A Bayesian goodness-of-fit test for regression

Andrés F. Barrientos[a,*], Antonio Canale[b]

[a]*Department of Statistics, Florida State University, 214 Rogers Building (OSB), 117 N. Woodward Ave. Tallahassee, Florida, 32306-4330, USA*
[b]*Department of Statistical Sciences, University of Padova, via C. Battisti 241, 35121, Padova, Italy*

Abstract

Regression models are widely used statistical procedures, and the validation of their assumptions plays a crucial role in the data analysis process. Unfortunately, validating assumptions usually depends on the availability of tests tailored to the specific model of interest. A novel Bayesian approach goodness-of-fit hypothesis testing approach is presented for a broad class of regression models the response variable of which is univariate and continuous. The proposed approach relies on a suitable transformation of the response variable and a Bayesian prior induced by a predictor-dependent mixture model. Hypothesis testing is performed via Bayes factor, the asymptotic properties of which are discussed. The method is implemented by means of a Markov chain Monte Carlo algorithm, and its performance is illustrated using simulated and real data sets.

Keywords: Bayes factor; Density regression; Dirichlet process mixture; Rosenblatt's transformation; Universal residuals

1. Introduction

Regression models are amongst the most extensively used statistical procedures. Despite the relatively wide range of nonparametric alternatives available, parametric regression models are the preferred modeling choice in many applications for their

☆Supplementary materials online include a description of the MCMC algorithm used in Section 3 and its R implementation. The R implementation of this algorithm is also available online at `https://anfebar.github.io/Software/BayesianGOFRegression/Bayesian-goodness-of-fit-testing-for-regression.html`.

*Corresponding author

Email addresses: `abarrientos@fsu.edu` (Andrés F. Barrientos), `canale@stat.unipd.it` (Antonio Canale)

ease of interpretation and estimation. However, the specific assumptions on which these parametric models rely are often questionable. Validating these assumptions is necessary since misspecified models can lead to erroneous inference and conclusions. Nonetheless, the ability to check assumptions usually depends on the availability of *ad hoc* formulations tailored to test specific aspects of the model. For example, the Shapiro-Wilk (Shapiro & Wilk, 1965) and the Jarque-Bera (Jarque & Bera, 1980) procedures test for normality and the Reset test of Ramsey (1969) assesses the linearity of the regression function. Hidalgo et al. (2018) propose a test to check whether the lack-of-fit comes from the incorrect parametric or nonparametric modelling of the regression function. Peña & Slate (2006) propose a test for globally testing the four assumptions of Gaussian linear regression models (i.e., linearity, homoskedasticity, uncorrelatedness, and normality). Despite addressing different aspects jointly, Peña & Slate's test still lacks generality as it is tailored to a specific class of models.

In this article, we aim to provide a global approach for testing the goodness-of-fit of general regression models. To this end, we present a novel Bayesian procedure applicable to a broad class of regression models whose response variable is univariate and continuous. The proposed approach departs from the ideas motivating standard approaches and exploits a suitable transformation of the response variable and a Bayesian nonparametric predictor-dependent mixture model.

There are very few Bayesian nonparametric contributions that propose general goodness-of-fit tests and that are applicable in a wide set of situations (see Tokdar et al. (2010) for a review). Furthermore, the majority of such works do not consider the regression framework, but rather focus on proposing tests for predictor-independent densities (Verdinelli & Wasserman, 1998; Berger & Guglielmi, 2001; Carota & Parmigiani, 1996; Robert & Rousseau, 2002; Basu & Chib, 2003; McVinish et al., 2009; Tokdar & Martin, 2013). To our knowledge, the only works focusing on regression models are those by Basu & Chib (2003) and Lu (2012), who compare parametric and nonparametric models via Bayes factors. Basu & Chib (2003) introduce an algorithm to approximate the marginal likelihood of Dirichlet process mixture models (Ferguson, 1973; Lo, 1984). Lu (2012) proposes a method to approximate a calibrated version of the Bayes factor between a parametric model and a Dirichlet process mixture model

alternative. Unfortunately, both of these proposals are difficult to implement when the goal is to test the fit of several regression models, as they require the analyst to derive or approximate the marginal likelihood of each model.

While Bayesian tests for goodness-of-fit are scarce, the frequentist literature on the topic is quite large (see, e.g., Miller & Neill, 2016, and references therein). Frequentist goodness-of-fit tests usually rely on either likelihood ratios or residual analysis. In the former case, the idea is to use likelihood ratio tests to compare the model of interest with a saturated version of it (Lindsey, 1997). Failure to detect differences between the two models suggests the model of interest fits the data well. Frequentist goodness-of-fit tests that rely on analysis of residuals, commonly defined as the response minus an estimate of the corresponding conditional mean (Eubank & Spiegelman, 1990; Fan & Huang, 2001), require a full characterization of the residuals' distribution. Unfortunately, such characterization is available only in a few cases (e.g., when the response is normally distributed).

Brockwell (2007, 2011) proposes a more general approach for defining the residuals of a regression model using Rosenblatt's transformation (see Rosenblatt, 1952). These residuals, referred to as universal residuals, take values in $(0, 1)$ and, under a correct model specification, are uniformly distributed. Universal residuals represent a powerful tool for defining Bayesian nonparametric goodness-of-fit tests. In fact, some Bayesian goodness-of-fit tests employ a simplified version of universal residuals (e.g., Verdinelli & Wasserman, 1998; Robert & Rousseau, 2002). Consistent with these approaches, our proposal exploits the fact that, under correct model specifications, universal residuals are not only uniformly distributed but also independent from the predictors. We propose using a Bayesian nonparametric approach to model universal residuals conditionally on predictors and look for deviations from both the uniformity and independence assumptions jointly. We assess these deviations in terms of the Bayes factor. Specifically, we use a mixture model based on predictor-dependent stick-breaking mixtures (MacEachern, 2000; De Iorio et al., 2004; Dunson & Park, 2008; Chung & Dunson, 2009; Jara et al., 2010; Barrientos et al., 2017). This class of models satisfies appealing properties in terms of flexibility (Barrientos et al., 2012) and large sample behavior (Pati et al., 2013; Norets & Pelenis, 2014).

The rest of the paper is organized as follows. In Section 2 we describe Rosenblatt's transformation and its relation to universal residuals. Then, we introduce our Bayesian goodness-of-fit test and discuss its properties. Section 3 describes a practical specification of our approach and discusses the related computational implementation. Illustrations of our proposal based on simulated and real data sets are provided in Section 4. Section 5 summarizes our findings and provides some directions for future work. All the proofs of our results are reported in the Appendix.

2. Goodness-of-fit test for regression

We consider a regression setting where $Y_i \in \mathbb{Y}$ is the response variable and $X_i \in \mathbb{X}$ is a vector of p predictors with $\mathbb{Y} \subseteq \mathbb{R}$ and $\mathbb{X} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$. Let $\{(y_i, x_i)\}_{i=1}^n$ be a collection of independent response and predictor values, realization of $\{(Y_i, X_i)\}_{i=1}^n$. Let $\mathcal{F} = \{F_x(\cdot) : x \in \mathbb{X}\}$ be the unknown true data generating mechanism, where $F_x(\cdot)$ denotes the cumulative distribution function of the response variable given the predictors, i.e., $Y_i|X_i = x_i \overset{\text{ind}}{\sim} F_{x_i}$, for $i = 1, \ldots, n$. Assuming $\mathcal{F}_0 = \{F_{0,x}(\cdot) : x \in \mathbb{X}\}$ to be a set of known conditional distribution functions, we consider the problem of testing whether $\mathcal{F}_0$ corresponds to the true conditional data generating mechanism. Specifically, we want to test

$$H_0 : \mathcal{F} = \mathcal{F}_0, \quad H_1 : \mathcal{F} \in \{\mathcal{F}_0\}^c, \tag{1}$$

where $\{\mathcal{F}_0\}^c = \mathcal{F}^{\mathbb{X}} \backslash \mathcal{F}_0$ and $\mathcal{F}^{\mathbb{X}}$ is the infinite dimensional set of all possible conditional data generating models of the form $\{\tilde{F}_x(\cdot) : x \in \mathbb{X}\}$. We aim to propose a Bayesian nonparametric procedure that controls the prior probability on the null hypothesis H_0 and computes its posterior probability. Thus, we can perform Bayesian hypothesis testing via Bayes factor,

$$\text{BF}_n = \frac{\pi\left(H_0|\{(y_i, x_i)\}_{i=1}^n\right)}{\pi\left(H_1|\{(y_i, x_i)\}_{i=1}^n\right)} \times \frac{\pi\left(H_1\right)}{\pi\left(H_0\right)}, \tag{2}$$

where π is a prior distribution on $\mathcal{F}^{\mathbb{X}}$ and $\pi\left(\cdot|\{(y_i, x_i)\}_{i=1}^n\right)$ is the corresponding posterior distribution.

