analytics

MDPI

Article

A New Semiparametric Regression Framework For Analyzing Non-Linear Data

Wesley Bertoli [1,*], Ricardo P. Oliveira [2] and Jorge A. Achcar [3]

[1] Department of Statistics, Federal University of Technology-Paraná, Curitiba 80230-901, Brazil
[2] Department of Statistics, Maringá State University, Maringá 87020-900, Brazil; rpuziol.oliveira@gmail.com
[3] Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto 14049-900, Brazil; achcar@fmrp.usp.br
* Correspondence: wbsilva@utfpr.edu.br

Abstract: This work introduces a straightforward framework for semiparametric non-linear models as an alternative to existing non-linear parametric models, whose interpretation primarily depends on biological or physical aspects that are not always available in every practical situation. The proposed methodology does not require intensive numerical methods to obtain estimates in non-linear contexts, which is attractive as such algorithms' convergence strongly depends on assigning good initial values. Moreover, the proposed structure can be compared with standard polynomial approximations often used for explaining non-linear data behaviors. Approximate posterior inferences for the semiparametric model parameters were obtained from a fully Bayesian approach based on the Metropolis-within-Gibbs algorithm. The proposed structures were considered to analyze artificial and real datasets. Our results indicated that the semiparametric models outperform linear polynomial regression approximations to predict the behavior of response variables in non-linear settings.

Keywords: Bayesian inference; non-linear data; non-linear regression modeling; polynomial models; semiparametric models

Citation: Bertoli, W.; Oliveira, R.P.; Achcar, J.A. A New Semiparametric Regression Framework For Analyzing Non-Linear Data. *Analytics* **2022**, *1*, 15–26. https://doi.org/10.3390/analytics1010002

Academic Editor: Fabio Postiglione

Received: 11 May 2022
Accepted: 7 June 2022
Published: 16 June 2022

1. Introduction

Non-linear models are often applied in many areas of quantitative research, such as biology, chemistry, epidemiology, and physics, among many others. These models have appeal in these areas as they have straightforward interpretability (regarding the nature of the underlying process) and typically provide excellent predictions for the response variable [1]. Alternatively to non-linear parametric models, researchers may consider approximating unknown non-linear functions using linear polynomial models. However, adopting such linear approximations to describe non-linear behaviors may be cumbersome as it could involve estimating more (and not easily interpretable) parameters [2]. On the other hand, non-linear models' are sensitive and should be carefully chosen depending on the application, and their parameters generally do not have analytical forms for their estimators. The absence of analytical solutions implies the need for numerical algorithms whose convergence strongly depends on the initial values chosen for the iterative procedures.

Several proposals for non-linear models can be found in the literature. However, modern techniques can also be used for non-linear modeling and predictions, such as nonparametric regression based on spline smoothing [3–5] and Generalized Additive Models (GAMs) [6,7]. In the context of agricultural applications, ref. [8] classifies non-linear models into six groups, as detailed in Table 1. Moreover, ref. [9] provides an excellent review on Gaussian Processes (GPs) and Relevance Vector Machines (RVMs), discussing how those nonparametric methods can be applied in non-linear frameworks for regressing over large datasets and how they can be effective for dealing with sequential data. The primary advantage of RVMs is that one can choose more general basis functions, and GPs present excellent behavior for predicting variances, although more restrictive regarding the kernel function choice. Ref. [10] has also studied such methods and concluded that the difficulty

of GPs is learning by maximizing the evidence, where the hyperparameters could be learned, which is distinct from RVMs with fixed basis functions where inputs are the learning targets.

Table 1. Groups of non-linear models in agricultural applications.

Group	Models	Published Works
I	Exponential Functions	[11,12]
II	Sigmoids (e.g., Logistic Function)	[13–17]
III	Asymptotic Exponential Modified Logistic Photosynthesis	[18,19]
IV	Modified Arrhenius Temperature Dependencies va not (Q_{10} Function)	[20–23]
V	Bell-shaped Curves Gaussian Function	[24]
VI	Michaelis–Menten Modified Hyperbola Power Functions Rational Functions Ricker Curve	[8,25–28]

In the presented context, this work aims to introduce a semiparametric non-linear regression model framework that can be very useful for obtaining highly accurate fits to non-linear datasets. Our goal is to provide an alternative to the existing linear approximations and nonparametric models. The proposed approach does not require numerical methods that strongly depend on precise initial values to reach convergence. By adopting the proposed framework, one could derive accurate inferences and predictions under a fully Bayesian approach [29] using standard MCMC (Markov Chain Monte Carlo) methods [30–32] (e.g., Gibbs Sampling, Metropolis–Hastings, and Metropolis-within-Gibbs (MwG), among others). In this paper, we have chosen to work with the MwG algorithm [33] to draw pseudo-random samples from the approximate posterior distribution of model parameters.

This paper is organized as follows. In Section 2, we present fundamental concepts regarding the formulation and estimation of parametric non-linear regression models. In Section 3, we analyze and discuss the results obtained using the proposed methodology for modeling artificial and real datasets featuring non-linear relationships. Model comparisons regarding linear polynomial approximations are also presented. General comments and concluding remarks are addressed in Section 4.

