

Bayesian Isotonic Regression for Discrete Outcomes

David B. Dunson^{1,*}

¹Biostatistics Branch

National Institute of Environmental Health Sciences

MD A3-03, P.O. Box 12233

Research Triangle Park, NC 27709, U.S.A.

* dunson1@niehs.nih.gov

SUMMARY. This article proposes a semiparametric Bayesian approach for inference on an unknown isotonic regression function, $f(x)$, characterizing the relationship between a continuous predictor, X , and a response variable, Y , adjusting for covariates, Z . A novel prior formulation is used, which avoids parametric assumptions on $f(x)$, while enforcing the non-decreasing constraint and assigning positive prior probability to the null hypothesis of no association between X and Y conditional on Z . Through the use of carefully tailored hyperprior distributions, we allow for borrowing of information across different regions of X in estimating of $f(x)$ and in assessing hypotheses about local increases in the function. Due to conjugacy properties, posterior computation is straightforward in a variety of settings, including log-linear models for Poisson data and logistic regression for binary outcomes. The methods are illustrated using a series of simulated data examples.

Key Words: Generalized additive model; Log linear; Logistic regression; Mixture prior; Multiple testing; Nonparametric regression; Smoothing; Trend test.

1. Introduction

There is commonly interest in the association between a continuous predictor, X , and a response variable, Y , adjusting for covariates, $\mathbf{Z} = (Z_1, \dots, Z_q)'$. In many applications, the mean of Y can be assumed to be non-decreasing with increases in X for a fixed value of \mathbf{Z} . This may be the case, for example, if Y is an indicator of the occurrence of a disease, and X is the level of a potentially-adverse environmental exposure. In such settings, one typically does not know *a priori* whether Y is associated with X adjusting for the confounding effects of \mathbf{Z} , and the functional form of the regression function relating the mean of Y to X is unknown.

Our interest focuses on assessing evidence of an association, while also estimating the regression function subject to the monotonicity constraint but allowing for flat regions. Flat regions of the curve correspond to ranges of the predictor, X , across which there is no effect on the response variable, Y . Assessing whether there is an increase in the function within particular regions of X is a primary goal in applications, such as toxicologic and epidemiologic studies assessing dose response. Since response data in such studies are typically non-Gaussian, we focus on count and dichotomous outcomes.

A variety of frequentist approaches have been proposed for monotone curve estimation (Mammen, 1991; Lee, 1996; Ramsay, 1998; Hall and Huang, 2001; Mammen et al., 2001) and testing of monotonicity (Ghosal, Sen, and van der Vaart, 2000; Doveh, Shapiro, and Feigen, 2002). Although most approaches have focused on normal outcomes with a single predictor, models for multiple predictors (Bachetti, 1989) and non-Gaussian outcomes (Morton-Jones et al., 2000; Wang, 2000) have been considered. In addition, Bayesian approaches for monotone curve estimation for normal (Lavine and Mockus, 1995; Holmes and Heard, 2003; Neelon and Dunson, 2003) and binary data (Gelfand and Kuo, 1991; Ramgopal, Laud, and Smith, 1993) have been proposed, though only the articles of Holmes and Heard (2003) and Neelon and Dunson (2003) allow for an uncertain association.

To our knowledge, currently available methods do not allow one to simultaneously conduct inferences on an uncertain association, while also estimating the regression function nonparametrically subject to the monotonicity constraint but allowing for flat regions. The methods of Dunson and Neelon (2003) and Holmes and Heard (2003) allow for estimation and testing, but are based on parametric piecewise linear and piecewise constant models, respectively, for normal means.

In contrast, the approach proposed in this article is fully nonparametric in its treatment of the unknown isotonic regression function, and the focus is on Poisson and Bernoulli distributed outcomes. We incorporate the constraint and allow for flat regions through a carefully structured prior for multiplicative increments on the exponentiated regression function. In particular, the increments are assigned independent prior densities consisting of mixtures of point masses at one and truncated gamma densities. A hyperprior structure is then defined in order to borrow information about the point mass probabilities and the magnitude of the changes across the increments, effectively smoothing the curve. The prior is conditionally-conjugate, resulting in simplified posterior computation.

Section 2 proposes the methodology in the case where data consist of Poisson counts, and discusses theoretical properties. Section 3 describes modifications for additive logistic regression modeling of Bernoulli data. Section 4 contains results from simulated examples, and Section 5 discusses the results.

2. Isotonic Regression for Count Data

2.1 Data Structure and Model

For subject i ($i = 1, \dots, n$), let y_i denote a Poisson distributed outcome variable, let $x_i \in [x_L, x_U]$ denote a continuous predictor, and let $\mathbf{z}_i = (z_{i1}, \dots, z_{iq})'$ denote a vector of covariates. Our initial focus is on the semiparametric regression model,

$$\log E(y_i | x_i, \mathbf{z}_i) = f(x_i) + \mathbf{z}_i' \boldsymbol{\beta}, \tag{1}$$

where $f(\cdot)$ is an unknown non-decreasing function, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$ are regression coefficients for the covariates \mathbf{z}_i .

