A Fully Nonparametric Modelling Approach to Binary Regression

Maria DeYoreo and Athanasios Kottas [*]

April 16, 2014

Abstract

We propose a general nonparametric Bayesian framework for binary regression, which is built from modelling 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 modelling strategy provides substantial simplifications in implementation of Markov chain Monte Carlo posterior simulation. We illustrate the utility of the modelling approach with two data examples, and discuss extensions to incorporate multivariate ordinal responses, as well as mixed ordinal-continuous responses.

KEY WORDS: 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 modelling perspective, interest centers on determining the regression relationship between the response and covariates. Standard approaches to this problem in both classical and Bayesian settings involve potentially restrictive distributional assumptions as well as those of linearity in relating the response to covariates.

[*] M. DeYoreo (mdeyoreo@ams.ucsc.edu) is Ph.D. student, and A. Kottas (thanos@ams.ucsc.edu) is Professor, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, USA. This research was supported in part by the National Science Foundation under award DMS 1310438.

In an effort to create more flexible models to combat overdispersion and asymmetry, several Bayesian semiparametric approaches to binary regression 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 & Mukhopadhyay, 2000), or linearity, for instance, by viewing the intercept of the linear predictor as arising from an unknown distribution (Follmann & Lambert, 1989; Mukhopadhyay & Gelfand, 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 & Muliere (2009) assumed each binary response to arise from a random colored tesselation, and placed a Dirichlet process prior (Ferguson, 1973) on the space of colored tesselations.

Shahbaba & Neal (2009), Dunson & Bhattacharya (2011), and Hannah et al. (2010), have proposed nonparametric solutions to the regression problem with categorical responses, building off the work of Müller et al. (1996), which modelled the joint distribution of continuous responses y and covariates x with a Dirichlet process mixture of normal distributions, inducing a flexible model for $E(y|x)$. 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, developing a flexible model for fully nonparametric binary regression. The foundation of the proposed methodology is different from the existing nonparametric modelling approaches. We elaborate further in Section 4, but here note that the key distinguishing feature of the proposed model involves the introduction of latent continuous responses, in similar spirit to parametric probit models; see, for instance, Albert & Chib (1993). Let $\{(y_i, x_i) : i = 1, ..., n\}$ denote the data, where each observation consists of a binary response y_i along with a vector of covariates, $x_i = (x_{i1}, ..., x_{ip})$. 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 Dirichlet process mixture of multivariate normal distributions, which induces a flexible model for the regression relationship, $\mathrm{pr}(y = 1|x)$. In addition to providing a general modelling 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 posterior simulation while enabling the use of priors more flexible than the inverse-Wishart distribution.

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. In Section 3, the methodology is applied to problems from environmetrics and evolutionary biology, using two data sets from the literature for illustration. Section 4 contains further discussion to place our contribution within the existing literature, and to indicate 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 modelling approach

Focusing on p continuous covariates, $x = (x_1, ..., x_p)$, 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 Dirichlet process is then used as a prior for the random mixing distribution G, to create a mixture model of the form: $f(z, x; G) = \int \mathrm{N}_{p+1}(z, x; \mu, \Sigma)\mathrm{d}G(\mu, \Sigma)$, $G \mid \alpha, \psi \sim \mathrm{DP}(\alpha, G_0(\cdot; \psi))$, where α is the Dirichlet process precision parameter, and ψ the parameters of the Dirichlet process centering distribution.

According to the Dirichlet process 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 $\mathrm{beta}(1, \alpha)$, $m = 1, 2, ...$, and define $p_1 = \zeta_1$, and $p_l = \zeta_l \prod_{r=1}^{l-1}(1 - \zeta_r)$, for $l = 2, 3, ...$; moreover, $\{\zeta_m : m = 1, 2, \dots\}$ and $\{\nu_l : l = 1, 2, \dots\}$ 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

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

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

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

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

which have the probit form with component-specific intercept and slope parameters.

The dependence structure of the mixture kernel in $f(z, x; G)$ is key to obtaining general inference for the implied binary regression function. However, is it possible to estimate all elements of the covariance matrix Σ? Recall the probit regression setting, which contains a similar binary and latent response relationship; in particular, the latent responses arise from normally distributed random variables with standard deviation σ fixed at 1, $z \mid \beta \sim \mathrm{N}(x^t\beta, 1)$. The reason for this restriction can be seen from $\mathrm{pr}(y = 1|\beta, \sigma) = \Phi(x^t\beta/\sigma)$, which implies that different combinations of σ and β result in the same value for the probability of positive response. In addition, the scale of the unobserved continuous responses is a quantity that we should not expect the binary observed responses to be able to inform. Standard probit regression as well as intuition suggests that some restrictions are necessary for identifiability in the kernel of the mixture.

