Bayesian Analysis (2015) **10**, Number 4, pp. 821–847

A Fully Nonparametric Modeling Approach to Binary Regression

Maria DeYoreo* and Athanasios Kottas†

Abstract. We propose a general nonparametric Bayesian framework for binary regression, which is built from modeling for the joint response–covariate distribution. The observed binary responses are assumed to arise from underlying continuous random variables through discretization, and we model the joint distribution of these latent responses and the covariates using a Dirichlet process mixture of multivariate normals. We show that the kernel of the induced mixture model for the observed data is identifiable upon a restriction on the latent variables. To allow for appropriate dependence structure while facilitating identifiability, we use a square-root-free Cholesky decomposition of the covariance matrix in the normal mixture kernel. In addition to allowing for the necessary restriction, this modeling strategy provides substantial simplifications in implementation of Markov chain Monte Carlo posterior simulation. We present two data examples taken from areas for which the methodology is especially well suited. In particular, the first example involves estimation of relationships between environmental variables, and the second develops inference for natural selection surfaces in evolutionary biology. Finally, we discuss extensions to regression settings with ordinal responses.

Keywords: Bayesian nonparametrics, Dirichlet process mixture model, identifiability, Markov chain Monte Carlo, ordinal regression.

1 Introduction

Binary responses measured along with covariates are present in several problems in science and engineering. From a modeling perspective, interest centers on determining the regression relationship between the response and covariates. Standard approaches to this problem – both classical and Bayesian – involve potentially restrictive distributional assumptions as well as those of linearity in relating the response to the covariates. Common modeling techniques result in a limited range of monotonic, symmetric trends for the probability response curve, and assume that covariate effects are additive.

There has been substantial effort devoted to relaxing the functional form of the linear predictor through the use of basis functions, including spline based approaches (e.g., Denison et al., 2002), and generalized additive models (Hastie and Tibshirani, 1990), applied in a Bayesian context by Wood and Kohn (1998). Under these approaches, the linear predictor is modified by applying a smoothing function to each covariate separately and assuming the transformed covariates are additive in their effects. However, the underlying distributional assumption is still present through the link function.

*Department of Statistical Science, Duke University, Durham, NC, USA, maria.deyoreo@stat.duke.edu

†Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA, USA, thanos@ams.ucsc.edu

 DOI: 10.1214/15-BA963SI

The motivation for Bayesian nonparametric methodology lies in the notion that the model should support a wide range of distributional shapes and regression relationships. In an effort to create more flexible models to combat overdispersion and asymmetry, which the standard links cannot, several Bayesian semiparametric binary regression methods have been developed. Early work has targeted either the link, treating it as a random function with a nonparametric prior (Newton et al., 1996; Basu and Mukhopadhyay, 2000), or linearity, for instance, by viewing the intercept of the linear predictor as arising from an unknown distribution (Follmann and Lamberdt, 1989; Mukhopadyay and Gelfand, 1997; Walker and Mallick, 1997). More recently, Choudhuri et al. (2007) relaxed the linearity assumption by placing a Gaussian process prior on the argument of the inverse link. Trippa and Muliere (2009) assumed each binary response to arise from a random colored tessellation, and placed a Dirichlet process (DP) prior (Ferguson, 1973) on the space of colored tessellations.

Shahbaba and Neal (2009), Dunson and Bhattacharya (2011), and Hannah et al. (2011) have proposed nonparametric solutions to the regression problem with categorical responses. These approaches build off the work of Müller et al. (1996), which modeled the joint distribution of continuous responses y and covariates x with a DP mixture of normal distributions, inducing a flexible model for $\mathrm{E}(y \mid x)$. The idea of inducing a regression model through the joint response–covariate distribution is attractive, since in many settings the covariates are not fixed prior to sampling, including several applications in the environmental, biomedical, and social sciences.

We target problems of this type, proposing a flexible model for fully nonparametric binary regression, in which the responses and covariates arise together as random vectors, requiring a model for their joint distribution. The foundation of the proposed methodology is different from the existing nonparametric models for binary regression. We elaborate further in Section 2.4, but here note that a key feature of the proposed model involves the introduction of latent continuous responses, in similar spirit to parametric probit models; see, for instance, Albert and Chib (1993). Let $\{(y_i, x_i) : i = 1, \ldots, n\}$ denote the data, where each observation consists of a binary response y_i along with a vector of covariates, $x_i = (x_{i1}, \ldots, x_{ip})$. The continuous auxiliary variables, z_i, determine the observed binary responses y_i by their sign, such that $y_i = 1$ if and only if $z_i > 0$. Instead of seeking a nonparametric model for the regression function, we estimate the joint distribution of latent responses and covariates, $f(z, x)$, using a DP mixture of multivariate normal distributions, which induces a flexible model for the regression relationship, $\Pr(y = 1 \mid x)$. In addition to providing a general modeling platform, the latent responses are conceptually meaningful in many applications. The proposed model is shown to be identifiable provided the variance of z within each mixture component is fixed, a restriction implemented through a square-root-free Cholesky decomposition of the mixture kernel covariance matrix. This aspect of the model formulation retains computational efficiency in Markov chain Monte Carlo (MCMC) posterior simulation while enabling the use of priors more flexible than the inverse-Wishart distribution. We develop two approaches to prior specification for the covariance parameters, one which involves prior simulation and can be used for problems with a small number of covariates, and a second which is more straightforward to apply as the dimension of the covariate space increases.

In Section 2, we formulate the mixture model for binary regression. We discuss identifiability for the parameters of the mixture kernel distribution, as well as prior specification approaches, and give details for posterior inference. We also discuss related work on Bayesian nonparametric regression to place our contribution within this large body of literature. In Section 3, the methodology is applied to problems from environmetrics and evolutionary biology, using two data sets from the literature for illustration. The latter example involves estimation of fitness surfaces, a problem for which our method is particularly powerful relative to existing approaches. Section 4 concludes with a summary and discussion of possible extensions. Technical details on the identifiability result, prior specification and posterior simulation, and the expressions for the model comparison criterion used in Section 3 are provided in the appendices.

2 Methodology

2.1 The modeling approach

Focusing on p continuous covariates, $x = (x_1, \ldots, x_p)$, and a single binary response y, with corresponding latent continuous response z, a normal distribution is a natural choice for the kernel in a mixture representation for $f(z, x)$. The DP is then used as a prior for the random mixing distribution G, to create a mixture model of the form: $f(z, x; G) = \int \mathrm{N}_{p+1}(z, x; \mu, \Sigma)\mathrm{d}G(\mu, \Sigma)$, with $G \mid \alpha, \psi \sim \mathrm{DP}(\alpha, G_0(\cdot; \psi))$, where α is the DP precision parameter, and ψ the parameters of the DP centering distribution.

According to the DP constructive definition (Sethuraman, 1994), a $\mathrm{DP}(\alpha, G_0)$ realization G is almost surely of the form $\sum_{l=1}^{\infty} p_l \delta_{\nu_l}$, with ν_l independent realizations from G_0, and p_l arising through stick-breaking from beta random variables. In particular, let ζ_m be independent $\mathrm{beta}(1, \alpha)$, $m = 1, 2, \ldots$, and define $p_1 = \zeta_1$, and $p_l = \zeta_l \prod_{r=1}^{l-1}(1-\zeta_r)$, for $l = 2, 3, \ldots$; moreover, $\{\zeta_m : m = 1, 2, \ldots\}$ and $\{\nu_l : l = 1, 2, \ldots\}$ are independent sequences of random variables. Applying the constructive definition with $\nu_l = (\mu_l, \Sigma_l)$, the model admits a representation as a countable mixture of multivariate normals, $f(z, x; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_{p+1}(z, x; \mu_l, \Sigma_l)$.

For the normal kernel distribution, let μ^z denote the mean of z, μ^x denote the mean of x, and partition the covariance matrix such that $\Sigma^{zz} = \mathrm{var}(z)$, $\Sigma^{xx} = \mathrm{cov}(x)$, a $p \times p$ matrix, and $\Sigma^{zx} = \mathrm{cov}(z, x)$, a row vector of length p. Then, integrating over the latent response z, the induced model for the observables assumes the form

$$f(y, x; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_p(x; \mu_l^x, \Sigma_l^{xx}) \mathrm{Bern}\left(y; \Phi\left(\frac{\mu_l^z + \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(x - \mu_l^x)}{(\Sigma_l^{zz} - \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\Sigma_l^{zx})^t)^{1/2}}\right)\right), \quad (1)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Flexible inference for the binary regression functional can be obtained through $\Pr(y = 1 \mid x; G) = \Pr(y = 1, x; G)/f(x; G)$. Marginalizing over z in $f(z, x; G)$, the marginal distribution for x is $f(x; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_p(x; \mu_l^x, \Sigma_l^{xx})$, and thus the conditional regression function can be expressed as a weighted sum of the form $\sum_{l=1}^{\infty} w_l(x)\pi_l(x)$,

with weights $w_l(x) = p_l \mathrm{N}_p(x; \mu_l^x, \Sigma_l^{xx}) / \sum_{j=1}^{\infty} p_j \mathrm{N}_p(x; \mu_j^x, \Sigma_j^{xx})$, and probabilities

$$\pi_l(x) = \Phi\left(\frac{\mu_l^z + \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(x - \mu_l^x)}{(\Sigma_l^{zz} - \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\Sigma_l^{zx})^t)^{1/2}} \right). \tag{2}$$

Hence, the modeling approach implies a nonparametric mixture of probit regressions for the conditional response distribution. The probit probabilities, $\pi_l(x)$, are defined through component-specific intercept and slope parameters. The corresponding mixture weights, $w_l(x)$, depend on the covariates. As a consequence, the model structure allows for general binary regression relationships by favoring sets of probit regressions with varying weights depending on the location in the covariate space.

The dependence structure of the mixture kernel in $f(z, x; G)$ is key for general inference about the implied binary regression function. However, is it sensible to leave all elements of the kernel covariance matrix Σ unrestricted? In the case of a single mixture component, which arises in the limit as $\alpha \to 0^+$, the regression function $\Pr(y = 1 \mid x; G)$ reduces to a normal cumulative distribution function, as given in (2). This function takes the same value for any x when μ^z and Σ^{zx} are scaled by a positive constant c, and Σ^{zz} by c^2, indicating that different combinations of μ and Σ result in the same probability of positive response. Hence, there is an identification problem if μ and Σ are unrestricted. This limiting case of our model is a parametric probit model, albeit with random covariates. In this setting, if identification constraints are not imposed, prior distributions become increasingly important yet difficult to specify, and the use of noninformative priors can be problematic and create computational difficulties (Hobert and Casella, 1996; McCulloch et al., 2000; Koop, 2003). In addition, empirical evidence based on simulated data – generated through multivariate normal mixtures for the latent response and covariates – suggests that, without parameter restrictions, the correlations implied by the covariance matrices Σ_l are not representative of the correlations that generated the data. Moreover, the uncertainty bands for the estimated binary regression function become excessively wide at the boundaries of the covariate space. Computational issues involving the latent continuous responses diverging to $\pm\infty$ were also reported by Di Lucca et al. (2013), where the kernel variance in a DP mixture of normal autoregressive models was fixed to avoid this problem. For these reasons, and the fundamental belief that within a particular mixture component the corresponding parameters should be identifiable, we now focus on appropriately restricting the kernel of the mixture.

Here, we employ the standard definition of likelihood identifiability, such that a parameter θ for a family of distributions $\{f(x \mid \theta) : \theta \in \Theta\}$ is identifiable if distinct values of θ correspond to distinct probability density functions, that is, if $\theta \neq \theta'$, then $f(x \mid \theta)$ is not the same function of x as $f(x \mid \theta')$. Under our setting, the focus is on the kernel of the mixture model for the observed data, $f(y, x; G)$, which has the form

$$k(y, x; \eta) = \mathrm{N}_p(x; \mu^x, \Sigma^{xx}) \mathrm{Bern}\left(y; \Phi\left(\frac{\mu^z + \Sigma^{zx}(\Sigma^{xx})^{-1}(x - \mu^x)}{(\Sigma^{zz} - \Sigma^{zx}(\Sigma^{xx})^{-1}(\Sigma^{zx})^t)^{1/2}} \right) \right), \tag{3}$$

with $\eta = (\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zz}, \Sigma^{zx})$. Note that if z and x are independent in the normal mixture kernel, the probability in the Bernoulli response becomes $\Phi(\mu^z / (\Sigma^{zz})^{1/2})$; hence,

a restriction – for instance, on Σ^{zz} – is required for identifiability. This is in fact the only restriction necessary to obtain an identifiable kernel, and we thus retain the ability to estimate Σ^{zx}, which is significant in capturing the dependence of y on x under the mixture distribution. The specific result is given in the following lemma whose proof can be found in Appendix A.

