A Bayesian goodness-of-fit test for regression

Andrés F. Barrientos* and Antonio Canale†

February 5, 2018

Abstract

Regression models are one of the most widely used statistical procedures and the validation of their assumptions plays a crucial role during the data analysis process. Unfortunately, this validation usually relies on the availability of tests tailored to the specific model of interest. In this work, we present a novel Bayesian approach to perform hypothesis testing of goodness-of-fit for a broad class of regression models whose response variable is univariate and continuous. We base our proposal on a suitable transformation of the response variable and a Bayesian prior induced by a predictor-dependent mixture model. We perform hypothesis testing via the Bayes factor and discuss its asymptotic properties. The proposed method is implemented by means of a Markov Chain Monte Carlo algorithm. Simulated and real datasets, illustrate the performance of the proposed approach.

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

*Andrés F. Barrientos is Postdoctoral Associate, Department of Statistical Science, Duke University, Durham, NC, USA (email: andres.barrientos@duke.edu);
†Antonio Canale is Assistant Professor of Statistics, Department of Statistical Sciences, University of Padua, Padua, Italy.
1 Introduction

Regression models are one of the most widely used statistical procedures. Despite the relatively wide range of nonparametric alternatives available, parametric regression models are the preferred modeling choice in many applications. This is due to their ease of interpretability and estimation. However, parametric models rely on specific assumptions which are most of the times unrealistic. Validating these assumptions should be a must since misspecified models can lead to mistaken conclusions and decisions. However, the ability to check assumptions usually depends on the availability of tests specific to the model. Examples includes the Shapiro-Whilk test (Shapiro and Wilk [1965]) and the Jarque-Bera test (Jarque and Bera, [1980]) for normality or the Reset test of Ramsey ([1969]) for functional form of the regression function.

A Bayesian nonparametric goodness-of-fit test might be a more practical tool for validating assumptions both for its generality—i.e., applicability to a wide variety of regression models—and interpretability—exploiting the Bayesian interpretation of the posterior probability of the null hypothesis. Consistent with this goals, this paper presents a novel Bayesian nonparametric procedure applicable to the 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 predictor-dependent mixture model.

The Bayesian nonparametric literature proposing general goodness-of-fit tests applicable on a wide set of situations is small and few contributions have been made in recent years. See Tokdar et al. (2010) for a review. The majority of contributions, however, do not consider the regression framework and rather focus on proposing tests for predictor-independent densities. Some of these contributions exploit log Gaussian process (Verdinelli and Wasserman, 1998), Polya tree processes (Berger and Guglielmi, 2001), and mixtures of densities (Carota and Parmigiani, 1996; Robert and Rousseau, 2002; Basu and Chib, 2003; McVinish et al., 2009; Tokdar and Martin, 2013). Other related contributions deal with right censored data by using beta-Stacy
processes (Al Labadi and Zarepour, 2013), design tests based on distances between a Dirichlet process and its base measure (Al Labadi and Zarepour, 2014), and provide theoretical results for the Bayes factor’s consistency (Dass and Lee, 2004).

Bayesian goodness-of-fit tests for regression models are scarce. The only works we are aware of are that by Basu and Chib (2003) and Lu (2012) who make model comparisons via Bayes factor and a calibrated version of it, respectively. Basu and Chib (2003) introduce an algorithm to approximate the marginal likelihood of Dirichlet process mixture models. Lu (2012) proposes a method to approximate a calibrated version of the Bayes factor between a parametric model and a Dirichlet process mixture model. Unfortunately, both of these proposals are difficult to implement when the goal is to test the fit of several regression models. This is because users will require either deriving or approximating the marginal likelihood for each of the considered regression models. Moreover, Basu and Chib (2003) and Lu (2012) only provide illustrations of their algorithms using mixture of normals. Implementing their algorithms for other kernels can be a tedious task that requires experience programming algorithms for Bayesian nonparametric models. Our proposal aims to alleviate these user-related issues. Although we use a different strategy motivated by ease of implementation in routine applications, using Basu and Chib’s or Lu’s approaches would be equally appropriate from a methodological standpoint.

While Bayesian tests for goodness-of-fit are scarce, the frequentist literature dealing with this type of tests is quite large (see, e.g., Miller and 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). A failure to detect differences between the two models suggests the model of interest fits well the data. In the second case, goodness-of-fit tests are based on the analysis of residuals commonly defined as the response minus an estimate of the corresponding conditional mean (Eubank and Spiegelman, 1990; Fan and Huang, 2001). These tests require a full characterization of the residuals distribution which is usually available when the response is normally distributed. If this is not the case but the regression model belongs to the generalized
linear model class (McCulloch and Nelder, 1989), there are several types of residuals, such as the Pearsons residuals or the Anscombes residuals, whose distributions are characterized using large-sample arguments or distributional approximations.

Brockwell (2007) takes a different and totally general approach to define the residuals of a regression model. Brockwell’s universal residuals take values on $(0, 1)$ and, under a correct model specification, are uniformly distributed. Universal residuals correspond to the transformation proposed by Rosenblatt (1952) to extend the Kolmogorov-Smirnov test in different contexts. Those extensions not only include goodness-of-fit tests for regression (Brockwell, 2007) but also for multivariate distributions (Justel et al., 1997), time series (Kim et al., 1998), Markov random fields (Kaiser et al., 2012), copulas (Hofert and Mächler, 2014), tests for conditional independence (Song, 2009), and many others.

