Shrinkage priors for isotonic probability vectors and binary data modeling

Philip S. Boonstra[1]*, Daniel R. Owen[2], and Jian Kang[1]

[1] Department of Biostatistics, University of Michigan, USA
[2]Department of Radiation Oncology, Ann Arbor, Michigan, U.S.A.

Abstract

This paper outlines a new class of shrinkage priors for Bayesian isotonic regression modeling a binary outcome against a predictor, where the probability of the outcome is assumed to be monotonically non-decreasing with the predictor. The predictor is categorized into a large number of groups, and the set of differences between outcome probabilities in consecutive categories is equipped with a multivariate prior having support over the set of simplexes. The Dirichlet distribution, which can be derived from a normalized cumulative sum of gamma-distributed random variables, is a natural choice of prior, but using mathematical and simulation-based arguments, we show that the resulting posterior can be numerically unstable, even under simple data configurations. We propose an alternative prior motivated by horseshoe-type shrinkage that is numerically more stable. We show that this horseshoe-based prior is not subject to the numerical instability seen in the Dirichlet/gamma-based prior and that the posterior can estimate the underlying true curve more efficiently than the Dirichlet distribution. We demonstrate the use of this prior in a model predicting the occurrence of radiation-induced lung toxicity in lung cancer patients as a function of dose delivered to normal lung tissue.
keywords: Dirichlet; gamma distribution; horseshoe; monotone

1 Introduction

Isotonic regression is a type of constrained modeling that imposes a monotonic, i.e. non-decreasing or non-increasing, assumption on the fitted functional relationship between a predictor and the expected outcome without making assumptions about the specific form of that relationship (Barlow et al., 1972). Except for the ordering constraint, isotonic regression is otherwise less presumptive than standard generalized linear models (GLMs). This is especially true in the case of modeling a binary outcome, where GLMs fit to a continuous predictor yield predicted probabilities of 0 and 1 for extreme enough values of the predictor, regardless of the observed data configuration.

Isotonic assumptions are common when modeling dose-response or dose-toxicity curves. For example, when planning radiation treatment for lung cancer, it is of interest to maximize radiation

*1415 Washington Heights, Ann Arbor, Michigan, USA, 48109-2029; Tel: +1 734 615 1580; philb@umich.edu

delivered to the tumor while limiting the amount of radiation delivered to the surrounding normal lung tissue. Typical radiation treatment plans generally assume all normal lung tissue is functionally equal, but in most patients, some parts of the lung will be lower functioning, e.g. due to emphysema or tumor burden. Based on the hypothesis that accounting for the functionality of normal lung tissue in treatment planning could potentially allow for a higher dose of radiation delivered to the tumor without increasing the risk of toxicity, Owen et al. (2020) conducted a retrospective analysis that assessed the correlation between *ventilation/perfusion single photon emission computed tomography* ("SPECT V/Q ") lung function metrics and the incidence of radiation-induced lung toxicity (RILT). The treatment plan actually used in these patients was not based upon lung function; however, as part of an IRB-approved clinical trial, patients received a SPECT V/Q scan prior to starting radiation treatment, which Owen et al. (2020) used to quantify lung function and subsequently the dose of radiation delivered to low-, moderate-, and high-functioning lung. Fitting a logistic regression, they found that the percent of a patient's low-functioning lung tissue receiving more than 20 Gray (Gy) of radiation (LF20), predicted grade 2+ RILT. See Owen et al. (2018) and Owen et al. (2020) for complete details. Although it can be reasonably assumed that the risk of RILT increases with LF20, given the novelty of this dosimetric, the exact nature of this dose-toxicity curve, e.g. where it begins to meaningfully increase and how it is shaped, is not known. An isotonic approach to this problem is therefore well-suited.

There are a number of Bayesian variants of isotonic regression, including piecewise linear functions with autoregressive mixture priors (Neelon & Dunson, 2004), restricted splines (Shively et al., 2009), and Gaussian process projections (Lin & Dunson, 2014). Moreover, several approaches have been developed specifically in the context of dose-response studies, including Bayesian model averaging (Ohlssen & Racine, 2015), a Bayesian isotonic regression dose-response model (BIRD) approach tailored to estimate important clinical parameters from the dose-response curve (Li & Fu, 2017), discrete mixture of parametric distribution functions (Bornkamp & Ickstadt, 2009), and isotonic regression for dose-schedule finding, in which both the dosage and the frequency of administration are being explored Li et al. (2008).

Our proposed approach is qualitatively different from these methods and is best described as an extension and generalization of a more classical Bayesian isotonic regression first proposed in Ramsey (1972) and later extended in e.g. Shaked & Singpurwalla (1990), Gelfand & Kuo (1991), and Ramgopal et al. (1993). In these papers, given a binary outcome and a categorical predictor, the set of differences between probabilities of the outcome in contiguous categories, also called the set of increments, is modeled using as a prior the Dirichlet distribution, equipped with a vector of concentration hyperparameters.

The Dirichlet distribution is likely the most well-known distribution having support over the unit simplex, making it a natural choice of prior in this context. However, to allow for potentially very small jumps in the fitted isotonic regression curve, the concentration hyperparameters must be correspondingly very small. In such a setting, the fitted model may be susceptible to numerical challenges – specifically, underflow. The cause of this shortcoming is illustrated by noting that the Dirichlet distribution can be equivalently characterized as a set of normalized gamma random variables, with the Dirichlet concentration hyperparameters becoming gamma shape parameters. When the shape parameter is close to zero, it becomes extremely difficult to sample gamma random variables accurately (e.g., Liu et al., 2013). In this paper, we propose a new isotonic prior for the binary outcome setting that is computationally robust and statistically efficient.

With this background, the objectives of this manuscript are (i) to mathematically elucidate the

above claim; (ii) to propose a novel alternative prior also having support over the unit simplex; and (iii) to show, both mathematically and through simulation, that our alternative prior is computationally robust to this issue of numerical underflow. As we will show, the key difference is that the Dirichlet/gamma prior results in a posterior distribution with mass increasingly concentrating in a neighborhood around zero as the shape parameter becomes smaller. Our novel prior, which avoids this undesirable behavior, is called the 'isotonic horseshoe prior' because it derives from the horseshoe distribution, which has been used as a continuous shrinkage prior in many regression contexts (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2015, 2017a,b). Bayesian estimates using a horseshoe prior satisfy the three desiderata of a shrinkage estimator outlined in Fan & Li (2001), namely, near-unbiasedness, sparsity, and continuity (this last characteristic meaning that the estimates are not sensitive to small changes in the data). To our knowledge, it has not yet been applied to the isotonic regression context.

The remainder of this paper is organized as follows. In Section 2, we present the classical Bayesian isotonic regression for binary outcomes considered in Shaked & Singpurwalla (1990). We then present our alternative prior that extends the horseshoe prior and give our main results, which are (i) that the horseshoe density always diverges to infinity as its hyperparameter approaches zero at a slower rate than the gamma density and (ii) that the posterior distribution based on our horseshoe prior does not concentrate around zero, whereas for the gamma prior it does. This difference impacts both the statistical efficiency of these isotonic regressions as well as the computational efficiency, which we demonstrate both through a simulation study (Section 3) and the re-analysis of our motivating lung cancer data (Section 4). We conclude with a discussion in Section 5.

2 Methods

We begin with notations in the article. Let $\text{Bernoulli}(\pi)$ represent a Bernoulli distribution with probability π. Let $\text{Gamma}(a, b)$ represent a gamma distribution with shape a and rate b. Let $\text{Dirichlet}(\boldsymbol{s})$ be a Dirichlet distribution with concentration parameter vector $\boldsymbol{s}$. Let $\text{Beta}(a, b)$ represent a beta distribution with shapes a and b. Denote by $I(\mathcal{A})$ an event indicator where $I(\mathcal{A}) = 1$ if event $\mathcal{A}$ occurs and $I(\mathcal{A}) = 0$, otherwise. Denote by $\Pr(\cdot)$, $E(\cdot)$ and $\text{Var}(\cdot)$ the probability measure, the expectation operator and the variance operator, respectively. For any two continuous functions $f(x)$ and $g(x)$ with $\lim_{x\downarrow 0} f(x) = \lim_{x\downarrow 0} g(x) = \infty$, write $f(x) = o\{g(x)\}$ if $\lim_{x\downarrow 0}\{f(x)/g(x)\} = 0$ and $f(x) = O(g(x))$ if there exists $M > 0$ such that $\lim_{x\downarrow 0} |f(x)/g(x)| = M$. For a smooth function $f(x)$, let $f'(x)$ and $f''(x)$ represent the first and second derivatives of f, respectively.

Suppose the data consists of n observations. Let $i (i = 1, \ldots, n)$ be the index of the observations. Let Y_i denote a binary outcome taking the value 0 or 1. Let X_i be the corresponding predictor, which is an ordered categorical variable taking one of the values in $\{1, \ldots, K\}$. Given the predictors, the binary outcomes are independent and marginally follow a Bernoulli distribution, i.e.

$$[Y_i \mid X_i] \sim \text{Bernoulli}\left\{\sum_{j=1}^{K} \xi_j I(X_i = j)\right\}, \tag{1}$$

where $\xi_j = \Pr(Y_i = 1 \mid X_i = j) \in [0,1]$, for $j = 1, \ldots, K$. Write $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^T$, which is the parameter vector of interest. We impose the monotonic non-decreasing assumption on $\boldsymbol{\xi}$, i.e., $0 \leq \xi_1 \leq \xi_2 \leq \cdots \leq \xi_K \leq 1$. This implies that the set of increments $\{\xi_j - \xi_{j-1}\}_{j=1}^{K+1}$, where $\xi_0 \equiv 0$ and $\xi_{K+1} \equiv 1$, forms a simplex. We refer to $\boldsymbol{\xi}$ as the isotonic probability vector (IPV). To specify the priors, we represent $\boldsymbol{\xi}$ as a function of $K+1$ non-negative parameters using an IPV transformation. Let $\mathbb{P}^K = \{(\pi_1, \ldots, \pi_K) : 0 \leq \pi_j \leq \pi_k \leq 1, \text{ for } 1 \leq j < k \leq 1\}$ be a K-dimensional IPV space. Let $\bar{\mathbb{R}}_+^{K+1} = \{(a_1, \ldots, a_{K+1}) : a_j \geq 0, \text{ for } 1 \leq j \leq K+1\}$ be a $K+1$-dimensional non-negative Euclidean space.

Definition 1 *The K-dimensional isotonic probability vector transformation is a function mapping from the non-negative $(K+1)$-dimensional Euclidean space onto the K-dimensional probability vector space. In particular, we have $\boldsymbol{F} : \bar{\mathbb{R}}_+^{(K+1)} \to \mathbb{P}^K$, that is, for any $\boldsymbol{a} = (a_1, \ldots, a_{K+1})^T \in \bar{\mathbb{R}}_+^{(K+1)}$ and $\boldsymbol{a} \neq \boldsymbol{0}$,*

$$\boldsymbol{F}(\boldsymbol{a}) = \{F_1(\boldsymbol{a}), \ldots, F_K(\boldsymbol{a})\}^T \in \mathbb{P}^K \quad \textit{with} \quad F_j(\boldsymbol{a}) = \frac{\sum_{k=1}^{j} a_k}{\sum_{k=1}^{K+1} a_k}, \quad \textit{for } j = 1, \ldots, K \tag{2}$$

With this transformation, we represent the IPV in (1) as

$$\boldsymbol{\xi} = \boldsymbol{F}(\boldsymbol{\alpha}) = \left(\sum_{j=1}^{K+1} \alpha_j\right)^{-1} \left(\alpha_1, \alpha_1 + \alpha_2, \ldots, \sum_{j=1}^{K} \alpha_j\right)^T, \tag{3}$$

where $\boldsymbol{\alpha} \in \bar{\mathbb{R}}_+^{K+1} \backslash \{0\}^{K+1}$. We then can construct the prior for the isotonic probability vector by placing priors on $\boldsymbol{\alpha}$ with support restricted to $\bar{\mathbb{R}}_+^{(K+1)} \backslash \{0\}^{K+1}$, that is, the $K+1$ dimensional space of non-negative numbers excluding the origin. In this article, we mainly discuss two types of priors: the gamma isotonic probability vector (GAIPV) distribution and the horseshoe isotonic probability vector (HSIPV) distribution

2.1 Gamma isotonic probability vector distribution

Following Ramsey (1972), a member of this family arises from placing independent gamma priors on $\boldsymbol{\alpha}$ with a set of positive shape parameters and a common positive rate parameter. We define the GAIPV distribution as follows:

Definition 2 *Suppose $\alpha_j \sim \text{Gamma}(s_j, 1)$ with $s_j > 0$ for $j = 1, \ldots, K+1$ and $\alpha_1, \ldots, \alpha_{K+1}$ are mutually independent, then $\boldsymbol{F}(\boldsymbol{\alpha})$ follows a K-dimensional gamma isotonic probability vector distribution. Denote $\boldsymbol{F}(\boldsymbol{\alpha}) \sim \text{GAIPV}(\boldsymbol{s})$ where $\boldsymbol{s} = (s_1, \ldots, s_{K+1})^T$.*

The GAIPV is a well-defined distribution since the gamma distribution has support on the positive numbers. Thus there is no probability mass on all α_j being zero-valued. In addition, we have the follow properties.

Proposition 1 *Let $\boldsymbol{s} = (s_1, \ldots, s_{K+1})^T$, $\tilde{n} = \sum_{k=1}^{K+1} s_k$ and $\tilde{p}_j = \sum_{k=1}^{j} s_k / \tilde{n}$, for $j = 1, \ldots, K+1$. We have the following results:*

a. *For any $r > 0$, if $\alpha_j(r) \sim \text{Gamma}(s_j, r)$ and $\alpha_1(r), \ldots, \alpha_{K+1}(r)$ are mutually independent, then $\boldsymbol{F}\{\boldsymbol{\alpha}(r)\} \sim \text{GAIPV}(\boldsymbol{s})$.*

b. *For any length-K probability vector $\boldsymbol{\pi}$, then $\boldsymbol{\pi} \sim \text{GAIPV}(\boldsymbol{s})$ if and only if $\{\pi_j - \pi_{j-1}\}_{j=1}^{K+1} \sim \text{Dirichlet}(\boldsymbol{s})$, where π_0 is defined to be 0 and π_{K+1} is defined to be 1.*

c. *If $\boldsymbol{\xi} \sim \text{GAIPV}(\boldsymbol{s})$, then the marginal distribution $\xi_j \sim \text{Beta}\{\tilde{p}_j\tilde{n}, (1-\tilde{p}_j)\tilde{n}\}$ for $j = 1, \ldots, K$. This further implies that*

$$E(\xi_j) = \tilde{p}_j, \quad \textit{and} \quad \text{Var}(\xi_j) = \frac{\tilde{p}_j(1-\tilde{p}_j)}{\tilde{n}+1}. \tag{4}$$

When we assign the GAIPV prior to $\boldsymbol{\xi}$, we ensure the isotonic ordering in $\boldsymbol{\xi}$ with prior probability one. Although the posterior of ξ_j will, for $K > 1$, no longer be a beta distribution, the isotonic ordering guarantees that any individual draw from the joint posterior distribution will satisfy $\xi_1 < \cdots < \xi_K$. One approach for eliciting values of $\boldsymbol{s}$ would be to specify anticipated outcome probabilities satisfying the monotonic relationship, $0 \equiv \tilde{p}_0 < \tilde{p}_1 < \cdots < \tilde{p}_K < \tilde{p}_{K+1} \equiv 1$, and an effective number of historical observations, $\tilde{n}$, and match quantities to the first two moments of the underlying beta distribution: solving for the hyperparameters, this gives $s_j = \tilde{n}(\tilde{p}_j - \tilde{p}_{j-1})$ for $j = 1, \ldots, K+1$.