Here, we employ the standard definition of likelihood identifiability, such that a parameter θ for a family of distributions $\{f(x|\theta) : \theta \in \Theta\}$ is identifiable if distinct values of θ correspond to distinct probability density functions, that is, if $\theta \neq \theta'$, then $f(x|\theta)$ is not the same function of x as $f(x|\theta')$. 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 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, due to the single degree of freedom parameter in the inverse-Wishart distribution, which does not allow for one element of Σ to be fixed while freely estimating the rest of the matrix. This problem is overcome by aid of a square-root-free Cholesky decomposition of Σ. This decomposition is useful for modelling longitudinal data (Daniels & Pourahmadi, 2002), as well as specifying conditional independence assumptions for the elements of a normal random vector (Webb & Forster, 2008). Let β be a unit lower triangular matrix, and let Δ be a diagonal matrix with positive elements, $(\delta_1, ..., \delta_{p+1})$, 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, $(\delta_2, ..., \delta_{p+1})$. 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, 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 $\delta_2, ..., \delta_{p+1}$ result in full conditional distributions which are multivariate normal and inverse-gamma, respectively. Therefore, G_0 comprises independent components for $\mu, \tilde{\beta}$, and $\delta_2, ..., \delta_{p+1}$, 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

In order to simulate from the full posterior distribution, we utilize the blocked Gibbs sampler (Ishwaran & Zarepour, 2000; Ishwaran & James, 2001). As a consequence of the constructive definition of the Dirichlet process, any distribution it generates can be represented as a countable mixture of point masses. This definition motivates the blocked Gibbs sampler, as it is based on a finite truncation approximation to G. 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 original Dirichlet process definition, whereas $p_N = 1 - \sum_{l=1}^{N-1} p_l$. Introducing configuration variables $L = (L_1, ..., L_n)$, each taking values in $\{1, ..., N\}$, the hierarchical version of the Dirichlet process mixture model for the data given the latent continuous responses, $z = (z_1, ..., z_n)$, becomes

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

where $W = (W_1, \ldots, W_N)$, and the implied prior for the vector $p = (p_1, ..., p_N)$ is a generalized Dirichlet distribution (Connor & Mosimann, 1969). 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 $i = 2, ..., p+1$. 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 modelling approach is that simulation from the full posterior distribution, $p(W, L, p, \psi, \alpha, z|\mathrm{data})$, 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 & Zarepour (2000). Each L_i, $i = 1, ..., n$, is sampled from a discrete distribution on $\{1, ..., N\}$, with probabilities proportional to $p_l \mathrm{N}_{p+1}(z_i, x_i; \mu_l, \Sigma_l)$, for $l = 1, ..., N$. 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 $z_i \leq 0$ if $y_i = 0$.

Letting $\{L^*_j, j = 1, ..., n^*\}$ be the vector of distinct values of L, the full conditional distribution for W_l is proportional to $G_0(W_l|\psi)\prod_{j=1}^{n^*}\prod_{\{i:L_i=L^*_j\}}\mathrm{N}_{p+1}(z_i, x_i; \mu_{L^*_j}, \beta^{-1}_{L^*_j}\Delta_{L^*_j}(\beta^{-1}_{L^*_j})^t)$. If $l \notin \{L^*_j : j = 1, ..., n^*\}$, then $W_l \sim G_0(\cdot|\psi)$. If $l \in \{L^*_j : j = 1, ..., n^*\}$, then the full conditional distribution for each element of $W_l = (\mu_l, \tilde{\beta}_l, \delta_{2,l}, ..., \delta_{p+1,l})$ 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 $j = 1, ..., n^*$, the full conditional for μ_l is multivariate normal with mean vector $(V^{-1} + M_l\Sigma^{-1}_l)^{-1}(V^{-1}m + \Sigma^{-1}_l\sum_{\{i:L_i=l\}}(z_i, x_i)^t)$ and covariance matrix $(V^{-1} + M_l\Sigma^{-1}_l)^{-1}$, 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 $i = 2, ..., p + 1$. 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 $k = 1, ..., r-1$. Let T_i be a block diagonal matrix of dimension $q \times q$ with $r - 1$ blocks, which can be constructed from square matrices $T^1_i, ..., T^{r-1}_i$ of dimensions $1, ..., r - 1$. Matrix T^j_i 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:*

