

Comparison of Feature Selection Methods in Regression Modeling: A Simulation Study

Vera Afreixo(✉), Jorge Cabral, and Pedro Macedo

CIDMA – Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
{vera,jorgecabral,pmacedo}@ua.pt

Abstract. This simulation study explores the impact of different undesirable scenarios (e.g., collinearity, Simpson's paradox, variable interaction, Freedman's paradox) on feature selection and coefficients' estimation using traditional methodologies, such as automatic selection (e.g., stepwise using Akaike information criterion and Bayesian information criterion) and penalized regression (e.g., least absolute shrinkage and selection operator (LASSO), elastic net, relaxed LASSO, adaptive LASSO, minimax concave penalty and smoothly clipped absolute deviation penalty, penalized regression with second-generation p-values). Specifically, we compare wrapper and embedded methods regarding the feature selection, coefficients' estimation and models' performance. Our results show that the choice of the methodology can affect the number and the type of selected features, as well as accuracy and precision of coefficients' estimates. Furthermore, we find that the performance can also depend on the characteristics of the data.

Keywords: Simulation · Feature Selection · Coefficient Estimation

1 Introduction

Feature selection and coefficients' estimation are two fundamental topics in statistical modeling, particularly in regression analysis. Feature selection is the process of selecting a subset of the most important features (i.e., variables) that have a significant impact on the outcome variable, while ignoring irrelevant or redundant variables. The feature selection contributes to less complex models and to reduce the potential overfitting, in addition to promoting estimation stability. There are several feature selection methodologies that can be used, including: filter methods (e.g., association measures or test, information gain); wrapper methods (stepwise linear models); embedded methods (penalized regression models, extreme gradient boosting, random forest); genetic algorithms. There is no one-size-fits-all approach to feature selection. The choice of the method may depend on the type of data, the specific problem, and the outcome variable.

Automatic selection, as the name itself indicates, is a family of techniques that automatically select a subset of predictors to include in a regression model. These techniques aim to improve the adequacy of the regression model by selecting the most relevant predictors and removing irrelevant or redundant ones. In some way, they can be considered

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14112, pp. 150–159, 2023.
https://doi.org/10.1007/978-3-031-37129-5_13

traditional techniques, based on t tests and F tests, with some of their weaknesses being well-known, such as the computational burden and the possible incorrect identification of relevant predictors, usually known as the Freedman's paradox (Freedman 1983).

Penalized regression methods are a class of statistical techniques that are commonly used in regression analysis to address problems such as overfitting and collinearity (Belsley et al. 2004). These methods add a penalty term to the objective function being minimized, which promotes simpler models and more parsimonious solutions, where some coefficients are shrunk towards zero depending on the norm (or combination of norms) used. These methods have become increasingly popular in recent years, especially in high-dimensional data settings where the number of predictors is much larger than the number of observations. These methods can help to improve the predictive performance of models while also providing insight into the most important predictors (Hastie et al. 2009; Efron and Hastie 2016).

2 Methods

We consider a simulation study with different scenarios to demonstrate the performance of nine feature selection methods: stepwise selection of variables (Efroymson 1960; Hocking 1976) based on the Akaike information criterion (StepAIC) (Akaike 1974) and on the Bayesian information criterion (StepBIC) (Schwarz 1978), least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), adaptive LASSO (Zou 2006), relaxed LASSO (Hastie et al. 2017), elastic net (ENet) (Zou and Hastie 2005), smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), minimax concave penalty (MCP) (Zhang 2010), and second-generation p-values (ProSGPV) (Zuo et al. 2022).