Universal residuals are a potential alternative to define Bayesian nonparametric goodness-of-fit tests. The proposals of Verdinelli and Wasserman (1998) and Robert and Rousseau (2002) are indeed in this direction. In predictor-independent density estimation, the Rosenblatt’s transformation corresponds to the cumulative distribution function of the estimated model. In this article, instead, we use universal residuals to define a goodness-of-fit test for regression models. One of our contributions relates to the characterization provided by Brockwell (2007). We find that, under the correct specification, universal residuals are not only uniformly distributed but also independent of the predictors. Hence, tests based on universal residuals need to account for these two assumptions. We propose to use a Bayesian nonparametric approach to model universals residual conditional on predictors and, thus, to look for deviations from the uniformity and independence assumption, jointly. These deviations are assessed 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 and Park, 2008; Chung and Dunson, 2009; Jara et al., 2010; Barrientos et al., 2017). This class of mixture models satisfies appealing properties in terms of flexibility (Barrientos et al., 2012) and large sample behavior (Pati et al., 2013; Norets and Pelenis, 2014).
This paper is organized as follows. In Section 2, we describe the Rosenblatt’s transformation and its relationship with universal residuals. Then, we introduce the proposed Bayesian nonparametric goodness-of-fit test and discuss its consistency properties. Illustrations of our proposal considering simulated and real datasets are provided in Section 3 and 4, respectively. Section 5 summarizes our findings and provides some direction for future work.

2 Goodness-of-fit test for regression

Let \( \{(y_i, x_i)\}_{i=1}^n \) be a collection of independent regression data where \( y_i \in \mathbb{Y} \subseteq \mathbb{R} \) is the response (or dependent) variable and \( x_i \in \mathbb{X} \subseteq \mathbb{R}^p \) is a vector of \( p \) predictors. For each \( i = 1, \ldots, n \), let \( \mathcal{F} = \{F_x(\cdot) : x \in \mathbb{X}\} \) be the unknown data generating mechanism, where \( F_x(\cdot) \) denotes the cumulative distribution function of the response variable given the predictors \( x \), i.e., \( y_i|x_i \sim F_{x_i} \). 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 or not \( \mathcal{F}_0 \) provides a reasonable approximation of the true data generating mechanism. Specifically, we want to test the hypotheses

\[
    H_0 : \mathcal{F} = \mathcal{F}_0, \quad H_1 : \mathcal{F} \in \{\mathcal{F}_0\}^c,
\]

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 data generating models of the form \( \{\tilde{F}_x(\cdot) : x \in \mathbb{X}\} \). We aim to propose a Bayesian nonparametric procedure that allows to control the prior probability on the null hypothesis \( H_0 \) and compute its posterior probability easily. Thus, we can perform Bayesian hypothesis testing via the Bayes factor,

\[
    BF_n = \frac{\pi(H_0 | \{(y_i, x_i)\}_{i=1}^n)}{\pi(H_1 | \{(y_i, x_i)\}_{i=1}^n)} \times \frac{\pi(H_1)}{\pi(H_0)}
\]

where \( \pi \) is a prior distribution on \( \mathcal{F}^\mathbb{X} \) and \( \pi(\cdot | \{(y_i, x_i)\}_{i=1}^n) \) is its corresponding posterior distribution. In the remaining of this section, we first describe in detail the concept of universal residuals and provide a new result in the context of regression. Then, we introduce the proposed model and prior specification used to test goodness-of-fit based on universal residuals and study...
the asymptotic properties of the proposed solution.

2.1 Rosenblatt’s transformation and universal residuals

Let $Z = (Z_1, \ldots, Z_k)$ be a continuous random vector with absolutely continuous distributions $G$. Rosenblatt (1952) proposes a transformation for $Z$ under a $k$-dimensional continuous distribution $G$ by letting $R_G(z_1, \ldots, z_k) = (r_1, \ldots, r_k)$ where

$$
    r_1 = \text{pr}_G(Z_1 \leq z_1),
$$

$$
    r_2 = \text{pr}_G(Z_2 \leq z_2|Z_1 = z_1),
$$

$$
    \vdots
$$

$$
    r_n = \text{pr}_G(Z_k \leq z_k|Z_1 = z_1, \ldots, Z_{n-1} = z_{k-1}),
$$

and $\text{pr}_G$ is the probability under $G$. We refer to $R_G$ as Rosenblatt’s transformation under the distribution $G$. Rosenblatt (1952) shows that, if $Z \sim G$, $R_G(Z)$ is a random vector uniformly distributed on $[0, 1]^k$, and uses this result to extend the Kolmogorov-Smirnov test to a multivariate setting. Given a multivariate distribution $G_0$, one can check whether $G_0$ is the distribution of $Z$ by testing whether $R_{G_0}(Z)$ is uniformly distributed on $[0, 1]^k$.

We can also define Rosenblatt’s transformation starting from a distribution function $G_X$ that depends from some $p$-dimensional random vector $X$. For example $G_x(\cdot)$ can be the conditional distribution of $Z$ given $X = x$. If $G_x(\cdot)$ is continuous for all $x$, then $R_{G_X}(Z)|X = x$ is also uniformly distributed on $[0, 1]^k$. The next proposition formalizes this statement.

**Proposition 1.** Let $Z$ be a continuous $k$-dimensional random vector and $X$ be a $p$-dimensional random vector. Let $G_x$ be the conditional distribution of $Z$ given $X = x$. Then $R_{G_X}(Z)$ is uniformly distributed on $[0, 1]^k$ and independent of $X$.

This proposition paves the way to define a method to perform goodness-of-fit testing for conditional distributions. Specifically, to see if a given a conditional distribution $G_{0,x}$ is exactly
the distribution of $Z$ given $X = x$, it is equivalent to check whether $R_{G_0,x}(Z)$ is uniformly distributed on $[0, 1]^k$ and independent of $X$. Brockwell (2007) uses part of Proposition 1 to define a goodness-of-fit test for univariate regression models. For a given dataset $\{(y_i, x_i)\}_{i=1}^n$ and model of interest $F_0$, Brockwell first defines the universal residuals $u_i$ as the Rosenblatt's transformation of $y_i$ given $x_i$ under $F_0$, that is, $u_i = R_{F_0, x_i}(y_i)$. Since the response is univariate, $R_{F_0, x_i}(y_i) = F_0, x_i(y_i)$. Under this proposal, the goodness-of-fit of $F_0$ is equivalent to test the uniformity of the universal residuals. Brockwell (2007) only focuses on testing uniformity and does not provide any insight about the independence between universal residuals and predictors.

