Journal of Data Science 19 (1), 96–110
January 2021

DOI: 10.6339/21-JDS1003
Statistical Data Science

Validation of Stepwise-Based Procedure in GAMLSS

Thiago G. Ramires[1,*], Luiz R. Nakamura[2], Ana J. Righetto[3], Rodrigo R. Pescim[4], Josmar Mazucheli[5], Robert A. Rigby[6], and Dimitrios M. Stasinopoulos[6]

[1] *Universidade Tecnológica Federal do Paraná, Apucarana, Brazil*
[2] *Universidade Federal de Santa Catarina, Florianópolis, Brazil*
[3] *Alvaz Agritech, Londrina, Brazil*
[4] *Universidade Estadual de Londrina, Londrina, Brazil*
[5] *Universidade Estadual de Maringá, Maringá, Brazil*
[6] *London Metropolitan University, London, United Kingdom*

Abstract

One of the key features in regression models consists in selecting appropriate characteristics that explain the behavior of the response variable, in which stepwise-based procedures occupy a prominent position. In this paper we performed several simulation studies to investigate whether a specific stepwise-based approach, namely Strategy A, properly selects authentic variables into the generalized additive models for location, scale and shape framework, considering Gaussian, zero inflated Poisson and Weibull distributions. Continuous (with linear and nonlinear relationships) and categorical explanatory variables are considered and they are selected through some goodness-of-fit statistics. Overall, we conclude that the Strategy A greatly performed.

Keywords *backward; forward; model selection; smoothing*

1 Introduction

In the past decades, statistical regression models have been greatly improved with the development of extremely sophisticated models in order to deal with an increasing amount of complex datasets. Back in the days, Sir Francis Galton introduced the concept of regression toward the mean with his experiments on the size of the seeds of successive generations of sweet peas (Stanton, 2001). Since then, some well-known extensions based on the same concept were developed, such as the generalized linear models (Nelder and Wedderburn, 1972) and the generalized additive models (Hastie and Tibshirani, 1990).

However, depending on the complexity of the data in study, we may have to consider more flexible models that are able to explain not only the mean of the response (target) variable distribution, but in fact all of its parameters, i.e. a beyond mean regression model (Kneib, 2013). In this sense, the generalized additive models for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos, 2005) seem to be a great alternative.

The GAMLSS framework involves a distribution for the response variable (that does not necessarily belong to the exponential family) (Rigby et al., 2019) and may involve parametric linear and/or nonparametric smoothing terms when modelling any or all of the parameters of the distribution as functions of the explanatory variables (Stasinopoulos et al., 2018). Within GAMLSS, any distribution parameter can be modeled as a function of explanatory variables and hence different regression structures might be selected for each of them. Alternative ap-

[*] Corresponding author Email: thiagogentil@gmail.com.

Received November, 2020; Accepted January, 2021

proaches, including criterion-based, regularization and dimension-reduction validation methods are discussed in Stasinopoulos et al. (2017) and available in *gamlss* package (Stasinopoulos and Rigby, 2007) in R software (R Core Team, 2020).

The main used method to select the explanatory variables in each of the regression structures in the GAMLSS framework is a stepwise-based procedure performed in each of distribution parameters called Strategy A (Stasinopoulos et al., 2017). Recent examples of its application can be seen in Ayuso et al. (2020), Righetto et al. (2019), De Bastiani et al. (2018), Ramires et al. (2018), Leroy et al. (2016), among others. Nonetheless, no formal studies regarding this methodology are presented in the literature apart from two quite simple and specific studies: considering one predictor only (Voncken et al., 2019) and in a specific distribution on the unit interval (Nakamura et al., 2019). Hence, the aim of this paper is to study the behavior of the Strategy A procedure within the *gamlss* package and validate it through a set of simulation studies, considering different response variable distributions (Gaussian, zero inflated Poisson and Weibull), structures (linear and nonlinear relationships between a parameter and explanatory variables) and sample sizes. It is noteworthy that these distributions were considered since they are commonly applied in a wide range of problems. Nonetheless, more complex distributions, characteristics and behaviors might be considered in future papers.

This paper is organized as follows. In Section 2, we present the GAMLSS framework and the adopted strategy regarding model selection. All simulation studies are presented in Section 3. Finally, Section 4 ends the paper with some concluding remarks.

2 GAMLSS Framework

Consider Y a univariate response variable which follows a specific distribution $\mathcal{D}(y;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is its vector of parameters (e.g. for a four parameter distribution $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^\top$). Mathematically, a GAMLSS can be written as

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} s_{jk}(\boldsymbol{x}_{jk}), \quad k = 1, \dots, p, \tag{1}$$

where $g_k(\boldsymbol{\theta}_k)$ is a link function usually determined by the range of $\boldsymbol{\theta}_k$ (for further details, please check De Bastiani et al., 2018; Stasinopoulos and Rigby, 2007), $\boldsymbol{X}_k$ is a design matrix, $\boldsymbol{\beta}_k$ is the parameter vector associated to $\boldsymbol{X}_k$ and each s_{jk} function is a smooth nonparametric function (e.g. a P-spline. See Eilers and Marx, 1996; Eilers et al., 2015) of an explanatory variable $\boldsymbol{x}_{jk}$. If, for $k = 1, \dots, p$, $J_k = 0$, i.e. if no smooth functions are fitted in model (1), then we have the fully parametric GAMLSS given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{X}_k\boldsymbol{\beta}_k. \tag{2}$$

In order to estimate $\boldsymbol{\beta}_k$ and also the parameters associated to the second term (say $\boldsymbol{\gamma}_{jk}$) in model (1), we shall fix the smoothing hyperparameters $\boldsymbol{\lambda}$ and maximize the penalized log-likelihood function. An example for a four-parameter distribution model is given by

$$l_p = \sum_{i=1}^{n} \log f(y_i; \mu_i, \sigma_i, \nu_i, \tau_i) - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \lambda_{jk}\boldsymbol{\gamma}_{jk}^\top \boldsymbol{P}_{jk}\boldsymbol{\gamma}_{jk}, \tag{3}$$

where $\boldsymbol{P}_{jk}$ is a symmetric matrix that may depend on a vector of smoothing parameters (Stasinopoulos and Rigby, 2007). The smoothing hyperparameters $\boldsymbol{\lambda}$ can be estimated through the penalized quasi likelihood method (Lee et al., 2006) and is implemented in the `pb()` function in the *gamlss* package (for further details, please check Rigby and Stasinopoulos, 2014). If the fully parametric GAMLSS is being considered, then we only have to maximize the first term in (3).

