Biometrika (2020), **107**, 1, *pp.* 173–190
Printed in Great Britain
doi: 10.1093/biomet/asz064
Advance Access publication 5 December 2019

A conditional density estimation partition model using logistic Gaussian processes

By R. D. PAYNE

Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285, U.S.A.
richard.payne@lilly.com

N. GUHA

Department of Mathematical Sciences, University of Massachusetts Lowell, 220 Pawtucket St, Lowell, Massachusetts 01854, U.S.A.
nilabja_guha@uml.edu

Y. DING

Department of Industrial and Systems Engineering, Texas A&M University, 3131 TAMU, College Station, Texas 77843-3131, U.S.A.
yuding@tamu.edu

and B. K. MALLICK

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, Texas 77843-3143, U.S.A.
bmallick@stat.tamu.edu

Summary

Conditional density estimation seeks to model the distribution of a response variable conditional on covariates. We propose a Bayesian partition model using logistic Gaussian processes to perform conditional density estimation. The partition takes the form of a Voronoi tessellation and is learned from the data using a reversible jump Markov chain Monte Carlo algorithm. The methodology models data in which the density changes sharply throughout the covariate space, and can be used to determine where important changes in the density occur. The Markov chain Monte Carlo algorithm involves a Laplace approximation on the latent variables of the logistic Gaussian process model which marginalizes the parameters in each partition element, allowing an efficient search of the approximate posterior distribution of the tessellation. The method is consistent when the density is piecewise constant in the covariate space or when the density is Lipschitz continuous with respect to the covariates. In simulation and application to wind turbine data, the model successfully estimates the partition structure and conditional distribution.

Some key words: Bayesian conditional density estimation; Laplace approximation; Logistic Gaussian process; Partition model; Reversible jump Markov chain Monte Carlo.

1. INTRODUCTION

Conditional density estimation, sometimes referred to as density regression, is used to estimate the conditional distribution of a response variable, y, given a vector of covariates, x. Regression methods are a type of conditional density estimation which usually focus on modelling a location change while assuming few, or no, distributional changes in the spread or shape of y. The term conditional density estimation as used in this paper refers to more general methods that model changes in location, spread and shape of the distribution of y throughout the covariate space. These methods are useful when data violate standard parametric assumptions, and the shape of the distribution of y is an important part of inference or prediction.

There are a number of existing frequentist approaches to performing conditional density estimation, including kernel methods (Fan et al., 1996; Fu et al., 2011), spline methods (Kooperberg & Stone, 1991; Stone et al., 1997) and mixtures of experts (Jacobs et al., 1991). There are also Bayesian approaches to conditional density estimation, a popular one being to use mixture models for the conditional distribution of $p(y \mid x)$ and allow the mixing weights and parameters to depend on the covariates (Griffin & Steel, 2006; Dunson et al., 2007; Dunson & Park, 2008; Chung & Dunson, 2009). An alternative is to apply the logistic Gaussian process model in the conditional density estimation setting (Tokdar et al., 2010). Latent variable models have been utilized by Bhattacharya & Dunson (2010) and Kundu & Dunson (2011). A multivariate spline-based method (Shen & Ghosal, 2016) and an optional Pólya tree-based method (Ma & Wong, 2011) have recently been proposed.

The method proposed in this paper provides a novel Bayesian partition model (Denison et al., 2002a; Holmes et al., 2005) to perform conditional density estimation using logistic Gaussian processes. The data are partitioned using a Voronoi tessellation on the covariate space x and the distribution of y within each partition region is modelled using a univariate logistic Gaussian process which is independent of x. The primary goal of the partition model is to infer the partition structure and the distribution of y within each partition element. In parametric partition models, this is typically done through Markov chain Monte Carlo; priors are selected such that the parameters in each partition element can be integrated out analytically, providing the marginal probability for the data given the partition and allowing for an efficient search of the partition's posterior (Denison & Holmes, 2001; Kim et al., 2005). The logistic Gaussian process model does not have an analytical form for the marginal of y, but this can be estimated using a Laplace approximation (Riihimäki & Vehtari, 2014), allowing an effective search of the approximate posterior tessellation structure via Markov chain Monte Carlo methods.

Logistic Gaussian processes have been used in other contexts to perform conditional density estimation. Tokdar et al. (2010) use the logistic Gaussian process to model the joint distribution of the response y and the covariates x, and use a subspace projection method to reduce the dimension of the covariates. Conversely, we are fitting univariate logistic Gaussian processes within each region of x, hence avoiding the curse of dimensionality. Furthermore, in joint modelling approaches, y and x are not identified as the response and covariates. Our experience is that the distribution of the response y is highly influenced by the distribution of x, especially when the dimension of x is high.

Most conditional density estimation techniques assume that the density of y changes smoothly over the covariate space, but this may be false. One motivating example for the present method is wind turbine data, where it is known that the distribution of power output changes sharply over wind speed and wind direction. Figure 1 plots normalized power output against wind speed and illustrates the sharp change in power output as wind speeds increase. The density of power output is also known to change sharply over different wind directions due to the terrain around the

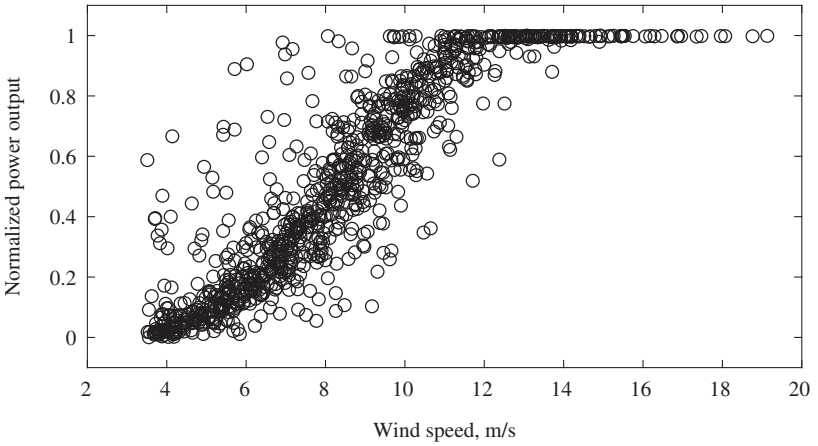

Fig. 1. A plot of normalized power output against wind speed (m/s).

turbine and wake effects from upwind turbines. Conditional density estimation methods which assume that the density of y changes smoothly throughout the covariate space are generally not designed to handle these sharp boundaries, nor do they provide information as to where they occur.

A few partition models exist for performing conditional density estimation. Petralia et al. (2013) used multiscale dictionaries and a tree decomposition to make density predictions using a large number of features. Their method constructs the partition independent of y, whereas we use y to influence the partitioning. Ma (2017) used optional Pólya trees to construct the partition and estimate the densities, but this method fails to model the densities adequately for small sample sizes. Tree models often have a nice interpretation, but they can become overly complex even in low dimensions, if the true partition structure is not oriented to the block partitions of the tree model. The tessellation structure in this paper is more flexible than tree methods since it is constructed using a weighted Euclidean distance, which allows the model to partition the covariate space into regions of varying shapes, rather than sets of hyper-rectangles.

The main advantage of partition models over smooth conditional density estimation is the inference of the partition structure. In smooth models, the user must specify at which points in the covariate space to view the density of y. It is possible when exploring the relationship between y and x that interesting relationships may be missed due to the volume of x, even in small dimensions. On the other hand, the partition model is designed to determine where important changes in y occur throughout x, providing greater interpretability. Indeed, the model is similar in spirit to the Bayesian classification and regression tree model (Chipman et al., 1998; Denison et al., 1998) as a decision tool to understand how and where the density of y changes for different x. By using logistic Gaussian processes to model the densities, the method can flexibly model the densities in each partition element while simultaneously providing interpretability through the tessellation's posterior.

An important contribution of this paper is to provide theoretical properties of the proposed partition model. To our knowledge, this is the first paper that considers the posterior consistency in estimating conditional distributions for partition models with Voronoi tessellations. Indeed, there are a few papers which have considered theoretical properties of conditional density estimation models using other modelling frameworks (Tokdar & Ghosh, 2007; Bhattacharya & Dunson, 2010; Norets & Pelenis, 2012; Pati et al., 2013).