2. Materials and Methods

Non-linear models are similar to linear regressions [34] in the sense of outlining the functional relationship between a continuous response variable Y and a set of covariates, thus providing a statistical prediction tool. Linear regressions are used to build purely empirical models, while non-linear models are typically applied when biological or physical interpretations imply relationships between responses and covariates that are not linear [35,36]. It is important to establish that either linearity or non-linearity is related to the unknown parameters and not the response–covariates relationship. In this context, a non-linear regression model for representing a response variable Y_i $(i = 1, \ldots, n)$ has the general form

$$Y_i = f(z_i, \boldsymbol{\alpha}) + \epsilon_i,$$

where f is a known function of the designed covariate z_i, and $\boldsymbol{\alpha}^\intercal = (\alpha_1 \ldots, \alpha_p)$ is a p-dimensional vector of non-linear parameters indexing f. Moreover, ϵ_i denotes the random error, which is typically assumed to be normally distributed with zero mean and constant variance. It is also usual to assume that the errors are uncorrelated, that is, $\mathbb{C}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

The most popular method for estimating $\boldsymbol{\alpha}$ is the non-linear least squares, which is based on minimizing

$$S(\boldsymbol{\epsilon}) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - f(z_i, \boldsymbol{\alpha})]^2, \tag{1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. It is worth mentioning that if $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, then the least squares and maximum likelihood estimators of $\boldsymbol{\alpha}$ are the same.

Typically, point estimates for non-linear regression coefficients are obtained from iterative optimization processes based on techniques to minimize the error sum of squares. A widespread iterative method to derive least-squares estimates for non-linear models is the Gauss–Newton algorithm. In this context, if $f(z_i, \boldsymbol{\alpha})$ in Equation (1) is continuously differentiable at $\boldsymbol{\alpha}$, then f can be linearized locally at $\boldsymbol{\alpha}_0$ as

$$f(z_i, \boldsymbol{\alpha}) = f(z_i, \boldsymbol{\alpha}_0) + \mathbf{Z}_0(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0),$$

where $\mathbf{Z}_0$ is the $n \times p$ Jacobian matrix whose elements

$$\frac{\partial f(z_i, \boldsymbol{\alpha})}{\partial \alpha_j}$$

are evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$. Thus, the iterative algorithm to estimate $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} + \left(\mathbf{Z}_0^\intercal \mathbf{Z}_0\right)^{-1} \mathbf{Z}_0^\intercal \boldsymbol{\epsilon},$$

where $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}_0$ is the vector of initial values for $\boldsymbol{\alpha}$, and $\boldsymbol{\epsilon}$ is evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(k)}$. If the errors are independent and normally distributed, then the Gauss–Newton algorithm is an application of the Fisher Scoring method.

Implementations of the Gauss–Newton algorithm are available in most of the existing statistical software, but, in practice, there is no guarantee that the algorithm will converge from initial values that are far from the solution. In this sense, some improvements for this method can be found in the literature, such as the Gradient Descent and Levenberg–Marquart algorithms [36].

After obtaining point estimates for $\boldsymbol{\alpha}$, one may derive confidence intervals and conduct hypothesis tests by assuming

$$\hat{\boldsymbol{\alpha}} \overset{a}{\sim} \mathcal{N}_p\left[\boldsymbol{\alpha}, \sigma_\epsilon^2 (\mathbf{Z}_0^\intercal \mathbf{Z}_0)^{-1}\right],$$

where σ_ϵ^2 can estimated by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-p} \sum_{i=1}^{n} [y_i - f(z_i, \hat{\boldsymbol{\alpha}})]^2.$$

2.1. The Semiparametric Non-Linear Regression Model

Suppose that a random experiment is conducted with n subjects. The primary response in this setting is described by a random variable Y_i denoting the outcome for the i-th subject $(i = 1, \ldots, n)$. The full response vector of the experiment is given by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, and we assume that the behavior of Y_i can be partially explained by a non-linear relationship involving a designed covariate z_i through a known function f. Simultaneously, we can consider that part of the variability of Y_i can also be linearly modeled by a k-dimensional

vector $\boldsymbol{x}_i^\top = (x_{1i}, \ldots, x_{ki})$ of fixed covariates [37–39]. In this context, we have the non-linear regression model

$$Y_i = f(z_i, \boldsymbol{\alpha}) + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{2}$$

where $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_k)$ is a k-dimensional vector of regression coefficients related to $\boldsymbol{x}_i$, and ϵ_i is the random error of the i-th observation. Here, we assume that the errors are uncorrelated and normally distribution with zero mean and constant variance (σ_ϵ^2).

A particular case arising from Equation (2) is the p-order polynomial regression model, which can be obtained by taking

$$f(z_i, \boldsymbol{\alpha}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j z_i^j.$$