Further, for survival data, which is characterized by censored data, the response variable of n-independent observations is defined by $y_i = \min\{\log(t_i), \log(c_i)\}$, where t_i and c_i represent the times of failure or censure, respectively. Let $\delta_i = 1$ and $\delta_i = 0$ the indicator variable for uncensored and censored observations, respectively. Considering non-informative censoring, i.e., lifetimes and censoring times independent, we maximize (3) replacing its first term by $\sum_{i=1}^{n} [\delta_i \log f(y_i; \mu_i, \sigma_i, \nu_i, \tau_i) + (1 - \delta_i) \log[1 - F(y_i; \mu_i, \sigma_i, \nu_i, \tau_i)]]$.

The numerical maximization of (3) can be achieved in the *gamlss* (or *gamlss.cens* for survival data) package in R using a generalization of the Cole and Green (CG) algorithm (Cole and Green, 1992), the Rigby and Stasinopoulos (RS) algorithm (Rigby and Stasinopoulos, 2005) or even through a combination of both. Both methods are well described in Rigby and Stasinopoulos (2005), Stasinopoulos and Rigby (2007) and Stasinopoulos et al. (2017), in which the main difference is that RS algorithm does not use the cross derivatives of the log-likelihood. In this paper we are using only the RS algorithm (default in *gamlss* package) which is generally more stable and, in most cases, faster than CG (Stasinopoulos et al., 2017).

2.1 Selecting Explanatory Variables

According to (Stasinopoulos et al., 2017), there are currently 13 different functions (methods) in the *gamlss* package that may assist us to select different subsets of explanatory variables for each of the parameters of a given response variable distribution (boosting is a different approach from these techniques and is discussed in Mayr et al., 2012; Hofner et al., 2016).

As mentioned in Section 1, the most used methodology to select covariates in GAMLSS is the Strategy A, a stepwise-based procedure, that can be accessed in the *gamlss* package through the `stepGAICAll.A()` function. For a four parameter distribution (i.e. $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^\top$), the steps of this approach are described in Nakamura et al. (2017) and Stasinopoulos et al. (2017) as follow:

1. Use a forward selection procedure to select an appropriate model for μ, with σ, ν and τ fitted as constants.
2. Given the model for μ obtained in step 1 and for ν and τ fitted as constants, use a forward selection procedure to select an appropriate model for σ.
3. Given the models for μ and σ obtained in steps 1 and 2 respectively and with τ fitted as constant, use a forward selection procedure to select an appropriate model for ν.
4. Given the models for μ, σ and ν obtained in steps 1, 2 and 3 respectively, use a forward selection procedure to select an appropriate model for τ.
5. Given the models for μ, σ and τ obtained in steps 1, 2 and 4 respectively, use a backward selection procedure to select appropriate model for ν,
6. Given the models for μ, ν and τ obtained in steps 1, 5 and 4 respectively, use a backward selection procedure to select appropriate model for σ.
7. Given the models for σ, ν and τ obtained in steps 6, 5 and 4 respectively, use a backward selection procedure to select an appropriate model for μ and then stop.

By the end of these steps, the final model may contain different subsets of covariates for $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$. The criterion used to add (or remove) a variable in each regression structure is based on the generalized Akaike information criterion (GAIC; Voudouris et al., 2012), which is given by $GAIC(\kappa) = -2\hat{l}(\boldsymbol{\theta}) + \kappa \times df$, where $\hat{l}(\boldsymbol{\theta})$ is the fitted log-likelihood (3), κ is the degree of penalty and df are the effective degrees of freedom of the fitted model. If $\kappa = 2$ or $\kappa = \log(n)$, GAIC reduces to the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978), respectively.

Looking specifically to the GAIC definition, we may note that the selection of a given variable will depend on how much such variable increases the log-likelihood function $\hat{l}(\boldsymbol{\theta})$ (consequently, decreasing the $GAIC(\kappa)$ value), weighted by the increasing on the number of degrees of freedom.

In this paper we will consider both AIC and BIC criteria in order to select the best fitted models. Furthermore, we will also consider in our simulation studies a third penalty which is given by the average of AIC and BIC penalties, i.e., $\kappa = [2 + \log(n)]/2$, denoted here as GAICav. This additional penalty is considered here since, as described in Hossain et al. (2016), AIC and BIC can lead to overfitting (undersmoothing) and underfitting (oversmoothing), respectively.

3 Simulation Studies

In this section we perform all simulation studies regarding the performance of the Strategy A approach to select explanatory variables for different regression structures in the GAMLSS framework. Three different response types that claim varied distributions with particular characteristics will be considered here (for further details regarding each of these distributions, please check Rigby et al., 2019):

- Continuous response: Gaussian distribution, i.e. $Y \sim N(\mu, \sigma)$, where $-\infty < y < \infty$, $-\infty < \mu < \infty$ is the mean and $\sigma > 0$ the standard deviation parameter;
- Zero inflated discrete response: zero inflated Poisson (ZIP) distribution, i.e. $Y \sim ZIP(\mu, \sigma)$, where $y = \{0, 1, \ldots\}$, $\mu > 0$ is the mean of the Poisson component and $0 < \sigma < 1$ is the exactly probability of $Y = 0$;
- Censored response: Weibull distribution, i.e. $Y \sim Wei(\mu, \sigma)$, where $Y > 0$, $\mu > 0$ is the mean and $\sigma > 0$ represents a scale parameter.

All link functions were chosen based on the range of each parameter (De Bastiani et al., 2018; Stasinopoulos et al., 2017). When θ_k is defined in the real support, the identity link function was used, when θ_k is positive the logarithm function was adopted and, finally, for θ_k on the unit interval, the logit link function was applied. It is noteworthy that in all scenarios we are using the default initial values in `stepGAICAll.A()` function in the *gamlss* package for the response variable distribution parameter vectors. As stated in Stasinopoulos et al. (2017), although we must use initial values for the distribution parameter vectors, the estimation algorithm is stable and fast using simple starting values (e.g. constants) for the parameter vectors. Nonetheless, users can easily set any different values as needed (Stasinopoulos and Rigby, 2007; Stasinopoulos et al., 2017).

Based on the three above mentioned distributions, we will consider two different main scenarios, one considering a linear structure (the simulated data consider only linear relationships between covariates and each of the distribution parameters) and one with both linear and nonlinear structures (nonlinear relationships are also considered). The sample sizes are generated by taking $n = 150$ and $n = 300$ and, for each scenario, all results are obtained from 1,000 Monte Carlo replications.

In all scenarios, eight different covariates will be used in the model selection process.

- Binary variable: $X_1 \sim X_2 \sim X_3 \sim X_4 \sim Bernoulli(0.5)$
- Continuous variable: $X_5 \sim X_6 \sim X_7 \sim X_8 \sim U(0, 1)$

Furthermore, different coefficient values associated to x_5 and x_6 were considered in μ and σ structures, respectively, while holding all other coefficients as constants. The same above mentioned data generation process was conducted.

