for cautionary tales for other latent variable models.

We are thus back to numerically evaluating the normalization integral in (1), one for each observed data point, which may be carried out with the trapezoidal rule of quadrature. However, these evaluations are needed only for the saved draws of the Gibbs sampler, whose number is typically orders of magnitude smaller than the total number of iterations the sampler is run for. In other words, even though one is forced to use quadrature for WAIC calculation, a substantial amount of computational time is saved by adopting the data augmentation strategy for posterior computation.

3 Numerical experiments

We report here results from two numerical experiments where we analyzed statistical performance of the proposed estimation method and the model selection strategy based on WAIC, with special attention to estimating the discontinuity regression coefficients $\boldsymbol{\alpha}$. In both experiments, we considered a two-dimensional covariate vector $\mathbf{x} = (1, u)$, with $u \sim \mathrm{N}(0, 1)$, and used the model specification 1 to simulate values of the endogenous variable Y. Experiment 1 used a true $K = 2$ with the true model parameters chosen in order to resemble the corporate voting proposal data that motivated the work. Experiment 2 considered a more challenging scenario with $K = 4$ and some sharper features of the density of Y away from the cutoff.



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

Figure 2: (a) Histogram of simulated y_i values in one replicate data set with $n = 20,000$ from Experiment 1. (b) True and estimated $p(y|\mathbf{x})$ (with $K = 2$) at select $\mathbf{x} = (1, u)$ values with $u \in \{-2, -0.85, 0.5, 1.75\}$. In each case, red lines are posterior samples of $p(y|\mathbf{x})$, their mean is shown in blue and the black line is the truth. The accompanying histogram, added for illustrative purposes, is based on the sub-sample with $|u_i - u| < 0.25$.

3.1 Experiment 1

In this experiment, the endogenous variable Y was simulated from a density given by (1) with $K = 2$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{tr}}$ consisting of

$$\boldsymbol{\gamma}^{\text{tr}} = (4, 1),\ \boldsymbol{\beta}_1^{\text{tr}} = (1.26, 0),\ \boldsymbol{\beta}_2^{\text{tr}} = (-1.65, -1.33),\ \boldsymbol{\alpha}_{\text{tr}} = (1, 4),\ \lambda^{\text{tr}} = 0.16.$$

The rejection sampling algorithm in Section 2.1 was used to simulate y values. A total of 100 synthetic data sets were generated, each with $n = 20,000$. Fig 2 shows data histogram from one such replicate, along with true and estimated $p(y|\mathbf{x})$ curves for select values of $\mathbf{x}$.

Density discontinuity regression analysis was run on the 100 replication data sets for each choice of $K = 2, 3, 4$, with same hyperparameters as described in Section 2.2. Posterior means and 95% credible intervals were used as point and interval estimates of all model parameters, and compared against true values (for analyses with $K > 2$, we took $\boldsymbol{\beta}_k^{\text{tr}} = (0, 0)$ for each $k > 2$). Table 1 reports, for each coice of K, the average length and coverage of the estimated intervals, the mean squared error (MSE) of the point estimates, and the percentage of data replicates where that choice of K had the smallest WAIC. Although WAIC chose the correct model ($K = 2$) most of the time, there were occasions where a larger value of K provided a better fit to the particular data set. Overall, estimates of $\boldsymbol{\alpha}$ (and also of λ) were similar across different choices of K, but on average $K = 2$ produced better point estimates and tighter intervals with comparable or better coverage. Reassuringly, the performance of the WAIC-selected model fit was comparable to that of the model fit with the oracle choice of $K = 2$.

Table 1: Results from Experiment 1 with true $K = 2$. Last three rows shows the performance of WAIC-based model selection

		$K = 2$			$K = 3$			$K = 4$		
$\boldsymbol{\theta}$	Tr	Length	MSE	Coverage	Length	MSE	Coverage	Length	MSE	Coverage
α_1	1	0.75	**0.048**	94%	0.84	0.048	95%	0.87	0.051	95%
α_2	4	1.17	**0.15**	83	1.23	0.23	77	1.24	0.22	78
β_{11}	1.26	0.66	**0.03**	90	0.81	0.08	83	0.89	0.10	81
β_{12}	0	0.27	**0.00**	97	0.40	0.02	86	0.41	0.01	92
β_{21}	–1.65	0.42	**0.02**	90	0.77	0.11	72	1.02	0.17	78
β_{22}	–1.33	0.33	**0.01**	92	0.57	0.04	81	0.61	0.04	86
β_{31}	0				0.38	0.02	79	0.60	0.05	83
β_{32}	0				0.29	0.01	83	0.47	0.02	93
β_{41}	0							0.27	0.01	91
β_{42}	0							0.28	0.00	98
γ_1	4	0.36	**0.01**	96	0.48	0.01	95	0.57	0.02	96
γ_2	1	0.07	**0.00**	96	0.09	0.00	98	0.10	0.00	99
λ	0.16	0.04	**0.00**	95	0.04	0.00	96	0.05	0.00	94
WAIC			**63%**			24%			13%	
α_1^{adp}	1	0.78	**0.044**	**96**						
α_2^{adp}	4	1.19	0.18	81						