In the context of Model (2), let $\boldsymbol{z} = (z_1, \ldots, z_n)$ be the full vector of designed values. In order to obtain an approximation for f, we assume that $z_1 \leqslant z_2 \leqslant \cdots \leqslant z_n$, and then we associate these values to each Y_i non-linearly by $\boldsymbol{\alpha}$. Thus, for each point z_i $(i = 3, \ldots, n)$, we take $a = z_{i-1}$, $a + h = z_i$, and $a - h = z_{i-2}$ to express the approximation $f(z_i)$ to f as

$$f(z_i) = f(z_{i-1}) + \frac{[f(z_i) - f(z_{i-1})](z_i - z_{i-1})}{h_i} + \frac{[f(z_i) - 2f(z_{i-1}) + f(z_{i-2})](z_i - z_{i-1})^2}{2h_i^2},$$

which is based in a Taylor's series of the function $f(z_i)$ around z_{i-1}. Now, one can notice that replacing $f(z_i)$ with the observed data on the right side of the previous equation leads to the approximation

$$f(z_i) \approx y_{i-1} + g_1(z_i) + g_2(z_i),$$

where

$$g_1(z_i) = \frac{(y_i - y_{i-1})(z_i - z_{i-1})}{h_i} \quad \text{and} \quad g_2(z_i) = \frac{(y_i - 2y_{i-1} + y_{i-2})(z_i - z_{i-1})^2}{2h_i^2},$$

with $h_i = z_{i+1} - z_i$. Therefore, an alternative for Model (2) is the semiparametric non-linear regression model given by

$$Y_i = \alpha_1 Y_{i-1} + \alpha_2 g_1(z_i) + \alpha_3 g_2(z_i) + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{3}$$

which holds for $i \in \{3, \ldots, n\}$.

2.2. Bayesian Inference

In this subsection, we address the problem of estimating and making inferences from Model (2) under a fully Bayesian perspective. Firstly, the log-likelihood of vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \zeta)$ can be written as

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) \propto \frac{n}{2} \log(\zeta) - \frac{\zeta}{2} \sum_{i=1}^{n} \left[y_i - f(z_i, \boldsymbol{\alpha}) - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right]^2,$$

where $\zeta = \sigma_\epsilon^{-2}$ is the precision parameter.

For the p-order polynomial model, we have $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)$ and, specifically for the semiparametric non-linear regression model, we have $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$. In either case, the log-likelihood function of $\boldsymbol{\theta}$ can be expressed by

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) \propto \frac{n}{2} \log(\zeta) - \frac{\zeta}{2} \sum_{i=3}^{n} \left[y_i - \alpha_1 y_{i-1} - \alpha_2 g_1(z_i) - \alpha_3 g_2(z_i) - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right]^2.$$

In this work, we have adopted weakly informative Normal prior distributions for the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, that is

$$\boldsymbol{\alpha} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{1}_q) \quad \text{and} \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{1}_k),$$

where $\mathbf{1}_q$ and $\mathbf{1}_k$ are identity matrices of sizes q and k, respectively. For the p-order polynomial model, we have that $q = p + 1$. As for parameter ζ, we have adopted a Gamma prior distribution with both hyperparameters equal to 0.01. We further assume prior independence among all parameters.

Now, we can express the posterior distribution of $\boldsymbol{\theta}$ as

$$\pi(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) \propto \exp\{\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) + \log[\pi(\boldsymbol{\alpha})] + \log[\pi(\boldsymbol{\beta})] + \log[\pi(\zeta)]\}. \tag{4}$$

From the Bayesian point of view, inferences for the elements of $\boldsymbol{\theta}$ can be derived from their marginal posterior distribution. Here, we have opted to use a suitable iterative procedure to draw pseudo-random samples from the approximate posterior density (Equation (4)) in order to make inferences for $\boldsymbol{\theta}$. Thus, in order to generate N pseudo-random values for each element of $\boldsymbol{\theta}$, we have adopted the MwG algorithm.

The simulated sequences' convergence can be monitored using trace, autocorrelation plots, and statistical tests (e.g., Heidelberger and Welch [40] and Geweke [41]). After diagnosing convergence, some samples can be discarded as burn-in. The strategy to decrease the correlation between generated values is based on getting thinned steps, and so the final sample is supposed to have size $B \ll N$. After that, a descriptive summary of Equation (4) can be obtained through approximate Monte Carlo estimators using the generated chains. We choose the posterior expected value as the Bayesian point estimator for the elements of $\boldsymbol{\theta}$.

The next section illustrates the usefulness of the proposed semiparametric non-linear regression model using artificial and real datasets. All computations were performed using the R2jags package, which is available in the `R` environment [42]. The executable scripts can be made available by the authors upon justified request.

2.3. Model Comparison

There are many methods for Bayesian model selection that are useful for comparing competing models. The most popular method is the Deviance Information Criterion (DIC), which works simultaneously to measure the model's fit and complexity. The DIC criterion is defined as

$$\text{DIC} = \mathbb{E}_{\boldsymbol{\theta}}[\text{D}(\boldsymbol{\theta})] + p_{\text{D}} = \underline{\text{D}}(\boldsymbol{\theta}) + p_{\text{D}},$$

where $\text{D}(\boldsymbol{\theta}) = -2\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$ is the *deviance* function, and $p_{\text{D}} = \underline{\text{D}}(\boldsymbol{\theta}) - \text{D}(\hat{\boldsymbol{\theta}})$ is the effective number of model parameters, where $\hat{\boldsymbol{\theta}}$ is the posterior expected value.