The variables X_4 and X_8 will be considered as noise variables, i.e., although they are included in the model selection procedure, they will not be considered in the data-generating process. Moreover, different slope values (i.e. coefficient magnitude) are evaluated in the simulation study, as will be further highlighted. For each replication, the Strategy A method selects the model parameters, then, at the end of all replications, the percentage of correct/incorrect specification in each distribution parameter is calculated. A summary of the returned p-values for the selected variables in each scenario, considering the different used penalties, are displayed in Appendices A and B. It is noteworthy here that many recent works highlight the danger of using naive p-values after the model selection stage (Lee et al., 2016), however we are providing these values only to get a sense regarding the significant parameters.

3.1 Linear Structure

In the scenarios where only linear relationships are allowed, we have a discrete (X_3) and a continuous (X_7) variable affecting both μ and σ parameters simultaneously (Table 1). Moreover, variables X_1 and X_5 will only be considered in the generating process for μ, and variables X_2 and X_6 for parameter σ.

Normal Data Let us consider the random variable $Y \sim N(\mu, \sigma)$. Here we consider the following regression structures for the two parameters of the normal distribution

$$\mu = \beta_{01} + \beta_{11}x_1 + \beta_{31}x_3 + \beta_{51}x_5 + \beta_{71}x_7 \quad \text{and}$$
$$\sigma = \exp[\beta_{02} + \beta_{22}x_2 + \beta_{32}x_3 + \beta_{62}x_6 + \beta_{72}x_7],$$

where the true parameter values in the data-generating processes are

$$\mu = 40 + 5x_1 - 3x_3 + 2.5x_5 - 3x_7 \quad \text{and}$$
$$\sigma = \exp[1.6 + 0.6x_2 - 0.35x_3 + 0.03x_6 - 0.02x_7].$$

Note that this model (when both parameters are affected by explanatory variables) is also known in the literature as the heteroscedastic normal regression model.

Table 1: Continuous (cont) and discrete authentic variables considered on the data generating process (linear structure).

Parameter	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
μ	discrete		discrete		cont		cont	
σ		discrete	discrete			cont	cont	

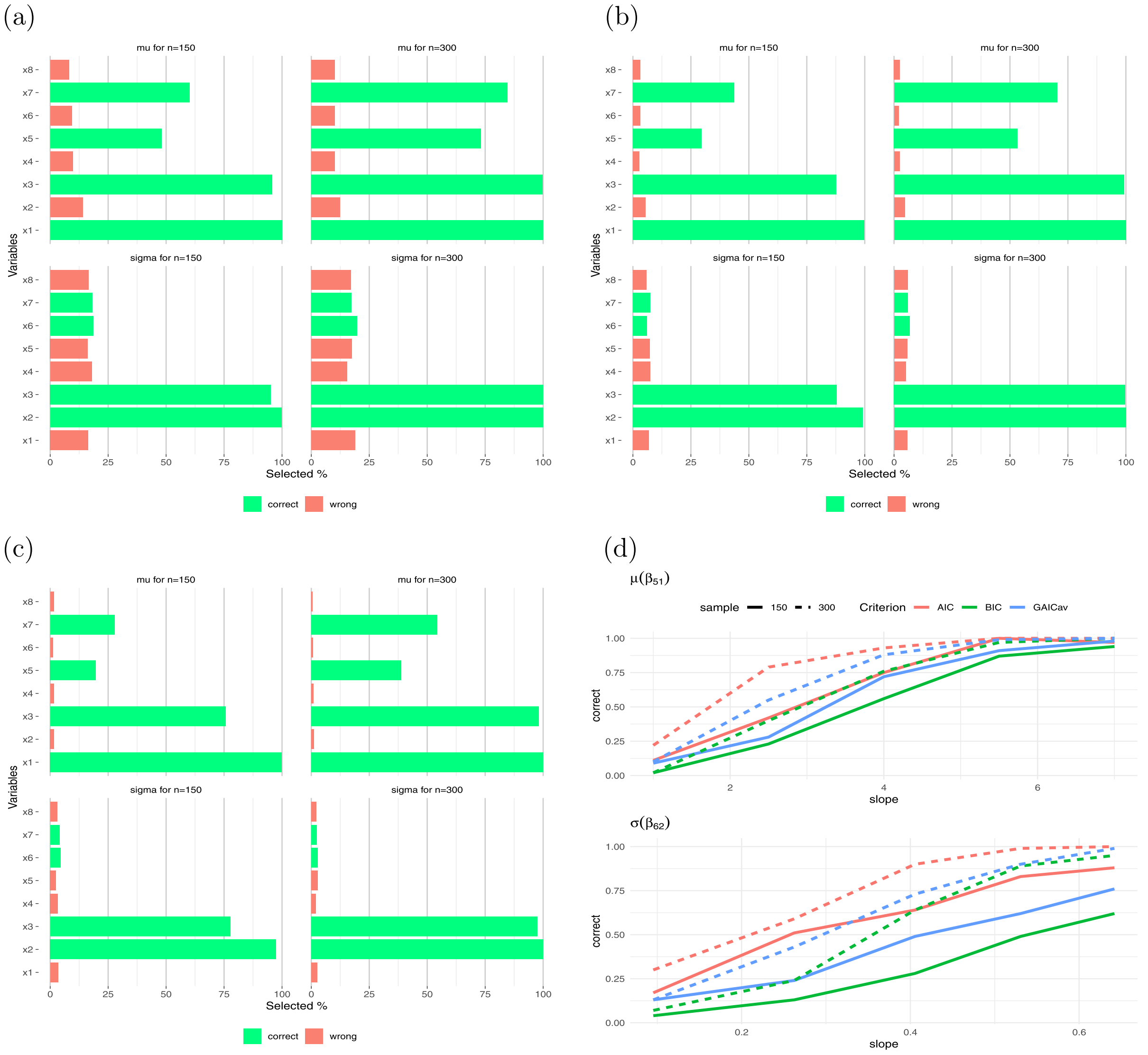

Figure 1: Selection of explanatory variables in the parametric normal heteroscedastic model. The correct/incorrect specification percentages, for different sample size, using (a) $\kappa = 2$ (AIC), (b) $\kappa = [2 + \log(n)]/2$ (GAICav) and (c) $\kappa = \log(n)$ (BIC) criteria, and (d) different values for slope.

Figure 1 displays the percentage of correct/incorrect variables selected for each of the model parameters through the Strategy A approach, based on the following goodness-of-fit measures: Panel (a) AIC, Panel (b) GAIC($k = [2 + \log(n)]/2$) and Panel (c) BIC. Further, Panel (d) displays the correct selection (using the three criteria) considering different values of slope.

We may conclude that, when compared to other variables, x_5 and x_7 presented the worst results for the mean μ. In its worst scenario (using BIC for $n = 150$ as can be seen in Panel (c)) the correct selection percentage of their selection were approximately 20% for x_5 and 27.5% for x_7. One possible explanation for this, as highlighted at the end of Section 2.1, is that the effect of continuous covariates in the log-likelihood function is smaller than the one exerted by a categorical variable (both with the same range in this example). Moreover, x_5 has the smallest coefficient.