We consider a number of simulation repetitions of $N = 50$ and different sample sizes of $n \in \{30, 100\}$. Continuous covariates (p) are drawn from a standard normal distribution considering $p \in \{5, 10, 15\}$, and random correlation matrices are generated based on the method proposed by Joe (2006), using values of $d \in \{0.1, 1, 100\}$ in order to generate different $n \times p$ matrices X. We firstly generate $Z = X\beta + \epsilon$, where ϵ is assumed to follow a normal distribution with a zero-mean and two different standard deviation (sd_ϵ), namely $sd_\epsilon \in \{1, 3\}$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of the true regression coefficients, where $(\beta_{p-4}, \beta_{p-3}, \ldots, \beta_p) \in \{(0, 0, 0, 0, 0), (1, -1, 1, -1, 1), (2, 2, 2, 2, 2), (2, 4, 6, 8, 10)\}$. Finally, when $p \in \{10, 15\}$, then $\beta_1 = \beta_2 = \ldots = \beta_{p-5} = 0$.

In scenarios where the Simpson's paradox (Simpson 1951) is present, Y is given by $Y = V_S + Z$, where V_S is defined by $V_S = \beta_{0S} + X_S\beta_{1S} + B_S\beta_{2S}$, X_S with a standard normal distribution, $N(0, 1)$, and B_S with a balanced Bernoulli distribution, $B(1, 0.5)$. V_S and X_S have a trend that disappear or reverses in the presence of B_S. We consider differences of means equal to 1 and 4 between the groups created by B_S in both V_S and X_S (Makowski et al. 2019).

In scenarios where the interaction between variables is present, each $Y_i \in Y$ is given by

$$Y_i = \begin{cases} 2 \times X_{Ii} + Z_i, \text{if } B_{Ii} = 0 \\ -2 \times X_{Ii} + Z_i, \text{if } B_{Ii} = 1 \end{cases},$$

with $i = 1, 2, \ldots, n$, $X_I \sim N(0, 1)$ and $B_I \sim B(1, 0.5)$. In the absence of either interaction or Simpson's paradox, then $Y = Z$.

The feature selection methods will be tested under different scenarios, namely: (1) in the absence of one of the variables (e.g., when $(\beta_{p-4}, \beta_{p-3}, \ldots, \beta_p) = (2,\ 4,\ 6,\ 8,\ 10)$ the methods will be evaluated in the presence of all the variables, in the absence of the variable with coefficient 2, and in the absence of the variable with coefficient 10); (2) the presence and absence of the variables implied in the Simpson's paradox; and (3) the variables interaction.

The 5-fold cross-validation root mean square error (CV-RMSE) will be determined for each scenario, method and repetition. Analyses will be performed in RStudio Version 2022.07.2+576 (Posit Team 2022) running R version 4.1.3 (R Core Team 2022). The R packages *glmnet* (Friedman et al. 2010), *MASS* (Venables and Ripley 2022), *ProSGPV* (Zuo et al. 2022a), *simstudy* (Goldfeld and Wujciak-Jens 2020), *clusterGeneration* (Qiu and Joe 2020) and *bayestestR* (Makowski et al. 2019) will be used. To organize the results, we develop a Shiny application that allows the visualization of the number of selected features and by each feature the percentage of selection, the coefficients' estimates, and the methods' performance through CV-RMSE (see the Appendix for two illustrations).

