Dirichlet process mixture models for regression discontinuity designs

Journal Title
XX(X):2–30

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Federico Ricciardi[1], Silvia Liverani[2,3] and Gianluca Baio[4]

Abstract

The regression discontinuity design is a quasi-experimental design that estimates the causal effect of a treatment when its assignment is defined by a threshold value for a continuous assignment variable. The regression discontinuity design assumes that subjects with measurements within a bandwidth around the threshold belong to a common population, so that the threshold can be seen as a randomising device assigning treatment to those falling just above the threshold and withholding it from those who fall just below.

Bandwidth selection represents a compelling decision for the regression discontinuity design analysis as the results may be highly sensitive to its choice. A few methods to select the optimal bandwidth, mainly originating from the econometric literature, have been proposed. However, their use in practice is limited.

We propose a methodology that, tackling the problem from an applied point of view, considers units' exchangeability, i.e., their similarity with respect to measured covariates, as the main criteria to select subjects for the analysis, irrespectively of their distance from the threshold. We carry out clustering on the sample using a Dirichlet process mixture model to identify balanced and homogeneous clusters. Our proposal exploits the posterior similarity matrix, which contains the pairwise probabilities that two observations are allocated to the same cluster in the Markov chain Monte Carlo sample. Thus we include in the regression discontinuity design analysis only those clusters for which we have stronger evidence of exchangeability.

We illustrate the validity of our methodology with both a simulated experiment and a motivating example on the effect of statins on cholesterol levels.

Keywords
Regression Discontinuity Design; Dirichlet Process Mixture Models; Causal Inference; Bayesian Inference

1 Introduction

The Regression Discontinuity Design (RDD) is a quasi-experimental design that estimates the causal effects of a treatment by exploiting the presence of a pre-determined treatment rule (either naturally occurring or regulated by ongoing policies). The first publication on RDD was an application in education by Thistlethwaite and Campbell[1]. Since then this framework has proved to be effective in a wide range of applications in other disciplines, including economics[2] and politics[3]. More recently there has been some interest in the RDD for epidemiology[4–6] and health and primary care applications[7–10].

The RDD can be applied in any context in which a particular treatment or intervention is administered according to a pre-specified rule linked to a continuous variable, referred to as the 'assignment' or 'forcing' variable: the treatment is then administered if the units' value for the assignment variable (X) lies above or below a certain threshold (x_0), depending on the nature of the treatment. If thresholds are strictly adhered to when assigning treatment, the design is termed *sharp*, while when this is not the case it is termed *fuzzy*.

The regression discontinuity design has become of particular interest in the definition of public health policies as it enables the use of routinely collected electronic medical records to evaluate the effects of drugs when these are prescribed according to well-defined decision rules. This is useful as government agencies such as the Food and Drug Administration (FDA) in the USA and the National Institute for Health and Care Excellence (NICE) in the UK are increasingly relying on guidelines for drug prescription in primary care. In fact

[1]Owlstone Medical, 183 Cambridge Science Park, Cambridge, UK
[2]School of Mathematical Sciences, Queen Mary University of London, UK
[3]The Alan Turing Institute, London, UK
[4]Department of Statistical Sciences, University College London, UK

Corresponding author:
Gianluca Baio, Department of Statistical Sciences, 1-19 Torrington Place, London, WC1E 7HB, UK.
Email: g.baio@ucl.ac.uk

we will use prescription of statins in the UK as our motivating example, but there is a wide range of potential applications including the prescription of anti-hypertensive drugs when systolic blood pressure exceeds 140mmHg or initiating antiretroviral therapy in patients with HIV-1 when their CD4 count has fallen to 350 cells/mm^3 or below.

The RDD can mimic a randomised experiment around the threshold and the treatment effect at the threshold can be obtained averaging the outcomes in 'small' bins in its proximity. The choice of the 'bandwidth' is an important decision for an RDD analysis since the results are highly sensitive to its choice, especially in all those cases in which the relationship between the assignment variable and the outcome, on both sides of the threshold, deviates from linearity.

In many applied studies[9,11,12], a standard strategy adopted to address the bandwidth issue is to produce local linear regression estimates obtained using data within a limited number of bandwidths (often not more than 3 or 4, sometimes defined with the guidance of experts in the field of study). Alternatively, more complex approaches can be adopted.

Historically, these methods find their roots in the econometric literature and have close connection with the non-parametric estimation of the effect for RDD. Their common rationale is that the 'optimal' bandwidth must be selected according to some criteria aimed to minimise an error term. The first proposal, by Ludwig and Miller[13], was based on a leave-one-out cross validation (CV) strategy in order to find the estimator minimising the mean integrated square error. Later, Imbens and Kalyanaraman[14] and Calonico et al.[15] demonstrated that the CV method was a potential source of bias and that it was not reliable in any case when the design is fuzzy, and hence devised two slightly different minimisation methods based on the asymptotic mean square error. Lee and Lemieux[16] give an overview of these approaches.

More recently, Local Randomization (LR) has been proposed by Cattaneo et al.[17] and used since in several applied papers[12,18,19] in an attempt to select a window around the threshold where the units can be seen as part of a randomised experiment. This approach, although motivated by a different intuition, shares a common trait with the other approaches outlined above (and further described in Section 3): they all aim at finding one bandwidth, having optimal properties under certain criteria and then use it within the RDD framework. As a consequence, they

rely on what we named 'all-or-nothing' selection mechanism: all units within the bandwidth are considered for the RDD analysis, but none of those outside.

In this paper, we propose an alternative approach to select the units to be included in a RDD analysis. Similarly to the LR method, our approach originates from a pragmatic and applied point of view, focusing on units' exchangeability, an attribute rooted in the unconfoundness assumption that guarantees that a RDD mimics a randomised control trial thanks to the similarity of the units above and below the threshold. However, our proposal has a more ambitious goal: not only do we aim at including units for the RDD analysis based on their mutual similarity and not on their proximity to the threshold, but we also want to overcome the need of an 'all-or-nothing' approach shared by all other methods existing in the literature.

Our novel proposal is motivated by the idea that that units can be grouped in an unknown yet finite number of clusters in which the available covariates are balanced among units above and below the threshold. Using a Dirichlet process mixture model (DPMM), we cluster the units using continuous and categorical covariates to account for potential sources of confounding. By quantifying the internal similarity of the clusters obtained, only units belonging to the most homogeneous clusters are then used in the RDD analysis, irrespective of their distance from the threshold. Our proposal aims to a more effective sample selection, as it searches for 'signal' in the data in farther regions from the threshold generally overlooked by the currently available bandwidth selection approaches and discards the 'noise' from data points closer to the cut-off.

The paper is organised as follows. Section 2 introduces the RDD and gives details about the Bayesian modelling framework we adopt for the analysis. Section 3 gives an overview of the current literature on bandwidth selection for regression discontinuity designs. Section 4 presents the methodological core of the paper, where we discuss the use of clustering based on Dirichlet Process Mixture Models (DPMM) within the RDD framework and Section 5 addresses the issue on cluster selection for the subsequent RDD analysis. Results on both a simulated experiment and a real dataset on the effect of statins on cholesterol level are given in Section 6. Finally a closing discussion is presented in Section 7.

2 Bayesian Inference for the Regression Discontinuity Design

In this section we introduce the basic framework and notation for the RDD. Our work is motivated by an application of the regression discontinuity design to statin prescription in primary care. In the past years other works from our broader research group have originated from the same practical application and data, every time exploring a different aspect of the RDD[8,9,20]. In the UK, according to guidelines given by the National Institution for Health and Care Excellence (NICE), statins must be prescribed to patients whose 10-year risk score of developing a cardiovascular disease, predicted using a logistic regression model with a number of clinical and lifestyle indicators as independent variables, exceeds 20%[21]. This threshold has been revised in 2014, lowering it to 10%, but we used pre-2014 data in this work and hence we applied the old cut-off value.

