A novel variable selection method in a nonlinear multivariate model using B-splines with an application to geoscience

Mary E Savino, Céline Lévy-Leduc

A NOVEL VARIABLE SELECTION METHOD IN A NONLINEAR MULTIVARIATE MODEL USING B-SPLINES WITH AN APPLICATION TO GEOSCIENCE

MARY E. SAVINO AND CÉLINE LÉVY-LEDUC

Abstract. In this article, we introduce a novel data-driven variable selection approach in a multivariate nonparametric regression model designed to capture only the variables on which the regression function depends. The core concept of our method consists in approximating the underlying function by a linear combination of B-splines of order M and their pairwise interactions. The coefficients of this linear combination are estimated by minimizing the standard least-squares criterion penalized by the sum of the ℓ_2-norms of the partial derivatives with respect to the different variables on which the function depends. We demonstrate that our proposed method can be formulated as a Group Lasso problem, aiming to discard irrelevant variables for which the corresponding coefficients are close to zero. We validate our approach through numerical experiments varying the number of observations, the noise level and the total number of variables and compared it to two other state-of-the-art methods. An application to a geochemical system based on calcite precipitation is also explored. In these different contexts, our approach exhibits better performance than the others. Our completely data-driven method is implemented in the `absorber` R package which is available on the Comprehensive R Archive Network (CRAN).

1. Introduction

The simulation of geochemical models that incorporate precipitation and dissolution reactions of minerals coupled to other physical processes represents a challenging task. Reactive transport modeling (RTM) serves as an illustration, striving to simultaneously consider geochemical reactions, fluid flow, heat transfer and solute transport, see (Steefel, 2019) for various applications. To achieve feasible computational times for these simulations, oversimplifications of the model are usually required. Despite significant improvements in the past few decades, solving three-dimensional-large-scale modelling of complex reactive transport over many time steps remains nearly impossible using standard computers. This challenge has led to the development of Machine Learning (ML)-based approaches aimed at estimating real solutions for full simulation models through the use of surrogate models. The main idea here consists in solving the transport equations explicitly and approximating solutions for geochemical reactions at equilibrium using surrogate models at each time step. A wealth of reviews and surveys on surrogate models for RTM is available in the works of (Razavi et al., 2012; Asher et al., 2015; Jatnieks et al., 2016; Lary et al., 2016). Among these models, Artificial Neural Networks (ANN) have gained prominence, see for instance (Guérillot and Bruyelle, 2020; Prasianakis et al., 2020; Laloy and Jacques, 2022; Demirer et al., 2023). These surrogate models can provide highly accurate approximations and with optimized hyperparameters, they can outperform other surrogate methods (Laloy and Jacques, 2019). However, the computational efficiency of ANN in reducing simulation time comes at the cost of requiring a large dataset and

Key words and phrases. Variable selection, nonparametric regression, B-splines, Group Lasso.

often demands extensive CPU times for training and tuning the hyperparameters (Karpatne et al., 2018). Furthermore, an approach based on an on-demand training algorithm allows to train the model at runtime to iteratively build the training dataset (Leal et al., 2017). Analogously, an active learning approach has also been introduced to RTM (Savino et al., 2022) to drastically diminish the dataset size while insuring good approximation accuracy. Finally, a novel approach based on B-splines and on an adaptive knot selection was proposed in (Savino and Lévy-Leduc, 2023) to improve the approximation accuracy while having only a few parameters to tune.

Another approach to improve the surrogate model accuracy while reducing the CPU times is to reduce the number of input variables to consider in the model. This can be reformulated as a variable selection problem in the following framework.

Let us consider that we have n observations satisfying the following nonparametric regression model:

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \in \mathbb{R}^p, \quad 1 \leq i \leq n \tag{1}$$

where f is an unknown real-valued function and where the ε_i's are i.i.d centered random variables of variance σ^2. We will also assume that f actually depends only on d variables instead of p, with $d < p$, which means that there exists a real-valued function $\widetilde{f}$ such that $f(x) = \widetilde{f}(\widetilde{x})$, where $x \in \mathbb{R}^p$ and $\widetilde{x} \in \mathbb{R}^d$. Variable selection consists in identifying the components of $\widetilde{x}$.

Efficient methodologies have been devised over the last few decades particularly when the variables x_is and Y_is in (1) are linearly related. Notable examples include the Lasso regression formulated by (Tibshirani, 1996) and one of its variant the Elastic Net defined by (Zou and Hastie, 2005). However, dealing with the nonlinearity of the relationship between the x_is and the Y_is in (1) poses a greater challenge.

In their paper, (Yamada et al., 2014) introduced a feature-wise kernelized Lasso method tailored for variable selection in nonlinear models: HSIC-Lasso and its variation NOCCO-Lasso. This approach employs the kernel trick within the regular Lasso problem in combination with kernel-based independence measures to discern and selectively choose relevant variables. While these methods offer scalability to high-dimensional variable selection problems, they exhibit sensitivity to the number of observations. Tree-based methods such as Random Forests introduced by (Breiman, 2001) are widely used for variable selection in regression models since they are well-suited to describe nonlinearity in (1). Numerous applications and challenges associated with its use for variable selection are discussed in (Genuer et al., 2010).

More recently, ANN have gained interest for variable selection and regression. We will present just a few examples of them. For instance, (Liang et al., 2018) developed a method based on a Bayesian neural network architecture to select variables for which the marginal inclusion probability exceeds a predefined threshold. Furthermore, regularized approaches with different ANN architectures were proposed as seen in the work of (Li et al., 2016) where a ridge regularization approach is considered for the weights of the first and hidden layers. Similarly, (Feng and Simon, 2017) introduced the SPINN method which is a single-layer neural network with a sparse group lasso regularization to shrink the weights corresponding to the units of the irrelevant variables. They later proposed another approach with deeper neural networks in (Feng and Simon, 2022). (Ye and Sun, 2018) adapted the SPINN method by adding a greedy elimination algorithm to iteratively drop one variable at a time and determine if the empirical loss decreases. Finally, (Lemhadri et al., 2021) presented LassoNet, a residual feed-forward neural network architecture as introduced in (He et al., 2016) which incorporates a regularization approach based on Lasso regression to selectively use a subset of the features

in the network. Similar work on variable selection with ANN can be found in (Chen et al., 2021; Lu et al., 2018; Zhu and Zhao, 2021). All these methods display interesting results especially for high-dimensional problems but often suffer from high training CPU times and a large number of hyperparameters to tune.

Another research direction focuses on developing flexible and interpretable methods using splines for piecewise polynomial fitting such as Multivariate Adaptive Regression Splines introduced by (Friedman, 1991). This method enables the description of interactions and nonlinear relations by automatically pruning the most irrelevant terms. In the same vein, (Lin and Zhang, 2006) developed COSSO, a regularization method for component selection and smoothing splines where the penalty term is the sum of the component norms. This approach can be considered as a more generalized form of the Lasso approach with a Reproducing Kernel Hilbert Space (RKHS) constraint. Compared to MARS, it shows better results except for higher dimensional cases with small datasets. Similarly, (Ravikumar et al., 2009) proposed sparse additive models (SpAM) based on a generalized additive model with a ℓ_2-norm regularization.

A few articles have proposed considering a sparse additive model using a linear combination of B-splines of order M ($M \geq 1$), introduced by (De Boor, 1978) in Chapter 9. Their ability to approximate complex functions without being significantly altered by the presence of noise has made them very attractive in the past few decades. As an illustration, (Huang et al., 2010) approximated the underlying function f in (1) using an additive B-spline estimator and subsequently, employed an adaptive group lasso approach for variable selection. In their study, they presented both numerical applications and theoretical results regarding the selection consistency of their proposed method. (Antoniadis et al., 2012) leveraged the benefits of B-splines by incorporating a penalized version known as P-splines, introduced by (Eilers and Marx, 1996), and compared their results with various adaptations of COSSO. Additional references on variable selection with P-splines can be found in the review by (Gijbels et al., 2015). While these approaches have proven to be efficient for high-dimensional nonparametric additive models, they fall short in describing interactions that may exist in real datasets. Therefore, (Radchenko and James, 2010) extended the SpAM approach to consider both single and interaction terms, aiming to construct a more interpretable approximation of f. This presented approach, known as VANISH, strongly penalizes interaction terms to simplify the model as much as possible and has demonstrated efficiency for small datasets. In parallel, (Rosasco et al., 2010) proposed a novel method for variable selection based on a regularized least-square estimator penalizing large values of the partial derivatives to select the most relevant variables in a multivariate nonlinear regression model with a RKHS constraint.

In this article, we propose a novel method for variable selection motivated by (Radchenko and James, 2010) using a multivariate nonparametric regression model to retrieve the d relevant variables on which f in (1) truly depends. Our approach involves approximating f using a linear combination of B-splines and their pairwise interactions. Additionally, drawing inspiration from the methodology of (Rosasco et al., 2010), the coefficients of the linear combination are estimated by minimizing the usual least-squares criterion penalized by the sum of the ℓ_2-norms of the partial derivatives with respect to the different variables on which f depends. We will demonstrate that our proposed method can be formulated as a Group Lasso problem defined by (Yuan and Lin, 2006) and thus, can be easily implemented. Two different approaches to choose the penalization parameter will be presented to the reader.

This paper is organized as follows. Section 2 presents the methodology that we propose for variable selection in nonlinear models. Section 3 investigates the performance of our

approach through numerical experiments. Finally, in Section 4, we apply our method to a real geochemical application that motivated this study.

2. Methodology

2.1. **Approximation of f using B-splines.** Let us first recall how the B-spline basis associated to a given dimension among the p, the ℓth for instance, is defined.