A more serious problem arose when the procedure selected the set of explanatory variables for the standard deviation parameter σ, where x_6 and (especially) x_7 were not selected very often. Again, this problem may be explained by the effects caused by these coefficients associated to these variables (this statement is true only if both continuous and categorical variables are within the same range and their associated coefficients are similar).

Regarding Figure 1(d), we may conclude that, greater values of slope imply a greater correct selection rate. Moreover, as expected, for a greater sample size, Strategy A will perform better, i.e., it will correctly select the given variable. Also, as we increase penalty κ, more rigorous is the selection of these variables.

Analyzing the p-values for the selected variables for different settings (see Supplementary Material), we may conclude that BIC criterion results in the smallest p-values, followed by GAICav and AIC criteria. As expected, when the sample size increases, the variables, when selected, become more significant. Looking at x_6 and x_7, for sigma parameter, we may note that, even for the smallest penalty (AIC), in more than 50% of the times they were correctly selected, they were not significant at a 5% level, due to their low coefficient values and also because they are continuous variables.

Count Data Let us consider the random variable $Y \sim ZIP(\mu, \sigma)$. The regression structures used to generate the data were

$$\mu = \exp[0.4 + 0.4x_1 + 0.04x_3 + 0.04x_5 + 0.05x_7] \quad \text{and}$$
$$\sigma = \text{logistic}[-2.11 + 0.75x_2 + 1.85x_3 + 0.50x_6 + 0.63x_7].$$

Figure 2 shows the percentage of selected covariates in each of the scenarios for the ZIP model. Panels (a, b, c) show that in all cases, the percentage of authentic variables that were in fact selected for the mean of the Poisson component μ (i.e., it should be included in the model for μ) using the Strategy A procedure is relatively low, except for x_1. However, this does not mean that there is a problem in the considered model, but actually this might be explained by the low coefficient value associated to x_3 (0.04) compared to the coefficient associated to x_1 (0.4) and, as explained in the previous scenario (normal data), the effect of continuous covariates (x_5 and x_7) since they are in the same range of the authentic discrete variables. Further, Panel (d) displays the correct selection rate for different values of slope, considering all other variables as fixed.

Regarding the probability of zero σ, all scenarios present a similar behavior verified in the normal data case, since the continuous variables (x_6 and x_7) were poorly selected, especially when BIC criterion was considered. Noise variables were selected very few times in the simulation study. In the worst scenario (considering AIC value and $n = 150$ as the selection criterion), noise variables were selected less than 20% of the times. All p-values related to this case are presented in Supplementary Material.

Censored Data Here, we focus on censored samples, characteristic found in survival analysis data. The regression structure considered here is based on the accelerated lifetime Weibull model,

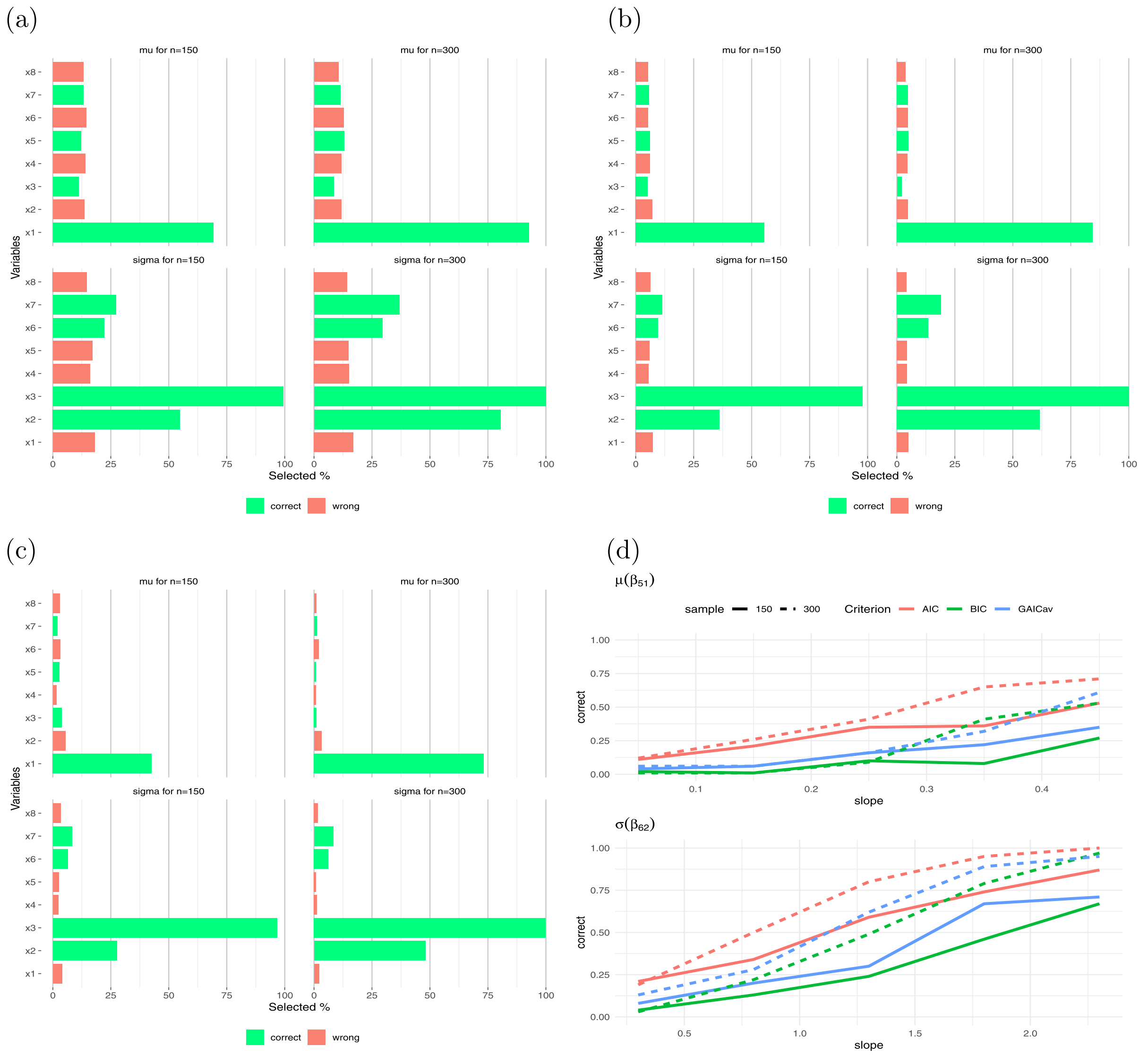

Figure 2: Selection of explanatory variables in the parametric ZIP model. The correct/incorrect specification percentages, for different sample size, using (a) $\kappa = 2$ (AIC), (b) $\kappa = [2 + \log(n)]/2$ (GAICav) and (c) $\kappa = \log(n)$ (BIC) criteria, and (d) different values for slope.

i.e. $Y \sim Wei(\mu, \sigma)$, and is given by

$$\mu = \exp[3.85 + 0.06x_1 + 0.09x_3 + 0.04x_5 + 0.08x_7] \quad \text{and}$$
$$\sigma = \exp[1.54 + 0.57x_2 - 0.27x_3 + 0.05x_6 - 0.03x_7].$$

Results of this simulation study are presented in Figure 3. The percentage of authentic explanatory variables selected to compose the regression structure for the mean parameter μ and σ present the same behavior of the previous cases, i.e. high selection rate for discrete variables and low percentage for continuous ones. Also, as expected, AIC criterion presents a greater correct selection rate (Panel (d)). Finally, the returned p-values of these scenarios are presented in Supplementary Material.