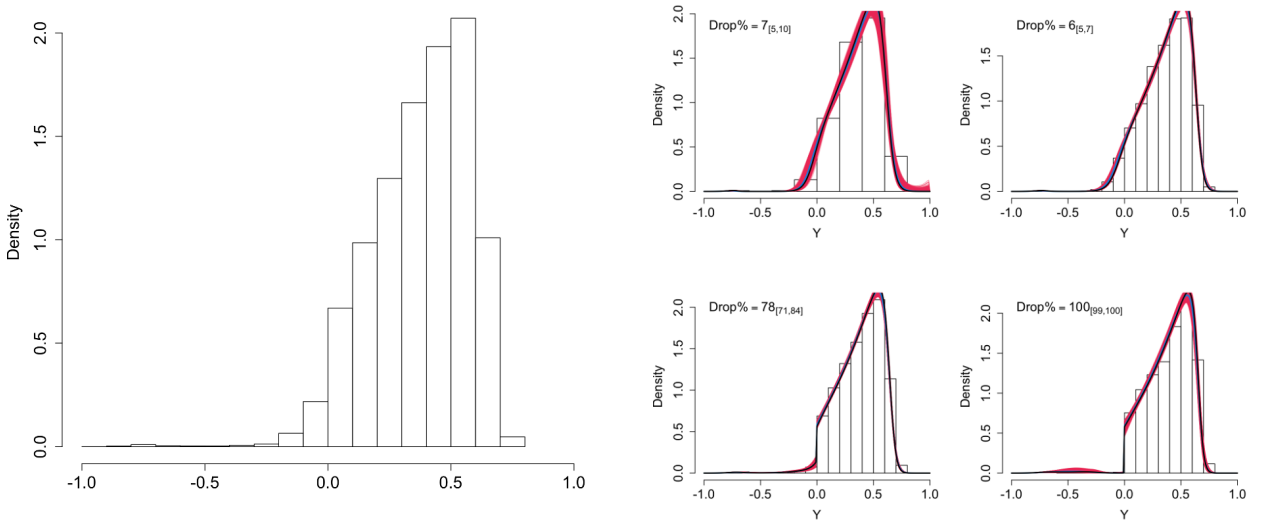
Absent any competing method, it is difficult to quantitatively evaluate how good the results of Experiment 1 truly are. Qualitatively speaking, estimation performance appears encouraging. The estimated intervals of α_j appear tight enough to recover the true sign with high accuracy. Small mean squared errors indicate good recovery of coefficient magnitudes. It does seem that estimating α_2 is harder than estimating α_1. This makes intuitive sense because α_1 relates to the jump size of the overall density discontinuity, whereas α_2 makes the more delicate but critical connection between jump size and covariates. The results here underline that density discontinuity regression is indeed a nontrivial problem and more research in this area is warranted.

3.2 Experiment 2

In a second experiment we tested model performance against a more challenging ground truth. Most model parameters, including $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and λ, were fixed at the same values as in Experiment 1, but a larger value of $K = 4$ was used to specify the true conditional densities, with corresponding $\boldsymbol{\beta}_k$ parameters chosen as

$$\boldsymbol{\beta}_1^{\text{tr}} = (-5, 0),\ \boldsymbol{\beta}_2^{\text{tr}} = (-1, -1.5), \boldsymbol{\beta}_{\text{tr}_3} = (-8, 0.5),\ \boldsymbol{\beta}_4^{\text{tr}} = (1, -1).$$

Fig 3 shows true and estimated $p(y|\mathbf{x})$ for select $\mathbf{x}$ values for one replicate data set. As can be seen, the true conditional densities show sharp drop beyond $y = 0.6$. As before, 100 replication data sets were simulated with $n = 20,000$ each. Each data set was analyzed by the density discontinuity regression method with two choices of $K \in \{3, 4\}$, with all other hyperparameters fixed as in the earlier study. Results are reported in Table 2. WAIC chose the correct model specification ($K = 4$) 100% of the time.