Let $\mathbf{t}_\ell = (t_{\ell,1}, \ldots, t_{\ell,K})$ be a set of K points called knots and let $\mathcal{S}_\ell$ be a compact subset of $\mathbb{R}$. Following (De Boor, 1978, p. 89-90) and (Hastie et al., 2009, p. 160), the augmented knot sequence $\boldsymbol{\tau}_\ell$ is defined as follows:

$$\begin{aligned}
\tau_{\ell,1} = \ldots = \tau_{\ell,M} &= x_{min}^{(\ell)}, \\
\tau_{\ell,j+M} &= t_{\ell,j}, \quad j = 1, \ldots, K, \\
\tau_{\ell,K+M+1} = \ldots = \tau_{\ell,K+2M} &= x_{max}^{(\ell)}, \\
\boldsymbol{\tau}_\ell = (\tau_{\ell,1}, \ldots, \tau_{\ell,K+2M}) &= (\underbrace{x_{min}^{(\ell)}, \ldots, x_{min}^{(\ell)}}_{\text{M times}}, \underbrace{t_{\ell,1}, \ldots, t_{\ell,K}}_{\mathbf{t}_\ell}, \underbrace{x_{max}^{(\ell)}, \ldots, x_{max}^{(\ell)}}_{\text{M times}}),
\end{aligned}$$

where $x_{min}^{(\ell)}$ and $x_{max}^{(\ell)}$ are the lower and upper bounds of $\mathcal{S}_\ell$, respectively.

Denoting by $B_{k,m}^{(\ell)}$ the kth B-spline basis function of order m with $m \leq M$ for the knot sequence $\boldsymbol{\tau}_\ell$ and for the dimension ℓ, B-splines are defined by the following recursion:

$$B_{k,1}^{(\ell)}(x^{(\ell)}) = \begin{cases} 1 & \text{if } \tau_{\ell,k} \leq x^{(\ell)} < \tau_{\ell,k+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } \; k = 1, \ldots, K + 2M - 1, \tag{2}$$

and for $2 \leq m \leq M$,

$$B_{k,m}^{(\ell)}(x^{(\ell)}) = \frac{x^{(\ell)} - \tau_{\ell,k}}{\tau_{\ell,k+m-1} - \tau_{\ell,k}} B_{k,m-1}^{(\ell)}(x^{(\ell)}) + \frac{\tau_{\ell,k+m} - x^{(\ell)}}{\tau_{\ell,k+m} - \tau_{\ell,k+1}} B_{k+1,m-1}^{(\ell)}(x^{(\ell)}), \tag{3}$$

for $k = 1, \ldots, (K + 2M - m)$.

Inspired by (Radchenko and James, 2010), we propose approximating the function $f(x^{(1)}, \ldots, x^{(p)})$ appearing in (1) by a linear combination of B-splines of each variable $x^{(1)}, \ldots, x^{(p)}$ and of pairwise interaction of them as follows:

$$F(x^{(1)}, \ldots, x^{(p)}) = \sum_{\ell=1}^{p} \sum_{k=1}^{K+M} \beta_k^{(\ell)} B_k^{(\ell)}(x^{(\ell)}) + \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \left(\sum_{k=1}^{K+M} \sum_{q=1}^{K+M} \beta_{k,q}^{(\ell,j)} B_k^{(\ell)}(x^{(\ell)}) B_q^{(j)}(x^{(j)}) \right), \tag{4}$$

where $B_k^{(\ell)} = B_{k,M}^{(\ell)}$ is defined in (2) and (3) and where $\beta_k^{(\ell)}$ and $\beta_{k,q}^{(\ell,j)}$ are unknown coefficients.

Observe that the column vector $(F(x_i^{(1)}, \ldots, x_i^{(p)}))_{1 \leq i \leq n}$ (4) can be rewritten as follows:

$$\sum_{\ell=1}^{p} \Psi_\ell \, \boldsymbol{\beta}_\ell + \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \, \boldsymbol{\beta}_{\ell,j} \, . \tag{5}$$

where Ψ_ℓ is a $n \times (K+M)$ matrix such that its ith row is equal to $(B_1^{(\ell)}(x_i^{(\ell)}), \ldots, B_{K+M}^{(\ell)}(x_i^{(\ell)}))$ and $\boldsymbol{\beta}_\ell = \left(\beta_1^{(\ell)} \; \ldots \; \beta_{K+M}^{(\ell)}\right)^T$ for $1 \leq \ell \leq p$, A^T denoting the transpose of the matrix A. Moreover, $\Phi_{\ell j}$ is an $n \times (K+M)^2$ matrix such that its ith row satisfies $(\Phi_{\ell j})_{i,\bullet} = ((\Psi_\ell)_{i,\bullet} \otimes$

$(\Psi_j)_{i,\bullet})$, $\otimes$ denoting the Kronecker product, $(\Psi_\ell)_{i,\bullet}$ denoting the ith row of Ψ_ℓ and $\boldsymbol{\beta}_{\ell,j} = \left(\beta_{1,1}^{(\ell,j)}\ \beta_{1,2}^{(\ell,j)}\ \dots\ \beta_{K+M,K+M}^{(\ell,j)}\right)^T$ for $1 \leq \ell < j \leq p$.

2.2. **Description of our variable selection method.** Inspired by the methodology of (Rosasco et al., 2010), we propose selecting the variables on which f depends by estimating the coefficients $\boldsymbol{\beta}_\ell$ and $\boldsymbol{\beta}_{\ell,j}$ appearing in (5) by minimizing the following regularized criterion:

$$\begin{aligned}&\left(\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\lambda}), \dots, \widehat{\boldsymbol{\beta}}_p(\boldsymbol{\lambda}), \widehat{\boldsymbol{\beta}}_{1,2}(\boldsymbol{\lambda}), \dots, \widehat{\boldsymbol{\beta}}_{(p-1),p}(\boldsymbol{\lambda})\right)\\ &= \underset{\substack{(\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_p)\\(\boldsymbol{\beta}_{1,2},\dots,\boldsymbol{\beta}_{(p-1),p})}}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell\, \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j}\, \boldsymbol{\beta}_{\ell,j} \right\|_2^2 + \boldsymbol{\lambda} \sum_{\ell=1}^{p} \sqrt{\sum_{i=1}^{n} \partial_\ell F(x_i)^2} \right),\end{aligned}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)$, the Y_i's being defined in (1), $\partial_\ell F(x_i)$ denotes the ℓth partial derivative of F defined in (4) at some observation point $x_i = (x_i^{(1)}, \dots, x_i^{(p)})$ and $||y||_2^2 = \sum_{i=1}^n y_i^2$. Note that the idea underlying this criterion is that when a function does not depend on a variable its partial derivative with respect to this variable is equal to zero.

Using the definition of F given in (5) the criterion can be rewritten as follows:

$$\begin{aligned}&\left(\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\lambda}), \dots, \widehat{\boldsymbol{\beta}}_p(\boldsymbol{\lambda}), \widehat{\boldsymbol{\beta}}_{1,2}(\boldsymbol{\lambda}), \dots, \widehat{\boldsymbol{\beta}}_{(p-1),p}(\boldsymbol{\lambda})\right)\\ &= \underset{\substack{(\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_p)\\(\boldsymbol{\beta}_{12},\dots,\boldsymbol{\beta}_{(p-1)p})}}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell\, \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j}\, \boldsymbol{\beta}_{\ell,j} \right\|_2^2 \right.\\ &\left. + \boldsymbol{\lambda} \sum_{\ell=1}^{p} \left\| \Psi'_\ell\, \boldsymbol{\beta}_\ell + \sum_{j=\ell+1}^{p} (\partial_\ell \Phi_{\ell j})\, \boldsymbol{\beta}_{\ell,j} + \sum_{1 \leq j < \ell} (\partial_\ell \Phi_{j\ell})\, \boldsymbol{\beta}_{j,\ell} \right\|_2 \right),\end{aligned} \tag{6}$$

where Ψ'_ℓ is the $n \times (K+M)$ matrix such that $(\Psi'_\ell)_{i,k} = B_k^{(\ell)'}(x_i^{(\ell)})$, $B_k^{(\ell)'}$ denoting the first derivative of $B_k^{(\ell)}$. The ith row of $(\partial_\ell \Phi_{\ell j})$ (resp. $(\partial_\ell \Phi_{j\ell})$) is defined by $(\partial_\ell \Phi_{\ell j})_{i,\bullet} = ((\Psi'_\ell)_{i,\bullet} \otimes (\Psi_j)_{i,\bullet})$ (resp. $(\partial_\ell \Phi_{j\ell})_{i,\bullet} = ((\Psi_j)_{i,\bullet} \otimes (\Psi'_\ell)_{i,\bullet})$). By denoting $(\partial_\ell \Phi_{\ell\bullet}) = \big((\partial_\ell \Phi_{\ell(\ell+1)}) \dots (\partial_\ell \Phi_{\ell p})\big)$, $(\partial_\ell \Phi_{\bullet\ell}) = \big((\partial_\ell \Phi_{1\ell}) \dots (\partial_\ell \Phi_{(\ell-1)\ell})\big)$, $\boldsymbol{\beta}_{\ell\bullet} = \left(\boldsymbol{\beta}_{\ell,(\ell+1)} \dots \boldsymbol{\beta}_{\ell,p}\right)$ and $\boldsymbol{\beta}_{\bullet\ell} = \left(\boldsymbol{\beta}_{1,\ell} \dots \boldsymbol{\beta}_{(\ell-1),\ell}\right)$, the penalty term can be written as:

$$\boldsymbol{\lambda} \sum_{\ell=1}^{p} \left\| \Psi'_\ell\, \boldsymbol{\beta}_\ell + (\partial_\ell \Phi_{\ell\bullet})\, \boldsymbol{\beta}_{\ell\bullet} + (\partial_\ell \Phi_{\bullet\ell})\, \boldsymbol{\beta}_{\bullet\ell} \right\|_2 =: \boldsymbol{\lambda} \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell)\, \boldsymbol{\gamma}_\ell \right\|_2, \tag{7}$$

where $\boldsymbol{\gamma}_\ell = \left(\boldsymbol{\beta}_\ell^T\ \boldsymbol{\beta}_{\ell,\ell+1}^T\ \dots\ \boldsymbol{\beta}_{\ell,p}^T\, \boldsymbol{\beta}_{1,\ell}^T\ \dots\ \boldsymbol{\beta}_{\ell-1,\ell}^T\right)^T$. Using that