The independence condition, however, is very important to verify in all applications as this allows to check whether all the information on the response contained in the predictors has been fully incorporated in the model. Our claim is that to verify the goodness-of-fit of $F_0$ based on universal residuals it is required to test both uniformity and independence, jointly. To stress this conjecture, consider the two following situations.

First, assume that $F_{0,x}$ is the true conditional cumulative distribution function of $Y$ given $X = x$, where $X$ is also a random variable. Let $Q$ be a bivariate distribution function defined on $\mathbb{Y} \times \mathbb{X}$ and define $Q_1(y) = \text{pr}_Q(Y \leq y)$ and $Q_2(y|x) = \text{pr}_Q(Y \leq y|X = x)$. Under this specification, and consistently with Brockwell's approach, if we want to 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, it is sufficient to test the uniformity of $R_{F^*_X}(Y)$. Unfortunately, $R_{F^*_X}(Y)$ is no-doubt uniformly distributed but not necessarily independent from $X$. It is clear that under Brockwell's strategy, the type II error related to $H_0 : \mathcal{F} = \mathcal{F}^*$ can remain the same regardless the sample size. This is a very unappealing property for a hypothesis test.

In addition, there are cases where the universal residuals are independent but not uniformly distributed. For example, take two univariate cumulative distribution function $Q_1$ and $Q_2$ defined on $\mathbb{Y}$. If $\mathcal{F}^* = \{F^*_x(\cdot) : x \in \mathbb{X}\}$ with $F^*_x(y) = Q_1(Q_2^{-1}(F_{0,x}(y)))$, then $R_{F^*_X}(Y)$ is independent of $X$ but not necessarily uniformly distributed.

This discussion lead us to question whether there exists another family of conditional distri-
butions $\mathcal{F}_0^* = \{ F_{0,x}^*(\cdot) : x \in \mathbb{X} \} \in \mathcal{F}^X$, $\mathcal{F}_0^* \neq \mathcal{F}_0$, such that $R_{x_0,0}^* (Y) \text{ is uniformly distributed and independent of } X$. In other words, we question the uniqueness of $\mathcal{F}_0$ defining universal residuals through $R_{F_0,0} (\cdot)$ with uniform distribution and independent of the predictors. The following theorem addresses this issue. Its proof is reported in the Appendix.

**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 $U = R_{F_0,X} (Y) = F_{0,X} (Y)$ and $U^* = R_{F_0,X}^* (Y) = F_{0,X}^* (Y)$ are uniformly distributed and are independent of $X$. Then $F_{0,X} = F_{0,X}^*$ almost surely $\mu_X$.

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

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

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

$$
P [X \in A, U \leq u] = P [X \in A] P[U \leq u] = u \mu_X (A) = \lambda (A),
$$

On the other hand, one has that

$$
P [X \in A, U \leq u] = P [X \in A, Y \leq F_{0,X}^{-1} (u)] = E \left[ \mathbb{I}_{\{X \in A\}} \mathbb{I}_{\{U^* \leq F_{0,X}^* (F_{0,X}^{-1} (u))\}} \right],
$$

$$
= E \left[ E \left[ \mathbb{I}_{\{X \in A\}} \mathbb{I}_{\{U^* \leq F_{0,X}^* (F_{0,X}^{-1} (u))\}} \right] \right],
$$

$$
= E \left[ \mathbb{I}_{\{X \in A\}} F_{0,X}^* (F_{0,X}^{-1} (u)) \right],
$$

$$
= \int_A F_{0,X}^* (F_{0,X}^{-1} (u)) d\mu_X.
$$

Hence,

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

8
which implies that $F_{0,X}(F_{0,X}^*(u)) = u$ almost surely $\mu_X$, i.e., $F_{0,X}(u) = F_{0,X}^*(u)$ almost surely $\mu_X$. □

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 $\mathcal{X}_0 \subseteq \mathcal{X}$ such that $\mu_X(\mathcal{X}_0) = 1$, then we have to assume that $\mathcal{F}_0^*$ and $\mathcal{F}_0$ should be indistinguishable. This assumption is met if we in turn assume that

i) for every $\mathcal{F} \in \mathcal{F}^\mathcal{X}$, the function $(y,x) \mapsto f_x(y)$ is continuous, where $f_x$ denotes the density function associated with $F_x$;

ii) $\mathcal{X}$ is the topological support of the unknown measure $\mu_X$, that is, $\mathcal{X}$ is the smallest closed set such that $\mu_X(\mathcal{X}) = 1$.

Assumption i) implies that if $\mathcal{F}_0^*, \mathcal{F}_0 \in \mathcal{F}^\mathcal{X}$ with $\mathcal{F}_0^* \neq \mathcal{F}_0$, then there exists an open set $\mathcal{O} \subseteq \mathcal{X}$ such that, for every $x \in \mathcal{O}$, $F_{0,x} \neq F_{0,x}^*$. Assumption ii) implies that $\mu_X(\mathcal{O}) > 0$. Thus, if $\mathcal{F}_0$ is the true model, then $\mathcal{F}_0$ is the unique element in $\mathcal{F}^\mathcal{X}$ such that the Rosenblatt’s transformation $R_{F_0}(\cdot)$ leads to uniformly distributed and predictor-independent random variables.

**Remark 1.** We can analogously define universal residuals when the predictors correspond to a fixed design. They are defined using the distribution function of the response conditioned on the values of the design. In that case, instead of testing independence between the residuals and the predictor, we test that the conditional distribution of the residuals is the same at any value of the design. Although there is a conceptual difference between random and fixed designs, there will not be any operational difference testing goodness-of-fit with the test we are proposing in this paper.

We use universal residuals to re-write the hypotheses outlined in (1). Assuming the $(y_i, x_i)$ for $i = 1, \ldots, n$ are realizations of the random elements $(Y_i, X_i)$, we have defined universal residuals $U_i$ as $F_{0,X_i}(Y_i)$. Thus, we can re-write the hypotheses in (1) as,

$$H_0 : U_i \sim U(0,1) \text{ and } U_i \perp X_i, \quad H_1 : U_i \not\sim U(0,1) \text{ or } U_i \not\perp X_i,$$
where $U_i \perp X_i$ denotes that $U_i$ and $X_i$ are independent. Notice that we are not forced to work with universal residuals taking values at $(0, 1)$. For example, for practical reasons we may map them to the real line using the inverse of the standard Gaussian distribution function, say $\Phi^{-1}$.