When there is little or no prior information on the probabilities in each category, an agnostic prior elicitation approach would assume that each individual increment is likely to be equally sized, i.e. $s_j \equiv s = \tilde{n}/(K+1)$ and set $\tilde{n}$ to be very small, reflecting a lack of actual prior information on the probabilities themselves. The resulting posterior will tend to have very small increases between consecutive categories but will be able to adapt to large increases where the data warrant.

The gamma distribution has the following properties:

Proposition 2 *Suppose $\alpha \sim \text{Gamma}(s, 1)$. Let $g(x)$ be the density function of α.*

a. *If $0 < s < 1$, then $g(x) = O(x^{s-1})$ and $|g'(x)| = O(x^{s-2})$ when $x \downarrow 0$.*

b. *For any $\kappa \in (0, 1)$, as $s \to 0$,*

$$\Pr\{\alpha \leq \exp\left(-s^{-\kappa}\right)\} \to 1, \tag{5}$$

This proposition implies when s gets close to zero, then with prior probability tending to one α_j takes very small values from the interval $(0, \exp\{-s^{-\kappa}\})$ of which length also goes to zero. For example, taking $\kappa = 0.95$ and $s = 0.001$, then $\exp(-s^{-\kappa}) < 3.5 \times 10^{-308}$. Although this prior shrinkage is a desirable statistical feature, Proposition 2 has negative numerical implications, as it is very difficult to accurately sample from a gamma distribution with a small shape parameter (e.g., Liu et al., 2013). To illustrate: mathematically, for a fixed $\tilde{n}$, at least one s_j will be at most $\tilde{n}/K$, which is 0.01 if, for example, $\tilde{n} = 0.5$ and $K = 50$. There is a 0.70 probability that a randomly sampled gamma-distributed random variable with shape 0.01 and rate 1 will be less than the machine precision of a 64-bit processor (namely $\epsilon \approx 2.2 \times 10^{-16}$), and a 0.10 probability that it will be less than 10^{-100}. This characteristic tends to cause numerical underflow when sampling from the prior, and it carries through to the posterior distribution of α_j, as stated in Theorem 1 below. Let $P_{\theta_0}^n$ be the actual distribution of data Y given the true parameter θ_0.

Theorem 1 *For any given $\kappa \in (0, 1)$, for any $j = 1, \ldots, K$, we have the following results:*

a. *For any* $n \geq 1$*, as* $s \to 0$*,*

$$\Pr\left\{\alpha_j \leq \exp\left(-s^{-\kappa}\right) \mid Y, X\right\} \to 1 \tag{6}$$

in $P_{\theta_0}^n$ *probability.*

b. *For any fixed* $M > 0$*, set* $s = (Mn)^{-1/\kappa}$*. when* $n \to \infty$*,*

$$\Pr\left\{\alpha_j \leq \exp(-Mn) \mid Y, X\right\} \to 1 \tag{7}$$

in $P_{\theta_0}^\infty$ *probability.*

In words, Theorem 1 gives that, regardless of data configuration, all of the posterior mass of α_j is bounded above by a number that goes to zero as the shape parameter decreases towards zero. This upper bound is expressed both as a function of the shape parameter itself (part a) or as function of the sample size n when the shape parameter $s = O(n^{-1/\kappa})$ (part b).

When model fitting using Hamiltonian Monte Carlo, the presence of numerical underflow is indicated by divergences from the Hamiltonian, which can be detected and therefore used as a model diagnostic (Section 14.5, Stan Reference Manual, Stan Development Team, 2019). Detection of such 'divergent transitions' indicates that the posterior distribution is not being fully explored and that the resulting inference may be biased (Betancourt, 2016). In the presence of divergences, the algorithm will adaptively and potentially dramatically decrease the step size of the proposal such that it appears to stop at an iteration, and thus another symptom can be long running times and/or extreme variation in total running time between independent chains of the sampler. Because the source of these divergences is readily identifiable in the present context, a simplistic workaround to prevent this underflow is to truncate the lower support of the gamma prior at some very small, positive number. In our simulation study below, we present results using three choices of this lower truncation on α_j, equal to and less than ϵ. As we will see, when this lower truncation is less than ϵ, both the typical number of divergent transitions and the total running time of the posterior sampler increase dramatically.

2.2 Horseshoe isotonic probability vector distribution

We consider an alternative and potentially better solution to the problem of underflow by modifying the so-called 'regularized horseshoe prior' (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2015, 2017a,b). For reasons presented below, this prior – even untruncated – does not encounter the same numerical difficulties as the gamma-based prior. We first construct the half horseshoe distribution for $K+1$ non-negative random variables. Let $\mathrm{N}^+(0, \sigma^2)$ be a half normal distribution with standard deviation σ and its density function proportional to $\exp\{-x/(2\sigma^2)\}I(x \geq 0)$. Let $\mathrm{C}^+(0,1)$ be the standard half-Cauchy distribution with the density function proportional to $(1+x^2)^{-1}$.

Definition 3 *Let* $c(c > 0)$ *be a constant. For* $j = 1, \ldots, K+1$*,*

$$[\alpha_j \mid \tau, \lambda_j\,] \sim \mathrm{N}^+\left(0, \frac{c^2\tau^2\lambda_j^2}{1 + c^2\tau^2\lambda_j^2}\right), \quad \lambda_j \sim \mathrm{C}^+(0,1), \quad \tau \sim \mathrm{C}^+(0,1). \tag{8}$$

Then $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{K+1})^T$ *follow a* $K+1$*-dimensional half-horseshoe distribution with a parameter* c*. Denote* $\boldsymbol{\alpha} \sim \mathrm{HS}^+(c)$*. Furthermore, we say* $\boldsymbol{F}(\boldsymbol{\alpha})$ *follows a* K*-dimensional horseshoe isotonic probability vector distribution with parameter* c*. and denote* $\boldsymbol{F}(\boldsymbol{\alpha}) \sim \mathrm{HSIPV}(c)$*.*

The half horseshoe distribution serves as a hierarchical shrinkage prior in a Bayesian model, in which the global shrinkage parameter τ controls overall shrinkage to zero, and the local shrinkage parameters λ_j can be large to offset this overall shrinkage as warranted by the data. The value of c in (8) is a user-supplied hyperparameter, the choice of which codifies an implicit assumption about the anticipated number of non-zero elements in $\boldsymbol{\alpha}$, or, in our case, non-zero jumps in the probability curve, with larger values corresponding to anticipating more non-zero elements (Piironen & Vehtari, 2017a).

Let $\theta_j \equiv (1 + 1/[c^2\tau^2\lambda_j^2])^{-1/2}$ denote the prior standard deviation of α_j given τ and λ_j in (8). The constant value 1 that is added to $1/[c^2\tau^2\lambda_j^2]$ in the expression for θ_j serves a two-fold purpose. First, it dominates the expression when $c^2\tau^2\lambda_j^2$ is very large, making $\theta_j \approx 1$ in such case and thinning the heavy tails that would otherwise result if no constant were added (Piironen & Vehtari, 2017b). Second, adding a constant identifies the parameters c and τ. If a constant was not added, θ_j would reduce to $c\tau\lambda_j$, and the expression $c\tau$ would cancel out in the numerator and denominator of equation (2). Piironen & Vehtari (2017a) proposed a more general recipe by adding $1/d^2$ instead of a constant 1, with d either a fixed constant or a hyperparameter with a hyperprior of its own. However, in our unique extension of the horseshoe prior, d and c cannot both independently vary, again, due to relative nature of each α_k in equation (2). Thus, we set $d = 1$. Also different from previous versions of the horseshoe prior, our formulation places a half normal prior on each α_j, with support only on the positive half of the real line, to match the context of the problem. To summarize, these modifications are unique to our application of the horseshoe prior to this problem and arise from the relative relationship of the scale parameters.

In the same way that gamma prior requires specifying one fixed value s, our implementation of the horseshoe prior in equations (8) also has only one fixed value, namely c, that is user-supplied or selected. Following Piironen & Vehtari (2017a), c can be derived by specifying an anticipated number of non-zero α_j's, $\tilde{m}_{\text{eff}}$, which here is interpreted as the anticipated number of jumps in the probability curve. This can be written as an approximate expression of c and n vis a vis

$$\tilde{m}_{\text{eff}} = E\left\{\sum_{j=1}^{K+1}\left(1 + 4n^{-1}\theta_j^{-2}\right)^{-1}\right\}, \tag{9}$$

where the expectation is with respect to the priors $\pi(\tau)$ and $\pi(\lambda_j)$. Consequently one can numerically solve for the value of c that corresponds to an anticipated number of jumps in the probability curve. In addition to the aforementioned horseshoe references, we have previously used a version of this shrinkage prior in Boonstra & Barbaro (2018), and the reader may refer there for additional details.

To better understand of the theoretical properties of the HSIPV distribution, we first study the half-horseshoe distribution and obtain the following results

Proposition 3 *Suppose* $\boldsymbol{\alpha} \sim \text{HS}^+(c)$. *When* $0 < c < 1/2$, *we have the following inequalities for the marginal mean and variance of* α_j*:*

$$\frac{c}{\sqrt{2\pi}} \leq \text{E}(\alpha_j) \leq 2\sqrt{\frac{c}{\pi}} \tag{10}$$

$$\text{Var}(\alpha_j) \leq \sqrt{2c} - \frac{c^2}{2\pi} \tag{11}$$

Analogous to Proposition 2 applied to the GAIPV distribution is Proposition 4 applied to the HSIPV distribution.

Proposition 4 *Suppose* $\boldsymbol{\alpha} \sim \mathrm{HS}^{+}(c)$ *for a constant* $c > 0$ *and let* $h(x)$ *be the marginal density of* α_j*, for* $j = 1, \ldots, K+1$*.*

a. *For any* $c > 0$*,* $h(x) \leq O(x^{-1})$ *and* $|h'(x)| = O\{-x^{-1}\log(x)\}$ *when* $x \downarrow 0$*.*

b. *For any* $\kappa, \upsilon \in (0,1)$*, as* $c \to 0$*,*

$$\Pr(c^{1/\kappa} < \alpha_j \leq c^{\upsilon}) \to 1. \tag{12}$$

Thus the rate at which the horseshoe-based density approaches infinity is always slower than the gamma-based density. More rigorously, let $g(x)$ and $h(x)$ be defined as in Propositions 2a and 4a, respectively. Then, from these propositions, we have that

$$\lim_{x \downarrow 0} \frac{|h'(x)|}{|g'(x)|} = \lim_{x \downarrow 0} \frac{-x^{-1}\log(x)}{x^{s-2}} = \lim_{x \downarrow 0} \frac{-\log(x)}{x^{s-1}} = \lim_{x \downarrow 0} \frac{x^{-1}}{x^{s-2}} = 0$$

In words, the rate at which the horseshoe density approaches infinity as x approaches zero from above is less than the rate at which the gamma density does so, and this holds for any $c > 0$ and $0 < s < 1$. As we will see in Section 3, the horseshoe density approaches infinity slowly enough to effectively resolve the divergences encountered by the gamma-based prior. When x is small, the values of s that make $|g'(x)|$ "most similar" to $|h'(x)|$, namely s just less than 1, will, for all intents and purposes, be no different than setting s equal to 1, at which point the gamma density is finite at $x = 0$.

We also have an analog to Theorem 1:

Theorem 2 *For any given* $\kappa, \upsilon \in (0,1)$*,* $j = 1, \ldots, K$*, then the following hold:*

a. *For any* $n \geq 1$*, as* $c \to 0$*,*

$$\Pr(c^{1/\kappa} < \alpha_j \leq c^{\upsilon} \mid Y, X) \to 1 \tag{13}$$

in $P_{\theta_0}^{n}$ *probability.*

b. *For any fixed* $M > 0$*, set* $c = (Mn)^{-1/\upsilon}$ *when* $n \to \infty$*, then*

$$\Pr(Mn^{-1/(\kappa\upsilon)} < \alpha_j \leq Mn^{-1} \mid Y, X) \to 1 \tag{14}$$

in $P_{\theta_0}^{\infty}$ *probability.*

Theorem 2 bounds the mass of the HSIPV-based posterior above and below, thus ensuring that it does concentrate near zero, which stands in contrast to the GAIPV-based posterior in Theorem 1.

2.3 HSIPV versus GAIPV

Both the HSIPV and GAIPV priors induce similar behavior in the posterior: there is substantial prior mass close to zero accompanied by relatively heavy tails. This first characteristic allows for the fitted model to force any given α_j to be very close to zero, yielding small increments between consecutive groups, and the heavy tails still admit large values of α_j. Interestingly, however, even though the un-truncated versions of both densities approach infinity as α_j approaches zero, the regularized horseshoe prior does not encounter the numerical underflow issues of its gamma counterpart.

The parameter c in the horseshoe prior reflects an assumption about how many α_j's are non-zero. In this way, its role is analogous to the parameter s in the gamma prior. However, c *scales* the horseshoe density and does not impact the limiting behavior of the slope of the density as $x \downarrow 0$, whereas s *shapes* the gamma density and has a substantial impact on this limiting behavior, potentially introducing numerical difficulties when actually fitting the model. Thus, it is not necessary to truncate the horseshoe prior at some small positive constant.

3 Simulation-based comparison of HSIPV to GAIPV

Here we empirically assess the computational and statistical implications of the above results using numerical studies. We divide this assessment into three subsections. The first subsection fixes a single simple dataset and demonstrates the implications of Theorems 1 and 2. The second subsection is also a fixed-data evaluation, now using three more complex datasets, and compares the priors with respect to diagnostics of the MCMC algorithm itself. Finally, in the third subsection, we fix the data-generating mechanism and assess the priors with respect to statistical metrics of predictive ability.