$$\sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j}\, \boldsymbol{\beta}_{\ell,j} = \sum_{j=2}^{p} \sum_{\ell=1}^{j-1} \Phi_{\ell j}\, \boldsymbol{\beta}_{\ell,j} = \sum_{\ell=2}^{p} \sum_{j=1}^{\ell-1} \Phi_{j\ell}\, \boldsymbol{\beta}_{j,\ell}\,,$$

the least-squares term can be rewritten as follows:

$$\begin{aligned}
&\Big\|\mathbf{Y}-\sum_{\ell=1}^{p}\Psi_\ell\,\boldsymbol{\beta}_\ell-\sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p}\Phi_{\ell j}\,\boldsymbol{\beta}_{\ell,j}\Big\|_2^2\\
&=\Big\|\mathbf{Y}-\sum_{\ell=1}^{p}\Psi_\ell\,\boldsymbol{\beta}_\ell-\frac{1}{2}\left(\sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p}\Phi_{\ell j}\,\boldsymbol{\beta}_{\ell,j}+\sum_{\ell=2}^{p}\sum_{j=1}^{\ell-1}\Phi_{j\ell}\,\boldsymbol{\beta}_{j,\ell}\right)\Big\|_2^2 \qquad (8)\\
&=:\Big\|\mathbf{Y}-\sum_{\ell=1}^{p}\Theta_\ell\,\boldsymbol{\gamma}_\ell\Big\|_2^2.
\end{aligned}$$

Equation (8) comes by setting $\Theta_1=\begin{pmatrix}\Psi_1 & \frac{1}{2}\Phi_{1\bullet}\end{pmatrix}$ and $\Theta_p=\begin{pmatrix}\Psi_p & \frac{1}{2}\Phi_{\bullet p}\end{pmatrix}$, where $\Phi_{\ell\bullet}=(\Phi_{\ell(\ell+1)}\dots\Phi_{\ell p})$ and $\Phi_{\bullet\ell}=(\Phi_{1\ell}\dots\Phi_{(\ell-1)\ell})$. Combining (7) and (8), (6) can be rewritten as:

$$\big(\widehat{\boldsymbol{\gamma}}_1(\boldsymbol{\lambda}),\dots,\widehat{\boldsymbol{\gamma}}_p(\boldsymbol{\lambda})\big)=\underset{(\boldsymbol{\gamma}_1,\dots,\boldsymbol{\gamma}_p)}{\operatorname{argmin}}\left(\Big\|\mathbf{Y}-\sum_{\ell=1}^{p}\Theta_\ell\,\boldsymbol{\gamma}_\ell\Big\|_2^2+\boldsymbol{\lambda}\sum_{\ell=1}^{p}\Big\|(\partial_\ell\Theta_\ell)\,\boldsymbol{\gamma}_\ell\Big\|_2\right). \qquad (9)$$

By defining $\boldsymbol{\alpha}_\ell=(\partial_\ell\Theta_\ell)\,\boldsymbol{\gamma}_\ell$ and $\widetilde{\mathbf{X}}_\ell=\Theta_\ell(\partial_\ell\Theta_\ell)^+$, A^+ being the Moore-Penrose inverse of matrix A, (9) can be rewritten as:

$$(\widehat{\boldsymbol{\alpha}}_1(\boldsymbol{\lambda}),\dots,\widehat{\boldsymbol{\alpha}}_p(\boldsymbol{\lambda}))=\underset{(\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_p)}{\operatorname{argmin}}\left(\Big\|\mathbf{Y}-\sum_{\ell=1}^{p}\widetilde{\mathbf{X}}_\ell\,\boldsymbol{\alpha}_\ell\Big\|_2^2+\boldsymbol{\lambda}\sum_{\ell=1}^{p}\Big\|\,\boldsymbol{\alpha}_\ell\Big\|_2\right). \qquad (10)$$

The last formulation of our variable selection criterion (10) can be seen as a group lasso problem introduced by (Yuan and Lin, 2006), where the size p_ℓ of each group ℓ belonging to $\{1,\dots,p\}$ is equal to n. This approach is implemented in numerous R packages such as the most recent one `sparsegl` developed by (Liang et al., 2022) that we used in the numerical experiments section. For a fixed number of parameters $\boldsymbol{\lambda}$, this package provides a set of penalization parameters Λ and the coefficients $\widehat{\boldsymbol{\alpha}}_\ell(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$ belonging to Λ. The coefficients $\widehat{\boldsymbol{\gamma}}_\ell(\boldsymbol{\lambda})$ are thus obtained as follows

$$\widehat{\boldsymbol{\gamma}}_\ell(\boldsymbol{\lambda})=(\partial_\ell\Theta_\ell)^+\widehat{\boldsymbol{\alpha}}_\ell(\boldsymbol{\lambda}). \qquad (11)$$

We then define the active variables for each $\boldsymbol{\lambda}$ in Λ as follows:

$$\mathcal{V}_{\boldsymbol{\lambda}}=\left\{\ell,\ \sum_{k\geq 1}|\widehat{\gamma}_{\ell,k}(\boldsymbol{\lambda})|\neq 0\right\}, \qquad (12)$$

where $\widehat{\gamma}_{\ell,k}(\boldsymbol{\lambda})$ is the kth coefficient of $\widehat{\boldsymbol{\gamma}}_\ell(\boldsymbol{\lambda})$.

We also introduce the set $\mathcal{V}_f$ of the indices of the d relevant variables on which f in (1) actually depends that we seek to select among the p variables and the set $\overline{\mathcal{V}_f}$ of the indices of the irrelevant variables on which f does not depend.

2.3. **Choice of K.** Our method relies on the definition of the B-spline basis for each ℓ in $\{1,\dots,p\}$ and thus on the choice of the set of knots $\mathbf{t}_\ell$ used for defining them. For simplifying this choice, we considered evenly spaced knots in the interval $[0,1]$. For regularity purposes, we use quadratic B-splines with $M=3$. Thus, we are only interested in optimizing the number of knots K. To find the best value of K, we use two sensitivity measures. Firstly, for each $\boldsymbol{\lambda}$ belonging to Λ, we computed the True Positive Rate (TPR) and the False Positive Rate

(FPR), defined as:

$$\mathrm{TPR}(\boldsymbol{\lambda}) = \frac{\mathrm{TP}(\boldsymbol{\lambda})}{d} = \frac{|\mathcal{V}_{\boldsymbol{\lambda}} \cap \mathcal{V}_f|}{d} \quad \text{and} \quad \mathrm{FPR}(\boldsymbol{\lambda}) = \frac{\mathrm{FP}(\boldsymbol{\lambda})}{p-d} = \frac{\left|\mathcal{V}_{\boldsymbol{\lambda}} \cap \overline{\mathcal{V}_f}\right|}{p-d},$$

where $d < p$, $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$, $\mathrm{TP}(\boldsymbol{\lambda})$ and $\mathrm{FP}(\boldsymbol{\lambda})$ are the number of true selected variables and the number of false selected variables for $\boldsymbol{\lambda}$, respectively. $\mathcal{V}_f$, $\overline{\mathcal{V}_f}$ and $\mathcal{V}_{\boldsymbol{\lambda}}$ are introduced in the previous section. We can then draw the ROC curve where each point has as coordinates $(\mathrm{FPR}(\boldsymbol{\lambda}), \mathrm{TPR}(\boldsymbol{\lambda}))$ for $\boldsymbol{\lambda}$ belonging to Λ.

In order to have an idea of the quality of our variable selection procedure, we calculate the Area Under Curve (AUC) of the ROC curves as well as a complementary indicator that we want to maximize:

$$\max(\mathrm{TPR} - \mathrm{FPR}) = \max_{\boldsymbol{\lambda} \in \Lambda} (\mathrm{TPR}(\boldsymbol{\lambda}) - \mathrm{FPR}(\boldsymbol{\lambda})).$$

To assess the quality of our variable selection procedure according to K, we define two functions on which our method is applied:

$$\begin{gathered} f_1(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}) = 2B_2^{(3)}(x^{(3)})B_4^{(5)}(x^{(5)}) + 2B_4^{(5)}(x^{(5)}) - 5B_2^{(3)}(x^{(3)}), \\ (x^{(1)}, \ldots, x^{(5)}) \in [0,1]^5, \end{gathered} \tag{13}$$

$$\begin{gathered} f_2(x^{(1)}, \ldots, x^{(10)}) = 1.8 \cos\left(x^{(1)}\right) \sin\left(x^{(7)} + 1\right) - 5 \ln\left(x^{(3)} + 1\right) - \frac{0.9}{{x^{(10)}}^2 + 1}, \\ (x^{(1)}, \ldots, x^{(10)}) \in [0,1]^{10}. \end{gathered} \tag{14}$$

In (13), the B-spline bases are defined using $\mathbf{t}_\ell = (0.2, 0.5, 0.6, 0.75, 0.8)$ for each ℓ belonging to $\{1, \ldots, 5\}$. Here, $\mathcal{V}_{f_1} = \{3, 5\}$ and $\mathcal{V}_{f_2} = \{1, 3, 7, 10\}$. Results for the two metrics defined above, AUC and max(TPR−FPR), are shown for $f = f_1$ and $f = f_2$ for 10 random samplings of the observation set and for $\sigma = 0$ and $\sigma = 0.5$ in Figure 1 and in Figure 3 for $n = 350$ and $n = 2000$, respectively.

Firstly, we can clearly see for $n = 350$ in Figure 1 that for f_1, all the values of K are satisfying as they allow us to get $\max(\mathrm{TPR} - \mathrm{FPR}) = 1$ and $\mathrm{AUC} = 1$.