Lemma 1. *The parameters* $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ *are identifiable in the model for observed data which has the form in (3), provided* Σ^{zz} *is fixed to a constant.*

While intuitively straightforward, fixing Σ^{zz} to a constant is challenging operationally. The usual conditionally conjugate inverse-Wishart choice for $G_0(\Sigma)$ does not offer the solution, since the single degree of freedom parameter of the inverse-Wishart distribution does not allow for one element of Σ to be fixed while freely estimating the rest of the matrix. One way to overcome this problem is to reparameterize Σ, using a square-root-free Cholesky decomposition. This decomposition is useful for modeling longitudinal data (Daniels and Pourahmadi, 2002), as well as specifying conditional independence assumptions for the elements of a normal random vector (Webb and Forster, 2008). Let β be a unit lower triangular matrix, and let Δ be a diagonal matrix with positive elements, $(\delta_1, \ldots, \delta_{p+1})$, such that $\Delta = \beta\Sigma\beta^t$. Hence, $\Sigma = \beta^{-1}\Delta(\beta^{-1})^t$, where β^{-1} is also lower triangular with all its diagonal elements equal to 1, and $\det(\Sigma) = \prod_{i=1}^{p+1} \delta_i$. Moreover, $\delta_1 = \Sigma^{zz}$, and thus the identifiability restriction can be implemented by setting the first element of Δ equal to a constant value; $\delta_1 = 1$ is used from this point forward. Instead of mixing directly on Σ, the mixing takes place on β and the p free elements of Δ, $(\delta_2, \ldots, \delta_{p+1})$. Hence, the mixture model for the joint density of the latent response and covariates is now written as:

$$f(z, x; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_{p+1}(z, x; \mu_l, {\beta_l}^{-1}\Delta_l(\beta_l^{-1})^t). \tag{4}$$

While this decomposition of Σ allows for the necessary flexibility in viewing only part of the covariance matrix as random, its real utility lies in the existence of a conditionally conjugate centering distribution G_0, which enables development of an efficient Gibbs sampler for posterior simulation. In particular, as shown in the next section, a multivariate normal G_0 component for the vector, $\tilde{\beta}$, of $p(p+1)/2 = q$ free elements of β, and independent inverse-gamma components for $\delta_2, \ldots, \delta_{p+1}$ result in full conditional distributions which are multivariate normal and inverse-gamma, respectively. Therefore, G_0 comprises independent components for $\mu, \tilde{\beta}$, and $\delta_2, \ldots, \delta_{p+1}$, such that it has the form $\mathrm{N}_{p+1}(\mu; m, V)\mathrm{N}_q(\tilde{\beta}; \theta, C) \prod_{i=2}^{p+1} \mathrm{IG}(\delta_i; \nu_i, s_i)$.

2.2 Posterior inference for binary regression

We implement posterior simulation using the blocked Gibbs sampler (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001), which is based on a finite truncation approximation to G, motivated by the DP constructive definition. Specifically, G is truncated to $G_N = \sum_{l=1}^{N} p_l \delta_{W_l}$, where $W_l = (\mu_l, \tilde{\beta}_l, \Delta_l)$, and $p_1, \ldots, p_{N-1}$ are defined through stick-breaking as in the DP definition, whereas $p_N = 1 - \sum_{l=1}^{N-1} p_l$. Then, intro-

ducing configuration variables $L = (L_1, \ldots, L_n)$, each taking values in $\{1, \ldots, N\}$, the hierarchical model for the data, given the latent continuous responses, $z = (z_1, \ldots, z_n)$, becomes

$$\begin{aligned} y_i \mid z_i &\overset{ind.}{\sim} 1_{(y_i=1)}1_{(z_i>0)} + 1_{(y_i=0)}1_{(z_i\le 0)}, \quad i = 1, \ldots, n, \\ (z_i, x_i) \mid W, L_i &\overset{ind.}{\sim} \mathrm{N}_{p+1}(z_i, x_i; \mu_{L_i}, \beta_{L_i}^{-1}\Delta_{L_i}(\beta_{L_i}^{-1})^t), \quad i = 1, \ldots, n, \\ L_i \mid p &\overset{ind.}{\sim} \sum_{l=1}^{N} p_l \delta_l(L_i), \quad i = 1, \ldots, n, \\ W_l \mid \psi &\overset{ind.}{\sim} \mathrm{N}_{p+1}(\mu_l; m, V)\mathrm{N}_q(\tilde{\beta}_l; \theta, C) \prod_{i=2}^{p+1} \mathrm{IG}(\delta_{i,l}; \nu_i, s_i) \quad l = 1, \ldots, N, \end{aligned}$$

where $W = (W_1, \ldots, W_N)$, and the prior density for $p = (p_1, \ldots, p_N)$ is given by a special case of the generalized Dirichlet distribution: $\alpha^{N-1}p_N^{\alpha-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1} \times \cdots \times (1-\sum_{l=1}^{N-2} p_l)^{-1}$. The full Bayesian model is completed with a gamma(a_α, b_α) prior for α, with mean a_α/b_α, and with conditionally conjugate hyperpriors for $\psi = (m, V, \theta, C, s_2, \ldots, s_{p+1})$, specifically: $m \sim \mathrm{N}_{p+1}(a_m, B_m)$, $V \sim \mathrm{IW}_{p+1}(a_V, B_V)$, $\theta \sim \mathrm{N}_q(a_\theta, B_\theta)$, $C \sim \mathrm{IW}_q(a_C, B_C)$, and $s_i \overset{ind.}{\sim}$ gamma(a_{s_i}, b_{s_i}), for $i = 2, \ldots, p+1$. Here, $S \sim \mathrm{IW}_k(a, B)$ indicates that the $k \times k$ positive definite matrix S follows an inverse-Wishart distribution with density proportional to $|S|^{-(a+k+1)/2}\exp\{-0.5\mathrm{tr}(BS^{-1})\}$. The notation $\delta_{i,l}$ is used for element i of the vector δ_l corresponding to the diagonal of Δ_l. Moreover, where convenient, we use the Σ notation for the structured covariance matrix, where the elements of Σ are computed through $\Sigma = \beta^{-1}\Delta(\beta^{-1})^t$.

A key feature of the modeling approach is that simulation from the full posterior distribution, $p(W, L, p, \psi, \alpha, z|\text{data})$, is possible via Gibbs sampling. We next discuss posterior simulation details focusing on a result that enables Gibbs sampling updates for the parameters that define the covariance matrices of the normal mixture components.

The updates for p and α are generic for any choice of mixture kernel; see Ishwaran and Zarepour (2000). Each L_i, $i = 1, \ldots, n$, is sampled from a discrete distribution on $\{1, \ldots, N\}$, with probabilities proportional to $p_l \mathrm{N}_{p+1}(z_i, x_i; \mu_l, \Sigma_l)$, for $l = 1, \ldots, N$. The full conditional distributions for the components of ψ are easily found using standard conjugate updating. The full conditional distribution for each z_i is a truncated version of the normal distribution $\mathrm{N}(\mu^z_{L_i} + \Sigma^{zx}_{L_i}(\Sigma^{xx}_{L_i})^{-1}(x_i - \mu^x_{L_i}), 1 - \Sigma^{zx}_{L_i}(\Sigma^{xx}_{L_i})^{-1}(\Sigma^{zx}_{L_i})^t)$, with the restriction $z_i > 0$ if $y_i = 1$, and $z_i \le 0$ if $y_i = 0$. The updates for the latent continuous responses are similar to the corresponding ones in Albert and Chib (1993), where data augmentation techniques were applied for MCMC inference in parametric models involving binary and ordinal data.

Letting $\{L_j^* : j = 1, \ldots, n^*\}$ be the vector of distinct values of L, the full conditional for W_l is proportional to $G_0(W_l \mid \psi)\prod_{j=1}^{n^*}\prod_{\{i:L_i=L_j^*\}} \mathrm{N}_{p+1}(z_i, x_i; \mu_{L_j^*}, \beta^{-1}_{L_j^*}\Delta_{L_j^*}(\beta^{-1}_{L_j^*})^t)$. If $l \notin \{L_j^* : j = 1, \ldots, n^*\}$, then $W_l \sim G_0(\cdot \mid \psi)$. If $l \in \{L_j^* : j = 1, \ldots, n^*\}$, then the full conditional distribution for each element of $W_l = (\mu_l, \tilde{\beta}_l, \delta_{2,l}, \ldots, \delta_{p+1,l})$ arises from the product of a normal likelihood component, based on $\{(z_i, x_i) : L_i = L_j^*\}$, and the base

distribution G_0. Therefore, when $l = L_j^*$, for $j = 1, \ldots, n^*$, the full conditional for μ_l is multivariate normal with mean vector $(V^{-1}+M_l\Sigma_l^{-1})^{-1}(V^{-1}m+\Sigma_l^{-1}\sum_{\{i:L_i=l\}}(z_i, x_i)^t)$ and covariance matrix $(V^{-1}+M_l\Sigma_l^{-1})^{-1}$, where $M_l = |\{i : L_i = l\}|$ is the size of mixture component l.

Lemma 2, whose proof can be found in Appendix A, provides the result for the posterior full conditional distributions of the $\tilde{\beta}_l$ and the $\delta_{i,l}$, for $i = 2, \ldots, p+1$. Before stating the lemma, we fix the required notation. As discussed earlier, vector $\tilde{\beta}$ consists of the lower triangle of free elements of matrix β. For instance, if $p = 2$, the mixture kernel is a trivariate normal, and the free elements of β are $(\beta_{21}, \beta_{31}, \beta_{32})$, corresponding to $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3)$. The matrix Δ contains vector δ on its diagonal. Let $r = p+1$ represent the dimension of the mixture kernel. Let d_i be a vector of length $r(r-1)/2 = q$, containing $r-1$ nonzero terms, occurring in elements $k(k+1)/2$ for $k = 1, \ldots, r-1$. Let T_i be a block diagonal matrix of dimension $q \times q$ with $r-1$ blocks, which can be constructed from square matrices $T_i^1, \ldots, T_i^{r-1}$ of dimensions $1, \ldots, r-1$. Matrix T_i^j occurs in rows and columns $j(j-1)/2+1$ to $j(j+1)/2$ of T_i.

Lemma 2. *Consider the following Bayesian probability model:*

$$(y_{i,1}, \ldots, y_{i,r}) \mid \mu, \tilde{\beta}, \delta \overset{ind.}{\sim} \mathrm{N}_r(\mu, \beta^{-1}\Delta(\beta^{-1})^t), \quad i = 1, \ldots, n,$$

with a multivariate normal prior for μ, independent inverse-gamma priors on the diagonal elements of Δ, $\delta_k \sim \mathrm{IG}(\nu_k, s_k)$, $k = 1, \ldots, r$, and a multivariate normal prior on the vector comprising the lower triangular elements of β, $\tilde{\beta} \sim \mathrm{N}_q(\theta, D)$. Then, the posterior full conditional distribution for δ_k, $k = 1, \ldots, r$, is an inverse-gamma distribution with shape parameter $\nu_k+0.5n$ and scale parameter $s_k+0.5\sum_{i=1}^n\{(y_{i,k}-\mu_k)+\sum_{j<k}\beta_{kj}(y_{i,j}-\mu_j)\}^2$. In addition, the posterior full conditional for $\tilde{\beta}$ is multivariate normal with mean vector $(D^{-1}+\sum_{i=1}^n T_i)^{-1}(D^{-1}\theta+\sum_{i=1}^n T_i d_i)$ and covariance matrix $(D^{-1}+\sum_{i=1}^n T_i)^{-1}$. Here, the non-zero elements of d_i are $-(y_{i,2}-\mu_2)/(y_{i,1}-\mu_1), \ldots, -(y_{i,r}-\mu_r)/(y_{i,r-1}-\mu_{r-1})$, and the (m,n)-th element of matrix T_i^j, for $j = 1, \ldots, r-1$, is given by $T_{i,mn}^j = (y_{i,m}-\mu_m)(y_{i,n}-\mu_n)/\delta_{j+1}$, for $m = 1, \ldots, j, n = 1, \ldots, j$.

