Simultaneous transformation and rounding (STAR) models for integer-valued data

Daniel R. Kowal

Department of Statistics
Rice University
Houston, TX
e-mail: daniel.kowal@rice.edu
url: www.danielrkowal.com

and

Antonio Canale

Department of Statistical Sciences
University of Padova
Padova, Italy
e-mail: canale@stat.unipd.it

Abstract: We propose a simple yet powerful framework for modeling integer-valued data, such as counts, scores, and rounded data. The data-generating process is defined by Simultaneously Transforming and Rounding (STAR) a continuous-valued process, which produces a flexible family of integer-valued distributions capable of modeling zero-inflation, bounded or censored data, and over- or underdispersion. The transformation is modeled as unknown for greater distributional flexibility, while the rounding operation ensures a coherent integer-valued data-generating process. An efficient MCMC algorithm is developed for posterior inference and provides a mechanism for adaptation of successful Bayesian models and algorithms for continuous data to the integer-valued data setting. Using the STAR framework, we design a new Bayesian Additive Regression Tree model for integer-valued data, which demonstrates impressive predictive distribution accuracy for both synthetic data and a large healthcare utilization dataset. For interpretable regression-based inference, we develop a STAR additive model, which offers greater flexibility and scalability than existing integer-valued models. The STAR additive model is applied to study the recent decline in Amazon river dolphins.

MSC 2010 subject classifications: Primary 62F15, 62G08; secondary 62M20.
Keywords and phrases: Additive models, BART, count data, nonparametric regression, prediction.

Contents

1. Introduction

A challenging scenario for prediction and inference occurs when the outcome variables are integer-valued, such as counts, (test) scores, or rounded data. Integer-valued data are common in many fields, including epidemiology (Osthus et al., 2018; Kowal, 2019), ecology (Dorazio et al., 2005), and insurance (Bening and Korolev, 2012), among many others (Cameron and Trivedi, 2013). Counts often serve as an indicator of demand, such as the demand for medical services (Deb and Trivedi, 1997), emergency medical services (Matteson et al., 2011), and call center access (Shen and Huang, 2008). Integer-valued data are discrete data, and exhibit a variety of complex distributional features including zero-inflation, skewness, over- or underdispersion, and in some cases may be bounded or censored. Consequently, prediction and modeling of integer-valued data—in the presence of predictors, over time intervals, and across spatial locations—remains a significant challenge.

The most widely-used models for count or integer-valued data build upon the Poisson distribution. However, the limitations of the Poisson distribution are well-known: the distribution is not sufficiently flexible in practice and cannot account for zero-inflation or over- and underdispersion. A common strategy is to generalize the Poisson model by introducing additional parameters, such as the quasi-Poisson (McCullagh and Nelder, 1989), negative-binomial (Hilbe, 2011), zero-inflated Poisson or negative-binomial (Cameron and Trivedi, 2013;

Neelon, 2019), lognormal Poisson (Zhou et al., 2012), restricted generalized Poisson (Famoye, 1993), and Conway-Maxwell Poisson models (Shmueli et al., 2005; Lord et al., 2008; Sellers and Shmueli, 2010). A fundamental limitation of these approaches is that the additional parameters can introduce formidable challenges for estimation and computational scalability, especially in conjunction with regression, temporal, or spatial models.

In practice, however, it is exceedingly common for the discrete nature of the data to be ignored. Practitioners often log- or square-root-transform the observed integer-valued data and subsequently apply methods designed for continuous or Gaussian data. However, transformations to Gaussianity are ineffective for small counts (Warton, 2018), while log-transformations introduce difficulties in the presence of zeros (O'Hara and Kotze, 2010). More broadly, these approaches are not well-defined for integer-valued data: the data-generating process for a (continuously-transformed) Gaussian model cannot produce discrete data, which immediately amplifies model misspecification, limits interpretability, and undermines the reliability of inference and predictive distributions.

To address these challenges, we propose a coherent modeling framework for integer-valued data. The process is defined by *simultaneously transforming and rounding* (STAR) a continuous-valued process. First, a *continuous-valued process* is specified to model the dependence between (latent) variables. We focus on conditionally Gaussian regression models, but the STAR framework applies more broadly. Second, the latent variables are *transformed* for greater distributional flexibility. While the transformation may be specified in advance, such as logarithmic or square-root, we develop both parametric and nonparametric approaches to learn the transformation from the data, which improves predictive accuracy. Lastly, the transformed latent variables are filtered through a *rounding operator* mapping them to the (nonnegative) integers. This construction is inspired by the popular approach of transforming count data and applying Gaussian models, yet produces a mathematically consistent and well-defined integer-valued process. Importantly, we show that STAR processes are not merely *valid* integer-valued distributions, but also offer modeling *flexibility*, and can account for zero-inflation, bounded or censored data, and over- or underdispersion.

Another major benefit of STAR is its computational modularity: using a simple and efficient data augmentation technique, existing computational tools for Bayesian inference under continuous data models can be used for Bayesian inference under STAR models. As a result, STAR provides a cohesive framework for seamlessly adapting state-of-the-art continuous data models and algorithms to the integer-valued data setting. Using the STAR framework, we design—among others—a new Bayesian Additive Regression Tree (BART) model for integer-valued data. The resulting BART-STAR model combines the integer-valued distributional flexibility provided by STAR with the predictive and computational advantages inherent to BART. For synthetic data and a large healthcare utilization dataset (Sections 4-5), the predictive performance of BART-STAR far exceeds that of competing methods which do not include *both* transformation and rounding in terms of out-of-sample predictive accuracy, model adequacy, and computational scalability.

We also apply STAR to study the recent decline in the tucuxi dolphin population, which inhabit the Amazon River. Using field survey data conducted by da Silva et al. (2018) from 1994 to 2017, we develop a STAR additive model for the number of observed tucuxi dolphins, which includes a smooth regression term for important predictor variables such as the year, day-of-year, and water level. The STAR additive model is interpretable yet flexible, and demonstrates favorable performance in model fit and computational efficiency relative to existing integer-valued models.

The remainder of the paper is organized as follows. Section 2 introduces the STAR framework, develops models for the unknown transformation, describes important properties, and discusses computational details for posterior inference. Section 3 provides example STAR models, which are applied to simulated data in Section 4 and real data in Sections 5 and 6. Section 7 concludes. Additional simulation results and empirical comparisons are in the Appendix. Methods are implemented in the R package rSTAR available on GitHub: www.github.com/drkowal/rSTAR.

2. Simultaneously transforming and rounding

Consider a count-valued stochastic process $y \colon \mathcal{X} \to \mathcal{N}$, where $\mathcal{X}$ may correspond to predictors, times, or spatial locations and $\mathcal{N} = \{0, \dots, \infty\}$. Although we focus on the nonnegative integers, our procedure may be trivially modified for integer-valued data and rounded data. Our goal is to introduce a flexible method for constructing a joint probability distribution for y that simultaneously builds upon successful approaches for continuous stochastic processes (observed on $\mathbb{R}$ or $\mathbb{R}^+$), yet produces well-defined distribution on $\mathcal{N}$.

To this end, we first introduce continuous-valued process $y^* \colon \mathcal{X} \to \mathcal{T}$, $\mathcal{T} \subseteq \mathbb{R}$ related to the observed count-valued data y via

$$y = h(y^*), \tag{1}$$

where $h \colon \mathcal{T} \to \mathcal{N}$ is a *rounding* operator that sets $y(x) = j$ when $y^*(x) \in \mathcal{A}_j$ and $\{\mathcal{A}_j\}_{j=0}^{\infty}$ is a known partition of $\mathcal{T}$. We select h to guarantee the correct support for y: for example, the floor function $\mathcal{A}_j = [a_j, a_{j+1}) = [j, j+1)$ for $j \in \mathcal{N}$ implies $y \in \mathcal{N}$; see Section 2.2 for other examples. The process y^* operates as a continuous proxy for the observed counts y, which is more convenient for modeling, yet has a simple mapping to the observable data in (1). Naturally, the properties of the count-valued process y will be determined by the rounding operator h and the distribution of the continuous-valued process y^*.

We propose to induce a distribution on y^* by *transforming* y^* and specifying a distribution Π_θ on the transformed scale:

$$g(y^*) = z^*, \quad z^* \sim \Pi_\theta, \tag{2}$$

where $g \colon \mathcal{T} \to \mathbb{R}$ is a strictly monotone function. Model (2) is inspired by the common practice of transforming count data prior to application of continuous (Gaussian) models. However, the STAR framework of (1)-(2) defines an integer-

valued process for y, in which the transformation g may be modeled as unknown for greater distributional flexibility.

While STAR is sufficiently general to incorporate any continuous family of stochastic process Π_θ for the latent z^*, an important special case of (2) is the conditionally Gaussian regression model:

$$z^*(x) = \mu(x) + \epsilon(x), \quad \epsilon(x) \overset{indep}{\sim} N(0, \sigma^2(x)), \tag{3}$$

where $\mu(x)$ is the conditional expectation of $z^*(x)$ and the errors $\epsilon(x)$ are independent but possibly heteroscedastic, conditional on $x \in \mathcal{X}$. Examples of (3) include linear and additive models (Section 3.1) and BART (Section 3.2), with extensions for mixed effects models, spatio-temporal models, dynamic linear models, and factor models, among others.

STAR deviates from Canale and Dunson (2011) and Canale and Dunson (2013), which pair (1) with a model directly on y^*. Our key innovation is the coupling of the *transformation* (2) with the regression model (3). The transformation g, which we model as unknown, endows the integer-valued process y with greater distributional flexibility, yet leaves model (3) unchanged. This construction allows seamless integration of Bayesian models *and* algorithms for continuous data of the common form (3) into the integer-valued STAR framework, with efficient posterior inference available via a general MCMC algorithm (Section 2.3). As demonstrated extensively in the simulations and applications (Sections 4-6), models that fail to include *both* rounding and transformation do not match the predictive performance of STAR models.

The distribution of y is completely determined by the rounding operator h, the transformation g, and the distribution Π_θ. Specifically, the probability mass associated to $y(x) = j$ for each integer $j \in \mathcal{N}$ is

$$\mathbb{P}\{y(x) = j\} = \mathbb{P}\{y^*(x) \in \mathcal{A}_j\} = \mathbb{P}\{z^*(x) \in g(\mathcal{A}_j)\}. \tag{4}$$

The distribution of z^* is given by Π_θ, while $g(\mathcal{A}_j)$ is determined by the transformation g and the rounding operator h. For model (3) and $\mathcal{A}_j = [a_j, a_{j+1})$, (4) simplifies to

$$\mathbb{P}\{y(x) = j\} = \Phi\left(\frac{g(a_{j+1}) - \mu(x)}{\sigma(x)}\right) - \Phi\left(\frac{g(a_j) - \mu(x)}{\sigma(x)}\right). \tag{5}$$

The distribution in (5) is related to, yet distinct from, ordinal regression (McCullagh, 1980; Dale, 1986; Agresti, 2013). In ordinal regression, each term $g(a_j)$ in (5) is replaced by an unknown latent threshold, say ω_j, with an ordering constraint $\omega_j \leq \omega_{j+1}$ for all j. However, the latent thresholds ω_j are based only on the *ranks* of the observed data, and therefore ignore the information contained in the numeric values of the observed counts. Furthermore, since each threshold ω_j is unknown, ordinal regression introduces a new parameter for each unique data value, and therefore produces a heavily-parametrized model that is challenging to estimate. By comparison, STAR is substantially more parsimonious: if g is known, no new parameters are needed, while if g is unknown, only a small number of parameters are needed (see Section 2.1).