Noticeably, we are not able to compute the expectation of $\text{D}(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$ analytically. Therefore, an approximate Monte Carlo estimator for such a measure is

$$\widehat{\underline{\text{D}}}(\boldsymbol{\theta}) = -\frac{2}{B}\sum_{i=1}^{B} \ell(\boldsymbol{\theta}_i; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}),$$

and so the DIC can be estimated by

$$\widehat{\text{DIC}} = 2\widehat{\underline{\text{D}}}(\boldsymbol{\theta}) - \text{D}(\hat{\boldsymbol{\theta}}).$$

The Expected Akaike (EAIC) and the Expected Bayesian (EBIC) information criteria can also be used when comparing Bayesian models [43,44]. Based on the approximation for the expected value of $\text{D}(\boldsymbol{\theta})$, these measures can be estimated by

$$\widehat{\text{EAIC}} = \widehat{\underline{\text{D}}}(\boldsymbol{\theta}) + 2k \quad \text{and} \quad \widehat{\text{EBIC}} = \widehat{\underline{\text{D}}}(\boldsymbol{\theta}) + k\log(n),$$

where $k = \dim(\boldsymbol{\theta})$.

Another widely used criterion is derived as a posterior measure of goodness-of-fit based on the observed and predicted values. This measure is given by

$$A[m] = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{\mu}_i|, \tag{5}$$

where $\hat{\mu}_i$ denotes the estimated mean of Y_i, which depends on the adopted model (m). For instance, under the semiparametric non-linear regression model in Equation (3), we have that

$$A[(3)] = \frac{1}{n-2}\sum_{i=3}^{n}\left|y_i - \hat{\alpha}_1 y_{i-1} - \hat{\alpha}_2 g_1(z_i) - \hat{\alpha}_3 g_2(z_i) - \boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}\right|,$$

since the first two observations are not considered when computing A under the semiparametric model in Equation (3).

3. Non-Linear Data Analysis

To illustrate the usefulness of the proposed methodology, we have considered three datasets and the non-linear models presented in Section 2.1. Using the MwG algorithm, a total of $N = 110{,}000$ pseudo-random values from the approximate posterior distribution in Equation (4) of $\boldsymbol{\theta}$ were obtained. After generating the values, the first 10,000 samples were discarded (burn-in period). Then, 1 out of every 100 generated values was kept, resulting in sequences of the size $B = 1000$ for each element of $\boldsymbol{\theta}$. Finally, trace plots were used to assess the stationarity of the obtained chains.

3.1. Artificial Data

Let us consider the artificial dataset displayed on Table 2. This dataset has $n = 21$ observations and is composed of a designed and a fixed covariate. For analyzing these data, we have adopted the two-order polynomial model

$$Y_i = \alpha_0 + \alpha_1 z_i + \alpha_2 z_i^2 + x_i\beta + \epsilon_i \quad (i = 1, \ldots, 21), \tag{6}$$

and the semiparametric non-linear regression model

$$Y_i = \alpha_1 Y_{i-1} + \alpha_2 g_1(z_i) + \alpha_3 g_2(z_i) + x_i\beta + \epsilon_i \quad (i = 3, \ldots, 21). \tag{7}$$

Table 2. Artificial dataset from a hypothetical experiment with 21 subjects.

y	x	z	y	x	z
12	1	10.0375	31	12	8.9171
14	2	11.4128	29	13	10.0933
15	3	9.8035	27	14	11.9097
18	4	9.9774	25	15	11.0709
20	5	9.0706	20	16	10.3041
21	6	10.8220	19	17	9.4895
22	7	9.6170	18	18	9.1792
25	8	9.1354	16	19	9.5295
28	9	10.0180	15	20	9.2414
30	10	10.1596	10	21	10.3354
34	11	10.3520	-	-	-

Table 3 presents the posterior parameter estimates and the 95% Credible Intervals (CIs) based on the fitted models. From the displayed results, one can make some conclusions. Firstly, one can notice that the CIs of parameter α_1 of both models do not contain the value zero, which constitute z and $g_1(z)$ as relevant covariates to explain part of the response's variability. The comparison procedure between the fitted models is presented in Table 4. One can notice that Model (6) has performed poorly compared with the proposed semiparametric non-linear regression model.

Table 3. Posterior parameter estimates and 95% credible intervals for the artificial dataset.

Model	Parameter	Mean	Std. Dev.	95% CI	
				Lower	Upper
(6)	α_0	1.3780	2.7771	−3.7411	7.0241
	α_1	3.6490	0.4037	2.8950	4.4070
	α_2	−0.1659	0.0181	−0.1990	−0.1297
	β	0.5975	0.3185	−0.0185	1.2420
	ζ	0.1697	0.0592	0.0756	0.3008
(7)	α_1	1.0000	0.0029	0.9944	1.0061
	α_2	0.9990	0.0121	0.9749	1.0230
	α_3	0.0012	0.0257	−0.0468	0.0518
	β	−0.0002	0.0064	−0.0131	0.0122
	ζ	85.2600	29.6000	39.1100	152.5000

Table 4. Posterior comparison criteria for the fitted models for the artificial dataset.

Model	k	p_D	DIC	EAIC	EBIC
(6)	5	4.826	103.986	109.160	114.382
(7)	5	5.022	59.858	69.413	79.857

Figure 1 illustrates the behavior of the predicted responses against the values of the designed covariate. When considering the goodness-of-fit measure in Equation (5), we have that $A[(6)] = 1.8061$ and $A[(7)] = 0.0007$, which indicates that the proposed semiparametric non-linear model has performed better in predicting the response variable. In order to reassure such a conclusion, these models were refitted considering only the first 20 observations, so we could predict the 21st outcome ($y_{21} = 10$). From Model (6), we have obtained $\hat{y}_{21} = 11.8$ (± 2.2), and from the semiparametric non-linear regression model, we obtained $\hat{y}_{21} = 10.4$ (± 3.3), which also suggest a better fit of Model (7).