3.1 Fixed-data evaluation 1

Theorems 1 and 2 establish probabilistic bounds on the posterior distribution of each α_j as the hyperparameters (s for GAIPV or c for HSIPV) go to zero. The theorems hold in general for any data configuration, including "well-behaved" datasets. Here we consider a dataset comprised of $n = 50$ observations of a binary outcome Y, all from a single category, i.e. $K = 1$, with exactly half of the observations having $Y = 1$ and half having $Y = 0$. With one category, this reduces to an inference on a single proportion, namely $\xi_1 \equiv \Pr(Y = 1)$, and a sensible inference should give that ξ_1 is close to 0.5. The IPV-based priors are placed on the length-two parameter $\boldsymbol{\alpha}$, where $\xi_1 = \alpha_1/(\alpha_1 + \alpha_2)$. We consider the GAIPV and HSIPV priors with hyperparameters decreasing from $1/2$ to $1/128$ by multiplicative scales of $1/4$. The shape parameter of the GAIPV prior is not interpreted equivalently to the scale parameter of the HSIPV prior; however, we are interested here in the relative change in the posterior as each hyperparameter goes to zero.

For each prior, we estimate the posterior distribution using Hamiltonian Monte Carlo as implemented in the STAN programming language (Stan Development Team, 2018, 2019). Each fitted model contains two independent chains running the No-U-Turn Hamiltonian Monte Carlo sampler for 5000 iterations each, with the first 2500 iterations discarded. The posterior medians of all parameters are in Table 1, as is a 95% credible interval for ξ_1.

Table 1: Posterior medians of parameters for the GAIPV and HSIPV priors as the hyperparameters decrease. The data are n observations of a binary outcome Y with $\bar{Y} = 1/2$.

			Posterior median		
n	Prior	Hyperparameter	α_1	α_2	ξ_1 (95% CI)
50	GAIPV	$s = 1/2$	0.35	0.36	0.50(0.37,0.64)
50	GAIPV	$s = 1/8$	0.029	0.029	0.50(0.36,0.63)
50	GAIPV	$s = 1/32$	2.7e-06	2.5e-06	0.50(0.36,0.64)
50	GAIPV	$s = 1/128$	1.7e-20	1.8e-20	0.50(0.37,0.64)
50	HSIPV	$c = 1/2$	0.35	0.35	0.50(0.36,0.64)
50	HSIPV	$c = 1/8$	0.098	0.099	0.50(0.37,0.64)
50	HSIPV	$c = 1/32$	0.022	0.022	0.50(0.37,0.64)
50	HSIPV	$c = 1/128$	0.0049	0.005	0.50(0.37,0.63)
500	GAIPV	$s = 1/2$	0.35	0.35	0.50(0.45,0.54)
500	GAIPV	$s = 1/8$	0.025	0.025	0.50(0.46,0.54)
500	GAIPV	$s = 1/32$	1.6e-05	1.6e-05	0.50(0.46,0.54)
500	GAIPV	$s = 1/128$	1.6e-21	1.5e-21	0.50(0.46,0.54)
500	HSIPV	$c = 1/2$	0.34	0.34	0.50(0.46,0.55)
500	HSIPV	$c = 1/8$	0.1	0.1	0.50(0.46,0.54)
500	HSIPV	$c = 1/32$	0.022	0.022	0.50(0.46,0.54)
500	HSIPV	$c = 1/128$	0.0057	0.0057	0.50(0.46,0.54)

From Table 1, the posterior medians of α_1 and α_2 both decrease with the value of the hyperparameter, but the the GAIPV-based values decrease by about 20 orders of magnitude whereas the HSIPV-based values decrease by just 2 orders of magnitude. As expected, however, given the amount of data, the posterior median of the main parameter of interest, ξ_1, is not affected in this scenario. An evaluation of the statistical implications of Theorems 1 and 2 on the estimation of $\boldsymbol{\xi}$ is left to Section 3.3.

3.2 Fixed-data evaluation 2

MCMC algorithms draw from a target posterior distribution through iterative sampling. Because this is a stochastic process that runs for a finite time period, each chain of iterations will follow a different trajectory and thus yield slightly different inference. In this section, we characterize the variation in empirical diagnostics for underflow, convergence, and running time across independent chains as well as the variation in posterior summaries across independent chains. Specifically, we create three datasets (described below) and, to each dataset, fit model (1) equipped with the GAIPV prior or the HSIPV prior. Because the GAIPV prior is susceptible to underflow, we also consider modifications to the GAIPV prior so that its support is truncated below at three different values. These five priors are described below:

GA$_1$ a modified GAIPV prior with shape parameter s chosen to represent $\tilde{n} = 0.5$ historical observations and the support of the distribution of each α_j truncated below at $\epsilon \approx 2.22 \times 10^{-16}$;

GA$_2$ a modified GAIPV prior with shape parameter s chosen to represent $\tilde{n} = 0.5$ historical observations and the support of the distribution of each α_j truncated below at $\epsilon/10 \approx 2.22 \times 10^{-17}$;

GA$_3$ a modified GAIPV prior with shape parameter s chosen to represent $\tilde{n} = 0.5$ historical observations and the support of the distribution of each α_j truncated below at $\epsilon/100 \approx 2.22 \times 10^{-18}$;

GA$_4$ the GAIPV prior with shape parameter s chosen to represent $\tilde{n} = 0.5$ historical observations (this is the standard GAIPV prior and thus, in contrast to the above priors, there is no truncation of α_j)

HS the HSIPV prior with scale parameter c chosen such that the number of anticipated jumps in the probability curve was $m_{\text{eff}} = 0.5$

For each of the 15 prior-by-dataset combinations, we fit 50 models using the same settings as in Section 3.1. Each of the 50 fitted models differs only by the starting seed for the Hamiltonian Monte Carlo algorithm, and we compare the variation in empirical diagnostics across these 50 seeds. For each fitted model, we calculate the number of divergent transitions (Section 2.1); the proportion of seeds in which the posterior calculation of ξ in (3) resulted in at least one numerical value of NaN, i.e. $0/0$; the Gelman-Rubin convergence statistic $\hat{R}$ (Gelman et al., 1992); and the running time of the longer of two chains. We also report the empiric distribution of the posterior median values of α_j and ξ_j, $j = 1, \ldots, K$ across the 50 different random seeds. The first dataset contains 80 $\{X, Y\}$ pairs, with the X's being split equally across $K = 10$ categories. That is, $n_j \equiv \sum_{i=1}^{n} I(X_i = j)$ is fixed at $n_j = 8$, $j = 1, \ldots, K$. The second dataset contains 80 $\{X, Y\}$ pairs split equally across $K = 5$ categories. And the third dataset contains 320 $\{X, Y\}$ pairs split equally across $K = 10$ categories.

For all datasets, the prevalence of outcomes within each category was set to be $\sum_{i=1}^{n} I(X_i = j, Y_i = 1) \equiv \text{round}(n_j(j-1)/(K-1))$, $j = 1, \ldots, K$, where the $\text{round}(\cdot)$ function rounds its argument to the nearest integer. Thus, the observed prevalence of the outcome is, to the closest extent possible, uniformly increasing in each category. The three datasets are plotted as diamonds in Figure 2 (one dataset per column).

These diagnostic results are summarized in Table 2. The most-truncated GAIPV priors, namely GA$_1$, and the unmodified HSIPV prior, HS, give similar numerical results: a median of 0 divergent transitions, no evidence of numeric underflow, and a relatively fast run time (median < 10 seconds per seed). In contrast, the least-truncated (GA$_3$) and un-truncated (GA$_4$) GAIPV priors both encountered hundreds of divergent transitions per seed, underflowed at least once in 8% to 26% of the seeds, and required a median across seeds of 529 to 721 seconds to run to completion. GA$_2$, which falls between GA$_1$ and GA$_3$ in terms of truncation, also falls between these priors in terms of typical diagnostics, suggesting that this loss of numeric stability occurs gradually.

Figure 1 shows the empiric distribution of the logarithm of the posterior medians of each α_j across the 50 starting seeds. The inner box gives the interquartile range of the posterior medians, and the longer line covers the entire observed range of the posterior medians. Because the data are fixed, what is shown here is not statistical variation due to sampling variability but rather numerical variation due to different starting seeds and the finitude of MCMC. Thus, ideally there would be no variation between iterations here. Noting the ranges of the y-axes, GA$_2$, GA$_3$, and GA$_4$ are clearly

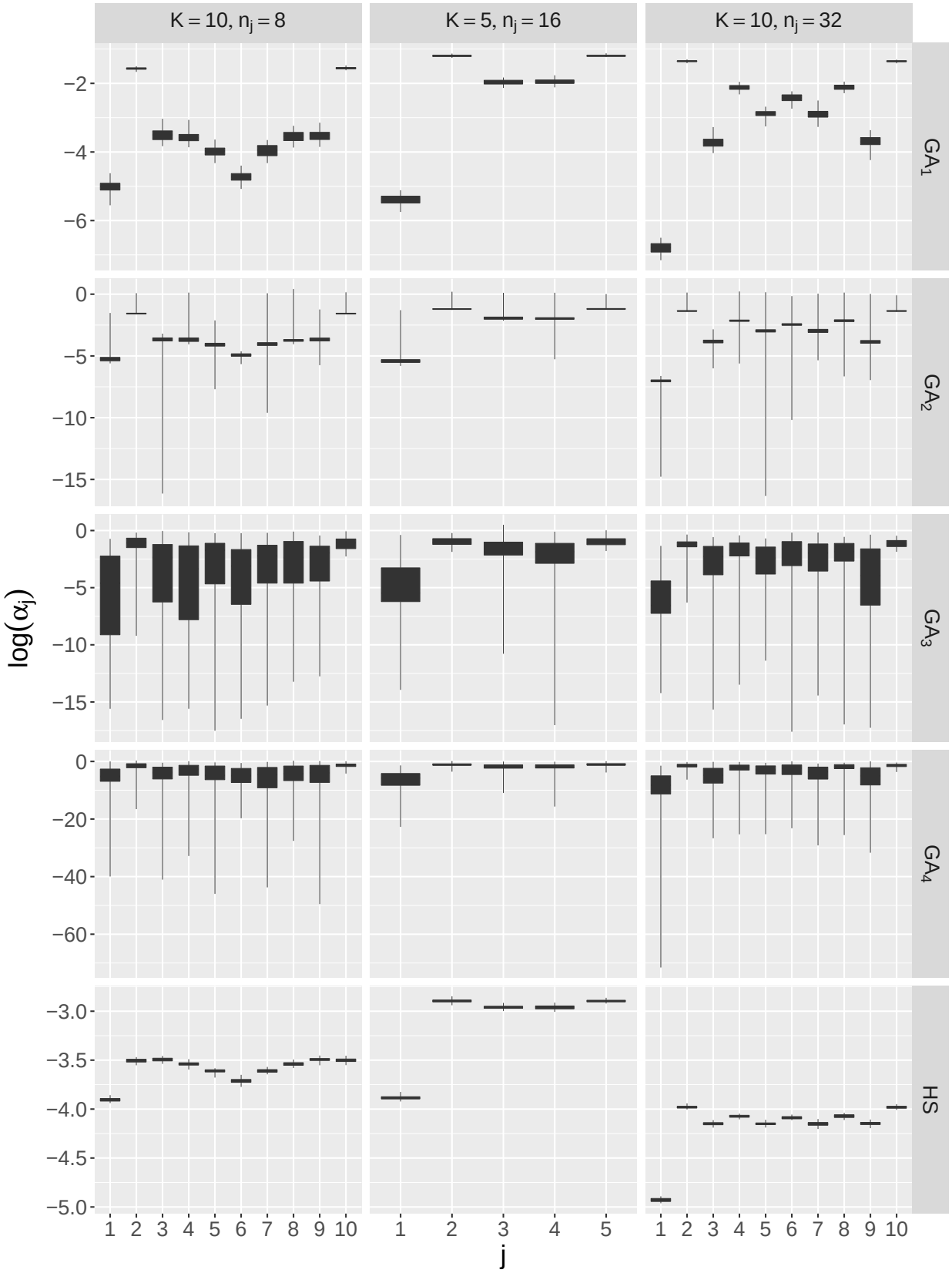

Figure 1: Boxplots giving the empirical distribution of the logarithm of the posterior median of α_j by dataset (columns) and prior (rows) across 50 different starting seeds.

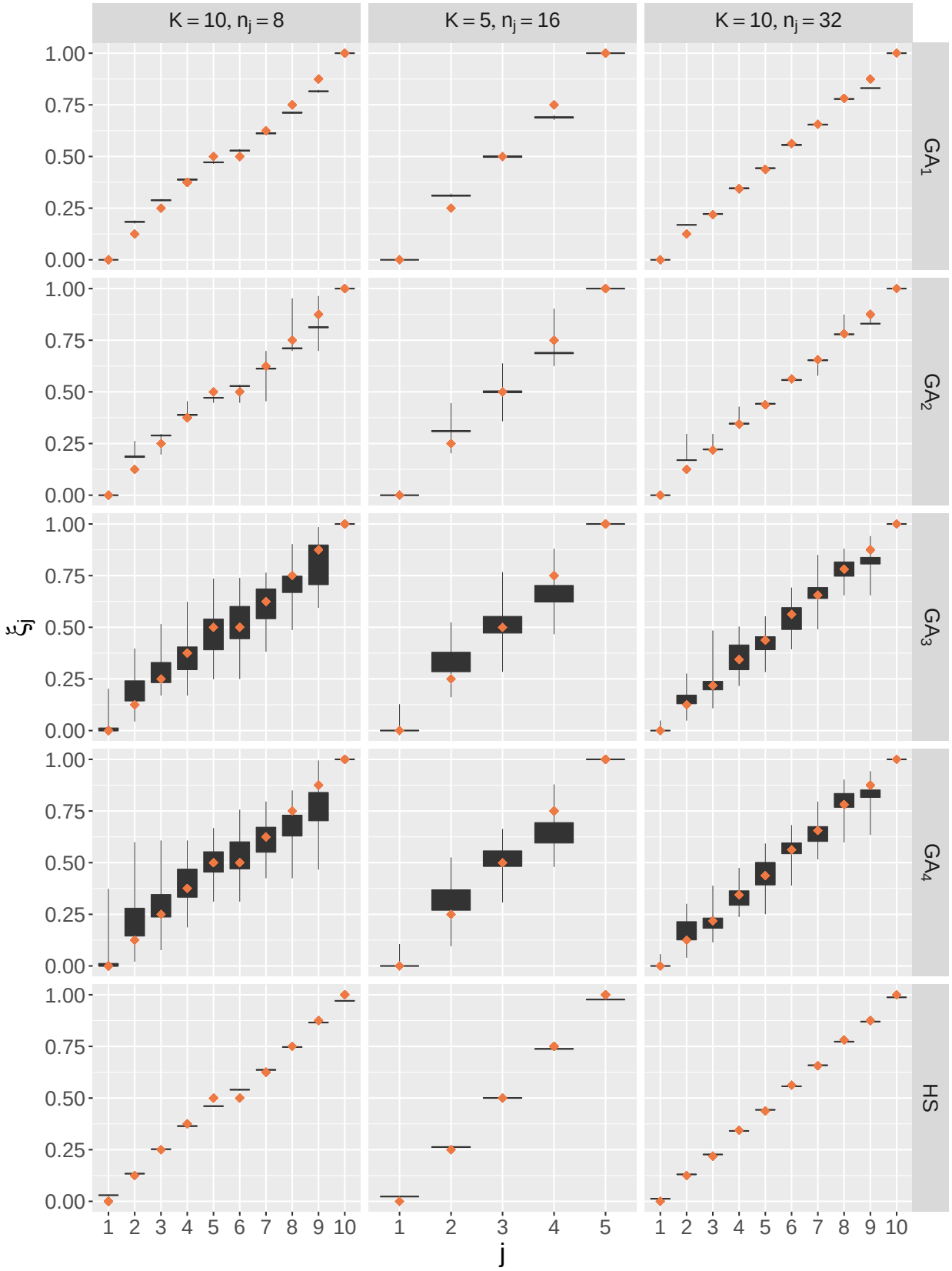

Figure 2: Boxplots giving the empirical distribution of the posterior median of ξ_j by dataset (columns) and prior (rows) across 50 different starting seeds. The red diamonds give the prevalence of the outcome in each dataset.