STAR is fundamentally different from simply rounding the predictions from a continuous data model Π_θ. *Post hoc* rounding ignores the discrete nature of the data in model-fitting, and consequently introduces a disconnect between the *fitted model* and the model used for *prediction*. STAR clearly avoids this issue, and maintains the benefits of using well-known models for continuous data while producing a coherent integer-valued predictive distribution.

2.1. The transformation g

The transformation g is a crucial component of STAR. When $g(t) = t$ and z^* is a draw from a Gaussian process, STAR simplifies to Canale and Dunson (2013). However, the identity transformation is suboptimal in many cases (see Sections 5 and 6). The popularity of log-linear models for count data, especially Poisson and negative-binomial models, suggests that regression effects $\mu(x)$ are often multiplicative for count data, and that the log-transformation $g(t) = \log(t)$ may be preferable for many applications. Similarly, the square-root transformation $g(t) = \sqrt{t}$ is the variance-stabilizing transformation of the Poisson distribution, and therefore is a common choice in applications of Gaussian methods to transformed count data. Empirically, the simulation studies and real data analyses in Sections 4-6 demonstrate that the transformation g provides substantial improvements in modeling flexibility and accuracy relative to an untransformed approach.

When g is fixed and known, the only unknowns are the parameters θ in the distribution Π_θ of the latent data z^* in (2). Relative to the analogous transformed continuous-valued model, say $g(y) \sim \Pi_\theta$, the number of parameters is the same, yet STAR produces a coherent integer-valued process. A fixed transformation g shares some characteristics with the link function of a generalized linear model (GLM) (McCullagh and Nelder, 1989). For GLMs, the link function maps the expectation of an exponential family distribution to $\mathbb{R}$, which is modeled using a linear predictor. By comparison, STAR maps the continuous-valued y^* to $\mathbb{R}$ and under (3) models $\mathbb{E}[g\{y^*(x)\}|x] = \mathbb{E}\{z^*(x)|x\} = \mu(x)$.

For general application of STAR, pre-specification of a transformation g is restrictive. By allowing the data to inform g, the implied distribution for y becomes more flexible, and the risk of model misspecification is lessened. For GLMs, Mallick and Gelfand (1994) similarly relax the assumption of a known link function, adopting a nonparametric approach. For STAR, we require that the functions g satisfy the following properties: (i) *monotonicity*, which preserves the ordering of the observed integers in the transformed latent space; (ii) *smoothness*, which provides regularization by encouraging information-sharing among nearby values; and (iii) *shrinkage* toward a pre-specified transformation, such as log or square-root.

A natural parametric specification for g satisfying the aforementioned criteria is the (signed) *Box-Cox* transformation (Box and Cox, 1964):

$$g(t; \lambda) = \{\mathrm{sgn}(t)|t|^\lambda - 1\}/\lambda, \quad \lambda > 0 \tag{6}$$

with $g(t; \lambda = 0) = \log(t)$. Box-Cox functions are a popular choice for transforming continuous data towards Gaussianity, which in the present setting is similar to (2) when Π_θ is Gaussian. Important special cases of (6) include the (shifted) identity transformation $g(t; \lambda = 1) = t - 1$, the (shifted and scaled) square-root transformation $g(t; \lambda = 1/2) = 2\sqrt{|t|} - 2$, and the log-transformation. To learn the shape of the transformation, we place a prior on λ: we recommend $\lambda \sim N(1/2, 1)$ truncated to $[0, 3]$, which shrinks g toward the (shifted and scaled) square-root transformation.

For additional flexibility, we also consider fully nonparametric specification for g. Consider an I-spline basis expansion (Ramsay, 1988) for g:

$$g(t) = b'_I(t)\gamma, \tag{7}$$

where b_I is an L-dimensional vector of I-spline basis functions and γ are the unknown basis coefficients. Since each I-spline basis function is monotone increasing, we ensure monotonicity of g by restricting the elements of γ to be positive.

We propose a prior for γ in (7) that simultaneously enforces monotonicity, smoothness, and shrinkage toward a pre-specified transformation. However, care must be taken to ensure identifiability of the STAR model and retain interpretability of the parameters θ in Π_θ. For model (3), arbitrary shifting and scaling of g can be matched by shifting and scaling of μ and σ. The parametric transformation (6) preserves identifiability: $g(1, \lambda) = 0$ for all λ (shift constraint) and the prior on λ is weakly informative (scale constraint). For nonparametric g in (7), we resolve the identifiability issue by fixing $g(0) = 0$ (shift constraint), which is satisfied automatically due to the I-spline construction, and $\lim_{t \to \infty} g(t) = 1$ (scale constraint), which is enforced by constraining $\sum_{\ell=1}^{L} \gamma_\ell = 1$. Specifically, let

$$\gamma_\ell = \tilde{\gamma}_\ell / \sum_{k=1}^{L} \tilde{\gamma}_k, \quad \tilde{\gamma}_\ell \overset{indep}{\sim} N_+(\mu_{\gamma_\ell}, \sigma_\gamma^2), \quad \ell = 1, \ldots, L, \tag{8}$$

where N_+ is the half-normal distribution. Clearly, $\gamma_\ell > 0$ for each ℓ and $\sum_{\ell=1}^{L} \gamma_\ell = 1$, which guarantees monotonicity of g and preserves identifiability of model (3). We select the prior mean $\mu_\gamma = (\mu_{\gamma_1}, \ldots, \mu_{\gamma_L})'$ such that $g(t)$ is *a priori* centered around a parametric function of interest, such as (6) with fixed $\lambda = \lambda_0$, and model σ_γ^2 with an inverse-Gamma prior to allow the data to determine the amount of shrinkage toward the parametric function of interest. Let $\boldsymbol{t}_g = (0, 1, \ldots, a_{\max y_i + 1})'$ and let B_I be the I-spline basis evaluated at $\boldsymbol{t}_g$, so $g(\boldsymbol{t}_g) = B_I \gamma$. We solve $\tilde{\mu}_\gamma = \arg\min_{\mu_\gamma} ||g(\boldsymbol{t}_g; \lambda_0) - B_I \mu_\gamma||^2$ subject to $\tilde{\mu}_{\gamma_\ell} > 0$ for $\ell = 1, \ldots, L$, which is a one-time cost, and normalize $\mu_{\gamma_\ell} = \tilde{\mu}_{\gamma_\ell} / \sum_{k=1}^{L} \tilde{\mu}_{\gamma_k}$. In the simulations and applications of Sections 4-6, we fix $\lambda_0 = 1/2$ and model $\sigma_\gamma^{-2} \sim \text{Gamma}(0.001, 0.001)$.

We use quadratic I-splines with $L = 2 + \min\{(\# \text{ unique } y_i)/4, 10\}$ knots, implemented using the `splines2` package in R (Wang and Yan, 2018). Boundary knots are placed at zero and $\max\{y_i\}$, while the $L - 2$ interior knots selected using the sample quantiles of $\{y_i\}$ excluding zero, one, and $\max\{y_i\}$, with an

interior knot placed at one to improve distributional flexibility near zero. Since Ramsay (1988) use $L = 3$ or $L = 5$ in all monotone spline examples, a small number of knots may be adequate in many cases.

2.2. Model properties

By design, STAR builds upon models for continuous data, such as those in Section 3, and adapts them for integer-valued data. Yet STAR is not merely a mechanism for producing valid integer-valued processes: STAR also provides important distributional properties for modeling integer-valued data in practice. By careful selection of the rounding operator h and the transformation g, STAR provides the capability to model zero-inflation, bounded or censored data, and over- or underdispersion.

In applications with count data, it is common to observe an abundance of zeros, $y(x) = 0$. STAR can be parametrized such that zero counts occur whenever $z^*(x)$ is negative:

Lemma 1 (Zero-inflation). *For any* STAR *model with* $g(\mathcal{A}_0) = (-\infty, 0)$, *we have (i)* $y(x) = 0$ *if and only if* $z^*(x) < 0$ *and (ii)* $\mathbb{P}\{y(x) = 0\} = \mathbb{P}\{z^*(x) \leq 0\}$.

Lemma 1 is valid for known or unknown transformations, and is easily satisfied for (6) letting $\mathcal{A}_0 = (a_0, a_1) = (-\infty, 1)$ for $\lambda \neq 0$ and $\mathcal{A}_0 = (a_0, a_1) = (0, 1)$ for $\lambda = 0$. For model (3), Π_θ is conditionally Gaussian, which may place substantial prior mass on $z^*(x) < 0$ and thus $y(x) = 0$. Therefore, STAR has a built-in and interpretable mechanism for handling zero counts, and does not require the addition of an artificial constant to the transformation, such as $\log(y+1)$. Furthermore, dependence among zero values is implicit in the model: $\mathbb{P}\{y(x) = 0, y(x') = 0\} = \mathbb{P}\{z^*(x) < 0, z^*(x') < 0\}$ depends on the joint distribution of $(z^*(x), z^*(x'))$, which is modeled by Π_θ.

Another common characteristic of count-valued data is a deterministic upper bound K. For instance, if y counts the number of days on which an event occurred in a given year, then $y(x) \in \{0, 1, \ldots, K\}$ and $K = 365$. STAR can easily incorporate this information into the distribution for y as formalized in the next lemma.

Lemma 2 (Bounded observations). *For any* STAR *model, letting* $g(\mathcal{A}_K) = [g(a_K), \infty)$ *implies* $\mathbb{P}\{y(x) \leq K\} = 1$.

The boundedness constraints in Lemma 2 are compatible with any choice of (unconstrained) continuous-valued model (2) and do not require modification of the algorithms for estimation and inference in Section 2.3. Similar to the case of zero values in Lemma 1, STAR allows for dependence among y values that attain the upper bound: $\mathbb{P}\{y(x) = K, y(x') = K\} = \mathbb{P}\{z^*(x) \geq g(a_K), z^*(x') \geq g(a_K)\}$, which again is modeled by (2) or (3). When $K = 1$, the STAR model (3) with $\mu(x) = x'\beta$ and $\sigma^2(x) = 1$ simplifies to probit regression.

