*Biostatistics* (2010), **11**, 3, *pp.* 419–431
doi:10.1093/biostatistics/kxq001
Advance Access publication on February 11, 2010

# Stochastically ordered multiple regression

BJÖRN BORNKAMP*, KATJA ICKSTADT

*Fakultät Statistik, Technische Universität Dortmund 44221 Dortmund, Germany*
bornkamp@statistik.uni-dortmund.de

DAVID DUNSON

*Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708-0251, USA*



## SUMMARY

In various application areas, prior information is available about the direction of the effects of multiple predictors on the conditional response distribution. For example, in epidemiology studies of potentially adverse exposures and continuous health responses, one can typically assume *a priori* that increasing the level of an exposure does not lead to an improvement in the health response. Such an assumption can be formalized through a stochastic ordering assumption in each of the exposures, leading to a potentially large improvement in efficiency in nonparametric modeling of the conditional response distribution. This article proposes a Bayesian nonparametric approach to this problem based on characterizing the conditional response density as a Gaussian mixture, with the locations of the Gaussian means varying flexibly with predictors subject to minimal constraints to ensure stochastic ordering. Theoretical properties are considered and Markov chain Monte Carlo methods are developed for posterior computation. The methods are illustrated using simulation examples and a reproductive epidemiology application.

*Keywords*: Conditional distribution estimation; Density regression; Isotonic regression; Nonparametric Bayes; Risk assessment; Stochastic order.

## 1. INTRODUCTION

In many biomedical applications, subject-specific knowledge suggests that the conditional distribution of a response variable $y \in \mathbb{R}$ given predictors $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^k$ increases (or decreases) stochastically with increasing $\boldsymbol{x}$. One example arises in epidemiology, where the exposure to toxic substances or environmental risk factors often can be assumed to be related to health risk in a monotonic way. A different example appears in clinical trials, where the effect of a pharmaceutical compound (or a combination of compounds or therapies) is assumed to be increasing with increasing dose level (or intensity of therapy). In these situations, it is natural to model the distribution of the response conditionally on covariates, such as age, as stochastically ordered with increasing value of the exposures. For ease of exposition, we focus on the increasing case, but a stochastic decrease can be considered analogously.

*To whom correspondence should be addressed.

Nonparametric modeling of stochastically increasing densities with respect to an ordered "categorical" covariate has recently been discussed quite extensively in a Bayesian framework by Gelfand and Kottas (2000), Hoff (2003), Karabatsos and Walker (2007), and Dunson and Peddada (2008), among others. The generalization to a multivariate continuous predictor is considerably more difficult. When normality and homoscedasticity are imposed on the residual density, the problem reduces to estimation of an isotonic regression in multiple predictors (e.g. Dykstra and Robertson, 1982). Mukarjee and Stern (1994) and Dette and Scheder (2006) proposed to monotonize an unconstrained nonparametric regression fit. To reduce dimensionality in modeling of the multivariate surface subject to monotonicity constraints, additivity constraints can be imposed as in Bacchetti (1989), Morton-Jones *and others* (2000), and Leitenstorfer and Tutz (2007) or more recently in Shively *and others* (2009) in a Bayesian framework (see also Cai and Dunson, 2007, for a Bayesian approach to monotonic regression with respect to a multivariate outcome).

Such methods focus on the mean of the response distribution, while in many applications, the distribution tails may be of greater interest. For example, in epidemiology, subjects in the right or left tail have an adverse health response. In order to assess how the entire conditional response distribution changes with predictors, it is important to avoid restrictive assumptions such as normality and homoscedasticity. Bayesian density regression methods, proposed by, Müller *and others* (1996) and Dunson *and others* (2007), among others, allow the conditional response density to change flexibly with predictors. To address the curse of dimensionality problem, such methods borrow strongly across different regions of the predictor space, relying on base parametric models and smoothing priors. Efficiency can be substantially improved through imposing stochastic ordering constraints. To our knowledge, Wang and Dunson (2009) is the only method to enforce stochastic ordering over a continuous predictor in nonparametric density regression. Our focus is on generalizing their approach to allow multiple predictors while incorporating ideas of Bornkamp and Ickstadt (2009).

Section 2 describes our model and discusses properties. Section 3 applies the methods to an epidemiology data set, and Section 4 concludes. A simulation study can be found in Section A of the supplementary material available at *Biostatistics* online.

## 2. METHODOLOGY

### 2.1 *Mixture priors*

Although there is a rich literature on multivariate stochastic ordering, the focus has been on multivariate responses. To our knowledge, we are the first to address the problem of nonparametric conditional distribution modeling subject to stochastic ordering in multiple predictors. We refer to the proposed order restriction as SO-$\mathcal{X}$, with $\mathcal{X}$ the (possibly multivariate) input space of the predictors. In particular, letting $F_{\boldsymbol{x}}(y)$ denote the conditional distribution function of $y$ given predictors $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^k$, restriction SO-$\mathcal{X}$ corresponds to

$$F_{\boldsymbol{x}}(y) \geqslant F_{\boldsymbol{x}'}(y), \quad \text{for all } y \in \mathbb{R} \text{ and } \boldsymbol{x} \leqslant \boldsymbol{x}',$$

where $\boldsymbol{x} \leqslant \boldsymbol{x}'$ if and only if $x_m \leqslant x'_m$ for all $m = 1, \ldots, k$.

Let $\mathcal{F}_{\mathcal{X}} = \{F_{\boldsymbol{x}}, \boldsymbol{x} \in \mathcal{X}\}$ denote an uncountable collection of continuous conditional distribution functions, with each $F_{\boldsymbol{x}}$ in $\mathcal{F}_{\mathcal{X}}$ having support on $\mathbb{R}$ and with $\mathcal{X} \subset \mathbb{R}^k$. We propose a prior $\mathcal{F}_{\mathcal{X}} \sim \mathcal{P}$, where $\mathcal{P}$ corresponds to a distribution over the set of all possible collections $\mathcal{F}_{\mathcal{X}}$ subject to restriction SO-$\mathcal{X}$. To induce such a prior, we propose to characterize each $F_{\boldsymbol{x}}$ as a location-scale mixture of Gaussians, with the variances constant with $\boldsymbol{x}$, while the conditional means vary according to unknown multivariate monotone functions. Such a restriction on the component-specific mean functions is sufficient to ensure SO-$\mathcal{X}$, as is shown formally below.