This lemma provides the information necessary to obtain the remaining full conditional distributions, which are available in closed form. Let $y_i^* = (z_i, x_i)$ denote the augmented latent response-covariate vector, such that $y_{i,1}^* = z_i$ and $y_{i,j+1}^* = x_{ij}$, for $j = 1, \ldots, p$. Then, when $l = L_j^*$, for $j = 1, \ldots, n^*$, the full conditional distribution for $\delta_{k,l}$ is inverse-gamma with shape parameter $\nu_k + 0.5M_l$ and scale parameter $s_k + 0.5\sum_{\{i:L_i=L_j^*\}}\{(y_{i,k}^* - \mu_{k,l}) + \sum_{j<k}\beta_{kj,l}(y_{i,j}^* - \mu_{j,l})\}^2$. The full conditional for $\tilde{\beta}_l$ is multivariate normal with covariance matrix $(C^{-1} + \sum_{\{i:L_i=L_j^*\}} T_i)^{-1}$, and mean vector $(C^{-1} + \sum_{\{i:L_i=L_j^*\}} T_i)^{-1}(C^{-1}\theta + \sum_{\{i:L_i=L_j^*\}} T_i d_i)$. The p non-zero terms in the vector d_i are $-(y_{i,2}^* - \mu_{2,l})/(y_{i,1}^* - \mu_{1,l}), \ldots, -(y_{i,p+1}^* - \mu_{p+1,l})/(y_{i,p}^* - \mu_{p,l})$, and for $j = 1, \ldots, p$, the matrix T_i^j contains elements $T_{i,mn}^j = (y_{i,m}^* - \mu_{m,l})(y_{i,n}^* - \mu_{n,l})/\delta_{j+1,l}$, $m = 1, \ldots, j, n = 1, \ldots, j$.

We reiterate the significance of Lemmas 1 and 2: the former provides the identifiability restriction for the mixture kernel covariance matrix; the latter yields computationally tractable inference under the reparameterization used to incorporate the restriction.

These results combined are key to implementation of the modeling approach. They allow us to avoid computational complications arising from a non-identifiable model, at the same time enabling MCMC posterior simulation that is no more complicated than for DP mixtures of normal distributions with unrestricted covariance matrices.

The mixing distribution G, approximated by $G_N = (p, W)$, is imputed as a component of the posterior simulation algorithm, enabling full inference for any functional of $f(y, x; G)$. The binary regression functional $\Pr(y = 1 \mid x; G)$ is the main quantity of interest, and is estimated by $\Pr(y = 1, x; G)/f(x; G)$, where $\Pr(y = 1, x; G) = \sum_{l=1}^{N} p_l \mathrm{N}_p(x; \mu_l^x, \Sigma_l^{xx}) \pi_l(x)$, with $\pi_l(x)$ given in (2), and $f(x; G) = \sum_{l=1}^{N} p_l \mathrm{N}_p(x; \mu_l^x, \Sigma_l^{xx})$. Therefore, full inference for $\Pr(y = 1 \mid x; G)$ can be readily obtained for any covariate value x, providing a point estimate along with uncertainty quantification for the binary regression function. Inference can also be obtained for the covariate distribution, $f(x; G)$, as well as the covariate distribution conditional on a particular value of y, $f(x \mid y; G)$, which we refer to as inverse inferences, discussed further in the context of the data example of Section 3.1.

The truncation level for G_N can be chosen to any desired level of accuracy, using standard DP properties. For instance, a simple way to specify N is through the expectation for the partial sum of the original DP weights, $\mathrm{E}(\sum_{l=1}^{N} p_l \mid \alpha) = 1 - \{\alpha/(\alpha+1)\}^N$. This expression can be averaged over the prior for α to estimate the marginal prior expectation $\mathrm{E}(\sum_{l=1}^{N} p_l)$, which is used to specify N given a tolerance level for the approximation. For both data examples of Section 3, we used a gamma$(1, 0.5)$ prior for α, and set $N = 75$ which results in $\mathrm{E}(\sum_{l=1}^{75} p_l) \approx 0.99997$. An alternative approach, which involves also the sample size, is available through Theorem 2 in (Ishwaran and James, 2001) which provides an (approximate) upper bound of $4n \exp\{-(N-1)/\alpha\}$ on the L_1 distance between the prior predictive probability of the sample under the infinite dimensional prior G and its truncated version G_N. Using the 90th percentile of the gamma$(1, 0.5)$ prior as a (conservative) proxy for α in the expression for the upper bound, and setting $N = 75$, we obtain that the L_1 distance is smaller than 0.000047 and 0.000061 for the data sets of Sections 3.1 and 3.2, respectively.

As also pointed out by one of the referees, we note that there are alternative MCMC methods for DP mixtures that avoid explicitly truncating the random mixing distribution G; see, e.g., Griffin and Holmes (2010). These include MCMC algorithms that build from the marginalized version of the model after integrating G over its DP prior (e.g., Neal, 2000), as well as slice sampling algorithms (e.g., Kalli et al., 2011). These methods are arguably preferable to the blocked Gibbs sampler if the inference objectives are limited to posterior predictive estimation, which does not require posterior samples for G. Evidently, under the joint modeling approach for the response–covariate distribution, posterior simulation must be extended to G to obtain full inference about regression functionals. Hence, in our context, truncation of G is necessary also under the alternative MCMC algorithms; for instance, such truncation can be developed based on the conditional posterior distribution for G if the model is fitted with a marginal MCMC sampler (Gelfand and Kottas, 2002; Ishwaran and Zarepour, 2002). Our contribution being on flexible mixture modeling for binary regression, we view the choice of the posterior simulation method as one that should balance efficiency with ease of

implementation. From a practical point of view, we have not observed any differences in inference results obtained with the blocked Gibbs sampler and with marginal MCMC algorithms augmented with posterior sampling for G.

2.3 Prior specification

We discuss two approaches to hyperprior specification considering the limiting case of the model as $\alpha \to 0^+$, which corresponds to a single mixture component. Both approaches use an approximate range and center of x, say r^x and c^x, both vectors of length p, with the objective being to center and scale the mixture kernel appropriately using only a small amount of prior information. Under the assumption of a single mixture component, the marginal moments are given by $\mathrm{E}((z,x)^t) = a_m$, and $\mathrm{cov}((z,x)^t) = \mathrm{E}(\Sigma) + B_m + (a_V - p - 2)^{-1}B_V$. We therefore set $a_m = (0, c^x)$, and let $B_m = 0.5\mathrm{diag}(1, (r_1^x/4)^2, \ldots, (r_p^x/4)^2)$, using c_j^x and $(r_j^x/4)^2$ as proxies for the marginal mean and variance of x_j, for $j = 1, \ldots, p$. We set $a_V = p + 3$, which yields a dispersed prior for V albeit with finite prior expectation, and determine B_V such that $(a_V - p - 2)^{-1}B_V = B_m$. Next, we must determine values for the prior hyperparameters associated with $\tilde{\beta}$ and the δ_i, and this is where the two approaches differ.

The first approach uses prior simulation to induce approximately uniform$(-1, 1)$ priors on all correlations of the mixture kernel covariance matrix, while appropriately centering the variances. Note that the number of correlations grows at a rate of $O(p^2)$, making this approach practically feasible only for a small number of covariates. In particular, with a single covariate the kernel covariance matrix comprises correlation, $\rho = -\tilde{\beta}(\tilde{\beta}^2 + \delta)^{-1/2}$, and variance, $\sigma^2 = \tilde{\beta}^2 + \delta$. Here, $\tilde{\beta}$ and δ are scalar parameters with G_0 components $\mathrm{N}(\theta, c)$ and $\mathrm{IG}(\nu, s)$, respectively, and the hyperpriors are: $\theta \sim \mathrm{N}(a_\theta, b_\theta)$, $c \sim \mathrm{IG}(a_c, b_c)$, and $s \sim \mathrm{gamma}(a_s, b_s)$. We set $\mathrm{E}(\tilde{\beta}) = a_\theta = 0$, and build the specification for the other hyperparameters from $\mathrm{E}(\sigma^2) = b_\theta + b_s^{-1}(\nu - 1)^{-1}a_s + (a_c - 1)^{-1}b_c$. We first fix the shape parameters ν, a_c and a_s to values that yield relatively large prior dispersion, for instance, $\nu = a_c = 2$ results in infinite prior variance for the inverse-gamma distributions. Next, using $(r^x/4)^2$ as a proxy for $\mathrm{E}(\sigma^2)$, we find constants k_1, k_2, k_3, where $k_1 + k_2 + k_3 = 1$, such that $k_1(r^x/4)^2 \approx b_\theta$, $k_2(r^x/4)^2 \approx b_s^{-1}(\nu - 1)^{-1}a_s$, and $k_3(r^x/4)^2 \approx (a_c - 1)^{-1}b_c$, while at the same time the induced prior on ρ is approximately uniform on $(-1, 1)$. Finally, with k_1, k_2, k_3 specified, b_θ, b_s, and b_c can be determined accordingly.

While this approach is attractive when a relatively noninformative prior is desired, it is difficult to implement with a moderate to large number of covariates. An alternative strategy arises from studying the distribution which is implied for (β, Δ) if Σ is inverse-Wishart distributed. Using properties of partitioned Wishart and inverse-Wishart matrices (Box and Tiao, 1973; Eaton, 2007), it can be shown that $\Sigma \sim \mathrm{IW}_{p+1}(v, T)$ implies inverse-gamma distributions for the δ_i, and a normal distribution for $\tilde{\beta}$ given the δ_i. It is customary to specify noninformative priors on the inverse-Wishart scale, usually fixing the degrees of freedom parameter to a small value, and the inverse scale parameter to be a diagonal matrix. Here, we use the smallest possible integer value for v that ensures a finite expectation for the $\mathrm{IW}_{p+1}(v, T)$ distribution, that is, $v = p + 3$, and set $\mathrm{E}(\Sigma) =$

$T = \text{diag}(T_1, \ldots, T_{p+1}) = \text{diag}(1, (r_1^x/4)^2, \ldots, (r_p^x/4)^2)$. Then, as shown in Appendix B, the distributions implied on δ_i, for $i = 2, \ldots, p+1$, are $\text{IG}(0.5(v + i - (p+1)), 0.5T_i)$. Hence, we let $\nu_i = 0.5(v + i - (p+1))$, and $\text{E}(s_i) = 0.5T_i$; for the data examples of Section 3, we worked with exponential priors for the s_i resulting in $b_{s_i} = 2/T_i$. Moreover, the $\text{IW}_{p+1}(v, T)$ distribution implies a normal distribution for the ith row of matrix β, given δ_i; see Appendix B. This can be translated into a distribution for $\tilde{\beta}$ conditionally on the δ_i, specifically, a normal distribution with zero mean vector and covariance matrix $\text{BD}(S_1, \ldots, S_p)$, which denotes a block diagonal matrix with elements $S_i = \delta_{i+1}\text{diag}(T_1^{-1}, \ldots, T_i^{-1})$, for $i = 1, \ldots, p$. Now, after marginalizing out θ, the G_0 prior component for $\tilde{\beta}$ becomes $\text{N}_q(a_\theta, B_\theta + C)$. We therefore specify a_θ to be equal to the zero mean vector, and since we have a further prior on C, and S_i is a function of δ_{i+1}, we set $B_\theta + \text{E}(C) = \text{BD}(\hat{S}_1, \ldots, \hat{S}_p)$, where $\hat{S}_i$ is a proxy for S_i obtained by replacing δ_{i+1} with its marginal prior mean. Finally, B_θ and $\text{E}(C)$ can be specified to be equal to each other or assigned different portions of $\text{BD}(\hat{S}_1, \ldots, \hat{S}_p)$.