Interestingly, the construction in Lemma 2 is coherent under right-censoring, which occurs when an observed count value of K implies that $y(x) \geq K$. Right-censoring is common in surveys, where large values are often grouped together.

The following lemma formalizes the properties of STAR subject to right-censoring of the observations.

Lemma 3 (Right-censoring). *Let* $\mathrm{STAR}(h, g, \Pi_\theta)$ *denote model* (1)-(2) *with rounding operator h, transformation g, and latent distribution Π_θ. For right-censored observations $y_c(x) = \min\{y(x), K\}$ with $y \sim \mathrm{STAR}(h, g, \Pi_\theta)$ and $y_c \sim \mathrm{STAR}(h', g, \Pi_\theta)$ such that h and h' satisfy $a_K = a'_K$ and $\mathcal{A}'_K = [a_K, \infty)$, we have $\mathbb{P}\{y(x) \geq K\} = \mathbb{P}\{z^*(x) \geq g(a_K)\} = \mathbb{P}\{y_c(x) = K\}$.*

For right-censored data, the likelihood includes terms of the form $\mathbb{P}\{y(x) \geq K\}$ for censored observations. Lemma 3 shows that the censored likelihood terms under a STAR model for y are equivalent to the non-censored likelihood terms $\mathbb{P}\{y_c(x) = K\}$ under a STAR model for $y_c = \min\{y(x), K\}$ with $\mathcal{A}'_K = [a_K, \infty)$. Remarkably, STAR preserves the correct right-censored likelihood for y by directly modeling the observed counts y_c and setting $\mathcal{A}'_K = [a_K, \infty)$, with no further modifications needed for the model specification or estimation procedure. By comparison, common parametric approaches for modeling count data, such as the Poisson model and its generalizations, require careful modifications of the likelihood and tailored algorithms for estimation and inference in the case of right-censoring. Naturally, a similar approach is available for left-censoring.

Lastly, we show that STAR processes are capable of modeling over- or under-dispersion. In Figure 1, we illustrate the relationships among the expectation $\mathbb{E}[y]$, the variance $\mathrm{Var}(y)$, and the probability of zeros $\mathbb{P}(y = 0)$ for a STAR process defined by $z^* \sim N(\mu, \sigma^2)$ and transformation (6) with $\lambda = 1/2$. For different values of the parameters the STAR process exhibits different features, including overdispersion, underdispersion, and zero-inflation.

2.3. Posterior inference

We develop a general Markov Chain Monte Carlo (MCMC) algorithm for Bayesian inference under STAR. The hierarchical construction of STAR in (1)-(2) is accompanied by a computationally convenient data augmentation strategy, which we leverage to incorporate existing sampling techniques for the unknown parameters θ in (2). To emphasize the modularity of the proposed approach, we omit model-specific details for sampling θ until Section 3.

Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ denote the observed pairs of points $x_i \in \mathcal{X}$ and integer-valued data $y_i = y(x_i)$. Consider a Bayesian specification of (2) with suitable prior on θ and an algorithm $\mathbb{A}$ which draws from the posterior distribution of θ given the (continuous) data. The sampling algorithm $\mathbb{A}$ is designed for continuous data, such as (Gaussian) additive models or BART, depending on the different choices for Π_θ (see Section 3 for details). The posterior sampling algorithm for STAR defines a Gibbs sampler by combining a data augmentation step with algorithm $\mathbb{A}$ as follows:

1. Sample $[z^*(x_i) \mid \mathcal{D}, \theta]$ from Π_θ truncated to $z^*(x_i) \in g(\mathcal{A}_{y_i})$ for $i = 1, \ldots n$;
2. Sample $[\theta \mid \boldsymbol{z}^*]$ using algorithm $\mathbb{A}$ conditioning on $\boldsymbol{z}^* = (z^*(x_1), \ldots, z^*(x_n))'$.
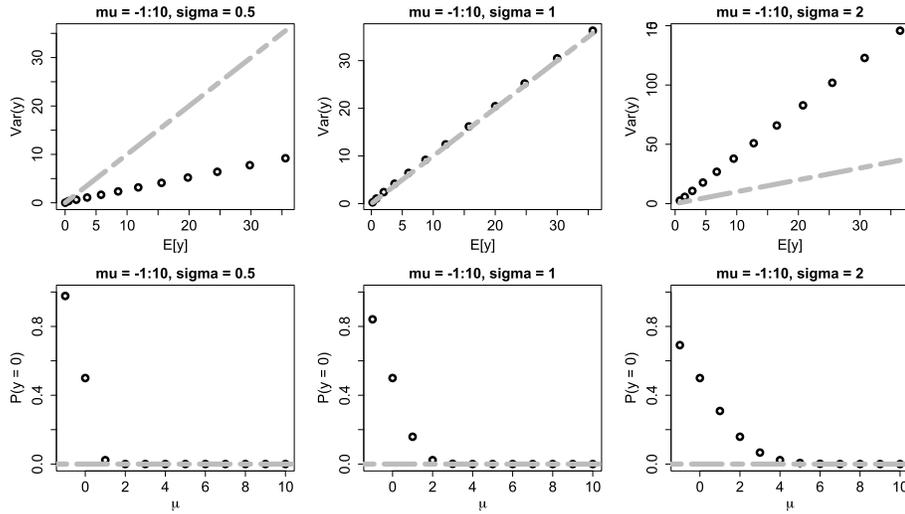
FIG 1. $\mathbb{E}[y]$ and Var(y) (top) and $\mathbb{P}(y = 0)$ (bottom) for a STAR process defined by $z^* \sim N(\mu, \sigma^2)$ and transformation (6) with $\lambda = 1/2$ for various (μ, σ) pairings. The dashed gray lines correspond to $\mathbb{E}[y] = Var(y)$ (top) and $\mathbb{P}(y = 0) = 0$ (bottom). STAR processes may include underdispersion (top left), overdispersion (top right), and zero-inflation (bottom).

In the case of model (3), the data augmentation step may be computed efficiently using a standard univariate truncated normal sampler. Specifically, for $\mathcal{A}_j = [a_j, a_{j+1})$, the full conditional distribution of the latent data is $[z^*(x_i) \mid \mathcal{D}, \theta] \sim N(\mu(x_i), \sigma^2(x_i))$ truncated to $[g(a_{y_i}), g(a_{y_i+1}))$. While the process y^* is useful for interpretability of the STAR model, it is not necessary for inference or sampling.

When the transformation g is unknown, an additional sampling step is required. For the parametric Box-Cox case (6) this translates to sampling the parameter λ from its full conditional posterior distribution, for which we use a slice sampler (Neal, 2003). For the nonparametric model (7) with prior (8), we sample $\xi_\gamma = \log(\tilde{\gamma})$ using Metropolis-Hastings and $\left[\sigma_\gamma^{-2} \mid -\right] \sim \text{Gamma}\{0.001 + L/2, 0.001 + \sum_{\ell=1}^{L}(\tilde{\gamma}_\ell - \mu_{\gamma\ell})^2\}$, and set $g(t) = b'_I(t)\gamma$ as defined in (7)-(8). The sampler for ξ_γ uses a Gaussian random walk proposal with covariance matrix tuned using the robust adaptive Metropolis (RAM) algorithm of Vihola (2012) during a preliminary burn-in period. Within the RAM algorithm we set a target acceptance rate of 30% with an adaptation rate of 0.75; see Vihola (2012) for details. We adapt the proposal covariance only during the first 50% of the burn-in period, so the MCMC draws we save for inference are generated from a (non-adaptive) Metropolis-within-Gibbs sampling algorithm.

To simulate from the posterior predictive distribution $[\tilde{y}(x)|\mathcal{D}]$, we additionally sample $[\tilde{z}^*(x) \mid \theta]$ from Π_θ using the current draw of θ and set $\tilde{y}(x) = h\left[g^{-1}\{\tilde{z}^*(x)\}\right]$ for each x. This step is extremely simple, yet provides inference for integer-valued predictions, model-based imputation of missing data at $x \in \mathcal{X}$, and useful model diagnostics. For parametric g in (6), the functions $g^{-1}(s; \lambda)$ are

known, while for nonparametric g we approximate $g^{-1}(s) \approx \arg\min_t |s - b'_I(t)\gamma|$, where the minimum t is computed over a grid of values.

The proposed framework for MCMC balances modularity and flexibility: it combines existing algorithms for continuous data models with a transformation to provide distributional flexibility for integer-valued data. The importance of modularity has been demonstrated recently for the negative-binomial distribution, for which Polson et al. (2013) developed a Pólya-Gamma data augmentation scheme for Gibbs sampling. This approach has allowed a variety of Gaussian models to be extended for negative-binomial data, including linear regression (Zhou et al., 2012), factor models (Klami, 2015), and functional time series models (Kowal, 2019), yet faces two important limitations: first, it is restricted to the negative-binomial distribution, and second, the resulting MCMC sampler is often inefficient (Duan et al., 2018). As demonstrated in Section 6 and Appendix A.2, the proposed STAR algorithm provides excellent MCMC efficiency, even for nonlinear versions of model (3).

3. Regression modeling with STAR

For inference and prediction of integer-valued data y observed with predictors x, we apply the STAR modeling framework to develop additive (Section 3.1) and BART (Section 3.2) regression models. Each model may be combined with a known or unknown transformation, and posterior inference proceeds using the general approach from Section 2.3. The additive and BART STAR models are evaluated for synthetic data in Section 4 and real data in Sections 5 and 6, with additional model comparisons and diagnostics in the Appendix.

3.1. Additive models

Suppose the p predictors are partitioned as $x' = (u', v')$ for linear predictors u and nonlinear predictors v. The STAR additive model is given by (3) with conditional mean

$$\mu(x) = u'\beta + \sum_j f_j(v_j), \tag{9}$$

where $f_j : \mathcal{X}_j \to \mathbb{R}$ is an unknown function of $v_j \in \mathcal{X}_j$. The unknown f_j are typically modeled as smooth nonparametric functions, and may capture nonlinearities in each v_j. The STAR linear model is a special case of (9) with $\mu(x) = x'\beta$. For the conditional variance of z^* in (3) under homoscedasticity $\sigma^2(x) = \sigma^2$, we use the conditionally conjugate prior $\sigma^{-2} \sim \text{Gamma}(0.001, 0.001)$.