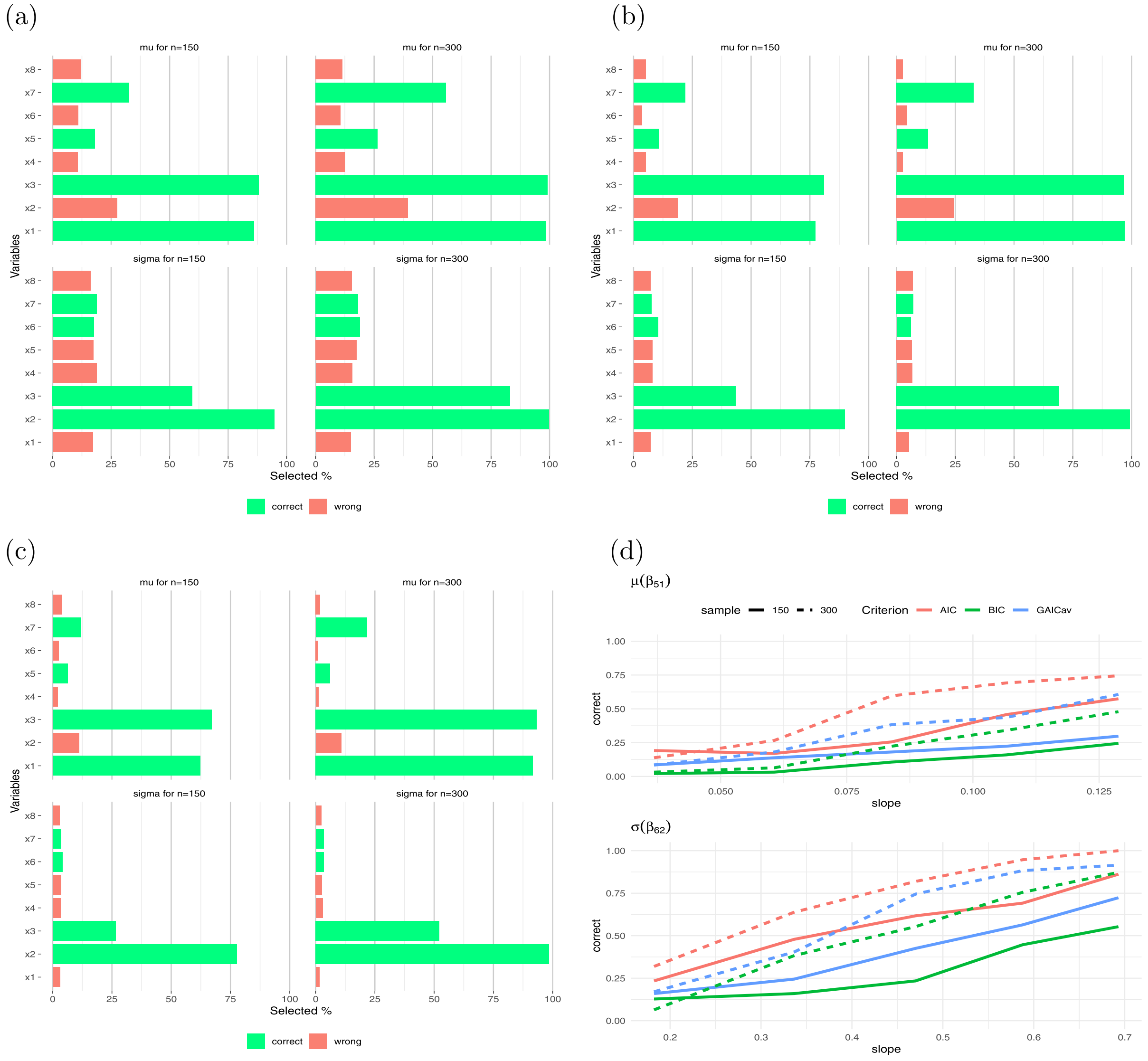

Figure 3: Selection of explanatory variables in the parametric Weibull model, for different sample size, using (a) $\kappa = 2$ (AIC), (b) $\kappa = [2 + \log(n)]/2$ (GAICav) and (c) $\kappa = \log(n)$ (BIC) criteria, and (d) different values for slope.

3.2 Linear and Nonlinear Structures

In the scenarios where both linear and nonlinear relationships are considered, we have a discrete variable (X_3) affecting both μ and σ parameters simultaneously (Table 2). Variables X_1, X_5 and X_7 will only be considered in the generating process for μ, and variables X_2 and X_6 for parameter σ. Moreover, X_7 will be generated in such a way that its effect in μ has an increasing-decreasing-increasing shape.

Normal Data Let us consider once again that $Y \sim N(\mu, \sigma)$. Now, assuming that X_7 has a nonlinear effect in μ, we will consider the following regression structures for the two parameters

Table 2: Continuous (cont) and discrete authentic variables considered on the data generating process (linear and nonlinear structures).

Parameter	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
μ	discrete		discrete		cont		cont	
σ		discrete	discrete			cont		

of the normal distribution

$$\mu = \beta_{01} + \beta_{11}x_1 + \beta_{31}x_3 + \beta_{51}x_5 + s(x_7) \quad \text{and}$$
$$\sigma = \exp[\beta_{02} + \beta_{22}x_2 + \beta_{32}x_3 + \beta_{62}x_6],$$

where $s(\cdot)$ stands for a P-spline (Eilers and Marx, 1996) considered to model the relationship between x_7 and the mean parameter μ. The true parameter values in the data-generating processes are

$$\mu = 40 + 5x_1 - 4x_3 + 2.5x_5 + 10\sin(0.2x_7\pi) \quad \text{and}$$
$$\sigma = \exp[1.6 + 0.6x_2 - 0.35x_3 + 0.03x_6].$$

As we can see in Figure 4, the correct selection rate for both parameters (μ and σ), including different values of slope (Panel (d)), of the heteroscedastic normal regression model, based on the Strategy A method, returned similar results (consequently, the same discussion can be applied here) to the ones presented when only linear structures were considered. Once again, as highlighted in Section 2.1, the effect of continuous covariates in the log-likelihood function is smaller than the one applied by a categorical variable, when they are within the same range.