Letting $f_{\boldsymbol{x}}$ denote the density corresponding to distribution function $F_{\boldsymbol{x}}$, we assume

$$f_{\boldsymbol{x}}(y) = \int \phi(y, \mu(\boldsymbol{x}), \sigma^2) P(\mathrm{d}\mu, \mathrm{d}\sigma^2) = \sum_h \pi_h \phi(x, \mu_h(\boldsymbol{x}), \sigma_h^2), \tag{2.1}$$

where $\phi(\mu, \sigma^2)$ is the normal density with mean $\mu$ and variance $\sigma^2$, the $\mu_h$: $\mathcal{X} \to \mathbb{R}$ are multivariate monotonic functions satisfying $\mu_h(\boldsymbol{x}) \leqslant \mu_h(\boldsymbol{x}')$ for all $\boldsymbol{x} \leqslant \boldsymbol{x}'$, and $P$ is a discrete mixing probability measure with support on $\mathcal{M} \times [0, \infty)$, where $\mathcal{M}$ is the space of multivariate monotonic functions mapping from $\mathcal{X} \to \mathbb{R}$. In the following, we take $\mathcal{X} = [0, 1]^k$ without loss of generality for bounded predictors. By assuming that the mixing measure is almost-surely discrete, we hence obtain a countable mixture with $\pi_h$ a probability weight on the $h$th component, which has associated mean function $\mu_h$ and variance $\sigma_h^2$. For each $\boldsymbol{x} \in \mathcal{X}$, the conditional density is expressed as a univariate Gaussian mixture, with the densities stochastically ordered due to the monotonicity of each $\mu_h$. We focus on Gaussian mixtures as they are well established and computationally tractable, note however, that most of the theory in this paper also applies to other kernels.

We are not aware of methods for estimating a model of form (2.1) in a non-Bayesian framework, although the mathematical form of the resulting estimator is related to the traditional kernel density estimator with a Gaussian kernel. We believe, however, that a naive classical approach (e.g. optimizing the likelihood function) could run into severe problems due to local optima in the likelihood surface and a possible overfitting problem. Instead, we follow a Bayesian approach by using a prior distribution for the mixing measure $P(\mathrm{d}\mu, \mathrm{d}\sigma^2)$. This has the advantage of an intrinsic regularization through prior distributions and implicit averaging over possible local optima of the likelihood.

As a general prior for the discrete mixing measure $P(\mathrm{d}\mu, \mathrm{d}\sigma^2)$, we focus on the class proposed by Ongaro and Cattaneo (2004), which includes a broad variety of priors as special cases (and is itself a special case of the general class of species sampling random probability measures due to Pitman, 1996). A random probability measure belongs to this class when its realizations can be represented as

$$P(\cdot) = \sum_{h=1}^{N} \pi_h \delta_{\boldsymbol{\xi}_h}(\cdot),$$

where $\boldsymbol{\xi}_h, \pi_h$ and $N$ are random variables specified as follows: The $\boldsymbol{\xi}_h$ are independent and identically distributed realizations of a nonatomic distribution $P_0$ on $\boldsymbol{\Xi}$ and are independent from $\pi_h, h = 1, \ldots, N$ and $N$. Note that $\boldsymbol{\Xi}$ can be a finite-dimensional space but also, for example, a function space. The weights $\pi_1, \ldots, \pi_N$ have a distribution $Q_N$ on the $N-1$ dimensional simplex $\mathbb{S}^N = \{\boldsymbol{\pi} \in \mathbb{R}^N : \sum_{h=1}^N \pi_h = 1, \boldsymbol{\pi} \geqslant \mathbf{0}\}$ and $N$ is a positive integer valued random variable with the value $\infty$ also being allowed. The Dirichlet process with parameter $MP_0$ is obtained by setting $N = \infty$ and using the so-called Griffiths–Engen–McCloskey distribution with parameter $M$ for the weights $\pi_h$ (see Ishwaran and Zarepour, 2002, for details). A truncated Dirichlet process has a fixed $N$ and a generalized Dirichlet distribution for the weights.

When a Dirichlet process is used as a prior for the mixing measure in (2.1), our model is a special case of the very popular dependent Dirichlet process model, see, for example, MacEachern (1999), De Iorio *and others* (2004), or Gelfand *and others* (2005), with the main innovation that multivariate monotone function $\mu_h(.)$ are used as atoms in the mixture. The following lemma establishes that the use of multivariate monotone functions in mixture model (2.1) induces the SO-$\mathcal{X}$ restriction on the conditional distributions. Additionally, we establish that any collection of continuous conditional distributions in SO-$\mathcal{X}$ can be approximated using (2.1).

LEMMA 2.1 (Support)

(i) Under model (2.1), the conditional distributions satisfy

$$F_{\boldsymbol{x}}(y) \geqslant F_{\boldsymbol{x}'}(y), \quad \text{for all } y \in \mathbb{R}, (\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X} \times \mathcal{X}, \boldsymbol{x} \leqslant \boldsymbol{x}';$$

(ii) Given a set $\tilde{\mathcal{F}}_{\mathcal{X}}$ of continuous distributions satisfying SO-$\mathcal{X}$ order, with conditional distribution functions $\tilde{F}_{\boldsymbol{x}}(y)$ on $\mathbb{R}$, there exist, for an arbitrarily small $\epsilon > 0$, $\pi_h$, $\mu_h(\boldsymbol{x})$, and $\sigma_h^2$ such that

$$\sup_{\boldsymbol{x} \in [0,1]^k} \left\{ \sup_{y \in \mathbb{R}} \left| \sum_{h=1}^{N} \pi_h \Phi(y, \mu_h(\boldsymbol{x}), \sigma_h^2) - \tilde{F}_{\boldsymbol{x}}(y) \right| \right\} \leqslant \epsilon + \frac{1}{N},$$

where $\Phi(., \mu, \sigma^2)$ is the distribution function of a normal distribution with mean $\mu$ and variance $\sigma^2$.