However, for $f = f_2$, we do not have necessarily $\mathrm{AUC} = 1$ when $\max(\mathrm{TPR} - \mathrm{FPR}) = 1$ for instance for $K = 3$, which does not imply that this method does not select properly the relevant variables. To illustrate this idea, one can relate to Figure 2 in which the ROC curves for $f = f_2$ are drawn for each value of K belonging to $\{1, \ldots, 10\}$. Here, we can observe that a good variable selection method will not necessarily lead to $\mathrm{AUC} = 1$ since $\mathrm{FPR}(\boldsymbol{\lambda}) < 1$ for every $\boldsymbol{\lambda}$ belonging to Λ, which indicates that the even smallest value of $\boldsymbol{\lambda}$ will not select all the irrelevant variables. These phenomena are even more visible for $n = 2000$ in Figure 3 for f_1 and f_2. The results for $n = 350$ allow us to discriminate a value of K which gives good selection for both $\sigma = 0$ and $\sigma = 0.5$ and for both functions f_1 and f_2. Indeed, for f_1 without any noise in the observation set we can see that only the true variables are selected for $K = 3$ since $\mathrm{AUC} = 0$ and $\max(\mathrm{TPR} - \mathrm{FPR}) = 1$. Moreover, $K = 3$ is the only case where $\max(\mathrm{TPR} - \mathrm{FPR}) = 1$ for nearly all the different samplings of the observation set when $\sigma = 0.5$. Higher values of K drastically deteriorate the AUC and the max(TPR−FPR), especially for noisy observation sets.

For all these reasons, we decide to only focus on using our method with $K = 3$ evenly spaced knots in the B-spline basis.

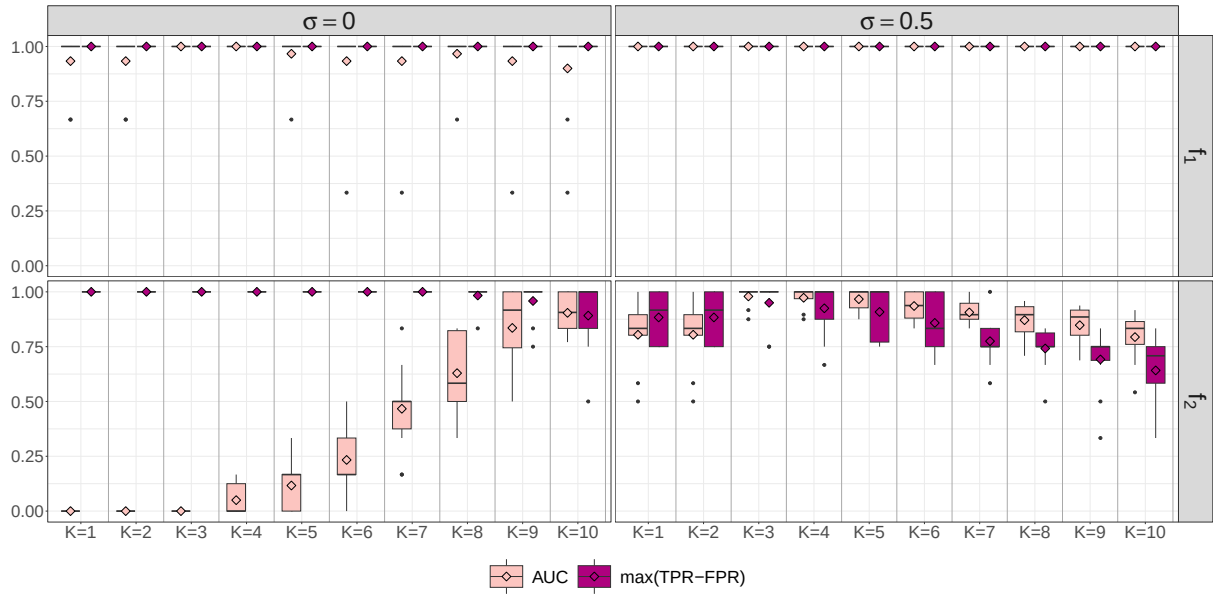

FIGURE 1. AUC and max(TPR − FPR) calculated for an increasing value of K for f_1 (top) and f_2 (bottom) with noise (right) or without noise (left) in the observation set $\mathbf{Y} = (Y_1, \ldots, Y_{350})$. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty bullets inside the boxplots represent the mean value and the plain bullets outside the boxplots are the extreme values.

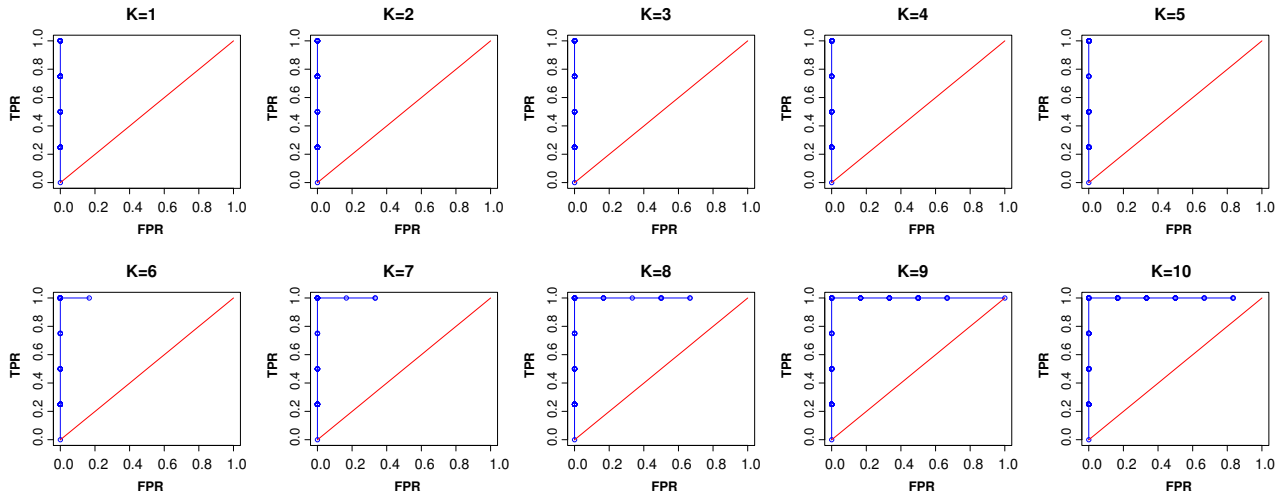

FIGURE 2. ROC curves for an increasing value of K for f_2 with no noisy observations $\mathbf{Y} = (Y_1, \ldots, Y_{350})$ ($\sigma = 0$ in (1)) and for one sampling of the observation set (blue line). The red line corresponds to the identity function TPR = FPR.

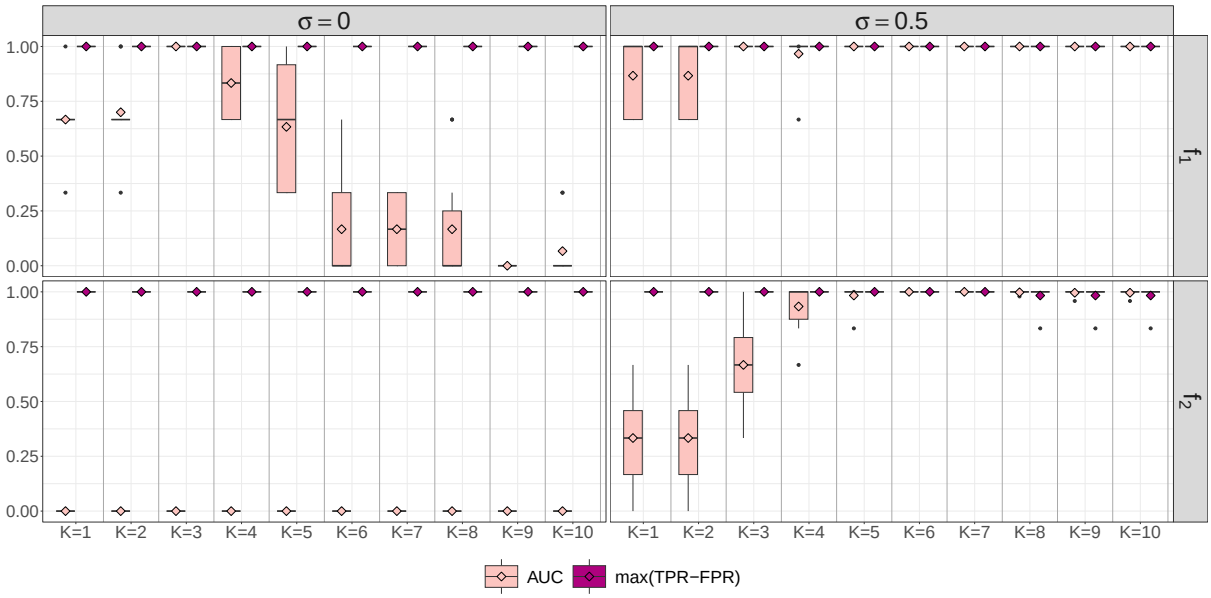

FIGURE 3. AUC and max(TPR − FPR) calculated for an increasing value of K for f_1 (top) and f_2 (bottom) with noise (right) or without noise (left) in the observation set $\mathbf{Y} = (Y_1, \ldots, Y_{2000})$. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty bullets inside the boxplots represent the mean value and the plain bullets outside the boxplots are the extreme values.

2.4. **Choice of $\boldsymbol{\lambda}$.** By following the previous method, we have a set of penalization parameters Λ and a set of indices of selected variables for each of them in $\mathcal{V}_\lambda$. Let us now propose two ways of selecting the final set of selected variables among the different sets $\mathcal{V}_\lambda$.

The first one is based on the percentage of variable selection defined for each variable ℓ belonging to $\{1, \ldots, p\}$ by:

$$P_\ell = \frac{100}{|\Lambda|} \sum_{\lambda \in \Lambda} \mathbb{1}\{\ell \in \mathcal{V}_\lambda\}, \tag{15}$$

where $|\Lambda|$ is the total number of parameters in Λ, $\mathbb{1}\{A\} = 1$ if the event A holds and 0 if not and $\mathcal{V}_\lambda$ is defined in (12).