(a) Density histogram of simulated y_i

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

Figure 3: (a) Histogram of simulated y_i and (b) true and estimated (with $K = 3$) conditional density $p(y|\mathbf{x})$ for select $\mathbf{x}$ values, with y values trimmed to $(-0.5, 0.5)$. See Fig 2 for more details.

Although WAIC was able to identify the correct model ($K = 4$) in all cases, and the corresponding estimates of $\boldsymbol{\alpha}$ are as good as in the simpler setting of Experiment 1, it is worthwhile to consider if an alternative estimation strategy could be adopted here. The Gibbs sampling algorithm adopted here exhibits slower mixing for larger K values. It is conceivable that substantial computational and statistical gains could be made by analyzing only a subset of the data with y_i values that lie within a truncated range around the cutoff. While a trimmed data analysis could be easily embedded within the modeling and estimation approach described here, it raises an interesting question of how one would select an ideal truncation level. At this point it is not clear how one could modify the WAIC formula to address the issue that for different trimming levels, we would be analyzing different sub-samples with different sizes. We leave this question for future consideration.

4 Corporate proposal voting analysis

4.1 Data

The original corporate voting dataset comprises all management-sponsored proposals in the ISS Voting Analytics database that were initiated by U.S. firms during the period 2003 to 2015. After adding in standard company-related variables from the Compustat and I/B/E/S databases, it had 52 variables and 30,566 proposals. For each proposal, the response value is the distance of the support fraction from the threshold, which is derived by subtracting the fraction of *yes* votes on the proposal from the pass threshold. We pick five important covariates based on domain knowledge of what covariates are likely to matter in this context and the extent to which they are not missing. After cleaning the data and

Table 2: Results from Experiment 2 with true $K = 4$

		$K = 3$			$K = 4$		
$\boldsymbol{\theta}$	Tr	Length	MSE	Coverage	Length	MSE	Coverage
α_1	1	0.82	0.18	88%	0.91	0.11	83%
α_2	4	1.36	1.04	73	1.28	**0.23**	76
β_{11}	-5	0.58	0.65	22	0.71	**0.11**	68
β_{12}	0	0.54	0.36	2	0.61	**0.03**	94
β_{21}	-1	0.66	0.60	3	0.93	**0.06**	94
β_{22}	1.5	0.55	1.06	0	0.84	**0.05**	97
β_{31}	-8	0.77	1.01	36	0.88	**0.20**	53
β_{32}	0.5	0.72	0.13	71	0.83	**0.07**	87
β_{41}	1				0.64	**0.02**	96
β_{42}	-1				0.57	**0.01**	98
γ_1	4	0.53	1.41	4	0.51	**0.03**	85
γ_2	1	0.31	0.94	3	0.31	**0.02**	77
λ	0.16	0.07	0.00	96	0.10	**0.00**	81
WAIC			0%			**100**%	
α_1^{adp}	1	0.91	0.11	83%			
α_2^{adp}	4	1.28	0.23	76			

dropping observations where important covariates are missing or the response cannot be reliably measured, we are left with 19,775 data points. A description of covariates is given in Table 3.

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

Variable Name	Explanation
from.requirement.threshold	Support fraction minus the passing threshold
ISS.against	Institutional Shareholder Services (ISS) is a company providing a subscription service that provides opinions on how shareholders should vote on each proposal. This is a binary variable that is 1 when ISS recommends voting against the proposal.
analyst.coverage	How many equity analysts actively track and publish opinions on a company and its stock.
past.stock.return	Return 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	The logarithm of the book value of assets of the company

4.2 Results

Density discontinuity regression analysis was carried out with $K \in \{2, 3, 4\}$. For each model, we ran 10 parallel chains from different initializing points, with a burn-in size of $15,000$ and posterior samples of $25,000$. For $K = 4$, we found the mixing of the Markov

Table 4: Estimates and 95% intervals of $\boldsymbol{\alpha}$ and λ from corporate voting data analysis

Parameter	Estimate	95% Interval
α: Intercept	1.55	(1.33, 1.77)
α: ISS.against	0.12	(–0.03, 0.26)
α: analyst.coverage	–0.30	(–0.55, –0.05)
α: past.stock.return	0.18	(–0.01, 0.37)
α: Q	0.11	(–0.09, 0.32)
α: firm.size	–0.04	(–0.29, 0.20)
λ	0.29	(0.25, 0.31)

chains less than satisfactory, and removed this case from further consideration. Mixing was good for both $K-2$ and 3, between which WAIC clearly favored the larger order (WAIC $= -6763.95$ for $K = 3$ versus -6724.98 for $K = 2$).