*Proof.* See Section B of the supplementary material available at *Biostatistics* online. □

Because the probability of having any observation exactly at a given $\boldsymbol{x}$ is zero for predictors having a continuous density, the ability to estimate $f_{\boldsymbol{x}}(y)$ necessarily relies on borrowing of information across different locations. We cannot simply define separate mixtures of normals for each location. Lemma 2 shows how the dependence arises through the prior, while also providing an expression for the prior expectation.

LEMMA 2.2 (Prior moments) Marginalizing out the random mixing measure $P$, the expectation of $F_{\boldsymbol{x}}(y)$, and the covariance of $F_{\boldsymbol{x}}(y)$ and $F_{\boldsymbol{x}}(y)$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ are given by

$$E\{F_{\boldsymbol{x}}(y)\} = \int \Phi(y, \mu(\boldsymbol{x}), \sigma^2) \mathrm{d}P_0,$$

$$\mathrm{Cov}\{F_{\boldsymbol{x}}(y), F_{\boldsymbol{x}'}(y)\} = k_0 \left\{ \int \Phi(y, \mu(\boldsymbol{x}), \sigma^2) \Phi(y, \mu(\boldsymbol{x}'), \sigma^2) \mathrm{d}P_0 - \int \Phi(y, \mu(\boldsymbol{x}), \sigma^2) \mathrm{d}P_0 \int \Phi(y, \mu(\boldsymbol{x}'), \sigma^2) \mathrm{d}P_0 \right\},$$

where $\Phi(y, \mu(\boldsymbol{x}'), \sigma^2)$ is the distribution function of a normal distribution with mean $\mu$ and variance $\sigma^2$, $P_0$ is a nonatomic probability distribution on $\mathcal{M} \times [0, \infty)$ and $k_0 \in [0, 1]$ is given by $E\left(\sum_{h=1}^{N} \pi_h^2\right)$.

*Proof.* The proof is along the lines of the proof of Lemma 1 in Bornkamp and Ickstadt (2009). □

Hence, the prior mean and the prior correlation structure is determined by the base measure $P_0$ alone, while the parameter $k_0$ of the random measure, jointly with $P_0$, determines the variability. In practice, we need to specify the base measure $P_0$ of the nonparametric prior, consisting of a prior distribution $H$ on the monotonic function space $\mathcal{M}$ as well as a prior distribution on $[0, \infty)$ for the variance parameter. Because standard choices can be used for the prior for the variance (e.g. inverse gamma), we focus in Section 2.2 on how to choose $H$.

### 2.2 *Prior for multivariate monotone functions*

Placing a prior on the space of multivariate monotonic functions is challenging. The use of multivariate basis expansions or tensor products of univariate bases quickly becomes infeasible as the dimension increases because more and more basis functions are needed to obtain an adequate approximation (Barron,

1993). Another challenging issue is how to impose monotonicity on the multivariate basis. A common strategy is to impose additional constraints to simplify the problem, with 2 such possibilities corresponding to additive models (where $\mu(x_1, \ldots, x_k) = \mu_1(x_1) + \cdots + \mu_k(x_k)$) or single-index models (where $\mu(\boldsymbol{x}) = \mu^*(\boldsymbol{a}'\boldsymbol{x})$, with $\mu^*: \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{a} \in \mathbb{S}^k$ or $\boldsymbol{a} \in \mathbb{R}^k$), see Antoniadis *and others* (2004) for a Bayesian approach to single-index models. For additive models, monotonicity is imposed through restricting each univariate function to be monotonic, while for single-index models, one can let $\boldsymbol{a} \in \mathbb{R}^k_+$ and $\mu^*$ be monotonic. Unfortunately, additive models do not allow interactions, and the single-index model is constant on hyperplanes of the form $\boldsymbol{a}'\boldsymbol{x} = \text{Const}$.

We propose to base our prior on linear combinations of ridge functions, $\mu(\boldsymbol{x}) = \sum c_j g_j(\boldsymbol{a}_j'\boldsymbol{x})$, where the $g_j: \mathbb{R} \to \mathbb{R}$ are univariate continuous functions and the $\boldsymbol{a}_j \in \mathbb{R}^k$ are direction vectors. Ridge functions form the building blocks of neural networks and projection pursuit regression. Linear combinations of sufficiently flexible ridge functions can approximate any multivariate continuous function in sup norm (Cheney and Light, 1999, Chapter 22) and are ideally suited to multivariate cases in requiring only a few ridge functions to characterize fairly complex relationships (Barron, 1993). As a sufficient but not necessary condition to ensure monotonicity, we assume $c_j \in \mathbb{R}_+$, the $g_j(.)$ to be monotonic and $\boldsymbol{a}_j \in \mathbb{R}^k_+$. We refer to the resulting class of functions as positive linear combination of monotonic ridge (plcmr) functions. As it is not straightforward to find simple, and hence computationally tractable, necessary restrictions for monotonicity and we find the plcmr class to be highly flexible, we restrict consideration to this class. It is straightforward to show that all plcmr functions are multivariate monotone, with monotone additive and monotone single-index models arising as special cases.