3 Results and Discussion

3.1 Only Noise Variables (Freedman's Paradox)

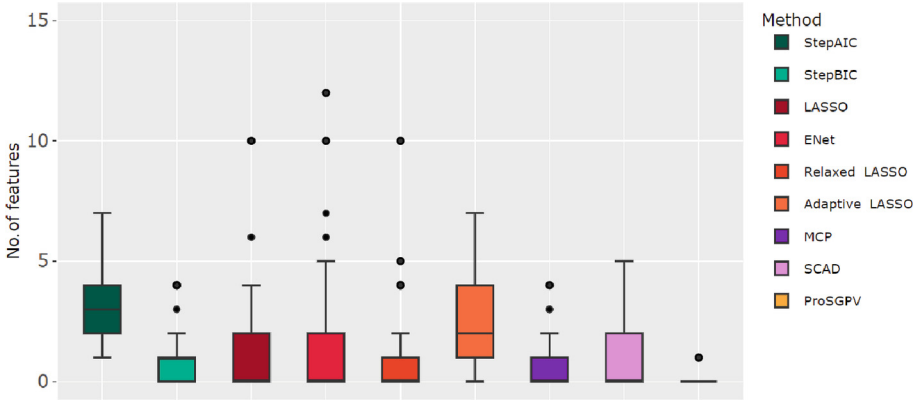

Fig. 1. Number of selected features by the methods under analysis in a complete noise scenario ($n = 100, p = 15, \beta = (0, \ldots, 0), sd_\epsilon = 1$) without relevant collinearity ($d = 100$).

In the scenario that only includes noise variables (i.e., only the variables with coefficients equal to zero), we observed that the more variables are included, the more variables tend to be selected. However, although several methods do not select any variable (in median), StepAIC, StepBIC and adaptive LASSO show a higher median number of selected features. ProSGPV is the method with the best overall performance and StepAIC the worst (see Fig. 1).

Adding collinearity to the models, the estimates obtained by the automatic selection methods are substantially different (and higher than) from the ones of the true regression coefficient vector (see Table 1).

Table 1. Descriptive analysis of the estimates (median [minimum, maximum]) by the methods under analysis in a complete noise scenario ($n = 100$, $p = 15$, $\beta = (0, \ldots, 0)$, $sd_\epsilon = 1$) with relevant collinearity ($d = 0.1$).

	StepAIC	StepBIC	LASSO	ENet	Relaxed LASSO	Adaptive LASSO	MCP	SCAD	ProSGPV
$\hat{\beta}_{15}$	0 [−6298.6, 7846.6]	0 [−4869.5, 2]	0 [−0.2, 0.1]	0 [−0.2, 0.1]	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.1]	0 [−0.3, 0.1]	0 [−0.4, 0]
$\hat{\beta}_{14}$	0 [−7916.1, 5399.7]	0 [−7916.1, 0]	0 [−0.1, 0.1]	0 [−0.1, 0.1]	0 [−0.2, 0.1]	0 [−0.2, 0.1]	0 [−0.2, 0]	0 [−0.3, 0]	0 [0, 0]
$\hat{\beta}_{13}$	0 [−13135.5, 11891.2]	0 [−684.4, 90.1]	0 [0, 0.3]	0 [0, 0.3]	0 [0, 0.3]	0 [−0.1, 0.3]	0 [0, 0.4]	0 [0, 0.4]	0 [0, 0]
$\hat{\beta}_{12}$	0 [−10191.6, 1419.3]	0 [−1.3, 1419.3]	0 [−0.1, 0.1]	0 [−0.1, 0.1]	0 [−0.1, 0.2]	0 [−0.2, 0.3]	0 [0, 0.2]	0 [−0.1, 0.1]	0 [0, 0.3]
$\hat{\beta}_{11}$	0 [−28.1, 25866.4]	0 [−1.4, 442.9]	0 [−0.3, 0.1]	0 [−0.3, 0.1]	0 [−0.3, 0.1]	0 [−0.3, 0.3]	0 [−0.4, 0.1]	0 [−0.4, 0.1]	0 [0, 0.2]
$\hat{\beta}_{10}$	0 [−5472.4, 7120.8]	0 [−370.8, 2077.9]	0 [−0.1, 0.2]	0 [−0.1, 0.2]	0 [−0.2, 0.3]	0 [−0.2, 0.4]	0 [0, 0.4]	0 [−0.1, 0.4]	0 [0, 0]
$\hat{\beta}_{9}$	0 [−3322.1, 13299.5]	0 [0, 5229.2]	0 [−0.1, 0]	0 [−0.1, 0]	0 [−0.1, 0]	0 [−0.2, 0.1]	0 [0, 0]	0 [0, 0]	0 [0, 0]
$\hat{\beta}_{8}$	0 [−8837, 3451.3]	0 [−8837, 0.5]	0 [0, 0]	0 [0, 0]	0 [0, 0.1]	0 [−0.2, 0.2]	0 [0, 0]	0 [0, 0]	0 [0, 0]
$\hat{\beta}_{7}$	0 [−13420.5, 2469.5]	0 [−1.3, 1419.4]	0 [−0.1, 0]	0 [−0.1, 0.1]	0 [−0.2, 0]	0 [−0.2, 0.2]	0 [−0.3, 0]	0 [−0.3, 0]	0 [0, 0]
$\hat{\beta}_{6}$	0 [−23867.7, 5269.8]	0 [−4178, 0.3]	0 [−0.1, 0.1]	0 [−0.1, 0.1]	0 [−0.2, 0.1]	0 [−0.2, 0.2]	0 [−0.2, 0.1]	0 [−0.2, 0.1]	0 [0, 0]
$\hat{\beta}_{5}$	0 [−6322.6, 3659.5]	0 [−6322.6, 194.3]	0 [−0.1, 0.1]	0 [−0.1, 0.1]	0 [−0.1, 0.1]	0 [−0.2, 0.2]	0 [−0.1, 0.1]	0 [−0.1, 0.1]	0 [0, 0]