Our proposal relies on the concept of universal residuals. The next subsections review this concept, explain how universal residuals help to re-formulate (1), and present the proposed method along with some large-sample guarantees.

2.1. Rosenblatt's transformation and universal residuals

The building block of our procedure is Rosenblatt's transformation of the random vector $Z = (Z_1, \ldots, Z_k)^T$, namely $R_G(z_1, \ldots, z_k) = (r_1, \ldots, r_k)$ where

$$r_1 = \text{pr}_G(Z_1 \leq z_1), \quad r_j = \text{pr}_G(Z_j \leq z_j | Z_1 = z_1, \ldots, Z_{j-1} = z_{j-1}), \tag{3}$$

with $\text{pr}_G(\cdot)$ the probability under G, the absolute continuous distribution of the random Z. As a result, $R_G(Z)$ is a random vector of independent random variables that are uniformly distributed (Rosenblatt, 1952). Based on the principles behind this transformation, we can define and study an analogous transformation for regression data, which is formalized in the next proposition.

Proposition 1. *Let (Y, X) be a random vector where Y and X denote the real-valued response and predictors, respectively. If $Y|X = x \sim G_x$, then*

$$U = G_X(Y), \tag{4}$$

is uniformly distributed on $[0, 1]$ and is independent from X.

The proof of this proposition is straightforward and relies on the fact that, for every $x \in \mathbb{X}$, the probability of the event $G_X(Y) < u$ conditional on $X = x$ is given by $\text{pr}_{G_x}(G_x(Y) < u) = u$. Brockwell (2007) uses part of Proposition 1 to define a goodness-of-fit test for univariate regression models. For a given data set $\{(y_i, x_i)\}_{i=1}^n$ and model of interest $\mathcal{F}_0$, Brockwell first defines the universal residuals u_i by means of the transformation in (4), that is, $u_i = F_{0,x_i}(y_i)$. Under this author's proposal, testing the goodness-of-fit of $\mathcal{F}_0$ is equivalent to testing the uniformity of $\{U_i\}_{i=1}^n$ based on $\{u_i\}_{i=1}^n$. While an important contribution, Brockwell's approach only focuses on testing uniformity and does not provide any insight regarding the independence between the universal residuals and predictors. This independence condition, however, is a key aspect that should be verified in all applications. By testing the independence

condition, we can determine whether all the information associated with the response that is contained in the predictors has been fully incorporated in the model. Our claim is that verifying the goodness-of-fit of $\mathcal{F}_0$ based on universal residuals requires jointly assessing deviations from uniformity and independence as opposed to only assessing uniformity.

To give the intuition of this, consider the following example. Assume that F_{0,x_i} is the true conditional cumulative distribution function of Y_i given $X_i = x_i$, for $i = 1, \ldots, n$. Let Q_0 be the bivariate distribution function of (Y, X) and define $Q_1(y) = \text{pr}_{Q_0}(Y \leq y)$ and $Q_2(y|x) = \text{pr}_{Q_0}(Y \leq y|X = x)$. Under this specification and using Brockwell's approach, we would assess whether $\mathcal{F}^* = \{F^*_x(\cdot) : x \in \mathbb{X}\}$ with $F^*_x(y) = Q_1(Q_2^{-1}(F_{0,x}(y)|x))$ is the true data generating process simply by testing the uniformity of $(U_1, \ldots, U_n)^T$, where $U_i = F^*_{x_i}(Y_i)$. Since F^*_x is not the true data generating process but the vector $(U_1, \ldots, U_n)^T$ is uniformly distributed, it is clear that by using Brockwell's strategy, the type II error in testing (1) will remain high regardless of the sample size.

This example leads us to question the uniqueness of $\mathcal{F}_0$ when defining universal residuals through (4) that are uniformly distributed and independent of the predictors. The following theorem addresses this issue.

Theorem 1. *Let (Y, X) be a random vector defined on a probability space $(\Omega, \mathcal{A}, P)$ and let $\mu_X(\cdot) = P[Y \in \mathbb{Y}, X \in \cdot]$, where Y is the response variable and X is a vector of predictors. Let $F_{0,x}$ and $F^*_{0,x}$ be conditional cumulative distribution functions such that $U = F_{0,X}(Y)$ and $U^* = F^*_{0,X}(Y)$ are uniformly distributed and are independent of X. Then $F_{0,X} = F^*_{0,X}$ almost surely μ_X.*

Theorem 1 implies that if $\mathcal{F}^*_0 \neq \mathcal{F}_0$ with $F_{0,x} = F^*_{0,x}$ for every x in a set $\mathbb{X}_0 \subseteq \mathbb{X}$ such that $\mu_X(\mathbb{X}_0) = 1$, then we have to assume that $\mathcal{F}^*_0$ and $\mathcal{F}_0$ are indistinguishable.

Remark 1. *Similarly, we can define universal residuals when the predictors correspond to a fixed design. More precisely, we assume that the distribution function of the response is indexed by the values of the design and use this distribution function to compute the universal residuals. In this case, instead of testing for independence between the residuals and predictors, we test that the distribution of the residuals (also*

indexed by predictors) is the same at any value of the design. Although there is a conceptual difference between random and fixed designs, there is no operational difference when testing goodness-of-fit with the proposed procedure.

We use the universal residuals to re-write the hypotheses (1) as

$$H_0 : U_i \sim \text{Unif}(0,1) \text{ and } U_i \perp\!\!\!\perp X_i, \qquad H_1 : U_i \not\sim \text{Unif}(0,1) \text{ or } U_i \not\perp\!\!\!\perp X_i,$$

where $U_i \perp\!\!\!\perp X_i$ denotes that U_i and X_i are independent. Note that we are not required to work with universal residuals that take values on $(0,1)$. For example, for practical reasons we may map them to the real line using the inverse of the standard Gaussian distribution function, say Φ^{-1}. Consistent with this choice, and with an abuse of notation we can re-define universal residuals as

$$U_i = \Phi^{-1}(F_{0,X_i}(Y_i)). \tag{5}$$

Although it is not strictly necessary, we prefer to re-write the hypotheses in (1) as,

$$H_0 : U_i \sim \text{N}(0,1) \text{ and } U_i \perp\!\!\!\perp X_i, \qquad H_1 : U_i \not\sim \text{N}(0,1) \text{ or } U_i \not\perp\!\!\!\perp X_i, \tag{6}$$

where $\text{N}(m,s)$ stands for the normal distribution with mean m and variance s. Hereafter, we will use the universal residuals defined in (5) and focus on testing the hypotheses stated in (6).

2.2. Bayesian nonparametric testing

Consistent with the discussion of Section 2.1, we test the hypotheses in (1) by examining whether the distribution of $U_i \mid X_i = x_i$ is a standard Gaussian distribution. To this end, we utilize a flexible Bayesian approach that estimates the conditional density of U_i given $X_i = x_i$ while assigning positive prior probability to the standard Gaussian model (i.e., the null hypothesis). More precisely, the Bayesian approach focuses on the definition of a prior probability π with large support on $\mathcal{P}\left(\mathbb{R}\right)^{\mathbb{X}}$ and such that $\pi(H_0) > 0$, where $\mathcal{P}\left(\mathbb{R}\right)^{\mathbb{X}}$ is the space of all predictor-dependent probability measures defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with continuous probability density function and $\mathcal{B}(\mathbb{R})$ being the Borel σ-algebra.

We define the prior π as the probability distribution induced by a stochastic process of the form $\mathcal{F} = \{F_x : x \in \mathbb{X}\}$, where $F_x(\cdot)$ is a probability distribution with density defined by the mixture

$$f_x(u) = \sum_{h=1}^{\infty} w_h \frac{1}{\sigma_h} \phi\left(\frac{u - \kappa_h(x)}{\sigma_h}\right), \tag{7}$$

w_h are random weights summing to one, and $\theta_h = (\kappa_h, \sigma_h) \sim P_0$ independently. The random weights are defined by means of a stick-breaking process (Sethuraman, 1994; Ishwaran & James, 2001), i.e., $w_h = V_h \prod_{l<h}(1 - V_l)$, with $V_h \sim \text{Beta}(a_h, b_h)$. The base probability measure P_0 has support on $K^{\mathbb{X}} \times \mathbb{R}^+$, with $K^{\mathbb{X}}$ the space of all $\mathbb{X} \to \mathbb{R}$ functions, and will rely on the assumption that κ_h and σ_h are independent.