Table 2: Results from numerical evaluation of three choices of prior fit to three exemplar datasets using 50 different random seeds to start the Hamiltonian Monte Carlo algorithm. GA_1, GA_2, and GA_3, have their support truncated below at ϵ, $\epsilon/10$, and $\epsilon/100$, respectively; GA_4 is the untruncated GAIPV prior.

Dataset Label	Prior	Divergent transitions (median (max))	at least one NaN (proportion)	$\hat{R}$ (median (max))	Run time, s (median (max))
1	GA_1	0(0)	0	1.00(1.01)	5(8)
	GA_2	44(70)	0	1.00(12.5)	5(1094)
	GA_3	366(2679)	0.22	4.88($> 10^9$)	721(1365)
	GA_4	498(3454)	0.12	7.89($> 10^9$)	711(1347)
	HS	0(0,0)	0	1.00(1.01)	2(3)
2	GA_1	0(0)	0	1.00(1.01)	1(2)
	GA_2	18(2495)	0	1.01(5.00)	2(850))
	GA_3	180(4351)	0.08	1.54($> 10^9$)	564(1122)
	GA_4	238(3026)	0.08	1.56($> 10^9$)	529(1359)
	HS	0(0,0)	0	1.00(1.01)	1(2)
3	GA_1	0(0)	0	1.00(1.02)	7(10)
	GA_2	61(2236)	0.06	1.00(9.09)	7(811))
	GA_3	540(4998)	0.26	3.74($> 10^9$)	650(1373)
	GA_4	511(2978)	0.22	2.99($> 10^9$)	658(1234)
	HS	0(0,0)	0	1.00(1.01)	3(4)

inferior, with the posterior median sometimes varying by several dozen logarithms between starting seeds. In contrast, GA_1 and HS exhibit substantially smaller variability.

The same phenomenon is exhibited in Figure 2 which gives the empiric distribution of the posterior medians of each ξ_j. The red diamonds in each panel give the actual data. For GA_1 and HS, in the first and fifth rows, respectively, the posterior medians of ξ_j are nearly constant between random seeds, as evidenced by the box plots having virtually no height. Again, and consistent with the previous results, GA_3 and GA_4 experience a great deal of undesirable between-seed variability.

3.3 Varying-data evaluation

Finally, we conducted a simulation study evaluating the statistical performance of the five priors above. Here, instead of fixing the data, we fix the data-generating mechanism, so that the underlying data varies from iteration to iteration. We consider two different generating scenarios, both based upon model (1). In both scenarios, there are $K = 10$ categories, with $\Pr(X_i = j) \propto 1$. In the first scenario, the outcome probabilities are given by $\xi_j = (j - 0.5)/K$, meaning that the probability of the outcome increases linearly with the category. In the second scenario, the outcome probabilities are given by

$$\xi_j = 0.35\frac{1}{1+\exp\{16.7 - 66.7(j-0.5)/K\}} + 0.65\frac{1}{1+\exp\{50 - 66.7(j-0.5)/K\}} \quad (15)$$

In this scenario, the outcome probability increases with j, then plateaus at $j = 4$, and then sharply increases again starting at $j = 8$. These true probabilities are plotted as dots in the two columns of

Figure 3. We crossed these two scenarios with two sizes of the training data, $n = 80$ and $n = 320$, yielding four unique data generating scenarios.

For each data generating scenario, we simulated 200 independent datasets and fit to each dataset model (1) with prior GA_1, GA_2, GA_3, GA_4, or HS, using the same settings as in the fixed-data example of Section 3.2. As a benchmark, we also compared our Bayesian models to standard isotonic regression (labeled Isoreg). The R package `cir` was used for fitting standard isotonic regression (Oron, 2017). All code developed for this article is available on the first author's github site (http://www.github.com/psboonstra)

Given the category of the predictor X, we summarized each fitted Bayesian model by calculating the posterior median of ξ_j, $j = 1, \ldots, K$. We construct these point estimates of ξ_j to facilitate direct comparison with the non-Bayesian method Isoreg. For each simulation of each of the data generating mechanisms, we assessed model fit on a testing dataset of size $n_{\text{test}} = 1000$, generated independently from the training data but arising from the same true generating mechanism. The ith observation in the testing data, $i = 1, \ldots, 1000$, consists of $j_i \in \{1, \ldots, K\}$, which is the predictor category that the ith testing observation falls in, and p_i, which is the true, unknown outcome probability arising from the same probability curve as the training data. For a given method, let $\hat{\xi}_{j_i}$ indicate the point estimate of the probability category for test observation i. With this notation, the pointwise Kullback-Leibler divergence (KL) is

$$\text{KL} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[p_i \log\left(\frac{p_i}{\hat{\xi}_{j_i}}\right) + (1 - p_i) \log\left(\frac{1 - p_i}{1 - \hat{\xi}_{j_i}}\right) \right].$$

To avoid infinitely valued KL divergences, we truncated all values of p_i and $\hat{\xi}_{j_i}$ below and above by ϵ and $1 - \epsilon$, respectively, where $\epsilon \approx 2.2 \times 10^{-16}$. Smaller values of KL divergences are better, with the best possible value being 0. We averaged KL across all independent datasets arising from the same generating scenario.

We also calculate the pointwise root mean-squared error (RMSE), defined as $\sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (p_i - \hat{\xi}_{j_i})^2}$. As with KL divergence, smaller values of RMSE are better.

The average pointwise KL divergences and RMSEs are given in Table 3. The methods that have a value within 10% of the best method are in **boldface**. Figure 3 plots the true probability curve and the distribution of posterior medians (or, for Isoreg, the distribution of point estimates) of each ξ_j across independent datasets, separately for each scenario; for clarity of presentation we do not show the GA_2 and GA_3 priors, which fall between GA_1 and GA_4. Visually, HS more faithfully captures the true probability curves, and this is formalized in Table 3, in which HS is generally preferred with regards to KL divergence. Consistent with our previous results, the GAIPV-type priors lose efficiency as the lower truncation decreases from ϵ (GA_1) to $\epsilon/10$ (GA_2) to $\epsilon/100$ (GA_3) to no truncation (GA_4).

Remark 1 The substandard performance of Isoreg with regard to the KL divergence metric (Table 3) is striking and is explained in part by the following. If the observations in the first category of the predictor all have $Y_i = 0$, then the fitted isotonic regression curve at that predictor value will have $\hat{\xi}_1 = 0$. As mentioned in Section 3.3, we truncate values of $\hat{\xi}_{j_i}$ above and below by ϵ and $1 - \epsilon$. Without this truncation, the resulting KL value would be infinite when any $\hat{\xi}_{j_i}$ is zero; *with* this truncation, it is very large but finite. A simple fix for Isoreg would be to add a small fractional observation, e.g. 1/64, 1/16, or 1/2, having $Y = 1$ and equal fractional observation

having $Y = 0$ to each category of the predictor, which, if the fraction used is 1/2, would reduce to a Jeffrey's prior in the case of $K = 1$. As evidenced by the bottom rows of Table 3. The relative performance of Isoreg improves somewhat when using RMSE as a metric.

Table 3: Average pointwise Kullback-Leibler (KL) divergence $\times 1000$ and average root mean square error (RMSE) $\times 1000$. **Bolded** entries highlight methods for which the pointwise value was within 10% of the rowwise minimum.

Design factors			Methods					
Curve	n	Metric	Isoreg	HS	GA_1	GA_2	GA_3	GA_4
1	80	KL Divergence	433	**17**	93	107	193	215
1	320	KL Divergence	78	**9**	34	36	57	59
2	80	KL Divergence	243	**34**	42	47	68	83
2	320	KL Divergence	14	12	**10**	**10**	16	16
1	80	RMSE	103	**70**	107	108	123	125
1	320	RMSE	62	**49**	71	72	75	76
2	80	RMSE	**94**	**90**	**96**	**98**	117	120
2	320	RMSE	**47**	54	55	56	69	66

4 Data analysis: Radiation-induced lung toxicity

We reanalyze the LF20-RILT data first reported in Owen et al. (2020) with our Bayesian isotonic regression models. For training data, there are 58 lung cancer patients enrolled in an early-phase clinical trial at the University of Michigan (ClinicalTrials.gov NCT00603057) between the years 2007 and 2013. The primary outcome for this analysis is grade 2+ RILT, which is a composite toxicity endpoint defined as the occurrence of either grade 2 or higher pneumonitis or grade 2 or higher fibrosis. Nine patients (15.5%) experienced grade 2+ RILT. The predictor is the dosimetric LF20, which is the percentage of a patient's normal lung tissue that both (i) is classified as low-functioning by SPECT V/Q *and* (ii) received greater than 20 Gy of radiation. The data are plotted as a rug in Figure 4. We fit the same priors considered in Sections 3.2 and 3.3, using a value of $\tilde{n} = 0.5$ historical patients (GAIPV-type priors) or $m_{\text{eff}} = 0.5$ anticipated jumps (HSIPV prior) to determine the value of the hyperparameters.

4.1 Categorizing LF20

Our theoretical development above is based upon the predictor X_i being distributed as an ordered categorical variable, whereas LF20 is a proportion taking on values in $[0, 1]$. The Bayesian priors can be applied after categorizing the predictor. We considered two algorithmic approaches for identifying an ordered vector of indices with K categories. Setting up notation, let ϵ be as defined above and let $\boldsymbol{\zeta}$ be a $K + 1$-length vector $\boldsymbol{\zeta} = \{\zeta_1, \ldots, \zeta_{K+1}\} \subset \{-\epsilon, X_1, X_2, \ldots, X_{n-1}, 1\}$, with $\zeta_1 \equiv -\epsilon$, $\zeta_{K+1} \equiv 1$, and $\zeta_j < \zeta_{j+1}$ for $j = 1, \ldots, K$. Then any value X_i for which $\zeta_j < X_i \leq \zeta_{j+1}$, $j = 1, \ldots, K$, is assumed to have a common outcome probability ξ_j.

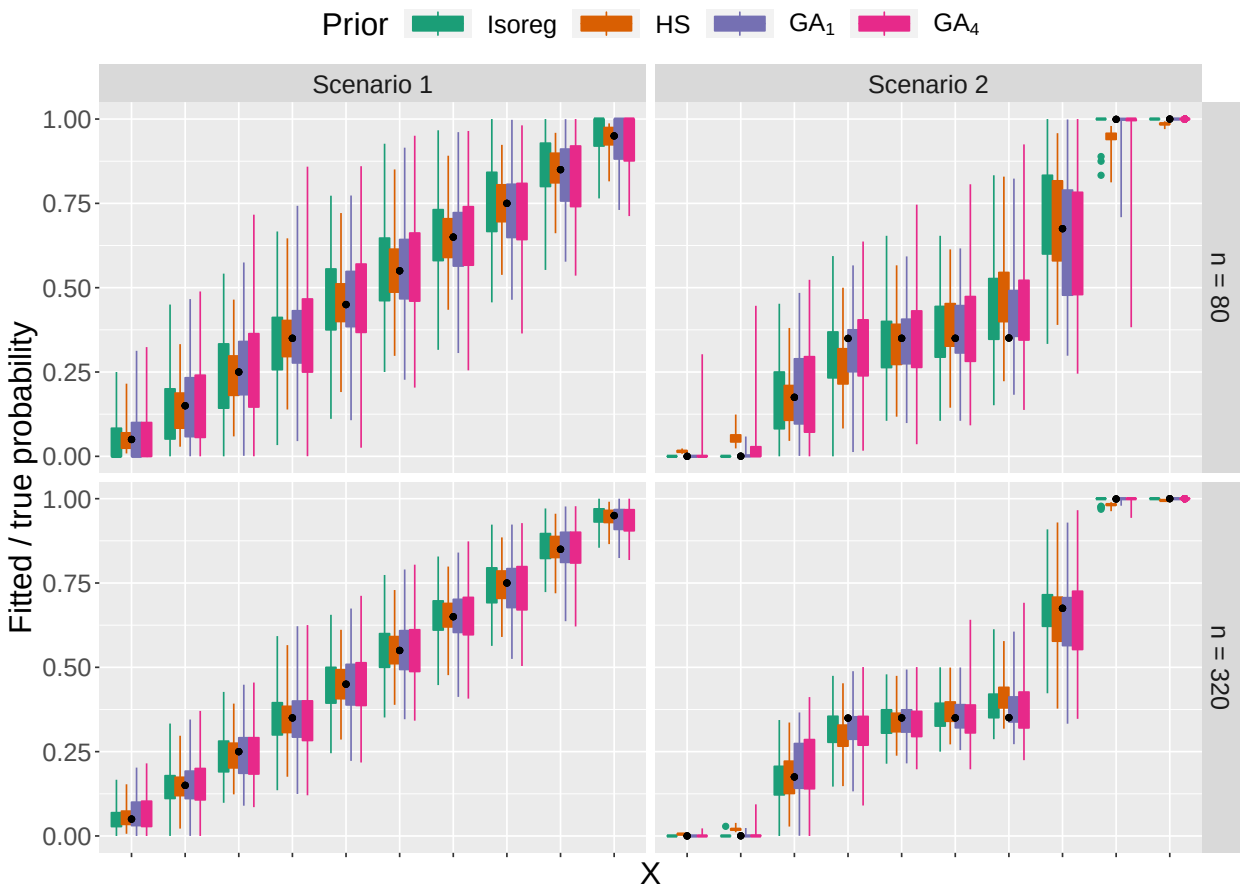

Figure 3: True probabilities of the outcomes across each category (dots) plotted on the distribution of posterior medians for five priors (boxplots) for two different true probability curves (columns) and two different sizes of the training dataset (rows). The online version of this figure is in color.

Using this, one approach, called "Quantiles", takes as input the desired number of categories, K, and sets $\zeta_j = X_{(f(j))}$, $j = 2, \ldots, K$, where $X_{(\ell)}$ is the ℓth order statistic from among the n observations and $f(j) = \text{round}(n[j-1]/K)$, i.e. the whole number closest to the quantity $n(j-1)/K$. Assuming a shrinkage-type prior is being used, K can be a large fraction of n, since the shrinkage will mitigate overfitting. We considered $K = 29$ for this approach in this data analysis, yielding $n/K = 58/29 = 2$ observations per category.