Results for the percentage of selection of variables are displayed in Figure 4 (resp. Figure 5) for f_1 (resp. for f_2). Firstly, we obtain a high percentage of selection for the relevant variables $\mathcal{V}_{f_1} = \{3, 5\}$ since they are selected for more than 75% of the $\boldsymbol{\lambda}$s belonging to Λ. Moreover, we can observe a huge gap between the frequency for the relevant ($\mathcal{V}_{f_1} = \{3, 5\}$) and the irrelevant ($\overline{\mathcal{V}_{f_1}} = \{1, 2, 4\}$) variables. This gap is amplified as we increase the number of observations from $n = 350$ to $n = 2000$. The noise of the observations does not seem here to deteriorate the results. For f_2, the percentage of the relevant variables ($\mathcal{V}_{f_2} = \{1, 3, 7, 10\}$) are lower than the previous function (40% for variable 7), however we can see similar frequencies for the irrelevant variables ($\overline{\mathcal{V}_{f_2}} = \{2, 4, 5, 6, 8, 9\}$) and a clear gap for the unnoisy observation sets ($\sigma = 0$) starting with only $n = 350$ as $\overline{\mathcal{V}_{f_2}} = \{2, 4, 5, 6, 8, 9\}$ are never selected. The noise has here an influence on the quality of selection but by increasing the number of observations we can circumvent this issue as the gap is visible for $\sigma = 0.5$ with $n = 2000$. We encourage the

user to add known fake variables in order to know which threshold of percentage of selection to use. All the variables having a percentage of selection close to the one of the added fake variables can then be considered as irrelevant as we can see in Figures 4 and 5 which are visualizations of the output of our method for one sample of the observation set for f_1 and f_2, respectively.

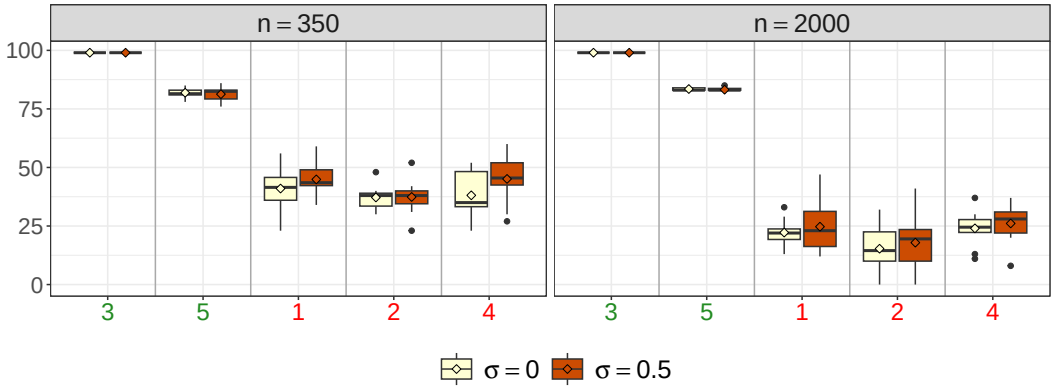

FIGURE 4. Percentage of selection of each variable for f_1 with $n = 350$ (left) and $n = 2000$ (right) and for $\sigma = 0$ or $\sigma = 0.5$. The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_1. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty bullets inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

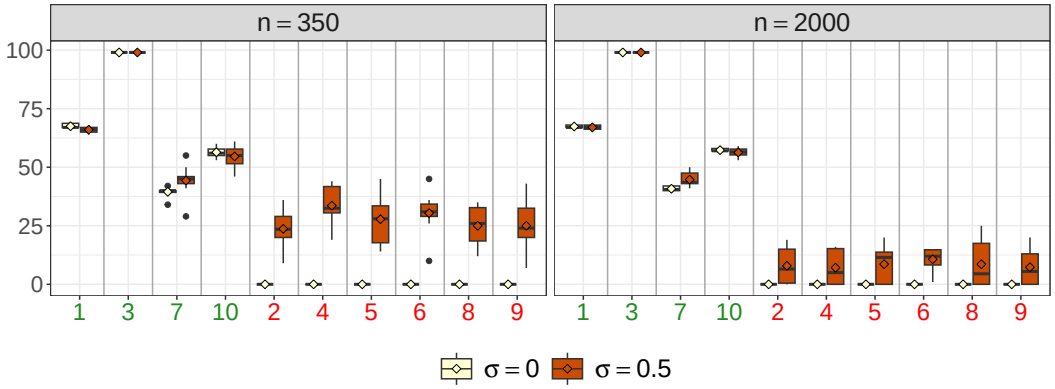

FIGURE 5. Percentage of selection of each variable for f_2 with $n = 350$ (left) and $n = 2000$ (right) and for $\sigma = 0$ or $\sigma = 0.5$. The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_2. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

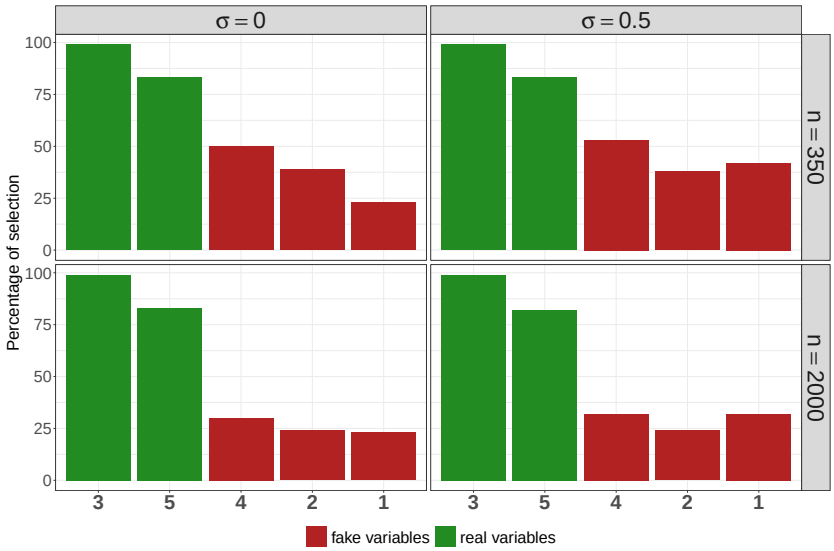

FIGURE 6. Percentage of selection of each variable for f_1 with $n = 350$ (top) and $n = 2000$ (bottom) and for $\sigma = 0$ (left) or $\sigma = 0.5$ (right). The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_1. Only one sampling of $\mathbf{Y}$ is used to obtain the results displayed.

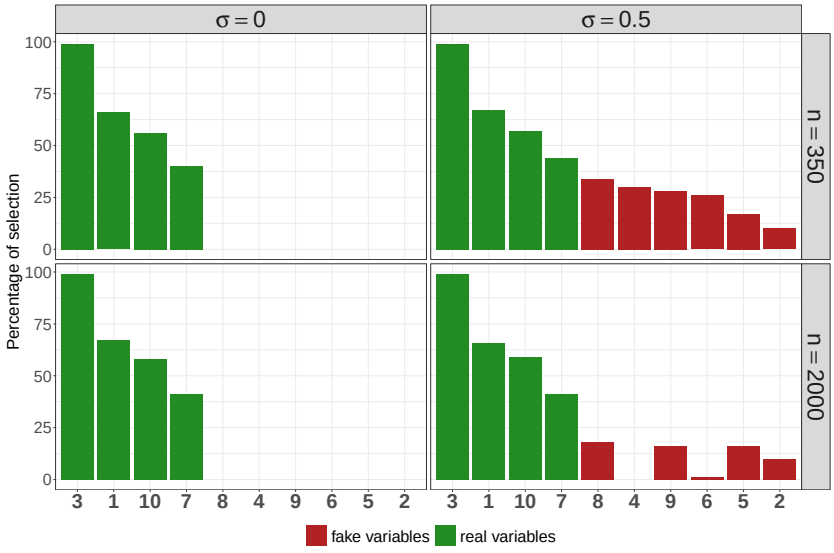

FIGURE 7. Percentage of selection of each variable for f_2 with $n = 350$ (top) and $n = 2000$ (bottom) and for $\sigma = 0$ (left) or $\sigma = 0.5$ (right). The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_2. Only one sampling of $\mathbf{Y}$ is used to obtain the displayed results.

We also propose another method to automatically choose $\boldsymbol{\lambda}$. Among several existing criteria for model selection, we initially assessed a cross-validation criterion using the mean-square error as a loss function. However this approach was not satisfactory as it tended to overestimate

the number of relevant variables. Hence, the suggested method leverages the popular Akaike Information Criterion (AIC) introduced in (Akaike, 1973) and defined by:

$$\mathrm{AIC}(\boldsymbol{\lambda}) = n \ln\left(\frac{\mathrm{RSS}(\boldsymbol{\lambda})}{n}\right) + 2T_{\boldsymbol{\lambda}}\,, \tag{16}$$

where $T_{\boldsymbol{\lambda}}$ is the number of terms appearing in (5) by keeping only the variables selected with $\boldsymbol{\lambda}$ and $\mathrm{RSS}(\boldsymbol{\lambda})$ is the residual sum of squares defined as follows:

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\lambda}) &= \left\| \mathbf{Y} - \widehat{\mathbf{Y}}(\boldsymbol{\lambda}) \right\|_2^2 , \\ \text{with} \quad \widehat{\mathbf{Y}}(\boldsymbol{\lambda}) &= \sum_{\ell=1}^{p} \Theta_\ell \widehat{\gamma}_\ell(\boldsymbol{\lambda}), \end{aligned} \tag{17}$$

where $\widehat{\gamma}_\ell(\boldsymbol{\lambda})$ is defined in (11). Then, the chosen $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\mathrm{AIC}}$ is such that:

$$\boldsymbol{\lambda}_{\mathrm{AIC}} = \underset{\boldsymbol{\lambda} \in \Lambda}{\operatorname{argmin}} \left(\mathrm{AIC}(\boldsymbol{\lambda})\right). \tag{18}$$