2. BAYESIAN HIERARCHICAL CONDITIONAL DENSITY ESTIMATION PARTITION MODEL

2.1. *Modelling the partition structure using a Voronoi tessellation*

The partition model divides the d-dimensional covariate space $\mathcal{D}$ into M distinct pieces where y is assumed to independently follow a different density $p_i(\cdot)$ $(i = 1, \ldots, M)$ within each partition. The partitioning of the covariate space is done through a Voronoi tessellation. The tessellation is defined by M centres $c_1^D, \ldots, c_M^D$ that divide the covariate space into M disjoint regions $R_1, \ldots, R_M$, where R_i consists of all the observed x that are closest to centre c_i^D. Formally, $R_i = \{x \in \mathcal{D} : ||x - c_i^D|| < ||x - c_j^D|| \text{ for all } i \neq j\}$. Here, $||x||^2 = ||(x^1, \ldots, x^d)||^2 = \sum_{i=1}^d w^i (x^i)^2$ where $w = (w^1, \ldots, w^d)$ is a normalized weighting vector, $\sum w^i = 1$, that places different weights on each of the covariates (Holmes et al., 2005). The weighting provides additional flexibility in the class of tessellations.

For simplicity, we assume that the possible centres of the tessellation are restricted to the observed covariate values. Next, we assign prior distributions for both the number of centres and the centre locations. The prior is $p(c^D, M, w) = p(c^D \mid M)p(M)p(w)$, where

$$\begin{aligned} p(M) &= \mathrm{Du}(M \mid 1, \ldots, M_{\max}), \\ p(c^D \mid M) &= \mathrm{Du}\big(c^D \mid 1, \ldots, [n!\{(n-M)!M!\}^{-1}]\big), \\ p(w) &= \mathrm{Dir}(w \mid 1, \ldots, 1), \end{aligned}$$

where $\mathrm{Du}(x \mid 1, \ldots, n)$ means discrete uniform on $1, \ldots, n$, and $M_{\max}$ is the maximum number of allowable centres, a hyperparameter chosen by the user. The prior on $p(c^D \mid M)$ gives equal weight to all possible combinations of M centres with possible centre locations corresponding to the n observed values of the covariates. This prior penalizes larger models as long as $M < \lfloor n/2 \rfloor$, which should always be the case since one cannot adequately model the density in each partition element with so few points. The vector w has the Dirichlet prior which is uniform on the simplex.

2.2. *Likelihood and prior*

Let $\tilde{y} = \{y_1, \ldots, y_n\}$ be the observed responses and $\tilde{y}_i = \{y_j : x_j \in R_i,\ j = 1, \ldots, n\}$ $(i = 1, \ldots, M)$ be the n_i observed response variables whose covariates are in the ith region of the tessellation. We assume that within the ith region, each observation is drawn independently from the same density, which is modelled by a univariate logistic-Gaussian model (Lenk, 1988). Given a partition, the density of y within each partition region is independent of x. The logistic Gaussian process models the density using an exponentiated Gaussian process over a bounded interval, $\mathcal{V}_i$. The density in the ith partition region is expressed as (Riihimäki & Vehtari, 2014)

$$p_i(y) = \frac{\exp\{f_i(y)\}}{\int_{\mathcal{V}_i} \exp\{f_i(s)\}\, \mathrm{d}s},$$

where $f_i(\cdot)$ is modelled with a latent Gaussian process as

$$f_i(\cdot) = \mu_i(\cdot) + g_i(\cdot), \quad g_i(\cdot) \sim \mathrm{GP}\{0, \kappa_{\theta_i}(\cdot, \cdot)\}, \quad \mu_i(\cdot) = h(\cdot)^{\mathrm{T}} \beta_i, \tag{1}$$

with $\kappa_{\theta_i}(\cdot, \cdot)$ a covariance function that depends on the smoothing hyperparameter vector θ_i. For simplicity in exposition, assume θ_i is known; its selection via empirical Bayes will be detailed later in this subsection. The function $h(z) = (z, z^2)^{\mathrm{T}}$ is used to encourage decreasing tails in the density function $p_i(\cdot)$ near the boundary where data may be sparse (Riihimäki & Vehtari, 2014).

By placing a Gaussian prior on $\beta_i \sim N(b, B)$, β_i can be integrated out to yield the marginal prior $f_i(\cdot) \sim \mathrm{GP}\{h(\cdot)^{\mathrm{T}}b,\ \kappa_{\theta_i}(\cdot,\cdot) + h(\cdot)^{\mathrm{T}}Bh(\cdot)\}$.

Although a full Gaussian process prior is flexible, it does present some computational challenges, particularly when trying to integrate out the latent function $f_i(\cdot)$ to obtain the marginal distribution of $\tilde{y}_i$. To aid in providing a computationally feasible solution, we follow Riihimäki & Vehtari (2014) and approximate the logistic-Gaussian density by discretization. The bounded region $\mathcal{V}_i$ is discretized into a regular grid of r_i subregions centred at $Z_i = (z_{i1}, \ldots, z_{ir_i})^{\mathrm{T}}$. This discretization occurs over the one-dimensional range of y, not x. The function $f_i(\cdot)$ is evaluated at r_i points in the vector $\mathrm{f}_i = \{f_i(z_{i1}), \ldots, f_i(z_{ir_i})\}^{\mathrm{T}}$, and the continuous density is approximated with the discrete version. That is, the likelihood of an observation falling into the j'th subregion would be

$$\frac{\omega_{j'} \exp(\mathrm{f}_{ij'})}{\sum_{j'=1}^{r_i} \omega_{j'} \exp(\mathrm{f}_{ij'})}, \tag{2}$$

where $\omega_{j'}$ is the width of the region centred at $z_{ij'}$ and $\mathrm{f}_{ij'} = f_i(z_{ij'})$, the j'th element of f_i. Since we are using a regular grid, the $w_{j'}$ cancel out in (2). This leads us to the joint likelihood of y_i as

$$p(\tilde{y}_i \mid \mathrm{f}_i) = \exp\left[y_i^{\star\mathrm{T}} \mathrm{f}_i - n_i \log\left\{ \sum_{j'=1}^{r_i} \exp(\mathrm{f}_{ij'}) \right\} \right],$$

where n_i is the length of $\tilde{y}_i$, and $y_i^{\star}$ is a column vector of length r_i with the j'th element as the number of elements of $\tilde{y}_i$ that fall into the subregion centred at $z_{ij'}$. Since we are modelling $f_i(\cdot)$ as a Gaussian process, f_i has a multivariate normal distribution $\mathrm{f}_i \sim N(H_i b,\ K_i + H_i B H_i^{\mathrm{T}})$, where K_i is an $r_i \times r_i$ matrix with (j', k')th element $\kappa_{\theta_i}(z_{ij'}, z_{ik'})$, and H_i is an $r_i \times 2$ matrix with j'th row $h(z_{ij'})^{\mathrm{T}}$. The partition model assumes independence between data in different partitions, therefore the posterior distribution can be expressed as a product:

$$p(T, \mathrm{f}_1, \ldots, \mathrm{f}_M \mid y_1, \ldots, y_n) \propto p(T) \prod_{i=1}^{M} p(\tilde{y}_i \mid \mathrm{f}_i) p(\mathrm{f}_i),$$

where $T = \{M, c^D, w\}$ denotes the tessellation parameters. In our case, as in many other partition models, interest lies in obtaining samples from the posterior of the tessellation, $p(T \mid y_1, \ldots, y_n)$. This is typically accomplished using reversible jump Markov chain Monte Carlo simulation (Green, 1995) after integrating out the partition-specific parameters. In this framework we seek

$$p(T \mid \tilde{y}) \propto p(T) p(\tilde{y} \mid T) = p(T) \prod_{i=1}^{M} \int p(\tilde{y}_i \mid \mathrm{f}_i) p(\mathrm{f}_i)\, \mathrm{d}\mathrm{f}_i = p(T) \prod_{i=1}^{M} p(\tilde{y}_i). \tag{3}$$

The integrals can be estimated by a Laplace approximation within each partition. Laplace's method requires finding $\hat{\mathrm{f}}_i = \arg\max_{\mathrm{f}_i} p(\tilde{y}_i \mid \mathrm{f}_i) p(\mathrm{f}_i)$ via Newton's method and using a Taylor expansion to construct the normal approximation. Indeed, the model presented is constructed primarily by embedding the univariate model of Riihimäki & Vehtari (2014) into a partition model framework to infer the effect of covariates via the partition structure. Riihimäki & Vehtari (2014) provided the form of the Laplace approximation to obtain $p(\tilde{y}_i)$ in equation (16) of their paper.