Using the risk score as our forcing variable ($X \in [0,1]$), a RDD analysis can assess whether binary statins treatment ($T \in \{0,1\}$) can cause a reduction in Low-Density Lipoprotein (LDL) cholesterol (our outcome, Y), evaluated at the threshold set to $x_0 = 0.20$. To complete the basic notation, let $X^c = (X - x_0)$ be the centred assignment variable and Z be the binary threshold indicator such that $Z = 1$ if the forcing variable $X \geq x_0$ and $Z = 0$ otherwise. Note that Z coincides with the observed treatment assignment variable T when the design is sharp, but when RDD is applied to health and medical data it is reasonable to expect the design to be fuzzy, and hence the two variables not to coincide. In our motivating example this can be due both to GPs not adhering to NICE guidelines and to patients failing to take statins although prescribed to do so.

It is widely known that the threshold indicator Z is a special case of binary Instrumental Variable (IV)[22]. For this reason, in order for the RDD analysis to be performed, a set of assumptions which can be derived from the IV literature must hold[9,23].

While further theoretical and technical aspects of the RDD would add very little to the scope of this paper, being extensively covered[24,25], we make use of the next subsection to provide a more detailed overview of the Bayesian modelling framework we aim to use for the the estimation of the causal effect at the threshold.

2.1 The causal effect

Motivated by our example, where GPs' prescribing behaviour may not adhere to NICE guideline, our primary focus is on fuzzy designs, hence the effect we are interested in is the *Local Average Treatment Effect* (LATE) at the threshold, defined as

$$\text{LATE} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(T|Z=1) - E(T|Z=0)}.$$

The LATE numerator is equal to the *Average Treatment Effect* (ATE). If the design is sharp, this can be proved to be an unbiased estimator for the jump in the outcome at the threshold. However, in the context of a fuzzy design, and aiming at estimating the effect at the threshold, such an estimate fails to account for the different probabilities of being treated above and below the threshold, due to the fact that the treatment assignment is greatly (but not deterministically) determined by the threshold indicator Z.

The denominator, obtained as the difference in the expected treatment probabilities above and below the threshold, scales the ATE to account for the fuzziness of the design. In our motivating example, the LATE quantifies the change in LDL cholesterol at the 10-year risk threshold of 20%. More details about the assumptions that allow the identification of the above effect under a fuzzy observational regime can be found in Constantinou and O'Keeffe[26].

2.1.1 Models for the ATE Let the index $l \in \{a, b\}$ specify whether a unit's forcing variable value lies above or below the threshold. We decided to model the outcome, i.e., LDL cholesterol, separately for $l = a$ and $l = b$ as

$$y_{il} \sim N(\mu_{il}, \sigma^2);$$
$$\mu_{il} = \beta_{0l} + \beta_{1l} x^c_{il},$$

where x^c_{il} is the centred distance of variable X from the threshold x_0 for the i-th individual belonging to l.

In our examples in Section 6, both for the simulated scenarios and for the real data analysis, the relatively large sample size reduces the impact on posterior inference of distributional assumptions, especially for σ which is likely dominated by information from observed data. With smaller samples or to ensure further robustness to prior on σ, other models are obviously possible, e.g., by considering an Half-Cauchy distribution[27].

For the regression parameters, their prior distributions are chosen to reflect plausible LDL cholesterol levels for the observed range of risk scores. Prior specifications are defined as follows

$$\begin{aligned}\beta_{0a} &= \beta_{0b} + \lambda;\\ \beta_{0b} &\sim N(3.7; \sigma^2_{0b} = 0.25);\\ \beta_{1l} &\sim N(0; \sigma^2_{1l} = 2);\\ \sigma &\sim \text{Uniform}(0, 5).\end{aligned}$$

To encode in the model some available information from the literature[28] about the effect of statins in lowering cholesterol levels, we specify the prior distribution of λ in order to be moderately informative, i.e.,

$$\lambda \sim N(-2, 1).$$

Finally the ATE is calculated as $\Delta_\beta = \beta_{0a} - \beta_{0b}$.

2.1.2 Models for the denominator of the LATE The total number of subjects treated on each side of the threshold is modelled, again separately for $l \in \{a, b\}$ as

$$\sum_{i=1}^{n_l} t_{il} \sim \text{Binomial}(n_l, \pi_l),$$

where n_l is the number of units either above or below the threshold.

Depending on the desired prior structure for (π_b, π_a), we specify two models which, analogously to those in Geneletti et al.[9], have been named *unconstrained* and *flexible difference* model.

For the unconstrained model we use vague Beta distributions, i.e.,

$$\pi_l^{unct} \sim \text{Beta}(1, 1),$$

with $l \in \{a, b\}$. Hence, we define the denominator for the LATE when using the unconstrained prior specification as

$$\Delta_\pi^{unct} = \pi_a^{unct} - \pi_b^{unct}.$$

For the flexible difference model, we impose a mild prior structure acknowledging an actual difference between the treatment probabilities above and below the threshold, defining

$$logit(\pi_a^{flex}) \sim N(2,1) \qquad \text{and} \qquad logit(\pi_b^{flex}) \sim N(-2,1).$$

These distributions keep the bulk on the prior probability of treatment distributions, above and below the threshold, reasonably separate from one another, limiting the possibility that they result to be similar, while not constraining them to have a fixed difference. For the flexible difference model, the denominator of the LATE is thus given by the difference

$$\Delta_\pi^{flex} = \pi_a^{flex} - \pi_b^{flex}.$$

***

Depending on the chosen model for the denominator we get two different local average treatment effects, namely

$$\text{LATE}^{unct} = \frac{\Delta_\beta}{\Delta_\pi^{unct}} \qquad \text{and} \qquad \text{LATE}^{flex} = \frac{\Delta_\beta}{\Delta_\pi^{flex}}$$

for the unconstrained and flexible difference model respectively.

3 A concise review of bandwidth selection methods

In recent years there has been a surge in the interest of researchers for the choice of the bandwidth, as accounted by Cattaneo and Vazquez-Bare[29] in their comprehensive review on the topic. In fact the definition of the bandwidth represents a fundamental decision for the RDD as there is both a clear link between the size of the bandwidth and the assumption of exchangeability and a trade-off with the precision of the estimates. If the bandwidth is small, units can be reasonably considered more similar to one another. If the bandwidth is too large, the converse is true, i.e., units could no longer be considered homogeneous.

In this section, we give an overview of the most prominent methods for neighbourhood selection in the literature.

3.1 Cross Validation based approach

The first approach found in the literature is based on a Cross Validation procedure as proposed by Ludwig and Miller[13*], also discussed by Imbens and Lemieux[24]. Let

$$\widehat{m}_h(X_i) = \begin{cases} \alpha_a + \beta_a X_i^c, & \text{if } X_i \geq x_0, \\ \alpha_b + \beta_b X_i^c, & \text{if } X_i < x_0 \end{cases}$$

be the predicted value, using a bandwidth equal to h, of the outcome Y regressed on the centred assignment variable X_i^c when the i-th unit is left out from the calculation. The Cross Validation criterion is defined as:

$$CV_{Y,\delta}(h) = \frac{1}{N} \sum_{i: q_{X,\delta,b} \leq X_i \leq q_{X,1-\delta,a}}^{N} \left(Y_i - \widehat{m}_h(X_i)\right)^2. \tag{1}$$

Here $\widehat{m}_h(X_i)$ is estimated using only observations on one side of X_i to mimic the fact that RDD estimates are based on regression estimates at the boundary. As a result, equation (1) is an average of boundary prediction errors. Furthermore $q_{X,\delta,b}$ and $q_{X,1-\delta,a}$ are the δ-th and $(1-\delta)$-th quantiles of the empirical distribution of X for the sub-samples 'below' and 'above' the threshold, respectively. Ludwig and Miller[30] suggest $\delta = 0.95$ to be appropriate, while other works[16,24] state that $\delta = 0.5$ represents a reasonable value, but the choice of an appropriate value varies according to the problem at hand and should be evaluated with care. The choice for the bandwidth given by this CV method is then represented by

$$h_{CV}^{opt} = \arg\min_h CV_{Y,\delta}(h).$$

This criterion leads to the bandwidth choice that minimises an approximation of the Mean Integrated Square Error (MISE):

$$\text{MISE}(h) = E\left[\int_x \left(\widehat{m}_h(x) - m(x)\right) f(x) dx\right]$$

where $m(x) = E[Y_i | X_i = x]$ and $f(x)$ is the density of the forcing variable.