The second approach is the 'pool adjacent violators algorithm' (PAVA) (Barlow et al., 1972). The usual PAVA algorithm is applied to the set of outcomes ordered with respect to LF20 to determine the cutpoints at which jumps occur in the traditional isotonic regression curve, and these indices are used directly as $\boldsymbol{\zeta}$. PAVA grouped the 58 observations in the training data into $K = 5$ categories.

4.2 Cross-validation and validation

We cross-validated the fitted models by holding out one randomly selected patient who did not experience grade 2+ RILT and, independently, one randomly selected patient who experienced

Table 4: Description of methods fit to the RILT data and their cross-validated and validated performance. The final three columns under the 'Cross validation' header give the predicted probability of grade 2+ RILT averaged across 150 held-out pairs of patients one of whom, in truth, experienced grade 2+ RILT ($\hat{p}_1$), one of whom did not experience grade 2+ RILT ($\hat{p}_0$), and their difference ($\hat{p}_1 - \hat{p}_0$), respectively.

			Cross validation					Validation
Method	Cut strategy	Hyper-parameter	Divergent transitions (median(max))	Run time, s (median(max))	$\hat{p}_1^{\dagger}$	$\hat{p}_0^{\ddagger}$	$\hat{p}_1 - \hat{p}_0^{\dagger}$	Brier Score‡ = $\frac{1}{30}\sum_{i=1}^{30}(Y_i - \hat{\xi}_{j_i})^2$
Isoreg	PAVA	–	–	$\ll 1 (\ll 1)$	0.247	0.113	0.134	0.149
HS	PAVA	$c = 0.0015$	0(0)	1.4(1.9)	0.237	0.121	0.116	0.146
GA_1	PAVA	$s = 0.083$	0(0)	1.9(3.3)	0.210	0.104	0.106	0.144
GA_2	PAVA	$s = 0.083$	11(24)	2.1(3.4)	0.210	0.105	0.105	0.144
GA_3	PAVA	$s = 0.083$	170(2548)	123(1091)	0.187	0.100	0.087	0.145
GA_4	PAVA	$s = 0.083$	2.72(974)	2.0(1263)	0.209	0.112	0.097	0.144
HS	Quantiles	$c = 0.00013$	0(0)	5.6(6.4)	0.256	0.147	0.110	0.136
GA_1	Quantiles	$s = 0.017$	0(0)	9.4(12.8)	0.240	0.147	0.112	0.138
GA_2	Quantiles	$s = 0.017$	0(0)	10.0(11.6)	0.239	0.127	0.112	0.137
GA_3	Quantiles	$s = 0.017$	0(0)	10.2(13.7)	0.237	0.125	0.112	0.138
GA_4	Quantiles	$s = 0.017$	2(7)	44.6(58.5)	0.222	0.119	0.103	0.139

† Larger is better
‡ Smaller is better

grade 2+ RILT. We refit the models on the remaining 56 patients (including re-calculating the LF20 dosimetric categories as described above) and estimated the probability of grade 2+ RILT for each held-out observation, given its value of LF20. There are $49 \times 9 = 441$ such possible pairings, and we repeated this 'leave-two out' validation step for 150 random, unique pairs. We report the average of the predicted probabilities, separately for the two outcomes, as well as the difference in these probabilities. We also calculated the number of divergences and the runtime across the 150 iterations.

We also validated the models fit to the the full 58 patient training cohort on a completely separate cohort of patients from a later clinical trial also at the University of Michigan (ClinicalTrials.gov NCT02492867), which began accruing patients in 2016. This validation cohort was comprised of 30 similar patients, of which six (20%) experienced the RILT outcome. We calculated the Brier score, which is the average squared difference between each patient's model-predicted risk, $\hat{\xi}_j$, and their eventual outcome, Y_i, namely $\frac{1}{30}\sum_{i=1}^{30}(Y_i - \hat{\xi}_{j_i})^2$. It is similar to the RMSE metric used in Section 3.3, but replacing the true unknown probability p_i with the observed outcome Y_i.

Figure 4 plots the model-based probabilities of grade 2+ RILT as a function of LF20 and Table 4 gives the cross-validated and validated assessments of each model. For clarity of presentation, we do not present results for GA_2 and GA_3 in the figure. From the figure, there were greater differences between methods at the higher dose levels, where there is less data. Specifically, using the PAVA categorization, GA_1 and GA_4 estimated the probability of toxicity at the highest dosimetric to be 0.3 and 0.4, respectively, versus 0.45 and 0.5 for HS and Isoreg. Under the quantiles-based categorization, HS estimates this probability to be 0.44, greater than GA_4 (0.37) and GA_1 (0.43).

Focusing on the cross-validated metrics, from Table 4, under the PAVA categorization, the

average difference between $\hat{p}_1$ and $\hat{p}_0$ was largest, i.e. best, for Isoreg and HS_1 (about 0.13 and 0.12 versus 0.11 or less for the remaining methods). Under the quantiles-based categorization, all methods except GA_4 had similar average differences between $\hat{p}_1$ and $\hat{p}_0$. With regard to running time, the non-Bayesian Isoreg can be fit virtually instantaneously in these data. HS_1, GA_1, and GA_2 required about 2 seconds per cross-validated dataset (PAVA categorization) or 5-10 seconds (quantiles-based categorization). The median running time for GA_3 was similarly short, but under the PAVA categorization its longest running time across cross-validated datasets was about 18 minutes. GA_4 was slower, particularly under the quantiles-based categorization, consistent with our numerical studies.

The untruncated/less-truncated GAIPV priors did encounter divergent transitions, although not to the extent observed in our simulation studies. GA_4 underflowed in one dataset (results not shown). Although the untruncated gamma-based prior (GA_4) is prone to underflow and/or may have long runtimes, this will not be the case in every data configuration.

Focusing on our validation in the separate cohort of 30 patients, from the last column of Table 4, the GAIPV-priors had the smallest, i.e. best, Brier scores under PAVA categorization (about 0.144-0.145), and HS had the smallest Brier score under the quantiles-based categorization (0.136).

5 Discussion

Extending the original Bayesian isotonic regression first considered in Ramsey (1972), we have shown that our novel prior for this model, based on the horseshoe distribution, is both computationally robust and statistically efficient. That the horseshoe-based prior is not subject to the numerical underflow that can ensnare the gamma-based prior is supported by our key theoretical result, given in Section 2.3, which gives that the rate at which the horseshoe density diverges as its argument goes to zero is always less than the rate at which the gamma density does so.

Apart from a head-to-head comparison of the horseshoe versus gamma-based prior, it is also noteworthy that the performance, running time, and diagnostics of the latter are sensitive to the value of the lower truncation. That is, the only difference between GA_1 and GA_4 is that in the former the support of the distribution of α_j is truncated to be no smaller than $\epsilon \approx 2.2 \times 10^{-16}$, whereas in the latter there is no truncation. This is a large relative difference, to be sure, but a small absolute difference nonetheless.

Funding

This work was supported by the National Institutes of Health [grant number P30 CA046592].

Appendix

5.1 Proof of Proposition 2

Proof 1 *It is straightforward to show part* ***a*** *by the definition of the density form:* $g(x) = \{\Gamma(s)\}^{-1}x^{s-1}\exp(-x) = O(x^{s-1})$ *and when* $x \to 0$,

$$|g'(x)| = \{\Gamma(s)\}^{-1}\{(s-1)x^{s-2} - x^{s-1}\}\exp(-x) = O(x^{s-2}).$$

For part ***b****, by the Theorem 1 by (Liu et al., 2013), we have* $-s\log(\alpha) \to \text{Exp}(1)$ *as* $s \to 0$*. This further implies that*

$$\begin{aligned}
&\lim_{s\downarrow 0}\Pr(\alpha < \delta) \\
&= \lim_{s\downarrow 0}\Pr\{\alpha < \exp\big(-s^{-\kappa}\big)\} \\
&= \lim_{s\downarrow 0}\Pr\{-s\log(\alpha) > s^{1-\kappa}\} \\
&= \lim_{s\downarrow 0}\exp\big(-s^{1-\kappa}\big) = 1
\end{aligned}$$

5.2 Proof of Theorem 1

Proof 2 *' Recall that* $\xi_k = \sum_{j=1}^{k}\alpha_j / \sum_{j=1}^{K}\alpha_j$*. Define* $m_k \equiv \sum_{i=1}^{n} I(X_i = k, Y_i = 1)$ *and* $n_k \equiv \sum_{i=1}^{n} I(X_i = k)$*. The posterior density of* $\boldsymbol{\alpha}$ *given* $\{m_k, n_k\}_{k=1}^{K}$ *is proportional to*

$$\begin{aligned}
h(\boldsymbol{\alpha}) &= \prod_{k=1}^{K-1}\xi_k^{m_k}(1-\xi_k)^{n_k-m_k}\prod_{l=1}^{K}\pi(\alpha_l) \\
&= \prod_{k=1}^{K-1}\left(\frac{\sum_{l=1}^{k}\alpha_l}{\sum_{l=1}^{K}\alpha_l}\right)^{m_k}\left(\frac{\sum_{l=k+1}^{K}\alpha_l}{\sum_{l=1}^{K}\alpha_l}\right)^{n_k-m_k}\prod_{l=1}^{K}\pi(\alpha_l) \\
&= g(\boldsymbol{\alpha}_{-j}, \alpha_j)\prod_{l=1}^{K}\pi(\alpha_l)
\end{aligned}$$

for any $1 \le j < K$*, where*

$$\begin{aligned}
g(\boldsymbol{\alpha}_{-j}, \alpha_j) &= \prod_{k=1}^{K-1}\left(\frac{\sum_{l=1}^{k}\alpha_l}{\sum_{l=1}^{K}\alpha_l}\right)^{m_k}\left(\frac{\sum_{l=k+1}^{K}\alpha_l}{\sum_{l=1}^{K}\alpha_l}\right)^{n_k-m_k} \\
&= \prod_{k\ge j}\left(\frac{\sum_{l=1,l\neq j}^{k}\alpha_l + \alpha_j}{\sum_{l=1,l\neq j}\alpha_l + \alpha_j}\right)^{m_k}\left(\frac{\sum_{l=k+1}^{K}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l + \alpha_j}\right)^{n_k-m_k} \times \\
&\quad \prod_{k<j}\left(\frac{\sum_{l=1}^{k}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l + \alpha_j}\right)^{m_k}\left(\frac{\sum_{l=k+1,l\neq j}^{K}\alpha_l + \alpha_j}{\sum_{l=1,l\neq j}^{K}\alpha_l + \alpha_j}\right)^{n_k-m_k}
\end{aligned}$$

Let $\delta = \exp(-s^{\kappa})$*. When* $0 < \alpha_j < \delta$*, for any* $\boldsymbol{\alpha}_{-j} = \{\alpha_l, l \neq j\}$*, it is straightforward to show that*

$$g_l(\boldsymbol{\alpha}_{-j}, \delta) \le g(\boldsymbol{\alpha}_{-j}, \alpha_j) \le g_u(\boldsymbol{\alpha}_{-j}, \delta), \tag{16}$$

where

$$g_l(\boldsymbol{\alpha}_{-j},\delta)=\prod_{k\geq j}\left(\frac{\sum_{l=1,l\neq j}^{k}\alpha_l}{\sum_{l=1,l\neq j}\alpha_l}\right)^{m_k}\left(\frac{\sum_{l=k+1}^{K}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l+\delta}\right)^{n_k-m_k}\times \prod_{k<j}\left(\frac{\sum_{l=1}^{k}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l+\delta}\right)^{m_k}\left(\frac{\sum_{l=k+1,l\neq j}^{K}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l}\right)^{n_k-m_k} \tag{17}$$

$$g_u(\boldsymbol{\alpha}_{-j},\delta)=\prod_{k\geq j}\left(\frac{\sum_{l=1,l\neq j}^{k}\alpha_l+\delta}{\sum_{l=1,l\neq j}\alpha_l+\delta}\right)^{m_k}\left(\frac{\sum_{l=k+1}^{K}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l}\right)^{n_k-m_k}\times \prod_{k<j}\left(\frac{\sum_{l=1}^{k}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l}\right)^{m_k}\left(\frac{\sum_{l=k+1,l\neq j}^{K}\alpha_l+\delta}{\sum_{l=1,l\neq j}^{K}\alpha_l+\delta}\right)^{n_k-m_k} \tag{18}$$

for any $\boldsymbol{\alpha}_{-j}$.

When $\alpha_j>\delta$, *we have*

$$g(\boldsymbol{\alpha}_{-j},\alpha_j)\leq g_u^*(\boldsymbol{\alpha}_{-j},\delta)$$

where

$$g_u^*(\boldsymbol{\alpha}_{-j},\delta)=\prod_{k\geq j}\left(\frac{\sum_{l=k+1}^{K}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l+\delta}\right)^{n_k-m_k}\times\prod_{k<j}\left(\frac{\sum_{l=1}^{k}\alpha_l}{\sum_{l=1,l\neq j}^{K}\alpha_l+\delta}\right)^{m_k}$$

In part ***a****, for any given* $n\geq 1$, *when* $s\to 0$, *then* $\delta=\exp\left(-s^{-\kappa}\right)\to 0$ *and*

$$g_l(\boldsymbol{\alpha}_{-j},\delta)\to g(\boldsymbol{\alpha}_{-j},0),$$
$$g_u(\boldsymbol{\alpha}_{-j},\delta)\to g(\boldsymbol{\alpha}_{-j},0).$$

Then the marginal posterior probability of α_j *being concentrated within a small neighborhood of zero is given by*

$$\begin{aligned}
&\Pr(\alpha_j<\delta\mid Y,X) \qquad (19)\\
&=\frac{\int_0^\delta\int_{(0,+\infty)^{K-1}}h(\boldsymbol{\alpha})d\alpha_j d\boldsymbol{\alpha}_{-j}}{\int_0^\infty\int_{(0,+\infty)^{K-1}}h(\boldsymbol{\alpha})d\alpha_j d\boldsymbol{\alpha}_{-j}}\\
&=\frac{\int_0^\delta\int_{(0,+\infty)^{K-1}}g(\boldsymbol{\alpha}_{-j},\alpha_j)\pi(\alpha_j)\prod_{l\neq j}\pi(\alpha_l)d\alpha_j d\boldsymbol{\alpha}_{-j}}{\int_0^\delta\int_{(0,+\infty)^{K-1}}g(\boldsymbol{\alpha}_{-j},\alpha_j)\pi(\alpha_j)\prod_{l\neq j}\pi(\alpha_l)d\alpha_j d\boldsymbol{\alpha}_{-j}+\int_\delta^\infty\int_{(0,+\infty)^{K-1}}g(\boldsymbol{\alpha}_{-j},\alpha_j)\pi(\alpha_j)\prod_{l\neq j}\pi(\alpha_l)d\alpha_j d\boldsymbol{\alpha}_{-j}}\\
&\geq\frac{\int_0^\delta\int_{(0,+\infty)^{K-1}}g_l(\boldsymbol{\alpha}_{-j},\delta)\pi(\alpha_j)\prod_{l\neq j}\pi(\alpha_l)d\alpha_j d\boldsymbol{\alpha}_{-j}}{\int_0^\delta\int_{(0,+\infty)^{K-1}}g_u(\boldsymbol{\alpha}_{-j},\delta)\pi(\alpha_j)\prod_{l\neq j}\pi(\alpha_l)d\alpha_j d\boldsymbol{\alpha}_{-j}+\int_\delta^\infty\int_{(0,+\infty)^{K-1}}g_u^*(\boldsymbol{\alpha}_{-j},\delta)\pi(\alpha_j)\prod_{l\neq j}\pi(\alpha_l)d\alpha_j d\boldsymbol{\alpha}_{-j}}\\
&=\frac{\left(\int_0^\delta\pi(\alpha_j)d\alpha_j\right)}{\int_0^\delta\pi(\alpha_j)d\alpha_j\left\{\frac{\int_{(0,+\infty)^{K-1}}g_u(\boldsymbol{\alpha}_{-j},\delta)\prod_{l\neq j}\pi(\alpha_l)d\boldsymbol{\alpha}_{-j}}{\int_{(0,+\infty)^{K-1}}g_l(\boldsymbol{\alpha}_{-j},\delta)\prod_{l\neq j}\pi(\alpha_l)d\boldsymbol{\alpha}_{-j}}\right\}+\left\{1-\int_0^\delta\pi(\alpha_j)d\alpha_j\right\}\left\{\frac{\int_{(0,+\infty)^{K-1}}g_u^*(\boldsymbol{\alpha}_{-j},\delta)\prod_{l\neq j}\pi(\alpha_l)d\boldsymbol{\alpha}_{-j}}{\int_{(0,+\infty)^{K-1}}g_l(\boldsymbol{\alpha}_{-j},\delta)\prod_{l\neq j}\pi(\alpha_l)d\boldsymbol{\alpha}_{-j}}\right\}}\\
&\to 1
\end{aligned}$$

in $P^n_{\theta_0}$ probability, as $s \to 0$ and $\delta \to 0$.