We will carefully specify our prior on the space of plcmr functions on $[0, 1]^k$ to facilitate interpretation and computation expressing the function $\mu(\boldsymbol{x})$ as

$$\mu(\boldsymbol{x}) = \beta_0 + \beta_1 \mu^0(\boldsymbol{x}), \tag{2.2}$$

with $\beta_0 \in \mathbb{R}$ the value at $\boldsymbol{x} = (0, \ldots, 0)'$, $\beta_1 \in \mathbb{R}_+$ the maximum change between $(0, \ldots, 0)'$ and $(1, \ldots, 1)'$, $\mu^0(\boldsymbol{x}) = \int G(\boldsymbol{\alpha}'\boldsymbol{x}, \boldsymbol{\theta}) Q(\mathrm{d}\boldsymbol{\alpha}, \mathrm{d}\boldsymbol{\theta}) = \sum_{j=1}^J w_j G(\boldsymbol{\alpha}_j'\boldsymbol{x}, \boldsymbol{\theta}_j)$, $Q$ a discrete probability measure, $\boldsymbol{w} \in \mathbb{S}^J$, $\boldsymbol{\alpha}_j \in \mathbb{S}^k$ and $G$ a univariate probability distribution function on $[0, 1]$ depending on parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Restricting $\boldsymbol{\alpha}_j$ to fall on the simplex has the advantage that automatically $\boldsymbol{\alpha}_j'\boldsymbol{x} \in [0, 1]$ for any $\boldsymbol{\alpha} \in \mathbb{S}^k$ and any $\boldsymbol{x} \in [0, 1]^k$. Hence, $\boldsymbol{\alpha}$ measures the proportions of the total increase in the function $\boldsymbol{\mu}^0(.)$ attributable to the different covariates.

We do not directly use a multivariate probability distribution function $G(\boldsymbol{x}, \boldsymbol{\theta})$ on $[0, 1]^k$ with parameters $\boldsymbol{\theta}$, as the base function to be mixed over, because this would be too restrictive: A multivariate cumulative distribution function (cdf) on $[0, 1]^k$ is equal to zero if one of the components of $\boldsymbol{x}$ is zero (independently of the other components) and hence the mean function $\mu(\boldsymbol{x})$ would be equal to $\beta_0$ in these cases. Our model of using linear combinations of (shifted and scaled) univariate probability distribution function evaluated at a linear combination is considerably more flexible. In fact Lemma 3 provides a condition on the base distribution $G$ under which a plcmr function can be approximated using (2.2).

LEMMA 2.3 Any plcmr function $\sum c_j g_j(\boldsymbol{a}_j'\boldsymbol{x})$ on $[0, 1]^k \to \mathbb{R}$ can be approximated arbitrarily well in supremum norm by a function of form (2.2), provided

$$\sup_{x \in [0,1]} \left| \sum_{j=1}^J w_j G(x, \boldsymbol{\theta}_j) - G^*(x) \right|$$

can be made arbitrarily small, for $\boldsymbol{w} \in \mathbb{S}^J$, $\boldsymbol{\theta}_j \in \Theta$ and any distribution function $G^*$ on $[0, 1]$.

*Proof.* Proof: See Section B of the supplementary material available at *Biostatistics* online. □

In order to induce smoothness in the collection of conditional distributions over the predictor space, it is appealing to focus on continuous multivariate monotonic functions. In this case, the prior is dense in the space of continuous plcmr functions when the base distribution function $G$ can approximate any continuous cdf on $[0, 1]$ arbitrarily well. Several choices fulfill this property. One example is the distribution function of the standard two-sided power (TSP) distribution of van Dorp and Kotz (2002),

$$G(x, m, \nu) = \begin{cases} m\left(\frac{x}{m}\right)^{\nu} & 0 \leqslant x \leqslant m, \\ 1 - (1-m)\left(\frac{1-x}{1-m}\right)^{\nu} & m \leqslant x \leqslant 1, \end{cases}$$

where $m \in [0, 1]$ is the mode of the distribution, while $\nu \in \mathbb{R}_+$ determines the steepness at the mode. The TSP cdf is sufficiently flexible (see Bornkamp and Ickstadt, 2009), numerically straightforward to evaluate and available in a closed form (unlike e.g. the beta cdf).

Assuming the functions follow (2.2), a specification of the prior $H$ is completed with parametric priors for $\beta_0$ and $\beta_1$ and a nonparametric prior for the mixing distribution $Q$ based on Ongaro and Cattaneo (2004). A typical choice is to use $J - 1 \sim \text{Poi}(\rho)$, while the components $(m, \nu, \boldsymbol{\alpha})'$ of the base measure $Q_0$ are chosen to match prior information and prior uncertainty. A useful tool is to simulate the prior distribution and assess whether the resulting simulations lead to *a priori* plausible results. A useful default in this setting are uniform distributions on reasonable subsets of the parameter space.

### 2.3 *Implementation*

In this section, we describe the implementation and specific priors used. We assume independently distributed data $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, $i = 1, \ldots, n$, where $y_i$ is a univariate response, $\boldsymbol{x}_i \in [0, 1]^k$ are the covariates which are in a multivariate monotonic relationship with respect to $y_i$ and $\boldsymbol{z}_i \in \mathbb{R}^p$ are additional unconstrained covariates we would like to adjust for in the analysis.

For the mixing measure $P$ (from (2.1)), we use the truncated Dirichlet process with parameter $MP_0$, which provides an accurate approximation to the Dirichlet process while facilitating an efficient implementation via a blocked Gibbs sampler (Ishwaran and James, 2001). We choose the truncation level $N = 20$, which provides a conservative upper bound on the number of mixture components occupied by individuals in the sample (see Walker, 2007 or Papaspiliopoulos and Roberts, 2008, for versions of the blocked Gibbs sampler that avoid truncation). The resulting model for the data is

$$P \sim DP_N(MP_0),\ P = \sum_{h=1}^{N} \pi_h \delta_{(\mu_h(\boldsymbol{x}), \sigma_h^{-2})}$$

$$y_i | \boldsymbol{x}_i, \boldsymbol{z}_i, P \overset{\text{i.i.d.}}{\sim} \sum_{h=1}^{N} \pi_h \phi(\mu_h(\boldsymbol{x}_i) + \boldsymbol{\gamma}'\boldsymbol{z}_i, \sigma_h^2),$$

where $DP_N(MP_0)$ denotes the truncated Dirichlet process with parameter $MP_0$ and $N$ components. The weights $\pi_h$ have the truncated stick-breaking representation $\pi_h = v_h \prod_{l<h}(1 - v_h)$ with $v_h \overset{\text{i.i.d.}}{\sim}$ Beta$(1, M)$ and $\pi_N = 1 - \sum_{h=1}^{N-1} \pi_h$. The atoms in the mixture $(\mu_h, \sigma_h^{-2})$ are i.i.d. realizations of the base measure $P_0$ with $P_0 = H \times \text{Exp}(\omega)$, $H$ is the prior on the space of plcmr functions as introduced in the last section, and the $\mu_h(\boldsymbol{x})$ are hence given by $\mu_h(\boldsymbol{x}) = \beta_{0h} + \beta_{1h} \sum_{j=1}^{J_h} w_{hj} G(\boldsymbol{\alpha}'_{hj}\boldsymbol{x}, m_{hj}, \nu_{hj})$. Within each Gaussian mixture component, we hence use a prior for multivariate functions with full support on the space of plcmr functions. This produces a fairly flexible model for the component-specific mean functions. The main advantage of this flexibility is the fact that complex relationships can be approximated with relatively few components in the Gaussian mixture. We adjust for the additional predictors $\boldsymbol{z}_i$ linearily.