The family $\mathcal{F}$ belongs to the class of predictor-dependent nonparametric mixture models commonly used in regression analysis (MacEachern, 2000; De Iorio et al., 2004; Dunson & Park, 2008; Chung & Dunson, 2009; Jara et al., 2010; Barrientos et al., 2017). These predictor-dependent models satisfy appealing properties in terms of flexibility (Barrientos et al., 2012) and large sample behavior (Pati et al., 2013).

The standard definition of Bayesian predictor-dependent mixture models assigns zero prior probability to the null hypothesis, which in turn implies the calculation of the Bayes factor is impossible. We now discuss how we modify the distribution of $\{\theta_h\}_{h\geq 1}$ such that the prior induced by (7) assigns positive mass to H_0. Under model (7), the null hypothesis is satisfied when the parameter θ_h is equal to $\theta_0 = (0, 1)$ for all h, i.e., $f_x(u) = \phi(u)$. We allow $\theta_h = \theta_0$, for all h, by introducing a binary variable ν that is equal to one with positive prior probability π_{H_0} and by assuming the following hierarchical structure for the component-specific parameter,

$$\{(\kappa_h, \sigma_h) \mid \nu = 1\} \sim \delta_{\theta_0}, \quad \{(\kappa_h, \sigma_h) \mid \nu = 0\} \sim P_0,$$

where δ_a is the Dirac measure. Under this formulation, $\theta_h = (\kappa_h, \sigma_h)$ can be either all equal to the specific value $\theta_0 = (0, 1)$ or all different (as long as P_0 is a nonatomic measure). Assuming prior distributions that assign point masses at some or all component specific parameters is a common approach in many other contexts, such as variable selection (Dunson et al., 2008; Yang, 2012; Barcella et al., 2016; Gutiérrez et al., 2018),

multiple testing (Bogdan et al., 2008; Do et al., 2005; Kim et al., 2009; Guindani et al., 2009), or functional clustering (Canale et al., 2017).

The prior measure π on $\mathcal{P}\left(\mathbb{R}\right)^{\mathbb{X}}$ is then identified by π_{H_0}, P_0, and the stick-breaking process' sequence of parameters (a_h, b_h). Under this formulation, we can express the Bayes factor as

$$\mathrm{BF}_n = \frac{\prod_{i=1}^{n} \phi(u_i)}{\int_{H_1} \prod_{i=1}^{n} f_{x_i}(u_i)\pi(d\mathcal{F})} = \frac{\pi(\nu = 1|\{(u_i, x_i)\}_{i=1}^{n})}{\pi(\nu = 0|\{(u_i, x_i)\}_{i=1}^{n})} \times \frac{\pi(\nu = 0)}{\pi(\nu = 1)}, \tag{8}$$

which turns out to be computationally manageable.

A desirable specification of π implies that BF_n remains consistent as the sample size increases. The following theorem provides sufficient conditions on π_{H_0}, P_0, and (a_h, b_h) that lead to consistency

Theorem 2. *Let $V_h \sim \mathrm{Be}(1, b_h)$, $\pi_{H_0} \in (0, 1)$, and $(\{b_h\}_{h\geq 1}, P_0)$ be defined as in Theorem 6.1 of Pati et al. (2013). Then $\mathrm{BF}_n \to \infty$ as $n \to \infty$ under H_0. Moreover, under H_1, $\mathrm{BF}_n \to 0$ as $n \to \infty$ if the proposed data-generating mechanism belongs to the class of elements of $\mathcal{P}\left(\mathbb{R}\right)^{\mathbb{X}} \setminus \mathcal{P}_0$ characterized by conditions A1-A5 in Pati et al. (2013).*

Remark 2. *In most real applications, users would be unable to propose a $\mathcal{F}_0$ without using an estimation procedure. In fact, users will have to deal with families of conditional distributions $\mathcal{F}_\Gamma = \{F_{\gamma,x}(\cdot), x \in \mathbb{X}, \gamma \in \Gamma\}$, indexed by a parameter space Γ. The approach described so far, however, assumes that the family of conditional distributions to be tested is fully specified. For this reason, when testing if $\mathcal{F}_\Gamma$ is the true data generating family, we apply our procedure to $\mathcal{F}_0 = \{F_{\hat{\gamma},x}(\cdot), x \in \mathbb{X}, \}$, where $\hat{\gamma}$ is an estimator of γ. Our conjecture is that if $\mathcal{F}_0$ is specified with a reasonable and consistent estimator of γ (e.g., using a maximum likelihood estimator), the Bayes factor will remain consistent. Section 3 provides empirical evidence in favor of this conjecture. A more formal and fully Bayesian approach would also assign a prior distribution to γ and obtain the joint posterior distribution of γ and the parameters in model* (7) *following the proposal of Verdinelli & Wasserman (1998). Unfortunately, this fully Bayesian approach leads to identifiability issues and would require specific Markov chain Monte Carlo (MCMC) implementations depending on the parametric structure of $\mathcal{F}_\Gamma$. Hence,*

such an approach would limit the applicability of our proposal.

Remark 3. *In Section 2.1, we show that the hypotheses of interest can equivalently focus on either the normality or uniformity of the universal residuals. Testing either one of these assumptions must produce same or similar results as long as the underlying modeling approaches share similar properties. For example, if we adapt the prior proposed in Barrientos et al. (2017) to test uniformity, we would expect to observe similar results. These similarities might depend on whether the sample size is large enough or the sampling strategies offer the same level of accuracy.*

3. Prior specification and computational details

To apply our method we first need to specify prior distributions for all the unknown parameters. In this section we describe our default choice of prior. The prior π on $\mathcal{P}\left(\mathbb{R}\right)^{\mathbb{X}}$ is induced through a simplified version of (7), where

$$f_x(u) = \sum_{h=1}^{H} w_h \frac{1}{\sigma_h} \phi\left(\frac{u - \kappa_h(x)}{\sigma_h}\right), \tag{9}$$

$w_h = V_h \prod_{l<h}(1 - V_l)$, $V_h \sim \text{Beta}(1, \alpha)$ for $h = 1, \ldots, H-1$, $V_H = 1$, and H is a conservative upper bound used to truncate the mixture to a finite number of components. We specify $\kappa_h(x)$ as

$$\kappa_h(x_i) = \sum_{j=1}^{p_1} \eta_{hj}(x_{i,j}) + (x_i^d)^T \phi_h^d, \quad \eta_{hj}(\cdot) = \sum_{l=1}^{L} \phi_{hjl} B_l(\cdot),$$

where $x_{i,j}$, $j = 1, \ldots, p_1$, denotes continuous predictors and x_i^d denotes a vector having the p_2 binary predictors. The function η_{hj} is defined using regression splines with B_l, a B-spline basis, and ϕ_{hjl}, a basis coefficient. The vector ϕ_h^d denotes the coefficients for the binary predictors. This construction induces a linear smoother. Hence, we can rewrite $\kappa_h(x_i)$ as $\kappa_h(x_i) = \tilde{x}_i^T \beta_h$, where $\tilde{x}_i$ is a vector containing all the basis expansions of the continuous predictors and the binary variables, and β_h is the vector of coefficients with suitable dimension p.

In our simulations, we use the simplest version of a cubic B-spline defined with zero knots. We specify the distribution for the atoms (β_h, σ_h) with a latent variable ν

that equals one if $\beta_1 = \cdots = \beta_H = \mathbf{0}$ and $\sigma_1 = \cdots = \sigma_H = 1$, and zero otherwise, where $\mathbf{0}$ denotes a vector with all components equal to zero.

Consistent with Section 2.2, we let

$$\sigma_h^2 = 1 \times \mathbb{1}_{\{\nu=1\}} + \sigma_h^{2*} \mathbb{1}_{\{\nu=0\}} \text{ and } \beta_h = \mathbf{0} \times \mathbb{1}_{\{\nu=1\}} + \beta_h^* \mathbb{1}_{\{\nu=0\}},$$

with $\sigma_h^{2*} \sim \text{InvGamma}(1.5, 0.5)$ and $\beta_h^* | \sigma_h^2, g_h \sim \text{N}_p(\mathbf{0}, \sigma_h^2 g_h (\tilde{x}^T \tilde{x})^{-1})$, where $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)^T$ denotes the design matrix. We also consider a conditional distribution for $g_h | \nu$ given by

$$g_h | \nu = 1 \sim \text{InvGamma}((p+1)/2, n_h/2), \quad g_h | \nu = 0 \sim \text{InvGamma}(1/2, n_h/2),$$

where $n_h = n \times E[w_h] + 1$ and $E[w_h]$ denotes the prior mean of w_h. The prior for β_h^* is known as the g-prior (Zellner, 1986), a prior commonly used in model selection. The g-prior shrinks towards zero as h increases and $E[w_h]$ decreases. We complete the prior specification by assuming $\nu \sim \text{Bernoulli}(\pi_{H_0})$, $\pi_{H_0} = 0.5$, and $\alpha \sim \text{Gamma}(0.25, 0.25)$. The truncation level H is fixed at 50. Under this specification, model (9) represents a practical approximation of model (7). Note that $H = 50$ can be safely considered a conservative choice as under H_0 we expect that a single mixture component (i.e. $H = 1$) should be sufficient. Larger values of H and moderate changes of the hyperparameters lead to similar results (not reported here) as model (9) is fitted to the universal residuals that, by definition, are somehow "standardized." Our empirical experiments suggest that this specific choice is stable and leads to consistent conclusions across different scenarios.