The $\mathrm{TPR}(\boldsymbol{\lambda})$ and $\mathrm{FPR}(\boldsymbol{\lambda})$ obtained with $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\mathrm{AIC}}$ for both functions f_1 and f_2 are displayed in Figure 8 for $n = 350$ and $n = 2000$ and for $\sigma = 0$ or $\sigma = 0.5$. We can observe for f_1 that with a sufficient number of observations this criterion allows us to get $\mathrm{TPR}(\boldsymbol{\lambda}_{\mathrm{AIC}}) = 1$ while having $\mathrm{FPR}(\boldsymbol{\lambda}_{\mathrm{AIC}}) = 0$ which means that the relevant variables are selected and not the irrelevant ones. The noise in the observation set has a stronger influence on the detection of the relevant variables of f_2 than for f_1. With $\sigma = 0.5$, we indeed have $\mathrm{TPR}(\boldsymbol{\lambda}_{\mathrm{AIC}}) < 1$ and $\mathrm{FPR}(\boldsymbol{\lambda}_{\mathrm{AIC}}) = 0$. By increasing the value of n from 350 to 2000, the value of $\mathrm{TPR}(\boldsymbol{\lambda}_{\mathrm{AIC}})$ is increased and thus the number of relevant selected variables. Moreover, for unnoisy set of observations the relevant variables are recovered from $n = 350$. Since we are interested in geochemical applications where the noise in the observation sets is very small, we will not be concerned by this issue.

3. Numerical experiments

In this section, we will assess the robustness of our method called ABSORBER implemented in the `absorber` R package when the variance of the noise σ^2 increases as well as the number of observations n. We will also study how this novel method behaves when the number of variables p grows. To demonstrate its efficiency, we will compare it to two state-of-the-art methods for feature selection: LassoNet introduced in (Lemhadri et al., 2021) and the widely used Random Forests (RF) introduced by (Breiman, 2001) and used in (Genuer et al., 2010) for variable selection.

LassoNet is an open-source package available on GitHub implemented in Python under the name `lassonet`. The algorithm generates a grid of penalization parameters and the corresponding selected variables. We can thus calculate the percentage of selection for each variable as defined in (15). In the following, the residual neural network is built by taking a hidden layer with 10 neurons as proposed in the notebook example given by this package which focuses on a 26-dimensional regression problem. Despite an increase of the number of neurons in the hidden layer and an extension of the maximal number of epochs, no significant improvement in the selection of relevant variables was observed. This is the reason why we did not explore more complex neural network architectures.

In order to apply Random Forests to our data, we used the R package `randomForest`, implemented by (Liaw et al., 2002), with 500 trees. This package provides the percentage of

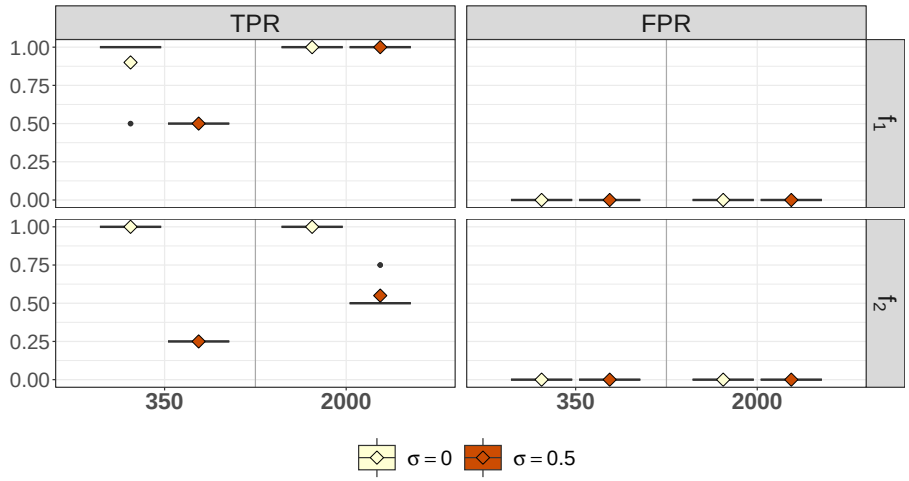

FIGURE 8. TPR($\boldsymbol{\lambda}$) and FPR($\boldsymbol{\lambda}$) values by choosing $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\text{AIC}}$ for f_1 (top) and f_2 (bottom) with an unnoisy ($\sigma = 0$) or noisy ($\sigma = 0.5$) set of observations. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

increased mean square error for each variable as the model excludes them one-by-one. This metric is then converted into a percentage of selection for each variable ensuring comparable results for the three methods.

3.1. **Influence of n and σ on the quality of variable selection.** In the following, we explore the impact of both the number of observations and the noise level in the observation sets on the efficiency of the three previously introduced variable selection methods.

As a result, the percentage of selection for each variable is calculated as described in (15) for 10 samplings of the observation set and the results are shown in Figure 9 (resp. Figure 10) for $f = f_1$ (resp. $f = f_2$) with n belonging to $\{350, 500, 700\}$ (resp. $\{350, 500\}$). The relevant variables are displayed in green and the irrelevant variables are displayed in red. Additional results are presented in the Appendix in Figure 16 (resp. Figures 17 and 18) for f_1 (resp. f_2). As we can see in Figure 9 and in Figure 16 of the Appendix, for a given n, the noise does not seem to have a significant influence on the percentage of variables of f_1 selected by our method ABSORBER. However, as we increase the value of n of the corrupted observation set with $\sigma = 0.5$, the percentage of irrelevant variables selected with our method drops from nearly 50% for $n = 350$ to 25% with $n = 1500$. As observed in these figures, LassoNet tends to select irrelevant variables since the variable selection percentage of the variables belonging to $\overline{\mathcal{V}_{f_1}}$ is nearly equal to that of variable 5 belonging to $\mathcal{V}_{f_1}$, both having a selection rate of 70% for $n = 350$ and $\sigma = 0.5$. Compared to our method and LassoNet, Random Forests selects variable 5 with only 30% of selection against 80% and 75%, respectively.

Let us now study the application of these methods to f_2. The corresponding results are displayed in Figure 10 and in Figures 17 and 18 of the Appendix. The noise has an effect on our method ABSORBER in this case since there is no selection of irrelevant variables regardless of the value of n with unnoisy observation sets against 25% of selection when $\sigma = 0.5$ for the smallest values of n. However, increasing the value of n in this case allows us to reduce

the percentage of selection for irrelevant variables to 10% while maintaining the minimum percentage of relevant variables up to 40%.

The two other methods appear to be unaffected by changes in both σ and n. Nevertheless, as observed previously with $f = f_1$, 50% of the penalization parameters of LassoNet select irrelevant variables. As a consequence, there is no distinct gap between these and relevant variable 7 since its percentage is very close to 50% as well. This statement suggests that using a high threshold on the percentage of selection obtained with LassoNet can result in omitting a relevant variable even for large n. Conversely, a low threshold includes irrelevant variables. In opposition to these two methods, the Random Forests approach tends to fail in detecting the relevant variables since variable 7 and 10 are selected nearly 0% and 5% of the time, respectively, regardless of σ and n. The same conclusion as with LassoNet can be drawn here, emphasizing that our method ABSORBER outperforms those two methods for variable selection while requiring only a few parameters to choose.

3.2. **Influence of p on the quality of variable selection.** In this section, we seek at studying the behavior of our method when the total number of variables p increases. To do so, we define two additional functions f_3 and f_4 such that:

$$\begin{aligned} f_3\left(x^{(1)},\dots,x^{(5)}\right) &= 1.8\cos\left(x^{(1)}\right)\sin\left(x^{(3)}+1\right) - 5\ln\left(x^{(3)}+1\right) - \frac{0.9}{(x^{(4)})^2+1} \\ &\left(x^{(1)},\dots,x^{(5)}\right) \in [0,1]^5, \end{aligned} \tag{19}$$

$$\begin{aligned} f_4\left(x^{(1)},\dots,x^{(50)}\right) &= 1.8\cos\left(x^{(1)}\right)\sin\left(x^{(7)}+1\right) - 5\ln\left(x^{(3)}+1\right) - \frac{0.9}{\left(x^{(10)}\right)^2+1}, \\ &\left(x^{(1)},\dots,x^{(50)}\right) \in [0,1]^{50}. \end{aligned} \tag{20}$$

We apply all three variable selection methods to f_2, f_3 and f_4 with observation sets of varying sizes n, all corrupted with the same noise levels as assessed in the previous section. Next, we compute the AUC and max(TPR−FPR) as defined in Section 2.3. The results for these comparisons are displayed in Figure 11 for $n = 350$ and $n = 2000$.

We can see from this figure that our method is affected by p when the observation set is corrupted with significant noise ($\sigma = 0.5$) and has a reduced size ($n = 350$). Specifically, $\max(\text{TPR} - \text{FPR}) = 1$ and $\text{AUC} = 0$ indicating that no irrelevant variables are selected, regardless of p with $\sigma \leq 0.25$. In contrast, the efficiency of the two other methods is impacted by the value of p as the metrics presented in Figure 11 are drastically deteriorated. For instance, with LassoNet $\max(\text{TPR} - \text{FPR}) = 1$ and $\text{AUC} = 1$ with $p = 5$ for $\sigma < 0.5$. However, $\max(\text{TPR} - \text{FPR})$ becomes less that 1 for $p = 10$ and it keeps decreasing as the values of p and σ increase. Even for the simpler case, $p = 5$ and $\sigma = 0$, Random Forests exhibit deteriorated results and this deterioration continues when both p and σ increase since $\max(\text{TPR} - \text{FPR})$ falls below 1. For both of these two methods, $\text{AUC} > 0$ indicating that $\text{FPR} > 0$ and suggesting that these methods select irrelevant variables while omitting one or more relevant ones. Increasing the number of observations up to $n = 2000$ improves the results for all three methods. Nevertheless, Random Forests continue to yield poor results, the only improvement being a reduction in variability between the different samplings. Our method is the only one showing satisfactory results as $\max(\text{TPR} - \text{FPR}) = 1$, regardless of p and σ. Globally, our method outperforms LassoNet and RF in this case and appears to be more robust when facing with higher dimensions p with or without noisy observation sets.