In the case of a fuzzy RDD, Imbens and Lemieux[24] suggest to use the smallest bandwidth selected by two CV criteria applied separately to the outcome and to

*This is a working paper, later published as peer-reviewed article in a shortened version[30]

the treatment:

$$h_{CV}^{opt} = \min\left(\arg\min_h CV_{Y,\delta}(h), \arg\min_h CV_{T,\delta}(h)\right),$$

where T denotes the treatment received and the formulation for $CV_{T,\delta}(h)$ is similar to that in (1).

3.2 MSE expansion bandwidth selection

Both Imbens and Kalyanaraman[14] and Calonico et al.[15] criticise the CV based approach, stating that this criterion relies on fitting the entire regression line between the δ-quantile for the observation on the left and the $(1-\delta)$-quantile for those on the right, so that the result is not optimal for the problem at hand, being the aim of a RDD to estimate the effect at the threshold.

Let $\widehat{\tau}$ be the estimated effect at the threshold for the RDD, the proposal of Imbens and Kalyanaraman is based on minimising its asymptotic Mean Squared Error (MSE), i.e., $(\widehat{\tau}-\tau)^2$. Hence the MSE is defined as:

$$\text{MSE}(h) = E[(\widehat{\tau}-\tau)^2] = E[((\widehat{\mu}_a - \mu_a) - (\widehat{\mu}_b - \mu_b))^2]$$

where $\widehat{\mu}_b = \lim_{x\uparrow x_0} \widehat{m}_h(x)$ and $\widehat{\mu}_a = \lim_{x\downarrow x_0} \widehat{m}_h(x)$, i.e, the two regression estimators for the 'true' models on the two sides of the threshold, i.e., $\mu_b = \lim_{x\uparrow x_0} m(x)$ and $\mu_a = \lim_{x\downarrow x_0} m(x)$.

To overcome some issues arising when trying to minimise the MSE(h) directly, the authors use a first-order approximation around $h=0$ of the above quantity, which they term Asymptotic Mean Squared Error or AMSE(h). The optimal bandwidth is therefore:

$$h_{IK} = \arg\min_h \text{AMSE}(h) = C_K \left(\frac{\sigma_a^2(x_0) + \sigma_b^2(x_0)}{f(x_0)(m_a''(x_0) + m_b''(x_0))^2}\right)^{1/5} N^{-1/5}$$

where C_K is a constant value depending on the choice of the kernel function $K(\cdot)$; $\sigma_b^2(x_0)$ and $\sigma_a^2(x_0)$ are the left and right limit at the threshold of the variance $\sigma^2(x) = Var(Y_i|X_i=x)$; $f(x)$ is the density of the forcing variable; $m_a''(x_0)$ and $m_b''(x_0)$ are the right and left limits of the second derivative of $m(x) = E[Y_i|X_i=x]$. The authors propose a data-dependent method to estimate h_{IK} in three steps.

Calonico et al.[15] considered that both previous methods produce bandwidths that are too wide, leading to confidence intervals with poor asymptotic coverage. The authors prove that correct asymptotic coverage is reached only if the bandwidth can satisfy the bias condition $nh_n^5 \to 0$, a requirement that none of the above mentioned methods can guarantee, leading to a first order bias in the distributional approximation. As a result, the conventional confidence intervals may substantially over-reject the null hypothesis of no treatment effect.

The authors propose a bias correction to address this problem that is able to improve the performance in finite samples. The final result is a generalisation of h_{IK}, which we term h_{CCT}, which allows for higher order polynomial to be used for the inference and provides more robust confidence interval estimators.

3.3 Local Randomization

The Local Randomization (LR) approach selects a window around the cutoff in which the randomization assumption is likely to hold[17,31–33].

The rationale behind LR is that, because treatment assignment is assumed to be randomised by the threshold inside the window, the distribution of pre-intervention covariates should be the same for treated and untreated units. This observation is directly related to the non-testable unconfoundness assumption needed for the RDD to infer valid causal estimators. For the RDD framework to be useful, the distribution of these covariates for treated and untreated units should be unaffected by the treatment T within the bandwidth h but should be affected by the treatment outside the window.

To find such desired bandwidth an iterative selection method is implemented. Starting from a arbitrary 'small' bandwidth $\overline{h}_1$, for each one of the covariates, multiple tests of the null hypothesis of no effect of the treatment on the covariates is conducted and the minimum p-value taken.

If the minimum p-value obtained, p_1, is less than some pre-specified level the initial window was too large, hence one should decrease the initial window and start over. Otherwise, if p_1 is greater than the selected significance level, choose a larger window $\overline{h}_2 \supseteq \overline{h}_1$, and go back to calculate a second iteration minimum p-value, p_2. The process continues until the minimum p-value is smaller than the desired level and a final bandwidth h_{LR} is defined.

* * *

The limited literature available and the lack of an unequivocal methodology for the bandwidth selection motivates our work: in the following we develop a more general RDD framework in which the choice of the bandwidth is not required, with positive effect on our results.

4 Dirichlet Process Mixture Models

In this paper, we propose a Dirichlet process mixture model to identify units that are similar (and so will be treated as exchangeable), above and below the threshold. We propose to identify these units by exploiting the characteristics of the clusters obtained with a Dirichlet process mixture model.

The Dirichlet process mixture model is a Bayesian nonparametric method for (unsupervised) clustering and applied in a variety of areas, such as retail analysis[34], language processing and classification[35–37], medical imaging[38,39], epidemiology[40–46] and genetics[47]. The Dirichlet process was first introduced by Ferguson[48] and is defined as a probability distribution over random probability measures. The distribution of a Dirichlet process is (almost surely) discrete, in that a random sample drawn from a Dirichlet process has a non zero probability that multiple draws will have identical values. It is this discreteness property which makes the Dirichlet process ideal for clustering, as it avoids the need to determine the number of clusters a priori[49]. The basic Dirichlet process mixture model is formulated as follows:

$$\begin{aligned} w_i|\theta_i &\sim p(w_i|\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, G_0). \end{aligned}$$

The Dirichlet process models the distribution from which data $w_1, \ldots, w_n$ are drawn as a mixture of distributions, $p(w_i|\theta_i)$, where each parameter θ_i is drawn from a mixing distribution G[49]. G_0 is the base distribution, that is the prior expectation of G, i.e., $E[G] = G_0$, and the concentration parameter α acts as an inverse variance where larger values of α result in smaller variances. Posterior inference from a DPMM utilises Markov chain Monte Carlo (MCMC) posterior simulation and our implementation uses the slice sampling procedure[50]. Moreover, due to the nature of the stick-breaking construction of the Dirichlet process[51],

label-switching moves are also implemented, to prevent the slice sampler from getting stuck in local modes[52].