*In part **b**, we consider the large sample property of the posterior distribution of α_j. Under the actual distribution of data given the true parameters $\theta_0 = \{\eta_{0k}, \xi_{0k}\}_{k=1}^K$, where $\eta_{0k} = \Pr(X_i = k)$ and $\xi_{0k} = \Pr(Y_i = 1 \mid X_i = k)$. By the law of the large numbers, as $n \to \infty$,*

$$\frac{m_k}{n} \to \eta_{0k}\xi_{0k} \text{ and } \frac{n_k - m_k}{n} \to \eta_{0k}(1 - \xi_{0k})$$

in $P^\infty_{\theta_0}$ probability. Since $s = o(n^{-1/\kappa})$, then $\delta = o\{\exp(-n)\}$ and for any $\boldsymbol{\alpha}_{-j}$,

$$\frac{g_l(\boldsymbol{\alpha}_{-j}, \delta)}{f_l(\boldsymbol{\alpha}_{-j}, n)} \to 1 \quad \text{and} \quad \frac{g_u(\boldsymbol{\alpha}_{-j}, \delta)}{f_u(\boldsymbol{\alpha}_{-j}, n)} \to 1$$

in $P^\infty_{\theta_0}$ probability, where

$$f_l(\boldsymbol{\alpha}_{-j}, n) = \prod_{k \geq j} \left(\frac{\sum_{l=1, l\neq j}^k \alpha_l}{\sum_{l=1, l\neq j} \alpha_l} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1}^K \alpha_l}{\sum_{l=1, l\neq j}^K \alpha_l + \exp(-n)} \right)^{\eta_{0k}(1-\xi_{0k})n} \times \prod_{k < j} \left(\frac{\sum_{l=1}^k \alpha_l}{\sum_{l=1, l\neq j}^K \alpha_l + \exp(-n)} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1, l\neq j}^K \alpha_l}{\sum_{l=1, l\neq j}^K \alpha_l} \right)^{\eta_{0k}(1-\xi_{0k})n}, \tag{20}$$

and

$$f_u(\boldsymbol{\alpha}_{-j}, n) = \prod_{k \geq j} \left(\frac{\sum_{l=1, l\neq j}^k \alpha_l + \exp(-n)}{\sum_{l=1, l\neq j} \alpha_l + \exp(-n)} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1}^K \alpha_l}{\sum_{l=1, l\neq j}^K \alpha_l} \right)^{\eta_{0k}(1-\xi_{0k})n} \times \prod_{k < j} \left(\frac{\sum_{l=1}^k \alpha_l}{\sum_{l=1, l\neq j}^K \alpha_l} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1, l\neq j}^K \alpha_l + \exp(-n)}{\sum_{l=1, l\neq j}^K \alpha_l + \exp(-n)} \right)^{\eta_{0k}(1-\xi_{0k})n}. \tag{21}$$

This implies that as $n \to \infty$,

$$\frac{f_l(\boldsymbol{\alpha}_{-j}, n)}{f_u(\boldsymbol{\alpha}_{-j}, n)} \to 1 \tag{22}$$

in $P^\infty_{\theta_0}$ probability.

$$\frac{\int_{(0,+\infty)^{K-1}} f_u(\boldsymbol{\alpha}_{-j}, n) \prod_{l\neq j} \pi(\alpha_l) d\boldsymbol{\alpha}_{-j}}{\int_{(0,+\infty)^{K-1}} f_l(\boldsymbol{\alpha}_{-j}, n) \prod_{l\neq j} \pi(\alpha_l) d\boldsymbol{\alpha}_{-j}} \to 1 \tag{23}$$

in $P^\infty_{\theta_0}$ probability.

$$\frac{\int_{(0,+\infty)^{K-1}} g_u(\boldsymbol{\alpha}_{-j}, \delta) \prod_{l\neq j} \pi(\alpha_l) d\boldsymbol{\alpha}_{-j}}{\int_{(0,+\infty)^{K-1}} g_l(\boldsymbol{\alpha}_{-j}, \delta) \prod_{l\neq j} \pi(\alpha_l) d\boldsymbol{\alpha}_{-j}} \to 1 \tag{24}$$

in $P^\infty_{\theta_0}$ probability. Thus, by (19), we have

$$\Pr(\alpha_j < \exp(-Mn) \mid Y, X) \to 1 \tag{25}$$

in $P^\infty_{\theta_0}$ probability.

5.3 Proof of Proposition 3

Proof 3 *Let $\theta_j = \lambda_j \tau$. Then the density of θ_j is given by*

$$\pi(t) = \frac{2}{\pi^2} \frac{\log(t^2)}{t^2 - 1}$$

and

$$\alpha_j \mid \theta_j \sim \mathrm{N}_+ \left(0, \frac{c^2\theta_j^2}{1 + c^2\theta_j^2}\right)$$

with

$$\mathrm{E}(\alpha_j \mid \theta_j, c) = \sqrt{\frac{2}{\pi}} \frac{c\theta_j}{\sqrt{1 + c^2\theta_j^2}}, \quad \mathrm{Var}(\alpha_j \mid \theta_j, c) = \frac{c^2\theta_j^2}{1 + c^2\theta_j^2} \left(1 - \frac{2}{\pi}\right)$$

This further implies that

$$\mathrm{E}(\alpha_j \mid c) = \mathrm{E}\{\mathrm{E}(\alpha_j \mid \theta_j, c)\} \tag{26}$$

$$= \sqrt{\frac{2}{\pi}} \int_0^\infty \frac{ct}{\sqrt{1 + c^2t^2}} \frac{2}{\pi^2} \frac{\log(t^2)}{t^2 - 1} dt. \tag{27}$$

Not that for any $t \geq 0$ and $0 \leq c \leq 1$,

$$\frac{ct}{t+1} \leq \frac{ct}{ct+1} \leq \frac{ct}{\sqrt{1 + c^2t^2}} \leq \min\left\{\sqrt{\frac{ct}{2}}, 1\right\}, \tag{28}$$

Also,

$$\int_0^\infty \sqrt{t} \frac{\log(t^2)}{t^2 - 1} dt = \pi^2 \quad \textit{and} \quad \int_0^\infty \frac{t}{t+1} \frac{\log(t^2)}{t^2 - 1} dt = \frac{\pi^2}{4}.$$

Thus,

$$\frac{c}{\sqrt{2\pi}} \leq \mathrm{E}(\alpha_j \mid c) \leq \sqrt{\frac{2}{\pi}} \min(\sqrt{2c}, 1) \tag{29}$$

When $0 < c < 1/2$, then

$$\frac{c}{\sqrt{2\pi}} \leq \mathrm{E}(\alpha_j \mid c) \leq 2\sqrt{\frac{c}{\pi}} \tag{30}$$

In addition,

$$\mathrm{Var}(\alpha_j \mid c) = \mathrm{Var}\{\mathrm{E}(\alpha_j \mid \theta_j, c)\} + \mathrm{E}\{\mathrm{Var}(\alpha_j \mid \theta_j, c)\} \tag{31}$$

$$= \mathrm{Var}\left(\sqrt{\frac{2}{\pi}} \frac{c\theta_j}{\sqrt{1 + c^2\theta_j^2}}\right) + \mathrm{E}\left\{\frac{c^2\theta_j^2}{1 + c^2\theta_j^2}\left(1 - \frac{2}{\pi}\right)\right\} \tag{32}$$

$$= \mathrm{E}\left(\frac{c^2\theta_j^2}{1 + c^2\theta_j^2}\right) - \left[\mathrm{E}\left(\sqrt{\frac{2}{\pi}} \frac{c\theta_j}{\sqrt{1 + c^2\theta_j^2}}\right)\right]^2 \tag{33}$$

Note that

$$\mathrm{E}\left(\frac{c^2\theta_j^2}{1+c^2\theta_j^2}\right) \leq \sqrt{\frac{c}{2}}\mathrm{E}\left(\sqrt{\theta_j}\right)$$

$$\mathrm{E}\left(\frac{c^2\theta_j^2}{1+c^2\theta_j^2}\right) \leq \sqrt{2c}$$

Note that

$$\{\mathrm{E}(\alpha_j \mid c)\}^2 \geq \frac{c^2}{2\pi}. \tag{34}$$

When $0 < c < (2\pi\sqrt{2})^{2/3}$, *then*

$$\mathrm{Var}(\alpha_j \mid c) \leq \sqrt{2c} - \frac{c^2}{2\pi}. \tag{35}$$

5.4 Lemmas for Proposition 4

Lemma 1 *For any* $c > 0$, *the marginal half-horseshoe prior density evaluated at* $\alpha_j = x$, $h(x) \equiv \iint \pi(\alpha_j = x \mid \lambda_j, \tau)\pi(\tau)\pi(\lambda_j)\mathrm{d}\lambda_j\mathrm{d}\tau$, *can be written as*

$$h(x) = \exp\left(-\frac{x^2}{2}\right)\sqrt{\frac{2}{\pi^5 c^2}}\int_0^\infty \sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t.$$

Proof 4 (of Lemma 1) *Let* $\theta_j = c\tau\lambda_j$. *Then* $\pi(\alpha_j \mid \theta_j) = N^+\{\alpha_j \mid 0, \theta_j^2/(1+\theta_j^2)\}$. *In this case, Carvalho et al. (2010) give that the density of* $\tilde{\theta}_j \equiv \theta_j/c$ *is* $\pi(\tilde{\theta}_j) = 2\pi^{-2}\log\left(\tilde{\theta}_j^2\right)/\left(\tilde{\theta}_j^2 - 1\right)$. *This implies that the marginal prior density evaluated at* $\alpha_j = x$ *is the function*

$$\begin{aligned} h(x) &= \int_0^\infty \pi(\alpha_j = x \mid \tilde{\theta}_j)\pi(\tilde{\theta}_j)\mathrm{d}\tilde{\theta}_j \\ &= \left(\frac{2}{\sqrt{2\pi c^2}}\right)\left(\frac{2}{\pi^2}\right)\int_0^\infty \frac{\sqrt{1+u^2c^2}}{u}\exp\left\{-\frac{x^2(1+c^2u^2)}{2c^2u^2}\right\}\frac{\log u^2}{u^2-1}\mathrm{d}u \\ &= \exp\left(-\frac{x^2}{2}\right)\left(\frac{2}{\sqrt{2\pi c^2}}\right)\left(\frac{2}{\pi^2}\right)\int_0^\infty \frac{\sqrt{1+u^2c^2}}{u}\exp\left(-\frac{x^2}{2c^2u^2}\right)\frac{\log u^2}{u^2-1}\mathrm{d}u. \end{aligned}$$

Letting $t = u^{-2}$, *so that* $-\frac{1}{2t}\mathrm{d}t = \frac{1}{u}\mathrm{d}u$ *gives*

$$\begin{aligned} h(x) &= \exp\left(-\frac{x^2}{2}\right)\sqrt{\frac{2}{\pi^5 c^2}}\int_0^\infty \sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{-\log t}{t(1/t-1)}\mathrm{d}t \\ &= \exp\left(-\frac{x^2}{2}\right)\sqrt{\frac{2}{\pi^5 c^2}}\int_0^\infty \sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t \end{aligned}$$

Lemma 2 *Let the function $f(x,y) : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be defined as*

$$\alpha(x,t) = \sqrt{1 + \frac{c^2}{t}\frac{\log(t)}{t-1}} \exp\left(-\frac{x^2}{c^2}t\right)$$

Then,

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^\infty \alpha(x,t)\mathrm{d}t = \int_0^\infty \frac{\partial\alpha(x,t)}{\partial x}\mathrm{d}t$$

Proof 5 (of Lemma 2) *We have that*

$$\frac{\partial\alpha(x,t)}{\partial x} = -\frac{2t}{c^2}\sqrt{1 + \frac{c^2}{t}\frac{\log(t)}{t-1}}x\exp\left(-\frac{x^2}{c^2}t\right)$$

Also, for any $t > 0$,

$$0 < \frac{\log(t)}{t-1} < \frac{1}{\sqrt{t}}$$

This implies that

$$\left|\frac{\partial\alpha(x,t)}{\partial x}\right| = 2t\sqrt{1 + \frac{c^2}{t}\frac{\log(t)}{t-1}}x\exp\left(-\frac{x^2}{c^2}t\right) < 2x\sqrt{t+c^2}\exp\left(-\frac{x^2}{c^2}t\right)$$

Note that

$$\begin{aligned}
\int_0^\infty 2x\sqrt{t+c^2}\exp\left(-\frac{x^2}{c^2}t\right)\mathrm{d}t &= \exp\left(x^2\right)\int_{c^2}^\infty 4xu^2\exp\left(-\frac{x^2}{c^2}u^2\right)\mathrm{d}u \\
&\le c^3\exp\left(x^2\right)\int_{-\infty}^\infty 2xz^2\exp\{-x^2z^2\}\mathrm{d}z \\
&= 2\sqrt{\pi}c^3\exp\left(x^2\right)\int_{-\infty}^\infty z^2\frac{\sqrt{2}x}{\sqrt{2\pi}}\exp\left\{-\frac{(\sqrt{2}x)^2}{2}z^2\right\}\mathrm{d}z \\
&= 2\sqrt{\pi}c^3\exp\left(x^2\right)E(Z^2), \quad Z \sim N\{0, 1/(2x^2)\} \\
&= \frac{\sqrt{\pi}c^3\exp(x^2)}{x^2}.
\end{aligned}$$

Then, for any two positive numbers a, b with $0 < a < b < \infty$, we have

$$\int_a^b\int_0^\infty \left|\frac{\partial\alpha(x,t)}{\partial x}\right| dtdx \le \sqrt{\pi}c^3\int_a^b \frac{\exp(x^2)}{x^2}dx \le \sqrt{\pi}c^3\exp\left(b^2\right)\left(\frac{1}{a} - \frac{1}{b}\right) < \infty,$$

which gives that $\frac{\partial\alpha(x,t)}{\partial x}$ is locally integrable (Theorem 4, Talvila, 2001). This completes the proof.