Lastly, we discuss the form of $\kappa_{\theta_i}(\cdot,\cdot)$ and the selection of θ_i in each partition. For the purposes of this paper, we assume that the covariance function $\kappa_{\theta_i}(\cdot,\cdot)$, which depends on hyperparameters $\theta_i = (\sigma_i, l_i)$, is the stationary squared exponential covariance function

$$\kappa_{\theta_i}(z,z') = \sigma_i^2 \exp\left\{-\frac{1}{2l_i^2}(z-z')^2\right\}, \quad z,z' \in \mathbb{R}, \quad \sigma_i, l_i \in (0,\infty),$$

where σ_i is the magnitude hyperparameter and l_i is a length-scale hyperparameter; together, these govern the smoothness properties of $f_i(\cdot)$. The value of θ_i is chosen by empirical Bayes by maximizing the posterior $p(\sigma_i, l_i \mid \tilde{y}_i)$ proportional to $p(\sigma_i)p(l_i)\int p(\tilde{y}_i \mid \mathrm{f}_i)p(\mathrm{f}_i)\,\mathrm{df}_i$, which we shall denote $\hat{\theta}_i = (\hat{\sigma}_i^2, \hat{l}_i^2)$ in each partition. We choose temporary hyperpriors, $p(\sigma_i)$ and $p(l_i)$, to guide the selection of θ_i. We place a weakly informative half Student t distribution with one degree of freedom and a variance equal to 10 for the magnitude parameter, and the same prior with a variance of 1 for the length-scale hyperparameter (Riihimäki & Vehtari, 2014).

Thus, computing $p(y_1,\ldots,y_n \mid T)$ in (3) is done by: (i) Determining $\tilde{y}_i$ $(i = 1,\ldots,M)$, based on covariate values; (ii) For each $\tilde{y}_i$, finding $\hat{\theta}_i$ by maximizing $p(\tilde{y}_i)$ subject to the temporary hyperprior on θ_i and setting $\theta_i = \hat{\theta}_i$; (iii) Computing $p(\tilde{y}_i)$ in each partition region via the Laplace approximation using the fixed values of θ_i.

2.3. *Reversible jump Markov chain Monte Carlo algorithm*

The reversible jump Markov chain Monte Carlo algorithm to sample from the posterior of the tessellation has four possible moves: birth, death, move and change. A birth step adds a new tessellation centre randomly from x which is not currently a tessellation centre. A death step randomly removes an existing tessellation centre. A move step randomly moves an existing tessellation centre to another value of x which is not currently a tessellation centre. A change step randomly selects an element of w, modifies it according to $q(\cdot,\cdot)$, and normalizes the resulting vector.

In our implementation we place an equal probability of 1/4 for each of the moves in the algorithm, but these probabilities can be changed if desired. Algorithm S1 in the Supplementary Material shows the pseudocode for performing the reversible jump algorithm. After choosing an initial tessellation, the algorithm consists of iteratively choosing one of the four moves with probability 1/4, and accepting the proposed moves from the current tesselation T to a proposed tesselation T' with probability

$$\alpha = \min\left\{1, \frac{q(w \mid w')p(\tilde{y} \mid T')}{q(w' \mid w)p(\tilde{y} \mid T)}\right\}, \tag{4}$$

where $q(w' \mid w)$ is the proposal used to modify an element of w. In our implementation we set $q(w' \mid w)$ to be a truncated normal distribution centred at w and truncated to allow only positive values; it applies only to the randomly selected single element of w which is modified in a change step. The ratio involving $q(\cdot \mid \cdot)$ is 1 if a birth, death or change step is proposed, since w remains unchanged. The only tuning parameters for the algorithm are the probabilities of proposing birth, death, move and change steps and the proposal variance of $q(\cdot \mid \cdot)$.

For most nonboundary cases, the prior on the tessellation structure does not appear in α due to cancellations with itself and the proposal distribution for the birth, death, move and change steps in the reversible jump algorithm. When M is at or near the boundary, however, adjustments to α need to be made in order to maintain the reversibility of the Markov chain. When $M = 1$

and a birth step is proposed, or when $M = M_{\max}$ and a death step is proposed, we must multiply the ratio in (4) by 3/4. When $M = 2$ and a death step is proposed, or $M = M_{\max} - 1$ and a birth step is proposed, the ratio must be multiplied by 4/3.

The mixing of a single Markov chain may be poor due to multiple posterior nodes, so we use parallel tempering similar to that of Gramacy (2007) and Gramacy & Taddy (2010) to better explore the posterior distribution. See the Supplementary Material.

3. Some results on convergence

3.1. *Preliminaries*

In this section we establish consistency of our method. If the true data-generating model is a partition model, then as n, the number of observations, goes to infinity, the posterior density concentrates on a small total variation neighbourhood around the true density. We state the result for the Euclidean metric $d(\cdot, \cdot)$. The details for the general weighted norm are given in the Supplementary Material. First, we show consistency under the true model where the target density has partition form. Later we show consistency under model misspecification where the true conditional density is Lipschitz continuous. Under smoothness conditions and if the underlying true model corresponds to a Voronoi partition, then the fitted joint density lies in an ϵ_n-radius Hellinger ball around the true density with high probability, where $\epsilon_n = n^{-(0.5-c_\delta)}$ and $c_\delta > 0$ can be any constant.

Under model misspecification, we show that the posterior probability of any small total variation neighbourhood around the true joint density goes to unity as the number of observations goes to infinity, under Lipschitz-type conditions. Further, it can be shown that the fitted conditional distribution would be arbitrarily close to the true conditional distribution for all values of the covariates outside a small probability set of the covariate space.

3.2. *Consistency under the true model*

First, we show that the partitions formed by a Voronoi tessellation can adequately approximate the true partition. We then establish that we have sufficient prior probability for the approximating partition around any small neighbourhood of the true Gaussian process paths in the supremum norm. Finally, if we have sufficient prior mass around the true density, the likelihood pulls the posterior density towards the data-generating density under the true model.

Let $c_1^D, \ldots, c_m^D$ be the centres of some Voronoi tessellation and $R_1, \ldots, R_m$ be the corresponding Voronoi regions in Ω, a subset of $\mathbb{R}^d$ with associated Lebesgue measure $\mathscr{L}$. Let $V_1, \ldots, V_k$ be any given partition of Ω, which is the covariate space $\mathcal{D}$ in our case. We assume that each region V_i is a finite union of rectangular regions. Our result holds for a general region approximated by a finite union of rectangles.

Proposition 1. *Given $\epsilon_1 > 0$ there exist M, $c_1^D, \ldots, c_M^D$ and a partition $J_1, \ldots, J_k$ of $\{1, \ldots, M\}$ such that $U_l = \cup_{i \in J_l} R_i$ and so forth, and $\sum_{l=1}^k \mathscr{L}(U_l \Delta V_l) \leqslant \epsilon_1$, where Δ denotes the symmetric differences of sets.*

In our proposed method we use the observed values of the covariates for the centres of the tessellation. Next, we show that a small perturbation of $c_1^D, \ldots, c_M^D$ from Proposition 1 does not change the partition dramatically and provides an approximation for regions $V_1, \ldots, V_k$. Then, we show that any small neighbourhood of $c_1^D, \ldots, c_M^D$ contains observed covariates with probability 1 as n goes to infinity, since we assume that the probability measure on x, $H(\cdot)$, has

a strictly positive, bounded away from zero, density function. We summarize these two results in the two following propositions. Let $x_1, \ldots, x_n$ be the observed covariate vectors, which are independent and $x_j \sim H(\cdot)$.

