Bayesian isotonic density regression

BY LIANMING WANG

*Department of Statistics, University of South Carolina, 1523 Greene Street, Columbia,
South Carolina 29208, U.S.A.*

wang99@mailbox.sc.edu

AND DAVID B. DUNSON

*Department of Statistical Science, Duke University, Box 90251, Durham,
North Carolina 27708, U.S.A.*

dunson@stat.duke.edu

SUMMARY

Density regression models allow the conditional distribution of the response given predictors to change flexibly over the predictor space. Such models are much more flexible than nonparametric mean regression models with nonparametric residual distributions, and are well supported in many applications. A rich variety of Bayesian methods have been proposed for density regression, but it is not clear whether such priors have full support so that any true data-generating model can be accurately approximated. This article develops a new class of density regression models that incorporate stochastic-ordering constraints which are natural when a response tends to increase or decrease monotonely with a predictor. Theory is developed showing large support. Methods are developed for hypothesis testing, with posterior computation relying on a simple Gibbs sampler. Frequentist properties are illustrated in a simulation study, and an epidemiology application is considered.

Some key words: Conditional density estimation; Dependent Dirichlet process; Hypothesis test; Isotonic regression; Nonparametric Bayes; Quantile regression; Stochastic ordering.

1. INTRODUCTION

In studying the relationship between a continuous predictor x and a response variable y , it is common to have prior knowledge of stochastic ordering. For example, in environmental epidemiology studies, one can assume that the severity of an adverse health outcome is stochastically nondecreasing with dose of a potentially adverse exposure. Incorporation of stochastic-ordering constraints can improve the efficiency in estimating conditional distributions and in testing for local or global changes as x increases. The focus of this article is on addressing these problems using nonparametric Bayesian methods.

Letting $F_x(y) = \text{pr}(Y \leq y \mid X = x)$ denote the conditional distribution function of the response given the predictor, we assume prior knowledge is available that F_x is stochastically smaller than $F_{x'}$ for any two points $x \leq x'$ in $\mathcal{X}$, with $\mathcal{X}$ the real line or an interval on the real line. Such a restriction is natural when the response is known to increase monotonically with the predictor. We otherwise have substantial uncertainty about $\{F_x, x \in \mathcal{X}\}$, with the true conditional distribution potentially having any form subject to the stochastic-ordering restriction. It is appealing to

choose a prior with full support on the space of stochastically ordered conditional distributions. Full support means that the prior can generate conditional distributions within arbitrarily small neighbourhoods of any true conditional distribution. If the true conditional distribution is not in the support of the prior, the posterior will not concentrate on small neighbourhoods of the truth, and hence an unrealistic and inconsistent estimate will be obtained.

Although using a prior with large support is critical in obtaining robust inferences and appropriately accounting for uncertainty in the true data-generating model, few results are available characterizing the support of the prior in Bayesian density regression models even in the unconstrained case. An important exception is the recent article of [Tokdar et al. \(2010\)](#), who showed large support and posterior consistency in unconstrained density estimation based on a class of logistic Gaussian process priors. Their approach was specific to the logistic Gaussian process, relying heavily on theoretical properties of Gaussian processes. [Norets \(2010\)](#) showed large Kullback–Leibler support for finite mixtures of normal regressions, with the mixing weights potentially dependent on predictors, and unpublished work by D. Pati and D. B. Dunson from Duke University and A. Norets and J. Pelenis from Princeton University independently obtained large support and posterior consistency results for various density regression priors. All of the previously proposed Bayesian models for incorporating monotonicity constraints in nonparametric regression with a continuous predictor violate the full support condition, and make restrictive assumptions about the nature of dependence on the predictor.

For example, the typical focus in the literature has been on nonparametric regression models of the form $y_i = f(x_i) + \epsilon_i$, where f is an unknown nondecreasing function and $\epsilon_i \sim F$ is a residual. One can then induce a prior on the conditional distribution through independent priors for the regression function f and residual distribution F . Using this general idea, [Lavine & Mockus \(1995\)](#) allowed both f and F to be unknown using Dirichlet process priors ([Ferguson, 1973, 1974](#)); [Neelon & Dunson \(2004\)](#) instead used adaptive linear splines for f ; [Holmes & Held \(2006\)](#) used a piecewise constant model; [Wang \(2008\)](#) used approximated monotone functions with cubic splines allowing an unknown number and location of knots; [Bornkamp & Ickstadt \(2009\)](#) modelled the monotone function with a mixture of parametric probability distributions with the mixing measure assigned a stick-breaking prior ([Ongaro & Cattaneo, 2004](#)); [Shively et al. \(2009\)](#) proposed Bayesian nonparametric approaches relying on the characterization of monotone functions proposed by [Ramsay \(1998\)](#) and quadratic regression splines. Even choosing flexible priors for f and F leads to a prior on $\{F_x, x \in \mathcal{X}\}$ with small support, as conditional distributions having residual variance or shape varying with x are not supported.

There has been some work on more flexible nonparametric Bayes methods for categorical x . [Arjas & Gasbarra \(1996\)](#) characterized stochastically ordered survival functions in two groups using a piecewise constant hazard model on a fixed grid, with stochastic ordering implying a restriction on the interval-specific hazards. [Gelfand & Kottas \(2001\)](#) instead expressed the distribution functions in two groups as $F_1 = G_1$ and $F_2 = G_1 G_2$, with independent Dirichlet process ([Ferguson, 1973, 1974](#)) mixture priors used for G_1 and G_2 . Although this approach enforces stochastic ordering, the implied constraint is more restrictive and many stochastically ordered distributions are not supported. [Karabatsos & Walker \(2007\)](#) proposed alternative models for stochastically ordered distributions based on Pólya tree ([Ferguson, 1974](#)) or Bernstein polynomial ([Petrone, 1999](#)) priors, but did not show large support. Without support results for the above priors, it is not clear whether the priors rule out a large subset of the set of stochastically ordered distributions a priori.

[Hoff \(2003a\)](#) proposed a latent variable specification in which individual i has potential outcomes in each of K groups, with these outcomes following a partial ordering. By placing a Dirichlet process prior on the joint distribution of the potential outcomes, he induced a prior on the

marginal distributions in each group having full support on the space of distributions satisfying a partial stochastic ordering. [Dunson & Peddada \(2008\)](#) generalized this approach to a mixture specification by proposing a restricted dependent Dirichlet process, while also developing methods for hypothesis testing. The models, theory of large support and algorithms for computation are not straightforward to modify to accommodate continuous predictors. The continuous case is much more challenging, since instead of a finite collection of unknown distributions, we are faced with an uncountable collection.

2. STOCHASTICALLY ORDERED DENSITIES

2.1. The restricted dependent Dirichlet process mixture

Let $f_{\mathcal{X}} = \{f_x, x \in \mathcal{X}\}$ denote a collection of conditional densities indexed by the predictor $x \in \mathcal{X}$. To induce a prior on $f_{\mathcal{X}}$, we consider the mixture model

$$f_x(y; P) = \int \tau^{1/2} \phi \left[\tau^{1/2} \{y - \mu(x)\} \right] dP_x \{ \mu(x), \tau \}, \quad (1)$$

where $\phi(\cdot)$ denotes the standard normal density, $\mu = \{\mu(x), x \in \mathcal{X}\}$ is a real-valued stochastic process over $\mathcal{X}$, $\tau \in \mathbb{R}^+$ is a scalar precision parameter and $P = (P_x, x \in \mathcal{X})$ is a collection of predictor-dependent mixing measures.