In the following section, posterior summaries are obtained from $1{,}000$ draws from the posterior distribution of the parameters using the MCMC algorithm described in the Supplemental Material with a burn-in period of $5{,}000$ iterations and thinning of 10. In running the MCMC procedure for a finite number of iterations, it is possible that the posterior probability for the null equals one. When this happens, we could set the Bayes factors equal to infinity. However, this would make any graphical illustration unfeasible. Instead, we subtract a random uniform noise between 0 and 0.001 to such probabilities, resulting in finite Bayes factors. The proposed approach is summarized in Algorithm 1.

Step 0: Provide the following inputs,

data: $\{(y_i, x_i)\}_{i=1}^n$;

conditional model under H_0: $\{F_x(\cdot) : x \in \mathbb{X}\}$;

posterior sample size: N;

Step 1: Compute the universal residuals

$u_i \leftarrow \Phi(F_{x_i}(y_i)), \quad$ for $i = 1, \ldots, n$;

Step 2: Fit model (7) to $\{(u_i, x_i)\}_{i=1}^n$ via MCMC and obtain N posterior draws

Step 3: Approximate the posterior probability of H_0

$\pi_{H_0|-} \leftarrow \frac{1}{N} \sum_{t=1}^N \nu^{(t)}$;

Step 4: Approximate the Bayes factor

$\text{BF}_n \leftarrow \frac{\pi_{H_0|-}}{(1-\pi_{H_0|-})} \times \frac{(1-\pi_{H_0})}{\pi_{H_0}}$;

Algorithm 1: Step by step procedure for the proposed Bayesian goodness-of-fit test. The notation $\nu^{(t)}$ stands for the t-th draw of the parameter ν.

4. Illustrations

In this section, we illustrate the use of the proposed procedure by means of different simulated and real data sets. Throughout this section, we use the term "residual" (unaccompanied by other terms) with its more classical meaning, i.e. observed response minus predicted response.

4.1. Synthetic data

To illustrate the performance of our approach and to provide a fair comparison with different competing methods under different model assumptions we conduct a simulation study, simulating synthetic data from different scenarios described later. For each of scenario, we consider three sample sizes, namely $n = 100$, 250, and 500, and simulate ten predictors (five continuous and five discrete) using the same mechanism. Specifically, let $x_i = (1, x_{i,1}, \ldots, x_{i,10})^T$, where $x_{i,1} \sim \text{N}(0, 1)$, $(x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5}) \sim \text{N}(0, \Sigma_x)$, $\Sigma_x = \{\sigma_{j,j'}\}$, $\text{cov}(x_{i,j}, x_{i,j'}) = 0.7^{|j-j'|}$, and $x_{i,j} \sim \text{Bernoulli}(0.5)$, $j = 6, \ldots, 10$. We generate $N = 100$ sets of predictors so that there are one hundred

simulated data replicates for each scenario and sample size.

We evaluate the performance of our procedure under four scenarios. In the first scenario, universal residuals are simulated from $\mathrm{N}(0,1)$, i.e., we assume that H_0 holds. A second scenario considers the most common multiple linear regression, i.e., cases when the response is equal to a linear combination of the predictors plus error. The third and fourth scenarios deal with beta and gamma regressions, respectively. For scenarios two, three, and four, we test the goodness-of-fit for models that are correctly and incorrectly specified.

For the first scenario, let $\mathcal{M}_{0,0}$ be the model that simulates $u_i \sim \mathrm{N}(0,1)$. When testing H_0, we expect high Bayes factors to provide strong evidence in favor of the null hypotheses. This is indeed what we see from the boxplots in panel (a) of Figure 1. These boxplots represent the distribution of the N Bayes factors for H_0 when the universal residuals are simulated from $\mathcal{M}_{0,0}$ for different sample sizes.

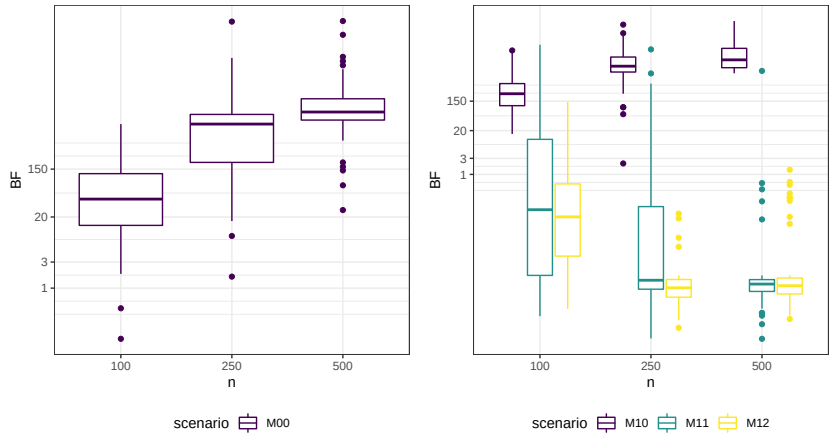
In the second scenario, we simulate data from the following models

$$\mathcal{M}_{1,0} = \{F_{x_i}(\cdot) : y_i \sim \mathrm{N}(x_i^T\beta, \sigma^2)\},$$
$$\mathcal{M}_{1,1} = \{F_{x_i}(\cdot) : y_i \sim \mathrm{N}(x_i^T\beta + .4[5\exp(x_{i,1}) + 4\exp(x_{i,2}) + \ldots + \exp(x_{i,5})], \sigma^2)\},$$
$$\mathcal{M}_{1,2} = \{F_{x_i}(\cdot) : y_i = x_i^T\beta + \epsilon_i, \;\; \epsilon_i \sim 0.25\mathrm{N}(0,4) + 0.5\mathrm{N}(0,1) + 0.25\mathrm{N}(0,0.25)\},$$

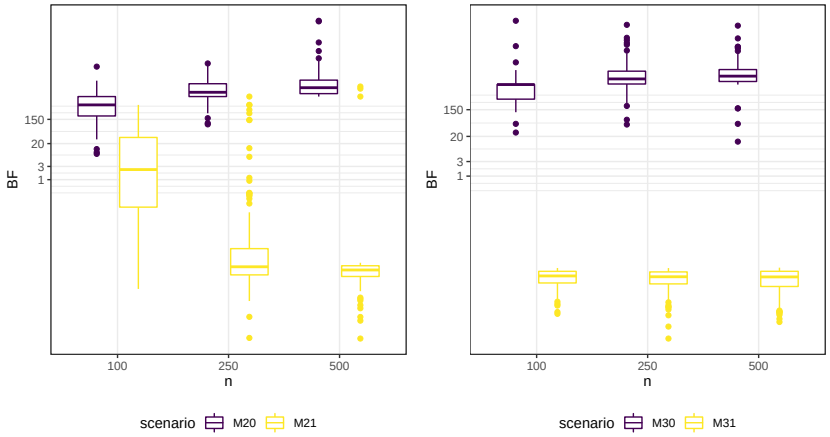
fixing β to a vector of ones and $\sigma^2 = 1$. Then, we test the goodness-of-fit for the simplest specification, i.e.,

$$H_0 : \mathcal{F} = \{F_{x_i}(\cdot) : y_i \sim \mathrm{N}(x_i^T\beta, \sigma^2)\},$$

where both the regression coefficients β and the error variance σ^2 are estimated via maximum likelihood from the simulated data. $\mathcal{M}_{1,1}$ represents a scenario where the conditional mean is "almost" linear while $\mathcal{M}_{1,2}$ represents the correct specification of the mean function with a misspecified distribution of errors. Panel (b) of Figure 1 reports the distributions of the Bayes factors in these different situations.



(a) (b)



(c) (d)

Figure 1: Monte Carlo distribution of the N Bayes factors in the simulation study as a function of the sample size; panel (a) represents $H_0 : \mathcal{F} = M_{0,0}$, panel (b) $H_0 : \mathcal{F} = M_{1,0}$, panel (c) $H_0 : \mathcal{F} = M_{2,0}$, and panel (d) $H_0 : \mathcal{F} = M_{3,0}$. In each panel, the y-axis is in log10 scale.