In this paper, the DPMM is implemented to model both continuous and discrete data using a mixture of Gaussian and categorical random variables . Let S_i be the latent allocation variable so that if $S_i = c$ then individual i is in cluster $c \in \{1, C\}$, then conditional on each cluster c, the likelihood for observable data $\mathbf{D}_i = (\mathbf{D}_i^1, \mathbf{D}_i^2)$ is

$$p(\mathbf{D}_i | S_i = c, \Theta_c) = p(\mathbf{D}_i^1 | \mu_c^{DP}, \Sigma_c) p(\mathbf{D}_i^2 | \Phi_c)$$

where $\mathbf{D}_i^1 = (D_{i,1}^1, ..., D_{i,J_1}^1)$ is the subset of the J_1 continuous random variables in $\mathbf{D}_i$ and $\mathbf{D}_i^2 = (D_{i,1}^2, ..., D_{i,J_2}^2)$ is the subset of the J_2 categorical random variables in $\mathbf{D}_i$. Note that we are assuming independence between continuous and categorical data conditional on the cluster allocations. The cluster specific parameters are given by $\Theta_c = (\mu_c^{DP}, \Sigma_c, \Phi_c)$, which are defined in detail below.

For the continuous random variables, we have

$$p(\mathbf{D}_i^1 | \mu_c^{DP}, \Sigma_c) = (2\pi)^{-\frac{J_1}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (D_i^1 - \mu_c^{DP})^\top \Sigma_c^{-1} (D_i^1 - \mu_c^{DP}) \right\}$$

and we choose $\mu_c^{DP} \sim \text{Normal}(\mu_0^{DP}, \Sigma_0)$ and $\Sigma_c \sim \text{InvWishart}(R_0, \kappa_0)$ (for each c) for our prior model to obtain a conjugate model, permitting Gibbs updates for the parameters μ^{DP} and $\mathbf{\Sigma}$.

For the discrete random variables, we have

$$p(\mathbf{D}_i^2 | \Phi_c) = \prod_{j=1}^{J_2} \phi_{S_i, j, X_{i,j}}.$$

For each individual i, $\mathbf{D}_i^2 = (D_1^2, \ldots, D_{J_2}^2)$ is a vector of J_2 locally independent discrete categorical random variables, where the number of categories for covariate $j = 1, 2, \ldots, J_2$ is R_j. Then we can write $\Phi_c = (\Phi_{c,1}, \ldots, \Phi_{c,J_2})$ with $\Phi_{c,j} = (\phi_{c,j,1}, \phi_{c,j,2}, \ldots, \phi_{c,j,R_j})$. Letting $\mathbf{a} = (\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_{J_2}})$, where $a_j = (a_{j,1}, \ldots, a_{j,R_j})$ and adopting conjugate Dirichlet priors $\Phi_{c,j} \sim \text{Dirichlet}(a_j)$, each $\Phi_{c,j}$ can be updated directly using Gibbs iterations.

As each iteration of the MCMC Gibbs sampler provides an estimate of the cluster labels, Partitioning Around Medoids (PAM) was used to obtain an overall estimate of the optimal number of clusters[53]. As the number of clusters varies

between iterations, the proposed method uses the posterior similarity matrix $\mathbf{P}$. The best clustering is selected by maximising an associated clustering score[54]. The Dirichlet process mixture model described above is available in the R package `PReMiuM`[55].

5 Cluster Ranking and Selection

Once we have identified units that are similar to one another using a Dirichlet process mixture model, above and below the threshold, we must identify the most suitable clusters for the RDD analysis. We propose to identify clusters that are *balanced* and *homogeneous*. These concepts have been extensively exploited in several branches of statistics, most notably by the Propensity Score Weighting literature[56,57], where overlap in covariates between treatment groups is a desired feature to estimate average treatment effects for sub-populations defined according to the propensity score.

A cluster is *balanced* when it has enough units on both sides of the threshold. As many small clusters are usually fully above, or below, the threshold, it is important to ensure that we consider balanced clusters for the RDD analysis. We call π_c^Z the proportion of units in cluster c with $Z_i = 1$, i.e., for which the assignment variable is greater than the threshold x_0. We then empirically set a constant value ζ , deeming a cluster *balanced* if the proportion π_c^Z falls within an acceptable range, i.e., $\frac{1}{\zeta} \leq \pi_c^Z \leq \frac{\zeta - 1}{\zeta}$. Empirical evidence, based on our experience, suggests that at least 10% of the units in a cluster should be treated, so $\zeta \leq 10$, and for symmetry we suggest $\zeta > 2$. These settings will guarantee that the acceptable range for π_c^Z is always centered around 0.5 and its width cannot exceed 0.8. Finally, we discard unbalanced clusters, leaving us with $C' \leq C$ clusters. For $c' = 1, \ldots, C'$ let $\mathcal{K}' = \{K_1, \ldots, K_{C'}\}$ with $K_{c'} \subseteq \mathcal{V}$ be the clustering set where the unbalanced clusters have been removed so that $n_{c'} = |K_{c'}|$ is the number of units in cluster c'.

A cluster is *homogeneous* (or compact) when the observations within it are very similar to one another. However, modelling with a mixture model does not always result in clusters of similar observations. For example, a Gaussian mixture model with a fully flexible covariance matrix may incur in large within-cluster dissimilarities compared to a model in which covariance matrices are assumed to be equal or spherical: observations that are modelled well by a common probability

distribution are not necessarily close. For example, in the case of a 2-dimensional Gaussian distribution with a high correlation, the maximum distance between the further observations can be significant. Generally, the mixture model does not come with implicit conditions that ensure the separation of clusters[58]. Therefore, we employ the Dirichlet process mixture model to exploit its flexibility, but we must take a close look to the homogeneity of each cluster.

We propose to rank clusters based on their homogeneity. The concept of homogeneity is widely explored in the clustering literature[59] and relies on the idea that if properly identified, units in a cluster must have a cohesive structure. The most straightforward way to formalise that all objects within a cluster should be similar to each other is the average within-cluster distance, a commonly used index for cluster internal validation[60]. We employ a version of this within-cluster index based on the posterior similarity matrix $\mathbf{P}$ obtained post-processing the output of the Dirichlet process mixture model. The values in $\mathbf{P}$ are the pairwise probabilities that two observations are allocated to the same clusters in the MCMC sample. As such, adapting the definition of dissimilarity from Henning[60], we can define a similarity function $s : \mathcal{V}^2 \longmapsto \mathbb{R}_0^+$ so that $s(v_1, v_2) = s(v_2, v_1) \geq 0$ and $s(v_1, v_1) = 1$, where v_1 and v_2 are elements from $\mathcal{V}$, the space of observations that we are clustering. This similarity function can be used to compute the within-cluster homogeneity.

Let $p_{l,v}$ be the elements of the similarity matrix $\mathbf{P}$. For each cluster this within-cluster homogeneity index can be calculated as:

$$I_{c'} = \frac{2}{n_{c'}(n_{c'} - 1)} \sum_{l=1}^{n_{c'}} \sum_{v \leq l}^{n_{c'}} p_{l,v}.$$

A lower within-cluster index is an indicator of a more homogeneous cluster, with 0 being the minimum value for $I_{c'}$. We exploit this measure of homogeneity to rank the clusters from the least homogeneous to the most homogeneous. We relabel the index as $I_{(c')}$ for $c' = 1, \ldots, C'$ such that $I_{(1)} < I_{(2)} < \ldots < I_{(C')}$.

Among the balanced clusters, we propose to use homogeneity to select the clusters to include in our model. We propose the following four criteria.

1. We include clusters until the relative difference between the homogeneity for the c'-th and $(c'+1)$-th ordered clusters is within a 10% margin, that

is, all ordered clusters from 1 to c' such that

$$\frac{I_{(c'+1)} - I_{(c')}}{I_{(c')}} < 0.10$$

for $c' = 1, \ldots, C'$. We refer to this criteria as *inc10*.