A prior for $f_{\mathcal{X}}$ can be induced through a prior for P . One such prior is the fixed p -dependent Dirichlet process ([MacEachern, 1999](#)), which can be expressed in stick-breaking form as

$$P_x = \sum_{h=1}^{\infty} p_h \delta_{\Theta_h(x)}, \quad p_h = V_h \prod_{l < h} (1 - V_l), \quad (2)$$

where δ_a is a Dirac probability measure concentrated at a , $V_h \sim \text{Be}(1, \alpha)$ and $\Theta_h = (\tilde{\mu}_h, \tilde{\tau}_h) \sim M$ are mutually independent for $h = 1, \dots, \infty$, and $M = H \otimes L$ is a baseline probability measure. For this choice of prior for $P = (P_x, x \in \mathcal{X})$, (1) becomes

$$f_x(y; P) = \sum_{h=1}^{\infty} p_h \tilde{\tau}_h^{1/2} \phi \left[\tilde{\tau}_h^{1/2} \{y - \tilde{\mu}_h(x)\} \right], \quad (3)$$

and the conditional density f_x is marginally assigned a Dirichlet process location-scale mixture of Gaussians prior ([Lo, 1984](#); [Escobar & West, 1995](#)). The f_x and $f_{x'}$ are not independent under this dependent Dirichlet process mixture prior but are correlated due to the common mixture weights $\{p_h\}$ and scale parameters $\{\tilde{\tau}_h\}$ and to dependence in the kernel locations $\tilde{\mu}_h(x)$ and $\tilde{\mu}_h(x')$. By letting $\tilde{\tau}_h = \tau$ to fix the kernel bandwidth across the mixture components, the model simplifies to a location mixture. Allowing the bandwidth to vary across components has some well-known advantages in terms of sparsely characterizing unknown densities.

The above model specification closely follows that proposed in previous work. In an as-yet unpublished paper, D. Pati and D. B. Dunson recently provided conditions for large support and weak and strong posterior consistency for fixed p -dependent Dirichlet processes for unconstrained density regression. Their theory does not apply automatically to the constrained case, and it is not clear whether the prior can be specified in such a way as to restrict support to a large subset of the space of stochastically nondecreasing densities.

Assumption 1. Assume the following continuity condition holds: $f_x(\cdot) \rightarrow f_{x'}(\cdot)$ as $x \rightarrow x'$, for any $x, x' \in \mathcal{X}$.

From an applied perspective, it is appealing to focus on conditional densities satisfying the continuity condition stated in Assumption 1. This regularity condition rules out sudden jumps in f_x as x changes.

Letting (y_i, x_i) denote the data for subject i , for $i = 1, \dots, n$, (3) implies that

$$y_i \sim N\{\mu_i(x_i), \tau_i^{-1}\}, \quad (\mu_i, \tau_i) \sim P, \quad P = \sum_{h=1}^{\infty} p_h \delta_{(\tilde{\mu}_h, \tilde{\tau}_h)}. \quad (4)$$

Hence, to sample y_i first generate a random function $\mu_i : \mathcal{X} \rightarrow \mathfrak{R}$ and precision $\tau_i \in \mathfrak{R}^+$ from a discrete distribution, which is assigned a Dirichlet process prior. Then, draw y_i from a normal distribution with precision τ_i^{-1} centred on the realization of μ_i at predictor value x_i . The function μ_i can be viewed as a latent trajectory defining the mean of the normal kernel for individual i at each possible value of the predictor. With this interpretation, a natural strategy for restricting the induced conditional density to be nondecreasing with x is to restrict the μ_i s to be nondecreasing functions. To also incorporate the continuity condition, one can draw $\tilde{\mu}_h \sim H$ independently from $\tilde{\tau}_h \sim L$, with H chosen as a probability measure on $S_{\mathcal{X},c}$, the space of continuous nondecreasing $\mathcal{X} \rightarrow \mathfrak{R}$ functions.

Theorem 1 states that restricting H to have support on $S_{\mathcal{X},c}$ is a necessary and sufficient condition for the collection of conditional densities to be stochastically nondecreasing in the index x and subject to the continuity condition with probability one.

THEOREM 1. *Let $D_{\mathcal{X}}$ denote the space of collections of densities indexed by $x \in \mathcal{X}$ and subject to Assumption 1 and the nondecreasing stochastic ordering constraint. Under (2) and (3), the following statements are equivalent: (a) $f_{\mathcal{X}} \in D_{\mathcal{X}}$ with probability one; and (b) H is a probability measure with support on $S_{\mathcal{X},c}$.*

Theorem 1 also holds for a simplified version of (3) with common unknown variance in all of the normal components.

2.2. Hypothesis testing

We also develop methods for testing equalities in conditional distributions against stochastically ordered alternatives. Letting $D(x, x') = \sup_{y \in \mathfrak{R}} |F_x(y) - F_{x'}(y)|$ as a measure of distance between $F_x(y) = \int_{-\infty}^y f_x(z) dz$ and $F_{x'}(y) = \int_{-\infty}^y f_{x'}(z) dz$ for $x < x'$, we consider the hypotheses $H_0(x, x') : D(x, x') < \epsilon$ versus $H_1(x, x') : D(x, x') \geq \epsilon$, where ϵ is a small positive constant chosen so that $D(x, x') < \epsilon$ implies that F_x and $F_{x'}$ are not significantly different. We recommend $\epsilon = 0.05$ as a default value. As a weight of evidence in favour of $H_1(x, x')$, we use the Bayes factor. Fixing x at the minimum predictor value observed and varying x' , we obtain a log Bayes factor function, which is useful for identifying threshold predictor values, a common interest in biomedical studies. Global hypothesis tests of near equality in the conditional distributions across a region $A = [a, b] \subset \mathcal{X}$ can be formulated as $H_0(A) = \bigcap_{x, x' \in A} H_0(x, x')$. We use the supremum of the log Bayes factor function as a weight of evidence in favour of $H_1(A)$.

3. PRIOR SPECIFICATION AND POSTERIOR COMPUTATION

To complete a specification of the prior under the model proposed in § 2.1, first let L correspond to the $\text{Ga}(a_\tau, b_\tau)$ distribution and choose $\alpha \sim \text{Ga}(a_\alpha, b_\alpha)$ to allow the Dirichlet process concentration parameter to be unknown. In addition, H is chosen to induce a prior for $f_{\mathcal{X}}$ with

support on a large subset of $D_{\mathcal{X}}$ by letting

$$\tilde{\mu}_h(x) = \sum_{l=0}^k \beta_{hl} b_l(x) \quad (5)$$

for all $x \in \mathcal{X}$, with $b_0(x) = 1$, $\{b_l\}_{l=1}^k$ monotone spline basis functions (Ramsay, 1988), and $\{\beta_{hl}\}_{l=0}^k$ basis coefficients having nonnegative constraint for $l \geq 1$. We focus in our applications on splines of degree 1, leading to piecewise linearity. It is well known that any continuous function can be approximated arbitrarily well by a piecewise linear function with sufficiently many knots.