In the third scenario, simulated data are generated from the models

$$\mathcal{M}_{2,0} = \{F_{x_i}(\cdot) : y_i \sim \text{Beta}(\mu_i\gamma, (1-\mu_i)\gamma), \;\; \mu_i = (1+\exp(-x_i^T\beta))^{-1}\},$$
$$\mathcal{M}_{2,1} = \{F_{x_i}(\cdot) : y_i \sim \text{Beta}(\mu_i\gamma_i, (1-\mu_i)\gamma_i), \;\; \mu_i = (1+\exp(-m_i))^{-1},$$
$$\gamma_i = \exp(\tilde{m}_i), m_i = 2 - 0.15x_i^T 1_{p+1}, \; \tilde{m}_i = 1 + x_{i,1} + x_{i,10}\}$$

where, for $\mathcal{M}_{2,0}$, we let $\beta = (1.85, -0.15, \ldots, -0.15)^T$ and $\gamma = 1.5$. We test the goodness-of-fit for model $\mathcal{M}_{2,0}$ (i.e., $H_0 : \mathcal{F} = \mathcal{M}_{2,0}$) estimating its parameters via maximum likelihood from the simulated data. Model $\mathcal{M}_{2,0}$ is then correctly specified while model $\mathcal{M}_{2,1}$ requires adding predictors to the precision parameter. Panel (c) of Figure 1 shows the distributions of the corresponding Bayes factors.

Finally, for the fourth scenario, we consider a gamma regression and simulate data according to the following models

$$
\begin{aligned}
\mathcal{M}_{3,0} &= \{F_{x_i}(\cdot) : y_i \sim \text{Gamma}(\mu_i, \gamma_i/\mu_i),\ \mu_i = \exp(x_i^T \tilde{\beta}_1),\ \ \gamma_i = \exp(x_i^T \tilde{\beta}_2)\} \\
\mathcal{M}_{3,1} &= \{F_{x_i}(\cdot) :\ y_i \sim \text{Gamma}(\mu_i, \gamma_i/\mu_i),\ \mu_i = \exp(1 + m_i), \\
&\qquad \gamma_i = \exp(-5 + \tilde{m_i}), m_i = 0.2x_{i,1} + 3x_{i,10} + \exp(1 + x_{i,2}), \\
&\qquad \tilde{m}_i = 0.1x_i^T 1_{p+1} + \exp(0.3(x_{i,2} + 3))\},
\end{aligned}
$$

where, for $\mathcal{M}_{3,0}$, we let $\tilde{\beta}_1 = (1, 0.2, 0, \ldots, 0, 3)$ and $\tilde{\beta}_2 = (-4.9, 0.1, \ldots, 0.1)$. We then test the goodness-of-fit of model $\mathcal{M}_{3,0}$ estimating its parameters via maximum likelihood. Panel (d) of Figure 1 reports the results.

As expected, when the model is correctly specified, the Bayes factor increases as n increases. Conversely, when the model is incorrectly specified, the Bayes factor decreases as n increases. This behavior provides empirical evidence in favor of the Bayes factor's consistency provided in Theorem 2 and aligns with the conjecture in Remark 2.

The Bayes factors obtained in the first scenario (i.e., without a plug-in estimate) are slightly smaller than those obtained in the remaining scenarios. This is unsurprising because in those scenarios H_0 is partially specified by the data, which might increase the evidence in favor of this hypothesis.

The comparison of the proposed Bayesian test with frequentist competitors is not straightforward as the Bayes factor is intrinsically different from the idea of p-values. However, if the goal is to reach a decision, we can decide to reject the null hypothesis if the Bayes factor is below a given threshold or the p-value is below a significance level. We can then assess the performance of these rules using the frequentist operating characteristics. To this end, we consider competitors that can test assumptions on the

universal residuals. For Brockwell's proposal, we test the normality of the residuals through Kolmogorov-Smirnov and Shapiro-Wilk tests. Another strategy is to regress the residuals on the predictors using a linear model with normal errors and then check the corresponding underlying assumptions. This strategy is implemented using the global procedure proposed by Peña & Slate (2006), which tests the four assumptions (linearity, homoscedasticity, uncorrelatedness, normality) of the linear model while accounting for multiple testing. We also use this global approach to assess goodness of fit when the underlying proposed model for the original data is linear with normal errors (as in the second scenario). Notice that, as in our proposal, these competitors involve a parameter estimation step for all but the first scenario. We decide to reject H_0 if the Bayes factor < 1, or the p-value < 0.05.

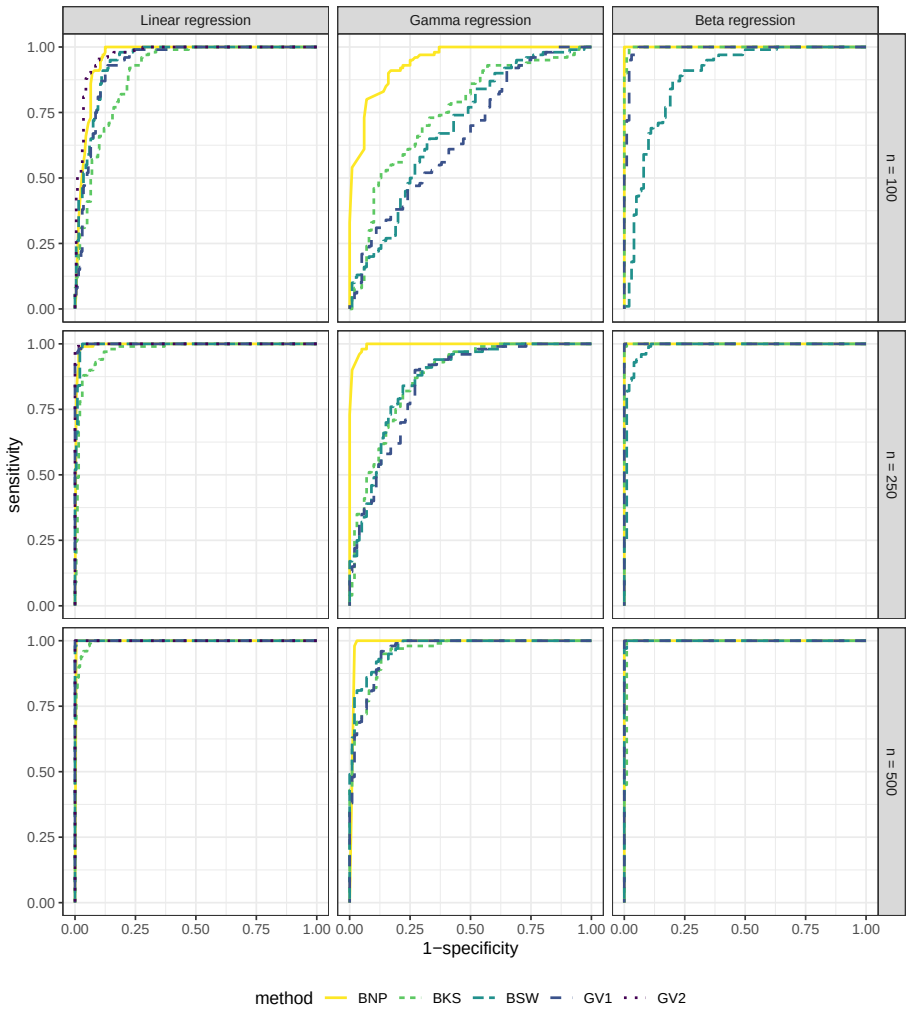
For each simulation scenario, we compute the type I or type II errors (depending on whether H_0 is true or false) simply counting the number of false rejections/acceptances over the $N = 100$ simulated data sets. Table 1 reports the results labeled as BNP for our proposal, and BKS and BSW for the Kolmogorov-Smirnov and Shapiro-Wilk tests, respectively, under Brockwell's strategy. The table also reports GV1 and GV2 for the global validation test of Peña & Slate (2006) applied to the universal residuals and to the original data, respectively. As already discussed, the procedure proposed by Peña & Slate (2006) is designed for models similar to those considered in the second scenario. Nonetheless, using the results in Section 2.1 and the strategy described in the previous paragraph, we can apply the global validation test to a broader class of models, e.g., to gamma and beta regression models.

Type I error becomes smaller as the sample size increases for all scenarios, with our proposed approach reporting errors comparable or smaller than the competitors. Regarding the type II error, the performance of the proposed test is comparable or better than the frequentist tests in most situations. As an exception, the type II error of the proposed procedure is higher when testing $\mathcal{M}_{10}$ and, particularly, using the approach of Peña & Slate (2006) with small sample sizes. This is unsurprising since the approach of Peña & Slate (2006) is specifically tailored to test $\mathcal{M}_{10}$.