PROPOSITION 2. *Given $\epsilon_1 > 0$, $c_1^D, \ldots, c_M^D$ and $R_1, \ldots, R_M$ from Proposition 1, we can have $\delta > 0$ and Voronoi centres $c_1^{D'}, \ldots, c_M^{D'}$ and corresponding $R'_1, \ldots, R'_M$ such that if $d(c_i^D, c_i^{D'}) < \delta$ then $\sum_{l=1}^k \mathscr{L}(U'_l \Delta V_l) \leqslant 2\epsilon_1$, where $U'_l = \cup_{i \in J_l} R'_i$ and $d(\cdot, \cdot)$ denotes the distance under the Euclidean norm.*

PROPOSITION 3. *Under the set-up of Proposition 1, as $n \to \infty$, for any $\delta > 0$, there exists x_{j_i} for some $1 \leqslant j_i \leqslant n$ in the δ-radius ball around c_i^D, for all $1 \leqslant i \leqslant M$, with probability one.*

Let

$$g_x(y) = \sum_i \mathrm{e}^{\mu_i + \eta_i(y)} \mathbf{1}_{x \in U_i}, \quad p(y \mid x) \propto g_x(y), \tag{5}$$

where the μ_i are the mean function given in (1), U_i denote partitions of the covariate space and $p^*(y \mid x)$ denotes the true conditional density, which we assume to be bounded away from zero and infinity. Let $|\eta_i^*(y)| < k_0/2$ and $|\mu_i^*| < k_0/2$, $k_0 > 1$, corresponding to p^*, and we have corresponding coefficients β_i^*. Let $\epsilon_1 < \epsilon_2$. Let $V_1^*, \ldots, V_k^*$ be the true underlying partition of Ω; from Proposition 1, there exist $U_1^*, \ldots, U_k^*$ from the Voronoi approximation. Consider the following neighbourhood in the supremum norm ($\| \cdot \|_\infty$):

$$N_1 = \{\|\eta_i(y) - \eta_i^*(y)\|_\infty < \epsilon_2 \text{ in } x \in U_i^* \cap V_i^* \text{ and } \|\eta_i(y)\|_\infty < k_0 \text{ for } x \in U_i^* \Delta V_i^*\},$$
$$N_2 = \{\|\beta_i - \beta_i^*\|_\infty < \epsilon_2 \text{ in } x \in U_i^* \cap V_i^* \text{ and } \|\mu_i(y)\|_\infty < k_0 \text{ for } x \in U_i^* \Delta V_i^*\}.$$

Without loss of generality, we assume here $R_l = U_l^*$ in Proposition 1. Otherwise, η_i, β_i can be replaced with η_{i_j}, β_{i_j} in N_1, N_2 where $i_j \in J_i$, and we have $\|\eta_{i_j} - \eta_i^*\|_\infty, \|\beta_{i_j} - \beta_i^*\|_\infty < \epsilon_2$ in $U_i^* \cap V_i^*$ for all j.

PROPOSITION 4. *For $\{\mu_i, \eta_i\}$ pairs such that $\mu_i \in N_2$ and $\eta_i \in N_1$ for all i, we have $\int |p(y \mid x) - p^*(y \mid x)| \, \mathrm{d}H(x) \, \mathrm{d}y < k\epsilon_2$ and $E_{p^*(y,x)}[\log\{p^*(y,x)/p(y,x)\}] < k\epsilon_2$, for some $k > 0$, where E_{p^*} implies expectation with respect to p^*.*

The covariance kernel between points s and t can be written as $K(s,t) = \sigma_i^2 K_0(s/l_i, t/l_i) = \sigma_i^2 K_{l_i}(s,t) = \kappa_{\theta_i}(s,t)$ for the ith region R_i, where $K_0(\cdot,\cdot)$ is a smooth kernel, which in our case is the Gaussian kernel, and $K_{l_i}(\cdot,\cdot)$ corresponds to a kernel with roughness parameter l_i. Let $\theta_i \sim \pi(\cdot)$ be independent priors on θ_i.

Let $\eta(\cdot)$, or in particular $\eta_i(\cdot)$, be any Gaussian process path under the proposed Gaussian kernel on the closed interval I that is the support of y. Under the smoothness of the covariance kernel the paths are smooth and the derivative process is again a Gaussian process (Ghosal & Roy, 2006). For any density based on $m \leqslant M$ partitions we have an m-dimensional product function space. We construct sieves on the function space where the probability outside the sieves decreases exponentially with n, and establish an entropy bound for the sieves. We use this construction to prove our following convergence results. We also require enough prior probability around the truth.

Prior probability around the $\eta_i^*(\cdot)$ depends on the underlying reproducing kernel Hilbert space structure (van der Vaart & van Zanten, 2007, 2008a). For a Gaussian process prior with covariance

kernel $\mathscr{K}(\cdot,\cdot)$, the reproducing kernel Hilbert space is constructed by completing the linear span of the $\mathscr{K}(\cdot,s_l)$ with inner product $\langle \mathscr{K}(\cdot,s_l), \mathscr{K}(\cdot,s_m)\rangle = \mathscr{K}(s_l,s_m)$. We assume that $\eta_i^*(\cdot)$ lies in the closure of the underlying reproducing kernel Hilbert space support in the supremum norm for an open set of hyperparameters. Then, under a Gaussian process prior, any small supremum norm neighbourhood around $\eta_i^*(\cdot)$ has positive probability (Tokdar & Ghosh, 2007). Under a Gaussian kernel, the set of reproducing kernel Hilbert space elements over all the l_i will have continuous functions in its closure under the supremum norm (Tokdar & Ghosh, 2007, Theorem 4.4). The next assumption summarizes this support restriction.

Assumption 1. Let the closure of $A_{R_l} = \{\eta(t) = \sum_{i=1}^{k} a_i K_l(t,t_i^*) : a_i \in \mathbb{R},\ k \in \mathbb{N}$ and $t, t_i^* \in I\}$ corresponding to the Hilbert space norm be the reproducing kernel Hilbert space for roughness parameter $l > 0$. Assume that $\eta_i^*(\cdot)$ are in the sup-norm closure of $A_R = \cup_l A_{R_l}$. Here, $\mathbb{R}, \mathbb{N}$ are the sets of real and natural numbers, respectively.

Let $\Pi(\cdot) = \Pi_n(\cdot)$ denote the joint prior distribution on the model space from § 2, specified by

$$\eta_i(\cdot) \sim \mathrm{GP}\{0, \kappa_{\theta_i}(\cdot,\cdot)\}; \quad \mu_i(\cdot) = h(\cdot)^{\mathrm{T}}\beta_i,\ \beta_i \sim N(b,B); \quad \theta_i \sim \pi(\cdot), \tag{6}$$

and

$$M \sim \mathrm{Du}(M \mid 1,\ldots,M_{\max}); \quad c^D \mid M \sim \mathrm{Du}\big(c^D \mid 1,\ldots,[n!\{(n-M)!M!\}^{-1}]\big), \tag{7}$$

and $\Pi(\mathbf{S} \mid \cdot)$ denote the posterior probability of a measurable subset $\mathbf{S}$ of the model space given observed data $\{y_j, x_j\}$ $(j = 1,\ldots,n)$. We assume that the $\eta_i^*(\cdot)$ satisfy the reproducing kernel Hilbert space condition stated in the last paragraph.

Assumption 2. We assume the following regarding hyperparameters:

(a) $\log[\max\{\Pi(\sigma_i > \lambda_n), \Pi(1/l_i > \nu_n)\}] = O(-n)$;
(b) $M_n^2 \lambda_n^{-2} \nu_n^{-2\alpha}/n \to \infty$;
(c) $M_n^{1/\alpha} = O(n^{\gamma}), 0 < \gamma < 1$.

Here, M_n is polynomial, of order of n; $\{M_n\}_{n\geqslant 1}$, $\{\lambda_n\}_{n\geqslant 1}$ and $\{\nu_n\}_{n\geqslant 1}$ are sequences going to infinity; and $\alpha \geqslant 1$ is an integer.

THEOREM 1. *Let* $U_{\epsilon'} = \{p : \int |p(y \mid x) - p^*(y \mid x)|\,\mathrm{d}y\,\mathrm{d}H(x) < \epsilon'\}$, $\epsilon' > 0$. *Then, for* $M_{\max} > M_0$, *where* M_0 *is a constant, under Assumptions* 1 *and* 2 *and the Voronoi log Gaussian process prior given in* (6) *and* (7) *for the model in* (5), $\Pi(U_{\epsilon'} \mid \cdot) \to 1$ *with probability one as n, the number of observations, goes to infinity.*

Even though the main results focus on the neighbourhood of the estimated density, the prior favours smaller partitions. Heuristically, if the true partition is further partitioned into smaller partitions, then the true likelihood remains the same over the smaller partitions, but the prior puts $O(n^{-m})$ weight on a partition with m centres. Hence, extra subpartitions will reduce the posterior probability. Therefore, we should have higher posterior probability for the smaller number of Voronoi centres, as long as it can capture the true data-generating partition. We can use a prior satisfying Assumption 2, or truncate the hyperparameters at λ_n and ν_n.

If the true model is in partition form and has an underlying true data-generating Voronoi partition, and $\eta_i^*(\cdot)$ is smooth and in the reproducing kernel Hilbert space in each partition, we can achieve a convergence rate arbitrarily close to the minimax rate in terms of Hellinger

distance, as given in the following theorem. To prove this result we need to have sufficient prior probability in a small neighbourhood around the true $\eta_i^*(\cdot)$ under the Gaussian kernel, which can be established through the underlying reproducing kernel Hilbert space.