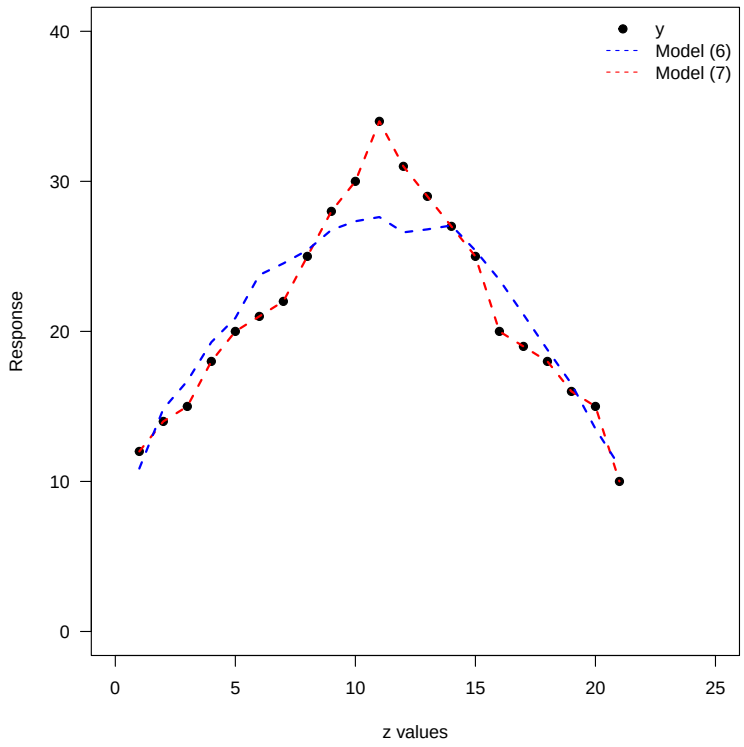

Figure 1. Predicted responses vs. designed covariate (z) for the artificial dataset.

3.2. COVID-19 Count Data

As a second application, we have considered data from $n = 358$ daily counts of cases and deaths caused by COVID-19 in Brazil (from 17 March 2020 to 21 March 2021). For analyzing these data, we have adopted a second-order autoregressive (AR) model with lagged effects given by

$$Y_i = \beta_0 + \beta_1 day_i + \beta_2 Y_{i-1} + \beta_3 Y_{i-2} + \epsilon_i \quad (i = 1, \ldots, 358), \tag{8}$$

and, for the moving averages (MA) of daily COVID-19 counts (average of last seven days), we have considered the following semiparametric non-linear model:

$$Y_i = \alpha_1 Y_{i-1} + \alpha_2 g_1(days_i) + \alpha_3 g_2(days_i) + \epsilon_i \quad (i = 1, \ldots, 358). \tag{9}$$

Table 5 presents the posterior parameter estimates and the 95% CIs based on the fitted models. From the fitted AR(2) model, it can be noticed that the covariate *day* is not relevant to describing the incidence behavior of cases and deaths by COVID-19 in the observed time frame. The comparison procedure between the fitted models is presented in Table 6. Noticeably, the semiparametric non-linear model outperformed the AR(2) model in both cases, which can be acknowledged as an excellent result since Model (9) has one less parameter.

Table 5. Posterior parameter estimates and 95% credible intervals for the COVID-19 dataset.

Count	Model	Parameter	Mean	Std. Dev.	95% CI	
					Lower	Upper
Cases	(8)	β_0	0.0371	1.0080	−1.8620	1.9800
		β_1	0.8959	0.7839	−0.6451	2.5610
		β_2	1.3070	0.0540	1.2100	1.4230
		β_3	−0.3089	0.0545	−0.4249	−0.2088
		ζ	<0.0001	<0.0001	<0.0001	<0.0001
	(9)	α_1	1.0060	0.0021	1.0021	1.0100
		α_2	−0.1662	0.0222	−0.2113	−0.1265
		α_3	−5.3710	0.5814	−6.5710	−4.3110
		ζ	<0.0001	<0.0001	<0.0001	<0.0001
Deaths	(8)	β_0	−0.0296	0.9695	−1.9500	1.8390
		β_1	0.0161	0.0185	−0.0214	0.0556
		β_2	1.2930	0.0550	1.1940	1.4080
		β_3	−0.2908	0.0558	−0.4049	−0.1893
		ζ	0.0009	<0.0001	<0.0001	<0.0001
	(9)	α_1	1.0101	0.0018	1.0060	1.0130
		α_2	−0.1681	0.0221	−0.2136	−0.1287
		α_3	−5.4050	0.5801	−6.5920	−4.3460
		ζ	0.0011	<0.0001	<0.0001	<0.0001

Table 6. Posterior comparison criteria for the fitted models for the COVID-19 dataset.

Count	Model	k	p_D	DIC	EAIC	EBIC
Cases	(8)	5	3.084	6237.000	6243.637	6263.040
	(9)	4	3.186	6194.920	6193.722	6197.602
Deaths	(8)	5	4.174	4104.000	4109.994	4129.397
	(9)	4	3.945	4057.920	4060.384	4063.967

Figure 2 illustrates the fitted means (moving averages for COVID-19 cases and deaths) across days. Noticeably, both models provide excellent fits for the COVID-19 counts, with the semiparametric non-linear model being slightly better than the AR(2) since we have

$A[(8)] = 994.03$ against $A[(9)] = 934.11$ for the number of cases and $A[(8)] = 21.59$ against $A[(9)] = 20.68$ for the number of deaths.