The results in Table 1 clearly depend on the specific value of thresholds used to reject the null (in this case, Bayes factor < 1 or p-value < 0.05). To assess the robust-

Table 1: Type I, and Type II errors for $N = 100$ replicates; first and second columns denote the data generating process and the model under H_0, respectively; column BNP denotes the decisions based on rejecting H_0 when the Bayes factor is < 1; columns BKS, BSW, GV1, and GV2 report the results of the frequentist tests and denote the decisions based on rejecting H_0 when the p-value is < 0.05

			Type I errors					Type II errors				
Truth	H_0	n	BNP	BKS	BSW	GV1	GV2	BNP	BKS	BSW	GV1	GV2
		100	.02	.06	.07	.07	-	-	-	-	-	-
$\mathcal{M}_{00}$	$\mathcal{M}_{00}$	250	.00	.04	.05	.05	-	-	-	-	-	-
		500	.00	.03	.06	.06	-	-	-	-	-	-
		100	.00	.00	.01	.06	.04	-	-	-	-	-
$\mathcal{M}_{10}$	$\mathcal{M}_{10}$	250	.00	.00	.06	.05	.06	-	-	-	-	-
		500	.00	.00	.05	.05	.02	-	-	-	-	-
		100	-	-	-	-	-	.38	.93	.35	.21	.03
$\mathcal{M}_{11}$	$\mathcal{M}_{10}$	250	-	-	-	-	-	.16	.60	.04	.02	.00
		500	-	-	-	-	-	.01	.28	.00	.00	.00
		100	-	-	-	-	-	.16	.87	.13	.15	.17
$\mathcal{M}_{12}$	$\mathcal{M}_{10}$	250	-	-	-	-	-	.00	.20	.00	.00	.00
		500	-	-	-	-	-	.01	.00	.00	.00	.00
		100	.00	.01	.05	.06	-	-	-	-	-	-
$\mathcal{M}_{20}$	$\mathcal{M}_{20}$	250	.00	.00	.08	.06	-	-	-	-	-	-
		500	.00	.02	.06	.03	-	-	-	-	-	-
		100	-	-	-	-	-	.57	.94	.69	.72	-
$\mathcal{M}_{21}$	$\mathcal{M}_{20}$	250	-	-	-	-	-	.10	.62	.34	.41	-
		500	-	-	-	-	-	.03	.27	.16	.14	-
		100	.00	.00	.03	.03	-	-	-	-	-	-
$\mathcal{M}_{30}$	$\mathcal{M}_{30}$	250	.00	.01	.04	.04	-	-	-	-	-	-
		500	.00	.02	.03	.04	-	-	-	-	-	-
		100	-	-	-	-	-	.00	.08	.39	.04	-
$\mathcal{M}_{31}$	$\mathcal{M}_{30}$	250	-	-	-	-	-	.00	.00	.08	.00	-
		500	-	-	-	-	-	.00	.01	.01	.00	-



(a)

Figure 2: Smoothed ROC curves for the three scenarios having null $\mathcal{M}_{10}$ (linear regression), $\mathcal{M}_{20}$, (gamma regression) and $\mathcal{M}_{30}$ (beta regression); BNP curves display the results obtained using the Bayes factor.

ness of the considered procedures to variations of the threshold choice, Figure 2 reports the receiver operating characteristic (ROC) curves for the three scenarios having null $\mathcal{M}_{10}$, $\mathcal{M}_{20}$, and $\mathcal{M}_{30}$. To obtain these ROC curves, we combine samples that have the same H_0 and label them according to whether H_0 is indeed true or not. Then, for

different thresholds, we calculate the specificity (correct acceptances of H_0) and sensitivity (correct rejections of H_0) using the posterior probability of H_0 and the p-values, respectively. These ROC curves are consistent with the results highlighted in Table 1 and show that our procedure's performance is uniformly better or comparable to that of Brockwell's strategy. This empirical evidence supports the need to assess not only normality but also independence. The ROC curves also show that, as expected, our procedure is outperformed by Peña & Slate's method when testing $\mathcal{M}_{10}$ and having small sample sizes.

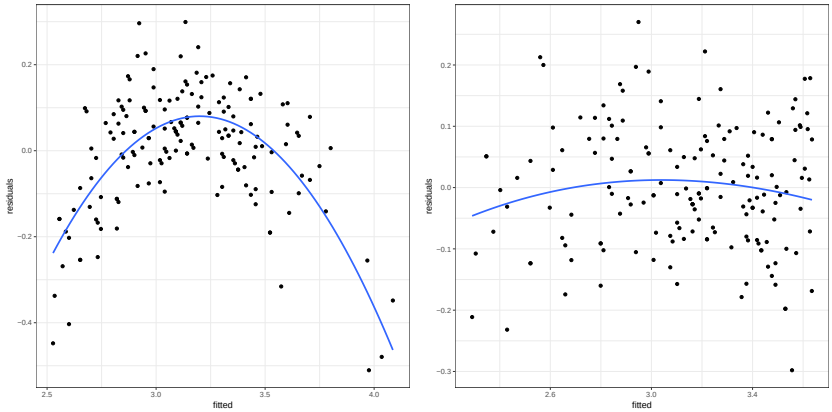
We conclude this section briefly discussing the specific choice discussed in Section 3 for what concerns the conservative upper bound H. We found that the a posteriori mixture concentrates 99% of the total mass among its first 45 components in each of the cases analyzed in this section, indicating the specific choice of $H = 50$ is robust.

4.2. Mandible length data

As a first illustration of our method on real data, we consider a simple data set concerning fetal growth that was first introduced by Chitty et al. (1993) and available in the R package `lmtest`. The data set contains the ultrasonographic measurements of mandible length in $n = 158$ fetuses along with the gestational age (in weeks) at which the measurement was taken. We attempt to model the well-known positive relationship between mandible length and gestational age. We first fit a linear model with log-length as response and gestational age as predictor with homoskedastic Gaussian errors. We use a maximum likelihood approach. While the logarithmic transformation of the response variable clearly stabilizes its variability, the analysis of residuals (reported in panel (a) of Figure 3) shows that the simple linear relation between log-length and gestational age is not sufficient to fully describe the relationship. After computing the universal residuals based on the estimated parameters, the proposed goodness-of-fit procedure reports a Bayes factor of 0, suggesting misspecification issues.

A second more realistic model assumes a quadratic relation between log-length and gestational age. Specifically, we include the squared gestational age as an additional predictor. Panel (b) of Figure 3 displays the residuals. The figure does not show signs of misspecification. Consistent with this, our procedure reports a Bayes factor of 59.6,

which can be interpreted as strong evidence in favor of the second model specification according to the Bayes factor's classification by Kass & Raftery (1995).



(a) (b)

Figure 3: Scatterplots of the residuals against fitted values along with nonparametric smoothing (continuous line) for the (a) linear regression model including only a linear term for the gestational age and (b) linear regression model including linear and quadratic terms for the gestational age for the Mandible length data set.

4.3. Australian Institute of Sport data

We consider now a data set comprising a sample of 202 Australian athletes who trained at the Australian Institute of Sport. For each athlete, 13 variables are recorded, but here we limit the analysis on modeling the (log) plasma ferritin concentration, henceforth `log-Fe`, as a function of the lean body mass (`LBM`) index and body mass index (`BMI`). The data set is available in the R package `sn`. Simple exploratory data analyses reveal mild linear relations between `log-Fe` and both `BMI` and `LBM`. Hence, as a first model, we fit a Gaussian linear regression assuming

$$\texttt{logFe} = \alpha + \beta_1 \times \texttt{BMI} + \beta_2 \times \texttt{LBM} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2). \tag{10}$$

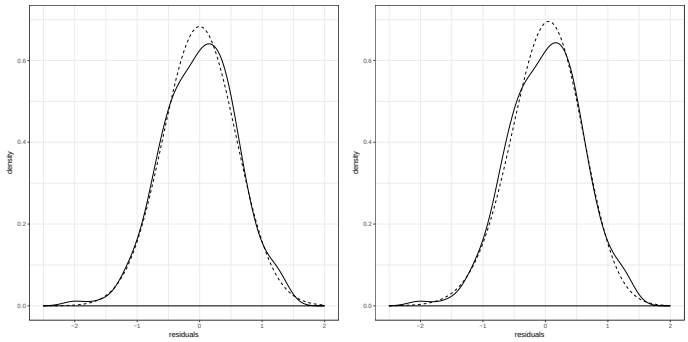
Panel (a) of Figure 4 presents the kernel density estimate of the residuals, obtained by fitting model (10) via maximum likelihood. Although the kernel density estimates show mild left skewness, the fitted (Gaussian) residuals' density (represented by the

dashed line) suggests a suitable fit. The goodness of fit is confirmed by the proposed method which returns a Bayes factor of 73, implying strong evidence in favor of the Gaussian assumption following the classification by Kass & Raftery (1995).