In our context, for a Gaussian process prior $\Pi^{\mathscr{G}}$ with covariance kernel $\mathscr{K}(\cdot,\cdot)$, let $\Pi^{\mathscr{G}}_{\epsilon,\tilde{\eta}} = \Pi^{\mathscr{G}}(\eta(\cdot) : \|\eta(\cdot) - \tilde{\eta}(\cdot)\|_\infty < \epsilon)$ be the probability of the sup-norm ϵ-neighbourhood of $\tilde{\eta}(\cdot)$. Then $\phi_{\tilde{\eta}(\cdot)}(\epsilon) = \inf_{\bar{h}\in\mathbb{H}:\|\bar{h}-\tilde{\eta}\|_\infty<\epsilon} \frac{1}{2}\|\bar{h}\|^2_{\mathbb{H}} - \log \Pi^{\mathscr{G}}(\|\eta(\cdot)\|_\infty < \epsilon)$ and $\phi_{\tilde{\eta}(\cdot)}(\epsilon) \leqslant -\log \Pi^{\mathscr{G}}_{\epsilon,\tilde{\eta}} \leqslant \phi_{\tilde{\eta}(\cdot)}(\epsilon/2)$. The preceding inequality is from van der Vaart & van Zanten (2008b). Here, $\mathbb{H}$ is the reproducing kernel Hilbert space associated with $\mathscr{K}(\cdot,\cdot)$. We assume all the $\eta_i^*(\cdot)$ belong to the reproducing kernel Hilbert space support on an open neighbourhood of σ_i, and the l_i are bounded away from zero and infinity, and have bounded norm induced by the reproducing kernel Hilbert space in that neighbourhood. See Smola et al. (1998) and van der Vaart & van Zanten (2011) for an explicit form for the reproducing kernel Hilbert space norm of the smooth function for the Gaussian kernel. Let $A^{\mathscr{C}}_{R_l}$ be the set containing the functions in A_{R_l} and the convolutions $\int_I w(s)K_l(t,s)\,\mathrm{d}s$ for bounded continuous convolution functions $w(\cdot)$ on I, the domain of y.

Assumption 3. For $A_{R_l} = \{\eta(t) = \sum_{i=1}^k a_i K_l(t,t_i^*) : a_i \in \mathbb{R},\ k \in \mathbb{N},\ \text{and } t, t_i^* \in I\}$, let $A^{\mathscr{H}}_{R_l}$ be the closure of A_{R_l} corresponding to the Hilbert space norm. Assume that $\eta_i^*(\cdot)$ are in $A^{\mathscr{H}}_{R_{l_i}}$ for l_i lying in an open subset of the positive real line with uniformly bounded Hilbert space norm, or $\eta_i^*(\cdot)$ are in $A^{\mathscr{C}}_{R_{l_i}}$s for some l_i.

Assumption 4. For a sequence ϵ_n^2 going to zero, we assume:

(a) $\epsilon_n = n^{-(0.5-c_\delta)}$, where $0 < c_\delta < 0.5$;
(b) ${\epsilon_n}^{-2/\alpha} {M_n}^{1/\alpha} = o(n\epsilon_n^2)$ for some integer α;
(c) $-\log\left[\max\left\{\Pi(\sigma_i > \lambda_n), \Pi\left(\frac{1}{l_i} > \nu_n\right)\right\}\right]/(n\epsilon_n^2) \to \infty$;
(d) $M_n^2 \lambda_n^{-2} \nu_n^{-2\alpha}/(n\epsilon_n^2) \to \infty$.

Assumption 5. For large n we have observations at the true Voronoi centres.

The above condition can be replaced with:

Assumption 6. First, $M \leqslant M_{\max}$ observation points are chosen from the discrete uniform prior from (7). Then, the M Voronoi centres are chosen with independent discrete uniform priors inside d-dimensional rectangles with centres of mass at M observation points and side length $2\epsilon_n^{2/d}$ on the grid points, induced by $\epsilon_n^2 d^{-1/2}$-distance equispaced grid points on each side of the rectangle. Here, d is the dimension of the covariate space. That is, we have less than $(2\epsilon_n^{2/d}\epsilon_n^{-2}d^{1/2})^d$ grid points inside one such rectangle.

Then we have the following result.

THEOREM 2. *Let* $U^h_{\epsilon_n} = \{p : \int[\{p(y \mid x)\}^{1/2} - \{p^*(y \mid x)\}^{1/2}]^2\,\mathrm{d}y\,\mathrm{d}H(x) < \epsilon_n^2\}$ *be the* ϵ_n*-radius Hellinger ball around the true density. Under Assumptions* 3 *and* 4*, and either of Assumptions* 5 *or* 6*, with the Voronoi log Gaussian process prior given in* (6) *and* (7) *for the model in* (5)*, if the true partition of the covariate space is induced by a Voronoi partition and* $M_{\max}$ *is greater than the total number of true Voronoi regions, then* $\Pi(U^h_{M_h\epsilon_n} \mid \cdot) \to 1$ *in probability for some large constant* M_h.

3.3. *Consistency under misspecification*

We have shown posterior consistency for densities where the true density has the partition form. Even though the motivation for this method is to model this class of densities by a Voronoi partition, here we extend our result to a more general case when the true conditional density is continuous and the covariate space is a d-dimensional rectangle. We consider a general Lipschitz condition for conditional densities in the following assumption.

Assumption 7. Assume $|p^*(y \mid x_1) - p^*(y \mid x_2)| \leqslant c_l d(x_1, x_2)$, for some constant $c_l > 0$.

We show that the class of Voronoi approximations of conditional density functions is dense in the class of Lipschitz continuous functions. For $U_{\epsilon'} = \{p : \int |p(y \mid x) - p^*(y \mid x)| \, \mathrm{d}H(x) \, \mathrm{d}y < \epsilon'\}$ we show that the posterior probability of $U_{\epsilon'}$ goes to one with probability one as n goes to infinity. We assume the following.

Assumption 8. The log of the conditional density $\log p^*(y \mid x)$ belongs to the sup-norm closure of A_R from Assumption 1.

THEOREM 3. *For $\epsilon' > 0$ and $U_{\epsilon'} = \{p : \int |p(y \mid x) - p^*(y \mid x)| \, \mathrm{d}y \, \mathrm{d}H(x) < \epsilon'\}$ there exists $M_0 > 0$ such that for $M_{\max} > M_0$, under Assumptions* 2, 7, *and* 8, *and the Voronoi log Gaussian process prior given in* (6) *and* (7) *for the model in* (5), *$\Pi(U_{\epsilon'} \mid \cdot) \to 1$ with probability one, as the number of observations, n, goes to infinity.*

Next, we show a result for conditional density convergence. We define the set $U^c_{\epsilon,\delta} = \left\{p : H(\nu^\epsilon_{x,p}) > \delta\right\}$, where $\nu^\epsilon_{x,p} = \{x \text{ such that } \int [\{p(y \mid x)\}^{1/2} - \{p^*(y \mid x)\}^{1/2}]^2 \, \mathrm{d}y > \epsilon\}$, as the set of densities that have conditional Hellinger distance from the true density more than ϵ on a set of points in the covariate space which has measure greater than δ under H.

Let $M^n_{\max}$ be the maximum number of Voronoi centres selected, depending on the number of observed data points. As we observe more and more data, we let the number of Voronoi regions go to infinity at a certain rate given by Assumption 10 below, and establish that under the posterior measure the set of densities having conditional Hellinger distance more than $\epsilon' > 0$ over the set of covariate points with measure δ_n or more will have small probability, where δ_n goes to zero. We show the following result under the following conservative assumptions.

Assumption 9. The log of the conditional density $p^*(y \mid x)$ is in $A^{\mathscr{H}}_{R_l}$ from Assumption 3 for l in an open subset of the positive real line, and assume that the norm induced by the reproducing kernel Hilbert space is uniformly bounded for all x, or we assume that $\log p^*(y \mid x)$ belongs to the convolution class $A^{\mathscr{C}}_{R_{l'}}$ for some $l' > 0$, with Hilbert space norm uniformly bounded over all x, for uniformly bounded convolution functions.

Assumption 10. We assume $M^n_{\max} = O(n^{\gamma'})$, where $0 < \gamma' \leqslant 1/(2d + 2)$, and $(M^n_{\max})^{1+1/d} M_n^{1/\alpha} = o(n^{\delta'_1})$ for $\delta'_1 = 1 - \{(2d + 2)d\}^{-1}$.