For $\beta_{0h}$, a normal distribution with parameter $m_0$ and variance $\nu_0^{-1}$ will be used. The parameter $m_0$ in turn has a normal prior with mean $\omega_0$ and variance $\tau_0$, while $\nu_0$ has a gamma prior with parameter $a_{\nu_0}$ and $b_{\nu_0}$. As a common focus is in assessing whether the predictors have any effect on the response distribution, it is important to allow a completely flat relationship. This can be accomplished through using a mixture of a point mass at 0 and an exponential distribution with parameter $\lambda$ as the prior for $\beta_{1h}$. The mixing probability $\pi_0$ is given a Beta$(a_{\pi_0}, b_{\pi_0})$ hyperprior, while $\lambda$ is given a gamma$(a_\lambda, b_\lambda)$ hyperprior. These hyperpriors induce a heavier tailed and hence a more robust specification.

In specifying the prior for the mixing distribution $Q$ in (2.2), we assign the number of components $J_h$ a Poisson$(\rho)$ distribution shifted by 1. The hierarchical prior for $\rho$ is given a gamma$(a_\rho, b_\rho)$ hyperprior. The weights $w_j$ in the mixture follow a uniform distribution on $\mathbb{S}^J$ for each $J$. For the base measure $Q_0$, we use the following distribution $U(0,1) \times U(1,20) \times D(\mathbf{1})$ for the parameter $(m, \nu, \boldsymbol{\alpha})$, where $D(\mathbf{1})$ is the $(k-1)$-dimensional Dirichlet distribution with parameter $(1,\ldots,1)'$, that is, a uniform distribution on the simplex. This corresponds to the prior assumption that all variables are equally important *a priori* and ensures an approximately linear increasing prior mean function for the univariate function $\mu$, with a reasonable variability.

For the precisions $\sigma_h^{-2}$, an exponential prior with parameter $\omega$ is used, where $\omega$ has a gamma hyperprior with parameters $a_\omega$ and $b_\omega$. The precision parameter $M$ of the truncated Dirichlet process is also treated as unknown and receives a conjugate gamma hyperprior with parameters $a_M$ and $b_M$. As a prior for the additional covariates $\boldsymbol{\gamma}$ a multivariate normal prior is used with mean $\boldsymbol{\mu}_\gamma$ and covariance matrix $\boldsymbol{\Sigma}_\gamma$.

To fit the model, Markov chain Monte Carlo (MCMC) techniques based on the blocked Gibbs sampler will be used. This algorithm introduces a latent class membership variable $K_i$ with categories $1,\ldots,N$ for each observation and iterates between updating the class memberships variables and the class-specific parameters. Most of the class-specific parameters can be updated by Gibbs steps, while a Reversible Jump MCMC step is used to update the functions $\mu_h^0(.)$. Additionally, the hyperparameters are updated in Gibbs steps, which is possible because conjugate hyperpriors were used. Section C of the supplementary material available at *Biostatistics* online contains a detailed description of the MCMC algorithm.

## 3. Application to epidemiologic data

In epidemiologic studies of the impact of potentially adverse environmental exposures on health responses, stochastic ordering restrictions are well motivated biologically. By including such biologically motivated restrictions, one can increase efficiency in estimating dose–response relationships, conducting inferences and risk assessments. Such restrictions are particularly helpful when there is more than 1 exposure, which is an increasing focus in epidemiology. The regression function is then multivariate, which makes it challenging to obtain precise dose–response estimates allowing for interactions unless strong parametric assumptions are made. Such parametric assumptions are difficult to justify scientifically, while order restrictions are natural. Using an order-restricted approach also provides an (at least partial) solution to the problem of extrapolation to low-dose risk. The current EPA guide (US Environmental Protection Agency (EPA), 2005) suggests linear extrapolation to low-dose risk because nonlinear models are, due to the sparsity of data, usually unreliable for extrapolation across low-dose regions. A nondecreasing constraint allows for the possibility of fitting a nonlinear model and can potentially improve the accuracy in quantitative risk assessment at low doses.

In this section, we apply our methodology to data from the US Collaborative Perinatal Project, which was conducted from 1959 to 1966. In the 1990s, a random sample of blood sera of the participants were reanalyzed for potential toxic substances (see Longnecker, Klebanoff, Zhou, and Brock, 2001 or Longnecker, Klebanoff, Brock, and Guo, 2005). We focus on the relationship between 2 exposures dichloro-diphenyl-dichloroethylene (DDE) (a metabolite of dichloro-diphenyl-trichloroethane [DDT]) and polychlorinated biphenyls (PCB) in the blood serum of the mother and gestational age of the newborn at

delivery (GAD). DDT is a pesticide that was primarily used as an agricultural insecticide and has mostly been banned in the 1970s, although it is still in use in some developing countries. PCB are organic compounds that were primarily used in electrical equipment and have been associated with a wide range of adverse health effects. Note that both the toxic substances were still in use in the United States when the data were collected.

Here, we focus on GAD (in weeks) in relationship to DDE (in μg/L) and the total serum PCB (in μg/L). For model fitting, we scaled these 2 predictors into the interval [0, 1]. As additional covariates, we include the serum triglycerides (originally measured in μg/L but standardized before model fitting) and the binary inputs smoking habit (1 = smoking) and race (1 = black). The parameters of these additional covariates were not constrained in the model. In the analysis, we excluded all values with length of gestation longer than 45 weeks for plausibility reasons and 68 cases with missing values, leaving a total sample size of 2312 for analysis.