Potentially, one can treat the number and locations of knots as unknown and adaptively determine their values by using reversible jump Markov chain Monte Carlo (Green, 1995). To facilitate efficient computation, we instead choose a large number of equally-spaced knots and then adaptively drop unnecessary basis functions by letting their coefficients have zero values (Smith & Kohn, 1996). This is accomplished through the following prior for β_h ,

$$\pi(\beta_h; m_0, v_0, \lambda, \pi_0) = N(\beta_{h0}; m_0, v_0^{-1}) \prod_{l=1}^k \{\pi_0 \delta_0(\beta_{hl}) + (1 - \pi_0) \text{Exp}(\beta_{hl}; \lambda)\},$$

where $m_0 \sim N(\omega_0, \eta_0)$, $v_0 \sim \text{Ga}(a_{v_0}, b_{v_0})$, $\pi_0 \sim \text{Be}(a_\pi, b_\pi)$ is the prior probability that a basis function is excluded from the model, $\lambda \sim \text{Ga}(a_\lambda, b_\lambda)$, and $\text{Exp}(\lambda)$ denotes the exponential distribution with mean λ^{-1} . The exponential distribution penalizes large values of the nonzero spline coefficients.

For posterior computation, we use the exact block Gibbs sampler (Yau et al., 2011), which bypasses the need to update the infinitely-many unknowns in the model by combining the ideas of retrospective sampling (Papaspiliopoulos & Roberts, 2008) and slice sampling (Walker, 2007). This algorithm avoids using truncation of the stick-breaking process as in Ishwaran & James (2001). Let N be the number of atoms sampled at the current iteration of the Gibbs sampler, $K = (K_1, \dots, K_n)^\top$ with $K_i = h$ if $\mu_i = \tilde{\mu}_h$ and $\tau_i = \tilde{\tau}_h$, and $c_h = \sum_{i=1}^n 1_{(K_i=h)}$ for $h = 1, \dots, \infty$. The Gibbs sampler cycles through the following steps.

Algorithm 1. The Gibbs sampler.

Step 1. Sample $\beta_h = (\beta_{h0}, \beta_{h1}, \dots, \beta_{hk})^\top$ ($h = 1, \dots, N$).

- (i) If $c_h = 0$ or $\sum_{i:K_i=h} b_l^2(x_i) = 0$, sample β_{h0} from $N(m_0, v_0)$, and sample β_{hl} from the mixture distribution $\pi_0 \delta_0 + (1 - \pi_0) \text{Exp}(\lambda)$ for $l = 1, \dots, k$.
- (ii) If $c_h > 0$ and $\sum_{i:K_i=h} b_l^2(x_i) > 0$, sample β_{h0} from $N(E_{h0}, W_{h0})$, where

$$W_{h0} = (\tilde{\tau}_h c_h + v_0)^{-1}, \quad E_{h0} = W_{h0} \left[\tilde{\tau}_h \sum_{i:K_i=h} \left\{ y_i - \sum_{l \neq 0} \beta_{hl} b_l(x_i) \right\} + v_0 m_0 \right].$$

For $l = 1, \dots, k$, sample β_{hl} from $\pi_{hl} \delta_0 + (1 - \pi_{hl}) N_+(E_{hl}, W_{hl})$, where

$$\pi_{hl} = \left\{ 1 + \frac{(1 - \pi_0) \lambda \Phi(W_{hl}^{-1/2} E_{hl})}{\pi_0 \phi(-W_{hl}^{-1/2} E_{hl}) W_{hl}^{-1/2}} \right\}^{-1}, \quad W_{hl} = \left\{ \tilde{\tau}_h \sum_{i:K_i=h} b_l^2(x_i) \right\}^{-1},$$

$$E_{hl} = W_{hl} \left[\tilde{\tau}_h \sum_{i:K_i=h} b_l(x_i) \left\{ y_i - \sum_{l' \neq l} \beta_{hl'} b_{l'}(x_i) \right\} - \lambda \right].$$

In the above, $\Phi(\cdot)$ denotes the cumulative distribution function associated with $N(0, 1)$, and $N_+(a, b)$ denotes the truncated normal distribution of $N(a, b)$ with support $(0, \infty)$.

Step 2. Sample $\tilde{\tau}_h$ from $\text{Ga}(a_\tau + c_h/2, b_\tau + \sum_{K_i=h} \{y_i - \sum_{l=0}^k \beta_{hl} b_l(x_i)\}^2/2)$ for $h = 1, \dots, N$.

Step 3. Sample V_h from $\text{Be}(1 + c_h, \alpha + n - \sum_{j=1}^h c_j)$ for $h = 1, \dots, N$.

Step 4. Sample U_i from $\text{Un}(0, p_{K_i})$ for $i = 1, \dots, n$.

Step 5. Sample $K_i, i = 1, \dots, n$. Let $U^* = \min(U_1, \dots, U_n)$.

(i) If $\sum_{h=1}^N p_h > 1 - U^*$, sample K_i from a multinomial distribution with

$$\text{pr}(K_i = j | \cdot) \propto \tilde{\tau}_j^{1/2} \phi \left[\tilde{\tau}_j^{1/2} \left\{ y_i - \sum_{l=0}^k \beta_{jl} b_l(x_i) \right\} \right] 1_{(p_j > U_i)} \quad (i = 1, \dots, n).$$

(ii) Otherwise, keep updating $N = N + 1$ and sampling (V_N, β_N, τ_N) from their priors until $\sum_{h=1}^N p_h > 1 - U^*$. Then sample K_i s as in (a).

Step 6. Sample π_0 from $\text{Be}(a_{\pi_0} + \sum_{h=1}^N \sum_{l=1}^k 1_{(\beta_{hl}=0)}, b_{\pi_0} + \sum_{h=1}^N \sum_{l=1}^k 1_{(\beta_{hl}>0)})$.

Step 7. Sample α from $\text{Ga}\{a_\alpha + N, b_\alpha - \sum_{h=1}^N \log(1 - V_h)\}$.

Step 8. Sample λ from $\text{Ga}(a_\lambda + \sum_{h=1}^N \sum_{l=1}^k 1_{(\beta_{hl}>0)}, b_\lambda + \sum_{h=1}^N \sum_{l=1}^k \beta_{hl})$.

Step 9. Sample v_0 from $\text{Ga}\{a_{v_0} + N/2, b_{v_0} + \sum_{h=1}^N (\beta_{h0} - m_0)^2/2\}$.

Step 10. Sample m_0 from $N\{(\tau_0^{-1} + N v_0)^{-1}(\tau_0^{-1} \omega_0 + v_0 \sum_{h=1}^N \beta_{h0}), (\tau_0^{-1} + N v_0)^{-1}\}$.

In Algorithm 1, steps 4 and 5 allow the number of clusters N to increase as needed during the sampling process. Because all the unknowns have full conditional posterior distributions that are easy to sample from directly, the algorithm is simple to implement and computation is fast per Gibbs iteration. Many functionals, such as conditional densities, distribution functions, quantile functions and posterior hypothesis probabilities, can be estimated easily from the Markov chain Monte Carlo samples.