Consistent with this choice, and with an abuse of notation we can re-define universal residuals

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

and optionally re-write the hypotheses in (1) as,

$$H_0 : U_i \sim N(0, 1) \text{ and } U_i \perp X_i, \quad H_1 : U_i \not\sim N(0, 1) \text{ or } U_i \not\perp X_i. \quad (4)$$

Hereafter, we will use the universal residuals defined in (3) and focus on testing the hypotheses prompted in (4).

### 2.2 Bayesian nonparametric testing

Let $\{(u_i, x_i)\}_{i=1}^n$ be the new dataset, where the $u_i$ are obtained via (3). Consistent with the discussion of Section 2.1, to test the null hypothesis in (1) it is sufficient to model the conditional density of $u_i \mid x_i$ and test if this is a standard Gaussian distribution. To this end we are going to define a flexible Bayesian approach to estimate the conditional density of $u_i$ given $x_i$ which gives positive prior probability to the standard Gaussian model. We will first focus on the definition of a prior probability $\pi$ with large support on $\mathcal{P}(\mathbb{R})^X$—the space of all predictor-dependent probability measures defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with continuous probability density function, where $\mathcal{B}(\mathbb{R})$ denotes the Borel $\sigma$-algebra—and then modify it in order to permit to place positive prior probability on the standard Gaussian model.

To this end let the prior $\pi$ be the probability distribution induced by a stochastic process of the form $\mathcal{F} = \{F_x : x \in \mathbb{R}\}$, where $f_x(\cdot) \equiv \text{conditional density function associated with the probability function } F_x(\cdot) \equiv \text{defined by the following mixture specification}$

$$f_x(u) = \sum_{h=1}^{\infty} w_h \frac{1}{\sigma_h} \phi \left( \frac{u - \kappa_h(x)}{\sigma_h} \right), \quad (5)$$
where $w_h$ are random weights summing to one and $\theta_h = (\kappa_h, \sigma_h) \sim P_0$ independently, with $P_0$ a base probability measure on $K \times \mathbb{R}^+$, where $K$ is the space of all the $\mathbb{X} \rightarrow \mathbb{R}$ functions. A typical assumption is that the base measure $P_0$ is made of independent marginals $\pi_\kappa$ and $\pi_\sigma$. The random weights, are defined by means of a stick-breaking process (Sethuraman, 1994; Ishwaran and James, 2001), i.e., $w_h = V_h \prod_{l<h} (1 - V_l)$, with $V_h \sim \text{Be}(\alpha_h, \beta_h)$. For a given $x \in \mathbb{X}$, and fixing $\alpha_h = 1$ and $\beta_h = \beta$ for all $h$, $f_x$ is a Dirichlet process mixture model (Lo, 1984; Escobar and West, 1995). 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 and Park, 2008; Chung and 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).

This standard nonparametric proposal, however, does not put positive mass on the null hypothesis which is represented by the atomic element $\mathcal{P}_0 = \{ \Phi(\cdot) : x \in \mathbb{X} \} \in \mathcal{P}(\mathbb{R}^\mathbb{X})$. This fact leads to the calculation of the Bayes factor not possible.

To perform our goodness-of-fit procedure, then, we need to modify the above construction by allowing the prior distribution $\pi$ to place positive mass at $H_0$. With the above construction, this is achieved when the parameters $\theta_h$ are equal to $\tilde{\theta} = (0, 1)$ for all $h$. We allow this introducing a binary variable $\nu$ which is equal to one with positive prior probability $\pi_{H_0}$ and assuming the following hierarchical structure for the component specific parameter $(\kappa_h, \sigma_h)$ of model (5)

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

where $\delta_a$ is the Dirach’s delta function. Under this formulation $\theta_h = (\kappa_h, \sigma_h)$ can be either all equal to the specific value $\tilde{\theta} = (0, 1)$ or all different —and not equal to $\tilde{\theta}$. The idea of assuming prior distribution assigning point masses at zero for some —or all— the component specific parameters arises in many other contexts such as variable selection (Dunson et al., 2008; Yang, 2012; Barcella et al., 2016), 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 \( \pi \) on \( \mathcal{P}(\mathbb{R})^X \) is then identified by \( \pi_{H_0}, P_0 \), and the stick-breaking process’ sequence of parameters \((\alpha_h, \beta_h)\). Under this formulation we can express the Bayes factor as

\[
BF_n = \frac{\prod_{i=1}^{n} \phi(u_i)}{\int_{H_1} \prod_{i=1}^{n} f_{x_i}(u_i) \pi(d\mathcal{F})} = \frac{\pi(\nu = 1|\{(u_i, x_i)\}_{i=1}^{n})}{\pi(\nu = 0|\{(u_i, x_i)\}_{i=1}^{n})} \times \frac{\pi(\nu = 0)}{\pi(\nu = 1)},
\]

which turns out to be computationally manageable.

A desirable specification of \( \pi \) implies that \( BF_n \) is consistent as the sample size increases. The following theorem provides sufficient conditions on \( \pi_{H_0}, P_0 = \pi_{\sigma} \times \pi_{\kappa} \) and \((\alpha_h, \pi_{\sigma}, \pi_{\kappa})\) defined as in Theorem 6.1 in Pati et al. (2013). Then \( BF_n \to \infty \) as \( n \to \infty \) under \( H_0 \). Moreover, under \( H_1 \), \( BF_n \to 0 \) as \( n \to \infty \) if the proposed data-generating mechanism belongs to the class of elements of \( \mathcal{P}(\mathbb{R})^X \setminus \mathcal{P}_0 \) characterized by conditions A1-A5 in Pati et al. (2013).