For the hyperpriors, we chose the following weakly informative setting $\omega_0 = 30$, $\tau_0 = 10\,000$, $a_{\nu_0} = 0.1$, $b_{\nu_0} = 0.1$, $a_\lambda = 0.01$, $b_\lambda = 0.01$, $a_\rho = 1$, $b_\rho = 1.5$, $a_{\pi_0} = 1$, $b_{\pi_0} = 1$, $a_M = 1$, $b_M = N$, $a_\omega = 0.01$ and $b_\omega = 0.01$. The components of the prior for the mixing distribution $Q$ (used in the prior for $\mu^0(.)$) are chosen exactly as specified in Section 2.3. The prior for $\boldsymbol{\gamma}$ was chosen as a multivariate normal with mean vector $\mathbf{0}$ and diagonal covariance matrix $6.7\mathbf{I}_{3\times 3}$, where 6.7 is an estimate of the approximate variance in the observations. The prior for $\boldsymbol{\gamma}$ hence approximately reflects the information obtained in one observation.

We ran 3 independent chains of the MCMC sampling algorithm of Section 2 for 110 000 iterations after using a burn-in of 10 000 iterations and a thinning 10, leaving a total of 10 000 iterations per chain. The results between the chains were consistent; hence the presented analysis is based on the last 3500 iterations per chain resulting in a total of 10 500 simulations.

Figure 1 plots the bivariate posterior median of the 50% and the 10% quantile of the conditional distribution against DDE and PCB, when the additional covariates are set to 0. There it can be seen that both substances seem to affect the gestational age at delivery only slightly, with a steeper decrease in the direction of DDE for both the 10% and the 50% quantile. Comparing the 10% and the 50% conditional quantile, it becomes obvious that the 10% conditional quantile is affected slightly stronger by an increasing DDE and PCB as the posterior median is decreasing steeper and stronger in overall effect for the conditional 10% quantile (in particular in the DDE direction).

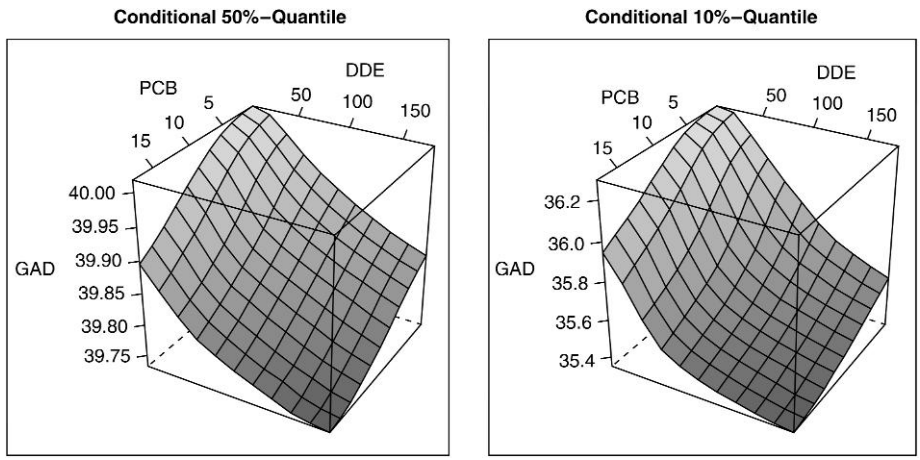


Fig. 1. Posterior median of the conditional 50% and 10% quantile.

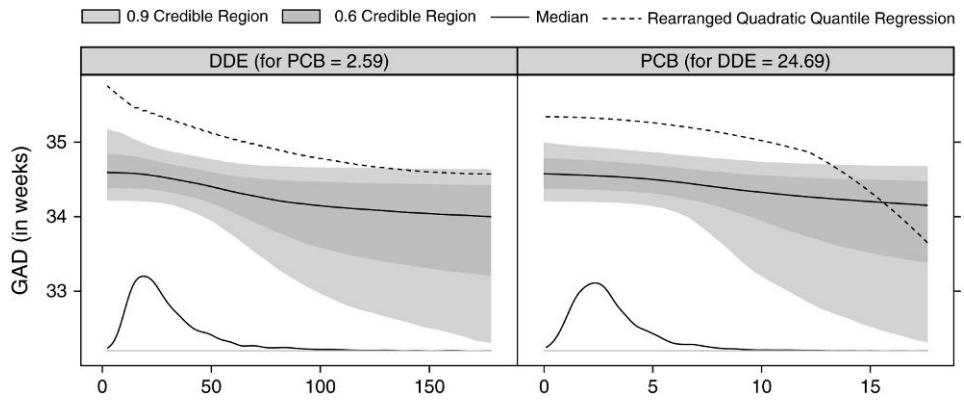


Fig. 2. Posterior of the conditional 5% quantile, (scaled and shifted) kernel density estimates of the covariate distribution, and a quadratic 5% quantile regression fit.

Figure 2 shows the posterior distribution of the conditional 5% quantile for DDE (holding PCB fixed at its median) and PCB (holding DDE at its median value), and all other covariates are set to 0. It can be seen that uncertainty in the estimate is quite large, in particular for DDE values larger than 50 and PCB values larger than 5. This can be attributed to the fact that most of the participants in the study had rather small PCB and DDE values, which is illustrated in the figure by including the (scaled and shifted) kernel density estimates of the covariates DDE and PCB. Primarily for DDE there seems to be an effect for persons with high exposure (i.e. larger than 40), but this effect cannot be estimated with high precision, as data are relatively sparse in this region.