2. We include the first quartile of the balanced clusters, that is, all clusters c' with

$$I_{(\lceil h \rceil)} \quad \text{such that} \quad h \leq C'/4.$$

We refer to this method as *c25*.

3. We include clusters starting from the most homogeneous until the sample includes at least half of the units from the entire cohort, that is, all clusters c' with $c' = 1, \ldots, C'$ such that

$$\sum_{c'=1}^{C'-1} n_{(c')} < N/2 \quad \text{and} \quad \sum_{c'=1}^{C'} n_{(c')} \geq N/2$$

where $n_{(c')}$ is the cardinality of the c'-th cluster, ordered according to the homogeneity index $I_{(s)}$. We refer to this criteria as *n50*.

4. We named to this final criteria as *n25* as it is similar to *n50*, but only considering one quarter of the units from the entire cohort, that is, all clusters c' with $c' = 1, \ldots, C'$ such that

$$\sum_{c'=1}^{C'-1} n_{(c')} < N/4 \quad \text{and} \quad \sum_{c'=1}^{C'} n_{(c')} \geq N/4.$$

The four strategies detailed above define four (possibly) different sub-samples of the partition obtained applying a Dirichlet process mixture model as in Section 4. RDD analysis, as detailed in Section 2, is hence performed for each of the sub-samples of units irrespective of their distance from the threshold (x_0). Note that we empirically observed that the *inc10* strategy tends to include fewer observations. This is due to the fact that it is the only method that does not provide any guarantee on the number of observations selected: *c25*, *n50* and *n25* are based on quartiles or sample size, while inc10 is only based on the homogeneity of the clusters and its rate of decrease.

6 Applications and results

We make use of our methodology for an application to primary care prescription: according to the guidelines given by the National Institute for Health and Care Excellence (NICE) between 2008 and 2014, statins should have been prescribed in the UK to patients with 10-year cardiovascular disease (CVD) risk scores, calculated via the so called Framingham Risk Score[61], in excess of 20%. To illustrate our methodology and check its performance, we use statins prescriptions data from The Health Improvement Network (THIN - `www.the-health-improvement-network.com`) a large primary care database that provides anonymised longitudinal general practice data on patients' diagnostic and prescribing records from more than 500 general practices across the UK. The database is broadly representative of the UK population[62]. Access to the dataset can be obtained by contacting the network.

In the following Sections we will present results obtained using our methodology both on a realistically simulated dataset (Section 6.1) and on a subset of data from THIN patients (Section 6.2). The simulated experiment makes a formal comparison between different methods (i.e., our DPMM clustering based approach and other relevant bandwidth selection criteria), while the real-data application showcases how our methodology can be useful in practice. In both cases, the values of three key covariates are used to cluster units with our DPMM: age, systolic blood pressure and high-density lipoprotein (HDL) cholesterol. With the same data, we have obtained results of RDD analyses using established bandwidth selection methods: those based on MSE (i.e., methods originating from Imbens and Kalyanaraman[14] and Calonico et al.[15], IK and CCT for short, respectively) and Local Randomization (LR) as detailed in Sections 3.2 and 3.3 as well as two arbitrarily selected windows, i.e., bandwidth of width 0.05 and 0.1 on each side of the threshold. Appropriate functions from R packages `rdd`, `rdrobust` and `rdlocrand` are used to estimate h_{IK}, h_{CCT} and h_{LR} respectively.

6.1 Simulated example

For this example we followed the same approach as Geneletti et al.[9] and used simulated data originated from the THIN database (details about the simulation algorithm can be found on the supplementary material of that paper). In particular data are obtained under a simulation scenario in which the risk score

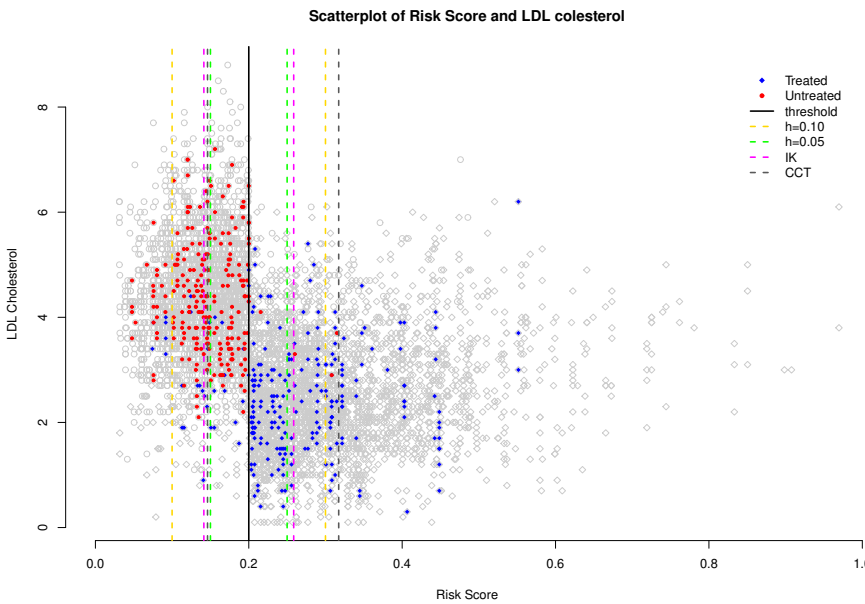

Figure 1. Scatterplot of 10-year CVD risk score vs. LDL cholesterol for one of the realistically simulated datasets, highlighting the units selected for the RDD analysis using the 'c25' strategy (treated (blue) and untreated (red)), compared with other bandwidth selection methods (LR bandwidths are not depicted as they are too close to the threshold line).

is a strong instrument for the treatment, the treatment effect size is equal to -2 and there is low level confounding. Both statins treatment status and the LDL cholesterol outcome are simulated to mimic realistic values.

We have simulated 100 datasets and for each of them, separately, we clustered the units using the DPMM approach. Then we selected the most homogeneous clusters based on the four criteria detailed in Section 5. The range for acceptable assignment probability for each cluster is $\frac{1}{10} \leq \pi_c^Z \leq \frac{9}{10}$, i.e., $\zeta = 10$. These boundaries are set in order to account for the fact that in most of the clusters the assignment probabilities are not very well balanced between observations below and above the threshold. Our aim is thus to define a reasonable way to discard extreme, ineligible clusters while, at the same time, preventing a too drastic exclusion of most of them. Finally we performed RDD Bayesian analysis and combined the results to obtain LATE^{unct} and LATE^{flex}.

Table 1. Results for the simulated example.

	method	Median	Mean	Lower	Upper
LATE^{flex}	inc10	-2.16	-2.19	-3.10	-1.47
LATE^{unct}		-2.26	-2.42	-3.55	-1.41
LATE^{flex}	c25	-1.96	-1.97	-2.28	-1.66
LATE^{unct}		-1.97	-1.97	-2.28	-1.67
LATE^{flex}	n50	-2.03	-2.03	-2.18	-1.89
LATE^{unct}		-2.03	-2.03	-2.18	-1.89
LATE^{flex}	n25	-1.96	-1.96	-2.16	-1.77
LATE^{unct}		-1.96	-1.96	-2.16	-1.77
LATE^{flex}	LR	-1.44	-1.46	-2.79	-0.27
LATE^{unct}		-1.54	-1.62	-3.52	-0.27
LATE^{flex}	CCT	-2.05	-2.05	-2.21	-1.90
LATE^{unct}		-2.05	-2.05	-2.21	-1.90
LATE^{flex}	IK	-2.08	-2.08	-2.24	-1.93
LATE^{unct}		-2.08	-2.09	-2.24	-1.93
LATE^{flex}	$h = 0.10$	-2.10	-2.10	-2.25	-1.94
LATE^{unct}		-2.10	-2.10	-2.25	-1.95
LATE^{flex}	$h = 0.05$	-2.10	-2.10	-2.27	-1.93
LATE^{unct}		-2.10	-2.10	-2.27	-1.93

Figure 1 gives a visual representation of how units are selected according to different bandwidth methods compared with our DPMM framework combined with the *c25* criteria: solid red dots and blue diamonds represent the selected units out of the whole initial sample, represented using void grey markers. Vertical lines show the bandwidths selected with some of the methods described in Section 3. Note that LR bandwidths are not shown to avoid confusion, as they are too close to the threshold.