**Theorem 2.** Let \( V_h \sim \text{Be}(1, \alpha_h) \), \( \pi_{H_0} \in (0, 1) \), and \((\alpha_h, \pi_{\sigma}, \pi_{\kappa})\) defined as in Theorem 6.1 in Pati et al. (2013). Then \( BF_n \to \infty \) as \( n \to \infty \) under \( H_0 \). Moreover, under \( H_1 \), \( BF_n \to 0 \) as \( n \to \infty \) if the proposed data-generating mechanism belongs to the class of elements of \( \mathcal{P}(\mathbb{R})^X \setminus \mathcal{P}_0 \) characterized by conditions A1-A5 in Pati et al. (2013).

**Proof.** We prove consistency using Theorem 1 and 3 in Dass and Lee (2004). Dass and Lee (2004) provides results in the single-density context; however, our proposed test is framed within a conditional/regression context. For this reason, we apply Dass and 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

\[
BF_n^* := \frac{\int_{H_0} \prod_{i=1}^{n} m(u_i, x_i) \pi(\mathcal{F})}{\int_{H_1} \prod_{i=1}^{n} m(u_i, x_i) \pi(\mathcal{F})} = \frac{\prod_{i=1}^{n} \phi(u_i)\lambda(x_i)}{\int_{H_1} \prod_{i=1}^{n} f_{x_i}(u_i)\lambda(x_i) \pi(d\mathcal{F})} = \frac{\prod_{i=1}^{n} \phi(u_i)}{\int_{H_1} \prod_{i=1}^{n} f_{x_i}(u_i) \pi(d\mathcal{F})}.
\]

Therefore, the Bayes factor for \( H_0 \) versus \( H_1 \) remains the same as in (6), i.e., \( BF_n^* \) and \( BF_n \) have the same limits.

Since \( \pi_{H_0} \in (0, 1) \), and by Theorem 1 in Dass and Lee (2004), we have that \( BF_n^* \to \infty \) as \( n \to \infty \) when \( \mathcal{F}_0 \) is the true conditional data-generating mechanism. The elements in \( \mathcal{P}(\mathbb{R})^X \) satisfying conditions A1-A5 in Pati et al. (2013) characterize the Kullback–Leibler support of \( \mathcal{G} \). Putting together the assumptions on \( \alpha \) and \((\kappa_h, \sigma_h)\), Theorem 3 in Dass and Lee (2004), and Theorem 6.1 in Pati et al. (2013), we have that \( BF_n^* \to 0 \) as \( n \to \infty \) when \( \mathcal{F}_0 \) is not the true conditional data-generating mechanism and satisfy condition A1-A5. \( \square \)
Remark 2. In any regression context, it is almost unrealistic to propose a $F_0$ without using an estimation procedure. In fact we will deal with families of conditional distributions $F_\theta = \{F_{\theta,x}(\cdot; \theta), x \in \mathbb{X}, \theta \in \Theta\}$, which depends from unknown parameter $\theta \in \Theta$. The approach described so far, however, assumes that the family of conditional distributions that we want to test is fully specified. For this reason, if our interest is to test if $F_{0,\theta} = \{F_{0,\theta,x}(\cdot; \theta), x \in \mathbb{X}, \theta \in \Theta\}$, is the true data generating family of distribution, we will apply our procedure to $F_0 = \{F_{0,\hat{\theta},x}(\cdot; \hat{\theta}), x \in \mathbb{X}\}$, where $\hat{\theta}$ is a consistent estimator of the model parameters. Our conjecture is that if $F_0$ is specified with a reasonable and consistent estimator of $\theta$ (e.g. using a maximum likelihood estimator), the Bayes factor is still consistent. Section 3 provides empirical evidence in favour of this conjecture. A more formal and fully Bayesian approach would put a prior on $\theta$ and consider the estimation of the posterior distribution of $\theta$ jointly with the estimation of the posterior for model (5) following the path of Verdinelli and Wasserman (1998). Unfortunately, a fully Bayesian approach would require specific Markov Chain Monte Carlo implementations depending on the parametric structure of $F_{0,\theta}$. Hence, such approach would limit the applicability of our proposal.

3 Simulation study

We now proceed to illustrate our approach by means of detailed simulation studies. In what follows we simulate different synthetic datasets and test different model assumptions. The latters give a quantification of our proposed method’s performance in a broad variety of situations.

Before going into the details of each situation, we first describe the specific prior specification that we use henceforth. The prior $\pi$ on $P(\mathbb{R}^\mathbb{X})$ is induced through a simplified version of model (5). Specifically, we use a truncated mixture model of the form

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

where $w_h = V_h \prod_{l<h}(1 - V_l)$, $V_h \sim Be(1, \alpha)$ for $h = 1, \ldots, H - 1$, $V_H = 1$, and $H$ is
a conservative upper bound for the number of mixture components. We assume that, for each observation, we have measures of $p_1$ continuous predictors and $p_2$ binary predictors. We specify $\kappa_h(x)$ with the following additive structure

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

where $x_{i,j}$ for $j = 1, \ldots, p_1$ are the continuous predictors with $\eta_{hj}$ a smooth function modelled with regressions splines with $b_j$ a B-spline basis, $\phi_{hjl}$ a basis coefficient, $x_i^d$ is the vector of dummy variables derived from a suitable map of the $p_2$ categorical predictors, and $\gamma_h$ a vector of coefficients related to the categorical predictors. B-splines expansions induce a linear smoother and 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 all the dummy variables related to the categorical predictors and $\beta_h$ is the related vector of coefficients of suitable dimension, say $p$.

In the following we use the simplest version of a cubic B-spline defined with zero knots which is equivalent to specifying an ordinary cubic polynomial regression. To specify the distribution for the atoms $(\beta_h, \sigma_h)$ we define a latent variable $\nu$ that is equal one if $\beta_1 = \cdots = \beta_H = 0$ and $\sigma_1 = \cdots = \sigma_H = 1$, and is equal to zero otherwise, where $0$ denotes a vector with all components equal to zero of suitable dimension. Different and simpler specifications of $\kappa_h(\cdot)$—not reported here—lead to qualitatively similar conclusions.