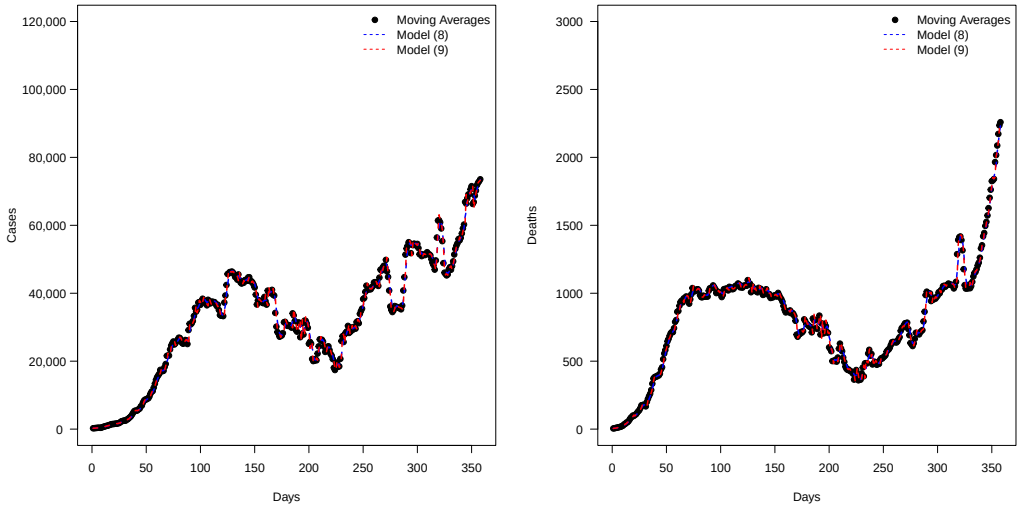

Figure 2. Fitted means for the daily number of COVID-19 cases (**left panel**) and deaths (**right panel**).

3.3. Tuberculosis Count Data

For the last application, we have considered data from $n = 216$ monthly counts of tuberculosis cases in Brazil (from January 2001 to December 2018). For analyzing these data, we have adopted the three-order polynomial regression model:

$$Y_i = \alpha_0 + \alpha_1 z_i + \alpha_2 z_i^2 + \alpha_3 z_i^3 + \beta\, year_i + \epsilon_i \quad (i = 1, \ldots, 216), \tag{10}$$

and the following semiparametric non-linear regression model:

$$Y_i = \alpha_1 Y_{i-1} + \alpha_2 g_1(z_i) + \alpha_3 g_2(z_i) + \beta(year_i - 2000) + \epsilon_i \quad (i = 1, \ldots, 216). \tag{11}$$

Table 7 presents the posterior parameter estimates and the 95% Credible Intervals (CIs) based on the fitted models. From the displayed results, one can make some conclusions. Firstly, one can notice that the CIs of parameter β of Model (10) do not contain the value zero, which constitute *year* as a relevant covariate to explain part of the response' variability.

Table 7. Posterior parameter estimates and 95% credible intervals for the tuberculosis dataset.

Model	Parameter	Mean	Std. Dev.	95% CI	
				Lower	Upper
(10)	α_0	0.7285	0.7262	−0.5683	1.9650
	α_1	0.0007	0.0008	<−0.0001	0.0023
	α_2	<−0.0001	<0.0001	<−0.0001	<−0.0001
	α_3	<0.0001	<0.0001	<0.0001	<0.0001
	β	0.0041	0.0004	0.0035	0.0047
	ζ	208.1000	19.8200	172.4000	249.7000
(11)	α_1	0.9998	0.0009	0.9982	1.0000
	α_2	−0.1817	0.0197	−0.2273	−0.1478
	α_3	−6.2030	0.5525	−7.3740	−5.1280
	β	0.0002	0.0007	−0.0012	0.0015
	ζ	340.1000	34.2600	275.5000	411.0000

The comparison procedure between the fitted models is presented in Table 8. One can notice that even having one less parameter, the proposed semiparametric non-linear model (Equation (11)) has performed much better than the polynomial model (Equation (10)).

Table 8. Posterior comparison criteria for the fitted models for the tuberculosis dataset.

Model	k	p_D	DIC	EAIC	EBIC
(10)	6	4.171	3316.000	3323.942	3344.194
(11)	5	4.121	1582.000	1596.344	1613.220

Figure 3 illustrates the fitted means for the daily number of tuberculosis cases. When considering the goodness-of-fit measure from Equation (5), we have that $A[(10)] = 407$ and $A[(11)] = 315.83$, which indicates that the semiparametric non-linear regression model (Equation (11)) has performed better in predicting the number of tuberculosis cases. In the following, these models were refitted considering only the first 215 observations, so we could predict the 216th outcome ($y_{216} = 6836$). From Model (10), we have obtained $\hat{y}_{216} = 7926.76$ (± 0.019), and from the semiparametric non-linear regression model we obtained $\hat{y}_{216} = 7030.41$ (± 0.003), which also suggest a better fit of Model (11).