Please check Supplementary Material for the p-values of the selected (authentic and noise) variables. Please note that the covariate x_7 is omitted from these plots since a P-spline is being considered to model its relationship with both parameters μ (authentic) and σ (noise), thus the resulting coefficients of each smoother and its standard error (consequently its p-value as well) refer only to the linear part of the smoother and not to the smoother's contribution as a whole (Stasinopoulos et al., 2017; Ramires et al., 2019). As in the linear structure scenario, we may conclude that BIC criterion returned the smallest p-values and, as expected, when the sample size increases, the variables, when selected, become more significant.

Finally, regarding noise variables, in the worst case scenario, variable x_7 was roughly selected more than 25% of the times to compose the regression structure of σ (considering AIC with a sample size equals to 150).

Count Data Considering $Y \sim ZIP(\mu, \sigma)$ and that x_7 has a nonlinear effect in μ, the following regression structures were used to generate the data

$$\mu = \exp[0.4 + 0.4x_1 + 0.04x_3 + 0.04x_5 + 0.2\sin(0.2\pi\, x_7)] \quad \text{and}$$
$$\sigma = \text{logistic}[-2.11 + 0.75x_2 + 1.85x_3 + 0.5x_6].$$

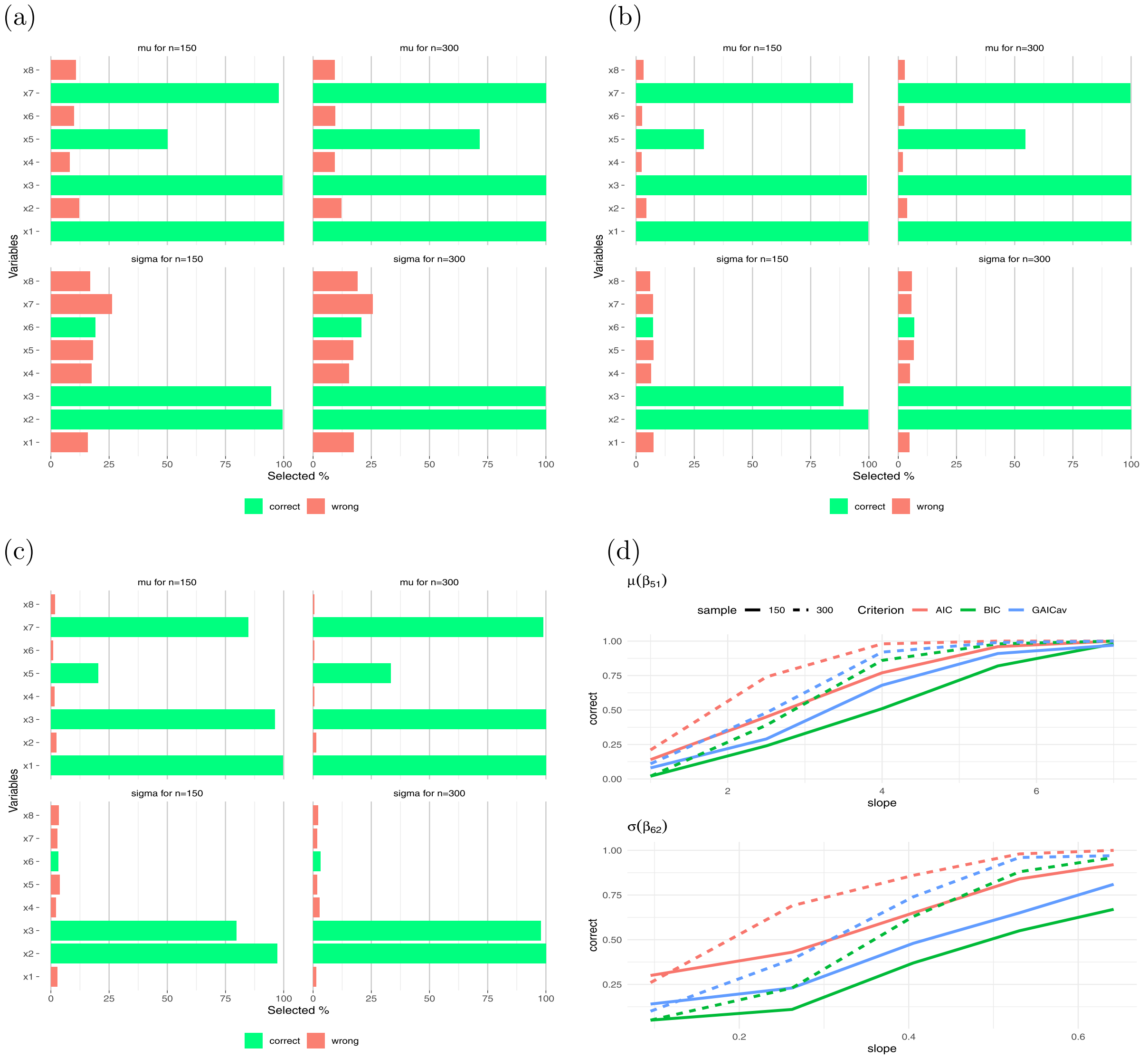

Figure 4: Selection of explanatory variables in the semiparametric normal heteroscedastic model. The correct/incorrect specification percentages, for different sample size, using (a) $\kappa = 2$ (AIC), (b) $\kappa = [2 + \log(n)]/2$ (GAICav) and (c) $\kappa = \log(n)$ (BIC) criteria, and (d) different values for slope.

Figure 5 displays the percentages of selected authentic and noise variables in this simulation study. The observed behavior for both parameters μ (mean of the Poisson component) and σ (exactly probability of $Y = 0$) is quite similar to the one in the scenario where only a linear structure was considered in the ZIP model, presented in Figure 2. All p-values from this simulation study are presented in Supplementary Material. Please note, once again, that covariate x_7 is omitted due to the P-spline considered to model its relationship in both regression structures.