Consistently with the discussion of previous section we let

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

with $\sigma_h^{2*} \sim \text{InvGa}(1.5, 0.5)$ and $\beta_h^{*} \sim N_p(0, \sigma_h^2 g_h(\tilde{x}^T \tilde{x})^{-1})$ where $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)^T$ denotes the design matrix and $g_h$ has conditional distributions given $\nu$ equal to

$$g_h|\nu = 1 \sim \text{InvGa}((p+1)/2, n_h/2), \quad g_h|\nu = 0 \sim \text{InvGa}(1/2, n_h/2),$$
where \( n_h = E[w_h]n \) and \( E[w_h] \) denotes the prior mean of \( w_h \). The prior for \( \beta^*_h \) corresponds to the \( g \)-prior commonly used in model selection procedures (Zellner, 1986). The prior specification is completed by assuming \( \nu \sim \text{Be}(\pi_{H_0}), \pi_{H_0} = 0.5, \) and \( \alpha \sim \text{Ga}(0.25, 0.25) \). The truncation level \( H \) is fixed at 15. Under this specification, we can think of model (7) as a practical approximation of model (5). Moderate changes of the hyperparameters lead to qualitatively similar results. Our experience suggests that this specific choice is stable and leads to consistent conclusions across different scenarios.

The Bayes factor is computed using a posterior sample of size 1,000. We draw the posterior sample using the Markov Chain Monte Carlo (MCMC) algorithm described in the Supplemental Material with a burning period of 5,000 iterations and thinning of 10.

For each simulation scenario, we consider three sample sizes, namely \( n = 100, 250, \) and \( 500 \), and simulate the predictors using the same mechanism. Specifically, we consider 10 predictors, five of them are continuous and five discrete. Specifically we let \( x_i = (1, x_{i,1}, \ldots, x_{i,10})^T \), where \( x_{i,1} \sim N(0, 1), (x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5}) \sim N(0, \Sigma_x), \Sigma_x = \{\sigma_{j,j'}\}, \text{cov}(x_{i,j}, x_{i,j'}) = 0.7|j-j'|, \) and \( x_{i,j} \sim \text{Bernoulli}(0.5), j = 6, \ldots, 10 \). The procedure is repeated for \( N = 100 \) replicates for each situation. Since we are using the specific MCMC procedure described in the Supplemental Material for a finite number of iterations, it may happen to obtain a posterior probability for the null equal to 0 or 1. For this reason and for graphical purposes, before computing the Bayes factors, in all the following results we add or subtract a random uniform noise between 0 and 0.001 to the posterior probabilities of the null equal to 0 and 1, respectively.

We test our procedure for four different specifications. In the first we simulate the data from the simplest situation, i.e. \( u_i \sim N(0, 1) \) and test the normality assumption in \( H_0 \). A second situation tries to emulate a simple linear regression’s setting with different types of model misspecifications. As third and fourth situations, we assess the performance of the procedure under beta and gamma regressions, respectively.

Call \( \mathcal{M}_{0,0} \) the model that simulates \( u_i \sim N(0, 1) \). In testing \( H_0 : \mathcal{F} = \mathcal{M}_{0,0} \), we would
expect to obtain high Bayes factors giving strong evidence for the null. This is indeed testified by looking at the boxplots of panel (a) of Figure 1 which represent, for each sample size, the distribution of the Bayes factors for $H_0 : \mathcal{F} = \mathcal{M}_{0,0}$ when the data are simulated from $\mathcal{M}_{0,0}$.

Following Kass and Raftery (1995) we interpret the Bayes factors above 1 but below 3 just worth of a bare mention of evidence in favour of $H_0$, those between 3 and 20 as positively in favour of $H_0$, those between 20 and 150 as strongly in favour of $H_0$, and those above 150 as very strongly in favour of $H_0$.

The second situation emulates a simple linear regression’s setting. We simulate data from the four following models

$$
\mathcal{M}_{1,0} = \{ F_{x_i} (\cdot) : y_i \sim N(x_i^T \beta, \sigma^2) \},
$$

$$
\mathcal{M}_{1,1} = \{ F_{x_i} (\cdot) : y_i \sim N(x_i^T \beta + x_i^2, \sigma^2) \},
$$

$$
\mathcal{M}_{1,2} = \{ F_{x_i} (\cdot) : y_i \sim N(x_i^T \beta + 0.4[5 \exp(x_{i,1}) + 4 \exp(x_{i,2}) + \ldots + \exp(x_{i,5})], \sigma^2) \},
$$

$$
\mathcal{M}_{1,3} = \{ F_{x_i} (\cdot) : y_i = x_i^T \beta + \epsilon_i, \ \epsilon_i \sim 0.25N(0, 4) + 0.5N(0, 1) + 0.25N(0, 0.25) \},
$$

fixing $\beta$ to a vector of ones and $\sigma^2 = 1$. In each situation we test the goodness-of-fit of the simplest specification, i.e.

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

where both the regression coefficients $\beta$ and the error variance $\sigma^2$ are estimated via maximum likelihood from the simulated data. The specification under $\mathcal{M}_{1,1}$ represents the traditional example where a predictor needs to be modelled along with its quadratic transformation; $\mathcal{M}_{1,2}$ represents a scenario where the conditional mean is “almost” linear; $\mathcal{M}_{1,3}$ represents a correct specification of the mean function but missspecifies the distribution of errors. The second panel of Figure 1 reports the distributions of the Bayes factors in these different situations.

As third situation we consider a beta regression. The data are simulated according to the
following models

\[ M_{2,0} = \{ F_{x_i}(\cdot) : y_i \sim \text{Beta}(\mu_i, (1 - \mu_i)\gamma), \; \mu_i = (1 + \exp(-x_i^T \beta))^{-1} \}, \]

\[ M_{2,1} = \{ F_{x_i}(\cdot) : y_i \sim \text{Beta}(\mu_i, (1 - \mu_i)\gamma_i), \; \mu_i = (1 + \exp(-m_i))^{-1}, \; \gamma_i = \exp(\tilde{m}_i), \]

\[ m_i = 2 - 0.15x_i^T 1_{p+1}, \; \tilde{m}_i = 1 + x_{i,1} + x_{i,10} \}

where under \( M_{2,0} \), we let \( \beta = (1.85, -0.15, \ldots, -0.15)^T \) and \( \gamma = 1.5 \). In these cases we test the goodness-of-fit of model \( M_{2,0} \), i.e. \( H_0 : \mathcal{F} = M_{2,0} \) estimating its parameters via maximum likelihood from the simulated data. Model \( M_{2,0} \) is then correctly specified while model \( M_{2,1} \) requires adding predictors to the precision parameter. Panel (c) of Figure 1 reports the distributions of the Bayes factors in the different situations.