Table 4 shows estimates of parameters associated with jump size from the model with $K = 3$. Our estimates indicate that analyst coverage is significantly negatively associated with jump size, whereas ISS recommendation against the proposal and past stock returns are weakly positively associated. Remaining covariates, namely the Q value and firm size, seem to have little association with density discontinuity. Estimate λ is on the larger side, suggesting that withholding may persist for proposals that land far below the passing benchmark. Among other model parameters, we note that the baseline shape parameter are estimated to be $\hat{\boldsymbol{\gamma}} = (3.94, 0.89)$. The corresponding beta density would suggest an overall average of 81.5% votes in favor, which closely matches the empirical figure of 82.8% we calculated from the data.

5 Concluding remarks

In this article we have formalized density discontinuity analysis as a regression problem and introduced a novel estimation strategy. Our method can quantitatively relate density discontinuity to measured covariates with reasonable statistical accuracy. In addition, we have provided a model selection criterion based on WAIC with moderate success in identifying the correct model complexity. In the process of developing the theory and interpreting results of the numerical experiments, we have highlighted that density discontinuity regression is not a trivial task; it has its unique set of challenges related to computation and statistical complexity.

In multiple ways, our efforts in this article barely scratch the surface of density discontinuity regression. For example, our illustrations restrict to a generic form of the jump function $j_\lambda(y)$, which may be justifiable for the voting proposal analysis. A more careful analysis should consider different form of the jump function motivated by the applied context and evaluate, through numerical experiments, whether WAIC or a similar model selection criterion could be used to pick the ideal function.

A second issue that we alluded to earlier is the question of data trimming. Some potential benefits of data trimming are as follows. When the sample size is large, the fit

of the polynomial model for the continuous piece $r_c(\mathbf{x}, y)$ could be strongly influenced by sharp density features away from the cutoff, and subsequently, may incur estimation bias around the cutoff. A biased estimation of the continuous piece could produce unreliable estimate of $\boldsymbol{\alpha}$, which is the central focus of our analysis. Trimming down the data to a shorter window around the cutoff could protect against non-local influence of model fit. Arguably an excellent local fit to trimmed data could be attained with a smaller value of K, for which Markov chain Monte Carlo approximation can attain high accuracy with fewer iterations than what is needed for models with larger K. Data trimming also reduces the time it takes to evaluate the likelihood function, thus also speeding up computation per iteration. However, our preliminary results (not reported here) suggest that comparing a trimmed data model fit to a full data model fit is far from straightforard; more work is needed in this direction.

Appendix

A Ensemble MCMC algorithm for λ

There are two facts being considered when we choose the MCMC method for λ. First, generating rejection history and Pólya-Gamma random variables requires some computation time. After this data augmentation, however, the computation of likelihood becomes much faster. This situation is sometimes called 'fast and slow variables'. Second, because we restrict λ to $(0, 0.32)$, it is easy to come up with a proper proposal distribution to sample from, e.g. a Beta distribution multiplied by 0.32. Based on these two facts, we use ensemble MCMC (Neal (2011)) for sampling λ to produce computational advantage.

Ensemble MCMC is a class of MCMC methods specifically designed for problems with fast and slow variables. A general ensemble MCMC includes three steps: 1. mapping from the original space χ to a new space χ^M; 2. performing updates on this new space; 3. mapping back to the original space. The term 'ensemble' comes from its use of an ensemble of M states. The stochastic mapping from χ to χ^M is defined as an ensemble base measure $\zeta(x_1, \ldots, x_M)$.

Neal (2011) shows that when $x_1, \ldots, x_M$ are independent and identically distributed under ζ, the ordering of states in the ensemble becomes irrelevant. The mapping from the original space to an ensemble is to combine the current state with $M-1$ states sampled from ζ. The mapping back to the original space is to randomly select a state from $x_1, \ldots, x_M$ with probabilities proportional to $\pi(x_m)/\zeta(x_m)$, where $\pi(x)$ is the target distribution and $\zeta(x)$ is the marginal distribution of all the x_k.

In our application, we can further simplify the computation by taking $\zeta(x)$ to be the same as the prior distribution of λ we set. Then probabilities are proportional to likelihood $L(\lambda|\alpha, \beta, \gamma, \omega)$, which can be computed based on result (**??**). Algorithm 1 shows steps to draw posterior sample of λ.