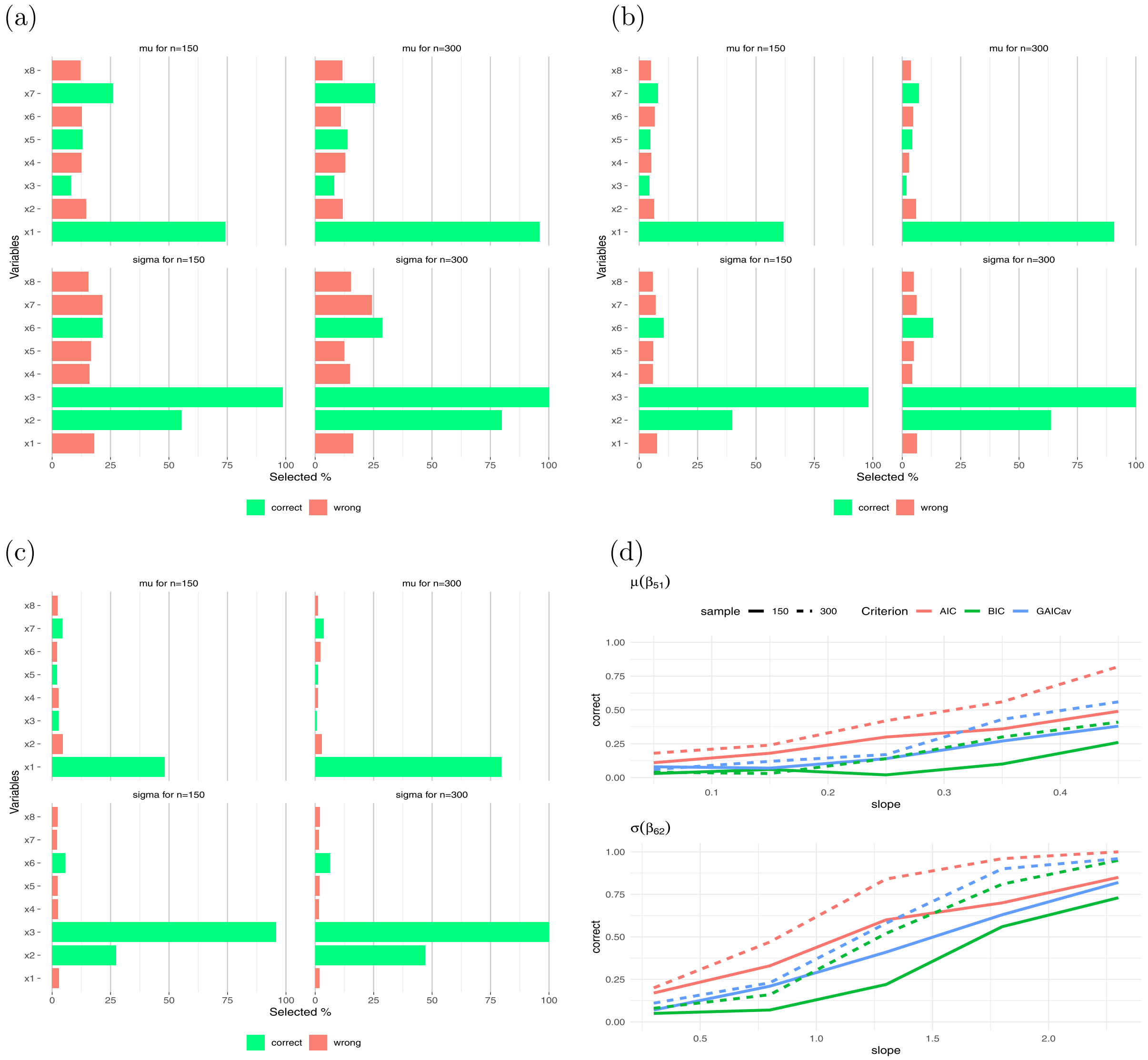

Figure 5: Selection of explanatory variables in the semiparametric ZIP model, for different sample size, using (a) $\kappa = 2$ (AIC), (b) $\kappa = [2 + \log(n)]/2$ (GAICav) and (c) $\kappa = \log(n)$ (BIC) criteria, and (d) different values for slope.

Censored Data In this final simulation study, let us consider $Y \sim Wei(\mu, \sigma)$ and a nonlinear effect in μ. The following regression structures were used to generate the data

$$\mu = \exp\left[3.98 + 0.09x_1 - 0.07x_3 + 0.04x_5 + \log\left(1 + 10\sin(0.2\pi\, x_7)/40\right)\right] \quad \text{and}$$
$$\sigma = \exp[1.53 + 0.49x_2 - 0.31x_3 + 0.05x_6].$$

The results are presented in Figure 6. As can be seen the observed percentages of variable selection in all scenarios are quite similar to the one presented in Figure 3, i.e. a high selection rate for authentic discrete variables, while the AIC criterion presented the greatest correct selection rate (Panel (d)). All p-values of this simulation study are presented in Supplementary Material.