Letting $x_{(1)} < \dots < x_{(k)}$ denote the $k \leq n$ unique values in $\mathbf{x} = (x_1, \dots, x_n)'$, so that $k = n$ when X is truly continuous and $k < n$ when there are ties, expression (1) implies

$$\begin{aligned} E(y_i | x_i = x_{(j)}, \mathbf{z}_i) &= \exp\{f(x_{(j)})\} \exp(\mathbf{z}'_i \boldsymbol{\beta}) = g(x_{(j)}) \exp(\mathbf{z}'_i \boldsymbol{\beta}), \\ &= \left\{ \prod_{h=1}^j \frac{g(x_{(h)})}{g(x_{(h-1)})} \right\} \exp(\mathbf{z}'_i \boldsymbol{\beta}) = \left\{ \prod_{h=1}^j \gamma_h \right\} \exp(\mathbf{z}'_i \boldsymbol{\beta}), \end{aligned} \quad (2)$$

where $g(x) = \exp\{f(x)\}$, $g(x_{(0)}) = 1$, and $\gamma_h = g(x_{(h)})/g(x_{(h-1)})$, for $h = 1, \dots, k$. Under this structure, the likelihood is proportional to

$$L(\mathbf{y}; \mathbf{x}, \mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^n \left[\left\{ \prod_{j=1}^k \gamma_j^{1(x_i \geq x_{(j)})} \right\} \exp(\mathbf{z}'_i \boldsymbol{\beta}) \right]^{y_i} \exp \left[- \left\{ \prod_{j=1}^k \gamma_j^{1(x_i \geq x_{(j)})} \right\} \exp(\mathbf{z}'_i \boldsymbol{\beta}) \right].$$

Note that we have made no assumptions about $f(x)$ in formulating the likelihood, and constraints on $f(x)$ will be incorporated through the prior.

2.2 Mixture Prior

The most common strategy for choosing a prior distribution for a function, $f(x)$, is to specify independent prior densities for increments on $f(x)$ across a finite number of intervals forming a partition of the support of x . A probability distribution on the space of $\{f(x)\}$ can be specified by assuming that the distributional assumptions hold for all possible partitions under the regularity condition that there is consistency between partitions. Special cases include the Dirichlet process (Ferguson, 1973; 1974) and the gamma process (Kalbfleisch, 1978). Unfortunately, finding a distribution that satisfies the regularity condition and that results in a prior for $f(x)$ that assigns positive probability to flat regions while maintaining the non-decreasing constraint is an unsolved problem.

For this reason, we limit ourselves to the problem of inference on $f(x)$ for the values of the predictor observed in the data set, $x \in \{x_{(1)}, \dots, x_{(k)}\}$. A prior distribution for $f(x_{(j)})$, for $j = 1, \dots, k$, can be specified by choosing a prior distribution for $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_k)'$, the

multiplicative increments on the exponentiated function across the observed data partition of $[x_L, x_U]$. The non-decreasing constraint on $f(x)$ implies that $\gamma_j = \exp\{f(x_{(j)}) - f(x_{(j-1)})\} \geq 1$, for $j = 2, \dots, k$, with $\gamma_j = 1$ corresponding to no change in $f(x)$ across the interval $x \in [x_{(j-1)}, x_{(j)}]$ and $\gamma_j > 1$ corresponding to a strict increase.

As a prior for $\boldsymbol{\gamma}$ with appropriate support, we choose:

$$\pi(\boldsymbol{\gamma}) = \mathcal{G}(\gamma_1; a_1, b_1) \prod_{j=2}^k \left\{ 1(\gamma_j = 1)p_j + 1(\gamma_j > 1) \frac{(1 - p_j)\mathcal{G}(\gamma_j; a_j, b_j)}{1 - F(1; a_j, b_j)} \right\}, \quad (3)$$

where $\mathcal{G}(\gamma; a, b)$ denotes the gamma density function with mean a/b and variance a/b^2 , $p_j = \Pr(\gamma_j = 1)$, and $F(\cdot; a, b)$ is the gamma distribution function. This prior assigns independent densities to the increments, $\gamma_2, \dots, \gamma_k$, with these densities consisting of mixtures of point masses at one and gamma densities truncated on the left by one.

Unlike the alternative approach of assigning independent gamma densities to increments on $f(x)$ according to a gamma process, our prior allows flat regions through incorporation of point masses. In addition, the multiplicative structure, which makes it necessary to truncate the gamma densities on the left by one to ensure monotonicity, results in conditional conjugacy (as we illustrate in Subsection 2.4).

A prior proposed in previous research on ordered categorical predictors in survival analysis (Dunson and Herring, 2003) has a related structure, but the dimension of $\boldsymbol{\gamma}$ is fixed in advance by the number of categories, which is typically small. In contrast, in prior (3) the dimension of $\boldsymbol{\gamma}$ typically increases with the sample size, necessitating a very different approach to prior elicitation.