(*continued*)

Table 1. (*continued*)

	StepAIC	StepBIC	LASSO	ENet	Relaxed LASSO	Adaptive LASSO	MCP	SCAD	ProSGPV
$\hat{\beta}_4$	0 [−2120.3, 9488.3]	0 [−1.4, 9488.3]	0 [0, 0.1]	0 [0, 0.1]	0 [−0.1, 0.1]	0 [−0.1, 0.2]	0 [0, 0]	0 [0, 0]	0 [0, 0]
$\hat{\beta}_3$	0 [−21954.6, 3774.7]	0 [−186.8, 3774.7]	0 [−0.1, 0]	0 [−0.1, 0]	0 [−0.2, 0]	0 [−0.2, 0]	0 [−0.1, 0]	0 [−0.1, 0]	0 [0, 0]
$\hat{\beta}_2$	0 [−42352.6, 3398]	0 [−2193.7, 145.2]	0 [0, 0]	0 [0, 0.1]	0 [0, 0.2]	0 [−0.2, 0.3]	0 [0, 0]	0 [0, 0.1]	0 [0, 0]
$\hat{\beta}_1$	0 [−19749.4, 5098.8]	0 [−5985.8, 45.1]	0 [0, 0]	0 [0, 0]	0 [0, 0]	0 [−0.1, 0.2]	0 [0, 0]	0 [0, 0]	0 [0, 0]
$\hat{\beta}_0$	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.2]	0 [−0.3, 0.2]

3.2 Five Explanatory Variables to Define the Outcome

When we have five explanatory variables (i.e., $\left(\beta_{p-4}, \beta_{p-3}, \ldots, \beta_p\right) \neq (0, 0, 0, 0, 0)$) to define the outcome (Y) without a relevant collinearity structure and without noise variables available (i.e., $p = 5$), we observed that, in general, the higher the magnitude of the coefficients, the easier the selection of variables and the better the estimation of coefficients, and the higher sd_ϵ the worse the performance of the methods under study (selection and estimation).

For the scenario with $\left(\beta_{p-4}, \beta_{p-3}, \ldots, \beta_p\right) = (1,\ -1,\ 1,\ -1,\ 1)$ with $sd_\epsilon = 3$, in general, the methods cannot select the five features (see Fig. 2). The ProSGPV presents in some cases the worst performance, selecting a smaller number of variables (less than five). For the scenario with the smallest standard deviation of the error ($sd_\epsilon = 1$), in general, all the methods present an adequate performance.