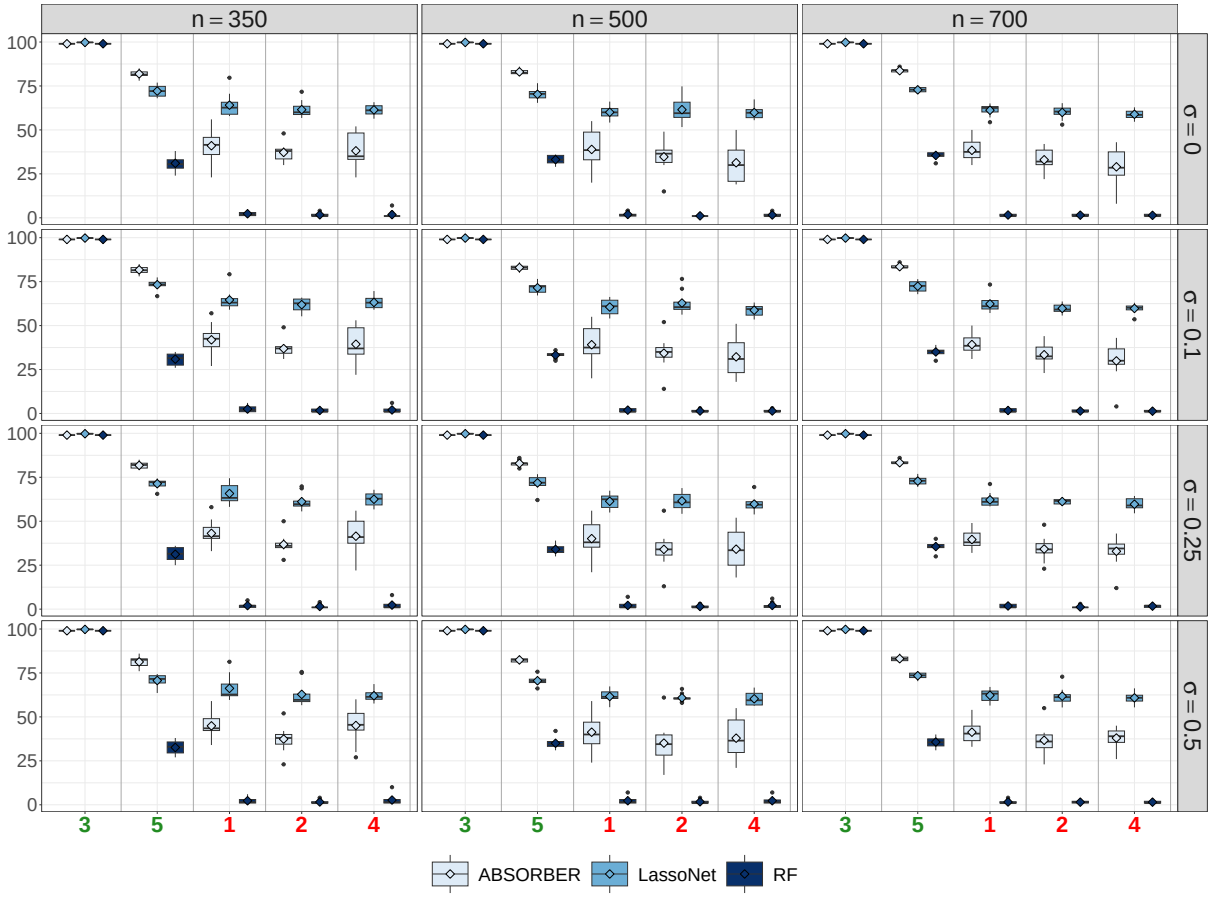

FIGURE 9. Percentage of selection of each variable of f_1 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

In Figure 12, the impact of the value p on our variable selection procedure using AIC is assessed. We can observe that only σ has a negative impact on the efficiency of our method. More precisely, the TPR values are smaller than 1, especially for $\sigma = 0.5$ regardless of n. However, for smaller noise levels ($\sigma < 0.25$) and with $n \geq 700$ our method enables $\text{TPR}(\lambda_{\text{AIC}}) = 1$ while maintaining $\text{FPR}(\lambda_{\text{AIC}}) = 0$. This means that no irrelevant variables are chosen, regardless of p, demonstrating the efficiency of our variable selection procedure.

3.3. **Numerical performance.** The goal of this section is to investigate the computational times of our variable selection approach implemented in the `absorber` R package. To this end,

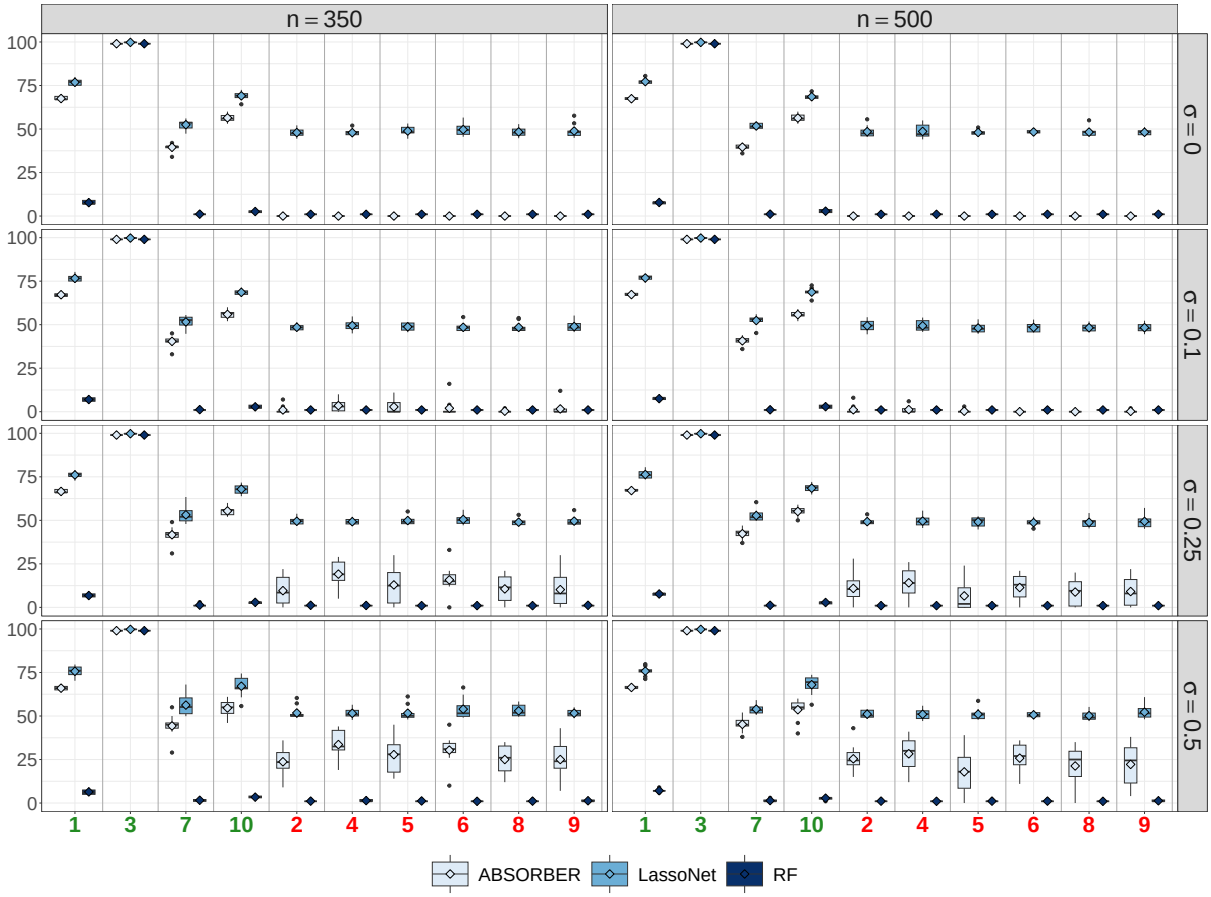

FIGURE 10. Percentage of selection of each variable of f_2 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

we introduce an additional function for which $p = 25$ defined as follows:

$$f_5\left(x^{(1)},\dots,x^{(25)}\right) = 1.8\cos\left(x^{(1)}\right)\sin\left(x^{(7)}+1\right) - 5\ln\left(x^{(3)}+1\right) - \frac{0.9}{\left(x^{(10)}\right)^2+1}, \qquad (21)$$
$$\left(x^{(1)},\dots,x^{(25)}\right) \in [0,1]^{25}.$$

Our variable selection method is applied to f_2, f_3, f_4 and f_5 (defined in (14), (19), (20) and (21), respectively) for $p = 5$, $p = 10$, $p = 25$ and $p = 50$, respectively. The timings were obtained on a workstation with 31.2GB of RAM and Intel Core i7 (3.8GHz) CPU. The $\log_{10}$-transformed average computational times and their standard deviation obtained from 20 independent executions are displayed in Figure 13. We can see from this figure that it only takes 250 seconds to perform variable selection on a function with $p = 50$ variables from