Finally, we consider a gamma regression. The data are simulated according to the following models

\[ M_{3,0} = \{ F_{x_i}(\cdot) : y_i \sim \text{Gamma}(\mu_i, \gamma_i/\mu_i), \; \mu_i = \exp(x_i^T \beta_1), \; \gamma_i = \exp(x_i^T \beta_2), \; \eta_3 = (\beta_1, \beta_2) \}

\[ M_{3,1} = \{ F_{x_i}(\cdot) : y_i \sim \text{Gamma}(\mu_i, \gamma_i/\mu_i), \; \mu_i = \exp(1 + m_i), \; \gamma_i = \exp(-5 + \tilde{m}_i), \]

\[ m_i = 0.2x_{i,1} + 3x_{i,10} + \exp(1 + x_{i,2}), \; \tilde{m}_i = 0.1x_i^T 1_{p+1} + \exp(0.3(x_{i,2} + 3)) \}

where, under \( M_{3,0} \), we let \( (\beta_{1,1}, \beta_{1,2}, \beta_{1,3}, \ldots, \beta_{1,10}, \beta_{1,11}) = (1, 0.2, 0, \ldots, 0, 3) \) and \( \beta_2 = (-4.9, 0.1, \ldots, 0.1) \). Also here, we test the goodness-of-fit of model \( M_{3,0} \) estimating its parameters via maximum likelihood. Panel (d) of Figure 1 reports the results.

As expected, we observe that, when the model is correctly specified, the Bayes factor increases as \( n \) increases. Viceversa, when is \( H_0 \) is not true, the Bayes factor converges to zero as \( n \) grows. This asymptotic behaviour provides empirical evidence in favour of the Bayes factor’s consistency based on model (7) and aligns with the results of Theorem 2.

Comparing the results of the first situation, where the procedure is applied directly, with that of the remaining situations where the parameters are estimated from the data, we can observe that the Bayes factors associated with the first situations are slightly smaller than those for the remaining ones. This is an expected result because in the remaining situations \( H_0 \) tends to...
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; columns BNP denote the decisions based on rejecting $H_0$ when the Bayes factor is $< 1$; columns Brockwell’s denote the decisions based on rejecting $H_0$ when the $p$-values of the frequentist test of Brockwell (2007) are $< 0.05$

<table>
<thead>
<tr>
<th>True model</th>
<th>$H_0$</th>
<th>$n$</th>
<th>Type I errors</th>
<th>Type II errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{00}$</td>
<td>$M_{00}$</td>
<td>100</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>$M_{10}$</td>
<td>$M_{10}$</td>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_{11}$</td>
<td>$M_{10}$</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$M_{12}$</td>
<td>$M_{10}$</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$M_{13}$</td>
<td>$M_{10}$</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$M_{20}$</td>
<td>$M_{20}$</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$M_{21}$</td>
<td>$M_{20}$</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$M_{30}$</td>
<td>$M_{30}$</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$M_{31}$</td>
<td>$M_{30}$</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>-</td>
<td>18</td>
</tr>
</tbody>
</table>
slightly favour more the data. Hence, if $H_0$ is defined based on estimates of the parameters, we should assume a more conservative position finding evidence in favour of $H_0$.

The comparison of the proposed Bayesian test with frequentist competitor is not straightforward as the Bayes factor is intrinsically different from the idea of $p$-values used in frequentist hypothesis testing. However, from a decision viewpoint, we can decide to reject the null if the Bayes factor is above a given value or the $p$-value is below a given value and assess the frequentist operating characteristics of these rules. To this end, we consider Brockwell’s (2007) proposal, i.e., we only verify whether the universal residuals are distributed according to a standard Gaussian distribution. We check this assumption using a Kolmogorov-Smirnov test and the `ks.test` function available in the R software (R Core Team, 2017). Note that also this procedure involves a first estimation step for all but the first scenarios as in our proposal. We decide to reject $H_0$ if the Bayes factor $< 1$, or the $p$-value is below 0.05 for our procedure and Brockwell’s procedure, respectively.

For the $N = 100$ replicated datasets we compute the type I or type II errors—depending if $H_0$ was true or not—simply counting how many false rejections/acceptances we get over $N$. Table 1 reports the results labelled as BNP for our proposal, and Brockwell’s for the frequentist test of Brockwell (2007).

Type I error become smaller for both as the sample size increases for all approaches with our proposed approach typically reporting a smaller error with respect to the frequentist alternative. For what concerns the type II errors, instead, the performance of the proposed test is comparable or better than the frequentist tests. This outperformance becomes more evident for smaller sample sizes.

The results of 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 check the robustness of these procedure to the threshold choice, Figure 2 reports the ROC curves for the three scenarios having null $\mathcal{M}_{10}$, $\mathcal{M}_{20}$, and $\mathcal{M}_{30}$. To obtain these (smoothed) ROC curves, we pulled together the samples
having the same $H_0$ and labelled accordingly to the fact that $H_0$ was indeed true or not. Then we calculate the values of specificity and sensitivity using the values of posterior probability of $H_0$ and the $p$-values, respectively by means of the \texttt{roc} function of the R package \texttt{pROC} (Robin et al., 2011). These ROC curves are consistent with the results highlighted in Table 1 and show that our procedure is always better or comparable to the Brockwell’s one. This empirical evidence supports the need to check not only normality but also independence.

4 Real data illustrations

In this section we illustrate the use of our proposal by means of two real datasets. The precise model specification, priors, and MCMC parameters are the same of the simulation study of Section 3.