For the scenario with $\left(\beta_{p-4}, \beta_{p-3}, \ldots, \beta_p\right) = (2,\ 4,\ 6,\ 8,\ 10)$ and with $sd_\epsilon = 1$, all methods select the five explanatory variables in the different simulation scenarios. In this scenario we added noise variables (5 or 10) and we clearly observed that some methods select more than five variables (StepAIC, StepBIC, LASSO and ENet). It is observed that all methods present at least one scenario of exaggerated selection, although they are all consistent in the selection of variables of interest. Best results are obtained by adaptive LASSO, MCP, SCAD and ProSGPV (see Fig. 3).

When five explanatory variables (scenario with $\left(\beta_{p-4}, \beta_{p-3}, \ldots, \beta_p\right) = (2,\ 4,\ 6,\ 8,\ 10)$ and $sd_\epsilon = 1$) are generating the outcome with a relevant collinearity structure, without adding noise variables ($p = 5$), all methods tend to select fewer variables than expected (median = 4 for all methods except ENet).

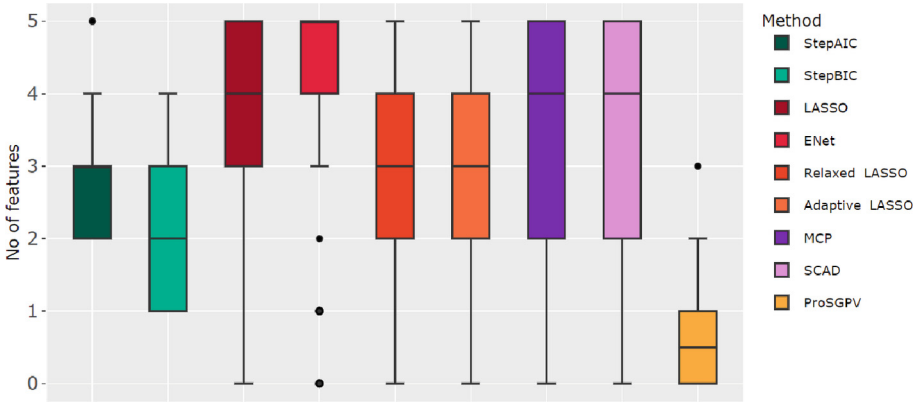

Fig. 2. Number of selected features by the methods under analysis for the scenario of five explanatory variables to define the outcome ($n = 30, p = 5, \beta = (1, -1, 1, -1, 1), sd_\epsilon = 3$) and without relevant collinearity ($d = 100$).

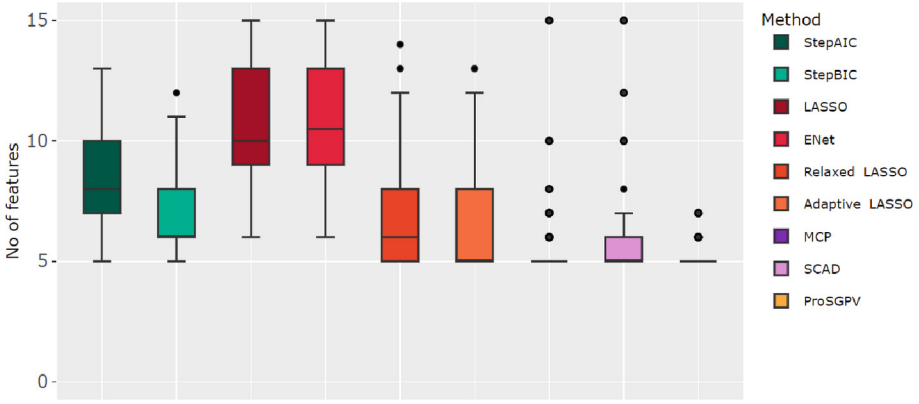

Fig. 3. Number of selected features by the methods under analysis for the scenario of five explanatory variables to define the outcome ($n = 30, p = 15, \left(\beta_{p-4}, \beta_{p-3}, \ldots, \beta_p\right) = (2, 4, 6, 8, 10)$, $sd_\epsilon = 1$) and without relevant collinearity ($d = 100$).