Nonparametric Bayes models involve infinitely-many parameters and are typically specified so that higher level parameters in the hierarchy are under-identified from a frequentist perspective and there are many values of these parameters that lead to similar values of the observed data likelihood. In this sense, the models share a close connection with parametric models that arise from redundant parameterizations (Gelman, 2004, 2006; Ghosh & Dunson, 2009). In both cases, we expect the redundancy and under-identification to lead to poor mixing for higher level unknowns that are typically not of inferential interest, while obtaining good mixing for unknowns that directly appear in the observed data likelihood. For example, we anticipate slow mixing for some of the hyperparameters sampled in Steps 6 through 10 and even for the basis coefficients sampled in Step 1. As we start with an overcomplete set of potential basis functions, we fully expect that there are many different subsets of these bases that lead to similar induced collections of conditional densities. This will intrinsically lead to slow mixing for the basis coefficients that should not degrade and may even aid mixing of the conditional density values $f(y | x)$. Indeed, this is exactly the behaviour we have observed in comprehensive simulation and sensitivity analyses that are summarized in the Supplementary Material; rates of convergence and mixing are good for $f(y | x)$ and functions of these conditional densities, but are poor for a number of the hyperparameters.

Prior elicitation in nonparametric models is challenging, since the impact of the hyperparameters on the induced conditional densities $f(y | x)$ is opaque, though some insight can be obtained through sampling from the prior. This opaqueness is not specific to our proposed model and is a general feature of most nonparametric Bayes models even in simple cases. We follow the strategy of choosing a weakly informative prior (Gelman et al., 2008) that leads to good performance in a wide variety of cases, with the posterior densities for $f(y | x)$ substantially different from the priors even in small sample sizes such as $n = 25$. Specifically, we suggest choosing $\alpha \sim \text{Ga}(1, 1)$, $\lambda \sim \text{Ga}(1, 1)$, $\nu_0 \sim \text{Ga}(1, 1)$, $\pi_0 \sim \text{Un}(0, 1)$, $m_0 \sim N(0, 1)$ and $L = \text{Ga}(1, 1)$ after normalizing y . The Supplementary Material contains a detailed sensitivity analysis.

4. SIMULATION STUDIES

Two simulation cases were investigated to evaluate the performance of the proposed method: a null case and an alternative case. For each case, we tried different sample sizes $n = 25, 50, 100, 200$ and 500 . All x_i s were independently sampled from $\text{Un}(0, 1)$. In the null case, y_i was generated from

$$f(y | x) = 0.6 N(y; 0, 0.3^2) + 0.4 N(y; 1, 0.2^2),$$

and in the alternative case from

$$f(y | x) = 0.6 N\{y; 3(x - 0.3)^2 1_{(x > 0.3)}, 0.3^2\} + 0.4 N\{y; 5(x - 0.8)^{3/2} 1_{(x > 0.8)}, 0.2^2\}.$$

We generated 100 datasets for each set-up. For each data set, we ran the Gibbs sampler described in § 4. In specifying the nondecreasing piecewise linear basis functions, we chose I-splines (Ramsay, 1988) with degree equal to 1 and 19 equally spaced interior knots in the predictor space $(0, 1)$. Markov chain Monte Carlo was started with initial parameter values sampled from the recommended priors and $N = 20$, and results were summarized based on 2000 iterations after deleting the first 500 iterations as a burn-in for each dataset. Using 49 equally spaced interior knots did not give noticeably different results. As shown in the Supplementary Material, we ran a smaller number of simulations using 1100 iterations with a 100 iteration burn-in and obtained essentially identical results, justifying the use of short 2500 iteration chains in these simulations.

Figure 1 shows the results of conditional density estimation at predictor values $x = 0.10, 0.50, 0.75$ and 0.90 with the sample sizes $n = 25$ and 50 in the null case shown by the panels on the left and the alternative case shown by the panels on the right respectively. In the null case, the true conditional densities are all the same, while in the alternative case, they are substantially different across $\mathcal{X}$. The proposed method does a good job in estimating the conditional densities, with the performance improving with increased sample size as expected.

For global testing, we reject the global null if the supremum of the log Bayes factor function is larger than $\log(10)$ for each dataset. The choice of $\log(10)$ corresponds to a standard threshold for strong evidence against the null hypothesis (Kass & Raftery, 1995). The proportion of rejected global hypotheses is interpreted as the type 1 error of the test in the null case and the power in the alternative case. Our simulations yield type 1 errors equal to 0.05, 0.07, 0.03, 0.04 and 0.02 and powers equal to 0.79, 0.98, 1.0, 1.0 and 1.0 for sample size $n = 25, 50, 100, 200$, and 500 , respectively.

We are also interested in local hypothesis testing of differences between F_x and F_0 , with F_0 denoting the distribution of y given $x = 0$. We estimate a threshold $\hat{x} = \min\{x : \text{BF}(0, x) > 10\}$, the smallest predictor value at which the Bayes factor is larger than 10. There is significant evidence in favour of the local alternative hypothesis for $x > \hat{x}$. Figure 2 shows the box plots

Fig. 1. Estimates of the conditional densities $f(y|x)$ at $x = 0.1, 0.5, 0.75$, and 0.9 , respectively. The left panels correspond to the null case and the right panels correspond to the alternative case. Each plot shows the true density (solid), the average of the 100 estimated densities when sample size $n = 25$ (dash), and the average of the 100 estimated densities when $n = 50$ (dot-dash).

of the estimated thresholds obtained from 100 datasets in the alternative case with sample size $n = 25, 50, 100, 200$, and 500 , respectively. The true threshold is 0.45 . From Fig. 2, the centre of the box plot moves closer to the true threshold and the variation becomes smaller as sample size increases. The averages of the thresholds are $0.89, 0.74, 0.65, 0.57$, and 0.46 when sample size is $25, 50, 100, 200$, and 500 , respectively. Hence, for small sample sizes, there is a tendency to overestimate the threshold, which is as expected.

5. AN EPIDEMIOLOGIC APPLICATION

We apply the proposed method to an epidemiologic study relating 1,1-dichloro-2,2-bis (*p*-chlorophenyl)-ethylene, commonly referred to as *p,p'*-DDE, to risk of premature delivery. The study was based on the U.S. Collaborative Perinatal Project (Longnecker et al., 2001). Dunson & Park (2008) previously showed a relationship between DDE and the density of gestational age at delivery without incorporating the stochastic ordering constraint. As it is reasonable to assume that gestational length does not increase as dose increases, the stochastic-ordering assumption is biologically plausible. The proposed method allows for testing and identification of a threshold DDE level, which is of substantial interest from a public health perspective.

Fig. 2. Boxplots of thresholds $\hat{x}$ identified from the local hypothesis tests for 100 datasets in the alternative case of simulation with sample size $n = 25, 50, 100, 200$, and 500 , respectively. The dashed horizontal line shows the true value of threshold, 0.45 .

We analysed the data using a modification of (4), which adds $z_i^T \gamma$ to $\mu_i(x_i)$ to include a parametric adjustment for the covariates z_i , with z_i including race and standardized cholesterol and triglyceride levels for subject i . Letting $z = (z_1, \dots, z_n)^T$ with sample size $n = 2313$, we chose the prior $\gamma \sim N\{0, n(z^T z)^{-1}\}$, with the prior otherwise specified as in § 3. The Gibbs sampler of § 3 can be applied with a minor modification. We used 36 I spline basis functions $\{\tilde{b}_l\}_{l=1}^{36}$ resulting from taking 35 equally spaced interior knots within $(0, 180)$ and specifying degree equal to 1, and took $b_l = 1 - \tilde{b}_l$ for $l = 1, \dots, 36$ as basis functions in (5) in order to model stochastically nonincreasing densities. We based our results on 10 000 Markov chain Monte Carlo iterations collected after discarding the first 10 000 iterations as a burn-in.