Using the rq function in the quantreg R package (Koenker, 2008), we also fitted a parametric quantile regression model to the data (using linear and quadratic effects for DDE and PCB and the same additional covariates) and applied a monotonic rearrangement methodology (see Chernozhukov *and others*, 2010) to obtain a monotonic fit. Univariate rearrangement was applied in both directions, and to eliminate order dependence, we take the average over both possibilities. The results are superimposed in Figure 2. Even though quantile regression is based on quite a different statistical model, results of rearranged quantile regression roughly agree with our results. Note that rearrangement is essential here to stabilize the fit: the unconstrained estimate was slightly u-shaped for DDE and umbrella-shaped for PCB, with these shapes implausible *a priori*. However, it is not surprising to estimate a nonmonotone curve even if the truth is monotone when one does not place restrictions on the quantile regression curves. There is substantial uncertainty in estimating these curves, and our approach should convey efficiency advantages through both the monotonicity restriction and the use of borrowing of information from different quantiles in estimating the entire conditional response distribution. Accurate estimation of a single quantile regression curve in isolation may require large samples, particularly when the quantile is in the tails.

Figure 3 shows the conditional densities at different locations in the predictor space. For this purpose, we are looking at the conditional distribution, when both DDE and PCB are at their median value and at 2 extreme quantiles (the 1% and the 99% quantile). There it can be seen that the shape of the residual distribution looks nonnormal, with a more pronounced left tail. In the simulations, typically 2 to 5 components were employed (with modal value 3). It is interesting to see that the shape of the residual density largely remains identical throughout the predictor space, only the uncertainty intervals are larger in parts, where the data are sparser (see also Figure 2). It also seems that there is a tendency that the left tail gets slightly more pronounced, in particular at the extreme quantiles of the predictor space. This is in accordance with the results in Figure 1, where we observed that the 10% conditional quantile is more effected by DDE and PCB than the conditional 50% quantile.

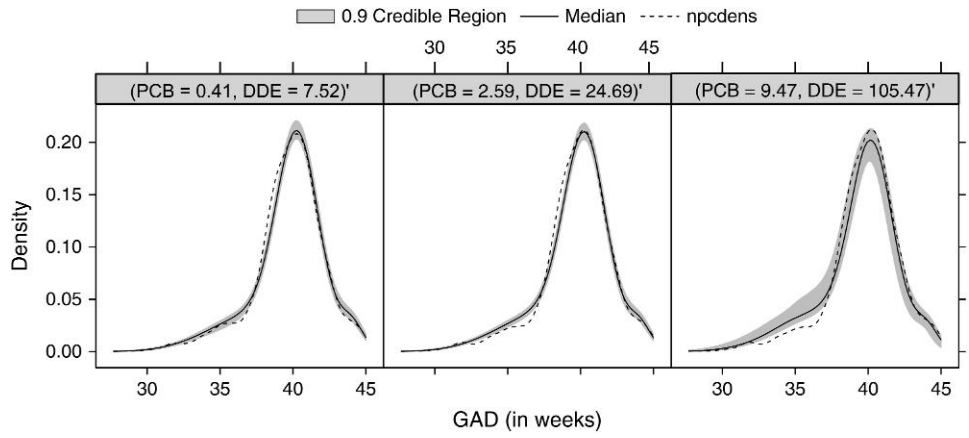


Fig. 3. Posterior distribution of the conditional densities at 3 locations in the input space, together with the estimate of the npcdens function.

Table 1. *Posterior summaries of additional covariates*

| Covariate | 0.05 quantile | Median | 0.95 quantile |
|---|---|---|---|
| Triglycerides | −0.29 | −0.22 | −0.14 |
| Smoking habit | −0.27 | −0.12 | 0.02 |
| Race | −0.78 | −0.62 | −0.47 |

Superimposed one can find an (unconstrained) conditional density estimate using the npcdens function in the np R package (Hayfield and Racine, 2008), implementing methods of Hall *and others* (2004). The fixed bandwidth was selected by maximum likelihood cross-validation. Both methods obtain rather similar results, with the main difference being in the left tail. Here, the Bayesian approach is less wiggly, which is at least partially due to the implicit averaging over the posterior simulations in the Bayesian approach (rather than using one particular point estimate), additionally the conditional density is considerably smaller in the left tail for larger values of the input. This is most likely due to the fact that stochastic ordering is imposed in our methodology, while the alternative approach is unconstrained.

It is also interesting to compare the results with those obtained by Wang and Dunson (2009), who modeled the conditional density of GAD versus DDE with univariate monotonic density regression. The posterior medians for the conditional densities are quite similar between the approaches, while the variability intervals for the conditional densities are wider in Wang and Dunson (2009). This is probably due to the fact that the bivariate shape constraint employed in this article restricts the conditional density considerably more than in the 1D case and hence reduces uncertainty in estimation.

Table 1 contains the credibility intervals for the (unconstrained) parameter estimates $\boldsymbol{\gamma}$ corresponding to the additional covariates. There it can be seen that both race and triglycerides have an impact on gestational age at delivery, while for smoking habit there seems to be a less pronounced negative effect, as its credibility interval contains zero.

## 4. CONCLUSIONS

In this paper, we have proposed a model for estimating conditional densities under the SO-$\mathcal{X}$ stochastic order, that is, the stochastic ordering is assumed with respect to multivariate continuous predictors. The

model relies on representing the conditional distributions as a location-scale mixture of normal distributions and the stochastic ordering constraint is imposed by assuming that the means of the components in the mixture are multivariate monotonically increasing functions. This type of model is extremely flexible, in particular we show that any collection of conditional densities under SO-$\mathcal{X}$ stochastic order can be approximated arbitrarily well by the proposed model. The model relies on a prior distribution for multivariate monotonic functions and we used positive linear combinations of monotonic ridge functions for this purpose. This class is quite flexible (compared to monotonic additive or single-index models for example) and seems well suited for sparse representation of multivariate functions.