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

with a multivariate normal prior for μ, independent inverse-gamma priors on the diagonal elements of Δ, $\delta_k \sim \mathrm{IG}(\nu_k, s_k)$, $k = 1, ..., r$, 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,

$k = 1, ..., r$, 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 $-(y_{i,2} - \mu_2)/(y_{i,1} - \mu_1), ..., -(y_{i,r} - \mu_r)/(y_{i,r-1} - \mu_{r-1})$, and the (m, n)-th element of matrix T_i^j, for $j = 1, ..., r-1$, is given by $T_{i,mn}^{j} = (y_{i,m} - \mu_m)(y_{i,n} - \mu_n)/\delta_{j+1}$, for $m = 1, ..., j, n = 1, ..., j$.

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 $j = 1, ..., p$. Then, when $l = L_j^*$, for $j = 1, ..., n^*$, 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 $-(y_{i,2}^* - \mu_{2,l})/(y_{i,1}^* - \mu_{1,l}), ..., -(y_{i,p+1}^* - \mu_{p+1,l})/(y_{i,p}^* - \mu_{p,l})$, and for $j = 1, ..., p$, the matrix T_i^j contains elements $T_{i,mn}^{j} = (y_{i,m}^* - \mu_{m,l})(y_{i,n}^* - \mu_{n,l})/\delta_{j+1,l}$, $m = 1, ..., j, n = 1, ..., j$.

The mixing distribution $G_N \equiv (p, W)$ is imputed as a component of the posterior simulation algorithm, enabling full inference for any functional of $f(y, x; G_N)$. The binary regression functional is the main quantity of interest, and is estimated as $\text{pr}(y = 1|x; G_N) = \text{pr}(y = 1, x; G_N)/f(x; G_N)$, where $f(x; G_N) = \sum_{l=1}^{N} p_l \text{N}_p(x; \mu_l^x, \Sigma_l^{xx})$, and $\text{pr}(y = 1, x; G_N) = \sum_{l=1}^{N} p_l \text{N}_p(x; \mu_l^x, \Sigma_l^{xx})\pi_l(x)$, with $\pi_l(x)$ given in (2). Therefore, full inference for $\text{pr}(y = 1|x_0; G_N)$ can be readily obtained for any covariate value x_0, providing a point estimate along with uncertainty quantification for the binary regression function.

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 (Taddy & Kottas, 2010). 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 $\text{E}((z, x)^t) = a_m$, and $\text{cov}((z, x)^t) = \text{E}(\Sigma) + B_m + (a_V - p - 2)^{-1} B_V$. We therefore set $a_m = (0, c^x)$, and let $B_m = 0.5\text{diag}(1, (r_1^x/4)^2, ..., (r_p^x/4)^2)$, using c_j^x and $(r_j^x/4)^2$ as proxies for the marginal mean and variance of x_j, for $j = 1, ..., p$. 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 & 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 $\mathrm{E}(\Sigma) = T = \mathrm{diag}(T_1, \ldots, T_{p+1}) = \mathrm{diag}(1, (r_1^x/4)^2, ..., (r_p^x/4)^2)$. Then, as shown in Appendix B, the distributions implied on δ_i, for $i = 2, \ldots, p + 1$, are $\mathrm{IG}(0.5(v + i - (p + 1)), 0.5T_i)$. Hence, we let $\nu_i = 0.5(v + i - (p + 1))$, and $\mathrm{E}(s_i) = 0.5T_i$; 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 $\mathrm{IW}_{p+1}(v, T)$ distribution implies a normal distribution for the i-th 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 $\mathrm{BD}(S_1, \ldots, S_p)$, which denotes a block diagonal matrix with elements $S_i = \delta_{i+1}\mathrm{diag}(T_1^{-1}, \ldots, T_i^{-1})$, for $i = 1, ..., p$. Now, after marginalizing out θ, the G_0 prior component for $\tilde{\beta}$ becomes $\mathrm{N}_q(a_\theta, B_\theta + C)$. 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 $B_\theta + \mathrm{E}(C) = \mathrm{BD}(\hat{S}_1, \ldots, \hat{S}_p)$, 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 $\mathrm{E}(C)$ can be specified to be equal to each other or assigned different portions of $\mathrm{BD}(\hat{S}_1, \ldots, \hat{S}_p)$.

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.