Algorithm 1 Ensemble MCMC algorithm for λ

Input: At iteration i, the augmented data, and the current parameter value θ_i

Output: Updated posterior sample λ_{i+1} of iteration i+1

1: **for** j in 1 to $M-1$ **do**

2: Draw $\tilde{x}_j \sim$ Beta(1,2), take $x_j^* = 0.32 * \tilde{x}_j$ and compute $w_j = L(x_j^*|\alpha_i, \beta_i, \gamma_i, \omega_i)$

3: **end for**

4: Set $x_M^* = \lambda_i$ and compute $w_M = L(x_M^*|\alpha_i, \beta_i, \gamma_i, \omega_i)$

5: Select an index m from $(1, \ldots, M)$ with probabilities proportional to $(w_1, \ldots, w_M)$

6: Update $\lambda_{i+1} = x_m^*$

B. Additional details of the corporate proposal voting analysis

Fig 4 shows 95% intervals for the coordinates of $\boldsymbol{\alpha}$ for the corporate proposal voting analysis with $K = 2$ and 3. For each analysis, ten Markov chain samplers were used, with random initial points. The fact that the ten results intervals (shown with different colors) are almost indistinguishable from one another suggests good mixing of the Markov chains, and hence highly accurate and reproducible posterior approximation.

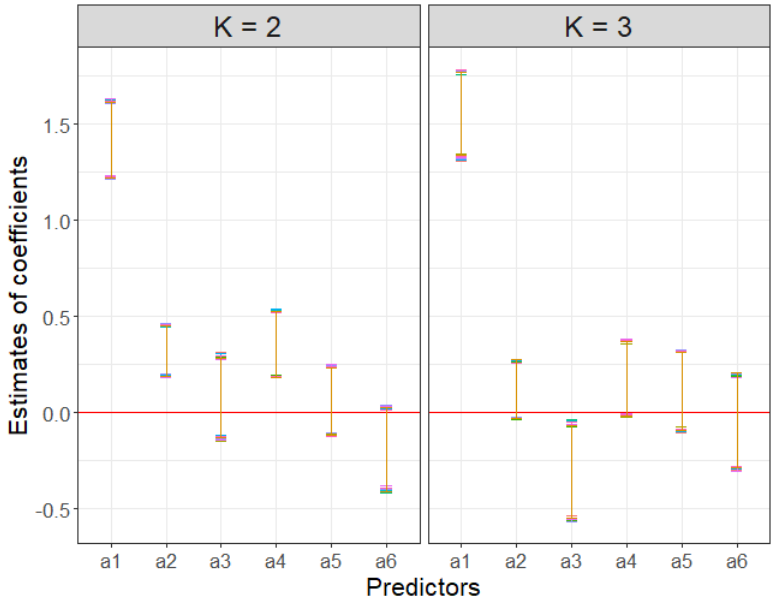


Figure 4: 95% credible intervals of $\alpha_1, \ldots, \alpha_6$ with different color denoting different chains

References

Adams, R. P., I. Murray, and D. J. C. MacKay (2009, June). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In L. Bottou and M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning (ICML)*, Montreal, pp. 9–16. Omnipress.

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

Berger, J. O., B. Liseo, and R. L. Wolpert (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical science*, 1–22.

Caimo, A. and A. Mira (2015). Efficient computational strategies for doubly intractable problems with applications to bayesian social networks. *Statistics and Computing 25*, 113–125.

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

Chen, J. (2017). Consistency of the mle under mixture models. *Statistical Science 32*(1), 47.

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.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pp. 287–302. Elsevier.

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing 24*(6), 997–1016.

Good, I. and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities. *Biometrika 58*(2), 255–277.

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

Kalbfleisch, J. and D. Sprott (1973). Marginal and conditional likelihoods. *Sankhyā: The Indian Journal of Statistics, Series A*, 311–328.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society. Series B (Methodological) 40*, 113–146.

Li, L., S. Qiu, B. Zhang, and C. X. Feng (2016). Approximating cross-validatory predictive evaluation in bayesian latent variable models with integrated is and waic. *Statistics and Computing 26*, 881–897.

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.

Murray, I., Z. Ghahramani, and D. J. MacKay (2006). Mcmc for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pp. 359–366. AUAI Press.

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*.

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.

Shen, W., S. T. Tokdar, and S. Ghosal (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika 100*(3), 623–640.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 795–810.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

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 (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics 37*(5B), 2655–2675.

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