THEOREM 4. *Let, $U^c_{\epsilon,\delta} = \left\{p : H(\nu^\epsilon_{x,p}) > \delta\right\}$, where $\nu^\epsilon_{x,p} = \{x$ such that $\int [\{p(y \mid x)\}^{1/2} - \{p^*(y \mid x)\}^{1/2}]^2 \, \mathrm{d}y > \epsilon\}$. For $\epsilon' > 0$, under Assumptions* 2, 7, 9 *and* 10, *and the Voronoi log Gaussian process prior given in* (6) *and* (7) *for the model in* (5), *there exists δ_n going to zero such that $\Pi(U^c_{\epsilon',\delta_n} \mid \cdot) \to 0$ in probability, as the number of observations, n, goes to infinity.*

The application of this model in practice differs slightly from the theory. The applied methodology uses two approximations: the discretized version of p_i in each partition, and the Laplace

approximation of the marginal of y. The theory is not based on these practical approximations, but the validity of practical results depends on a reasonable approximation. A measure of the closeness of these approximations to the true underlying model is not undertaken here, but the empirical results of Riihimäki & Vehtari (2014) indicate that these approximations are reasonable in practice for density estimation.

4. SIMULATIONS AND APPLICATIONS

4.1. *Preliminaries*

For our applications, we choose the maximum number of partitions to be $M_{\max} = 10$ to maintain simplicity in interpretation. We set the hyperparameters for the prior on β as $b = (0,0)^{\mathrm{T}}$ and $B = 10^2 I$, where I is the 2×2 identity matrix. Each Z_i $(i = 1, \ldots, M)$, the grid on which the density of y is discretized in each partition element, is a subset of a larger regular grid of 400 points $Z^\star = \{z_1^\star, \ldots, z_{400}^\star\}$ with $z_1^\star = \min\{\min(\tilde{y}), \bar{y} - 3\hat{\sigma}^2\}$ and $z_{400}^\star = \max\{\max(\tilde{y}), \bar{y} + 3\hat{\sigma}^2\}$, where $\bar{y}$ and $\hat{\sigma}^2$ are the sample mean and variance of $\tilde{y}$, the set of responses. Specifically, $Z_i = \{z_{j'}^\star : z_{j'}^\star \in [\min\{\min(\tilde{y}_i), \bar{y}_i - 3\hat{\sigma}_i{}^2\}, \max\{\max(\tilde{y}_i), \bar{y}_i + 3\hat{\sigma}_i{}^2\}]\ (j' = 1, \ldots, 400)\}$, where $\bar{y}_i$ and $\hat{\sigma}_i^2$ are the mean and standard deviation of the data $\tilde{y}_i$. We standardize the covariates and the response variable to have a mean of 0 and standard deviation of 1. This standardization allows us to use the priors selected generally without worrying about the range of the data.

4.2. *Simulation*

A total of 500 observations were simulated from the following model: $X_1, X_2 \sim \mathrm{Un}(0,1)$ and

$$Y \mid X_1, X_2 \sim \begin{cases} \mathrm{Ga}(10,2) & \text{if } X_1 > X_2,\ X_2 < 0.75, \\ \zeta N(1,1) + (1-\zeta)N(5,1),\ \zeta \sim \mathrm{Ber}(0.5) & \text{if } X_1 < X_2,\ X_1 < 0.75, \\ N(1, 0.5^{1/2}) & \text{if } X_1 > 0.75,\ X_2 > 0.75, \end{cases}$$

where $\mathrm{Ga}(a,b)$ is a gamma distribution with shape parameter a and rate parameter b, and $N(\mu,\sigma)$ denotes a normal distribution with mean μ and standard deviation σ. The Markov chain Monte Carlo algorithm was run for 10^5 iterations. Details regarding the mixing of the Markov chain for the partition model can be found in the Supplementary Material. The partition with the highest posterior probability is plotted in Fig. 2(a). The method performed extremely well in this setting and correctly captured the partition. The model also does well in capturing the distribution within each partition element; see Fig. 2(b).

This dataset was also analysed by three other methods, including dependent Bernstein polynomials (Barrientos et al., 2017), linear dependent tail-free processes (Jara & Hanson, 2011) and the Dirichlet process mixtures of normals method (Müller et al., 1996). All three methods are implemented in the `DPpackage` in `R` (R Development Core Team, 2020), and generally assume that the density of y changes smoothly as a function of the covariates, a very different assumption than the data-generating model. The methods were run for a 10^4-iteration burn-in period and kept every other draw for the next 40 000 iterations for a total of 20 000 posterior draws. We chose to evaluate the posterior density at four different covariate locations, $\{(0.76, 0.76), (0.9, 0.9), (0.1, 0.8), (0.8, 0.1)\}$. The $(0.76, 0.76)$ location is on the boundary of all three partition elements, whereas the other three locations are away from the boundary. Of these smooth methods, the Dirichlet process mixture of normals performed best, and the posterior densities at the four selected points are plotted with the true density in Fig. 3. We generally see that this method failed at the boundary, but performed adequately, albeit with wide credible intervals,

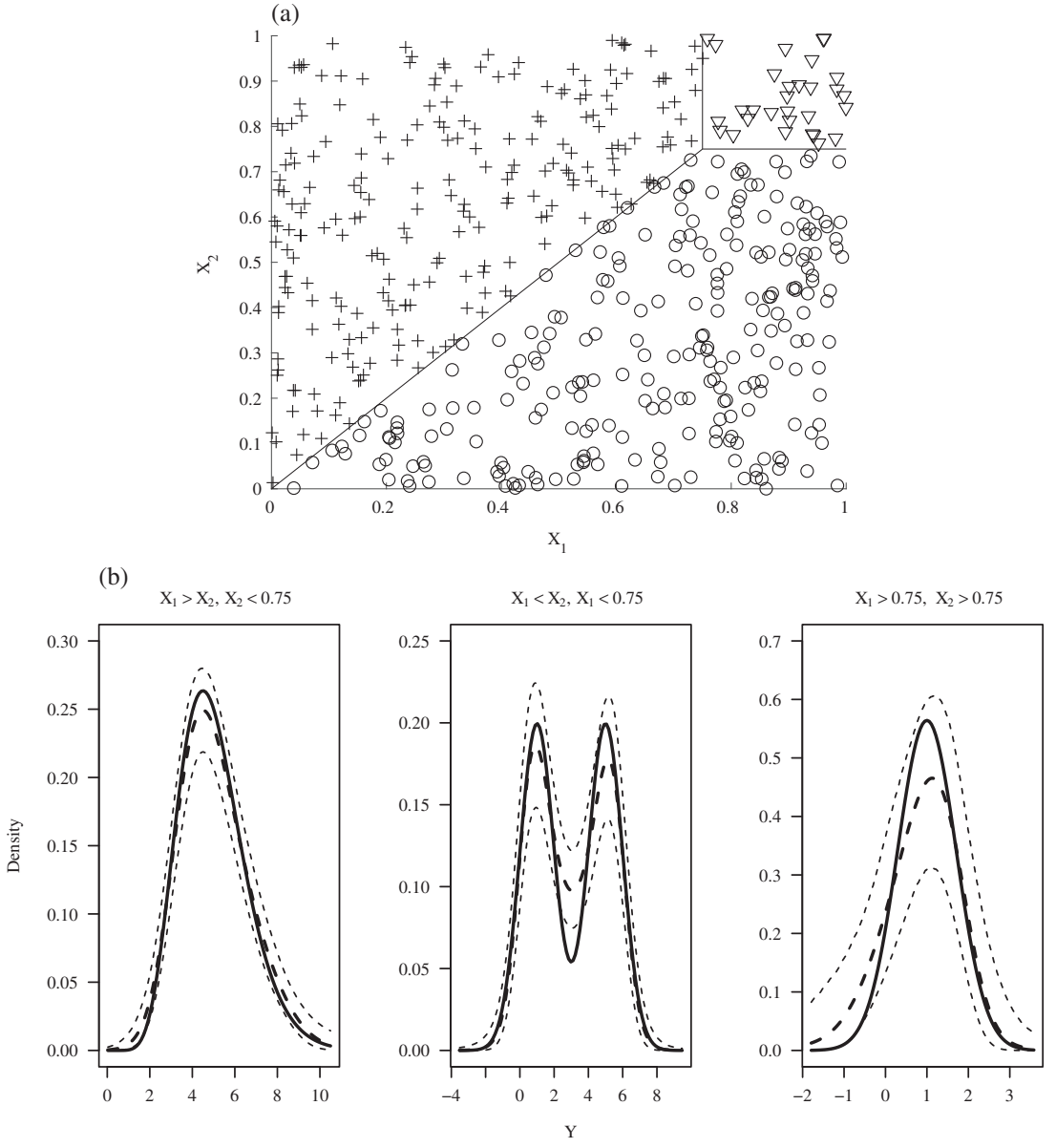

Fig. 2. (a): The partition of the covariate space with the highest posterior probability from the simulated dataset in § 4.2. The true partition boundaries are denoted by the lines. The shape of the points designate the posterior partition to which each point belongs. (b): The posterior mean (dashed, bold), 95% credible intervals (dashed) and the true density (solid) in each partition element from the partition model.

at the points away from the boundary. The other two smooth methods had poorer fits, and their plots are provided in the Supplementary Material.