Figure 3(a) shows the estimated conditional densities of gestational age at delivery for a DDE level equal to 2.5, 24.68, 36.53, 53.75 corresponding to the minimum, the 50th, 75th, and 90th percentiles of the observed DDE values, controlling the other covariates equal to zero. These density plots show an increasingly heavy left tail as DDE increases, implying an increasing risk of premature delivery. We also obtain substantially narrower credible intervals for the density estimates compared with Dunson & Park (2008), likely due to improved efficiency attributable to inclusion of the order constraint.

In order to assess the weight of evidence in the data in favour of stochastic ordering and to estimate a threshold dose, we apply the approach implemented in § 4. Let F_0 denote the distribution function of gestation length given $x = 2.5$, the minimum of the observed DDE levels, and let F_x denote the distribution function at DDE = x . Figure 3(b) plots the log Bayes factor curve for testing $H_0(2.5, x)$ versus $H_1(2.5, x)$ for increasing x . Treating Bayes factors greater than 10 as significant, we conclude in favour of $H_1(2.5, x)$ for all $x > 52$.

In addition to density estimation and hypothesis testing, the proposed method also yields a mean function estimate, as well as various quantile curves of gestational age at delivery changing with the DDE values. The proposed mean function estimate has much narrower 95% intervals than the estimate obtained by fitting the generalized additive model fitted with the gam function in R (R Development Core Team, 2011). We also estimate different quantile curves by using the proposed approach and by running the R function `rqss` with the nonincreasing constraint for a

Fig. 3. Estimated conditional densities of gestational age at delivery and the log Bayes factor curve used for global hypothesis testing. (a) The estimated conditional densities of gestational age at delivery at DDE values $x = 2.5$ (solid), 24.68 (dot), 36.53 (dash), 53.75 (dot-dash), respectively. (b) the log Bayes factor curve $\log\text{-BF}(2.5, x)$ (solid). The three dashed horizontal lines in the right panel are $\log(10)$, $\log(3)$ and 0 for reference.

nonparametric additive quantile regression model. The estimates from both methods look similar overall, but the estimates from the proposed method are much smoother. The proposed method should be more efficient in estimating different quantile curves since it estimates all of them jointly while existing methods usually estimate them separately. Matlab codes for the simulation study and data analysis are available upon request from the first author.

6. STOCHASTICALLY ORDERED DISCRETE DISTRIBUTIONS

When the response variable is discrete, it is useful to consider a simplification in which $y_i \sim P_{x_i}$, with $P = (P_x, x \in \mathcal{X})$ assigned the prior in (2) simplified to let $\Theta_h = \tilde{\mu}_h \sim H$. A restricted dependent Dirichlet process prior for P results by constraining H to have support on the space of nondecreasing functions. We show this leads to a prior with large support on the space of nondecreasing stochastically ordered distributions.

Remark 1. When $y_i \in \{0, 1, \dots, \infty\}$ is a count, H should have support on the space of nondecreasing $\mathcal{X} \rightarrow \{0, 1, \dots, \infty\}$ functions. This can be accomplished by starting with an H^* having support on nondecreasing $\mathcal{X} \rightarrow \mathbb{R}^+$ functions. One such H^* corresponds to an exponentiated Gaussian process. Then, to obtain realizations $\tilde{\mu}_h \sim H$, draw $\mu_h^* \sim H^*$ and let $\tilde{\mu}_h(x) = \lfloor \mu_h^*(x) \rfloor$. The resulting P_{x_i} will have nonnegative integer atoms.

Let $P = (P_x, x \in \mathcal{X})$ denote an uncountable collection of stochastically ordered probability measures. The nondecreasing stochastic ordering constraint means $P_x \leq P_{x'}$, for any two points $x \leq x'$ in $\mathcal{X}$. Let $\mathcal{P}$ denote the set of probability measures on $\{\mathcal{Y}, \mathcal{B}(\mathcal{Y})\}$, with $\mathcal{Y}$ a subset of the real line and $\mathcal{B}(\mathcal{Y})$ the Borel subsets of $\mathcal{Y}$. Let $P_{\mathcal{X}} = \{P = (P_x, x \in \mathcal{X}) : P_x \in \mathcal{P} \text{ for all } x \in \mathcal{X}\}$, and $C_{\mathcal{X}} = \{P = (P_x, x \in \mathcal{X}) : P_x \in \mathcal{P}, P_x \leq P_{x'} \text{ for any } x \leq x', x, x' \in \mathcal{X}\}$, with $P_{\mathcal{X}}$ the set of uncountable collections of probability measures on $\{\mathcal{Y}, \mathcal{B}(\mathcal{Y})\}$ indexed by $x \in \mathcal{X}$, and $C_{\mathcal{X}}$ the subset of collections satisfying nondecreasing stochastic order.

In addition, let $R_{\mathcal{X}}$ denote the $\mathcal{X} \rightarrow \mathcal{Y}$ function space and $S_{\mathcal{X}}$ the nondecreasing $\mathcal{X} \rightarrow \mathcal{Y}$ function space. Define a metric d in $R_{\mathcal{X}}$ with $d(s_1, s_2) = \sup_{x \in \mathcal{X}} |s_1(x) - s_2(x)|$ for any $s_1, s_2 \in R_{\mathcal{X}}$. This metric induces a topology on $R_{\mathcal{X}}$ and thus on $S_{\mathcal{X}}$. Let $\mathcal{Q}$ denote the set of probability measures on $S_{\mathcal{X}}$.

THEOREM 2. *Consider the prior π_P for P defined in (2), with $\Theta_h \sim H$ and $H \in \mathcal{Q}$ a probability measure that has full support on $S_{\mathcal{X}}$. Then π_P has full support on $C_{\mathcal{X}}$ with respect to the topology of weak convergence defined in the Appendix.*

From Theorem 2, the constructed prior π_P in (2) can assign a positive probability to any open set containing the true P of interest. Theorem 2 follows directly from a more general result in Theorem 3 to be presented below. In the Appendix, Lemma A1 establishes a generalized Choquet's theorem by showing that every $P \in C_{\mathcal{X}}$ can be represented as a mixture over the extreme points of $C_{\mathcal{X}}$ as follows

$$P(B) = T(Q)(B) = \int_{S_{\mathcal{X}}} \{\delta_{s(x)}(B), x \in \mathcal{X}\} dQ(s) \quad (6)$$

for any $B \in \mathcal{B}(\mathcal{Y})$, where Q is a mixing measure over $S_{\mathcal{X}}$, T is an integral operator mapping Q to its marginals and $\{\delta_{s(x)}(\cdot), x \in \mathcal{X}\}$ is an extreme point of $C_{\mathcal{X}}$. The mapping operator T in (6) is shown to be continuous in Lemma A2.

The integral representation (6) allows us to induce a prior π_P for P by introducing a prior π_Q for the mixing measure Q . Based on Lemmas 1 and 2, we establish the following theorem.

THEOREM 3. *Through (6), the prior π_P for P has full support on $C_{\mathcal{X}}$ with respect to the topology of weak convergence if and only if the prior π_Q for Q has full support on $\mathcal{Q}$.*

The proofs of Lemmas A1 and A2 and Theorem 3 are sketched in the Appendix. Theorem 2 follows directly from Theorem 3 by taking π_Q to be $\text{DP}(\alpha H)$, a Dirichlet process with a precision parameter α and a base probability measure H having full support on $S_{\mathcal{X}}$, after noting that $\text{DP}(\alpha H)$ has a weak support on all of $\mathcal{Q}$ (Hoff, 2003a) with respect to the topology induced by the metric d .