2.3 Prior Elicitation and Hyperprior Specification

We follow the approach of choosing hyperprior distributions in order to borrow information across the different increments, simplify prior elicitation, and reduce the sensitivity of inferences to subjectively-chosen hyperparameters. In particular, we first express the hyper-

parameters p_j, a_j, b_j as follows:

$$p_j = \exp\{-\alpha(x_{(j)} - x_{(j-1)})\}, \quad a_j = \frac{c \exp\{f_0(x_{(j)})\}}{\exp\{f_0(x_{(j-1)})\}}, \quad \text{and} \quad b_j = c, \quad \text{for } j = 2, \dots, k, \quad (4)$$

where α measures the rate of occurrence of increases in $f(x)$, $f_0(x)$ is one's best guess for $f(x)$ given that the function is increasing at x , and c is a precision parameter measuring the degree to which $f(x)$ follows the trajectory predicted by $f_0(x)$ during increasing regions. Although $\alpha, f_0(x)$ and c can potentially be chosen subjectively, we prefer to use hyperprior distributions in order to allow the data to inform about their values.

In particular, focusing on the special case where $f_0(x)$ is linear with slope r so that $a_j = c \exp\{r(x_{(j)} - x_{(j-1)})\}$, for $j = 2, \dots, k$, we choose the following hyperprior distributions:

$$\pi(\alpha) = \mathcal{G}(\alpha; a_\alpha, b_\alpha), \quad \pi(r) = \mathcal{G}(r; a_r, b_r), \quad \text{and} \quad \pi(c) = \mathcal{G}(c; a_c, b_c), \quad (5)$$

where the intercept, $f(x_L)$, is characterized by γ_1 . The constants $a_\alpha, b_\alpha, a_r, b_r, a_c, b_c$ are fixed in advance by the investigator. We suggest reasonable default choices, and assess the degree to which inferences are driven by the data in Section 4.

One of our primary interests is in assessing evidence in favor of the null hypothesis, $H_0 : f(x)$ is constant for all $x \in [x_L, x_U]$, against the alternative hypothesis, $H_1 : f(x)$ has at least one increase so that $f(x_U) > f(x_L)$. To perform Bayesian inferences, it is necessary to choose hyperparameters consistent with a prespecified value for $\Pr(H_0)$, the prior probability of the null hypothesis, with a typical default value being $\Pr(H_0) = 0.5$.

Under the above specification, the probability of H_0 conditional on α is $\Pr(H_0 | \alpha) = \exp\{-\alpha(x_U - x_L)\}$. Then, integrating across the prior density for α , we have

$$\begin{aligned} \Pr(H_0) &= \int_0^\infty \Pr(H_0 | \alpha) \pi(\alpha) d\alpha \\ &= \int_0^\infty \exp\{-\alpha(x_U - x_L)\} \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha-1} \exp(-\alpha b_\alpha) d\alpha \\ &= \left(\frac{b_\alpha}{b_\alpha + x_U - x_L} \right)^{a_\alpha}. \end{aligned} \quad (6)$$

Therefore, as a reasonable approach to prior elicitation, one can choose $\Pr(H_0)$ (e.g., equal to 0.5) and then solve $b_\alpha = (x_U - x_L)/\{\Pr(H_0)^{-1/a_\alpha} - 1\}$, with a_α fixed in advance to control the adjustment for multiple testing and the degree of correlation in the local null hypotheses. The role of a_α is considered in detail in the following subsection.

2.4 Properties of Prior in Assessing Local Hypotheses

In addition to comparing the global null and alternative hypotheses, H_0 and H_1 , it is typically of interest to investigate local hypotheses about increases in $f(x)$ within different subregions of $[x_L, x_U]$. Formally, letting $\mathcal{D}_j = (d_{j-1}, d_j]$ for $j = 1, \dots, h$ denote a sequence of intervals forming a partition of $[x_L, x_U]$, interest focuses on comparing the local null hypothesis, $H_{0j} : f(x)$ is constant for all $x \in \mathcal{D}_j$, to the local alternative hypothesis, $H_{1j} : f(x)$ increases at least once within \mathcal{D}_j . In considering this sequence of local hypotheses from a frequentist perspective, one would need to adjust for an inflated type I error rate due to multiple testing, e.g., by using a Bonferroni correction.

A Bayesian version of the Bonferroni correction would be to fix $\Pr(H_0) = 0.5$, with

$$\prod_{j=1}^h \exp\{-\alpha(d_j - d_{j-1})\} = \Pr(H_0) = 0.5, \quad (7)$$

which implies that $\alpha = -\log \Pr(H_0)/(x_U - x_L)$. Letting $\delta_j = d_j - d_{j-1}$ and $\Delta = x_U - x_L$, this strategy assumes *a priori* independence between the different local hypotheses, with $\Pr(H_0) = \exp(-\alpha\Delta)$ and $\Pr(H_{0j}) = \exp(-\alpha\delta_j) = \Pr(H_0)^{\delta_j/\Delta}$. When the \mathcal{D}_j 's are equal width intervals, this corresponds to the $\Pr(H_0)^{1/h}$ correction proposed previously (Kass and Raftery, 1995). As the width of the subregions \mathcal{D}_j decrease and the number of comparisons increase, $\Pr(H_{0j})$ moves rapidly towards one, reflecting the rather severe penalty for multiple comparisons imposed by this approach. For an interesting article on the Bayesian Bonferroni adjustment, refer to Westfall, Johnson, and Utts (1997).