The mild skewness of the residuals in this first regression model, however, motivates further investigations. A possible extension consists in going beyond the normality assumption of the error terms by assuming a skew distribution of the error terms. A straightforward solution consists of assuming a skew-normal (Azzalini, 1985) distribution being a good compromise between mathematical tractability and flexibility in modeling the skewness but other parametric families may lead to similar conclusions. Consistent with this, we fit the following model

$$\log\texttt{Fe} = \alpha + \beta_1 \times \texttt{BMI} + \beta_2 \times \texttt{LBM} + \eta, \quad \eta \sim \mathrm{SN}(0, \omega^2, \alpha), \tag{11}$$

where $\mathrm{SN}(\xi, \omega^2, \alpha)$ denotes a skew-normal distribution with location ξ, scale ω^2, and shape α. Panel (b) of Figure 4 presents the kernel density estimates of the residuals obtained through model (11) via maximum likelihood. The mild skewness is now estimated and the newer formulation seems to provide a slightly better fit to the data at hand. This is confirmed by the Bayes factor of 124, which is higher than that under model (10) and provides stronger evidence in favor of the skew-normal assumption.



(a) (b)

Figure 4: Kernel density estimates (continuous lines) of the residuals and corresponding estimated Gaussian density (dashed lines) for (a) the Gaussian regression model and (b) the skew-normal regression model for the Australian Institute of Sport data set.

4.4. Children reading accuracy and dyslexia data

We now replicate the analysis of the children reading accuracy data set of Smithson & Verkuilen (2006) conducted by Cribari-Neto & Zeileis (2010) to illustrate the beta regression package `betareg`. The goal is to investigate whether `dyslexia`, a dummy variable separating a dyslexic and a control group of children, contributes to explaining children's reading accuracy (`accuracy`), a continuous score in the open unit interval, while controlling for a nonverbal intelligence quotient index (`IQ`). Figure 5 shows that the data are clearly asymmetric and heteroskedastic, particularly in the control group. Despite this, and consistent with Smithson & Verkuilen (2006), we first fit a Gaussian regression model using the logit transformation of accuracy as the response variable to account for the fact that accuracy ranges from 0 to 1. Specifically, we fit the model

$$\text{logit}(\texttt{accuracy}) = \beta_0 + \beta_1 \times \texttt{IQ} + \beta_2 \times \texttt{dyslexia} + \beta_3 \times \texttt{(IQ:dyslexia)} + \epsilon, \tag{12}$$

where the coefficient β_3 accounts for the interaction between the two regressors and $\epsilon \sim \text{N}(0, \sigma^2)$. It is not surprising that the model provides a poor fit to the data. The poor fit is confirmed when applying our procedure: we obtained a Bayes factor of 4.58. Given the poor performance of model (12), Smithson & Verkuilen (2006) suggest to fitting a more appropriate beta regression model. As in Cribari-Neto & Zeileis (2010), we let

$$\begin{aligned} \texttt{accuracy} &\sim \text{Be}(\mu\phi, (1-\mu)\phi) \\ \text{logit}(\mu) &= \beta_0 + \beta_1 \times \texttt{IQ} + \beta_2 \times \texttt{dyslexia} + \beta_3 \times \texttt{(IQ:dyslexia)} \\ \log(\phi) &= \gamma_0 + \gamma_1 \times \texttt{IQ} + \gamma_2 \times \texttt{dyslexia}, \end{aligned} \tag{13}$$

where the parameter μ represents the expectation of `accuracy` and ϕ its precision. While the beta regression fit does not differ from the normal regression fit for the group with dyslexia, the fit is much better for the control group. This improvement is confirmed by our goodness of fit approach, which returns a Bayes factor of 43.44 (thus providing strong evidence in favor of H0).

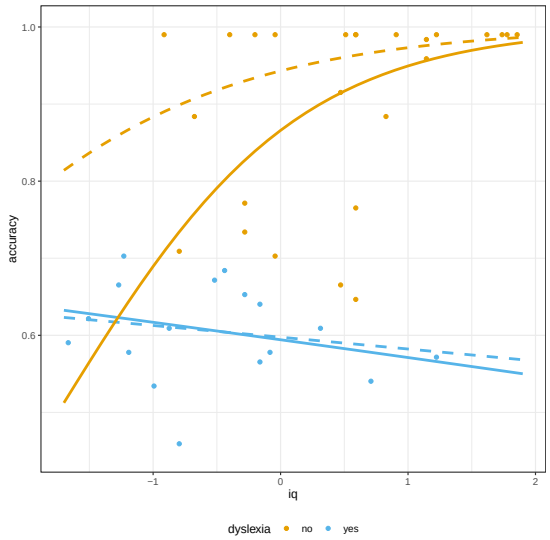


Figure 5: Reading skills and dyslexia data (Smithson & Verkuilen, 2006): Linearly transformed reading accuracy by IQ score and dyslexia status (control, orange vs. dyslexic, blue). Fitted curves correspond to beta regression (solid) and normal linear regression with logit-transformed dependent variable (dashed).

5. Discussion

We present a novel Bayesian approach to test goodness-of-fit of regression models via the Bayes factor. The approach can be applied to models with univariate and continuous response. The implementation of the proposed method is not specific to the regression model and only requires predictors and universal residuals as input. As discussed in Section 2.2, our proposal has desirable asymptotic properties; simulation studies and the real data analyses reported in Section 3 show that the procedure performs well in different scenarios.

Future research is needed to study the asymptotic properties of the Bayes factor when the proposed $\mathcal{F}_0$ is defined through an estimation procedure as discussed in Remark 2. Furthermore, extensions of the proposed procedure to more general regression models, such as those with a multivariate response, are subject to ongoing research.

Acknowledgments

The authors would like to thank Pierpaolo De Blasi, David Dunson, Victor Peña, and Mauricio Sadinle for helpful comments on previous versions of this article. AFB work is partially supported by grant FYAP program at Florida State University. AC work is supported by the University of Padova under the STARS Grant.

Appendix

Proof of Theorem 1. Let $\lambda(\cdot) = u\mu_X(\cdot)$, where $u \in [0,1]$. Then, λ is absolutely continuous with respect to μ_X and, by the Lebesgue decomposition theorem,

$$\lambda(A) = \int_A T d\mu,$$

where T is unique up to sets of μ_X-measure zero. It follows that $T = u$ almost surely μ_X. The assumption of uniformity and independence implies that,

$$\text{pr}\left[X \in A, U \leq u\right] = \text{pr}\left[X \in A\right] \times \text{pr}[U \leq u] = u\mu_X(A) = \lambda(A).$$

On the other hand, one has that

$$\begin{aligned}
\text{pr}\left[X \in A, U \leq u\right] &= \text{pr}\left[X \in A, Y \leq F_{0,X}^{-1}\left(u\right)\right], \\
&= E\left[\mathbb{I}_{\{X \in A\}}\mathbb{I}_{\left\{U^* \leq F_{0,X}^*\left(F_{0,X}^{-1}(u)\right)\right\}}\right], \\
&= E\left[E\left[\mathbb{I}_{\{X \in A\}}\mathbb{I}_{\left\{U^* \leq F_{0,X}^*\left(F_{0,X}^{-1}(u)\right)\right\}} \mid X\right]\right], \\
&= E\left[\mathbb{I}_{\{X \in A\}}E\left[\mathbb{I}_{\left\{U^* \leq F_{0,X}^*\left(F_{0,X}^{-1}(u)\right)\right\}} \mid X1\right]\right], \\
&= E\left[\mathbb{I}_{\{X \in A\}}F_{0,X}^*\left(F_{0,X}^{-1}\left(u\right)\right)\right] = \int_A F_{0,X}^*\left(F_{0,X}^{-1}\left(u\right)\right) d\mu_X.
\end{aligned}$$

Hence,

$$\lambda(A) = \int_A u d\mu = \int_A F_{0,X}^*\left(F_{0,X}^{-1}\left(u\right)\right) d\mu,$$

which implies that $F_{0,X}\left(F_{0,X}^{*-1}(u)\right) = u$ almost surely μ_X, i.e. $F_{0,X}(u) = F_{0,X}^*(u)$ almost surely μ_X. □

Proof of Theorem 2. We prove consistency using Theorem 1 and 3 in Dass & Lee (2004). Dass & Lee (2004) provide results in the single-density context. However,

our proposed test is framed within a conditional/regression context. For this reason, we apply Dass & Lee's theorems assuming that the element of interest is the joint distribution $m(u, x) = f_x(u)\lambda(x)$, where $\lambda(x)$ is the unknown but fixed distribution of the predictors. Under this assumption, one has that

$$\begin{aligned}\mathrm{BF}_n^* &= \frac{\int_{H_0} \prod_{i=1}^n m(u_i, x_i)\pi(dm)}{\int_{H_1} \prod_{i=1}^n m(u_i, x_i)\pi(dm)} = \frac{\prod_{i=1}^n \phi(u_i)\lambda(x_i)}{\int_{H_1} \prod_{i=1}^n f_{x_i}(u_i)\lambda(x_i)\pi(d\mathcal{F})} \\ &= \frac{\prod_{i=1}^n \phi(u_i)}{\int_{H_1} \prod_{i=1}^n f_{x_i}(u_i)\pi(d\mathcal{F})}.\end{aligned}$$

Therefore, the Bayes factor for H_0 versus H_1 remains the same as in (8), i.e., BF_n^* and BF_n have the same limits.