Table 1 and Figure 2 show the results of these scenarios. It is worth noticing that flexible and unconstrained estimators give very similar results. Among the four cluster selection strategies we propose, *c25*, *n50* and *n25* all show a reduced or similar bias than those obtained using other established methods - i.e., CCT, IK, LR and arbitrarily-selected fixed-width bandwidths. Our proposed DPMM clustering method is performing as well existing methods in certain

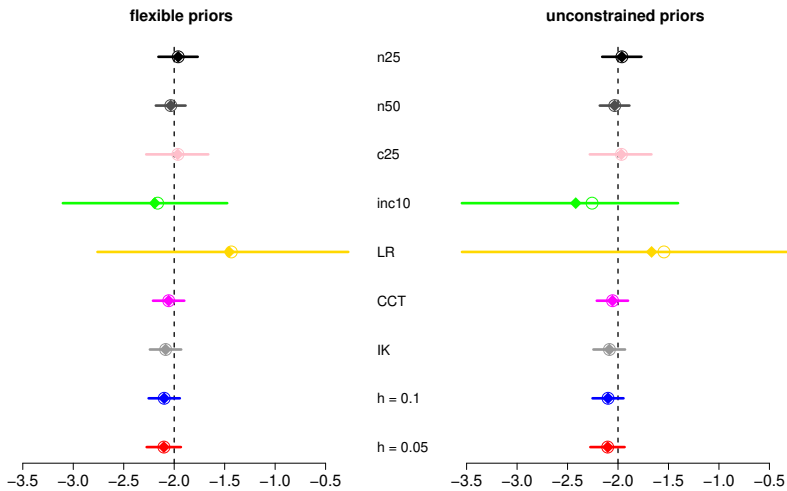

Figure 2. Comparison of results for the simulated example.

circumstances, so it can prove valuable in all those RDD applications where exchangeability is regarded as a key feature and where traditional methods do not offer viable solutions to tackle it. On the other hand, results for strategy *inc10* are considerably less reliable. Precision of all estimators is comparable for all strategies but *inc10*, which shows wider credible intervals.

6.2 Real data - Statins prescription in the UK

In this second example, we considered a subset of patients from THIN: male individuals aged from 50 to 70 who had not previously received a statin prescription nor suffered from a CVD event and for whom the Framingham risk score was recorded by the GP during the time between 1 January 2007 and 31 December 2008. We further restricted the analysis to non-diabetic and non-smoking patients, so that the total number of units is 1386.

Figure 3 shows why we believe an RDD is appropriate for the data at hand. On the left-hand side, the scatterplot highlights a discontinuity for the LDL level, which is visibly higher for data points with a risk score lower than 0.2 and drops sensibly when the risk score is higher than the threshold. The decrease is also shown by the black dots representing mean values within equally spaced bins. On

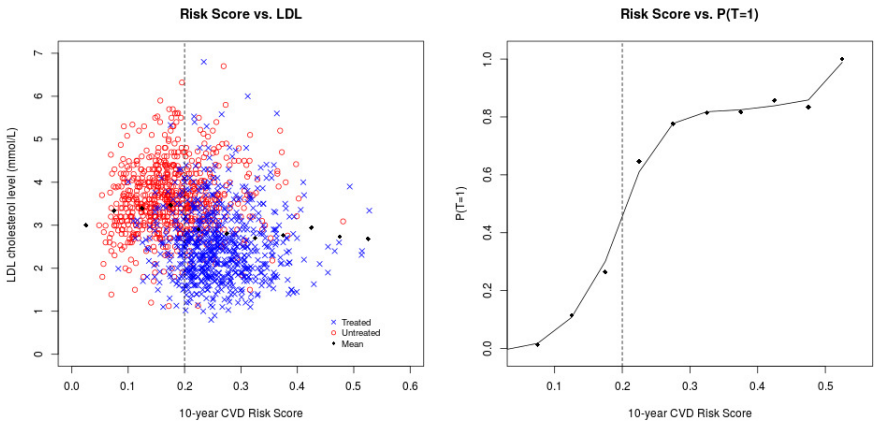

Figure 3. The left-hand plot shows 10-year CVD risk score vs. LDL cholesterol for treated (blue) and untreated (red), and the mean cholesterol lever within some equally spaced bins (black); the right-hand side plot shows risk score vs. the estimated probability of treatment, within the same bins. The dashed line indicates the threshold of 0.2.

the right-hand side, the probability of being assigned with the statins treatment displays a characteristic S-shape, with a rapid increase when crossing the risk score threshold and more stable values far from it. Moreover, there is substantial fuzziness in the data around the threshold, suggesting that LATE estimators are appropriate in this setting, and we calculated them using the Bayesian methods detailed in Section 2 and relying on both flexible and unconstrained prior specifications. The data also shows that the assumption of monotonicity, required for an RDD, is tenable. The monotonicity assumption states that no decision-maker systematically defies the guidelines - i.e., no GPs would prescribe statins only to those patients with their risk score lower than the threshold and withhold treatment only to those with a risk greater than 0.2. [8].

For the clustering selection process, the range for acceptable assignment probability for each cluster is set to $\frac{1}{10} \leq \pi_c^Z \leq \frac{9}{10}$, i.e., $\zeta = 10$. Similarly to Figure 1 for the simulated example, Figure 4 depicts unit selection according to different established bandwidth methods (i.e., IK, CCT, LR and arbitrarily selected) compared with our proposed framework based on the DPMM and the *c25* strategy. Once again it is possible to notice how the solid red and blue dots and diamonds are picked across most of the Risk Score range when exchangeability is the privileged criteria for unit selection.

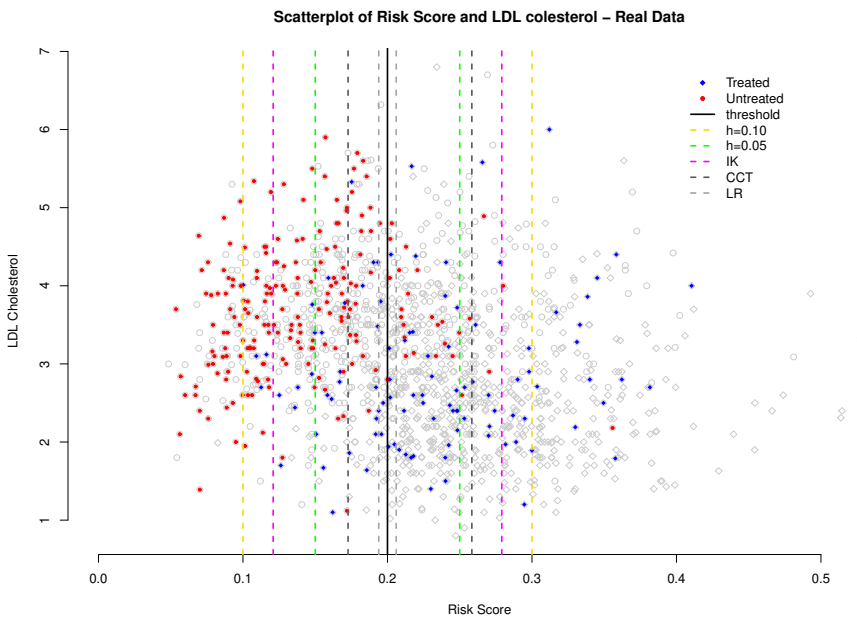

Figure 4. Scatterplot of 10-year CVD risk score vs. LDL cholesterol for Real case, highlighting the units selected for the RDD analysis using the 'c25' strategy (treated (blue) and untreated (red)), compared with other bandwidth selection methods.