The flexibility of the proposed model comes at the cost of being quite complex, and in some situations simpler models might be appropriate. When there is, for example, strong *a priori* evidence that the conditional residual density does not change in shape with $\boldsymbol{x}$, a semiparametric model can be adequate. One approach would be to model $y_i = \mu(\boldsymbol{x}_i) + \epsilon_i$, for a multivariate monotonic function $\mu(.)$, with the $\epsilon_i$ sampled i.i.d. from a homoscedastic residual density with mode zero. A related model was proposed by Lavine and Mockus (1995) in the univariate case. A mean zero constraint in the residual distribution would lead to an easier to interpret regression function, but such a constraint is not straightforward to include without also assuming symmetry about zero. In addition, regression models focused on characterizing predictor effects only on the center of the response distribution are not adequate in quite a few applications. This is particularly the case when the tails of the response distribution are of primary interest. For example, in many applications, the greatest interest is in the extremes corresponding to unusual health responses, pollution levels, financial events, or weather conditions, where the semiparametric model described above would not be appropriate. Hence, in such settings, most of the literature has focused on either using quantile regression models that focus on a single quantile (e.g. 95th) or models for extremes that effectively discard all information below a certain quantile. By using density regression methods, one simultaneously models all quantiles and hence allows inferences on differing predictor effects on the center and extreme quantiles while using all the available data. A concern in density regression is the curse of dimensionality as it is challenging to allow the response distribution to change flexibly over the predictor space. The incorporation of stochastic ordering constraints in multiple predictors is a highly effective strategy for reducing the effective dimensionality of the problem. An interesting direction for future research is the incorporation of high-dimensional predictors. In such cases, sparse maximum *a posteriori* estimation or sequential Monte Carlo methods may be preferred to MCMC.

## Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

## Acknowledgment

*Conflict of Interest:* None declared.

## Funding

Research Training Group “Statistical modeling” of the German Research Foundation (Deutsche Forschungsgemeinschaft) to B.B. and K.I.

## References

Antoniadis, A., Grégoire, G. and McKeague, I. W. (2004). Bayesian estimation in single-index models. *Statistica Sinica* **14**, 1147–1164.

Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association* **84**, 289–294.

BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, 930–945.

BORNKAMP, B. AND ICKSTADT, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics* **65**, 198–205.

CAI, B. AND DUNSON, D. B. (2007). Bayesian multivariate isotonic regression splines. *Journal of the American Statistical Association* **102**, 1158–1171.

CHENEY, W. AND LIGHT, L. (1999). *A Course in Approximation Theory*. Boston, MA: Brooks and Cole.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. AND GALICHON, A. (2010). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* doi: 10.1093/biomet/asp030.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. AND MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

DETTE, H. AND SCHEDER, R. (2006). Strictly monotone and smooth nonparametric regression for two or more variables. *The Canadian Journal of Statistics* **34**, 535–561.

DUNSON, D. AND PEDDADA, S. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika* **95**, 859–874.

DUNSON, D., PILLAI, N. AND PARK, J. (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B* **69**, 163–183.

DYKSTRA, R. L. AND ROBERTSON, T. (1982). An algorithm for isotonic regression for two or more independent variables. *Annals of Statistics* **10**, 708–716.

GELFAND, A. AND KOTTAS, A. (2000). Nonparametric Bayesian modeling for stochastic order. *Annals of the Institute of Statistical Mathematics* **53**, 865–876.

GELFAND, A. E., KOTTAS, A. AND MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

HALL, P., RACINE, J. S. AND LI, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99**, 1015–1026.

HAYFIELD, T. AND RACINE, J. S. (2008). Nonparametric econometrics: the np package. *Journal of Statistical Software* **27**, 1–32.

HOFF, P. (2003). Bayesian methods for partial stochastic orderings. *Biometrika* **90**, 303–317.

ISHWARAN, H. AND JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

ISHWARAN, H. AND ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *The Canadian Journal of Statistics* **30**, 269–283.

KARABATSOS, G. AND WALKER, S. G. (2007). Bayesian nonparametric inference of stochastically ordered distributions, with Polya trees and Bernstein polynomials. *Statistics and Probability Letters* **77**, 907–913.

KOENKER, R. (2008). *quantreg: Quantile Regression*. R package version 4.26. http://CRAN.R-project.org/package=quantreg.

LAVINE, M. AND MOCKUS, A. (1995). A nonparametric Bayes method for isotonic regression. *Journal of Statistical Planning and Inference* **46**, 235–248.

LEITENSTORFER, F. AND TUTZ, G. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8**, 654–673.

LONGNECKER, M., KLEBANOFF, M., BROCK, J. AND GUO, X. (2005). Maternal levels of polychlorinated biphenyls in relation to preterm and small-for-gestational-age birth. *Epidemiology* **16**, 641–647.

LONGNECKER, M. P., KLEBANOFF, M. A., ZHOU, H. AND BROCK, J. W. (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *The Lancet* **358**, 110–114.

MACEACHERN, S. N. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association, pp. 50–55.

MORTON-JONES, T., DIGGLE, P., PARKER, L., DICKINSON, H. O. AND BINKS, K. (2000). Additive isotonic regression models in epidemiology. *Statistics in Medicine* **19**, 849–859.

MUKARJEE, H. AND STERN, S. (1994). Feasible nonparametric estimation of multiargument monotone functions. *Journal of the American Statistical Association* **89**, 77–80.

MÜLLER, P., ERKANLI, A. AND WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

ONGARO, A. AND CATTANEO, C. (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics and Probability Letters* **67**, 33–45.

PAPASPILIOPOULOS, O. AND ROBERTS, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186.

PITMAN, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In: Ferguson, T. S., Shapley, L. S. and MacQueen, J. B. (editors), *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*. Hayward, CA: Institute of Mathematical Statistics, pp. 245–267.

SHIVELY, T. S., SAGER, T. W. AND WALKER, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society, Series B* **71**, 159–175.

US ENVIRONMENTAL PROTECTION AGENCY (2005). *Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F. Risk Assessment Forum*. Washington, DC.

WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation* **36**, 45–54.

WANG, L. AND DUNSON, D. B. (2009). Bayesian isotonic density regression. *Technical Report*. Biometrika (under revision).

VAN DORP, J. AND KOTZ, S. (2002). The standard two-sided power distribution and its properties: with applications in financial engineering. *The American Statistician* **56**, 90–99.

[*Received March 6, 2009; first revision July 16, 2009; second revision November 24, 2009; third revision December 18, 2009; accepted for publication January 5, 2010*]