2.4 Related work

The joint response–covariate modeling approach to construct general conditional response distributions was presented originally by Müller et al. (1996), and has been more recently studied in different settings; see, e.g., Müller and Quintana (2010), Park and Dunson (2010), Taddy and Kottas (2010, 2012), and Wade et al. (2014). Existing joint models which can be used for binary regression include those of Shahbaba and Neal (2009), Dunson and Bhattacharya (2011) and Hannah et al. (2011). In Section 3.1, we discuss the special case of our model arising from $\Sigma^{zx} = 0$ in the mixture kernel covariance matrix, which has been previously proposed with the further restriction that Σ^{xx} is diagonal (Dunson and Bhattacharya, 2011). There, the simplicity of independence among covariates within mixture components was viewed as appealing, and the response was modeled as independent of the covariates within the kernel, resulting in what was termed a product-kernel. In a related approach, Shahbaba and Neal (2009) also build a model for the joint distribution $f(y, x)$ by separately estimating $f(x)$ and $f(y \mid x)$, where the latter is assumed to be a multinomial logit model within a mixture component. Due to the difficulties arising from estimation of full covariance matrices unless the inflexible inverse-Wishart prior is used, they too take $x_1, \ldots, x_p$ to be independent within each component. This idea was generalized by Hannah et al. (2011) to allow any standard generalized linear model to replace the multinomial logit model.

The independence assumptions discussed above are, in general, restrictive. The proposed justification is that because independence is imposed only within each component, dependence arises when more than one component is contained in the mixture. Therefore, the ability of product-kernel models to approximate the regression relationship and the covariate distribution is enhanced through the mixture. However, in order to capture the covariate distribution and the dependence of y on x in complex problems, there is need for models which allow for dependence within clusters. Dunson and Bhattacharya (2011) note that if interest centers on quantifying dependence, then there is no need to introduce a response, and the method for joint modeling can still be used in this case. If estimation of dependence is, in fact, the goal, this is clearly more adequately achieved

when random variables are allowed to depend on one another through more than just clustering. In this work, the introduction of latent variables and reparameterization of the covariance matrix allow these assumptions to be relaxed.

The other main class of nonparametric regression models is based on a conditional approach, in which $f(y \mid x)$ is modeled directly through nonparametric mixtures; see, e.g., the review in Müller and Mitra (2013). These models require a prior on the mixing distribution indexed by x, and include the dependent Dirichlet process models of MacEachern (2000), DeIorio et al. (2004), Gelfand et al. (2005), Griffin and Steel (2006), Taddy (2010), as well as other approaches which do not retain the Dirichlet process marginally (e.g., Dunson and Park, 2008; Rodriguez and Dunson, 2011). In general, these modeling approaches require replication in the "covariate" x (e.g., space, time, or a categorical covariate), which prohibits their use in our setting.

The nonparametric mixture regression model of Antoniano-Villalobos et al. (2014) expresses the mixture kernel for $y \mid x$ as a normal linear regression model, with weights $w_j(x) \propto p_j k(x \mid \psi_j)$ for some kernel function k which has support on the covariate space. The interpretation for $w_j(x)$ is through the probability that an observation with covariate value x comes from the jth regression model, and this is not related to whether the covariates are fixed or random. Our modeling formulation is fundamentally different in terms of its motivation, as we begin with a model for the joint latent response-covariate distribution, which implies a mixture form for $f(z \mid x)$ with covariate-dependent weights and kernels. This happens to have a similar form to the model of Antoniano-Villalobos et al. (2014) when a normal kernel is chosen for $k(x \mid \psi_j)$.

Regarding other nonparametric models for categorical responses built through latent variables, Di Lucca et al. (2013) present a model for a time series of binary responses induced from a DP mixture of normal autoregressive models for the latent responses conditional on fixed covariates, with mixture weights that are not covariate-dependent. While not a regression model, Canale and Dunson (2011) also use discretized latent variables, modeled with a DP mixture of normals, to model discrete count variables.

3 Data illustrations

3.1 Ozone data

Ozone is a gas which has detrimental consequences when it occurs near the Earth's surface. Ground-level ozone is a harmful pollutant, making up most of the smog which is visible in the sky over large cities. Because of the effects ozone has on the environment and our health, its concentration is monitored by environmental agencies. Rather than recording the actual concentration, presence or absence of an exceedance over a given ozone concentration threshold may be measured, and the number of ozone exceedances in a particular area is of interest.

To develop an illustrative example, we work with data set `ozone` from the "ElemStatLearn" R package. The data set includes measurements of ozone concentration in parts per billion, wind speed in miles per hour, temperature in degrees Fahrenheit, and

radiation in langleys, recorded over 111 days from May to September of 1973 in New York. To construct a binary ozone exceedance response, we define an exceedance as an ozone concentration which is larger than 70 parts per billion, a threshold determined to provide an appropriate degree of public health protection (The Environmental Protection Agency, 2014). Therefore, we can model the probability of an ozone exceedance as a function of wind speed, temperature, and radiation, using the DP mixture binary regression model. In addition, the modeling approach is evidently appropriate here, since it is natural to estimate conditional relationships between the four environmental variables through modeling the stochastic mechanism for their joint distribution. We are not suggesting dichotomizing a continuous response in practice, but use this example to illustrate a practically relevant setting in which a binary response may arise as a discretized version of a continuous response. Moreover, the existence of the continuous ozone concentrations enables comparison of inferences from the binary regression model with a model based on the actual continuous responses.

Prior specification was performed using the first approach discussed in Section 2.3 that favors uniform priors for the correlations of the kernel covariance matrix. Although the corresponding priors were not all close to the uniform on $(-1, 1)$ under the inverse-Wishart prior specification approach, both methods resulted in prior mean estimates for $\Pr(y = 1 \mid x_j)$ that were, for each of the three random covariates, roughly constant around 0.5, with 90% interval bands that essentially span the unit interval. All posterior inference results discussed below were robust to the prior choice.

The marginal binary response curves for the probability of exceedance as a function of wind speed, temperature, and radiation, are shown in the top row of Figure 1. There is a decreasing trend in probability as wind speed increases, with the probability being essentially 0 when wind speed is greater than 15 mph. The opposite trend is observed with temperature, as the probability of exceedance is near 0 when temperature is less than 75 degrees, and above 0.8 when temperature exceeds 90 degrees. A non-monotonic unimodal response curve is obtained as a function of radiation, with peak probability occurring at moderate values of radiation, and declining with higher and lower values. Bivariate surfaces indicating probability of exceedance as a function of temperature and wind speed, as well as radiation and wind speed, are shown in Figure 2. An attractive feature of the joint modeling approach, relative to models that treat the covariates as fixed, is that interactions and dependence between covariates are naturally accounted for, without the need to make simplifying assumptions (such as additivity in covariate effects) or to accommodate interactions with additional terms.

For this illustrative data example, the continuous ozone concentration responses are also available. We can therefore compare the binary regression model inferences for $\Pr(y = 1 \mid x_j)$ with the ones for $\Pr(z > 70 \mid x_j)$, under the corresponding density estimation model – a DP mixture based on a four-dimensional normal kernel with unrestricted covariance matrix – applied to the original data set $\{(z_i, x_i) : i = 1, \ldots, 111\}$. Results are shown in the bottom row of Figure 1, based on a prior choice for the density estimation model that induces prior estimates for the $\Pr(z > 70 \mid x_j)$ curves that are similarly diffuse to the ones for $\Pr(y = 1 \mid x_j)$. Save for some differences in the uncertainty bands, the density estimation model reveals similar trends for the regression functions to the ones uncovered by the binary regression model.

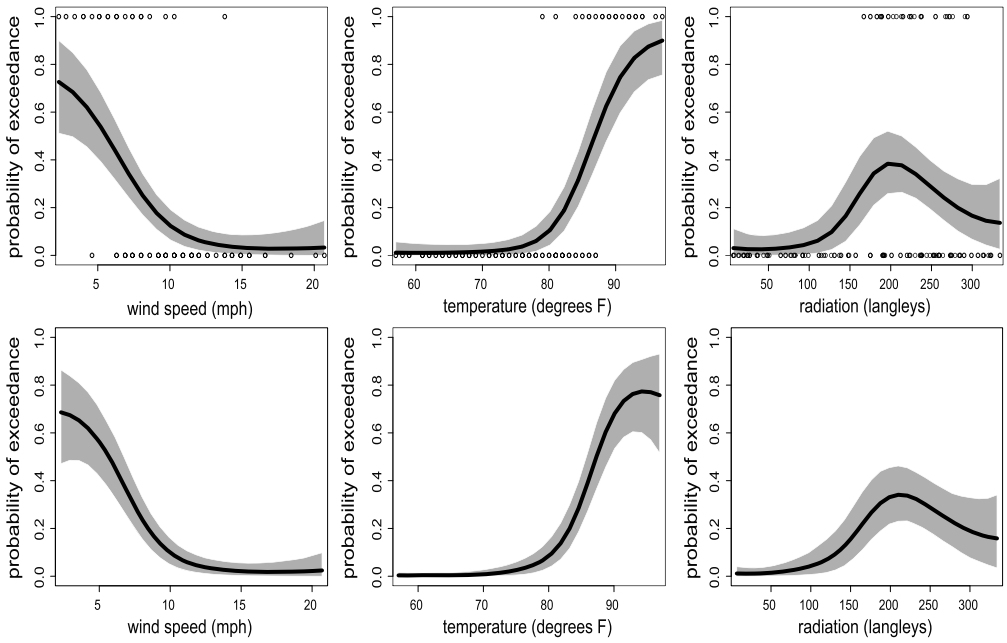

Figure 1: Ozone data. Posterior mean (solid line) and 90% uncertainty bands (in gray) for probability of exceedance versus wind speed (left panels), temperature (middle panels), and radiation (right panels). The top row plots results under the binary regression model, including the binary response data in each panel. The bottom row shows results under the density estimation model. Refer to Section 3.1 for further details.

As another appealing consequence of estimating the joint response–covariate distribution, we can obtain inference for the distribution of covariates conditional on a particular value of y. These inverse inferences may be of interest in many settings, as they indicate how the covariate distribution differs given a positive versus a negative binary response. Such inferences are not possible under a model directly for the conditional response distribution (with the implicit assumption of fixed covariates). Figure 3 shows estimates for the density of each covariate conditional on the binary exceedance response, $f(x_j \mid y = 1)$ and $f(x_j \mid y = 0)$, for $j = 1, 2, 3$. Note that when an exceedance occurs, temperature is generally higher and wind speed lower. In addition, the conditional densities associated with an exceedance have smaller dispersion than those associated with a non-exceedance, indicating that a smaller range of covariate values are supported when an exceedance occurs.

Recall from Section 2.1 that if we make the simplifying assumption $\Sigma^{zx} = 0$ for the covariance matrix of the kernel in $f(z, x; G)$, we obtain a kernel for $f(y, x; G)$ that comprises independent components $\mathrm{N}_p(x; \mu^x, \Sigma^{xx})$ and $\mathrm{Bern}(y; \Phi(\mu^z))$. The implied conditional regression function is again a weighted sum of probabilities with the same covariate-dependent weights as the proposed model, but probabilities which are not functions of x; the probability $\pi_l(x)$ in expression (2) reduces to $\pi_l = \Phi(\mu_l^z)$. This

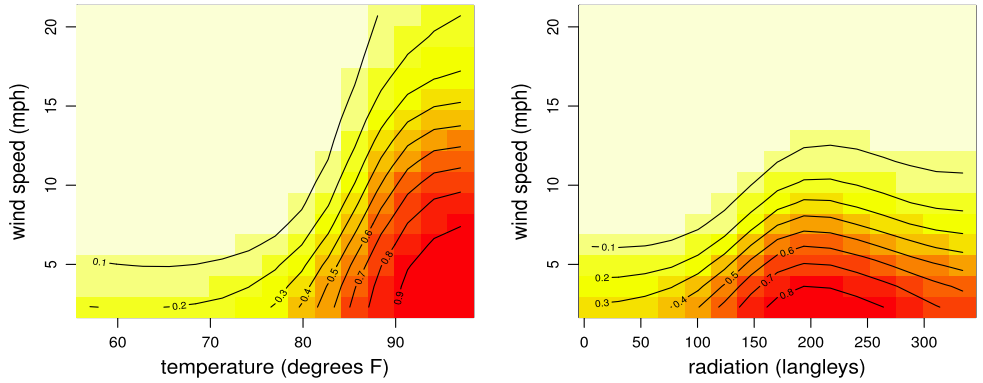

Figure 2: Ozone data. Posterior mean surface for probability of exceedance versus temperature and wind speed (left panel), and radiation and wind speed (right panel). Probabilities ranging from 0 to 1 are indicated by a spectrum of colors from white to red.