4.1 Mandible length data

As first illustration, we consider a simple dataset related to fetal growth data first presented in Chitty et al. (1993) and available in the R package \texttt{lmtest} (Zeileis and Hothorn, 2002). The dataset contains the ultrasonographic measurements of mandible length in $n = 158$ fetuses along with the gestational age (in weeks) in which the measurement was taken. Clearly, there is a well known positive relationship between mandible length and gestational age that we try to model here. A first linear model with log-length as response and gestational age as predictor with homoskedastic Gaussian errors is fitted to the dataset with maximum likelihood approach. While the logarithmic transformation of the response variable clearly stabilize 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 their relations. After computing the universal residuals obtained fixing the estimated parameters, our goodness-of-fit procedure confirm this conjecture reporting a Bayes factor of 0. A second more realistic
model assumes a quadratic relation between the log-length and the gestational age. Specifically we include in the linear predictor another predictor which is simply the squared gestational age. The related residuals (reported in panel (b) of Figure 3) do not show signs of apparent misspecification, at least at the mean level. Consistently with this, our procedure reports a Bayes factor of 59.6 which can be interpreted as strong evidence in favour of the second model specification according to the Bayes factor’s classification of Kass and Raftery (1995).

4.2 Australian Institute of Sport data

To illustrate the performance of the proposed goodness-of-fit method to detect lack of fit on the scale of the residual distribution, we consider a dataset containing a sample of 202 elite Australian athletes who trained at the Australian Institute of Sport. For each athlete 13 variables are recorded but here we limit the analysis on modelling the (log) plasma ferritin concentration —henceforth \( \log - \text{Fe} \)—in function of the the lean body mass (LBM) index and body mass index (BMI). The data have been made firstly publicly available by the book by Cook and Weisberg (1994) and are included in the R package sn by Azzalini (2017). Simple exploratory data analysis reveal mild linear relations between \( \log - \text{Fe} \) and both BMI and LBM. Hence, as first model, we fit a Gaussian linear regression assuming

\[
\log \text{Fe} = \alpha + \beta_1 \times \text{BMI} + \beta_2 \times \text{LBM} + \epsilon, \quad \epsilon \sim N(0, \sigma^2).
\]

The kernel density estimates of the residuals obtained fitting model (9) via maximum likelihood, is reported in panel (a) of Figure 4. Despite a mild left skewness can be appreciated from the kernel density estimates, the fitted (Gaussian) residuals density (represented with the dashed line) seems to provide a decent fit. This is confirmed by our method which returns a Bayes factor of 54.6 when applying the method to the related universal residuals which implies strong evidence in favour of the Gaussian assumption following Kass and Raftery (1995) classification.

The mild skewness of the residuals of this first regression model, however, motivates further investigations. A possible solution which remains in the context of parametric models, consists
in assuming a skew-normal (Azzalini, 1985) distribution for the error term. Consistent with this we fit a regression model assuming

\[
\log\text{Fe} = \alpha + \beta_1 \times \text{BMI} + \beta_2 \times \text{LBM} + \eta, \quad \eta \sim \text{SN}(0, \omega^2, \alpha),
\]  

(10)

where \(\text{SN}(\xi, \omega^2, \alpha)\) denotes a skew-normal distribution with location \(\xi\), scale \(\omega^2\), and shape \(\alpha\). For a detailed account on this family of distributions see the book by Azzalini and Capitanio (2014). The kernel density estimates of the residuals obtained fitting model (10) via maximum likelihood, is reported in panel (b) of Figure 4. The mild skewness is now estimated and the latter formulations seems to provide a slightly better fit to the data at hand. This is confirmed by our method which returns a Bayes factor of 85.9 which is higher that that obtained fixing the parameters estimated under model (9) thus implying stronger evidence in favour of the skew-normal assumption.

5 Discussion

We presented a novel Bayesian approach to perform hypothesis testing of goodness-of-fit for regression models based on Bayes factor. The proposed procedure can be applied for a large class of regression models, namely those with univariate and continuous response. The proposed solution enjoys desirable asymptotic properties as discussed in Section 2.2. Consistent with this, the finite-sample simulation study performed in Section 3 shows that the proposed test performs well in a variety of scenarios. Despite not being a closed form solution, the proposed solution is easy to implement since it only requires the predictors and the universal residuals as input features. For all these reasons we foreseen it to be routinely implemented in many applications.

Some future research is needed to study the asymptotic properties of the Bayes factor when the null hypothesis is defined based on estimates of the parameters of a given parametric model. In addition, extensions of the proposed procedure to more general regression models, e.g. those with a multivariate response, are subject to ongoing research.
Supplementary Material

The online supplementary material contains a description of the MCMC algorithm used in Sections 3–4. The R code with the implementation of this algorithm is also available online.

Acknowledgments

We would like to thank Pierpaolo De Blasi, David Dunson, Víctor Peña, and Mauricio Sadinle for helpful comments that improved the content of the article. This work is supported by grants from the National Science Foundation ACI 1443014 and SES 1131897.

References


Azzalini, A. (2017). *The R package sn: The Skew-Normal and Related Distributions such as the Skew-t (version 1.5-1).* Università di Padova, Italia.


Figure 1: Montecarlo distribution of the $N$ Bayes factors in the simulation study in function of the sample sizes; 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 in in log10 scale.
Figure 2: Smoothed ROC curves for the three scenarios having null $\mathcal{M}_{10}$ (linear regression scenarios), $\mathcal{M}_{20}$, (gamma regression scenarios) and $\mathcal{M}_{30}$ (beta regression scenarios). Continuous, dotted, and dashed lines represent the curves for $n = 100, 250$, and $500$, respectively for the BNP and Brockwell’s procedures.

Figure 3: Scatterplots of the residuals against fitted values 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 dataset along with nonparametric smoothing (continuous line).
Figure 4: Kernel density estimates (continuous lines) of the residuals and related estimated residuals densities (dashed lines) for (a) the Gaussian regression model and (b) the skew-normal regression model for the Australian Institute of Sport dataset.