Note that this approach considers α to be a constant chosen to yield a particular value for $\Pr(H_0)$. In contrast, the approach proposed in subsection 2.3 considers α to be an unknown

parameter, which is assigned a hyperprior distribution. This strategy induces dependency in the different local hypotheses and results in a penalty for multiple comparisons which tends to be much less severe than the Bayesian Bonferroni approach. This property is formalized in the following theorem:

Theorem 1. Under the prior specification proposed in subsection 2.3, for any given value of the global null hypothesis H_0 , the prior probability of the local null hypothesis H_{0j} is bounded below by $\Pr(H_0)$ and above by $\Pr(H_0)^{\delta_j/\Delta}$, the value obtained under the Bonferroni-type prior (7).

The marginal probability of H_{0j} integrating out α follows the form $\Pr(H_{0j}) = E_\alpha[(e^{-\alpha\Delta})^{\delta_j/\Delta}]$ under the approach of subsection 2.3, and the form $\Pr(H_{0j}) = E_\alpha[e^{-\alpha\Delta}]^{\delta_j/\Delta} = \Pr(H_0)^{\delta_j/\Delta}$ under the Bonferroni approach. Theorem 1 follows directly from Jensen's inequality.

The penalty for multiple testing induced by our approach can be formalized as follows:

$$\Pr(H_{0j}) = \left(\frac{1}{1 + \delta_j/b_\alpha} \right)^{a_\alpha} = \left\{ \left(1 - \frac{\delta_j}{\Delta} \right) \Pr(H_0)^{1/a_\alpha} + \frac{\delta_j}{\Delta} \right\}^{-a_\alpha} \Pr(H_0), \quad (8)$$

which is greater than $\Pr(H_0)$ by a factor which increases with a_α and $(1 - \delta_j/\Delta)$. Under the Bonferroni approach, the term in $\{\cdot\}$ is instead $\Pr(H_0)^{(1-\delta/\Delta)/a_\alpha}$. Letting $Q = \Pr(H_0)^{1/a_\alpha}$ and $P = (1 - \delta_j/\Delta)$, an alternative proof to Theorem 1 follows from the inequality $PQ + 1 - P > Q^P$, for all $P, Q : 0 < P < 1$ and $0 < Q < 1$.

The upper bound on $\Pr(H_{0j})$ provided by $\Pr(H_0)^{\delta_j/\Delta}$ converges to an equality for large a_α . Hence, the penalty for multiple testing is smaller for values of a_α closer to zero, and the performance is similar to the Bayesian-Bonferroni approach for large a_α . In the special case where $a_\alpha = 1$, we have

$$\Pr(H_{0j}) = \frac{\Pr(H_0)}{\Pr(H_0) + (\delta_j/\Delta)\{1 - \Pr(H_0)\}}, \quad (9)$$

which further simplifies to $\Pr(H_{0j}) = \Delta/(\delta_j + \Delta)$ for $\Pr(H_0) = 0.5$. Also note that in the limit as $\delta_j/\Delta \rightarrow 0$, the width of D_j becomes very small and $\Pr(H_{0j}) \rightarrow 1$. This is intuitively

reasonable, since the probability of an increase occurring within an infinitesimal interval should be vanishingly small.

An additional property of the prior is that it accounts for dependency in the different hypotheses, H_{01}, \dots, H_{0h} . We formalize this in Theorem 2:

Theorem 2. The prior specification in subsection 2.3 induces positive correlation in the local hypotheses H_{01}, \dots, H_{0h} , with the level of correlation increasing with $1/a_\alpha$.

To demonstrate this, we calculate the prior probability of H_{0j} given that $H_{0j'}$ is true, for $j' \neq j$, divided by the prior probability of H_{0j} (letting $P = \Pr(H_0)$):

$$\begin{aligned} \frac{\Pr(H_{0j} | H_{0j'})}{\Pr(H_{0j})} &= \frac{1}{\Pr(H_{0j'})\Pr(H_{0j})} \int_0^\infty \Pr(H_{0j}, H_{0j'} | \alpha) \pi(\alpha) d\alpha \\ &= \frac{1}{\Pr(H_{0j'})\Pr(H_{0j})} \int_0^\infty \exp(-\alpha\delta_j) \exp(-\alpha\delta_{j'}) \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha-1} \exp(-\alpha b_\alpha) d\alpha \\ &= \left[\frac{\left\{1 + \frac{\delta_j}{\Delta}(P^{-1/a_\alpha} - 1)\right\} \left\{1 + \frac{\delta_{j'}}{\Delta}(P^{-1/a_\alpha} - 1)\right\}}{1 + \left(\frac{\delta_j + \delta_{j'}}{\Delta}\right)(P^{-1/a_\alpha} - 1)} \right]^{a_\alpha} > 1. \end{aligned} \quad (10)$$

This expression is interpretable as the multiplicative increase in the prior probability of H_{0j} given the additional information that $H_{0j'}$ is true. In the special case where $\delta_j = \delta_{j'}$, this expression further simplifies to

$$\frac{\Pr(H_{0j} | H_{0j'})}{\Pr(H_{0j})} = \left\{ 1 + \frac{\left(\frac{\delta_j}{\Delta}\right)^2 (P^{-1/a_\alpha} - 1)^2}{1 + \frac{2\delta_j}{\Delta}(P^{-1/a_\alpha} - 1)} \right\}^{a_\alpha} > 1. \quad (11)$$

Since (10) and (11) are greater than one, the local null hypotheses are positive correlated *a priori*. It is clear from the above expressions that the correlation structure is exchangeable when the intervals \mathcal{D}_j are equally-spaced, and otherwise depends on the relative widths δ_j/Δ . The level of correlation decreases as a_α increases, which is consistent with the above observation that large values of a_α induce a penalty for multiple testing, which is similar to the Bonferroni adjustment. Although the degree of *a priori* dependency in the different hypotheses is driven by the subjectively-chosen parameter a_α , the data are informative about

the value of α . Hence, as we demonstrate in Section 4, the posterior density for the unknown function $f(x)$ tends to be driven primarily by the data even for moderate sample sizes such as $n = 25$.