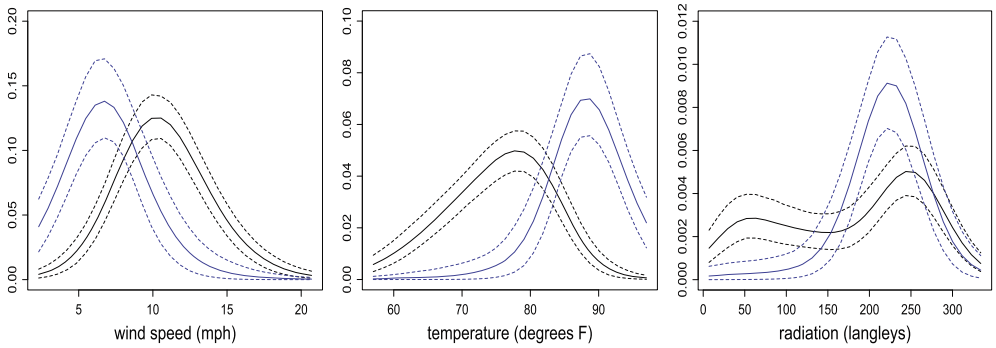

Figure 3: Ozone data. Posterior mean estimates (solid lines) and 90% uncertainty bands (dashed lines) for the density of wind speed (left panel), temperature (middle panel), and radiation (right panel), given an ozone concentration exceedance (blue) and non-exceedance (black).

is a limitation, as all dependence must be captured through clustering. As discussed in Section 2.4, mixtures of this product-kernel form have been previously proposed in the literature (Dunson and Bhattacharya, 2011).

We fitted the simpler product-kernel model to the ozone data, using hyperpriors that induce similarly diffuse prior estimates for the regression functions with the general binary regression model. Differences in the response probabilities produced by the product-kernel mixture model (not shown here) tend to occur at peaks or low points of the curves in Figure 1. In general, the product-kernel model underestimates the probability surface or curve when it takes a high value, and overestimates regions of low probability. In addition, the uncertainty bands from the product-kernel model are generally wider than those produced by the proposed model.

For a more formal comparison, we use the posterior predictive loss criterion of Gelfand and Ghosh (1998). The criterion favors the model m that minimizes the predictive loss measure $D_k(m) = P(m) + \{k/(k+1)\}G(m)$, with penalty term $P(m) = \sum_{i=1}^{n} \text{var}^{(m)}(y_{new,i} \mid \text{data})$, and goodness of fit term $G(m) = \sum_{i=1}^{n}\{y_i - \text{E}^{(m)}(y_{new,i} \mid \text{data})\}^2$. Here, $\text{E}^{(m)}(y_{new,i} \mid \text{data})$ is the mean under model m of the posterior predictive distribution for replicated response $y_{new,i}$ with corresponding covariate value x_i. The variance is similarly defined. Details involving expressions contributing to $D_k(m)$ for each model are given in Appendix C, but note that computations are based on the conditional posterior predictive distribution of y given x. The penalty term under the product-kernel model is 10.17, while it is 7.95 under the proposed model, and the goodness of fit terms are 4.17 and 4.08, respectively. Hence, regardless of the choice for constant k, the criterion favors the general DP mixture binary regression model.

3.2 Estimating natural selection functions in song sparrows

In addition to enabling more general modeling of binary regression relationships, the latent variables may be practically relevant in specific applications. Often, we may only observe whether or not some event occurred, although there exists an underlying continuous response which drives the binary observation. The ozone data was used to illustrate an environmental application for which the latent continuous responses are actually present. In applications in biology, the latent response may represent maturity, which is recorded on a discretized scale, or an unobservable trait or measure of health. In general, the continuous responses may be latent either because they are actually unobservable, or as consequence of recording taking place on a discretized scale. As an example of the former scenario, consider a binary response which represents survival. While we only observe survival on a binary scale, it is meaningful to conceptualize an underlying process which drives survival. Quantifying the probability of survival as a function of phenotypic traits is of great interest in evolutionary biology (Lande and Arnold, 1983; Schluter, 1988; Janzen and Stern, 1998). Survival can be thought of as a measure of fitness, and the fitness surface describes the relationship between phenotypic traits and fitness. The proposed methodology is particularly well-suited for this area of application, as it allows flexible inference for the shape of the fitness surface and for the distribution of population traits under a joint modeling framework that incorporates the scientifically relevant latent fitness responses.

As an illustration, we consider a standard data set from the relevant literature that records overwinter mortality along with six morphological traits in a population of 145 female song sparrows (Schluter and Smith, 1986). The traits measured consist of weight, wing length, tarsus length, beak length, beak depth, and beak width. Our initial analysis included four traits – weight, wing length, tarsus length, and beak length – as beak width and depth are highly discretized, correlated with beak length, and did not appear to be associated with a trend in survival. This analysis revealed tarsus length and beak length to be the main targets of selection, which is consistent with the findings of Schluter and Smith (1986). A key objective in this example is to obtain inferences for functionals used to assess the strength and form of natural selection acting on phenotypic traits, and we thus focus on the two traits associated with survival.

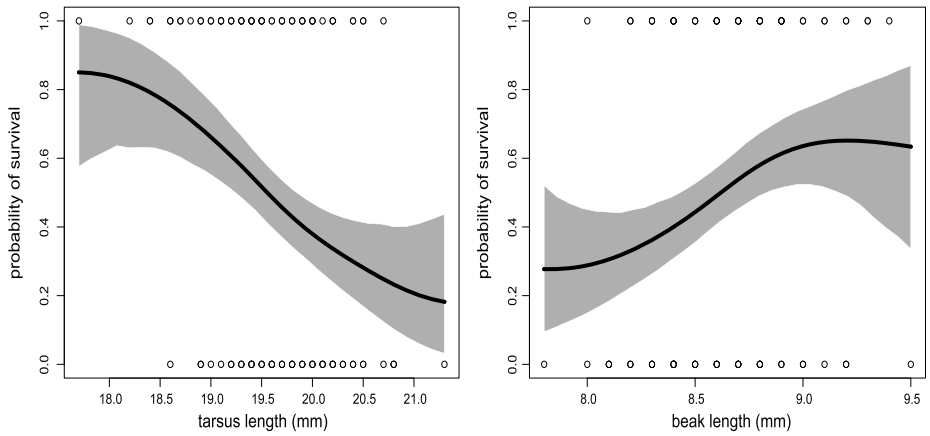

Figure 4: Song sparrows data. Posterior mean (solid line) and 90% uncertainty bands (in gray) for the probability of survival as a function of tarsus length (left panel) and beak length (right panel). Plotted in each panel are the corresponding observations.

The model was applied with standardized covariates tarsus length (x_1) and beak length (x_2), measured in millimeters, using the second approach to prior specification involving the inverse-Wishart distribution. The estimated selection curves are shown in Figure 4, revealing a strong decreasing trend in fitness over tarsus length, in which a sparrow with tarsus length 20.55 millimeters has a 10% lower probability of surviving overwinter than a sparrow with tarsus length just 0.5 mm shorter. The opposite trend in fitness is present over beak length, as longer beaks are associated with higher probabilities of survival. The posterior median estimate for the probability of survival as a function of both traits (Figure 5, left panel) confirms that the combination of long beaks and short tarsi is optimal for fitness; importantly, it also indicates that a short tarsus provides the more significant contribution to higher probability of survival. The corresponding posterior interquartile range estimate (Figure 5, right panel) depicts more uncertainty in the survival probability surface for sparrows having both a short beak and short tarsus, and those with both a long beak and long tarsus.

For each of the two traits, we estimated the standardized directional selection differential, $\bar{x}_j^* - \bar{x}_j$, which provides a measure of selection intensity representing the change in mean value of a phenotype produced by selection (Lande and Arnold, 1983). Here, $\bar{x}_j = \int x_j f(x_j) \mathrm{d}x_j$ is the mean value of phenotypic trait x_j before selection, and $\bar{x}_j^* = \int x_j f(x_j \mid y = 1) \mathrm{d}x_j = \{\Pr(y = 1)\}^{-1} \int x_j \Pr(y = 1, x_j) \mathrm{d}x_j$ is the mean value after selection; the marginal probability $\Pr(y = 1)$ is referred to as mean absolute fitness. Under our model, $\bar{x}_j = \sum_{l=1}^{N} p_l \mu_l^{x_j}$, the mean absolute fitness is given by $\sum_{l=1}^{N} p_l \Phi(\mu_l^z)$, and $\int x_j \Pr(y = 1, x_j; G_N) \mathrm{d}x_j$ is approximated with a Riemann sum. The posterior mean estimate for the standardized selection differential for tarsus length was -0.31, with a 90% posterior credible interval of $(-0.46, -0.18)$. For beak length, the posterior mean and 90% credible interval for the standardized selection differential were 0.22 and $(0.09, 0.36)$. Note that these intervals do not contain zero. Combined with the esti-

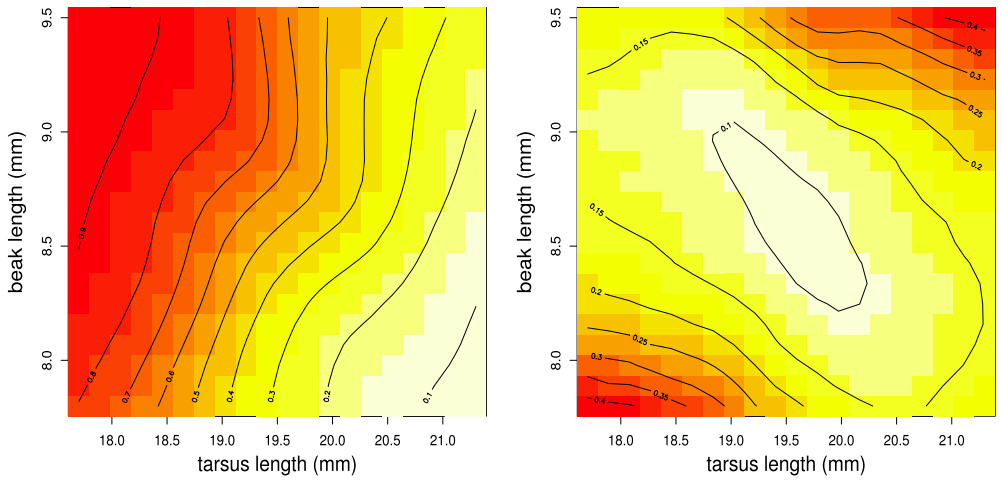

Figure 5: Song sparrows data. Posterior median surface (left panel) and interquartile range surface (right panel) for the probability of survival as a function of tarsus length and beak length.

mated regression curves, these results give strong evidence that directional selection is acting on tarsus length and beak length, favoring sparrows with long beaks and short tarsi.