This dataset was also analysed with a Voronoi partition model which assumes normality within each partition element, as described in Denison et al. (2002b, Ch. 7). The Markov chain ran for 80 000 iterations following a 20 000-iteration burn-in period. Since this partition model assumes normality within each partition element, model averaging was employed to obtain the posterior distribution at the same locations as the methods previously discussed. A plot of the result is

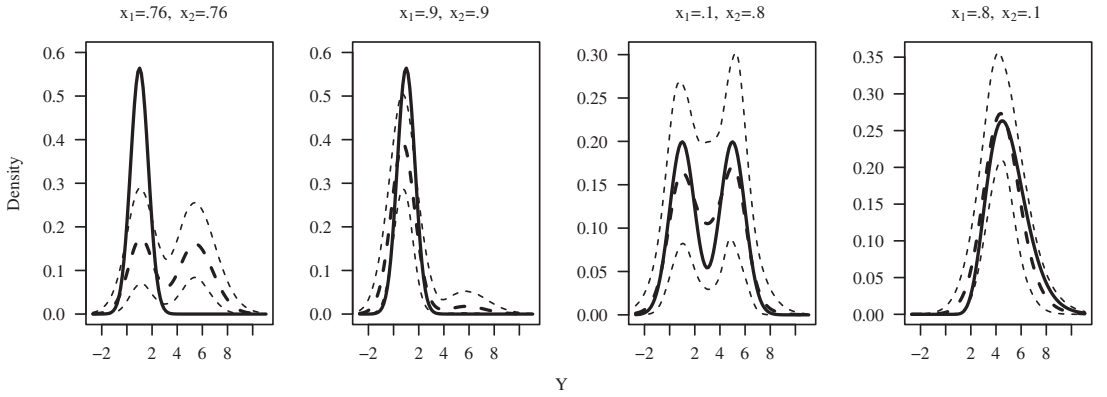

Fig. 3. The posterior mean density (dashed, bold), 95% credible intervals (dashed) and the true density (solid) of the Bayesian Dirichlet process mixture of normals model.

provided in the Supplementary Material and indicates that the normality modelling assumption is too restrictive.

A direct comparison of this dataset to other partition models was not possible since Petralia et al. (2013) did not provide public code, and the code provided by Ma (2017) lacks adequate documentation. However, we replicate one of the scenarios in Ma (2017) where, for a small sample size, our method outperforms Ma (2017). The details of this comparison are given in the Supplementary Material.

4.3. *Smooth changes*

To demonstrate the partition model's performance on data where the density changes smoothly as a function of the covariates, 500 realizations of the vector (Y, X_1, X_2) were simulated from a trivariate normal distribution with mean $(1, 5, 7.5)$ and covariance matrix Σ, with $\Sigma_{i,i} = 1$, $\Sigma_{i,j} = 0.5$ if $|i-j| = 1$, and $\Sigma_{i,j} = 0.1$ if $|i-j| = 2$.

The Markov chain was run for 10^5 iterations. Figure 4(a) shows the partition with the highest posterior probability; the other three panels plot the posterior mean, 95% credible intervals and the true distribution at the tessellation centre. Since we are estimating the marginal density of y in each region, there is no guarantee that the distribution at the tessellation centre will be contained in the credible intervals, but the densities are used as a rough benchmark for where the marginal density should lie. The partition model performs as expected: it identifies that both X_1 and X_2 influence the distribution of y and captures general shifts in the distribution over the covariate space.

4.4. *Wind turbine data*

Conditional density estimation is particularly useful when it is unclear how the density of y changes with respect to the predictors, x. The exact relationship of electrical power output in wind turbines to covariates measuring wind speed, wind direction, air density, wind shear and turbulence intensity is unknown, but is believed to be nonlinear (Lee et al., 2015). Furthermore, the conditional density of power output is known to have sharp changes as a function of wind speed and direction. For instance, modern wind turbines employ a pitch control mechanism to protect the generator under high wind. When the wind speed is high enough, the turbine blades start turning more parallel into the wind to reduce the energy absorption capability. Thus, for very high wind speeds, the power output is concentrated near a maximum power level. This change

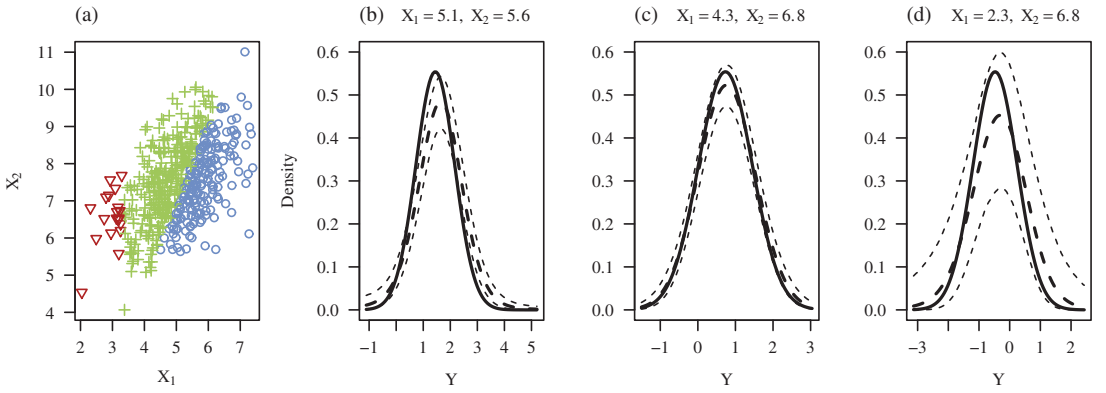

Fig. 4. Panel (a) shows the partition of the covariate space with the highest posterior probability. The colours/shapes designate partition assignments. The remaining three panels show the posterior mean density (dashed bold), 95% credible intervals (dashed) and the density at the corresponding Voronoi centre (solid) in each of the three posterior partition regions.

from a very broad and often skewed distribution to a very narrow distribution can happen within a wind speed range of just 1 m/s. The change in the spread of power output as a function of wind speed can be seen in Fig. 1.

Wind speed is not the only covariate which induces sharp changes in the distribution of power output. When terrain is not flat and smooth, there may also be sharp changes in power output as a function of wind direction, as the wind is affected by the terrain surrounding the turbine. When a wind turbine is downwind of another turbine, the upwind turbine creates a narrow wake region and when the downwind turbine is in the wake of the upwind turbine, it produces noticeably less power, due to the reduction of kinetic energy in the air flow after the rotor of the upwind turbine. The difference in power outputs of the downwind turbine in and out of the wake region also presents a sharp change in the power output's distribution. Combining all of these effects together, we expect sharp changes in power output to occur in both wind speed and wind direction.

The wind turbine dataset consists of aggregated measurements of wind turbine power output and several covariates over 10 minute increments from a single turbine, consisting of 10 000 observations randomly sampled from a larger dataset. We analyse the wind turbine dataset in order to estimate the distribution of power output in various regions of the covariate space and thus better understand where the most important changes occur. The partition model was fitted to the data using five covariates: wind speed, wind direction, air density, wind shear and turbulence intensity. The reversible jump Markov chain Monte Carlo algorithm was run for 100 000 iterations.

Figure 5 shows the tessellation with the highest posterior probability. The partition structure identifies major changes in power output across wind speed, and also identifies one interesting change across wind direction at about 120°. This change in power output is believed to be caused by a wake effect from another turbine which caused the instruments measuring wind speed to underestimate wind speed when the turbine is oriented at about 120°. The partition model also successfully captures the wind speed at which the maximum power output is achieved, and can be seen at the partition boundary located near 12 m/s. Overlaps and blurred edges of the partitions indicate the role of other covariates in determining the partition structure. Figure 6 demonstrates the various forms the posterior densities take in four of the ten tessellation regions.