2.5 Posterior Computation

An appealing feature of the above prior specification is that the full conditional posterior distribution of γ_j follows the same one-inflated truncated gamma form as the prior shown in expression (3), but with updated parameters that depend on the data. This conditionally-conjugate form facilitates posterior computation using an MCMC algorithm with Gibbs and Metropolis steps. Our MCMC algorithm alternates between the following steps:

Step 1. Sample γ_1 from its full conditional distribution, which is

$$\mathcal{G}\left(\gamma_1; a_1 + \sum_{i=1}^n y_i, b_1 + \sum_{i=1}^n \left\{ \prod_{j=2}^k \gamma_j^{1(x_i \geq x_{(j)})} \right\} \exp(\mathbf{z}'_i \boldsymbol{\beta})\right). \quad (12)$$

Step 2. Sample γ_j , for $j = 2, \dots, k$, from its full conditional posterior distribution, which is the mixture of a point mass at one with probability $\hat{\pi}_j =$

$$\Pr(\gamma_j = 1 | -) = \frac{p_j \exp\{- (B_j - b_j)\}}{p_j \exp\{- (B_j - b_j)\} + (1 - p_j) \frac{C(a_j, b_j)\{1 - F(1; A_j, B_j)\}}{C(A_j, B_j)\{1 - F(1; a_j, b_j)\}}}, \quad (13)$$

and a $\mathcal{G}(\gamma_j; a_j, b_j)$ density truncated below by one, where $A_j = a_j + \sum_{i=1}^n y_i 1(x_i \geq x_{(j)})$, $B_j = b_j + \sum_{i=1}^n \left\{ \prod_{h:h \neq j}^n \gamma_h^{1(x_i \geq x_{(j)})} \right\} \exp(\mathbf{z}'_i \boldsymbol{\beta})$, and $C(a, b)$ is the constant in the $\mathcal{G}(\cdot; a, b)$ density.

Step 3. Introduce the latent variable $M_j \sim \text{Poisson}(\alpha)$, where $M_j = 0$ if $\gamma_j = 1$ and $M_j > 0$ if $\gamma_j > 1$. Sample M_j from its full conditional distribution by letting $M_j = 0$ if $\gamma_j = 1$ and otherwise sampling from $\text{Poisson}(-\log \hat{\pi}_j)$ truncated so that $M_j > 0$. Then, sample α from its full conditional distribution, which is $\mathcal{G}(\alpha; a_\alpha + \sum_{j=2}^k M_j, b_\alpha + k - 1)$.

Step 4. Update $\boldsymbol{\beta}, r, c$ using Metropolis steps.

Step 5. Repeat steps 1-3 until apparent convergence, and calculate posterior summaries based on a large number of additional iterates.

This algorithm is easy to program and has had good computational efficiency in simulated and real data applications, with rapid convergence and good mixing of the Markov chain. To estimate the posterior probability of the global null hypothesis, we use

$$\widehat{\Pr}(\text{H}_0 \mid \text{data}) = \frac{1}{S} \sum_{s=1}^S \mathbf{1}(\gamma_1^{(s)} = \dots = \gamma_k^{(s)} = 1),$$

where $\gamma_j^{(s)}$ is the value of γ_j at iteration s of the MCMC algorithm, with $s = 1$ corresponding to the first iteration after burn-in. To estimate the probability of the j th local null hypothesis (i.e., $f(x)$ is constant for all $x \in \mathcal{D}_j$), we use

$$\widehat{\Pr}(\text{H}_{0j} \mid \text{data}) = \frac{1}{S} \sum_{s=1}^S \mathbf{1}(\gamma_l^{(s)} = 1 \text{ for all } l : x_{(l)} \in \mathcal{D}_j),$$

where we assume that $d_j \in \{x_{(1)}, \dots, x_{(k)}\}$ for $j = 0, \dots, h$ (without loss of generality since the vector $x_{(1)}, \dots, x_{(k)}$ can be defined as the union of the unique values of \mathbf{x} and $\mathbf{d} = (d_0, d_1, \dots, d_h)'$).