The average gradient of the selection surface, weighted by the phenotype distribution, is given under our model by the vector

$$\left(\int \frac{\partial \Pr(y = 1 \mid x; G_N)}{\partial x_1} f(x; G_N) \mathrm{d}x, \int \frac{\partial \Pr(y = 1 \mid x; G_N)}{\partial x_2} f(x; G_N) \mathrm{d}x \right)^t.$$

Under a linear regression structure with a multivariate normal distribution for the phenotypic traits, the selection gradient is equivalent to the vector of linear regression slopes (Lande and Arnold, 1983). Janzen and Stern (1998) do not incorporate in their approach a distributional assumption for $f(x)$, and approximate the jth selection gradient by $n^{-1} \sum_{i=1}^{n} \partial \Pr(y = 1 \mid x)/\partial x_j \mid_{x=x_i}$. Our joint mixture modeling approach avoids the assumption of normality for the phenotypic distribution, as well as the need to estimate the integral by assuming the sample represents the population distribution. The integrand of the jth component of the selection gradient vector can be written as $\{\partial \Pr(y = 1, x; G_N)/\partial x_j\} - \{\Pr(y = 1 \mid x; G_N) \partial f(x; G_N)/\partial x_j\}$, for $j = 1, 2$. We omit the specific expressions for each of these two terms, but note that both are analytically available as a consequence of the mixture of normals representation for $f(z, x; G_N)$. Finally, the average gradient of the relative selection surface, also referred to as the directional selection gradient by Lande and Arnold (1983), is obtained by dividing each element of the selection gradient vector by mean absolute fitness. We obtained posterior mean estimates of -0.27 and 0.18, with corresponding 90% credible intervals of $(-0.40, -0.14)$ and $(0.06, 0.31)$, for the directional selection gradient associated with tarsus length and beak length, respectively.

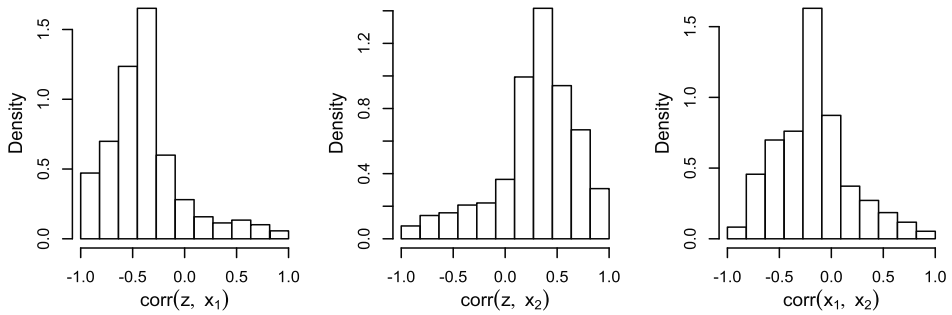

Figure 6: Song sparrows data. Posterior predictive samples for $\text{corr}(z, x_1)$ (left panel), $\text{corr}(z, x_2)$ (middle panel), and $\text{corr}(x_1, x_2)$ (right panel).

The presence of stabilizing or disruptive selection can be explored by considering the change in the phenotypic variance–covariance matrix due to selection, that is, the change from the pre-selection covariance matrix P, with elements $\int (x_1 - \bar{x}_1, x_2 - \bar{x}_2)^t (x_1 - \bar{x}_1, x_2 - \bar{x}_2) f(x) \mathrm{d}x$, to the post-selection covariance matrix P^*, with elements $\int (x_1 - \bar{x}_1^*, x_2 - \bar{x}_2^*)^t (x_1 - \bar{x}_1^*, x_2 - \bar{x}_2^*) f(x \mid y = 1) \mathrm{d}x$. The stabilizing selection differential matrix is given by $P^* - P + (\bar{x}_1^* - \bar{x}_1, \bar{x}_2^* - \bar{x}_2)^t (\bar{x}_1^* - \bar{x}_1, \bar{x}_2^* - \bar{x}_2)$ (Lande and Arnold, 1983), where negative values for a particular trait indicate the presence of stabilizing selection, while positive values indicate disruptive selection. The posterior mean for the matrix element corresponding to tarsus length is 0.038, that for beak length is -0.020, and the off-diagonal element has a posterior mean of -0.018. The 90% posterior credible intervals for each element of the matrix all include zero, indicating lack of significant evidence for stabilizing or disruptive selection acting on either trait.

Finally, as a means to check if a kernel with independent components for x and y would be adequate for this data example, we study in posterior predictive space the correlations between the latent response and the two traits. Denoting by Θ the vector comprising all model parameters, the joint posterior predictive distribution is given by $p(z, x \mid \text{data}) = \int \sum_{l=1}^{N} p_l \mathrm{N}_3(z, x; \mu_l, \Sigma_l) p(\Theta \mid \text{data}) \, \mathrm{d}\Theta$, which requires sampling one of $(\Sigma_1, \ldots, \Sigma_N)$ with probabilities $p_1, \ldots, p_N$ for each set of posterior samples. The correlations resulting from these posterior predictive draws for the kernel covariance matrix are plotted in Figure 6. These results suggest that it would be restrictive to force uncorrelated mixture kernel components, since the distribution of correlations for (z, x_1) is right-skewed and centered on negative values, while that for (z, x_2) is mainly focused on positive values and left-skewed, a pattern which is consistent with the shape of the estimated binary regression curves.

4 Discussion

We have presented a flexible method for estimating the regression relationship between binary responses and continuous covariates, which is built from a DP mixture model for the latent response–covariate distribution. Identifiability was established for the

parameters of the mixture kernel. In order to impose the restriction which is necessary for identifiability, the covariance matrix of the normal kernel was reparamaterized in such a way that allows for viewing only part of the matrix as random, while retaining the desirable features of conjugacy. Full conditional distributions were derived for the random elements of the covariance matrix, providing the key component of an efficient Markov chain Monte Carlo algorithm for posterior simulation. Two strategies for prior specification were discussed. The methodology was illustrated with two data examples that were chosen to indicate the practical utility of the modeling approach for problems in the environmental sciences and in population biology. In particular, estimation of fitness and natural selection functions is an active research area in evolutionary biology that can substantially benefit from Bayesian nonparametric joint modeling techniques, due to the need to model the fitness function flexibly as well as estimate the distribution of the (random) phenotypic traits associated with specific selection functionals.

As demonstrated with the two data examples, treating the covariates as random, for applications where this is appropriate, is beneficial in achieving flexible inference for regression functionals. However, if x is fixed by experimental design, then we do not recommend the joint approach as unnecessary resources may be spent on capturing the marginal distribution of x. Moreover, as observed by Wade et al. (2014), when x is very high dimensional, its likelihood may dominate the posterior and lead to an unnecessarily large number of mixture components in order to capture the conditional response distribution.

The proposed modeling approach relies on the choice of the multivariate normal distribution for the mixture kernel. This choice can accommodate essentially any type of continuous covariate, possibly through use of appropriate transformation. It can also handle ordinal categorical covariates x by incorporating in the model associated continuous variables, x_c, such that x arises from x_c through discretization. In particular, although in this case inferences were not affected, beak length in the data example of Section 3.2 was recorded only to the nearest tenth, and it could therefore be treated as a discrete covariate.

This work lays in place the foundations for a variety of extensions to ordinal regression problems involving data of different types. In particular, extensions of the modeling approach to incorporate ordinal and mixed ordinal-continuous responses follow naturally. In analogy with the binary setting, a univariate ordinal response y with K categories may be thought to arise as a discretized version of an underlying continuous response z, such that $y = k$ if and only if $\gamma_{k-1} < z \leq \gamma_k$, for $k = 2, \ldots, K-1$, and $y = 1$ or $y = K$ if and only if $z \leq \gamma_1$ or $z > \gamma_{K-1}$. A normal DP mixture model can again be used for (z, x). However, extending the argument in Kottas et al. (2005), it can be shown that if $K \geq 3$ all elements of the kernel covariance matrix are identifiable when the cut-off points, $\gamma_1, \ldots, \gamma_{K-1}$, are fixed. A key feature of the nonparametric mixture modeling framework is that we can obtain general inference with fixed cut-off points, resulting in a great advantage over parametric models, the implementation of which involves computationally challenging cut-off point estimation. In the case of multivariate ordinal regression, each response may be assumed to arise from its own underlying continuous response. Modeling these latent continuous responses jointly with the covariates in the

kernel sets the stage for flexible inference on the relationship between the multivariate ordinal response and the covariates, as well as among the ordinal responses. Finally, we can consider mixed ordinal-continuous responses, using a multivariate normal kernel for the latent responses, continuous responses, and covariates. We will report on these modeling extensions in a future manuscript.

Appendix A: Proofs of Lemmas 1 and 2

Proof of Lemma 1

Recall the kernel distribution in (3) for which we wish to prove that parameters $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ are identifiable, fixing $\Sigma^{zz} = 1$. Assume that

$$k(y, x; \mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = k(y, x; \mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx}). \tag{5}$$

If this implies $(\mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = (\mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx})$, then $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ are identifiable.

From (5), it must be the case that $\mathrm{N}_p(x; \mu_1^x, \Sigma_1^{xx}) = \mathrm{N}_p(x; \mu_2^x, \Sigma_2^{xx})$. This follows from summing each side of (5) over the two possible values of y. Because the mean vector and covariance matrix are identifiable for the multivariate normal likelihood, it can be concluded that $\mu_1^x = \mu_2^x$, and $\Sigma_1^{xx} = \Sigma_2^{xx}$. Now, after this simplification, each side of the equality in (5) consists of a Bernoulli distribution for $y \mid x$, and since y is either 0 or 1, the corresponding Bernoulli probabilities must be equal. Since Φ is a monotonically increasing function of its argument, the arguments of Φ are equal, that is,

$$\frac{\mu_1^z + \Sigma_1^{zx}(\Sigma^{xx})^{-1}(x - \mu^x)}{(1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^t)^{1/2}} = \frac{\mu_2^z + \Sigma_2^{zx}(\Sigma^{xx})^{-1}(x - \mu^x)}{(1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^t)^{1/2}}.$$

This can be written in the form $a^t x + b = 0$, and in order to hold true for all x, each element of vector a must be 0, and scalar b must be 0. The two equations $a = 0$ and $b = 0$ require

$$\frac{\Sigma_1^{zx}}{(1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^t)^{1/2}} = \frac{\Sigma_2^{zx}}{(1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^t)^{1/2}}, \tag{6}$$

$$\frac{\mu_1^z - \Sigma_1^{zx}(\Sigma^{xx})^{-1}\mu^x}{(1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^t)^{1/2}} = \frac{\mu_2^z - \Sigma_2^{zx}(\Sigma^{xx})^{-1}\mu^x}{(1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^t)^{1/2}}. \tag{7}$$

Using (6), (7) can be replaced by $\mu_1^z \Sigma_2^{zx} = \mu_2^z \Sigma_1^{zx}$. Writing these two equations component-wise, and letting Σ_{ji}^{zx} denote element i of the vector Σ_j^{zx}, results in two systems of p equations:

$$\frac{(\Sigma_{1i}^{zx})^2}{1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^t} = \frac{(\Sigma_{2i}^{zx})^2}{1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^t}, \quad i = 1, \ldots, p, \tag{8}$$

$$\mu_1^z \Sigma_{2i}^{zx} = \mu_2^z \Sigma_{1i}^{zx}, \quad i = 1, \ldots, p. \tag{9}$$

When $p = 1$ such that Σ^{zx} is a scalar, (8) becomes $|\Sigma_1^{zx}| = |\Sigma_2^{zx}|$, which has only the solution $\Sigma_1^{zx} = \Sigma_2^{zx}$, since $\Sigma_1^{zx} = -\Sigma_2^{zx}$ would violate (6). Then from (9) we conclude $\mu_1^z = \mu_2^z$.

In general, with p covariates, (8) can be written as

$$(\Sigma_{1i}^{zx})^2 - (\Sigma_{1i}^{zx})^2 \sum_{k=1}^{p}\sum_{j=1}^{p} \Sigma_{2j}^{zx}\Sigma_{2k}^{zx}(\Sigma^{xx})_{jk}^{-1} = (\Sigma_{2i}^{zx})^2 - (\Sigma_{2i}^{zx})^2 \sum_{k=1}^{p}\sum_{j=1}^{p} \Sigma_{1j}^{zx}\Sigma_{1k}^{zx}(\Sigma^{xx})_{jk}^{-1}$$

for $i = 1, \ldots, p$. Because (9) implies $\Sigma_{1l}^{zx}\Sigma_{2m}^{zx} = \Sigma_{1m}^{zx}\Sigma_{2l}^{zx}$ for any $l, m = 1, \ldots, p$, the equation reduces to $(\Sigma_{1i}^{zx})^2 = (\Sigma_{2i}^{zx})^2$. The constraint $\Sigma_{1l}^{zx}\Sigma_{2m}^{zx} = \Sigma_{1m}^{zx}\Sigma_{2l}^{zx}$ leaves only $\Sigma_1^{zx} = -\Sigma_2^{zx}$ and $\Sigma_1^{zx} = \Sigma_2^{zx}$ as possible solutions. The first can be eliminated as well, since this contradicts (6). This leaves as the only feasible solution $\Sigma_1^{zx} = \Sigma_2^{zx}$, which implies $\mu_1^z = \mu_2^z$ from (9).