Again, it is observed that, in the presence of collinearity, the automatic selection methods can provide unstable estimates, with quite different values and with opposite signs when compared to the real coefficients (the range of values in the different simulations is very high). The methods' performance (measured through CV-RMSE) is better with ENet and worse with both automatic selection methods. Adding noise variables to the previous scenario, the best methods regarding feature selection are MCP, SCAD and ProSGPV. The well-known problems of the automatic selection methods are highlighted here, however the CV-RMSE of the different methods regarding the generation model are similar.

Table 2. Descriptive analysis of the estimates (median [minimum, maximum]) by the methods under study using models with five explanatory variables and with relevant collinearity. top) estimates for the model with the five explanatory variables; bottom) estimates for the model removing one of the five explanatory variables.

	StepAIC	StepBIC	LASSO	ENet	Relaxed LASSO	Adaptive LASSO	MCP	SCAD	ProSGPV
$\hat{\beta}_5$	1.7 [−35756.3, 986.6]	1.6 [−3004.8, 24.6]	1 [−1.9, 4.8]	1.4 [−1.9, 4.8]	1 [−2.7, 5.1]	1 [−1.8, 4.9]	1 [−1.9, 8]	1 [−1.9, 5.3]	0.9 [−2, 5.5]
	1.3 [−21.8, 24.6]	1.3 [−21.8, 24.6]	1.3 [−8.2, 10.7]	1.3 [−10, 10.7]	1.3 [−8.3, 10.8]	1.1 [−8.2, 10.7]	1.3 [−8.3, 10.9]	1.3 [−8.3, 10.9]	1.3 [−10.2, 10.9]
$\hat{\beta}_4$	1.3 [−23100.4, 34147.9]	1.3 [−2964.1, 74.2]	1.3 [0, 6.3]	1.7 [0, 4.8]	1.3 [−0.1, 6.3]	1.3 [0, 6.3]	1.7 [−0.9, 6.7]	1.8 [−0.9, 6.7]	1 [0, 6.3]
	3 [−5.7, 74.2]	3 [−5.7, 74.2]	2.9 [−5.6, 16.8]	2.9 [−5.6, 16.8]	2.9 [−5.6, 16.9]	2.4 [−5.6, 16.8]	2.8 [−5.7, 17]	2.9 [−5.7, 17]	3 [−6.7, 17.1]
$\hat{\beta}_3$	1.7 [−43167.1, 62593.6]	1.7 [−7222, 26.4]	1.5 [−0.3, 4.5]	1.8 [−0.7, 4.4]	1.5 [−0.7, 4.6]	1.6 [−0.4, 4.6]	1.7 [−0.2, 4.7]	1.7 [−0.4, 4.7]	1.5 [0, 4]
	2.6 [−40.6, 8.9]	2.1 [−40.6, 5.4]	2.1 [−2.3, 7.8]	2.1 [−19.8, 7.5]	2.1 [−2.4, 5.4]	2.1 [−2.3, 7.7]	2.1 [−3.6, 8.9]	2.1 [−3.6, 5.4]	2 [−3.3, 5.4]
$\hat{\beta}_2$	1 [−47710.6, 70377.6]	1 [−1139.4, 13.9]	1.6 [−0.3, 5]	1.9 [−0.3, 5]	1.6 [−0.3, 6.6]	1.8 [−0.9, 5]	1.9 [0, 7.1]	1.8 [0, 5.5]	1.7 [0, 5.3]
	2.1 [−61.4, 8.6]	2.1 [−61.4, 8.6]	2.1 [−12.4, 8.4]	2 [−12.4, 8.5]	2.1 [−12.5, 8.1]	2.3 [−12.4, 8.5]	2.3 [−12.6, 8.6]	2.3 [−12.6, 8.6]	2.1 [−12.7, 8.6]
$\hat{\beta}_1$	1.4 [−82648.3, 5876.5]	1.4 [−11.8, 5876.5]	1.6 [−0.6, 5.3]	1.7 [−0.8, 5.3]	1.3 [−0.6, 5.4]	1.6 [0, 5.3]	1.4 [0, 5.8]	1.3 [−1.2, 5.9]	1.3 [0, 6.6]
	–	–	–	–	–	–	–	–	–
$\hat{\beta}_0$	0.1 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.7, 0.5]	0 [−0.7, 0.5]	0 [−0.7, 0.5]	0 [−1.1, 0.7]
	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−0.6, 0.5]	0 [−1.1, 0.7]