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains a sensitivity analysis, the influence of the recommended prior specification for the hyperparameters, the Markov chain Monte Carlo mixing performance in the simulation and DDE data analysis, and the estimated covariate effects in the DDE data analysis.

ACKNOWLEDGEMENT

This research was partially supported by a grant from the National Institute of Environmental Health Sciences of the National Institutes of Health. The authors wish to thank the editor, an associate editor, and two reviewers for their critical and constructive comments that greatly improved the original presentation.

APPENDIX

Proof of Theorem 1. Under (3), the conditional cumulative distribution functions are

$$F_x(y) = \sum_{h=1}^{\infty} p_h \Phi[\tilde{\tau}_h^{1/2}\{y - \tilde{\mu}_h(x)\}]$$

for any $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Then for any two predictor values x and x' with $x \leq x'$,

$$F_x(y) - F_{x'}(y) = \sum_{h=1}^{\infty} p_h (\Phi[\tilde{\tau}_h^{1/2}\{y - \tilde{\mu}_h(x)\}] - \Phi[\tilde{\tau}_h^{1/2}\{y - \tilde{\mu}_h(x')\}]).$$

First, we show that (b) implies (a). It is straightforward to see $F_x(y) - F_{x'}(y) \geq 0$ for any $y \in \mathcal{Y}$ and $x < x' \in \mathcal{X}$ with probability 1 based on the facts that Φ is an increasing function and that $\tilde{\mu}_h$ s are nondecreasing random functions sampled from H . This suggests that $f_{\mathcal{X}}$ is subject to the nondecreasing stochastic ordering constraint. Now we show $f_{\mathcal{X}}$ under (3) are subject to the continuity condition. First, for any $y_1, y_2 \in \mathfrak{R}$ and $b > 0$, one has $\phi(b^{1/2}y_1) - \phi(b^{1/2}y_2) = (2\pi)^{-1/2}by_0 \exp(-by_0^2/2)(y_1 - y_2)$ for some y_0 between y_1 and y_2 using the mean value theorem. One further obtains $\phi(b^{1/2}y_1) - \phi(b^{1/2}y_2) \leq (2\pi)^{-1/2}b^{1/2}|y_1 - y_2|$ since $b^{1/2}y_0 \exp(-by_0^2/2) \leq 1$ for any y_0 . Fixing $x' \in \mathcal{X}$, for any $y \in \mathcal{Y}$ and any $x \in \mathcal{X}$, one then has

$$\begin{aligned} f_x(y) - f_{x'}(y) &= \sum_{h=1}^{\infty} p_h \tilde{\tau}_h^{1/2} (\phi[\tilde{\tau}_h^{1/2}\{y - \tilde{\mu}_h(x)\}] - \phi[\tilde{\tau}_h^{1/2}\{y - \tilde{\mu}_h(x')\}]) \\ &\leq (2\pi)^{-1/2} \sum_{h=1}^{\infty} p_h \tilde{\tau}_h |\tilde{\mu}_h(x) - \tilde{\mu}_h(x')|. \end{aligned}$$

Then $f_x(y) - f_{x'}(y) \rightarrow 0$ for any $y \in \mathcal{Y}$ as $x \rightarrow x'$ follows directly by noting that $\sum_{h=1}^{\infty} p_h \tilde{\tau}_h |\tilde{\mu}_h(x) - \tilde{\mu}_h(x')| \rightarrow 0$ as $x \rightarrow x'$ since $\tilde{\mu}_h$ s are continuous random functions sampled from H .

Now we show (a) implying (b) by showing if (b) does not hold, then (a) does not hold. Suppose $H(S_{\mathcal{X},c}) < 1$. Denote $B_{x,x'} = \{s \in R_{\mathcal{X}} : s(x) \leq s(x')\}$ for a pair of rational numbers $x, x' \in \mathcal{X}$ and $x < x'$, where $R_{\mathcal{X}}$ is the space of $\mathfrak{R}$ functions. Being the space of continuous nondecreasing functions on $\mathcal{X}$, $S_{\mathcal{X},c}$ can be written as a countable intersection of $B_{x,x'}$ over all possible rational pairs (x, x') in $\mathcal{X}$. Thus, $R_{\mathcal{X}} \setminus S_{\mathcal{X},c}$ can be written as a countable union of $B_{x,x'}^c$ over all possible rational pairs (x, x') in $\mathcal{X}$, where $B_{x,x'}^c = R_{\mathcal{X}} \setminus B_{x,x'}$. Since $H(R_{\mathcal{X}} \setminus S_{\mathcal{X},c}) = 1 - H(S_{\mathcal{X},c}) > 0$, there exist a pair of rational numbers (x_1, x_2) with $x_1 < x_2$ such that $H(B_{x_1,x_2}^c) > 0$ by using the subadditivity of a probability measure. Taking this pair (x_1, x_2) and fixing a $y \in \mathcal{Y}$, we consider

$$C = \{(s, t) \in B_{x_1,x_2}^c \times \mathfrak{R}^+ : \Phi[t^{1/2}\{y - s(x_1)\}] - \Phi[t^{1/2}\{y - s(x_2)\}] < 0\}.$$

It is straightforward to show $C = B_{x_1,x_2}^c \times \mathfrak{R}^+$ and

$$\text{pr}\{(\Theta_1, \tilde{\tau}_1) \in C\} = M(C) = H \otimes L(B_{x_1,x_2}^c \times \mathfrak{R}^+) = H(B_{x_1,x_2}^c)L(\mathfrak{R}^+) = H(B_{x_1,x_2}^c) > 0$$

due to the independence between H and L and $L(\mathfrak{R}^+) = 1$. Write

$$C = \bigcup_{k=1}^{\infty} C_k = \bigcup_{k=1}^{\infty} \{(s, t) \in B_{x_1,x_2}^c \times \mathfrak{R}^+ : \Phi[t^{1/2}\{y - s(x_1)\}] - \Phi[t^{1/2}\{y - s(x_2)\}] < -1/k\}.$$

Then there exists a positive integer k_0 such that $\text{pr}(C_{k_0}) > 0$ since $0 < \text{pr}(C) \leq \sum_{k=0}^{\infty} \text{pr}(C_k)$.