It has been shown that if $k(y, x; \mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = k(y, x; \mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx})$, then this implies $(\mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = (\mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx})$. Therefore, applying directly the definition, the parameters $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ are identifiable in the kernel of the mixture.

Proof of Lemma 2

Consider $y = (y_1, \ldots, y_r)|\mu, \beta, \Delta \sim \mathrm{N}_r(\mu, \beta^{-1}\Delta(\beta^{-1})^t)$, such that the likelihood for β is proportional to $\exp\{-(y-\mu)^t\beta^t\Delta^{-1}\beta(y-\mu)\}$. First, focus on determining the likelihood for $\tilde{\beta}$, a vector of length $q = r(r-1)/2$. Write $\beta(y-\mu)$ as $M(1, \tilde{\beta}^t)^t$, for a matrix M, of dimension $r \times (q+1)$ which has row i containing i nonzero elements, the first being $(y_i - \mu_i)$, occurring in column 1, and the rest being $(y_1 - \mu_1), \ldots, (y_{i-1} - \mu_{i-1})$, occurring in columns $2+(i-1)(i-2)/2$ to $i+(i-1)(i-2)/2$. Then, the likelihood for $\tilde{\beta}$ can be written proportional to $\exp\{-(1, \tilde{\beta}^t)M^t\Delta^{-1}M(1, \tilde{\beta}^t)^t\}$. Let $C = M^t\Delta^{-1}M$. If there exists a symmetric, positive definite matrix T and vector d for which $(1, \tilde{\beta}^t)C(1, \tilde{\beta}^t)^t = \tilde{\beta}^tT\tilde{\beta} - 2\tilde{\beta}^tTd + R$, where R is a constant that does not depend on $\tilde{\beta}$, then the likelihood for $\tilde{\beta}$ corresponds to a normal distribution with mean vector d and covariance matrix T^{-1}. The left side of the above equation is $C_{11} + 2\sum_{j=2}^{q+1}\tilde{\beta}_{j-1}C_{1j} + \sum_{j=2}^{q+1}\sum_{i=2}^{q+1}\tilde{\beta}_{j-1}\tilde{\beta}_{i-1}C_{ij}$, and the last of these terms is just $\tilde{\beta}^tC_{q\times q}\tilde{\beta}$, where $C_{q\times q}$ denotes the $q \times q$ submatrix of C obtained by deleting the first row and column of C. Therefore, with $T = C_{q\times q}$, we seek d such that $-\tilde{\beta}^tTd = \sum_{j=2}^{q+1}\tilde{\beta}_{j-1}C_{1j}$. Equating the coefficient associated with $\tilde{\beta}_i$, $i = 1, \ldots, q$, on each side of the equation results in a system of q equations:

$$-\sum_{j=1}^{q} d_jT_{i-1,j} = C_{1i}, \quad i = 2, \ldots, q+1. \tag{10}$$

As explained in Section 2.2, T is a block diagonal matrix which can be constructed from square matrices $T^1, \ldots, T^{r-1}$, of dimensions $1, \ldots, r-1$, where

$$T_{mn}^j = (y_m - \mu_m)(y_n - \mu_n)/\delta_{j+1}, \quad m = 1, \ldots, j, \quad n = 1, \ldots, j. \tag{11}$$

The symmetry of T follows from the symmetry of C, but it remains to be shown that T is positive definite. For a non-zero vector v, we must have $v^tTv > 0$. When $r = 2$,

v^tTv becomes $v_1^2(y_1-\mu_1)^2/\delta_2$. When $r=3$, v^tTv is the sum of the result for $r=2$ and the term $(v_2(y_1-\mu_1)+v_3(y_2-\mu_2))^2/\delta_3$. For $r=4$, the term $(v_4(y_1-\mu_1)+v_5(y_2-\mu_2)+v_6(y_3-\mu_3))^2/\delta_4$ is added to the result for $r=3$. In general, a term of the form $(v_{q-r+2}(y_1-\mu_1)+\cdots+v_q(y_{r-1}-\mu_{r-1}))^2/\delta_r$ is added in going from $r-1$ to r dimensions. Clearly, T is positive semidefinite. However, to have $v^tTv>0$, and all elements of T strictly positive, it must be the case that $y_i\neq\mu_i$, for $i=1,\ldots,r-1$, which holds true with probability 1, since μ is a continuous random vector.

We now derive the form of the mean vector d. Because T is sparse, the system of q equations (10) can be divided into $r-1$ sets of equations, where set j consists of j equations with j unknowns, $d_{1+j(j-1)/2},\ldots,d_{j(j+1)/2}$. Let the index $1+j(j-1)/2$ be denoted by (1) and let the index $j(j+1)/2$ be denoted by (j). Set the first $j-1$ of these elements equal to 0, so that $d_{1+j(j-1)/2}=\cdots=d_{j(j+1)/2-1}=0$. Then the j equations become

$$-d_{(j)}T_{(1),(j)}=C_{1,(1)+1},\ldots,-d_{(j)}T_{(j),(j)}=C_{1,(j)+1}. \tag{12}$$

The solution $d_{(j)}=-(y_{j+1}-\mu_{j+1})/(y_j-\mu_j)$ satisfies these j equalities (12), since the elements $C_{1,(1)+1},\ldots,C_{1,(j)+1}$ are $(y_1-\mu_1)(y_{j+1}-\mu_{j+1})/\delta_{j+1},\ldots,(y_j-\mu_j)(y_{j+1}-\mu_{j+1})/\delta_{j+1}$, and the elements $T_{(1),(j)},\ldots,T_{(j),(j)}$ are $(y_1-\mu_1)(y_j-\mu_j)/\delta_{j+1},\ldots,(y_j-\mu_j)(y_j-\mu_j)/\delta_{j+1}$, as given in (11), so that

$$-C_{1,(1)+1}/T_{(1),(j)}=\cdots=-C_{1,(j)+1}/T_{(j),(j)}=-(y_{j+1}-\mu_{j+1})/(y_j-\mu_j).$$

With n data vectors, $(y_{i,1},\ldots,y_{i,r})$, for $i=1,\ldots,n$, the likelihood for $\tilde{\beta}$ is proportional to a normal with mean $(\sum_{i=1}^n T_i)^{-1}(\sum_{i=1}^n T_i d_i)$, and covariance matrix $(\sum_{i=1}^n T_i)^{-1}$, where T_i and d_i are computed using the ith observation. When combined with a normal prior for $\tilde{\beta}$, the full conditional is also normal.

Next, consider the likelihood for the δ_k, which up to the proportionality constant is given by $\prod_{k=1}^r \delta_k^{-1/2}\exp\{-\mathrm{tr}(\beta^t\Delta^{-1}\beta(y-\mu)(y-\mu)^t)/2\}$. By properties of trace, $\mathrm{tr}(\beta^t\Delta^{-1}\beta(y-\mu)(y-\mu)^t)=\mathrm{tr}(\beta(y-\mu)(y-\mu)^t\beta^t\Delta^{-1})$. Let $A=\beta(y-\mu)(y-\mu)^t\beta^t$. Since Δ is diagonal with δ on the diagonal, the likelihood for each δ_k is proportional to $\delta_k^{-1/2}\exp\{-A_{kk}/(2\delta_k)\}$. The diagonal elements of A are the squares of $\beta(y-\mu)$, which are $A_{kk}=\{(y_k-\mu_k)+\sum_{j<k}\beta_{kj}(y_j-\mu_j)\}^2$. Then, with n data vectors, $(y_{i,1},\ldots,y_{i,r})$, $i=1,\ldots,n$, the likelihood for δ_k, $k=1,\ldots,r$, is proportional to an inverse-gamma with shape parameter $(n/2)-1$ and scale parameter $0.5\sum_{i=1}^n\{(y_{i,k}-\mu_k)+\sum_{j<k}\beta_{kj}(y_{i,j}-\mu_j)\}^2$. When combined with an inverse-gamma prior, this results in a posterior full conditional distribution which is inverse-gamma.

Appendix B: Distributions implied by the inverse-Wishart

Assume $\Sigma\sim \mathrm{IW}_r(v,T)$, with $r=p+1$, and partition Σ into blocks, Σ_{11}, Σ_{12}, Σ_{21}, and Σ_{22}, of dimensions $q\times q$, $q\times(r-q)$, $(r-q)\times q$, and $(r-q)\times(r-q)$, respectively. Moreover, consider the corresponding partition for matrix T. Then, applying Propositions 8.7 and 8.8 of Eaton (2007), we obtain:

(a) $\Sigma_{11}\sim \mathrm{IW}_q(v-(r-q),T_{11})$.

(b) $\Sigma_{22\cdot 1} \sim \text{IW}_{r-q}(v, T_{22\cdot 1})$, where $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ and $T_{22\cdot 1} = T_{22} - T_{21}T_{11}^{-1}T_{12}$.

(c) $\Sigma_{11}^{-1}\Sigma_{12}|\Sigma_{22\cdot 1}^{-1} \sim \text{MN}_{q,r-q}(T_{11}^{-1}T_{12}, T_{11}^{-1}, \Sigma_{22\cdot 1})$, that is, a matrix normal distribution; thus, conditionally on $\Sigma_{22\cdot 1}$, $\text{vec}(\Sigma_{11}^{-1}\Sigma_{12}) \sim \text{N}_{q(r-q)}(\text{vec}(T_{11}^{-1}T_{12}), T_{11}^{-1} \otimes \Sigma_{22\cdot 1})$.

We now assume T is diagonal, with elements $(T_1, \ldots, T_{p+1})$, as this is the case relevant to our prior specification approach. Let $T^i = \text{diag}(T_1, \ldots, T_i)$. Applying result (b) with $q = p$, we obtain $\delta_{p+1} \sim \text{IG}(0.5v, 0.5T_{p+1})$. This uses the fact that $\Sigma_{22\cdot 1} = \delta_{p+1}$ as a consequence of the (β, Δ) parameterization, and the simplification of $T_{22\cdot 1}$ to $T_{22} = T_{p+1}$ when T is diagonal. Applying result (a) with $q = p$, we obtain the marginal distribution of the upper left p dimensional block of the covariance matrix Σ, which is $\Sigma_{1:p,1:p} \sim \text{IW}_p(v - 1, T^p)$. Next, using result (b) for matrix $\Sigma_{1:p,1:p}$ with $q = p - 1$, we have $\delta_p \sim \text{IG}(0.5(v - 1), 0.5T_p)$, since $(\Sigma_{1:p,1:p})_{22\cdot 1} = \delta_p$. Analogously, applying results (a) and (b) in succession, we obtain $\delta_i \sim \text{IG}(0.5(v+i-(p+1)), 0.5T_i)$, for $i = 2, \ldots, p+1$.

For each $i = 2, \ldots, p+1$, result (a) yields an $\text{IW}_i(v + i - (p+1), T^i)$ distribution for $\Sigma_{1:i,1:i}$, that is, for the upper left block of Σ of dimension i. Then, applying result (c) to $\Sigma_{1:i,1:i}$ with $q = i-1$, we obtain $(-\beta_{i,1}, \ldots, -\beta_{i,i-1})^t|\delta_i \sim \text{N}_{i-1}((0, \ldots, 0)^t, \delta_i(T^{i-1})^{-1})$, for $i = 2, \ldots, p+1$. This uses the fact that $(T^i)_{12} = (0, \ldots, 0)^t$, $\text{vec}((\Sigma_{1:i,1:i})_{11}^{-1}(\Sigma_{1:i,1:i})_{12})$ $= (-\beta_{i,1}, \ldots, -\beta_{i,i-1})^t$, and $(\Sigma_{1:i,1:i})_{22\cdot 1} = \delta_i$.