Currently, the choice of $M_{\max}$ is done in an ad hoc manner, taking into consideration computation and data size. It is important that $M_{\max}$ is large enough to capture the distributional changes

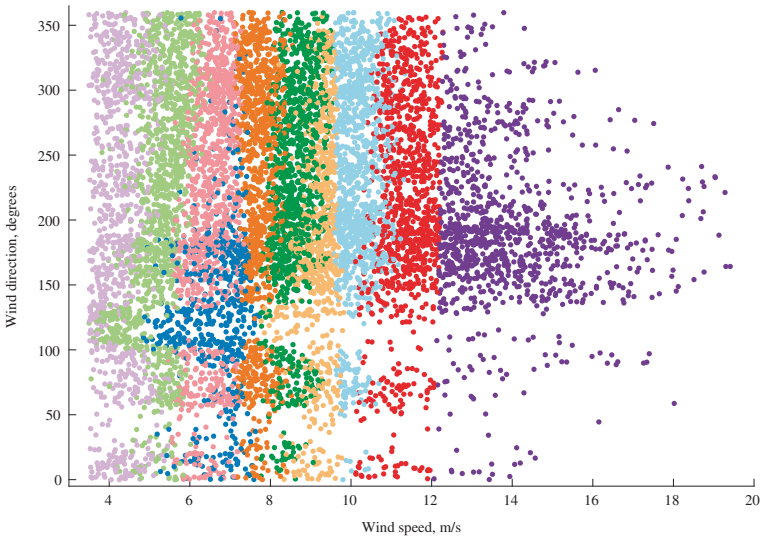

Fig. 5. Wind direction plotted against wind speed, with colours denoting the partition regions of the tessellation with the highest posterior probability.

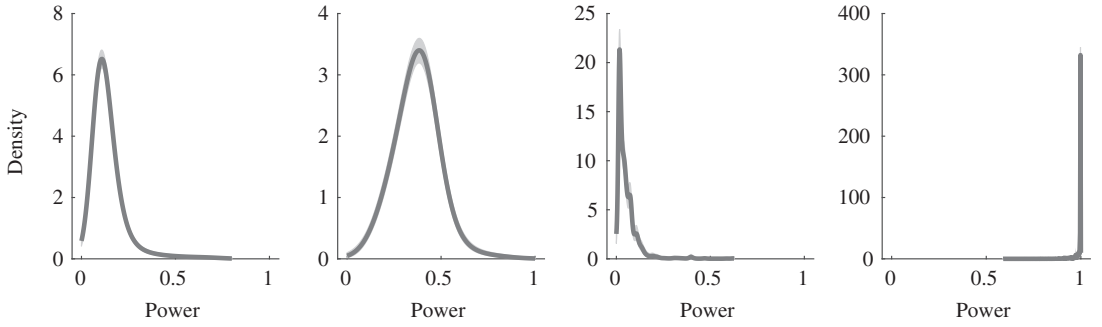

Fig. 6. Posterior densities with 90% credible intervals (shaded) in four of the ten partitions of the final wind turbine posterior tessellation.

in y, but not so large as to be computationally intractable. A theoretically optimum value of $M_{\max}$ for this partition model is still an open question.

ACKNOWLEDGEMENT

The authors thank the reviewers and associate editor for their detailed review, which greatly improved the final presentation of this material. This research was supported by the National Cancer Institute of the National Institutes of Health, the National Science Foundation and the University of Massachusetts Lowell. Payne and Guha contributed equally to this article.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains details of the proofs, the parallel tempering used in the reversible jump Markov chain Monte Carlo algorithm, details of the Markov chain performance, and additional information and comparisons. Code is available at

`https://github.com/richardbayes/bayes-cde` and utilizes the `GPstuff` Matlab code by Vanhatalo et al. (2013).

References

Barrientos, A. F., Jara, A. & Quintana, F. A. (2017). Fully nonparametric regression for bounded data using dependent Bernstein polynomials. *J. Am. Statist. Assoc.* **112**, 806–25.

Bhattacharya, A. & Dunson, D. B. (2010). Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika* **97**, 851–65.

Chipman, H. A., George, E. I. & McCulloch, R. E. (1998). Bayesian CART model search. *J. Am. Statist. Assoc.* **93**, 935–48.

Chung, Y. & Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Am. Statist. Assoc.* **104**, 1646–60.

Denison, D. G. T., Adams, N. M., Holmes, C. C. & Hand, D. J. (2002a). Bayesian partition modelling. *Comp. Statist. Data Anal.* **38**, 475–85.

Denison, D. G. T. & Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics* **57**, 143–9.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. (2002b). *Bayesian Methods for Nonlinear Classification and Regression*. New York: John Wiley & Sons.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85**, 363–77.

Dunson, D. B. & Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–23.

Dunson, D. B., Pillai, N. & Park, J.-H. (2007). Bayesian density regression. *J. R. Statist. Soc. B* **69**, 163–83.

Fan, J., Yao, Q. & Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.

Fu, G., Shih, F. Y. & Wang, H. (2011). A kernel-based parametric method for conditional density estimation. *Pat. Recog.* **44**, 284–94.

Ghosal, S. & Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann. Statist.* **34**, 2413–29.

Gramacy, R. B. (2007). tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *J. Statist. Software* **19**, 1–46.

Gramacy, R. B. & Taddy, M. (2010). Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed Gaussian process models. *J. Statist. Software* **33**, 1–48.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

Griffin, J. E. & Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Am. Statist. Assoc.* **101**, 179–94.

Holmes, C. C., Denison, D. G. T., Ray, S. & Mallick, B. K. (2005). Bayesian prediction via partitioning. *J. Comp. Graph. Statist.* **14**, 811–30.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp.* **3**, 79–87.

Jara, A. & Hanson, T. E. (2011). A class of mixtures of dependent tail-free processes. *Biometrika* **98**, 553–66.

Kim, H.-M., Mallick, B. K. & Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *J. Am. Statist. Assoc.* **100**, 653–68.

Kooperberg, C. & Stone, C. J. (1991). A study of logspline density estimation. *Comp. Statist. Data Anal.* **12**, 327–47.

Kundu, S. & Dunson, D. B. (2011). Latent factor models for density estimation. *arXiv:* 1108.2720v2.

Lee, G., Ding, Y., Genton, M. G. & Xie, L. (2015). Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *J. Am. Statist. Assoc.* **110**, 56–67.

Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Am. Statist. Assoc.* **83**, 509–16.

Ma, L. (2017). Recursive partitioning and multi-scale modeling on conditional densities. *Electron. J. Statist.* **11**, 1297–325.

Ma, L. & Wong, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *J. Am. Statist. Assoc.* **106**, 1553–65.

Müller, P., Erkanli, A. & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Norets, A. & Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *J. Economet.* **168**, 332–46.

Pati, D., Dunson, D. B. & Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multi. Anal.* **116**, 456–72.

Petralia, F., Vogelstein, J. T. & Dunson, D. B. (2013). Multiscale dictionary learning for estimating conditional distributions. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger, eds. Red Hook, NY: Curran Associates, Inc., pp. 1797–805.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

RIIHIMÄKI, J. & VEHTARI, A. (2014). Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Anal.* **9**, 425–48.

SHEN, W. & GHOSAL, S. (2016). Adaptive Bayesian density regression for high-dimensional data. *Bernoulli* **22**, 396–420.

SMOLA, A. J., SCHÖLKOPF, B. & MÜLLER, K.-R. (1998). The connection between regularization operators and support vector kernels. *Neural Networks* **11**, 637–49.

STONE, C. J., HANSEN, M. H., KOOPERBERG, C. & TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *Ann. Statist.* **25**, 1371–470.

TOKDAR, S. T. & GHOSH, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *J. Statist. Plan. Infer.* **137**, 34–42.

TOKDAR, S. T., ZHU, Y. M. & GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5**, 319–44.

VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Statist.* **1**, 433–48.

VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36**, 1435–63.

VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pp. 200–22. Bethesda, MD: Institute of Mathematical Statistics.

VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12**, 2095–119.

VANHATALO, J., RIIHIMÄKI, J., HARTIKAINEN, J., JYLÄNKI, P., TOLVANEN, V. & VEHTARI, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.* **14**, 1175–9.

[*Received on* 10 *February* 2017. *Editorial decision on* 16 *May* 2019]