Table 2 and Figure 5 show the results. Obviously there is no real value to compare the results of the estimators with, but there are a few aspects of interest nonetheless. All our DPMM estimators, including *inc10*, produce similar results, irrespective of which cluster selection method is used, with *n50* and *n25* strategies both producing more precise estimates. It is also interesting to note how, in this case, the LR method produces very wide credible intervals for both LATE^{unct} and LATE^{flex}, as this method is not able to pick a large enough subset of similar unit, being constrained to limit the search within nested windows. Results from both MSE based and arbitrary-selected bandwidths appear substantially different: CCT estimators are less precise that the DPMM based ones, and only $h = 0.10$ produces results similar, in median, to those obtained applying our DPMM and cluster selection.

Table 2. Results for example based on real data.

	method	**Median**	**Mean**	**Lower**	**Upper**
LATE^{flex}	inc10	-1.01	-1.03	-1.58	-0.57
LATE^{unct}		-1.10	-1.10	-1.74	-0.49
LATE^{flex}	c25	-1.02	-1.04	-1.59	-0.55
LATE^{unct}		-1.09	-1.09	-1.67	-0.49
LATE^{flex}	n50	-0.95	-0.96	-1.32	-0.67
LATE^{unct}		-0.97	-0.97	-1.30	-0.68
LATE^{flex}	n25	-1.12	-1.12	-1.49	-0.76
LATE^{unct}		-1.14	-1.14	-1.56	-0.71
LATE^{flex}	LR	-2.07	0.53	-21.36	26.04
LATE^{unct}		-1.79	3.52	-28.67	57.38
LATE^{flex}	CCT	-1.53	-1.56	-2.21	-0.95
LATE^{unct}		-1.55	-1.58	-2.44	-0.94
LATE^{flex}	IK	-1.17	-1.18	-1.63	-0.82
LATE^{unct}		-1.19	-1.19	-1.61	-0.83
LATE^{flex}	$h = 0.10$	-1.04	-1.05	-1.40	-0.74
LATE^{unct}		-1.09	-1.09	-1.42	-0.72
LATE^{flex}	$h = 0.05$	-1.39	-1.40	-1.98	-0.94
LATE^{unct}		-1.42	-1.43	-2.00	-0.99

7 Conclusions

We have proposed a novel, data-driven approach to deal with the bandwidth selection issue for the regression discontinuity design from a different perspective than those adopted in the available literature. Our approach originates from the idea that what matters the most in a regression discontinuity design is the exchangeability of the units included in the analysis, i.e., their homogeneity with respect to know observable characteristics. Following this rationale, it is reasonable to believe that subgroups of units might share common characteristics irrespective of their distance from the threshold. This, we believe, represents the most appealing aspect of this framework: instead of relying on the 'all-or-nothing' approach, which is implicit with any of the currently available bandwidth selection methodologies, we propose a tool that is capable of using all the

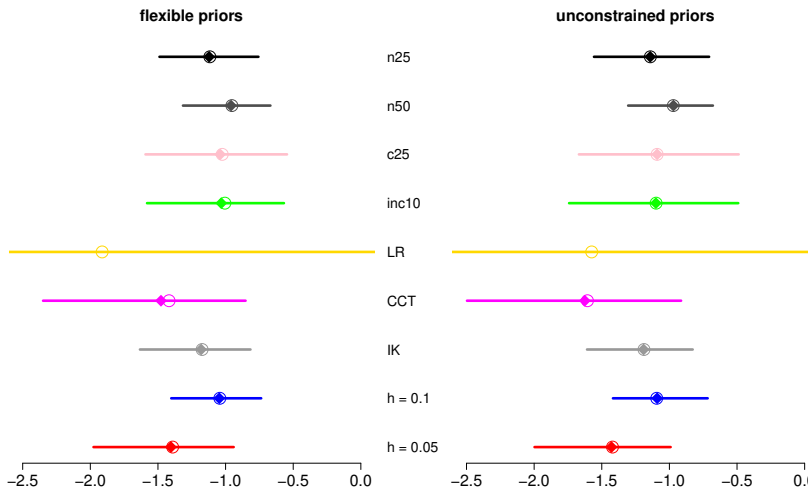

Figure 5. Comparison of results for the Real Case.

available information from all the individuals showing homogeneous covariates and balanced forcing variable.

Furthermore, when compared with the Local Randomization method which is similar in principle to ours, our DPMM clustering approach has the merit of tackling exchangeability more directly: while the former tests the null hypothesis of no effect of the treatment on each observed confounder separately in a univariate way, the latter, relying on clustering methods, evaluates the homogeneity of the considered covariates in a joint, more comprehensive approach.

The results of the RDD analysis using our DPMM clustering framework, especially in combination with *c25* cluster selection strategy, compared favourably in terms of bias with those obtained following other bandwidth selection approaches, i.e., CCT and IK (methods that are specifically designed to minimise the bias of the causal estimator), LR and with the arbitrarily-selected fixed-width bandwidths.

We are aware of the limitations to our approach. In particular we acknowledge the issue that, due to the complexity of the DPMM which involves the estimation of a latent clustering structure, our analysis is more time consuming than those based on other bandwidth selection methods, an issue that is amplified as the number of clustering covariates increases. Due to label switching and the lack of a

specific parameter to target, it is also hard to assess Bayesian DPMM convergence with the usual MCMC diagnostics. We remain convinced that these limitations are a reasonable price to pay in order to be able to overcome the 'all-or-nothing' bandwidth approach. A further limitation is the fact that our method relies on the availability of observed data or known confounders, although this issue is not exclusive of our approach as it is shared with Local Randomization method as well.

As a final remark, we think it is useful to note that we are not advocating the indiscriminate use of our methodology in any given RDD analysis. Expert assessment of any application and a proper evaluation of the plausibility of the RDD assumptions must always constitute the ground for subsequent analyses. Also, availability of covariates data and their role as potential confounders must be assessed beforehand. A certain degree of subjectivity remains in the choice of value ζ, for which an assessment of the balance of the forcing variable has been proposed as a way to deal with clusters with unbalanced representation on both sides of the threshold, but the magnitude of the reasonably allowed unbalance represents an application-specific feature and it is left for the practitioner to be determined.

Rather than being a 'one-size-fits-all' tool, our proposal offers an alternative approach to identify the units to be included in the RDD analysis in a more targeted way than the bandwidth selection methods currently available. Thoughtful use of our proposed DPMM clustering framework can prove valuable in all those RDD applications where exchangeability is regarded as a key feature and where traditional methods do not offer viable solutions to tackle it.

Acknowledgements

This research has been funded by a UK MRC grant MR/K014838/1. Approval for this study was obtained from the Scientific Review Committee of THIN in August 2014.