When we remove an explanatory variable from a model with a strong collinearity structure, it is observed that the different methods provide more stable estimates. This elimination of a variable affects the coefficients' estimates of the StepAIC and StepBIC methods, both assuming much lower dispersion (see for example Table 2).

3.3 Two Variables to Define the Outcome

Simpson's Paradox
Considering two variables (one quantitative and the other dichotomous) incorporating the structure of Simpson's paradox and five noise variables without relevant collinearity structure to define the outcome, we observe that the LASSO and ENet methods, in general, select more variables than the other methods and more than expected (in line with what we had observed in previous scenarios when we have noise variables in the models). In the actual scenario, in general, all methods estimate well the coefficients of the two variables. However, when excluding the binary variable, all methods reveal the Simpson's phenomenon, noting that the estimate of the coefficient of the quantitative variable changes sign.

Variables Interaction
Considering that in the definition of the outcome we used two variables (one quantitative and the other dichotomous) that describe an interaction phenomenon and five noise variables without relevant collinearity structure, we noticed that the methods generally perform similarly by not selecting (in mean/median) neither of the two variables that define the interaction.

4 Discussion

In general, the performance of the methods depends on the characteristics/structure of the dataset. None of the methods under analysis can overcome the phenomenon of Simpson's paradox. In real data it would be important to quantify for each dataset the possibility of Simpson's paradox phenomenon occurrence and intensity due to missing variables, stablishing intensity measures and methods to evaluate the impact of the inclusion of unknown variables with potential moderator effects. Variables that define the outcome through interaction will not be selected by the usual methods under analysis. So, the traditional main effects regression models are not able to correctly identify the true explanatory variables.

The methods StepAIC, StepBIC, LASSO and ENet were the most sensitive to the presence of collinearity, revealing that in this context these methods do not have an adequate performance. In the presence of strong collinearity, the performance of the methods can benefit from the elimination of variables.

Funding. This research was partially funded by Portuguese funds through CIDMA, The Center for Research and Development in Mathematics and Applications of University of Aveiro, and the Portuguese Foundation for Science and Technology (FCT–Fundação para a Ciência e a Tecnologia), within projects UIDB/04106/2020 and UIDP/04106/2020.

Appendix

See Fig. A1 and Fig. A2 for two illustrations.

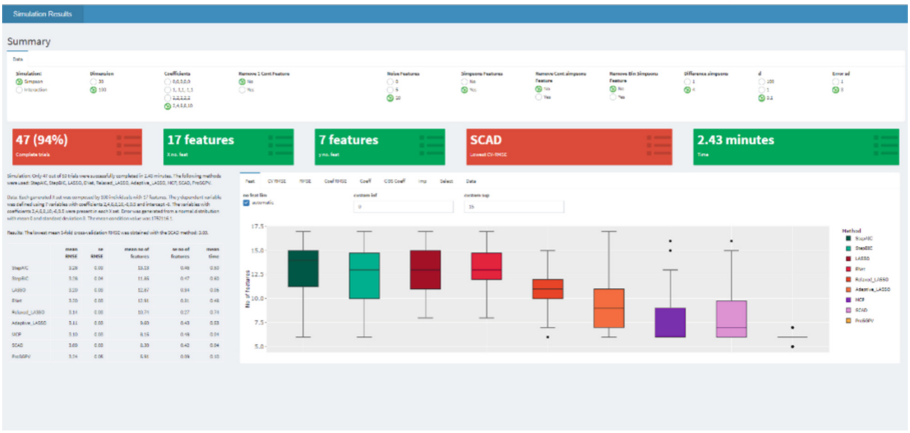

Fig. A1. Shiny app, view of the number of features.