Within the STAR framework, we apply flexible and computationally efficient parametrizations of (9) that have been well-developed for Gaussian and exponential family models. The linear regression coefficients are assigned conditionally Gaussian priors, $\beta \sim N(0, \Sigma_\beta)$, including the ridge prior $\Sigma_\beta = \sigma^2_\beta I$ and other shrinkage priors (Carvalho et al., 2010) as special cases. The nonlinear functions in (9) are modeled smoothly using a basis expansion $f_j(v_j) = b'_j(v_j)\alpha_j$, where b_j is an L_j-dimensional vector of known basis functions and α_j is a vec-

tor of unknown coefficients. We select a cubic P-spline basis with second-order difference penalty on the coefficients, which may be reparametrized such that $\alpha_j \sim N(0, \sigma_{\alpha_j}^2 I)$ is the smoothing prior and $B_j' B_j$ is diagonal, where $B_j = (b_j(v_{1,j}), \ldots, b_j(v_{n,j}))'$ is the basis matrix (Scheipl et al., 2012). The nonlinear terms are constrained such that $\sum_{i=1}^n f_j(v_{i,j}) = 0$ for identifiability, which is enforced in the reparametrization of the basis b_j. We let $\sigma_{\alpha_j}^{-2} \sim \text{Gamma}(0.1, 0.1)$, which allows the smoothness of each f_j to be learned from the data.

For observed predictors $x_i' = (u_i', v_i')$, let U denote the matrix of linear predictors and let $\boldsymbol{f}_j = B_j \alpha_j$. The Gibbs sampler for the STAR additive model iterates the following full conditional distributions:

1. Sample $[z^*(x_i) \mid -] \sim N(u_i'\beta + \sum_j f_j(v_{i,j}), \sigma^2)$ truncated to $z^*(x_i) \in g(\mathcal{A}_{y_i})$;
2. Sample $[\beta \mid -] \sim N\big(Q_\beta^{-1}\ell_\beta, Q_\beta^{-1}\big)$ where $Q_\beta = \sigma^{-2}U'U + \Sigma_\beta^{-1}$ and $\ell_\beta = \sigma^{-2}U'(\boldsymbol{z}^* - \sum_j \boldsymbol{f}_j)$;
3. For each j, sample $[\alpha_j \mid -] \sim N\big(Q_{\alpha_j}^{-1}\ell_{\alpha_j}, Q_{\alpha_j}^{-1}\big)$ where $Q_{\alpha_j} = \sigma^{-2}B_j'B_j + \sigma_{\alpha_j}^{-2}I$ and $\ell_{\alpha_j} = \sigma^{-2}B_j'(\boldsymbol{z}^* - U\beta - \sum_{k \neq j} \boldsymbol{f}_k)$ and set $\boldsymbol{f}_j = B_j\alpha_j$;
4. Sample $[\sigma^{-2} \mid -] \sim \text{Gamma}\big(0.001 + n/2, 0.001 + ||\boldsymbol{z}^* - U\beta - \sum_j \boldsymbol{f}_j||^2/2\big)$;
5. For each j, sample $[\sigma_{\alpha_j}^{-2} \mid -] \sim \text{Gamma}\big(0.1 + L_j/2, 0.1 + \sum_{\ell=1}^{L_j} \alpha_j^2/2\big)$.

The MCMC sampling algorithm for the STAR additive model is efficient, with empirical support provided in Section 6. The computational complexity for the nonlinear basis coefficients $\{\alpha_j\}$ is $\mathcal{O}(\sum_j L_j)$ due to the diagonality of $B_j'B_j$, while the linear coefficients β also may be sampled efficiently (Rue, 2001; Bhattacharya et al., 2016). Additional sampling steps for Σ_β depend on the model specification, but are often available in closed form.

3.2. *Bayesian additive regression trees*

While additive models are effective at capturing nonlinear marginal effects, they are often inadequate for modeling interactions among predictors. Specific pairwise or higher order interactions may be specified in advance, but including all possible interactions in an additive model requires a massive number of parameters. As a remedy, Chipman et al. (2010) proposed BART, which is a "sum-of-trees" model within a fully Bayesian framework. Tree-based regression models, such as Chipman et al. (1998), are designed to model complex interactions among predictors. Notably, BART utilizes many trees, where each individual tree is constrained via the prior to be a weak learner. As a result, BART provides the capability to capture nonlinear interactions yet features built-in mechanisms to guard against overfitting. For continuous and binary data, the predictive performance of BART is highly competitive with state-of-the-art statistical and machine learning models.

For integer-valued data, BART has been relatively underutilized. Adaptations of BART for negative-binomial data are feasible via Pólya-Gamma augmentation (Polson et al., 2013), similar to the probit implementation in Chipman et al. (2010) for binary data. However, this approach is limited in distribu-

tional flexibility, and the MCMC inefficiencies of Pólya-Gamma augmentation are unlikely to be ameliorated given the complexity of the (Gaussian) BART sampling algorithm. Recently, Murray (2017) proposed a log-linear BART model for count-valued and categorical data using a parameter expansion valid for certain likelihoods in log-linear models, in particular (zero-inflated) negative-binomial and Poisson. However, extensions to more flexible count distributions may require alternative computational strategies and appropriately modified prior distributions.

Within the STAR framework, we parametrize the BART model (BART-STAR) as in Chipman et al. (2010) and specifically

$$\mu(x) = \sum_{k=1}^{m} f(x; T_k, M_k), \tag{10}$$

where T_k is a binary tree comprised of interior splitting rules and terminal nodes and $M_k = \{\eta_{1,k}, \ldots, \eta_{b_k,k}\}$ is the value at each of b_k terminal nodes for tree T_k. For a given predictor x, each tree T_k in (10) assigns a value $\eta_{\ell,k} \in M_k$, and these values are summed across all trees $k = 1, \ldots, m$. Chipman et al. (2010) propose prior distributions that constrain each T_k to be shallow, thereby limiting the order of interactions, and constrain each $\eta_{\ell,k}$ to be small, thereby limiting the contribution of each tree. Both mechanisms guard against overfitting, and in combination produce a sum of weak learners. The joint prior distribution is specified as a prior for the tree, $p(T_k)$, which follows Chipman et al. (1998), and a prior for the terminal values given the tree, $p(\eta_{\ell,k}|T_k)$, which is Gaussian. A key feature of STAR is that, by transforming to Gaussianity, we inherit the same framework as the original BART, and therefore may directly incorporate the well-studied priors and hyperparameters from Chipman et al. (2010).

More careful consideration is required for the prior distribution of σ^2 in (3). Chipman et al. (2010) emphasize that an informative prior distribution is important to balance between overly aggressive and overly conservative model fits, and parametrize the prior for σ^2 as an inverse chi-square distribution calibrated using a data-based overestimate $\hat{\sigma}$ of σ. However, any statistics calculated from the data y are likely inappropriate for STAR, since the transformation g impacts the scale of z^*. As a remedy, we compute $\hat{\sigma}$ as the posterior median of σ from the STAR linear model (9), where the transformation in the linear model is chosen to match the transformation in BART-STAR. Given $\hat{\sigma}$, which indeed is a data-based overestimate of σ, we adopt the default hyperparameter suggestions of Chipman et al. (2010).

For posterior inference under BART-STAR, we combine a data augmentation step for $z^*(x_i)$ similar to Section 3.1 with a sweep from the original Chipman et al. (2010) algorithm to draw the BART parameters in (10) using z^* as data. The Chipman et al. (2010) BART sampler proceeds using backfitting: draws for the kth tree $[(T_k, M_k)|z^*, \{(T_{k'}, M_{k'})\}_{k \neq k'}, \sigma]$ are generated using Chipman et al. (1998) and σ^2 is sampled from an inverse-Gamma distribution. Incorporating these sampling steps into the larger Gibbs sampler in Section 2.3 is straightforward using the dbarts package (Dorie et al., 2018).

4. Simulation studies

The proposed STAR modeling framework is evaluated using simulated data and compared to existing methods for Poisson, negative-binomial, and Gaussian data. Synthetic data y_i for $i = 1, \ldots n$ and $n = 100$ are simulated from a negative-binomial distribution parametrized by conditional mean $\mathbb{E}\{y_i(x)|\lambda_i^*\} = \lambda_i^*(x)$ and variance $\mathrm{Var}\{y_i(x)|\lambda_i^*, r^*\} = \lambda_i^*(x)\{1 + \lambda_i^*(x)/r^*\}$ with dispersion parameter $r^* > 0$. As r^* decreases to zero, the variance increasingly dominates the mean while as $r^* \to \infty$, the distribution converges to a Poisson distribution with parameter $\lambda_i^*(x)$. We select $r^* = 1$ to simulate negative-binomial data with large overdispersion and $r^* = 1000$ to simulate approximate Poisson data. We consider linear (Section 4.1) and nonlinear (Section 4.2) parametrizations for the log-mean $\log \lambda_i^*(x)$. We emphasize that in all cases, the simulated datasets are *not* generated under the proposed STAR model: they are simulated from negative-binomial and (approximate) Poisson distributions.

Competing models are compared using the Watanabe-Akaike/widely-applicable information criteria (WAIC) (Watanabe, 2010). WAIC estimates out-of-sample predictive accuracy using a single model fit requiring only minimal additional computations, and is asymptotically equivalent to cross-validation. The WAIC for a model $\mathcal{M}$ is defined as $\mathrm{WAIC}_{\mathcal{M}} = -2\,(\mathrm{lpd}_{\mathcal{M}} - d_{\mathcal{M}})$, where $d_{\mathcal{M}}$ is the effective number of parameters for model $\mathcal{M}$ and $\mathrm{lpd}_{\mathcal{M}}$ is the log-predictive pointwise density defined by

$$\mathrm{lpd}_{\mathcal{M}}(y) = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p_{\mathcal{M}}(y_i \mid \theta^s)\right)$$

for θ^s drawn from its posterior distribution. For STAR models, we simply have

$$\mathrm{lpd}_{\mathrm{STAR}}(y) = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} \Phi\left\{\frac{g^s(a_{y_i+1}) - \mu^s(x_i)}{\sigma^s(x_i)}\right\} - \Phi\left\{\frac{g^s(a_{y_i}) - \mu^s(x_i)}{\sigma^s(x_i)}\right\}\right). \tag{11}$$

For the effective number of parameters, we follow the recommendation of Gelman et al. (2014) and use the sample variance of the pointwise log-likelihoods across MCMC simulations: $d_{\mathcal{M}} = \sum_{i=1}^{n} \mathrm{Var}\,(\log p_{\mathcal{M}}(y_i \mid \theta^s))$. The pointwise log-likelihood of STAR is simple and efficient to compute, and is sufficient for computing WAIC as well as other information criteria.

WAIC is used in Sections 4.1 and 4.2 for synthetic data, but exact out-of-sample metrics are provided for the real data in Section 5. Appendix A.1 also provides root mean square errors for estimating the conditional expectation $\lambda_i^*(x)$ to evaluate point estimation accuracy.