References

1. Thistlethwaite DL and Campbell DT. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology* 1960; 51: 309–317.

2. Cook TD. Waiting for life to arrive: a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics* 2008; 142(2): 636–654.
3. Lee DS. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* 2008; 142(2): 675–697.
4. Bor J, Moscoe E, Mutevedzi P et al. Regression Discontinuity Designs in Epidemiology: causal inference without randomized trials. *Epidemiology* 2014; 25(5): 729–737.
5. Deza M. The effects of alcohol on the consumption of hard drugs: Regression discontinuity evidence from the national longitudinal study of youth, 1997. *Health Economics* 2015; 24(4): 419–438.
6. Petersen I, Nicolaisen S, Ricciardi F et al. Impact of being eligible for type 2 diabetes treatment on all-cause mortality and cardiovascular events: regression discontinuity design study. *Clinical Epidemiology* 2020; Volume 12: 569–577.
7. Linden A and Adams JL. Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice* 2012; 18(2): 317–325.
8. O'Keeffe AG, Geneletti S, Baio G et al. Regression discontinuity designs: an approach to the evaluation of treatment efficacy in primary care using observational data. *BMJ* 2014; 349.
9. Geneletti S, O'Keeffe AG, Sharples LD et al. Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in Medicine* 2015; 34: 2334–2352.
10. Bor J, Fox MP, Rosen S et al. Treatment eligibility and retention in clinical HIV care: A regression discontinuity study in South Africa. *PLOS Medicine* 2017; 14(11): 1–20.
11. Broockman DE. Do congressional candidates have reverse coattails? Evidence from a regression discontinuity design. *Political Analysis* 2009; 17(4): 418–434.
12. Li F, Mattei A and Mealli F. Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics* 2015; 9(4): 1906–1931.
13. Ludwig J and Miller DL. Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, 2005. Discussion Paper No. 1311-05.

14. Imbens G and Kalyanaraman K. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies* 2012; 79(3): 933–959.
15. Calonico S, Cattaneo MD and Titiunik R. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 2015; 82(6): 2295–2326.
16. Lee DS and Lemieux T. Regression discontinuity designs in economics. *Journal of economic literature* 2010; 48(2): 281–355.
17. Cattaneo MD, Frandsen B and Titiunik R. Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference* 2015; 3(1): 1–24.
18. Skovron C and Titiunik R. A practical guide to regression discontinuity designs in political science. *American Journal of Political Science* 2015: 1–36.
19. Mattei A and Mealli F. Regression discontinuity designs as local randomized experiments. *Observational Studies* 2016; 2: 156–173.
20. Geneletti S, Ricciardi F, O'Keeffe AG et al. Bayesian modelling for binary outcomes in the regression discontinuity design. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2019; 183(3): 983–1002.
21. NICE. *Quick reference guide: Statins for the prevention of cardiovascular events.* NICE, 2008.
22. Didelez V, Meng S and Sheehan NA. Assumptions of IV methods for observational epidemiology. *Statistical Science* 2010; 25(1): 22–40.
23. Hahn J, Todd P and Van Der Klaauw W. Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica* 2001; 69: 201–209.
24. Imbens G and Lemieux T. Regression discontinuity designs - A guide to practice. *Journal of Econometrics* 2008; 142(2): 615–635.
25. Van Der Klaauw W. Regression-discontinuity analysis: A survey of recent developments in economics. *Labour* 2008; 22(2): 219–245.
26. Constantinou P and O'Keeffe AG. Regression discontinuity designs: A decision theoretic approach. 2015; ArXiv preprint arXiv:1601.00439.
27. Polson NG and Scott JG. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* 2012; 7(4): 887–902.
28. Ward S, Jones L, Pandor A et al. A systematic review and economic evaluation of statins for the prevention of coronary events. *Health Technology Assessment* 2007; 11(14): 1–160.
29. Cattaneo MD and Vazquez-Bare G. The choice of neighborhood in regression discontinuity designs. *Observational Studies* 2017; 3(2): 134–146.

30. Ludwig J and Miller DL. Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly journal of economics* 2007; 122(1): 159–208.
31. Sekhon JS and Titiunik R. On interpreting the regression discontinuity design as a local experiment. In Cattaneo MD and Escanciano J (eds.) *Regression Discontinuity Designs*. Bingley, U.K.: Emerald Group, 2017. pp. 1–28.
32. Cattaneo MD, Titiunik R and Vazquez-Bare G. Inference in regression discontinuity designs under local randomization. *The Stata Journal* 2016; 16(2): 331–367.
33. Calonico S, Cattaneo MD, Farrell MH et al. Regression discontinuity designs using covariates. *The Review of Economics and Statistics* 2019; 101(3): 442–451.
34. Pitkin J, Ross G and Manolopoulou I. Dirichlet process mixtures of order statistics with applications to retail analytics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2019; 68(1): 3–28.
35. Crook N, Granell R and Pulman S. Unsupervised classification of dialogue acts using a Dirichlet process mixture model. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. USA: Association for Computational Linguistics. ISBN 9781932432640, pp. 341–348.
36. Dreyer M and Eisner J. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11, USA: Association for Computational Linguistics. ISBN 9781937284114, pp. 616–627.
37. Zhang J, Ghahramani Z and Yang Y. A probabilistic model for online document clustering with application to novelty detection. In *Advances in neural information processing systems*. pp. 1617–1624.
38. da Silva A. A Dirichlet process mixture model for brain MRI tissue classification. *Medical Image Analysis* 2007; 11(2): 169–182.
39. Wachinger C and Golland P. Atlas-based under-segmentation. In (eds.) *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2014; Springer. ISBN 978-3-319-10404-1, pp. 315–322.
40. Hastie DI, Liverani S, Azizi L et al. A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. *BMC Medical Research Methodology* 2013; 13(1): 129.
41. Molitor J, Brown IJ, Chan Q et al. Blood pressure differences associated with optimal macronutrient intake trial for heart health (OMNIHEART)–like diet compared with a typical American diet. *Hypertension* 2014; 64(6): 1198–1204.

42. Pirani M, Best N, Blangiardo M et al. Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environment International* 2015; 79: 56–64.
43. Mattei F, Liverani S, Guida F et al. Multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the ICARE study. *Occupational and environmental medicine* 2016; 73(6): 368–377.
44. Liverani S, Lavigne A and Blangiardo M. Modelling collinear and spatially correlated data. *Spatial and Spatio-temporal Epidemiology* 2016; 18: 63–73.
45. Coker E, Liverani S, Ghosh JK et al. Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. *Environment International* 2016; 91: 1–13.
46. Coker E, Liverani S, Su JG et al. Multi-pollutant modeling through examination of susceptible subpopulations using profile regression. *Current Environmental Health Reports* 2018; 5(1): 59–69.
47. Papathomas M, Molitor J, Hoggart C et al. Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene x gene patterns. *Genetic Epidemiology* 2012; 36(6): 663–674.
48. Ferguson TS. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1973: 209–230.
49. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000; 9(2): 249–265.
50. Kalli M, Griffin JE and Walker SG. Slice sampling mixture models. *Statistics and computing* 2011; 21(1): 93–105.
51. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica sinica* 1994; 4(2): 639–650.
52. Hastie DI, Liverani S and Richardson S. Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and computing* 2015; 25(5): 1023–1037.
53. Kaufman L and Rousseeuw P. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2006.
54. Molitor J, Papathomas M, Jerrett M et al. Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics* 2010; 11(3): 484–498.
55. Liverani S, Hastie DI, Azizi L et al. `PReMiuM`: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. *Journal of Statistical Software* 2015;

64(7).

56. Crump RK, Hotz VJ, Imbens GW et al. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009; 96(1): 187–199.
57. Li F, Morgan KL and Zaslavsky AM. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* 2018; 113(521): 390–400.
58. Hennig C and Liao TF. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2013; 62(3): 309–369.
59. Everitt BS, Landau S, Leese M et al. *Cluster Analysis*. 5th ed. Wiley, 2011.
60. Hennig C. Cluster validation by measurement of clustering characteristics relevant to the user. In Skiadas CH and Bozeman JR (eds.) *Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining*. John Wiley & Sons, 2019. pp. 1–24.
61. D'Agostino RB, Vasan RS, Pencina MJ et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008; 117(6): 743–753.
62. Bourke A, Dattani H and Robinson M. Feasibility study and methodology to create a quality evaluated database of primary care data. *Journal of Innovation in Health Informatics* 2004; 12(3): 171–177.