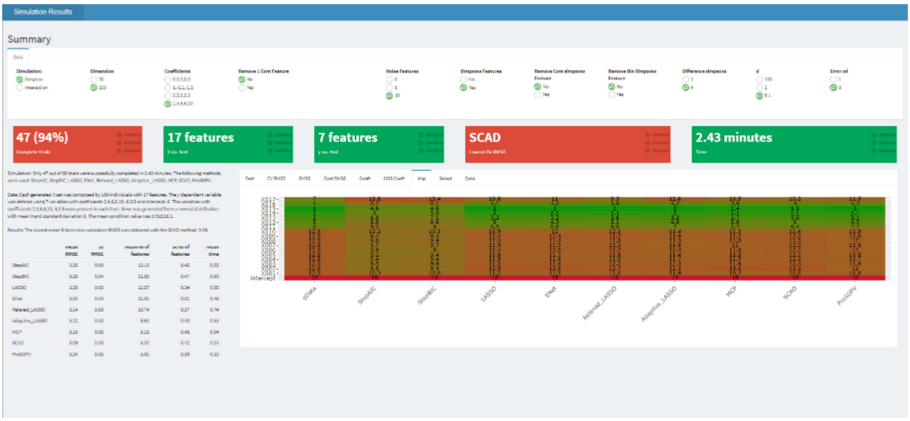

Fig. A2. Shiny app, view of the features' importance.

References

Freedman, D.A.: A note on screening regression equations. Am. Stat. **37**(2), 152–155 (1983)
Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics – Identifying Influential Data and Sources of Collinearity. Wiley, Hoboken, New Jersey (2004)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning – Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
Efron, B., Hastie, T.: Computer Age Statistical Inference. Cambridge University Press, Cambridge (2016)
Efroymson, M.A.: Multiple regression analysis. In: Ralston, A., Wilf, H.S. (eds.) Mathematical Methods for Digital Computers. Wiley, New York (1960)
Hocking, R.R.: The analysis and selection of variables in linear regression. Biometrics **32**(1), 1–49 (1976)
Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**(1), 267–288 (1996)
Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)
Hastie, T., Tibshirani, R., Tibshirani, R.J.: Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv:1707.08692 (2017)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. B **67**(2), 301–320 (2005)
Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
Zuo, Y., Stewart, T.G., Blume, J.D.: Variable selection with second-generation p-values. Am. Stat. **76**(2), 91–101 (2022)
Joe, H.: Generating random correlation matrices based on partial correlations. J. Multivar. Anal. **97**, 2177–2189 (2006)
Simpson, E.H.: The interpretation of interaction in contingency tables. J. R. Stat. Soc. B **13**, 238–241 (1951)
Posit Team. RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA (2022). http://www.posit.co/
R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2022). https://www.R-project.org/
Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2022)
Zuo, Y., Stewart, T.G., Blume, J.D.: ProSGPV: an R package for variable selection with second-generation p-values. F1000Research **11**, 58 (2022). https://doi.org/10.12688/f1000research.74401.1
Goldfeld, K., Wujciak-Jens, J.: simstudy: illuminating research methods through data generation. J. Open Source Softw. **5**(54), 2763 (2020)
Qiu, W., Joe, H.: clusterGeneration: random cluster generation (with specified degree of separation). R version 1.3.7 (2020). https://CRAN.R-project.org/package=clusterGeneration
Makowski, D., Ben-Shachar, M., Lüdecke, D.: bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. J. Open Source Softw. **4**(40), 1541 (2019)