4.1. Linear mean functions

We first consider a linear log-mean, $\log \lambda_i^*(x) = \beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j$, where the $p = 6$ predictors are drawn independently from $x_{i,j} \sim N(0, 1)$ and the coefficients

are $\beta_0 = \log(1.5)$, $\beta_1 = \beta_2 = \beta_3 = \log(2.0)$, and $\beta_4 = \beta_5 = \beta_6 = 0$. Under this specification, the expected count at $x_{i,j} = 0$ is 1.5, while each nonzero coefficient β_j for $j = 1, 2, 3$ increases the expected count by a factor of 2 per one unit change in each x_j.

For comparison, we consider a variety of Bayesian linear regression models. Among STAR models, we use the linear model (9) with prior $[\beta_j|\sigma_\beta] \sim N(0, \sigma_\beta^2)$ and $[\sigma_\beta] \sim \text{Uniform}(0, 10^4)$, and the following transformations: known transformation (6) with $\lambda = 0$ (LM-STAR-log), $\lambda = 1/2$ (LM-STAR-sqrt), and $\lambda = 1$ (LM-STAR-id); unknown parametric transformation (6) (LM-STAR-bc); and unknown nonparametric transformation (7) (LM-STAR-np). We also include the same linear model, but instead with the Gaussian model (3) applied directly to the counts y (LM) and the log-transformed counts $\log(y + 1)$ (LM-log). These models are natural competitors to STAR, since they incorporate the same model for $\mu(x)$ but omit the rounding step in (1) and therefore do not produce an integer-valued distribution. Lastly, we include Poisson (LM-Pois) and negative-binomial (LM-NegBin) linear regression models with a log-link, implemented using the `rstanarm` package (Goodrich et al., 2018). LM-Pois and LM-NegBin are widely used for modeling count data, and here correspond to the true data-generating process.

In Figure 2, we plot the relative WAICs across simulated datasets, defined as the ratio between the WAIC of the generic model over the WAIC for a baseline method, for which we select LM-log. Relative WAIC standardizes model performance across simulated datasets: methods with a relative WAIC less than one demonstrate improvement relative to the baseline method. The STAR models, particularly those with unknown transformations (LM-STAR-bc and LM-STAR-np), offer substantial improvements relative to LM-log, and are highly competitive with the true models LM-Pois and LM-NegBin. The STAR model improvements relative to the Gaussian models and the identity transformation LM-STAR-id definitively demonstrate the importance of both rounding and transformation.
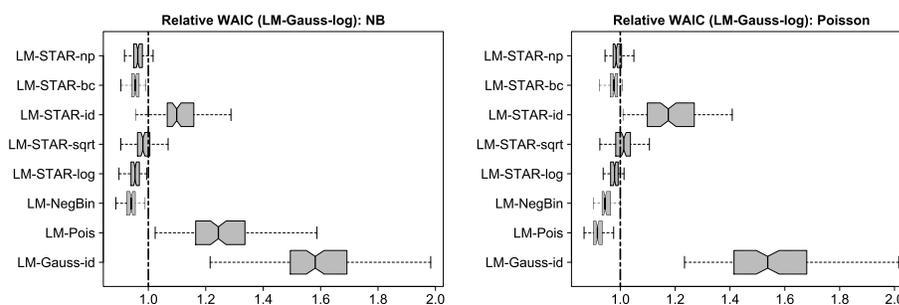

FIG 2. *Relative WAIC for negative-binomial (left) and Poisson (right) data with linear mean functions. Preferred models have smaller values, and models with values less than one are preferred to* LM-*log. The* STAR *models outperform the Gaussian models and are highly competitive with the true models* LM-*Pois and* LM-*NegBin.*

4.2. Nonlinear mean functions

To evaluate BART-STAR, we specify a nonlinear form for the log-mean, $\log \lambda_i^*(x) = \beta_0 + \beta_1 \tilde{f}(x)$, where $\tilde{f}(x)$ is the centered and scaled Friedman function (Friedman, 1991)

$$f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 \tag{12}$$

featured in the original BART simulations (Chipman et al., 2010). As in Chipman et al. (2010), we select $p = 10$ and simulate $x_{i,j} \overset{iid}{\sim} \text{Uniform}(0,1)$. We fix the parameters $\beta_0 = \log(1.5)$ and $\beta_1 = \log(5.0)$, which again corresponds to low counts with a moderate signal.

We combine the BART-STAR model of Section 3.2 with known transformation (6) for $\lambda = 0$ (BART-STAR-log), $\lambda = 1/2$ (BART-STAR-sqrt), and $\lambda = 1$ (BART-STAR-id); unknown parametric transformation (6) (BART-STAR-bc); and unknown nonparametric transformation (7) (BART-STAR-np). For competitors, we include the Gaussian BART model (BART-id) of Chipman et al. (2010) and a Gaussian BART model on the log-transformed counts $\log(y + 1)$ (BART-log). Lastly, we include the linear models LM-STAR-bc and LM-log from Section 4.1.

The relative WAICs are plotted in Figure 3, where again we use the log-transformed Gaussian model (BART-log) as the baseline. Both BART-STAR-id and BART-id are omitted as noncompetitive, and indicates the importance of an appropriate transformation. BART-STAR provides substantial improvements relative to BART and linear Gaussian models, with especially strong performance from the unknown transformation models (BART-STAR-bc and BART-STAR-np). Perhaps surprisingly, the STAR linear model LM-STAR-bc outperforms BART-log for negative-binomial data, despite the nonlinearity in (12). By comparison, BART-STAR-bc consistently outperforms LM-STAR-bc, which suggests that the proposed BART-STAR model is capable of detecting the nonlinear features in (12).
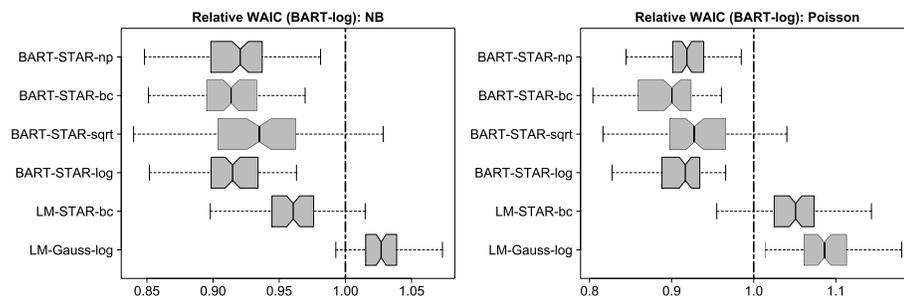

FIG 3. *Relative WAIC for negative-binomial (left) and Poisson (right) data with nonlinear mean functions. Preferred models have smaller values, and models with values less than one are preferred to* BART-*log. The identity models (*BART-*id,* BART-STAR-*id) are omitted since they are noncompetitive, with relative WAICs above 1.6. The* BART-STAR *models are clearly superior.*

5. Predicting the demand for healthcare utilization

Individualized prediction of healthcare utilization is critical both for assessing the health risks of an individual and for monitoring the aggregate stress on the healthcare system. By providing more accurate *predictive distributions* of individual healthcare utilization, it is possible to obtain uncertainty quantification for various measures of individual and aggregate demand, and consequently achieve more efficient allocation of medical resources and more informed patients. To assess the predictive ability of STAR models for this task, we use data from the National Medical Expenditure Survey (NMES) available in the AER package in R (Kleiber and Zeileis, 2008). Multiple measures of healthcare utilization are available, including physician office visits (visits), non-physician office visits (nvisits), physician hospital outpatient visits (ovisits), and non-physician hospital outpatient visits (novisits). Individualized predictors are also provided, including health measures, socioeconomic and demographic variables, and indicators of each patient's type of insurance. We consider a subset of $n = 4406$ elderly adults (aged 66 and older) covered by Medicare, which was previously analyzed by Deb and Trivedi (1997) and Cameron and Trivedi (2013).

The NMES data provides a unique opportunity for insightful out-of-sample prediction comparisons. Each measure of healthcare utilization (visits, nvisits, ovisits, and novisits) is count-valued with distinct characteristics: the probability mass functions in Figure 4 illustrate the differences in the marginal distributions, most notably the proportion of zeros and the degree of overdispersion. An adequate prediction of individual healthcare utilization may require prediction of one or more of these response variables, each of which presents unique count-valued distributional features, and which share a common set of $p = 17$ individual predictor variables.

We consider out-of-sample predictive distribution accuracy for each response variable in Figure 4. In all cases, we select $n_{train} = 3525$ (80%) individuals randomly for training and evaluate the predictive accuracy for the remaining $n_{test} = 881$ test individuals, and repeat this exercise for 100 iterations. Posterior predictive distributions were computed for each model in Section 4.2; for conciseness, we report results for BART-log, BART-STAR-id, BART-STAR-log, and BART-STAR-bc. The (untransformed) BART model of Chipman et al. (2010) was noncompetitive and is omitted from the subsequent results.

To evaluate the aggregate predictive distribution accuracy, we compute the log-predictive density score for the test data $\{\tilde{y}_i\}_{i=1}^{n_{test}}$ for each model $\mathcal{M}$:

$$\text{lpd}_{\mathcal{M}}(\tilde{y}) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \log p_{\mathcal{M}}(\tilde{y}_i|y) \approx \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \log\left\{\frac{1}{S} \sum_{s=1}^{S} p_{\mathcal{M}}(\tilde{y}_i \mid \theta^s)\right\} \quad (13)$$

where $\theta^s \sim p_{\mathcal{M}}(\theta|y)$ is a draw from the posterior under model $\mathcal{M}$. For STAR with model (3), we have $p_{\mathcal{M}}(\tilde{y}_i \mid \theta^s) = \Phi\big[\{g^s(a_{\tilde{y}_i+1}) - \mu^s(x_i)\}/\sigma^s(x_i)\big] - \Phi\big[\{g^s(a_{\tilde{y}_i}) - \mu^s(x_i)\}/\sigma^s(x_i)\big]$ similar to (11). The results for each response variable are in Figure 5; larger values indicate more accurate predictive distributions. The STAR
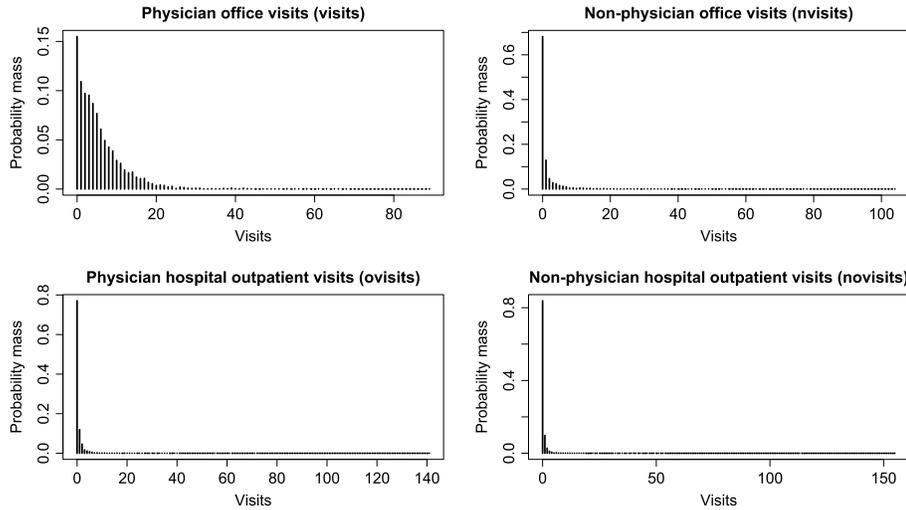
FIG 4. *Probability mass functions for each measure of medical care demand. Zero-inflation and overdispersion are present in each case to varying degrees.*

model with unknown transformation (BART-STAR-bc) performs best in all cases, closely followed by the STAR model with known transformation (BART-STAR-log). The transformation is impactful: the untransformed STAR model (BART-STAR-id), similar to Canale and Dunson (2013), is substantially less accurate than both transformed STAR models. Furthermore, the untransformed BART-STAR-id model produced infinite scores for some test points $\tilde{y}_i$; these points were excluded for computing (13), but imply that BART-STAR-id performs even worse than indicated.