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. Therefore, we can model the probability of an ozone exceedance as a function of wind speed, temperature, and radiation, using the Dirichlet process mixture binary regression model. This example illustrates a practically relevant setting in which a binary response may arise as a discretized version of a continuous response. In addition, the modelling approach is evidently appropriate here, since it is natural to estimate conditional relationships between the four environmental variables through modelling the stochastic mechanism for their joint distribution.

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 $\mathrm{pr}(y = 1|x_i)$ that were, for each of the three random covariates, 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.

For this illustrative data example, the continuous ozone concentration responses are also available. We can therefore compare the binary regression model inferences for $\mathrm{pr}(y = 1|x_i)$ with the ones for $\mathrm{pr}(z > 70|x_i)$, under the corresponding density estimation model – a Dirichlet process mixture based on a four-dimensional normal kernel with unrestricted covariance matrix – applied to the original data

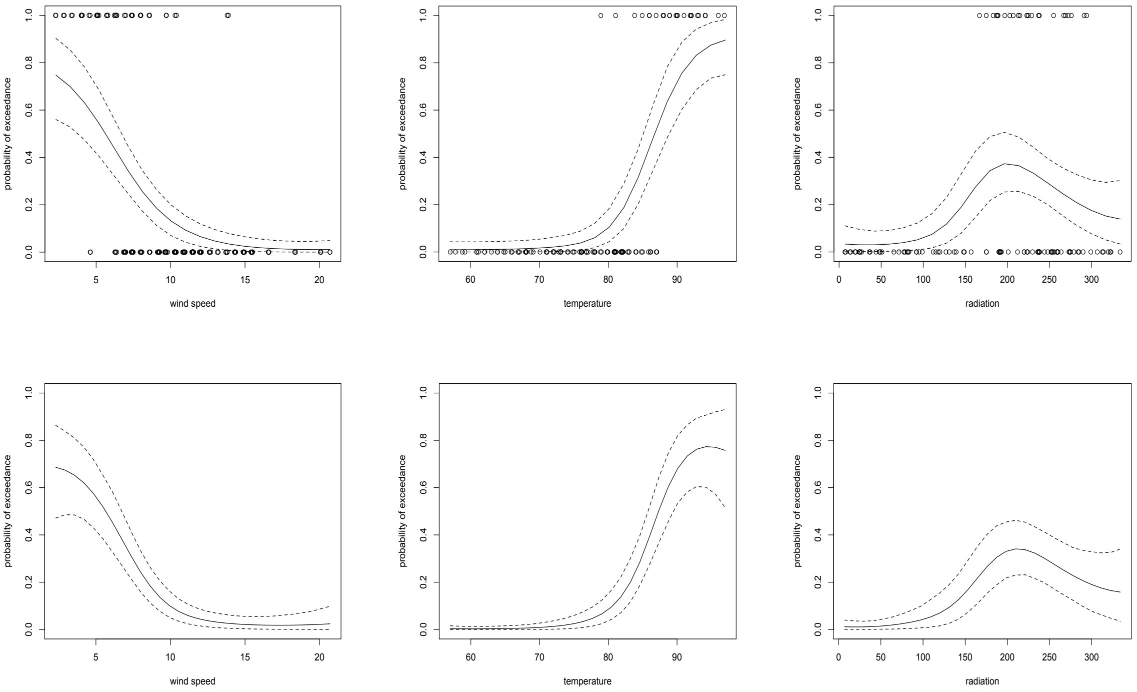

Figure 1: Ozone data. Posterior mean (solid line) and 90% uncertainty bands (dashed lines) for probability of exceedance versus wind speed (left panels), temperature (middle panels), and radiation (right panels). The top row plots results under the binary regression model, including the binary response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

set $\{(z_i, x_i) : i = 1, ..., 111\}$. 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 $\text{pr}(z > 70|x_i)$ curves that are similarly diffuse to the ones for $\text{pr}(y = 1|x_i)$. 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.