Choose $\epsilon \in \{0, 1/(1 + 2k_0)\}$ and a positive integer T_0 such that $T_0 > \log(\epsilon)/\log\{\alpha/(1 + \alpha)\}$. Define

$$A = \left\{ q = (q_h, h = 1, \dots, \infty) : \sum_{h=1}^{\infty} q_h = 1, \sum_{h=T_0+1}^{\infty} q_h \leq \epsilon, \text{ and } 0 < q_h < 1 \text{ for any } h \right\}.$$

Denote $g_h = \Phi[\tilde{\tau}_h^{1/2}\{y - \tilde{\mu}_h(x_1)\}] - \Phi[\tilde{\tau}_h^{1/2}\{y - \tilde{\mu}_h(x_2)\}]$ for each h . Clearly, g_h s are independent and identically distributed, and $\text{pr}(|g_h| \leq 2) = 1$ for each h since Φ is a cumulative distribution function. Then on $D = \{p = (p_h, h = 1, \dots, \infty) \in A, \text{ and } (\tilde{\mu}_h, \tilde{\tau}_h) \in C_{k_0} \text{ for } h = 1, \dots, T_0\}$,

$$\sum_{h=1}^{\infty} p_h g_h = \sum_{h=1}^{T_0} p_h g_h + \sum_{h=T_0+1}^{\infty} p_h g_h < -\frac{1}{k_0} \sum_{h=1}^{T_0} p_h + 2\epsilon < -\frac{1}{k_0}(1 - \epsilon) + 2\epsilon < 0. \quad (\text{A1})$$

By using Markov's inequality, one has

$$\text{pr}(p \in A) = \text{pr}\left(\sum_{h=T_0+1}^{\infty} p_h \leq \epsilon\right) \geq 1 - \frac{1}{\epsilon} E\left(\sum_{h=T_0+1}^{\infty} p_h\right) = 1 - \frac{1}{\epsilon} \left(\frac{\alpha}{1 + \alpha}\right)^{T_0} > 0.$$

Hence, one obtains $\text{pr}(D) = \text{pr}(A)\text{pr}(C_{k_0})^{T_0} > 0$ by using the mutual independence of $(\tilde{\mu}_h, \tilde{\tau}_h)$ s and p . This result together with (A1) indicates that F_{x_2} is not stochastically larger than F_{x_1} , and thus (a) cannot hold. $\square$

DEFINITION 1 (Weak convergence in $P_{\mathcal{X}}$ used in § 6). Define $G_Z(P)(\cdot) = \int_{\mathcal{X}} P_x(\cdot) dZ(x)$ for any $P = (P_x, x \in \mathcal{X}) \in P_{\mathcal{X}}$ and any probability measure Z on $\mathcal{B}(\mathcal{X})$, the Borel sets of $\mathcal{X}$. Then $G_Z(P) \in \mathcal{P}$ is a probability measure on $\mathcal{B}(\mathcal{Y})$. Define weak convergence in $P_{\mathcal{X}}$ as follows: a sequence $\{P^{(n)}\}_{n=1}^{\infty}$, with $P^{(n)} \in P_{\mathcal{X}}$ for all n , converges weakly to $P \in P_{\mathcal{X}}$, if and only if $G_Z(P^{(n)})$ converges weakly to $G_Z(P)$ for any probability measure Z on $\mathcal{B}(\mathcal{X})$.

The above weak convergence of $P^{(n)}$ to P in $P_{\mathcal{X}}$ implies pointwise weak convergence of $P_x^{(n)}$ to P_x for all $x \in \mathcal{X}$, which can be easily seen by taking $Z = \delta_x$. Essentially the weak convergence of P_n to P in $P_{\mathcal{X}}$ is equivalent to the uniform weak convergence of $P_x^{(n)}$ to P_x for all $x \in \mathcal{X}$. Another equivalent definition is that $d(\int f dP, \int f dP^{(n)}) \rightarrow 0$ for any bounded continuous f on $\mathcal{Y}$ with d being the uniform metric defined in § 6 and $\int f dP = (\int f dP_x, x \in \mathcal{X}) \in R_{\mathcal{X}}$.

The above weak convergence induces a weak topology in $P_{\mathcal{X}}$. Under this topology, it can be shown that $C_{\mathcal{X}}$ is a weakly closed convex set and that the set of extreme points of $C_{\mathcal{X}}$ is $\text{ex}C_{\mathcal{X}} = \{(\delta_{s(x)}, x \in \mathcal{X}) : s \in S_{\mathcal{X}}\}$. Lemma A1 generalizes Choquet's Theorem by considering an uncountable collection of probability measures.

LEMMA A1. For any probability measure $Q \in \mathcal{Q}$, $T(Q) \in C_{\mathcal{X}}$, where T is defined in (6). Also, for any $P \in C_{\mathcal{X}}$, there exists a probability measure $Q \in \mathcal{Q}$ such that $Q(\{s \in S_{\mathcal{X}} : s(x) \in \cdot\}) = P_x(\cdot)$ for all $x \in \mathcal{X}$, where s denotes a random function over $\mathcal{X}$.

Proof of Lemma A1. For a probability measure Q on $S_{\mathcal{X}}$, define $Q_x(\cdot) = \int_{S_{\mathcal{X}}} \delta_{s(x)}(\cdot) dQ(s)$. Then Q_x is a probability measure on $\mathcal{B}(\mathcal{Y})$. To show $T(Q) \in C_{\mathcal{X}}$, one only needs to show $(Q_x, x \in \mathcal{X})$ are subject to nondecreasing stochastic ordering constraint. For any two points $x, x' \in \mathcal{X}$ with $x \leq x'$, one has $s(x) \leq s(x')$ for $s \in S_{\mathcal{X}}$. Further one has $\{s \in S_{\mathcal{X}} : s(x) > a\} \subseteq \{s \in S_{\mathcal{X}} : s(x') > a\}$ for any $a \in \mathfrak{N}$. By the definition of Q_x and $Q_{x'}$, one obtains that $Q_x(a, \infty) = \int_{S_{\mathcal{X}}} \delta_{s(x)}(a, \infty) dQ(s) = Q[\{s \in S_{\mathcal{X}} : s(x) > a\}] \leq Q[\{s \in S_{\mathcal{X}} : s(x') > a\}] = \int_{S_{\mathcal{X}}} \delta_{s(x')}(a, \infty) dQ(s) = Q_{x'}(a, \infty)$. This suggests $Q_x \leq Q_{x'}$ and thus $T(Q) = (Q_x, x \in \mathcal{X}) \in C_{\mathcal{X}}$.

Suppose $P = (P_x, x \in \mathcal{X}) \in C_{\mathcal{X}}$. Let F_x denote the cumulative distribution function with respect to P_x and define $s_x(\omega) = F_x^{-1}(\omega) = \inf\{u : F_x(u) \geq \omega\}$ for any $x \in \mathcal{X}$ and $\omega \in [0, 1]$. For any $x, x' \in \mathcal{X}$ with $x \leq x'$, $P_x \leq P_{x'}$ leads to $\{u : F_{x'}(u) \geq \omega\} \subseteq \{u : F_x(u) \geq \omega\}$ and thus $s_x(\omega) \leq s_{x'}(\omega)$ for all ω . Let Q be the canonical measure of $s = (s(x), x \in \mathcal{X})$ on $(S_{\mathcal{X}}, \sigma_{S_{\mathcal{X}}})$ induced by a uniform distribution on ω , where

$\sigma_{S_{\mathcal{X}}}$ is a σ -algebra on $S_{\mathcal{X}}$. Then $\mathcal{Q}$ represents P with $\mathcal{Q}_x = P_x$ for any $x \in \mathcal{X}$ using similar arguments to the proof of Proposition 8 in Hoff (2003b). $\square$

LEMMA A2. *The integral operator T defined in (6) is continuous.*

Proof of Lemma A2. The continuity of T means that if $P^{(n)} = T(\mathcal{Q}^{(n)})$, $P = T(\mathcal{Q})$, and $\mathcal{Q}^{(n)}$ converges weakly to $\mathcal{Q}$, then $P^{(n)}$ converges weakly to P . By the definition of weak convergence in $P_{\mathcal{X}}$, we need to show $G_Z(P^{(n)})$ converges to $G_Z(P)$ for any probability measure Z on $\mathcal{B}(\mathcal{X})$. This is equivalent to showing that $\int_{\mathcal{Y}} f dG_Z(P^{(n)}) \rightarrow \int_{\mathcal{Y}} f dG_Z(P)$ for any bounded uniformly continuous function f on $\mathcal{Y}$ and any probability measure Z on $\mathcal{B}(\mathcal{X})$ by noting that $\mathcal{Y}$ is a separable metric space (Parthasarathy, 1967; Billingsley, 1999). Denote by $U(\mathcal{Y})$ the set of bounded uniformly continuous functions on $\mathcal{Y}$.