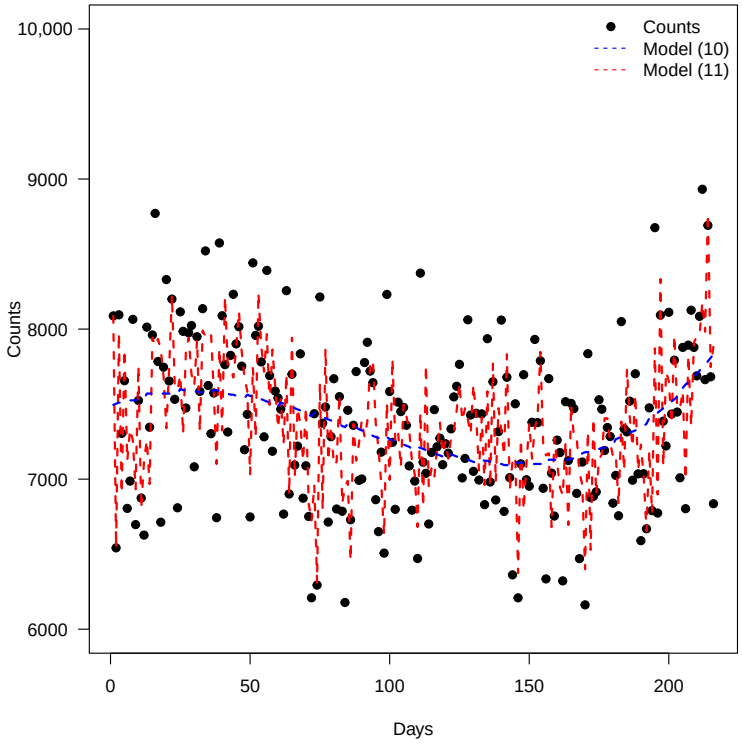

Figure 3. Fitted means for the daily number of tuberculosis cases.

4. Concluding Remarks

Parametric non-linear approaches typically involve choosing a model among many existing non-linear formulations, which can be a burden in many applications. Moreover, most numerical iterative methods for model fitting strongly depend on choosing precise initial values. However, non-linear models often provide insightful parameter (biological or physical) interpretations for many researchers. In this sense, we aimed to introduce a semiparametric non-linear regression framework as an alternative to standard non-linear models. The proposed model can be considered an excellent alternative to many existing nonparametric regression techniques based on spline smoothing and GAM. Approximate posterior inferences for the model parameters were obtained from a fully Bayesian approach based on MwG with weakly informative priors. The proposed model and some well-established non-linear models were considered for analyzing three datasets. Based on the prediction accuracy, we could conclude that the proposed semiparametric framework can be a powerful alternative for estimation and prediction in non-linear settings.

Supplementary Materials: The datasets information can be downloaded at: https://www.mdpi.com/article/10.3390/analytics1010002/s1.

Author Contributions: Conceptualization, W.B., R.P.O. and J.A.A.; Formal analysis, W.B., R.P.O. and J.A.A.; Methodology, W.B., R.P.O. and J.A.A.; Software, W.B., R.P.O. and J.A.A.; Writing– original draft, W.B., R.P.O. and J.A.A.; Writing–review and editing, W.B., R.P.O. and J.A.A. All authors equally contributed to developing this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this work were made available as Supplementary Materials.

Acknowledgments: We would like to thank the Associate Editor and the three anonymous referees for their careful reading and thoughtful suggestions, which helped us improve this work's content and presentation.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AR	Autoregressive
CI	Credible Interval
GAM	Generalized Additive Models
GP	Gaussian Process
MA	Moving Average
MCMC	Markov Chain Monte Carlo
MwG	Metropolis-within-Gibbs
RVM	Relevance Vector Machine

References

1. Bates, D.M.; Watts, D.G. *Nonlinear Regression Analysis and Its Applications*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2007.
2. Pinheiro, J.C.; Bates, D.M. *Mixed-Effects Models in S and S-Plus*; Springer: New York, NY, USA, 2000.
3. Eubank, R.L. *Spline Smoothing and Nonparametric Regression*; Marcel Dekker: New York, NY, USA, 1988.
4. Green, P.J.; Silverman, B.W. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*; Chapman & Hall: London, UK, 1994.
5. Gu, C. *Smoothing Spline ANOVA Models*, 2nd ed.; Springer: New York, NY, USA, 2013.
6. Hastie, T.; Tibshirani, R. *Generalized Additive Models*; Chapman & Hall: London, UK, 1990.
7. Hastie, T.; Tibshirani, R. Varying-coefficient models. *J. R. Stat. Soc. Ser. B* **1993**, *55*, 757–796. [CrossRef]
8. Archontoulis, S.V.; Miguez, F.E. Nonlinear regression models and applications in agricultural research. *Agron. J.* **2015**, *107*, 786–798. [CrossRef]
9. Martino, L.; Read, J. A joint introduction to Gaussian processes and relevance vector machines with connections to Kalman filtering and other kernel smoothers. *Inf. Fusion* **2021**, *74*, 17–38. [CrossRef]
10. Candela, J.Q. *Learning with Uncertainty-Gaussian Processes and Relevance Vector Machines*; Technical University of Denmark: Copenhagen, Denmark, 2004; pp. 1–152.
11. Dixon, B.L.; Sonka, S.T. A note on the use of exponential functions for estimating farm size distributions. *Am. J. Agric. Econ.* **1979**, *61*, 554–557. [CrossRef]
12. Shimojo, M.; Nakano, Y. An investigation into relationships between exponential functions and some natural phenomena. *J. Fac. Agric. Kyushu Univ.* **2013**, *58*, 51–53. [CrossRef]
13. Gompertz, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philos. Trans. R. Soc. B* **1825**, *115*, 513–585.
14. Verhulst, P.F. A note on population growth. *Corresp. Math. Phys.* **1838**, *10*, 113–121.
15. Weibull, W. A statistical distribution function of wide applicability. *J. Appl. Math.* **1951**, *18*, 293–297. [CrossRef]
16. Richards, F.J. A flexible growth function for empirical use. *J. Exp. Bot.* **1959**, *10*, 290–300. [CrossRef]
17. Yin, X.; Goudriaan, J.; Lantinga, E.A.; Vos, J.; Spiertz, J.H.J. A flexible sigmoid function of determinate growth. *Ann. Bot.* **2003**, *91*, 361–371. [CrossRef] [PubMed]