5.5 Proof of Proposition 4a

Proof 6 *According to Lemma 1, we can write*

$$h(x) \propto \exp\left(-x^2/2\right)\alpha(x),$$

where

$$\begin{aligned}\alpha(x) &= \int_0^\infty \sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t \\ &= \int_0^2 \sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t + \int_2^\infty \sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t \\ &\leq c\{\sqrt{2+4/c^2}+c\,\mathrm{csch}^{-1}(c/\sqrt{2})\}+\sqrt{1+\frac{c^2}{2}}\frac{\sqrt{2}c}{x}\end{aligned}$$

Note that

$$\begin{aligned}&\int_0^2 \sqrt{1+c^2/t}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t \\ &\leq \int_0^2 \sqrt{1+c^2/t}\mathrm{d}t = c\{\sqrt{2+4/c^2}+c\,\mathrm{csch}^{-1}(c/\sqrt{2})\},\end{aligned}$$

where $\mathrm{csch}^{-1}(z) = \log\left(\sqrt{1+1/z^2}+1/z\right)$ *and*

$$\begin{aligned}&\int_2^\infty \sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t \\ &\leq \int_2^\infty \frac{1}{\sqrt{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\mathrm{d}t \leq \frac{\sqrt{2}c}{x}\end{aligned}$$

This implies that

$$h(x) \leq O\left(\frac{1}{x}\right).$$

For the derivative of $h(x)$*, we have*

$$h'(x) = -xh(x) - 2x\exp\left(-x^2/2\right)\sqrt{\frac{2}{\pi^5 c^2}}\beta(x)$$

where

$$\begin{aligned}\beta(x) &= \int_0^\infty t\sqrt{1+\frac{c^2}{t}}\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t \\ &\leq \int_0^2 \sqrt{t(t+c^2)}\mathrm{d}t + \sqrt{1+c^2/2}\int_2^\infty t\exp\left(-\frac{x^2}{2c^2}t\right)\frac{\log t}{t-1}\mathrm{d}t \\ &= \frac{1}{4}\left\{\sqrt{2}\sqrt{2+c^2}(4+c^2) - c^2\sinh^{-1}\left(\frac{\sqrt{2}}{c}\right)\right\} + 2c^2\sqrt{1+c^2/2}\frac{\{\exp(-2x^2)\log(2)+\mathrm{E}_1(x^2/c^2)\}}{x^2}.\end{aligned} \tag{36}$$

The exponential integral function $E_1(z)$ is defined as

$$\mathrm{E}_1(z) = \int_z^\infty \frac{\exp(-t)}{t}\mathrm{d}t = -\gamma - \log(z) - \sum_{k=1}^{\infty} \frac{(-z)^k}{kk!},$$

and $\gamma \approx 0.577$ is the Euler-Mascheroni constant. Thus,

$$h'(x) \leq O\left(-x^{-1}\log(x)\right),$$

as $x \to 0$

5.6 Proof of Proposition 4b

Proof 7 *The conditional prior probability of α_j given λ_j and τ for any $\lambda_j > 0$ and $\tau > 0$,*

$$\begin{aligned}
&\Pr(c^{1/\kappa} < \alpha_j \leq c^{\upsilon} \mid \lambda_j, \tau) \\
&= 2\left\{\Phi\left(\frac{c^{\upsilon}}{c\tau\lambda_j/\sqrt{1+c^2\tau^2\lambda_j^2}}\right) - \Phi\left(\frac{c^{1/\kappa}}{c\tau\lambda_j/\sqrt{1+c^2\tau^2\lambda_j^2}}\right)\right\} \\
&= 2\left\{\Phi\left(\frac{c^{\upsilon-1}}{\tau\lambda_j/\sqrt{1+c^2\tau^2\lambda_j^2}}\right) - \Phi\left(\frac{c^{1/\kappa-1}}{\tau\lambda_j/\sqrt{1+c^2\tau^2\lambda_j^2}}\right)\right\} \uparrow 1, \textit{as } c \downarrow 0.
\end{aligned}$$

By the monotone convergence theorem, the result implies that, as $c \to 0$,

$$\Pr(c^{1/\kappa} < \alpha_j \leq c^{\upsilon}) = \mathrm{E}\{\Pr(c^{1/\kappa} < \alpha_j \leq c^{\upsilon} \mid \lambda_j, \tau)\} \to 1,$$

5.7 Proof of Theorem 2

Proof 8 *The joint posterior density of $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ and τ given $\{m_k, n_k\}_{k=1}^K$ is proportional to*

$$\begin{aligned}
h(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau) &= \prod_{k=1}^{K-1} \xi_k^{m_k}(1-\xi_k)^{n_k-m_k} \prod_{l=1}^{K} \pi(\alpha_l \mid \lambda_l, \tau) \prod_{l=1}^{K} \pi(\lambda_l)\pi(\tau) \\
&= \prod_{k=1}^{K-1} \left(\frac{\sum_{j=1}^{k}\alpha_l}{\sum_{l=1}^{K}\alpha_l}\right)^{m_k} \left(\frac{\sum_{l=k+1}^{K}\alpha_l}{\sum_{l=1}^{K}\alpha_l}\right)^{n_k-m_k} \prod_{l=1}^{K} \pi(\alpha_l \mid \lambda_l, \tau) \prod_{l=1}^{K} \pi(\lambda_l)\pi(\tau) \\
&= g(\boldsymbol{\alpha}_{-j}, \alpha_j) \prod_{l=1}^{K} \pi(\alpha_l \mid \lambda_l, \tau) \prod_{l=1}^{K} \pi(\lambda_l)\pi(\tau)
\end{aligned}$$

for any $1 \leq j < K$*, where*

$$
\begin{aligned}
g(\boldsymbol{\alpha}_{-j}, \alpha_j) &= \prod_{k=1}^{K-1} \left(\frac{\sum_{j=1}^{k} \alpha_l}{\sum_{l=1}^{K} \alpha_l} \right)^{m_k} \left(\frac{\sum_{l=k+1}^{K} \alpha_l}{\sum_{l=1}^{K} \alpha_l} \right)^{n_k - m_k} \\
&= \prod_{k \geq j} \left(\frac{\sum_{l=1, l \neq j}^{k} \alpha_l + \alpha_j}{\sum_{l=1, l \neq j} \alpha_l + \alpha_j} \right)^{m_k} \left(\frac{\sum_{l=k+1}^{K} \alpha_l}{\sum_{l=1, l \neq j}^{K} \alpha_l + \alpha_j} \right)^{n_k - m_k} \times \\
&\quad \prod_{k < j} \left(\frac{\sum_{l=1}^{k} \alpha_l}{\sum_{l=1, l \neq j}^{K} \alpha_l + \alpha_j} \right)^{m_k} \left(\frac{\sum_{l=k+1, l \neq j}^{K} \alpha_l + \alpha_j}{\sum_{l=1, l \neq j}^{K} \alpha_l + \alpha_j} \right)^{n_k - m_k}
\end{aligned}
$$

In part ***a****, when* $c^{1/\kappa} < \alpha_j < c^{\upsilon}$*, for any* $\boldsymbol{\alpha}_{-j} = \{\alpha_l, l \neq j\}$*, it is straightforward to show that*

$$
g_l(\boldsymbol{\alpha}_{-j}, c) \leq g(\boldsymbol{\alpha}_{-j}, \alpha_j) \leq g_u(\boldsymbol{\alpha}_{-j}, c), \tag{37}
$$

where

$$
\begin{aligned}
g_l(\boldsymbol{\alpha}_{-j}, c) &= \prod_{k \geq j} \left(\frac{\sum_{l=1, l \neq j}^{k} \alpha_l + c^{1/\kappa}}{\sum_{l=1, l \neq j} \alpha_l + c^{1/\kappa}} \right)^{m_k} \left(\frac{\sum_{l=k+1}^{K} \alpha_l}{\sum_{l=1, l \neq j}^{K} \alpha_l + c^{\upsilon}} \right)^{n_k - m_k} \times \\
&\quad \prod_{k < j} \left(\frac{\sum_{l=1}^{k} \alpha_l}{\sum_{l=1, l \neq j}^{K} \alpha_l + c^{\upsilon}} \right)^{m_k} \left(\frac{\sum_{l=k+1, l \neq j}^{K} \alpha_l + c^{1/\kappa}}{\sum_{l=1, l \neq j}^{K} \alpha_l + c^{1/\kappa}} \right)^{n_k - m_k}
\end{aligned} \tag{38}
$$

$$
\begin{aligned}
g_u(\boldsymbol{\alpha}_{-j}, c) &= \prod_{k \geq j} \left(\frac{\sum_{l=1, l \neq j}^{k} \alpha_l + c^{\upsilon}}{\sum_{l=1, l \neq j} \alpha_l + c^{\upsilon}} \right)^{m_k} \left(\frac{\sum_{l=k+1}^{K} \alpha_l}{\sum_{l=1, l \neq j}^{K} \alpha_l + c^{1/\kappa}} \right)^{n_k - m_k} \times \\
&\quad \prod_{k < j} \left(\frac{\sum_{l=1}^{k} \alpha_l}{\sum_{l=1, l \neq j}^{K} \alpha_l + + c^{1/\kappa}} \right)^{m_k} \left(\frac{\sum_{l=k+1, l \neq j}^{K} \alpha_l + c^{\upsilon}}{\sum_{l=1, l \neq j}^{K} \alpha_l + c^{\upsilon}} \right)^{n_k - m_k}.
\end{aligned} \tag{39}
$$

Note that for any $\boldsymbol{\alpha}_{-j}$*. Then*

$$
\lim_{c \to 0} g_u(\boldsymbol{\alpha}_{-j}, c) = \lim_{c \to 0} g_l(\boldsymbol{\alpha}_{-j}, c) = g(\boldsymbol{\alpha}_{-j}, 0). \tag{40}
$$

Let $\mathbb{R}_+^d = (0, \infty)^d$ *and the marginal posterior probability of* $c^{1/\kappa} < \alpha_j \leq c^{\upsilon}$ *is given by*

$$
\begin{aligned}
&\Pr(c^{1/\kappa} < \alpha_j \leq c^{\upsilon} \mid Y, X) \\
&= \frac{\int_0^{\infty} \int_{\mathbb{R}_+^K} \int_{\mathbb{R}_+^{K-1}} \int_{c^{1/\kappa}}^{c^{\upsilon}} h(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau) d\alpha_j d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau}{\int_0^{\infty} \int_{\mathbb{R}_+^K} \int_{\mathbb{R}_+^{K-1}} \int_0^{\infty} h(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau) d\alpha_j d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau} \\
&= \frac{\int_0^{\infty} \int_{\mathbb{R}_+^K} \int_{\mathbb{R}_+^{K-1}} \int_{c^{1/\kappa}}^{c^{\upsilon}} g(\boldsymbol{\alpha}_{-j}, \alpha_j) \pi(\alpha_j \mid \lambda_j, \tau) \prod_{l \neq j} \pi(\alpha_l \mid \lambda_l, \tau) \pi(\boldsymbol{\lambda}) \pi(\tau) d\alpha_j d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau}{\int_0^{\infty} \int_{\mathbb{R}_+^K} \int_{\mathbb{R}_+^{K-1}} \int_0^{\infty} g(\boldsymbol{\alpha}_{-j}, \alpha_j) \pi(\alpha_j \mid \lambda_j, \tau) \prod_{l \neq j} \pi(\alpha_l \mid \lambda_l, \tau) \pi(\boldsymbol{\lambda}) \pi(\tau) d\alpha_j d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau} \\
&= \frac{I(c^{1/\kappa}, c^{\nu})}{I(0, c^{1/\kappa}) + I(c^{1/\kappa}, c^{\nu}) + I(c^{\nu}, \infty)}
\end{aligned}
$$

where

$$I(a,b)$$
$$= \int_0^\infty \int_{\mathbb{R}_+^K} \int_{\mathbb{R}_+^{K-1}} \int_a^b g(\boldsymbol{\alpha}_{-j}, \alpha_j)\pi(\alpha_j \mid \lambda_j, \tau) \prod_{l \neq j} \pi(\alpha_l \mid \lambda_l, \tau)\pi(\boldsymbol{\lambda})\pi(\tau) d\alpha_j d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau.$$

This implies that

$$I_l(c) \leq I(c^{1/\kappa}, c^\nu) \leq I_u(c) \tag{41}$$

$$I_l(c) = \int_0^\infty \int_{\mathbb{R}_+^{K-1}} \int_{\mathbb{R}_+^K} \left\{ \int_{c^{1/\kappa}}^{c^\nu} \pi(\alpha_j \mid \lambda_j, \tau) d\alpha_j \right\} \left\{ g_l(\boldsymbol{\alpha}_{-j}, c) \prod_{l \neq j} \pi(\alpha_l \mid \lambda_l, \tau) \right\} \pi(\boldsymbol{\lambda})\pi(\tau) d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau$$
$$I_u(c) = \int_0^\infty \int_{\mathbb{R}_+^{K-1}} \int_{\mathbb{R}_+^K} \left\{ \int_{c^{1/\kappa}}^{c^\nu} \pi(\alpha_j \mid \lambda_j, \tau) d\alpha_j \right\} \left\{ g_u(\boldsymbol{\alpha}_{-j}, c) \prod_{l \neq j} \pi(\alpha_l \mid \lambda_l, \tau) \right\} \pi(\boldsymbol{\lambda})\pi(\tau) d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau$$
$$I_u(a,b) = \int_0^\infty \int_{\mathbb{R}_+^{K-1}} \int_{\mathbb{R}_+^K} \left\{ \int_a^b \pi(\alpha_j \mid \lambda_j, \tau) d\alpha_j \right\} \left\{ g_u(\boldsymbol{\alpha}_{-j}, c) \prod_{l \neq j} \pi(\alpha_l \mid \lambda_l, \tau) \right\} \pi(\boldsymbol{\lambda})\pi(\tau) d\boldsymbol{\alpha}_{-j} d\boldsymbol{\lambda} d\tau$$

Then

$$\lim_{c \to 0} \frac{I_l(c)}{I_u(c)} = 1, \quad \lim_{c \to 0} I_u(0, c^{1/\kappa}) = 0, \quad \lim_{c \to 0} I_u(c^\nu, \infty) = 0$$

Thus, for any $n \geq 1$, as $c \to 0$,

$$\Pr(c^{1/\kappa} < \alpha_j \leq c^\upsilon \mid Y, X)$$
$$\geq \frac{I_l(c)}{I_u(0, c^{1/\kappa}) + I_u(c) + I_u(c^\nu, \infty)} \to 1,$$

in $P_{\theta_0}^n$ probability.