Since $\pi_{H_0} \in (0, 1)$, and by Theorem 1 in Dass & Lee (2004), we have that $\mathrm{BF}_n^* \to \infty$ as $n \to \infty$ when $\mathcal{F}_0$ is the true conditional data-generating mechanism. The elements in $\mathcal{P}(\mathbb{R})^{\mathbb{X}}$ satisfying conditions A1-A5 in Pati et al. (2013) characterize the Kullback–Leibler support of $\mathcal{G}$. Putting together the assumptions on α and (κ_h, σ_h), Theorem 3 in Dass & Lee (2004), and Theorem 6.1 in Pati et al. (2013), we have that $\mathrm{BF}_n^* \to 0$ as $n \to \infty$ when $\mathcal{F}_0$ is not the true conditional data-generating mechanism and satisfies conditions A1-A5. □

Supplementary Material

The supplementary material contains a description of the MCMC algorithm used in Section 3. The R code with the implementation of this algorithm is also available online at `https://git.io/fj15t`.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat.*, (pp. 171–178).

Barcella, W., De Iorio, M., Baio, G., & Malone-Lee, J. (2016). Variable selection in covariate dependent random partition models: an application to urinary tract infection. *Stat. Med.*, *35*, 1373–1389.

Barrientos, A. F., Jara, A., & Quintana, F. A. (2012). On the support of MacEachern's dependent Drichlet processes and extensions. *Bayesian Anal.*, *7*, 277– 310.

Barrientos, A. F., Jara, A., & Quintana, F. A. (2017). Fully nonparametric regression for bounded data using dependent Bernstein polynomials. *J. Amer. Statist. Assoc.*, *112*, 806–825.

Basu, S., & Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Amer. Statist. Assoc.*, *98*, 224–235.

Berger, J. O., & Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *J. Amer. Statist. Assoc.*, *96*, 174–184.

Bogdan, M., Ghosh, J. K., & Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen* (pp. 211–230). Inst. Math. Statist., Beachwood, OH volume 1 of *Inst. Math. Stat. Collect.*. URL: http://dx.doi.org/10.1214/193940307000000158. doi:10.1214/193940307000000158.

Brockwell, A. (2007). Universal residuals: A multivariate transformation. *Statist. Probabil. Letters*, *77*, 1473–1478.

Brockwell, A. (2011). Acknowledgement of priority to: "universal residuals: A multivariate transformation. Stat. Probabil. Letters 27, 2007, 1473–1478". *Stat. Probabil. Letters*, *81*, 1822.

Canale, A., Lijoi, A., Nipoti, B., & Prünster, I. (2017). On the Pitman–Yor process with spike and slab base measure. *Biometrika*, *104*, 681–697.

Carota, C., & Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian statistics* (pp. 507–511).

Chitty, L. S., Campbell, S., & Altman, D. G. (1993). Measurement of the fetal mandiblefeasibility and construction of a centile chart. *Prenatal diagnosis*, *13*, 749–756.

Chung, Y., & Dunson, D. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.*, *104*, 1646–1660.

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of statistical software*, *34*, 1–24.

Dass, S. C., & Lee, J. (2004). A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *J. Statist. Plann. Inference*, *119*, 143–152.

De Iorio, M., Müller, P., Rosner, G. L., & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.*, *99*, 205–215.

Do, K.-A., Müller, P., & Tang, F. (2005). A Bayesian mixture model for differential gene expression. *J. Roy. Statist. Soc. Ser. C*, *54*, 627–644. URL: http://dx.doi.org/10.1111/j.1467-9876.2005.05593.x. doi:10.1111/j.1467-9876.2005.05593.x.

Dunson, D. B., Herring, A. H., & Engel, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *J. Amer. Statist. Assoc.*, *103*, 534–546. URL: http://dx.doi.org/10.1198/016214507000000554. doi:10.1198/016214507000000554.

Dunson, D. B., & Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, *95*, 307–323.

Eubank, R. L., & Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Statist. Assoc.*, *85*, 387–392.

Fan, J., & Huang, L.-S. (2001). Goodness-of-fit tests for parametric regression models. *J. Amer. Statist. Assoc.*, *96*, 640–652.

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

Guindani, M., Müller, P., & Zhang, S. (2009). A Bayesian discovery procedure. *J. Roy. Statist. Soc. Ser. B*, *71*, 905–925. URL: http://dx.doi.org/10.1111/j.1467-9868.2009.00714.x. doi:10.1111/j.1467-9868.2009.00714.x.

Gutiérrez, L., Barrientos, A. F., González, J., Taylor-Rodríguez, D. et al. (2018). A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control. *Bayesian Analysis*, .

Hidalgo, S. J. T., Wu, M. C., Engel, S. M., & Kosorok, M. R. (2018). Goodness-of-fit test for nonparametric regression models: Smoothing spline ANOVA models as example. *Comput. Statist. Data Anal.*, *122*, 135–155.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.*, *96*, 161–173.

Jara, A., Lesaffre, E., De Iorio, M., & Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Annals of Applied Statistics*, *4*, 2126–2149.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, *6*, 255–259.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, *90*, 773–795.

Kim, S., Dahl, D. B., & Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Anal.*, *4*, 707–732.

Lindsey, J. K. (1997). *Applying generalized linear models*. Springer Science & Business Media.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Annals of Statistics*, *12*, 351–357.

Lu, P. (2012). *Calibrated Bayes factors for model selection and model averaging*. PhD dissertation The Ohio State University.

MacEachern, S. N. (2000). *Dependent Dirichlet processes*. Technical Report Department of Statistics, The Ohio State University.

McVinish, R., Rousseau, J., & Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scand. J. Stat.*, *36*, 337–354.

Miller, F. R., & Neill, J. W. (2016). Lack of fit tests for linear regression models with many predictor variables using minimal weighted maximal matchings. *J. Mult. Anal.*, *150*, 14–26.

Norets, A., & Pelenis, J. (2014). Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory*, *30*, 606–646.

Pati, D., Dunson, D. B., & Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Mult. Analysis*, *116*, 456–472.

Peña, E. A., & Slate, E. H. (2006). Global validation of linear model assumptions. *J. Amer. Statist. Assoc.*, *101*, 341–354.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *J. Roy. Statist. Soc. Ser. B*, (pp. 350–371).

Robert, C. P., & Rousseau, J. (2002). *A Mixture Approach to Bayesian Goodness of Fit*. Technical Report Université Paris Dauphine.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, *23*, 470–472.

Sethuraman, J. (1994). A constructive definition of Dirichlet prior. *Statist. Sinica*, *2*, 639–650.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, *11*, 54.

Tokdar, S. T., Chakrabarti, A., & Ghosh, J. K. (2010). Bayesian nonparametrics and semi-parametrics. In Chen, Müller, P. and Sun, D. and Ye, K. and Dey, D. K. (Ed.), *Frontiers of Statistical Decision Making and Bayesian Analysis* (pp. 185–217). Springer New York.

Tokdar, S. T., & Martin, R. (2013). Bayesian test of normality versus a dirichlet process mixture alternative. *arXiv, 1108.2883*, .

Verdinelli, I., & Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, *26*, 1215–1241.

Yang, M. (2012). Bayesian variable selection for logistic mixed model with nonparametric random effects. *Comput. Statist. Data Anal.*, *56*, 2663–2674. URL: `http://dx.doi.org/10.1016/j.csda.2011.12.014`. doi:`10.1016/j.csda.2011.12.014`.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with gprior distributions. In P. Goel, & A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (pp. 233–243).