3. Isotonic Logistic Regression

With minor modifications, the approach described in Section 2 can be used for isotonic logistic regression for binary outcome data. In particular, letting y_i denote a 0/1 outcome variable, with the other notation as defined previously, we consider the model

$$\text{logitPr}(y_i = 1 \mid x_i = x_{(j)}, \mathbf{z}_i) = f(x_{(j)}) + \mathbf{z}_i' \boldsymbol{\beta}. \quad (14)$$

In order to utilize the results of Section 2, we note that the additive logistic regression model (14) has an equivalent underlying Poisson formulation:

$$\begin{aligned} y_i &= \mathbf{1}(y_i^* > 0) \\ y_i^* &\sim \text{Poisson}(y_i^*; \xi_i \exp\{f(x_{(j)}) + \mathbf{z}_i' \boldsymbol{\beta}\}) \\ \xi_i &\sim \mathcal{G}(\xi_i; 1, 1). \end{aligned} \quad (15)$$

The equivalence between (14) and (15) can be verified by calculating the marginal probability of $y_i = 1$ integrating out the latent y_i and ξ_i :

$$\begin{aligned}
\Pr(y_i = 1 | x_i = x_{(j)}, \mathbf{z}_i) &= \int_0^\infty \Pr(y_i^* > 0 | \xi_i, x_i = x_{(j)}, \mathbf{z}_i) \exp(-\xi_i) d\xi_i \\
&= 1 - \int_0^\infty \exp[-\xi_i \exp\{f(x_{(j)}) + \mathbf{z}'_i \boldsymbol{\beta}\}] \exp(-\xi_i) d\xi_i \\
&= 1 - [1 + \exp\{f(x_{(j)}) + \mathbf{z}'_i \boldsymbol{\beta}\}]^{-1} \\
&= \frac{\exp\{f(x_{(j)}) + \mathbf{z}'_i \boldsymbol{\beta}\}}{1 + \exp\{f(x_{(j)}) + \mathbf{z}'_i \boldsymbol{\beta}\}}.
\end{aligned}$$

The formulation shown in expression (15) makes it possible to directly apply the methods to the logistic regression case exactly as in Section 2, but with minor changes to the MCMC algorithm of Subsection 2.5. In particular, prior to step 1 of the algorithm we add the following data augmentation steps:

Step i. Sample y_i^* from its full conditional distribution by setting $y_i^* = 0$ if $y_i = 0$, and otherwise drawing from

$$\text{Poisson}(y_i^*; \xi_i \exp\{f(x_{(j)}) + \mathbf{z}'_i \boldsymbol{\beta}\}) \quad \text{truncated so that } y_i^* > 0.$$

Step ii. Sample ξ_i from its full conditional distribution, which is

$$\mathcal{G}(\xi_i; 1 + y_i^*, 1 + \exp\{f(x_{(j)}) + \mathbf{z}'_i \boldsymbol{\beta}\}).$$

Then, in steps 1,2 and 4 shown in Subsection 2.4, replace y_i with y_i^* and $\exp(\mathbf{z}'_i \boldsymbol{\beta})$ with $\xi_i \exp(\mathbf{z}'_i \boldsymbol{\beta})$. The underlying Poisson mixture representation of the logistic regression model, which is (to our knowledge) novel, should prove useful in other settings.

4. Simulation Examples

We checked the methodology through application to a number of simulated data sets. Although a full simulation study of the operating characteristics of the method (e.g., type I error rate, power, bias) is not computationally feasible, the examples considered provide evidence that the approach produces reasonable results in a variety of cases. We consider three

different forms for $f(x)$ in simulating the data: (i) *linear*, $f(x) = x$; (ii) *flat*, $f(x) = 0$; and (iii) *threshold*, $f(x) = 0$ for $x < 2/3$ and $f(x) = 2(x - 2/3)$ otherwise. For each choice of $f(x)$, we simulated data for count and binary data and for 3 different sample sizes $n = 25, 50, 100$, letting $x_i \sim U(0, 1)$, for $i = 1, \dots, n$.

Each data set was analyzed using the MCMC algorithm outlined in Subsection 2.5, with the modifications of Section 3 used for binary outcomes. We used a burn-in of 500 iterations and a collection interval of 2,500 iterations. The parameters in the hyperprior distributions were chosen to be $a_\alpha = 0.5$, $a_r = 0.5$, $b_r = 0.5$, $a_c = 1.0$, and $b_c = 1.0$, with b_α chosen to yield $\Pr(H_0) = 0.5$ and to correspond to a moderately vague specification in which the posterior distributions should be driven primarily by the data.

Figure 1 shows the results for the Poisson data simulations, with rows 1-3 corresponding to sample sizes $n = 25, 50, 100$, respectively, and with columns 1-3 corresponding to functions (i)-(iii). Points represent the simulated data values, the x-axis ranges from 0 to 1, and the y-axis ranges from 0 to 6. The dark solid lines represent pointwise posterior means of $f(x)$ and the dashed lines represent pointwise 95% credible intervals. For purposes of comparison, the unrestricted frequentist estimates for generalized additive models (GAM) fitted in S-PLUS using the gam function are also shown (dotted lines). In each case, the posterior mean provides a good estimate of the true curve, which essentially follows the GAM estimate except when the GAM estimate violates the non-decreasing constraint. Hence, as the same prior was used in each of these analyses, the proposed approach provides a smooth approximation, which is driven primarily by the data subject to the non-decreasing constraint.

Figure 2 shows the corresponding results for the Bernoulli data simulations. Again, the proposed approach provides good estimates of the true function in each case. Although the prior assigns 0.5 prior probability to the null hypothesis of no association, the posterior estimates appear to be driven primarily by the data subject to the non-decreasing constraint. In particular, the estimates consistently follow the trajectory of the frequentist GAM esti-

mates, but violations of the non-decreasing constraint are smoothed out. This suggests that the approach of using hyperprior distributions to ensure robustness to the prior specification was successful.