18. Blackman, F.F. Optima and limiting factors. *Ann. Bot.* **1905**, *19*, 281–295. [CrossRef]
19. Sinclair, T.R.; Horie, T. Leaf nitrogen, photosynthesis, and crop radiation use efficiency: A review. *Crop Sci.* **1989**, *29*, 90–98. [CrossRef]
20. van't Hoff, J.H. *Lectures on Theoretical and Physical Chemistry. Part 1: Chemical Dynamics*; Edward Arnold: London, UK, 1898.
21. Arrhenius, S. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Z. Für Phys. Chem.* **1889**, *4*, 226–248. [CrossRef]
22. Ratkowsky, D.A.; Olley, J.; McMeekin, T.A.; Ball, A. Relationship between temperature and growth rate of bacterial cultures. *J. Bacteriol.* **1982**, *149*, 1–5. [CrossRef]
23. Lloyd, J.; Taylor, J.A. On the temperature dependence of soil respiration. *Funct. Ecol.* **1994**, *8*, 315–323. [CrossRef]
24. Yin, X.; Kroff, M.J.; McLean, G.; Visperas, R.M. A nonlinear model for crop development as a function of temperature. *Agric. For. Meteorol.* **1995**, *77*, 1–16. [CrossRef]
25. Hu, Y.; Tao, V.; Croitoru, A. Understanding the rational function model: Methods and applications. *Int. Arch. Photogramm. Remote Sens.* **2004**, *20*, 119–124.
26. Braverman, E.; Kinzebulatov, D. On linear perturbations of the Ricker model. *Math. Biosci.* **2006**, *202*, 323–339. [CrossRef] [PubMed]
27. Nijland, G.O.; Schouls, J.; Goudriaan, J. Integrating the production functions of Liebig, Michaelis-Menten, Mitscherlich and Liebscher into one system dynamics model. *NJAS-Wagening. J. Life Sci.* **2008**, *55*, 199–224. [CrossRef]
28. Ye, Z.; Zhao, Z. A modified rectangular hyperbola to describe the light-response curve of photosynthesis of *Bidens pilosa* L. grown under low and high light conditions. *Front. Agric. China* **2010**, *4*, 50–55. [CrossRef]
29. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*; John Wiley & Sons: New York, NY, USA, 1994.
30. Gelfand, A.E.; Smith, A.F.M. Sampling based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **1990**, *85*, 398–409. [CrossRef]
31. Casella, G.; George, E.I. Explaining the Gibbs sampler. *Am. Stat.* **1992**, *46*, 167–174.
32. Chib, S.; Greenberg, E. Understanding the Metropolis-Hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
33. Gilks, W.R.; Best, N.G.; Tan, K.K. Adaptive rejection Metropolis sampling within Gibbs sampling. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1995**, *44*, 455–472. [CrossRef]
34. Seber, G.A.F.; Lee, A.J. *Linear Regression Analysis*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2003.
35. Ratkowsky, D.A. *Nonlinear Regression Modelling: A Unified Practical Approach*; Marcel Dekker: New York, NY, USA, 1983.
36. Seber, G.A.F.; Wild, C.J. *Nonlinear Regression*; John Wiley & Sons: New York, NY, USA, 1989.
37. Koop, G.; Poirier, D.J. Bayesian variants of some classical semiparametric regression techniques. *J. Econom.* **2004**, *123*, 259–282. [CrossRef]
38. Munkin, M.; Trivedi, P. Bayesian analysis of the ordered probit model with endogenous selection. *J. Econom.* **2008**, *143*, 334–348. [CrossRef]
39. Feng, L.; Munkin, M. Bayesian semiparametric analysis on the relationship between BMI and income for rural and urban workers in China. *J. Appl. Stat.* **2021**. [CrossRef]
40. Heidelberger, P.; Welch, P.D. Simulation run length control in the presence of an initial transient. *Oper. Res.* **1983**, *31*, 1109–1144. [CrossRef]
41. Geweke, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *J. R. Stat. Soc.* **1994**, *56*, 501–514.
42. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
43. Carlin, B.P.; Louis, T.A. *Bayes and Empirical Bayes Methods for Data Analysis*; Chapman & Hall: Boca Raton, FL, USA, 2001.
44. Brooks, S.P. Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2002**, *64*, 616–639.