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 $\text{N}_p(x; \mu^x, \Sigma^{xx})$ and $\text{Bern}(y; \Phi(\mu^z))$. 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 $\pi_l = \Phi(\mu_l^z)$. Mixtures of this product-kernel form have been previously proposed in the literature; see, for instance, Dunson & 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

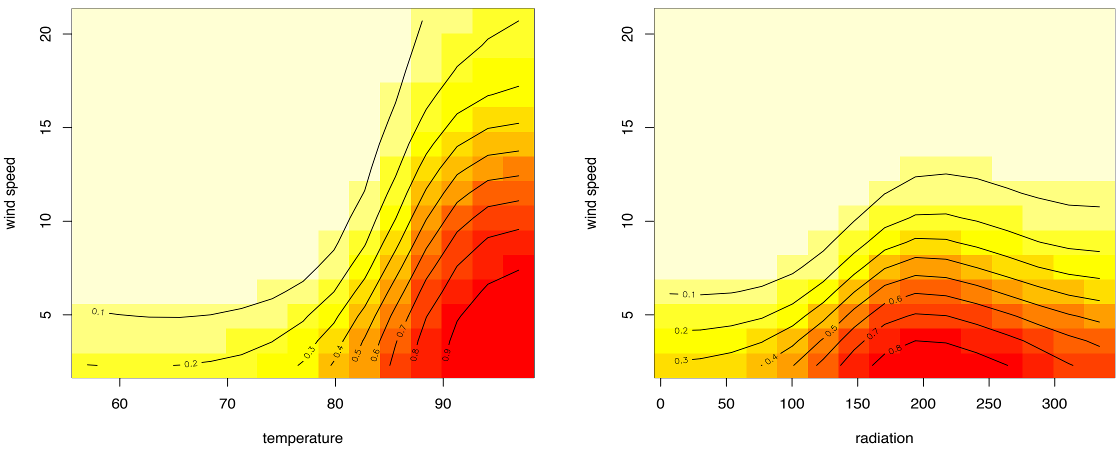

Figure 2: Ozone data. Posterior mean surface for probability of exceedance versus temperature and wind speed (left panel), and radiation and wind speed (right panel). Probabilities ranging from 0 to 1 are indicated by a spectrum of colors from white to red.

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 & 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 $P(m) = \sum_{i=1}^{n} \text{var}^{(m)}(y_{new,i}|\text{data})$, and goodness of fit term $G(m) = \sum_{i=1}^{n}\{y_i - \text{E}^{(m)}(y_{new,i}|\text{data})\}^2$. Here, $\text{E}^{(m)}(y_{new,i}|\text{data})$ 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 Dirichlet process binary regression model.

3.2 Estimating Natural Selection Functions in Song Sparrows

In addition to enabling more general modelling 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 & Arnold, 1983; Schluter, 1988; Janzen & 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 modelling 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.

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 3, 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.

For each of the two traits, we estimated the standardized directional selection differential, $\bar{x}_i^* - \bar{x}_i$, which provides a measure of selection intensity representing the change in mean value of a phenotype produced by selection (Lande and Arnold, 1983). Here, $\bar{x}_i = \int x_i f(x_i) \mathrm{d}x_i$ is the mean value of phenotypic trait x_i before selection, and $\bar{x}_i^* = \int x_i f(x_i | y = 1) \mathrm{d}x_i = \{\mathrm{pr}(y = 1)\}^{-1} \int x_i \mathrm{pr}(y = 1, x_i) \mathrm{d}x_i$ is the mean value after selection; the marginal probability $\mathrm{pr}(y = 1)$ is referred to as mean absolute fitness. Under our model, $\bar{x}_i = \sum_{l=1}^{N} p_l \mu_l^{x_i}$, mean absolute fitness is given by $\sum_{l=1}^{N} p_l \Phi(\mu_l^z)$, and $\int x_i \mathrm{pr}(y = 1, x_i; G_N) \mathrm{d}x_i$ 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