Table 1 presents estimated posterior probabilities of H_0 for each of the simulated examples. For simulations in which the true function was flat, the estimated posterior probability of the null hypothesis ranged from 0.18 to 0.59, values that are not low enough to suggest rejecting H_0 in favor of H_1 . In contrast, several of the simulations under the linear and threshold functions had low posterior probabilities of H_0 , though in many cases the results were inconclusive, which is as expected since sample sizes are small and the rates of increase modest. We purposely chose to focus on such cases to evaluate the degree of estimation bias in the presence of weak evidence in the data of an increasing function.

5. Discussion

This article has proposed a new strategy for assessing the association between a continuous predictor and a discrete response incorporating prior information that the regression function is non-decreasing. Key features distinguishing the proposed approach from earlier work on isotonic regression include the ability to assign positive probability to the null hypothesis of no association and to account for flat regions in the response function, while borrowing information about local changes in the function across different regions of the predictor. In many applications, there is interest in assessing whether the function is flat or increasing for values of the predictor falling in predefined categories. For example, one may be interested in whether the probability of a disease increases with body mass index overall and within the categories: normal, overweight, and obese.

From the perspective of multiple testing, the proposed prior structure has several appealing properties. First, unlike the Bayesian-Bonferroni strategy of inflating the prior probability of local null hypotheses under the assumption of *a priori* independence, we account for *a*

priori dependency in the local hypotheses through the use of a carefully-defined hyperprior distribution. This prior distribution has a single parameter to be chosen by the investigator, which controls the degree of positive correlation in the hypotheses. In general, under this prior specification, the penalty for multiple testing is less severe than that induced by the Bayesian-Bonferroni approach. In addition, inferences tend to be driven much more by the data, since the use of a hyperprior allows the data to inform about the degree of shrinkage towards the null hypothesis.

Local hypotheses that the regression function is flat within particular regions of the predictor essentially correspond to point null hypotheses, which have been the focus of a considerable amount of research in the Bayesian literature. We follow Westfall et al. (1997), Gopalan and Berry (1998), and Gönen, Westfall, and Johnson (2003) in treating the null hypotheses as if they could potentially be true (see George and McCulloch, 1993; Carlin and Chib, 1995; and Geweke, 1996 among others for a related view on the variable selection problem). Although one may not believe that a point null can be exactly true, this treatment results in a convenient approximation (Berger and Delampady, 1987). In the literature, point nulls are routinely considered to be independent *a priori*, though Gönen, Westfall, and Johnson (2003) proposed an approach that accounts for dependency in the setting of two-sample multivariate normal data.

An appealing feature of our prior is the ability to borrow information across different regions of the predictor with respect to the rate of increase in the function during increasing regions, which results in smoothing of the function. Also, the conditionally-conjugate form makes posterior computation feasible even though the number of parameters used to characterize the regression function increases with the sample size. These simplifications should also hold in other cases, such as for isotonic regression of survival data.

References

- Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association* **84**, 289-294.
- Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science* **2**, 317-352.
- Carlin, B. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **57**, 473-484.
- Doksum, K.A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability* **2**, 183-201.
- Doveh, E., Shapiro, A., and Feigin, P.D. (2002). Testing of monotonicity in parametric regression models. *Journal of Statistical Planning and Inference* **107**, 289-306.
- Dunson, D.B. and Herring, A.H. (2003). Bayesian inferences in the Cox model for order restricted hypotheses. *Biometrics*, in press.
- Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209-230.
- Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615-629.
- George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.
- Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds), 609-620. Oxford: Clarendon Press.

- Ghosal, S., Sen, A., and van der Vaart, W. (2000). Testing monotonicity of regression. *Annals of Statistics* **28**, 1054-1082.
- Gönen, M., Westfall, P.H., and Johnson, W.O. (2003). Bayesian multiple testing for two-sample multivariate endpoints. *Biometrics* **59**, 76-82.
- Gopalan, R. and Berry, D.A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association* **93**, 1130-1139.
- Hall, P. and Huang, L.S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Annals of Statistics* **29**, 624-647.
- Holmes, C.C. and Heard, N.A. (2003). Generalized monotonic regression using random change points. *Statistics in Medicine* **22**, 623-638.
- Kalbfleisch, J.D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society B* **40**, 214-221.
- Lee, C.I.C. (1996). On estimation for monotone dose-response curves. *Journal of the American Statistical Association* **91**, 1110-1119.
- Mammen, E. (1991). Estimating a smooth monotone regression function. *Annals of Statistics* **19**, 724-740.
- Mammen, E., Marron, J.S., Turlach, B.A., and Wand, M.P. (2001). A general projection framework for constrained smoothing. *Statistical Science* **16**, 232-248.
- Morton-Jones, T., Diggle, P., Parker, L., Dickinson, H.O., and Binks, K. (2000). Additive isotonic regression models in epidemiology. *Statistics in Medicine* **19**, 849-859.
- Neelon, B. and Dunson, D.B. (2003). Bayesian isotonic regression and trend analysis. *Institute of Statistics and Decision Sciences Discussion Paper* **03-15**, Duke University.

Ramsay, J.O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society B* **60**, 365-375.

Shiboski, S.C. (1998). Generalized additive models for current status data. *Lifetime Data Analysis* **4**, 29-50.