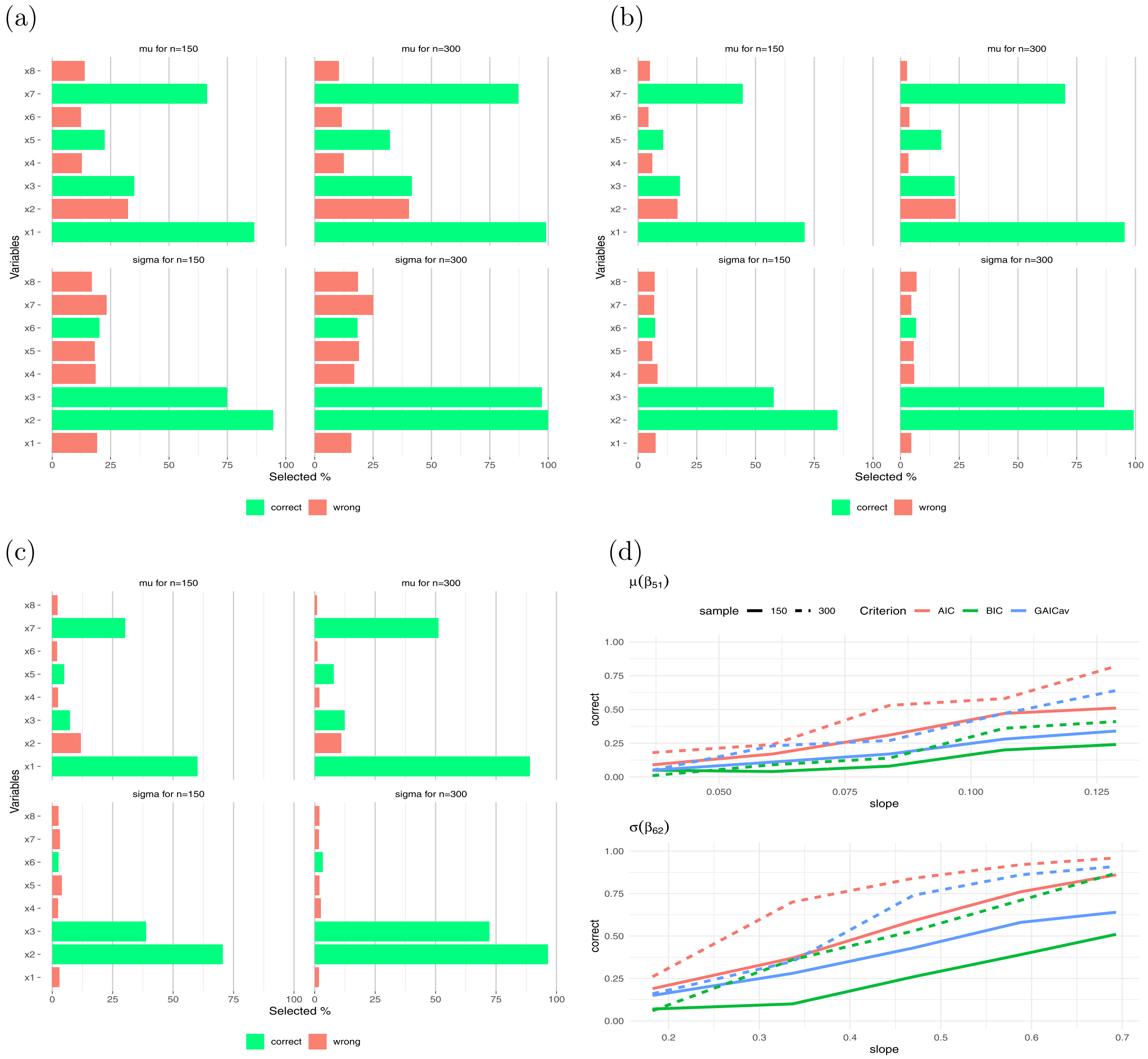

Figure 6: Selection of explanatory variables in the semiparametric Weibull model, for different sample size, using (a) $k = 2$ (AIC), (b) $k = [2 + \log(n)]/2$ (GAICav) and (c) $k = \log(n)$ (BIC) criteria, and (d) different values for slope.

4 Concluding Remarks

The Strategy A procedure, accessed through the `stepGAICAll.A()` function implemented in the *gamlss* package in R is a stepwise-based procedure to select variables based on the generalized Akaike information criterion (GAIC). Due to the GAIC definition, the selection of a given variable will directly depend on how much such variable increases the log-likelihood function (decreasing the GAIC value), weighted by the increasing on the number of degrees of freedom. Thus, continuous covariates tended to be selected fewer times than discrete ones if they have some similarity. Nonetheless, three levels of penalties κ were considered, showing that the AIC criterion tends to include more variables in the model (including those that should not be selected, i.e. noise variables), BIC tends to select only covariates with strong effects and GAICav was a moderate criterion, which is somewhat expected. Furthermore, as sample size increases, the results of the stepwise-based procedure become better and more realistic. Finally, as suggestions for future research, we can think in the behavior of the Strategy A procedure considering more complex distributions (three or more parameters), different shapes for the nonlinear term, the presence of interactions between covariates and others.

Supplementary Material

Please note that the following supplementary files are available online: i) `suppl_stepgaic.pdf`: p-values for the selected variables in each simulated scenario; and ii) `codes_stepgaic.zip`: all codes in R software that were used to conduct the simulation studies presented in this paper.

References

Akaike H (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19: 716–723.

Ayuso SV, Oñatibia GR, Maestre FT, Yahdjian L (2020). Grazing pressure interacts with aridity to determine the development and diversity of biological soil crusts in Patagonian rangelands. *Land Degradation & Development*, 31: 488–499.

Cole TJ, Green PJ (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, 11: 1305–1319.

De Bastiani F, Rigby RA, Stasinopoulos DM, Cysneiros AHMA, Uribe-Opazo M (2018). Gaussian Markov random field spatial models in gamlss. *Journal of Applied Statistics*, 45: 168–186.

Eilers PH, Marx BD (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11: 89–121.

Eilers PH, Marx BD, Durbán M (2015). Twenty years of p-splines. *SORT*, 39: 149–186.

Hastie TJ, Tibshirani RJ (1990). *Generalized Additive Models*. Chapman and Hall/CRC.

Hofner B, Mayr A, Schmid M (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, 74: 1–31.

Hossain A, Rigby RA, Stasinopoulos DM, Enea M (2016). Centile estimation for a proportion response variable. *Statistics in Medicine*, 35: 859–904.

Kneib T (2013). Beyond mean regression. *Statistical Modelling*, 13: 275–303.

Lee JD, Sun DL, Sun Y, Taylor J (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44: 907–927.

Lee Y, Nelder JA, Pawitan Y (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman and Hall/CRC.

Leroy B, Peatman T, Usu T, Caillot S, Moore B, Williams A, et al. (2016). Interactions between artisanal and industrial tuna fisheries: Insights from a decade of tagging experiments. *Marine Policy*, 65: 11–19.

Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012). Generalized additive models for location, scale and shape for high dimensional data – a flexible approach based on boosting. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 61: 403–427.

Nakamura LR, Cerqueira PHR, Ramires TG, Pescim RR, Rigby RA, Stasinopoulos DM (2019). A new continuous distribution on the unit interval applied to modelling the points ratio of football teams. *Journal of Applied Statistics*, 46: 416–431.

Nakamura LR, Rigby RA, Stasinopoulos DM, Leandro RA, Villegas C, Pescim RR (2017). Modelling location, scale and shape parameters of the Birnbaum-Saunders generalized t distribution. *Journal of Data Science*, 15: 221–237.

Nelder JA, Wedderburn RWM (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A. General*, 135: 370–384.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramires TG, Nakamura LR, Righetto AJ, Ortega EMM, Cordeiro GM (2018). Predicting survival function and identifying associated factors in patients with renal insufficiency in the metropolitan area of Maringá, Paraná state, Brazil. *Cadernos de Saúde Pública*, 34: 1–13.

Ramires TG, Nakamura LR, Righetto RR, Pescim AJ, Mazucheli J, Cordeiro GM (2019). A new semiparametric Weibull cure rate model: Fitting different behaviors within GAMLSS. *Journal of Applied Statistics*, 46: 2744–2760.

Rigby RA, Stasinopoulos DM (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 54: 507–554.

Rigby RA, Stasinopoulos DM (2014). Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research*, 23: 318–332.

Rigby RA, Stasinopoulos DM, Heller GZ, De Bastiani F (2019). *Distributions for Modeling Location, Scale and Shape: Using GAMLSS in R.* Chapman and Hall/CRC.

Righetto AJ, Ramires TG, Nakamura L, Castanho PLDB, Faes C, Savian TV (2019). Predicting weed invasion in a sugarcane cultivar using multispectral image. *Journal of Applied Statistics*, 46: 1–12.

Schwarz G (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464.

Stanton J (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9: 1–13.

Stasinopoulos DM, Rigby RA (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23: 1–46.

Stasinopoulos DM, Rigby RA, De Bastiani F (2018). GAMLSS: A distributional regression approach. *Statistical Modelling*, 18: 1–26.

Stasinopoulos DM, Rigby RA, Heller GZ, Voudouris V, De Bastiani F (2017). *Flexible Regression and Smoothing: Using GAMLSS in R.* Chapman and Hall/CRC.

Voncken L, Albers CJ, Timmerman ME (2019). Model selection in continuous test norming with GAMLSS. *Assessment*, 26: 1329–1346.

Voudouris V, Gilchrist R, Rigby R, Sedwick J, Stasinopoulos DM (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, 39: 1279–1293.