For a more targeted assessment of the predictive distributions, we compare the precision and coverage of the 90% prediction intervals for each method. Interval precision is measured by the mean prediction interval width (MPIW) averaged across all $\{\tilde{y}_i\}_{i=1}^{n_{test}}$: smaller intervals that provide at least 90% coverage are preferable. The MPIWs with empirical coverages are displayed in Figure 6. Note that the boxplots correspond to the MPIWs for each method across the out-of-sample testing sets, and do not indicate whether the prediction intervals overlap among the different methods. Across all responses, the STAR model with unknown transformation (BART-STAR-bc) provides the most precise prediction intervals with at least 90% coverage. For the responses with greater zero-inflation and overdispersion (`nvisits`, `ovisits`, and `novisits`), BART-STAR-bc reduces MPIWs by a median of 45%, 50%, and 59%, respectively, relative to BART-STAR-id, which indicates a substantial gain in predictive precision offered by the (unknown) transformation.

Lastly, we consider a specific prediction task of interest: estimating the probability that an individual will utilize the healthcare system, $\mathbb{P}(\tilde{y}_i > 0|y)$. We evaluate each method using logarithmic scoring on the event $\{\tilde{y}_i > 0\}$, which is a proper scoring rule for binary outcomes (Gneiting and Raftery, 2007). The
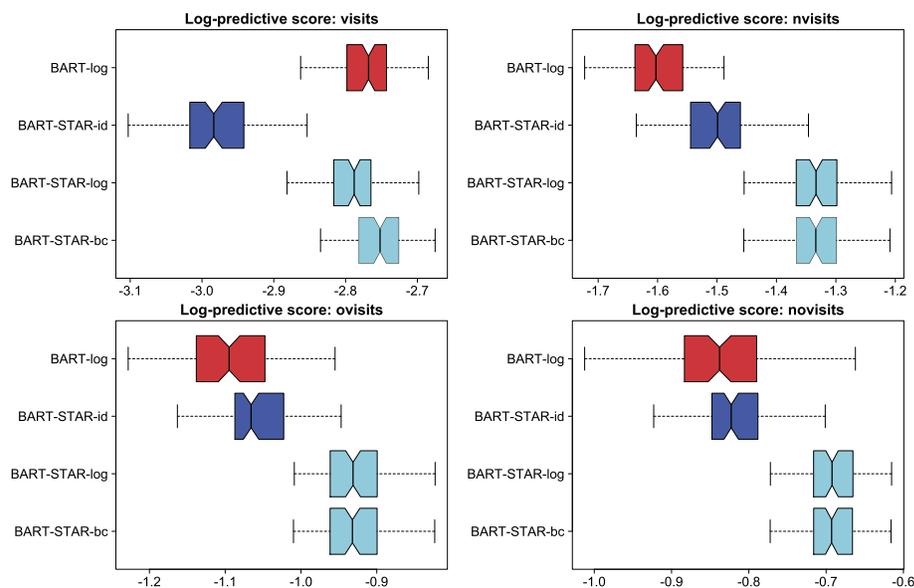
FIG 5. *Log-predictive density (LPD) score for the test data across 100 randomly selected test sets (20% of data). Large LPD score indicates better performance. The* STAR *models that include transformations—known (*STAR-*log) or unknown (*STAR-*bc)—are decisively favored.*

results are in Figure 7, where larger values indicate superior predictive accuracy. The BART-STAR models which include both transformation and rounding (BART-STAR-log and BART-STAR-bc) decisively outperform the competitors, especially the continuous-valued model (BART-log).

Based on multiple measures of predictive accuracy for each of four count-valued measures of healthcare utilization, the results from our out-of-sample comparison provide definitive confirmation of the predictive capabilities of STAR models, and in particular BART-STAR models which include *both* transformation and rounding.

6. Modeling the decline in Amazon river dolphins

The tucuxi dolphin (*sotalia fluviatilis*) is a small river dolphin that inhabits the Amazon River. While the tucuxi dolphin population was once stable, the progression of habitat degradation, dolphin fishing, and other human interference has to led to increased concerns of population decline. To assess the validity of these concerns, da Silva et al. (2018) gathered data from 1994 to 2017 using multiple observers to search for tucuxi dolphins along a particular segment of the Amazon River. In addition to the number of tucuxi dolphins observed, the data include the water level (in meters), the number of observers present, and the date for each of $n = 312$ surveys. While da Silva et al. (2018) fit a linear model to the logarithm of dolphin counts, we propose to leverage the STAR modeling
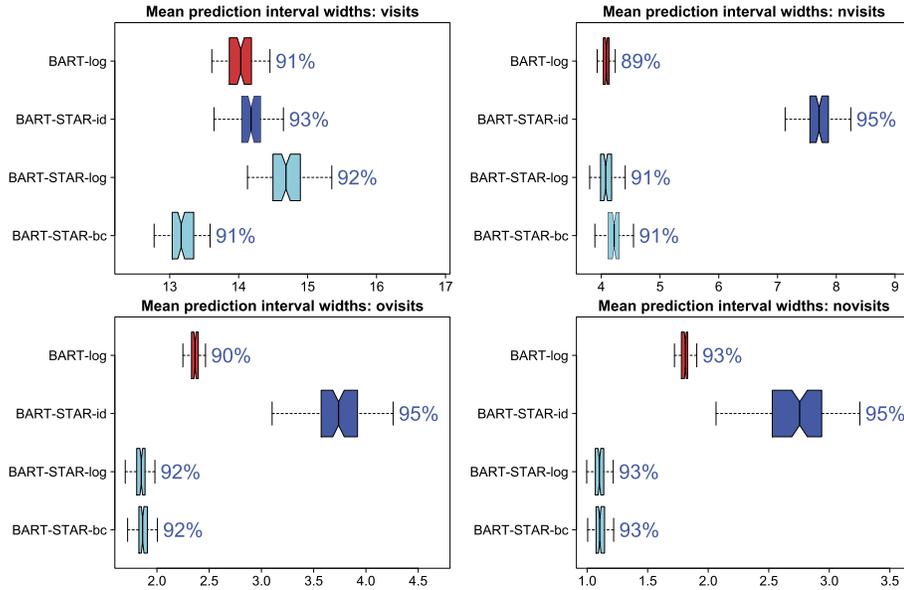
FIG 6. *Mean prediction interval widths and empirical coverage for the test data across 100 test sets. Preferred methods provide the narrowest intervals while maintaining the nominal (90%) coverage. The* BART-STAR *model with unknown transformation (* BART-STAR-bc*) is consistently competitive, followed by the* BART-STAR *model with known log transformation (* BART-STAR-log*). The methods that do not including rounding (* BART-log*) or include rounding* without *transformation (* BART-STAR-id; Canale and Dunson, 2013*) are decisively inferior.*

framework to investigate nonlinear effects and provide greater integer-valued distributional flexibility.

We use additive models to study the yearly evolution of tucuxi dolphin counts, which may be nonlinear, while adjusting for seasonal, water level, and observer effects. Specifically, for each survey we include the year (`year`), day-of-year (`doy`), and water level (`water`) as nonlinear predictors and the number of observers (`obs`) as a linear predictor. For comparisons, we implement a variety of Bayesian additive models: Gaussian additive models for the raw (AM-id) and $\log(y+1)$ transformed data (AM-log); STAR models with identity (AM-STAR-id), unknown parametric (AM-STAR-bc), and unknown nonparametric (AM-STAR-np) transformations; and Poisson (AM-Pois) and negative-binomial (AM-NB) additive models (using `rstanarm`). For each method, we jointly evaluate the model performance and the computational efficiency: performance is measured using WAIC, while efficiency is reported as seconds per 10000 effective samples. In particular, we compute multivariate effective sample sizes (Vats et al., 2019) for 10 randomly sampled points from the posterior predictive distribution using the `mcmcse` package in R (Flegal et al., 2017). The results are in Table 1. According to WAIC, the STAR models with unknown transformation (AM-STAR-bc and AM-STAR-np) are strongly preferred, while AM-NB is the closest competitor.
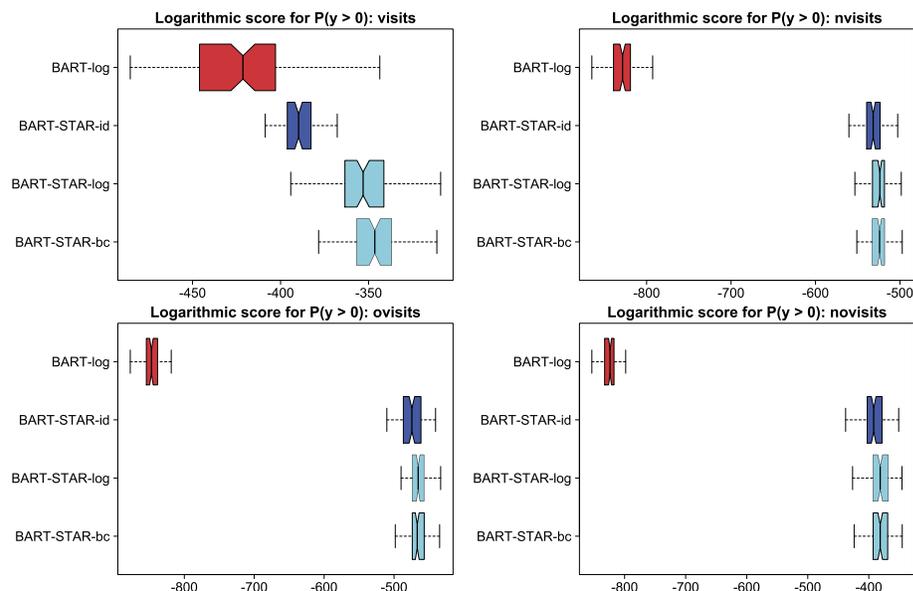
FIG 7. *Logarithmic scoring for estimation of* $\mathbb{P}(\tilde{y}_i > 0|y)$ *across 100 test sets. Large scores indicate better performance. The* STAR *models are clearly superior relative to* BART*-log, with transformed* STAR *models (*STAR*-log and* STAR*-bc) offering the best performance overall.*

TABLE 1
WAIC and seconds per 10000 effective samples for additive models (dolphins data).