Appendix C: Model comparison criterion

The predictive loss measure used for model comparison in Section 3.1 requires for each model m the posterior predictive mean, $\text{E}^{(m)}(y_{new,i}|\text{data})$, and posterior predictive variance, $\text{var}^{(m)}(y_{new,i}|\text{data})$, for replicated response $y_{new,i}$ with associated covariate vector x_i.

Denote generically by Θ the full parameter vector for either the product-kernel model or for the more general binary regression model developed in Section 2. For the former model, $\text{E}(y|x_i, \text{data}) = \{p(x_i|\text{data})\}^{-1} \int \sum_{l=1}^{N} p_l \text{N}_p(x_i; \mu_l^x, \Sigma_l^{xx}) \Phi(\mu_l^z)\, p(\Theta|\text{data})\text{d}\Theta$, with $p(x_i|\text{data}) = \int \sum_{l=1}^{N} p_l \text{N}_p(x_i; \mu_l^x, \Sigma_l^{xx})\, p(\Theta|\text{data})\text{d}\Theta$, and $\text{E}(y^2|x_i, \text{data})$ also has the same form. Under the proposed model, $\text{E}(y|x_i, \text{data})$ is given by

$$\{p(x_i|\text{data})\}^{-1} \int \sum_{l=1}^{N} p_l \text{N}_p(x_i; \mu_l^x, \Sigma_l^{xx}) \Phi\left(\frac{\mu_l^z + \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(x_i - \mu_l^x)}{(\Sigma_l^{zz} - \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\Sigma_l^{zx})^t)^{1/2}} \right) p(\Theta|\text{data})\text{d}\Theta$$

where $p(x_i|\text{data}) = \int \sum_{l=1}^{N} p_l \text{N}_p(x_i; \mu_l^x, \Sigma_l^{xx})\, p(\Theta|\text{data})\text{d}\Theta$, and, again, $\text{E}(y|x_i, \text{data}) = \text{E}(y^2|x_i, \text{data})$. Hence, under both models, straightforward Monte Carlo integration using the posterior samples for model parameters yields estimates for the required posterior predictive means and variances.

References

Albert, J. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88: 669–679. MR1224394. 822, 826

Antoniano-Villalobos, I., Wade, S., and Walker, S. G. (2014). "A Bayesian nonparametric regression model with normalized weights: A study of hippocampal atrophy in Alzheimer's disease." *Journal of the American Statistical Association*, 109: 477–490. MR3223726. doi: http://dx.doi.org/10.1080/01621459.2013.879061. 831

Basu, S. and Mukhopadhyay, S. (2000). "Bayesian analysis of binary regression using symmetric and asymmetric links." *Sankhya: The Indian Journal of Statistics, Series B*, 62: 372–387. MR1834162. 822

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, Massachusetts: Addison-Wesley. MR0418321. 829

Canale, A. and Dunson, D. (2011). "Bayesian kernel mixtures for counts." *Journal of the American Statistical Association*, 106: 1528–1539. MR2896854. doi: http://dx.doi.org/10.1198/jasa.2011.tm10552. 831

Choudhuri, N., Ghosal, S., and Roy, A. (2007). "Nonparametric binary regression using a Gaussian process prior." *Statistical Methodology*, 4: 227–243. MR2368147. doi: http://dx.doi.org/10.1016/j.stamet.2006.07.003. 822

Daniels, M. and Pourahmadi, M. (2002). "Bayesian analysis of covariance matrices and dynamic models for longitudinal data." *Biometrika*, 89: 553–566. MR1929162. doi: http://dx.doi.org/10.1093/biomet/89.3.553. 825

DeIorio, M., Müller, P., Rosner, G., and MacEachern, S. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99: 205–215. MR2054299. doi: http://dx.doi.org/10.1198/016214504000000205. 831

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Chichester: Wiley. MR1962778. 821

Di Lucca, M. A., Guglielmi, A., Müller, P., and Quintana, F. A. (2013). "A simple class of Bayesian nonparametric autoregression models." *Bayesian Analysis*, 8: 63–88. 824, 831

Dunson, D. and Park, J. (2008). "Kernel stick-breaking processes." *Biometrika*, 95: 307–323. MR2521586. doi: http://dx.doi.org/10.1093/biomet/asn012. 831

Dunson, D. B. and Bhattacharya, A. (2011). "Nonparametric Bayes regression and classification through mixtures of product kernels." In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9, Proceedings of the Ninth Valencia International Meeting*, 145–164. Oxford University Press. MR3204005. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0005. 822, 830, 834

Eaton, M. (2007). *Multivariate Statistics: A Vector Space Approach*. Beachwood, Ohio: Institute of Mathematical Statistics. MR2431769. 829

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1: 209–230. MR0350949. 822

Follmann, D. and Lamberdt, E. (1989). "Generalizing logistic regression by nonparametric modelling." *Journal of the American Statistical Association*, 84: 295–300. 822

Gelfand, A. and Ghosh, S. (1998). "Model choice: A minimum posterior predictive loss approach." *Biometrika*, 85: 1–11. MR1627258. doi: http://dx.doi.org/10.1093/biomet/85.1.1. 835

Gelfand, A. E. and Kottas, A. (2002). "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 11: 289–305. MR1938136. doi: http://dx.doi.org/10.1198/106186002760180518. 828

Gelfand, A. E., Kottas, A., and MacEachern, S. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100: 1021–1035. MR2201028. doi: http://dx.doi.org/10.1198/016214504000002078. 831

Griffin, J. and Holmes, C. (2010). "Computational issues arising in Bayesian nonparametric hierarchical models." In: Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds.), *Bayesian Nonparametrics*, 208–222. Cambridge University Press. MR2730664. 828

Griffin, J. and Steel, M. (2006). "Order-based dependent Dirichlet processes." *Journal of the American Statistical Association*, 101: 179–194. MR2268037. doi: http://dx.doi.org/10.1198/016214505000000727. 831

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011). "Dirichlet process mixtures of generalized linear models." *Journal of Machine Learning Research*, 1: 1–33. MR2819022. 822, 830

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall. MR1082147. 821

Hobert, J. and Casella, G. (1996). "The effect of improper priors on Gibbs sampling in hierarchical linear mixed models." *Journal of the American Statistical Association*, 61: 1461–1473. MR1439086. doi: http://dx.doi.org/10.2307/2291572. 824

Ishwaran, H. and James, L. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–173. MR1952729. doi: http://dx.doi.org/10.1198/016214501750332758. 825, 828

Ishwaran, H. and Zarepour, M. (2000). "Markov Chain Monte Carlo in approximate Dirichlet and Beta two-parameter process Hierarchical Models." *Biometrika*, 87: 371–390. MR1782485. doi: http://dx.doi.org/10.1093/biomet/87.2.371. 825, 826

Ishwaran, H. and Zarepour, M. (2002). "Exact and approximate sum representations for the Dirichlet process." *The Canadian Journal of Statistics*, 30: 269–283. MR1926065. doi: http://dx.doi.org/10.2307/3315951. 828

Janzen, F. and Stern, H. (1998). "Logistic regression for empirical studies of multivariate selection." *Evolution*, 52: 1564–1571. 835, 837

Kalli, M., Griffin, J., and Walker, S. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: http://dx.doi.org/10.1007/s11222-009-9150-y. 828

Koop, G. (2003). *Bayesian Econometrics*. Chichester: John Wiley and Sons. 824

Kottas, A., Müller, P., and Quintana, F. (2005). "Nonparametric Bayesian modeling for multivariate ordinal data." *Journal of Computational and Graphical Statistics*, 14: 610–625. MR2170204. doi: http://dx.doi.org/10.1198/106186005X63185. 839

Lande, R. and Arnold, S. (1983). "The measurement of selection on correlated characters." *Evolution*, 37: 1210–1226. 835, 837

MacEachern, S. (2000). "Dependent Dirichlet processes." Technical report, The Ohio State University Department of Statistics. 831

McCulloch, P., Polson, N., and Rossi, P. (2000). "A Bayesian analysis of the multinomial probit model with fully identified parameters." *Journal of Econometrics*, 99: 173–193. 824

Mukhopadyay, S. and Gelfand, A. (1997). "Dirichlet process mixed generalized linear models." *Journal of the American Statistical Association*, 92: 633–639. MR1467854. doi: http://dx.doi.org/10.2307/2965710. 822

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures." *Biometrika*, 83: 67–79. MR1399156. doi: http://dx.doi.org/10.1093/biomet/83.1.67. 822, 830

Müller, P. and Mitra, R. (2013). "Bayesian nonparametric inference – why and how (with discussion)." *Bayesian Analysis*, 8: 269–360. MR3066939. doi: http://dx.doi.org/10.1214/13-BA811. 831

Müller, P. and Quintana, F. (2010). "Random partition models with regression on covariates." *Journal of Statistical Planning and Inference*, 140: 2801–2808. MR2651966. doi: http://dx.doi.org/10.1016/j.jspi.2010.03.002. 830

Neal, R. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. MR1823804. doi: http://dx.doi.org/10.2307/1390653. 828

Newton, M., Czado, C., and Chappell, R. (1996). "Bayesian inference for semiparametric binary regression." *Journal of the American Statistical Association*, 91: 142–153. MR1394068. doi: http://dx.doi.org/10.2307/2291390. 822

Park, J.-H. and Dunson, D. B. (2010). "Bayesian generalized product partition model." *Statistica Sinica*, 20: 1203–1226. MR2730180. 830

Rodriguez, A. and Dunson, D. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian Analysis*, 6: 145–178. MR2781811. doi: http://dx.doi.org/10.1214/11-BA605. 831

Schluter, D. (1988). "Estimating the form of natural selection on a quantitative trait." *Evolution*, 42: 849–861. 835

Schluter, D. and Smith, J. (1986). "Natural selection on beak and body size in the song sparrow." *International Journal of Organic Evolution*, 40: 221–231. 835

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. MR1309433. 823

Shahbaba, B. and Neal, R. (2009). "Nonlinear modeling using Dirichlet process mixtures." *Journal of Machine Learning Research*, 10: 1829–1850. MR2540778. 822, 830

Taddy, M. (2010). "Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking the intensity of violent crime." *Journal of the American Statistical Association*, 105: 1403–1417. MR2796559. doi: http://dx.doi.org/10.1198/jasa.2010.ap09655. 831

Taddy, M. and Kottas, A. (2010). "A Bayesian nonparametric approach to inference for quantile regression." *Journal of Business and Economic Statistics*, 28: 357–369. MR2723605. doi: http://dx.doi.org/10.1198/jbes.2009.07331. 830

Taddy, M. and Kottas, A. (2012). "Mixture modeling for marked Poisson processes." *Bayesian Analysis*, 7: 335–362. MR2934954. doi: http://dx.doi.org/10.1214/12-BA711. 830

The Environmental Protection Agency (2014). "Policy Assessment for the Review of the Ozone National Ambient Air Quality Standards." Available online at: http://www.epa.gov/ttn/naaqs/standards/ozone/data/20140829pa.pdf. 832

Trippa, L. and Muliere, P. (2009). "Bayesian nonparametric binary regression via random tesselations." *Statistics and Probability Letters*, 79: 2273–2282. MR2591984. doi: http://dx.doi.org/10.1016/j.spl.2009.07.026. 822

Wade, S., Dunson, D. B., Petrone, S., and Trippa, L. (2014). "Improving prediction from Dirichlet process mixtures via enrichment." *Journal of Machine Learning Research*, 15: 1041–1071. MR3195338. 830, 839

Walker, S. and Mallick, B. (1997). "Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 59: 845–860. MR1483219. doi: http://dx.doi.org/10.1111/1467-9868.00101. 822

Webb, E. and Forster, J. (2008). "Bayesian model determination for multivariate ordinal and binary data." *Computational Statistics and Data Analysis*, 52: 2632–2649. MR2419531. doi: http://dx.doi.org/10.1016/j.csda.2007.09.008. 825

Wood, S. and Kohn, R. (1998). "A Bayesian approach to robust binary nonparametric regression." *Journal of the American Statistical Association*, 93: 203–213. 821

Acknowledgments

The authors wish to thank an Editor and two reviewers for comments that improved the presentation of the material in the paper. This research is part of the Ph.D. dissertation of Maria DeYoreo, completed at University of California, Santa Cruz, and was supported in part by the National Science Foundation under award DMS 1310438.