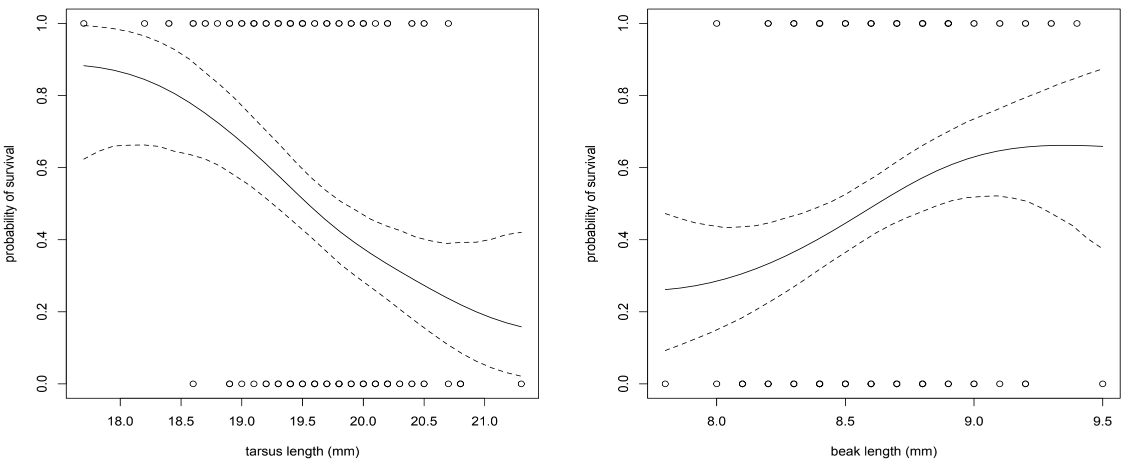

Figure 3: Song sparrows data. Posterior mean (solid line) and 90% uncertainty bands (dashed lines) for the probability of survival as a function of tarsus length and beak length.

contain zero. Combined with the estimated 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

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

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 j-th selection gradient by $n^{-1}\sum_{i=1}^{n} \partial \mathrm{pr}(y = 1|x)/\partial x_j|_{x=x_i}$. Our joint mixture modelling 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 i-th component of the selection gradient vector can be written as $\{\partial \mathrm{pr}(y = 1, x; G_N)/\partial x_i\} - \{\mathrm{pr}(y = 1|x; G_N)\partial f(x; G_N)/\partial x_i\}$, for $i = 1, 2$. 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.

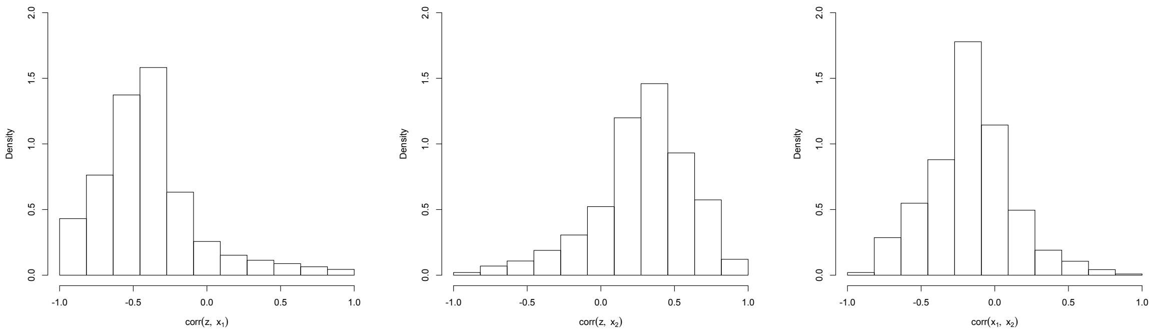

Figure 4: Song sparrows data. Posterior predictive samples for corr(z, x_1) (left panel), corr(z, x_2) (middle panel), and corr(x_1, x_2) (right panel).

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 $\int(x_1 - \bar{x}_1^*, x_2 - \bar{x}_2^*)^t(x_1 - \bar{x}_1^*, x_2 - \bar{x}_2^*)f(x|y = 1)\mathrm{d}x$. 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.

One way to check if a kernel with independent components for x and y would be adequate is to 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 $p(z, x|\mathrm{data}) = \int \sum_{l=1}^{N} p_l \mathrm{N}_3(z, x; \mu_l, \Sigma_l)p(\Theta|\mathrm{data})\,\mathrm{d}\Theta$, 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 4. 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 Dirichlet process 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 modelling approach for problems in the environmental sciences and in population biology.

We discussed the special case of the model arising from $\Sigma^{zx} = 0$ in the mixture kernel, which has been previously proposed with the further restriction that Σ^{xx} is diagonal (Dunson & Bhattacharya, 2011). There, the simplicity of independence among covariates within mixture components was viewed as appealing, and the response was modelled as independent of the covariates within the kernel, resulting in what was termed a product-kernel. In a related approach, Shahbaba & Neal (2009) also build a model for the joint distribution $f(y, x)$, but do so by separately estimating $f(x)$ and $f(y|x)$, 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 is used as a prior, they too assume $x_1, ..., x_p$ to be independent within each component. This idea was generalized by Hannah et al. (2010) to allow any standard generalized linear model to take the place of 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 correctly 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 & Bhattacharya (2011) note that if interest centers on quantifying dependence, then there is no need to introduce a response, and the method for joint modelling 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 proposed modelling 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 modelling approach to incorporate ordinal and mixed ordinal-continuous responses follow naturally. In analogy with the binary setting, a univariate ordinal response y 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 $k = 2, ..., K-1$, and $y = 1$ or $y = K$ if and only if $z \leq \gamma_1$ or $z > \gamma_{K-1}$. A Dirichlet process mixture of normals 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, $\gamma_1, ..., \gamma_{K-1}$, are fixed. A key feature of the nonparametric mixture modelling 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. Modelling 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 modelling 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