Wang, Z. (2000). An algorithm for generalized monotonic smoothing. *Journal of Applied Statistics* **27**, 495-507.

Westfall, P.H., Johnson, W.O. and Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84**, 419-427.

Table 1

Estimated posterior probabilities of H_0 in each of the simulations.

true $f(x)$	Outcome type	n	$\widehat{\Pr}(H_0 \mid \text{data})$
(i) linear	count	25	0.16
	count	50	0.24
	count	100	0.01
	binary	25	0.32
	binary	50	0.06
	binary	100	0.43
(ii) flat	count	25	0.53
	count	50	0.46
	count	100	0.59
	binary	25	0.23
	binary	50	0.18
	binary	100	0.36
(iii) threshold	count	25	0.15
	count	50	0.26
	count	100	0.01
	binary	25	0.27
	binary	50	0.17
	binary	100	0.46

Figure 1. Nonparametric isotonic curve estimates for Poisson data. Dark solid line is the estimated posterior mean, dashed lines are pointwise 95% credible intervals, light solid line is the true curve, and dotted line is the unrestricted frequentist GAM estimate.

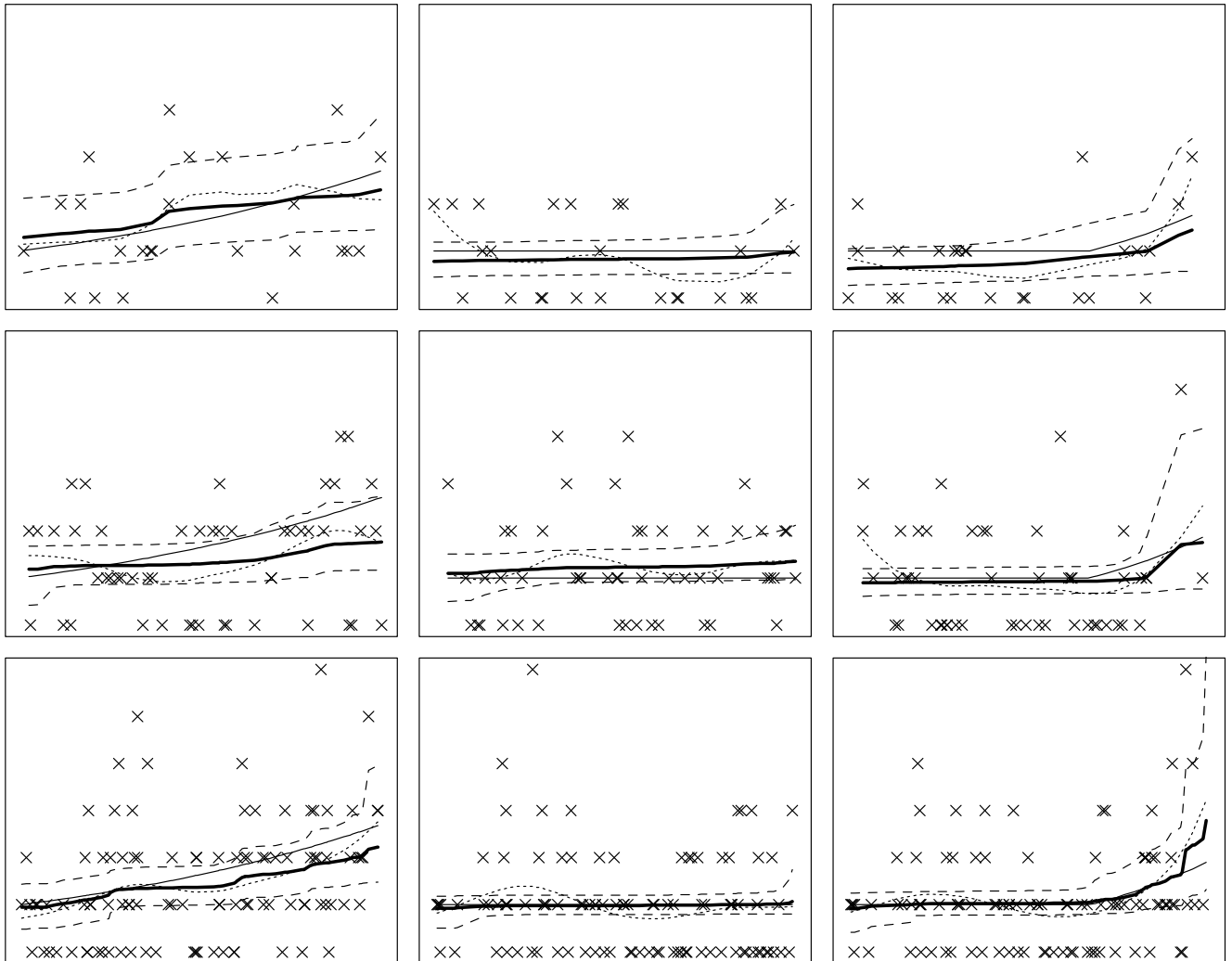


Figure 2. Nonparametric isotonic curve estimates for Bernoulli data. Dark solid line is the estimated posterior mean, dashed lines are pointwise 95% credible intervals, light solid line is the true curve, and dotted line is the unrestricted frequentist GAM estimate.