*In part **b**, when $Mn^{-1/(\kappa\nu)} < \alpha_j \leq Mn^{-1}$ for any fixed $M > 0$. Similar to Theorem 1**b**, under the actual distribution of data given the true parameters $\theta_0 = \{\eta_{0k}, \xi_{0k}\}_{k=1}^K$, where $\eta_{0k} = \Pr(X_i = k)$ and $\xi_{0k} = \Pr(Y_i = 1 \mid X_i = k)$. By the law of the large numbers, as $n \to \infty$,*

$$\frac{m_k}{n} \to \eta_{0k}\xi_{0k} \text{ and } \frac{n_k - m_k}{n} \to \eta_{0k}(1 - \xi_{0k})$$

in $P_{\theta_0}^\infty$ probability. Since $c = (Mn)^{-1/\nu}$, and for any $\boldsymbol{\alpha}_{-j}$, as $n \to \infty$,

$$\frac{g_l(\boldsymbol{\alpha}_{-j}, c)}{f_l(\boldsymbol{\alpha}_{-j}, n)} \to 1 \quad \text{and} \quad \frac{g_u(\boldsymbol{\alpha}_{-j}, c)}{f_u(\boldsymbol{\alpha}_{-j}, n)} \to 1$$

in $P_{\theta_0}^{\infty}$ probability, where

$$f_l(\boldsymbol{\alpha}_{-j}, n) = \prod_{k \geq j} \left(\frac{\sum_{l=1, l\neq j}^{k} \alpha_l + (Mn)^{-1/\kappa\nu}}{\sum_{l=1, l\neq j} \alpha_l + (Mn)^{-1/\kappa\nu}} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1}^{K} \alpha_l}{\sum_{l=1, l\neq j}^{K} \alpha_l + (Mn)^{-1}} \right)^{\eta_{0k}(1-\xi_{0k})n} \times \prod_{k < j} \left(\frac{\sum_{l=1}^{k} \alpha_l}{\sum_{l=1, l\neq j}^{K} \alpha_l + (Mn)^{-1}} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1, l\neq j}^{K} \alpha_l + (Mn)^{-1/\kappa\nu}}{\sum_{l=1, l\neq j}^{K} \alpha_l + (Mn)^{-1/\kappa\nu}} \right)^{\eta_{0k}(1-\xi_{0k})n}, \tag{42}$$

and

$$f_u(\boldsymbol{\alpha}_{-j}, n) = \prod_{k \geq j} \left(\frac{\sum_{l=1, l\neq j}^{k} \alpha_l + (Mn)^{-1}}{\sum_{l=1, l\neq j} \alpha_l + (Mn)^{-1}} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1}^{K} \alpha_l}{\sum_{l=1, l\neq j}^{K} \alpha_l + (Mn)^{-1/\kappa\nu}} \right)^{\eta_{0k}(1-\xi_{0k})n} \times \prod_{k < j} \left(\frac{\sum_{l=1}^{k} \alpha_l}{\sum_{l=1, l\neq j}^{K} \alpha_l + (Mn)^{-1/\kappa\nu}} \right)^{\eta_{0k}\xi_{0k}n} \left(\frac{\sum_{l=k+1, l\neq j}^{K} \alpha_l + (Mn)^{-1}}{\sum_{l=1, l\neq j}^{K} \alpha_l + (Mn)^{-1}} \right)^{\eta_{0k}(1-\xi_{0k})n}, \tag{43}$$

As $n \to \infty$, it can be shown that

$$\frac{f_l(\boldsymbol{\alpha}_{-j}, n)}{f_u(\boldsymbol{\alpha}_{-j}, n)} \to 1 \tag{44}$$

in $P_{\theta_0}^{\infty}$ probability. Thus

$$\frac{g_l(\boldsymbol{\alpha}_{-j}, 1/(Mn)^{-1/\nu})}{g_u(\boldsymbol{\alpha}_{-j}, 1/(Mn)^{-1/\nu})} \to 1 \tag{45}$$

in $P_{\theta_0}^{\infty}$ probability.

Let $J(n) = I(Mn^{-1/(\kappa\nu)}, Mn^{-1})$, $J_l(n) = I_l\{(Mn)^{-1/\nu}\}$ and $J_u(n) = I_u\{(Mn)^{-1/\nu}\}$. Then we have

$$J_l(n) \leq J(n) \leq J_u(n).$$

*Then by proposition 4**b**, as $n \to \infty$,*

$$\int_{(Mn)^{-1/\nu\kappa}}^{(Mn)^{-1/\nu}} \pi(\alpha_j \mid \lambda_j, \tau) d\alpha_j \to 1.$$

Then

$$\lim_{n\to\infty} \frac{J_l(n)}{J_u(n)} = \lim_{n\to\infty} \frac{E\left[\int_{(Mn)^{-1/\nu\kappa}}^{(Mn)^{-1/\nu}} \pi(\alpha_j \mid \lambda_j, \tau) d\alpha_j g_l(\boldsymbol{\alpha}_{-j}, 1/(Mn)^{-1/\nu}) \mid Y, X\right]}{E\left[\int_{(Mn)^{-1/\nu\kappa}}^{(Mn)^{-1/\nu}} \pi(\alpha_j \mid \lambda_j, \tau) d\alpha_j g_u(\boldsymbol{\alpha}_{-j}, 1/(Mn)^{-1/\nu}) \mid Y, X\right]} = 1$$

in $P_{\theta_0}^{\infty}$ probability, where $E(\cdot \mid Y, X)$ is taken with respect to the joint posterior distribution of $\tau, \lambda, \boldsymbol{\alpha}_{-j}$. In addition,

$$I_u\{0, (Mn)^{-1/\kappa\nu}\} \to 0, \quad I_u\{(Mn)^{-1}), \infty\} \to 0$$

in $P_{\theta_0}^{\infty}$ probability. Thus, for any fixed $M > 0$, as $n \to \infty$,

$$\begin{aligned}&\Pr\{(Mn)^{1/\kappa\nu}\alpha_j \leq (Mn)^{-1} \mid Y, X\} \\ &\geq \frac{J_l(n)}{I_u\{0, (Mn)^{-1/\kappa\nu}\} + J_u(n) + I_u\{(Mn)^{-1}), \infty\}} \to 1\end{aligned}$$

in $P_{\theta_0}^{\infty}$ probability.

References

BARLOW, R. E., BARTHOLOMEW, D., BREMNER, J. M. & BRUNK, H. D. (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression*. New York: John Wiley & Sons.

BETANCOURT, M. (2016). Diagnosing suboptimal cotangent disintegrations in hamiltonian monte carlo. *arXiv preprint arXiv:1604.00695* .

BOONSTRA, P. S. & BARBARO, R. P. (2018). Incorporating historical models with adaptive bayesian updates. *Biostatistics* **In press**.

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

CARVALHO, C., POLSON, N. & SCOTT, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

CARVALHO, C. M., POLSON, N. G. & SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, D. van Dyk & M. Welling, eds., vol. 5 of *Proceedings of Machine Learning Research*. Clearwater Beach, Florida USA: PMLR.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**, 1348–1360.

GELFAND, A. E. & KUO, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika* **78**, 657–666.

GELMAN, A., RUBIN, D. B. et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science* **7**, 457–472.

LI, W. & FU, H. (2017). Bayesian isotonic regression dose-response model. *Journal of Biopharmaceutical Statistics* **27**, 824–833.

LI, Y., BEKELE, B. N., JI, Y. & COOK, J. D. (2008). Dose–schedule finding in phase i/ii clinical trials using a bayesian isotonic transformation. *Statistics in medicine* **27**, 4895–4913.

LIN, L. & DUNSON, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101**, 303–317.

LIU, C., MARTIN, R. & SYRING, N. (2013). Simulating from a gamma distribution with small shape parameter. *arXiv preprint arXiv:1302.1884* .

NEELON, B. & DUNSON, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60**, 398–406.

OHLSSEN, D. & RACINE, A. (2015). A flexible bayesian approach for modeling monotonic dose-response relationships in drug development trials. *Journal of Biopharmaceutical Statistics* **25**, 137–156.

ORON, A. P. (2017). *cir: Centered Isotonic Regression and Dose-Response Utilities*. R package version 2.0.0.

OWEN, D. R., BOONSTRA, P. S., VIGLIANTI, B. L., BALTER, J. M., SCHIPPER, M. J., JACKSON, W. C., NAQA, I. E., JOLLY, S., HAKEN, R. K. T., KONG, F.-M. S. & MATUSZAK, M. M. (2018). Modeling patient-specific dose-function response for enhanced characterization of personalized functional damage. *International Journal of Radiation Oncology*Biology*Physics* **102**, 1265 – 1275.

OWEN, D. R., SUN, Y., BOONSTRA, P. S. et al. (2020). Investigating the spect dose-function metrics associated with radiation-induced lung toxicity risk in non-small cell lung cancer patient undergoing radiation therapy. *submitted* .

PIIRONEN, J. & VEHTARI, A. (2015). Projection predictive variable selection using Stan+R. Tech. rep. ArXiv preprint arXiv:1508.02502.

PIIRONEN, J. & VEHTARI, A. (2017a). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh & J. Zhu, eds., vol. 54 of *Proceedings of Machine Learning Research*. Fort Lauderdale, FL, USA: PMLR.

PIIRONEN, J. & VEHTARI, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11**, 5018–5051.

RAMGOPAL, P., LAUD, P. & SMITH, A. (1993). Nonparametric bayesian bioassay with prior constraints on the shape of the potency curve. *Biometrika* **80**, 489–498.

RAMSEY, F. L. (1972). A bayesian approach to bioassay. *Biometrics* **28**, 841–858.

SHAKED, M. & SINGPURWALLA, N. D. (1990). A bayesian approach for quantile and response probability estimation with applications to reliability. *Annals of the Institute of Statistical Mathematics* **42**, 1–19.

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

STAN DEVELOPMENT TEAM (2018). RStan: the R interface to Stan. R package version 2.19.2.

STAN DEVELOPMENT TEAM (2019). *Stan Modeling Language User's Guide and Reference Manual, Version 2.19*. Http://mc-stan.org/.

TALVILA, E. (2001). Necessary and sufficient conditions for differentiating under the integral sign. *The American Mathematical Monthly* **108**, 544–548.

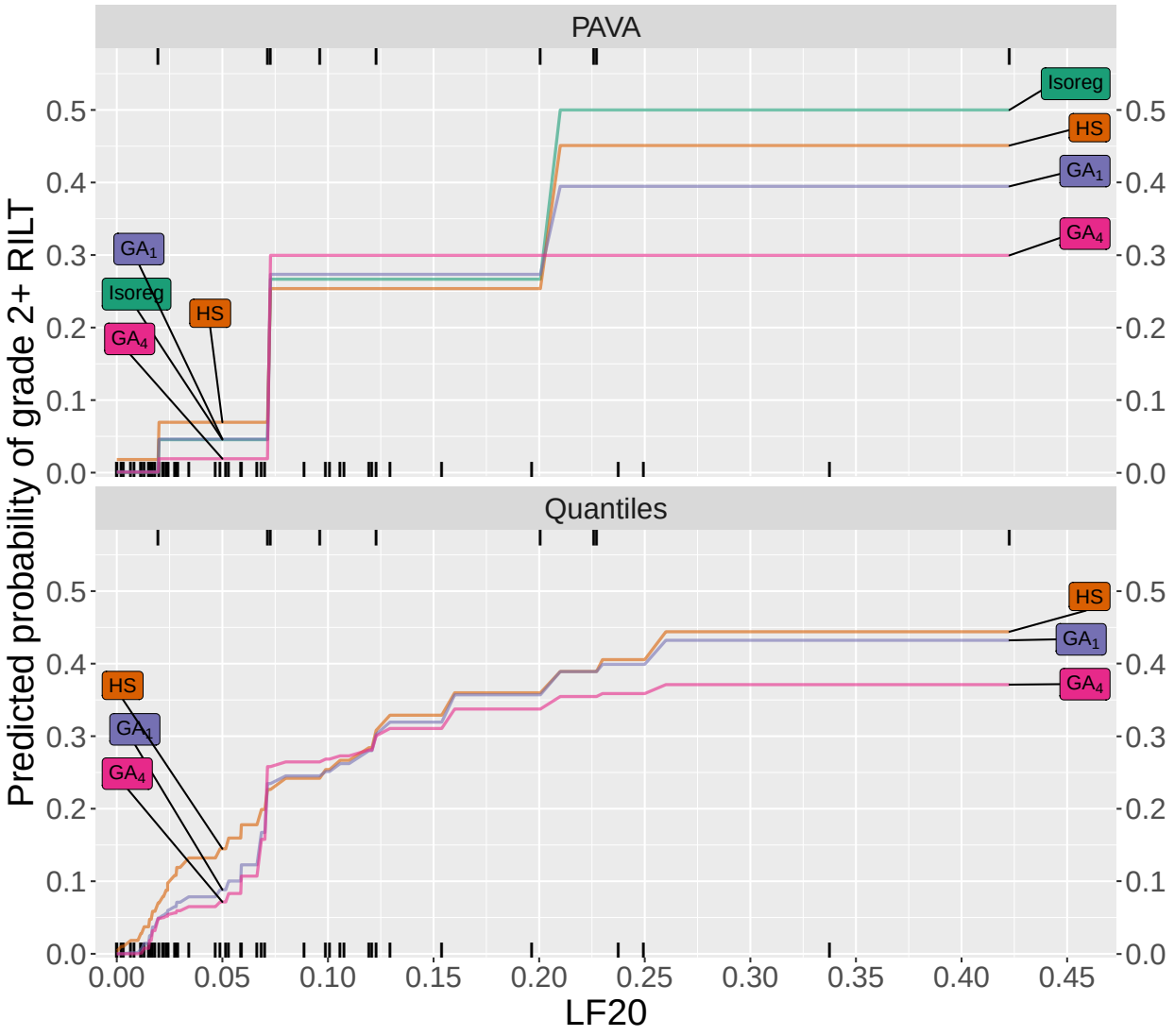

Figure 4: Model-based probabilities of grade 2+ radiation-induced lung toxicity (RILT) as a function of proportion of lung that is both low-functioning and received 20 Gy. The rugs at the top and bottom of each panel denote the data.