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

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 for this to be 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:

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

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

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

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

Because (9) implies $\Sigma_{1l}^{zx}\Sigma_{2m}^{zx} = \Sigma_{1m}^{zx}\Sigma_{2l}^{zx}$ for any $l,m=1,...,p$, 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 $y=(y_1,...,y_r)|\mu,\beta,\Delta \sim \mathrm{N}_r(\mu,\beta^{-1}\Delta(\beta^{-1})^t)$, 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 $(y_1-\mu_1),...,(y_{i-1}-\mu_{i-1})$, 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 $\tilde{\beta}_i$, $i = 1, ..., q$, on each side of the equation results in a system of q equations:

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

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

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

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 $v^t T v > 0$. When $r = 2$, $v^t T v$ becomes $v_1^2 (y_1 - \mu_1)^2/\delta_2$. When $r = 3$, $v^t T v$ 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 $(v_{q-r+2}(y_1 - \mu_1) + ... + v_q(y_{r-1} - \mu_{r-1}))^2/\delta_r$ is added in going from $r-1$ to r dimensions. Clearly, T is positive semidefinite. However, to have $v^t T v > 0$, and all elements of T strictly positive, it must be the case that $y_i \neq \mu_i$, for $i = 1, ..., r-1$, 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, $d_{1+j(j-1)/2}, ..., d_{j(j+1)/2}$. 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 $d_{1+j(j-1)/2} = ... = d_{j(j+1)/2-1} = 0$. Then the j equations become

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

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

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

With n data vectors, $(y_{i,1}, ..., y_{i,r})$, for $i = 1, ..., n$, 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 i-th 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 $\prod_{k=1}^r \delta_k^{-1/2} \exp\{-\text{tr}(\beta^t \Delta^{-1} \beta (y-\mu)(y-\mu)^t)/2\}$. By properties of trace, $\text{tr}(\beta^t \Delta^{-1} \beta (y-\mu)(y-\mu)^t) = \text{tr}(\beta (y-\mu)(y-\mu)^t \beta^t \Delta^{-1})$. 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, $(y_{i,1},...,y_{i,r})$, $i=1,...,n$, the likelihood for δ_k, $k=1,...,r$, 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) $\Sigma_{22\cdot 1}\sim \mathrm{IW}_{r-q}(v,T_{22\cdot 1})$, 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) $\Sigma_{11}^{-1}\Sigma_{12}|\Sigma_{22\cdot 1}^{-1}\sim \mathrm{MN}_{q,r-q}(T_{11}^{-1}T_{12},T_{11}^{-1},\Sigma_{22\cdot 1})$. Here, MN denotes the matrix normal distribution such that, conditionally on $\Sigma_{22\cdot 1}$, $\mathrm{vec}(\Sigma_{11}^{-1}\Sigma_{12})\sim \mathrm{N}_{q(r-q)}(\mathrm{vec}(T_{11}^{-1}T_{12}),T_{11}^{-1}\otimes\Sigma_{22\cdot 1})$.

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 $T^i=\mathrm{diag}(T_1,\ldots,T_i)$. Applying result (b) with $q=p$, we obtain $\delta_{p+1}\sim \mathrm{IG}(0.5v,0.5T_{p+1})$. 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 $\Sigma_{1:p,1:p}\sim \mathrm{IW}_p(v-1,T^p)$. Next, using result (b) for matrix $\Sigma_{1:p,1:p}$ with $q=p-1$, we have $\delta_p\sim \mathrm{IG}(0.5(v-1),0.5T_p)$, since $(\Sigma_{1:p,1:p})_{22\cdot 1}=\delta_p$. Analogously, applying results (a) and (b) in succession, we obtain $\delta_i\sim \mathrm{IG}(0.5(v+i-(p+1)),0.5T_i)$, for $i=2,\ldots,p+1$.