	AM-id	AM-log	AM-STAR-id	AM-STAR-bc	AM-STAR-np	AM-Pois	AM-NB
WAIC	2171	1964	2000	1924	**1916**	2821	1931
Sec / 10000 ES	32	30	52	67	95	843	295

Multivariate effective sample sizes are computed for 10 randomly selected points from the posterior predictive distribution. Both transformation *and* rounding via STAR are essential to achieve competitive performance. The STAR MCMC performance is strong, especially relative to the Poisson and negative-binomial models (using `rstanarm`).

However, the existing integer-valued additive models (AM-Pois and AM-NB) are noncompetitive in computational efficiency.

Proceeding with AM-STAR-np, which is selected by WAIC, we plot posterior expectations and credible intervals for each f_j in Figure 8. MCMC diagnostics for f_j show exceptional mixing with no lack of convergence (see Appendix A.2). The `doy` plot suggests a seasonal pattern, while the `water` plot exhibits an approximately quadratic effect. Most importantly, the `year` plot shows a near linear decline in tuxucis dolphins from 1994-2017, which interestingly has leveled off since 2013. These findings are partially consistent with the results of da Silva et al. (2018) which, assuming only a linear model, also report a significant decrement of dolphins since 1994. Posterior predictive diagnostics are in Appendix A.2, and indicate clear improvements in fit for AM-STAR-np relative to models that exclude transformation (AM-STAR-id) or rounding (AM-log).
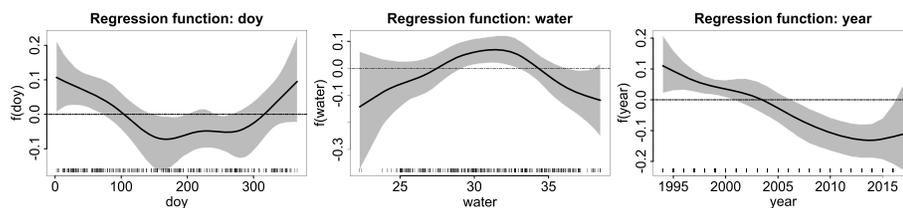
FIG 8. *Posterior expectation and 95% pointwise credible intervals for each $f_j(v_j)$ under* AM-STAR-*np for the dolphins data. The tick marks indicate the observation points for each predictor.*

7. Discussion

STAR processes provide a mathematically elegant and empirically successful framework to model count and integer-valued data. The approach seamlessly adapts state-of-the-art continuous data models and algorithms to the integer-valued data setting, and thus it offers remarkable modularity both in terms of model specification and computation, while providing ease of implementation and interpretability for practitioners. By incorporating known, unknown parametric, and unknown nonparametric transformations, STAR processes provide varying degrees of distributional flexibility, and are able to account for important distributional features such as zero-inflation, bounded or censored data, and over- or underdispersion. Empirically, STAR processes demonstrate goodness-of-fit, out-of-sample point and interval predictive accuracy, reliable inference, and computational scalability.

In addition to the healthcare utilization and animal abundance datasets considered here, we provide further empirical comparisons on three additional datasets in Appendix A.3. Among STAR, Gaussian, Poisson, and negative-binomial models, the STAR additive and BART-STAR models with unknown transformations consistently provide the best performance according to WAIC.

A variety of promising extensions exist for STAR. The modeling and computational modularity of STAR suggest that new multivariate, functional, and time series models may be developed for integer-valued data. Furthermore, STAR is capable of modeling rounded data, which is ubiquitous in practice yet rarely considered in modern statistical and machine learning methods. Lastly, the STAR model (1)-(2) does not strictly require a Bayesian modeling approach, and may be adapted for classical estimation and inference.

Appendix A: Supplemental results

A.1. Evaluating point accuracy for synthetic data

To accompany the WAIC comparisons from the main paper, we evaluate each method for point estimation accuracy. Specifically, we are interested in estimating the conditional expectation of the observed data, $\lambda_i^*(x)$. For an estimator

$\hat{y}_i(x)$, we compute the root mean squared error RMSE $= \sqrt{\sum_{i=1}^{n} \left\{ \lambda_i^*(x) - \hat{y}_i(x) \right\}^2}$.
The fitted values for STAR are computed using the conditional expectation of y
at $x \in \mathcal{X}$, that is

$$\mathbb{E}\{y(x)\} = \sum_{j=0}^{\infty} j\mathbb{P}\{y(x) = j\} \approx \sum_{j=1}^{J(x)} j\mathbb{P}\{y(x) = j\}, \tag{14}$$

where $\mathbb{P}\{y(x) = j\}$ is the STAR probability mass function and $J(x)$ is a finite
truncation. Since $\mathbb{P}\{y(x) = j\}$ depend on the parameters θ in Π_θ, the posterior
distribution of (14) may be computed by evaluating (14) for each draw of θ in
the MCMC algorithm. Conservatively, we select $J(x)$ to be the 99.99th quantile
of the distribution of $y(x)$ pointwise for each x, which is easily computable as
$h\left[g^{-1}\left\{z_q^*(x)\right\}\right]$ where $z_q^*(x)$ is the qth quantile of Π_θ. The point estimate is
computed as the posterior expectation of (14).

Figures 9 and 10 depict the relative RMSEs across simulated data sets for
the linear and nonlinear simulation designs in Sections 4.1 and 4.2, respec-
tively, defined as the ratio between the RMSE of the generic model over the
RMSE for a baseline method, and specifically the LM-log method which repre-
sents the common approach of modeling (log-) transformed counts using Gaus-
sian models. Relative RMSE standardizes model performance across simulated
datasets: methods with a relative RMSE less than 1.0 demonstrate superior
point estimation relative to the baseline method. As in Section 4, we find that
STAR-log and STAR-bc are consistently competitive and outperform other meth-
ods. Interestingly, STAR-np is much less competitive in RMSE than in WAIC,
which suggests that the additional distributional flexibility acquired by model-
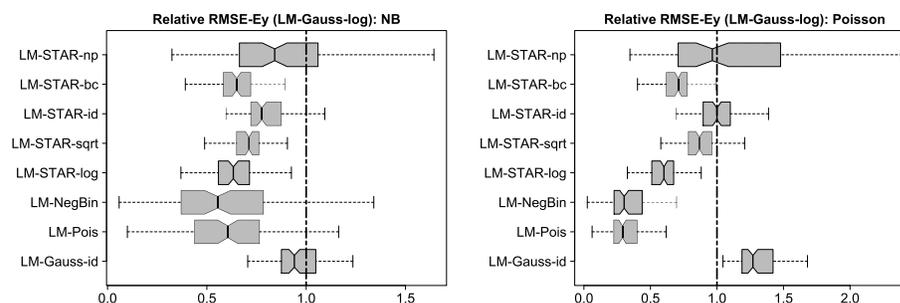ing g nonparametrically does not necessarily imply more accurate point estima-
tion.

FIG 9. *Relative RMSE under various distributions. Preferred models have smaller values,
and models with values less than 1.0 are preferred to* LM-*log. As expected, the* LM-*NegBin
performs well, since it closely matches the data-generating process. Notably, the* STAR *models
are highly competitive, and clearly superior to the Gaussian models, especially* LM-STAR-*bc
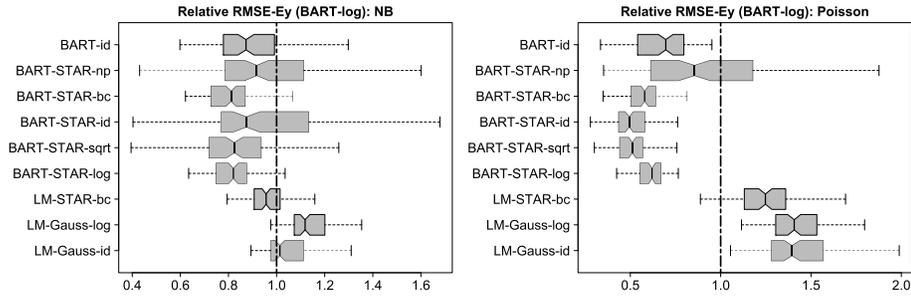and* LM-STAR-*log.*

Fig 10. *Relative RMSE under various distributions. Preferred models have smaller values, and models with values less than 1.0 are preferred to* BART-*log.*

A.2. Model and MCMC diagnostics for the dolphins data

Posterior predictive diagnostics for additive models fit to the tucuxis dolphins data are in Figure 11. The additive STAR-np model is adequate for the data, while the models which lack either rounding or transformation are incapable of capturing distributional features, including the variability and the proportion of zeros.

The MCMC convergence of the additive STAR-np model is assessed via traceplots in Figure 12. The traceplots indicate no lack of convergence and demonstrate exceptional mixing: effective sample sizes for all $f_j(v_j)$ exceed 2000.

A.3. Supplemental empirical examples

For further models comparisons, we apply the linear, additive, and BART models to several additional datasets, and again consider STAR, Gaussian, Poisson, and negative-binomial distributions. For the additive models, each continuous variable (i.e., variables with at least 10 unique observation points) is modeled nonlinearly.