First, for any $f \in U(\mathcal{Y})$ and any probability measure Z on $\mathcal{B}(\mathcal{X})$, we have the following facts: $G_Z(P)(\cdot) = \int_{\mathcal{X}} \mathcal{Q}_x(\cdot) dZ(x) = \int_{\mathcal{X}} \int_{S_{\mathcal{X}}} \delta_{s(x)}(\cdot) d\mathcal{Q}(s) dZ(x)$ and

$$\begin{aligned} \int_{\mathcal{Y}} f dG_Z(P) &= \int_{\mathcal{Y}} f(y) \int_{\mathcal{X}} \int_{S_{\mathcal{X}}} d\mathcal{Q}(s) dZ(x) \delta_{s(x)}(y) dy \\ &= \int_{S_{\mathcal{X}}} \int_{\mathcal{X}} \int_{\mathcal{Y}} f(y) \delta_{s(x)}(y) dy dZ(x) d\mathcal{Q}(s) = \int_{S_{\mathcal{X}}} \int_{\mathcal{X}} f(s(x)) dZ(x) d\mathcal{Q}(s). \end{aligned}$$

Let $g_{(Z,f)}(s) = \int_{\mathcal{X}} f(s(x)) dZ(x)$, a function on $S_{\mathcal{X}}$, given $Z \in \mathcal{B}(\mathcal{X})$ and $f \in U(\mathcal{Y})$. Then one has $\int_{\mathcal{Y}} f dG_Z(P) = \int_{S_{\mathcal{X}}} g_{(Z,f)}(s) d\mathcal{Q}(s)$ and $\int_{\mathcal{Y}} f dG_Z(P^{(n)}) = \int_{S_{\mathcal{X}}} g_{(Z,f)}(s) d\mathcal{Q}^{(n)}(s)$. Since $\mathcal{Q}^{(n)}$ converges weakly to $\mathcal{Q}$, one only needs to show that $g_{(Z,f)}$ is a bounded continuous function of s on $S_{\mathcal{X}}$ for any $f \in U(\mathcal{Y})$ and any measure Z on $\mathcal{B}(\mathcal{X})$. This is trivial based on the uniform metric d in $R_{\mathcal{X}}$. $\square$

Proof of Theorem 3. Suppose that $\pi_{\mathcal{Q}}$ has a full support on $\mathcal{Q}$. Let A be an open subset of $C_{\mathcal{X}}$, then one has $\pi_P(A) = \text{pr}\{T(\mathcal{Q}) \in A\} = \pi_{\mathcal{Q}}(T^{-1}A) > 0$ based on the facts that $T^{-1}A$ is an open set due to the continuity of T and that $\pi_{\mathcal{Q}}$ has the full support on $\mathcal{Q}$. It is straightforward to have $\pi_P(C_{\mathcal{X}}) = \pi_{\mathcal{Q}}(\mathcal{Q}) = 1$ based on Lemma A1. The other part can be proven similarly. $\square$

REFERENCES

- ARIAS, E. & GASBARRA, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *J. Am. Statist. Assoc.* **91**, 1101–9.
- BILLINGSLEY, P. (1999). *Convergence of Probability Measures*. New York: John Wiley & Sons, Inc.
- BORNKAMP, B. & ICKSTADT, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose response analysis. *Biometrics* **65**, 198–205.
- DUNSON, D. B. & PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–23.
- DUNSON, D. B. & PEDDADA, S. D. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika* **95**, 859–74.
- ESCOBAR, M. D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577–88.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.
- FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–29.
- GELFAND, A. E. & KOTTAS, A. (2001). Nonparametric Bayesian modeling for stochastic order. *Inst. Statist. Math.* **53**, 865–76.
- GELMAN, A. (2004). Parameterization and Bayesian modeling. *J. Am. Statist. Assoc.* **99**, 537–45.
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* **3**, 515–33.
- GELMAN, A., JAKULIN, A., PITTAU, M. G. & SU, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Statist.* **2**, 1360–83.
- GHOSH, J. & DUNSON, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *J. Comp. Graph. Statist.* **18**, 306–20.
- GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.
- HOFF, P. D. (2003a). Bayesian methods for partial stochastic orderings. *Biometrika* **90**, 303–17.
- HOFF, P. D. (2003b). Nonparametric estimation of convex models via mixtures. *Ann. Statist.* **31**, 174–200.
- HOLMES, C. C. & HELD, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **1**, 145–68.

- ISHWARAN, H. & JAMES, L. F. (2001). Gibbs sampling methods for stick breaking priors. *J. Am. Statist. Assoc.* **96**, 161–73.
- KARABATSOS, G. & WALKER, S. G. (2007). Bayesian nonparametric inference of stochastically ordered distributions, with Pólya trees and Bernstein polynomials. *Statist. Prob. Lett.* **77**, 907–13.
- KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.
- LAVINE, M. & MOCKUS, A. (1995). A nonparametric Bayes method for isotonic regression. *J. Statist. Plan. Infer.* **46**, 235–48.
- LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351–7.
- LONGNECKER, M. P., KLEBANOFF, M. A., ZHOU, H. B. & BROCK, J. W. (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *Lancet* **358**, 110–4.
- MACEachern, S. N. (1999). Dependent nonparametric processes. In *Proc. Bayesian Statist. Sci. Sect.*, pp. 50–5. Alexandria, VA: American Statistical Association.
- NEELON, B. & DUNSON, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60**, 398–406.
- NORETS, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *Ann. Statist.* **38**, 1733–66.
- ONGARO, A. & CATTANEO, C. (2004). Discrete random probability measures: A general framework for nonparametric Bayesian inference. *Statist. Prob. Lett.* **67**, 33–5.
- PAPASPILIOPOULOS, O. & ROBERTS, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–86.
- PARTHASARTHY, K. (1967). *Probability Measures on Metric Spaces*. New York: Academic Press Inc.
- PETRONE, S. (1999). Bayesian density estimation using Bernstein polynomials. *Can. J. Statist.* **27**, 105–26.
- RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3**, 425–41.
- RAMSAY, J. O. (1998). Estimating smooth monotone functions. *J. R. Statist. Soc. B* **60**, 365–75.
- SHIVELY, T. S., SAGER, T. W. & WALKER, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Statist. Soc. B* **71**, 159–75.
- SMITH, M. & KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Economet.* **75**, 317–43.
- TOKDAR, S., ZHU, Y. & GHOSH, J. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5**, 1–26.
- WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Commun. Statist. B* **36**, 45–54.
- WANG, X. (2008). Bayesian free-knot monotone cubic spline regression. *J. Comp. Graph. Statist.* **17**, 373–87.
- YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G. O. & HOLMES, C. (2011). Bayesian nonparametric hidden Markov models with applications in genomics. *J. R. Statist. Soc. B* **73**, 37–57.

[Received October 2008. Revised March 2011]