For each $i=2,\ldots,p+1$, result (a) yields an $\mathrm{IW}_i(v+i-(p+1),T^i)$ 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 $(-\beta_{i,1},\ldots,-\beta_{i,i-1})^t|\delta_i\sim \mathrm{N}_{i-1}((0,\ldots,0)^t,\delta_i(T^{i-1})^{-1})$, for $i=2,\ldots,p+1$. This uses the fact that $(T^i)_{12}=(0,\ldots,0)^t$, $\mathrm{vec}((\Sigma_{1:i,1:i})_{11}^{-1}(\Sigma_{1:i,1:i})_{12})=(-\beta_{i,1},\ldots,-\beta_{i,i-1})^t$, 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, $\mathrm{E}^{(m)}(y_{new,i}|\mathrm{data})$, and posterior predictive variance, $\mathrm{var}^{(m)}(y_{new,i}|\mathrm{data})$, 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, $\mathrm{E}(y|x_i,\mathrm{data})=\{p(x_i|\mathrm{data})\}^{-1}\int\sum_{l=1}^{N}p_l\mathrm{N}_p(x_i;\mu_l^x,\Sigma_l^{xx})\Phi(\mu_l^z)\,p(\Theta|\mathrm{data})\mathrm{d}\Theta$, with $p(x_i|\mathrm{data})=\int\sum_{l=1}^{N}p_l\mathrm{N}_p(x_i;\mu_l^x,\Sigma_l^{xx})\,p(\Theta|\mathrm{data})\mathrm{d}\Theta$, and $\mathrm{E}(y^2|x_i,\mathrm{data})$ also has the same form. Under the proposed model, $\mathrm{E}(y|x_i, \mathrm{data})$ is given by

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

where $p(x_i|\mathrm{data}) = \int \sum_{l=1}^{N} p_l \mathrm{N}_p(x_i; \mu_l^x, \Sigma_l^{xx})\, p(\Theta|\mathrm{data})\mathrm{d}\Theta$, and $\mathrm{E}(y|x_i, \mathrm{data}) = \mathrm{E}(y^2|x_i, \mathrm{data})$. 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. & S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669 – 679.

Basu, S. & S. Mukhopadyay (2000). Bayesian Analysis of Binary Regression Using Symmetric and Asymmetric Links. *The Indian Journal of Statistics Series B 62*, 373–387.

Box, G. & G. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Reading, Mass.: Addison-Wesley.

Choudhuri, N., Ghosal, S., & A. Roy (2007). Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology 4*, 227–243.

Connor, R.J. & J.E. Mosimann (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association 64*, 194–206.

Daniels, M. & M. Pourahmadi (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika 89*, 553–566.

Dunson, D.B. & A. Bhattacharya (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., & M. West (eds.), *Bayesian Statistics 9, Proceedings of the Ninth Valencia International Meeting*, 145–164. Oxford University Press.

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

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

Follmann, D.A. & E. Lambert. (1989). Generalizing logistic regression by nonparametric modelling. *Journal of the American Statistical Association 84*, 295–300.

Hannah, L.A., Blei, D.M. & W.B. Powell. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research 12*, 1923–1953.

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

Ishwaran, H. & M. Zarepour (2000). Markov Chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. *Biometrika 87*, 371–390.

Janzen, F. & H. Stern (1998). Logistic regression for empirical studies of multivariate selection. *Evolution 52*, 564–1571.

Kottas, A., P. Müller & F. Quintana (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics 14*, 610–625.

Lande, R. & S.J. Arnold (1983). The measurement of selection on correlated characters. *Evolution 37*, 1210–1226.

Mukhopadyay, S. & A.E. Gelfand (1997). Dirichlet Process Mixed Generalized Linear Models. *Journal of the American Statistical Association 92*, 633–639.

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

Newton, M.A., C. Czado & C. Chappell (1996). Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association 91*, 142–153.

Schluter, D. (1988). Estimating the Form of Natural Selection on a Quantitative Trait. *Evolution 42*, 849–861.

Schluter, D & J. Smith. (1986). Natural Selection on Beak and Body Size in the Song Sparrow. *International Journal of Organic Evolution 40*, 2211–231.

Sethuraman, J. (1994). A constructive definition of the Dirichlet prior. *Statistica Sinica 4*, 639–650.

Shahbaba, B., & R. Neal. (2009). Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research 10*, 1829–1850.

Taddy, M. & A. Kottas (2010). A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business and Economic Statistics 28*, 357–369.

Trippa, L. & P. Muliere (2009). Bayesian nonparametric binary regression via random tesselations. *Statistics and Probability Letters 79*, 2273–2282.