Ships data

The ships data, available in the MASS package in R, provides the number of damage incidents due to waves for $n = 34$ cargo-carrying vessels, as well as ship type (A-E), year of construction (1960-1964, 1965-1969, 1970-1974, or 1975-1979), the period of construction (1960-1974 or 1975-1979), and the aggregated months of service (ranging from 0 to 44882). We model the ship type, year of construction, and period of construction as factors, and center and scale the service variable. The data were analyzed in McCullagh and Nelder (1989) using a quasi-Poisson regression model to account for observed overdispersion, and subsequently re-analyzed in Mallick and Gelfand (1994) using a Poisson regression model with unknown link function, which suggests that additional distributional flexibility in the regression model may be important.
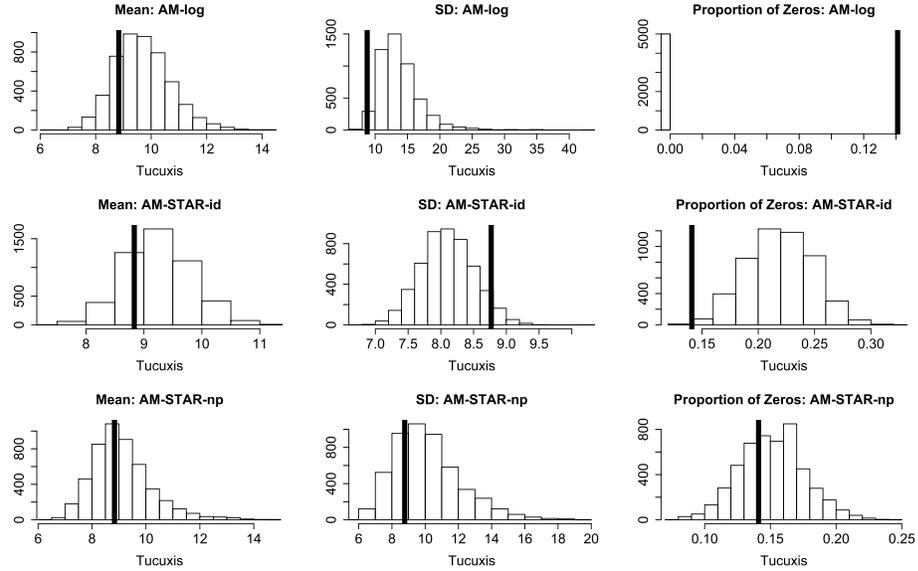
FIG 11. *Posterior predictive diagnostics for* AM-*log (top row),* AM-STAR-*id (middle row), and* AM-STAR-*np (bottom row). The mean (left), standard deviation (center), and proportion of zeros (right) were computed for each posterior predictive simulated dataset (histograms) and the observed data y (vertical lines). Only the model including both transformation and rounding (* AM-STAR-*np) is adequate by these measures for the dolphins data.*
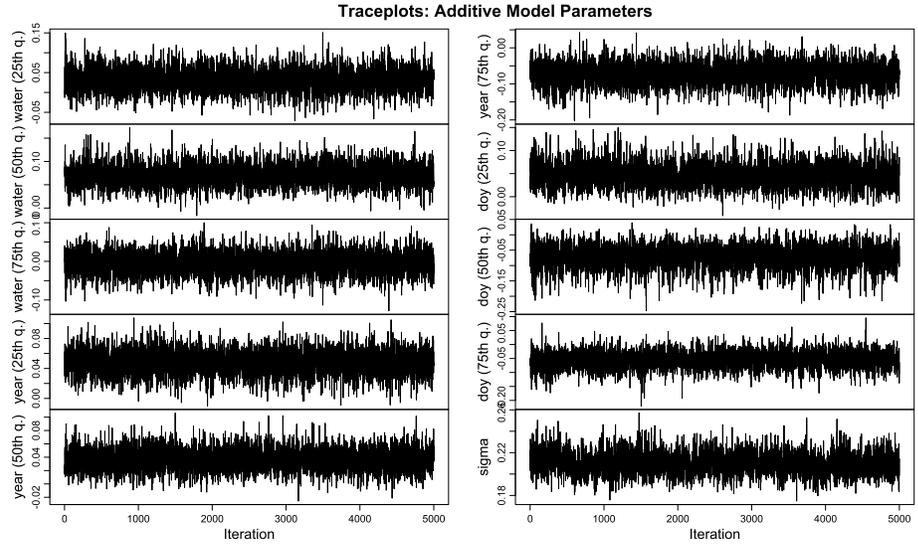

FIG 12. *Traceplots for $f_j(v_j)$ and σ under the additive* STAR-*np model for the dolphin data, where the functions f_j are evaluated at the 25th, 50th, and 75th sample quantiles of each $\{v_{i,j}\}_{i=1}^n$. The MCMC chain consisted of 5000 iterations (after discarding a burn-in of 5000 and retaining every 3rd sample). Effective sample sizes for all $f_j(v_j)$ exceeded 2000.*

Roaches data

Gelman and Hill ([2006](#)) consider a study of pest management for eliminating cockroaches in city apartments. The response variable, y_i, is the number of roaches caught in traps in apartment i, with $i = 1, \ldots, n = 262$. A pest management treatment was applied to a subset of 158 apartments, with the remaining 104 apartments receiving a control. Additional data are available on the pre-treatment number of roaches, whether the apartment building is restricted to elderly residents, and the number of days for which the traps were exposed. A notable feature of the data is zero-inflation: $y_i = 0$ for 94 (36%) of the apartments.

Highway data

The `Highway` data, available in the `carData` package in R, consists of the 1973 accident rate per million vehicle miles on $n = 39$ large sections of Minnesota highway. Important predictors include the number of access points per mile, the speed limit, the width of the outer shoulder on the roadway (in feet), and the number of signals per mile of roadway, among others. We consider the accident rate per 10,000 miles, which is the smallest rate for which the observations y_i are integer-valued. A notable feature of these data are that, despite being (scaled) accident counts, no two observations y_i and y_j are equal, and the counts themselves are large, ranging from 161 to 923. Therefore, it is unclear *a priori* whether an integer-valued model is necessary or advantageous.

Results

The WAICs for the supplementary datasets are reported in Table [2](#). The STAR models consistently perform well across all datasets, and in particular STAR-bc and STAR-np. Notably, STAR provides the best linear, additive, and BART models for all datasets with the exception of the `Highway` data, for which all BART models perform similarly. Interestingly, additive STAR models are preferred for both the `ships` data and the `Highway` data, while BART-STAR is slightly preferred to the additive STAR model for the `roaches` data.

TABLE 2
WAIC for the `ships`, `roaches`, and `Highway` datasets.

	Model	Gauss-id	Gauss-log	STAR-log	STAR-sqrt	STAR-id	STAR-bc	STAR-np	Pois	NB
ships	LM	292	204	204	195	220	196	**194**	239	203
	AM	240	188	193	173	182	175	**<u>171</u>**	–	–
	BART	242	204	200	**182**	191	196	187	–	–
roaches	LM	2722	1886	1791	1772	1952	**1756**	1759	12565	1793
	AM	2711	1838	1740	1732	1928	**1710**	1729	–	–
	BART	2700	1833	1736	1732	1927	**<u>1708</u>**	1719	–	–
Highway	LM	495	481	481	484	496	486	**479**	1587	496
	AM	487	465	465	473	484	478	**<u>460</u>**	–	–
	BART	466	**463**	465	465	469	476	464	–	–

The best method (lowest WAIC) for each model class is in bold, and the best overall model for each outcome is underlined. The STAR models dominate, with the additive and BART models typically outperforming the linear models.

Acknowledgements

References

Agresti, A. (2013). *Categorical data analysis*, 3rd edition. New York: Whiley. MR3087436   1748

Bening, V. E. and Korolev, V. Y. (2012). *Generalized Poisson models and their applications in insurance and finance*. Walter de Gruyter.   1745

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991. MR3620452   1755

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243. MR0192611   1749

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press. MR3155491   1745, 1760

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539. MR2896854 1748

Canale, A. and Dunson, D. B. (2013). Nonparametric Bayes modelling of count processes. *Biometrika*, 100(4):801–816. MR3142333   1748, 1749, 1761, 1763

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, pages 465–480. MR2650751   1754

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.   1755, 1756

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298. MR2758172   1755, 1756, 1759, 1760

da Silva, V. M., Freitas, C. E., Dias, R. L., and Martin, A. R. (2018). Both cetaceans in the Brazilian Amazon show sustained, profound population declines over two decades. *PloS one*, 13(5):e0191304.   1747, 1762, 1764

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, pages 909–917.   1748

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12(3):313–336. 1745, 1760

Dorazio, R. M., Jelks, H. L., and Jordan, F. (2005). Improving removal-based estimates of abundance by sampling a population of spatially distinct subpopulations. *Biometrics*, 61(4):1093–1101. MR2216203   1745

Dorie, V., Chipman, H., and R, M. (2018). *dbarts: Discrete Bayesian Additive Regression Trees Sampler*. R package version 3.1.   1756

Duan, L. L., Johndrow, J. E., and Dunson, D. B. (2018). Scaling up data augmentation MCMC via calibration. *The Journal of Machine Learning Research*, 19(1):2575–2608. MR3899766   1754

Famoye, F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 22(5):1335–1354. MR1225247   1746

Flegal, J. M., Hughes, J., Vats, D., and Dai, N. (2017). `mcmcse: Monte Carlo Standard Errors for MCMC`. Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN. R package version 1.3-2.   1763

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67. MR1091842   1759

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.   1769

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016. MR3253850   1757

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. MR2345548   1761

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2018). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.17.4.   1758

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press. MR2797563   1745

Klami, A. (2015). Pólya–Gamma augmentations for factor models. In *Asian Conference on Machine Learning*, pages 112–128.   1754

Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R*. Springer-Verlag, New York. ISBN 978-0-387-77316-2.   1760

Kowal, D. R. (2019). Integer-valued functional data analysis for measles forecasting. *Biometrics*. Forthcoming. MR4041833   1745, 1754

Lord, D., Guikema, S. D., and Geedipally, S. R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, 40(3):1123–1134.   1746

Mallick, B. K. and Gelfand, A. E. (1994). Generalized linear models with unknown link functions. *Biometrika*, 81(2):237–245. MR1294888   1749, 1767

Matteson, D. S., McLean, M. W., Woodard, D. B., and Henderson, S. G. (2011). Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, 5(2B):1379–1406. MR2849778   1745

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127. MR0583347   1748

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Vol. 37 of Monographs on Statistics and Applied Probability*. Chapman and Hall, London. MR3223057   1745, 1749, 1767

Murray, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. *arXiv preprint* arXiv:1701.01503.   1756

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, pages 705–741.

MR1994729   1753

Neelon, B. (2019). Bayesian zero-inflated negative binomial regression based on pólya-gamma mixtures. *Bayesian Analysis*, 14(3):849–875. MR3960773 1746

O'Hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122.   1746

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2018). Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy. *Bayesian Analysis*. MR3934087   1745

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349. MR3174712   1754, 1755

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical science*, 3(4):425–441.   1750, 1751

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338. MR1841418   1755

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532. MR3036413   1755

Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, pages 943–961. MR2758428   1746

Shen, H. and Huang, J. Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410.   1745

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142. MR2134602   1746

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337. MR3949306   1763

Vihola, M. (2012). Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008. MR2950080   1753

Wang, W. and Yan, J. (2018). `splines2`: *Regression Spline Functions and Classes*. R package version 0.2.8.   1750

Warton, D. I. (2018). Why you cannot transform your way out of trouble for small counts. *Biometrics*, 74(1):362–368. MR3777956   1746

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594. MR2756194   1757

Zhou, M., Li, L., Dunson, D., and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Proceedings of the International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access. 1746, 1754