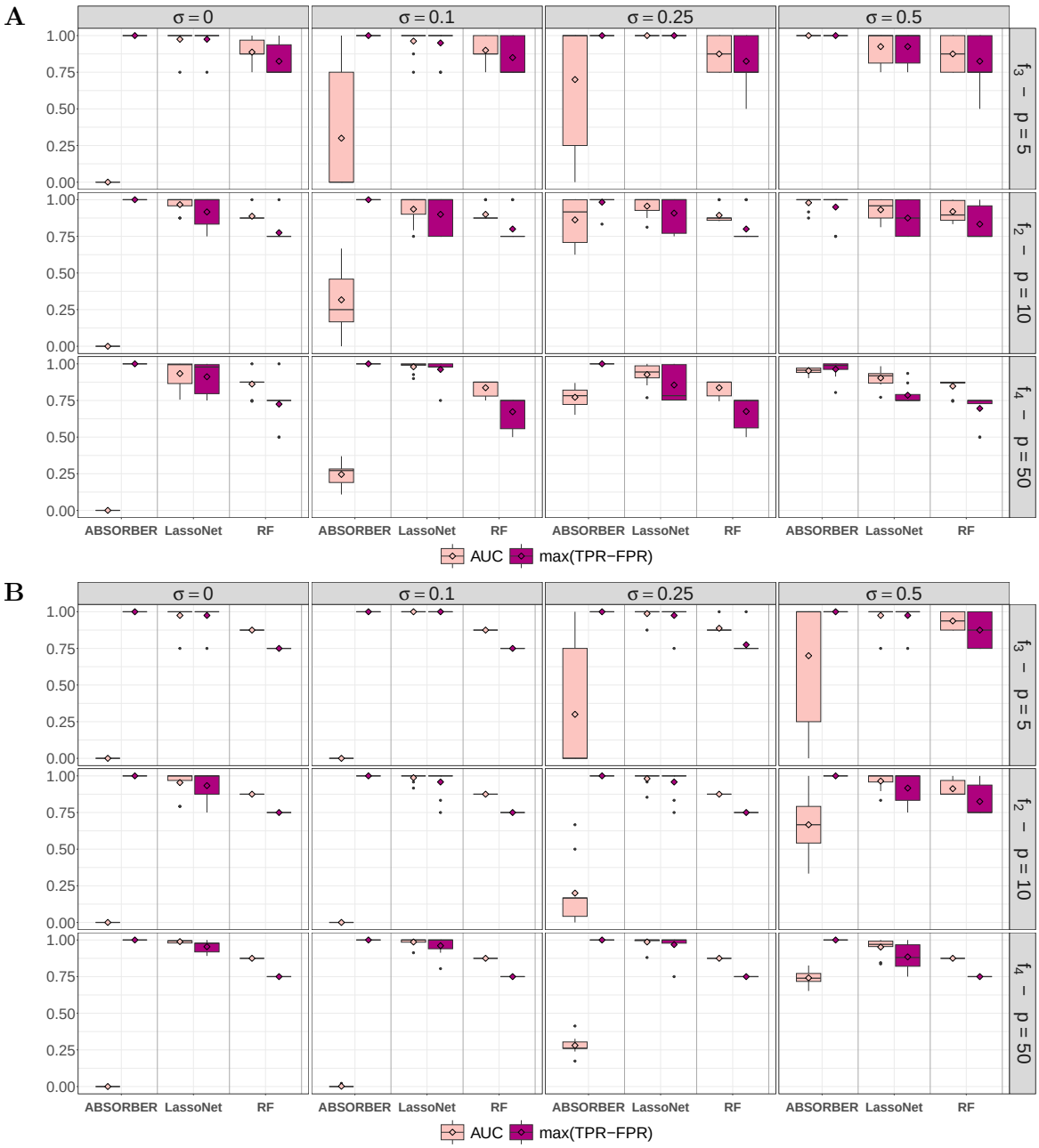

FIGURE 11. AUC and max(TPR − FPR) calculated for three different variable selection methods: ABSORBER, LassoNet and RandomForests (RF) applied to three functions (top to bottom) f_3, f_2 and f_4 with $n = 350$ (A) and $n = 2000$ (B) and an increasing value of σ (left to right). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

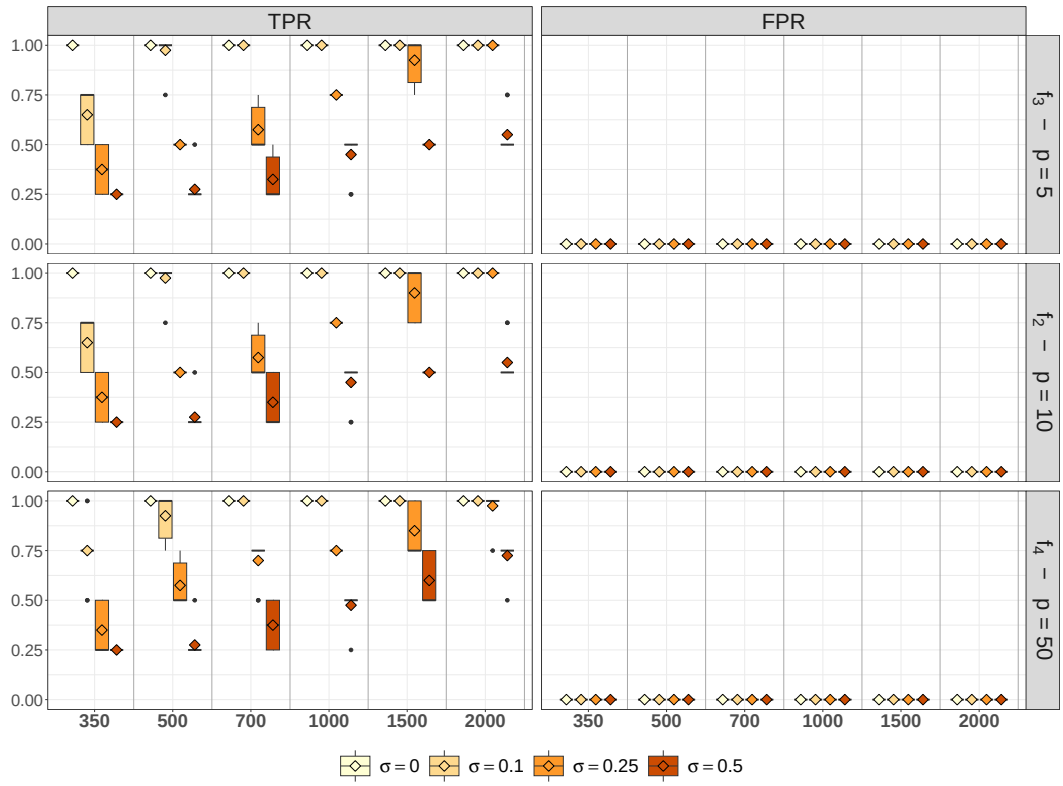

FIGURE 12. TPR($\boldsymbol{\lambda}$) and FPR($\boldsymbol{\lambda}$) values by choosing $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\text{AIC}}$ for f_3, f_2 and f_4 with three noise levels in the observation sets. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

$n = 2000$ observations. It has to be noticed that the execution times reported are mainly due to the use of the `sparsegl` package.

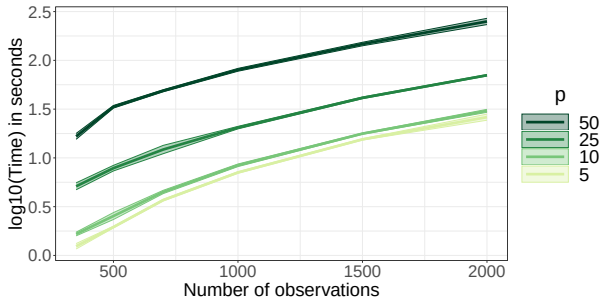

FIGURE 13. Execution times for an increasing number of observations n and values of p. The tested functions are here f_2, f_3, f_4 and f_5 with unnoisy observation sets ($\sigma = 0$).

4. Application to a multidimensional geochemical system

In this section, we apply our variable selection method to real geochemical systems. We start by defining the geochemical system derived from the calcite dissolution and precipitation study in (Kolditz et al., 2012). In the following, the observation sets are generated using the geochemical solver PHREEQC as in (Parkhurst and Appelo, 2013). The corresponding thermodynamic data for aqueous species and minerals are available in the Phreeqc.dat distributed with PHREEQC. The compositional system actually solved consists of 14 species in solution, 2 mineral components, 8 geochemical reactions and 2 mineral dissolution-precipitation reactions. For the purposes of this paper, we specifically focus on the calcite precipitation/dissolution:

$$\text{Calcite} \rightleftharpoons \text{Ca}^{2+} + \text{CO}_3^{2-},\ \text{logK} = -8.48.$$

The previous equation represents the dissolution reaction along with its corresponding $\log_{10}$-transformed equilibrium constant value logK. The amount of calcite can be here computed with PHREEQC as a function of the total elemental concentrations (C, Ca), the pH (as $-\log(\text{H}^+)$) and the amount of calcite initially present. The pH is here fixed at 9.8 and we introduce two additional concentrations (K, Cl) which do not participate in the calcite precipitation. This allows us to assess our variable selection on real datasets. We can thus formulate the problem as:

$$\begin{aligned}\text{Calcite} &= f_6(\text{C}^*, \text{Ca}^*, \text{K}^*, \text{Cl}^*, \text{Calcite}^*)\\ &= \widetilde{f_6}(\text{C}^*, \text{Ca}^*, \text{Calcite}^*),\end{aligned}$$

where C^*, Ca^*, K^*, Cl^*, Calcite^* are the normalized concentrations and quantities initially present of C, Ca, K, Cl and Calcite, respectively. The normalization of each variable is done by taking into account the minimal and the maximal bound of the values so that each variable belongs to $[0, 1]$. We also define another function f_7 which takes into account known fake variables to study the behavior of all three methods:

$$\begin{aligned}\text{Calcite} &= f_7(\text{C}^*, \text{Ca}^*, \text{K}^*, \text{Cl}^*, \text{Calcite}^*, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)})\\ &= \widetilde{f_6}(\text{C}^*, \text{Ca}^*, \text{Calcite}^*),\end{aligned}$$

where $x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}$ are synthetic irrelevant variables obtained through uniform sampling between 0 and 1. Hereafter, $\mathcal{V}_{f_6} = \mathcal{V}_{f_7} = \{\text{C}^*, \text{Ca}^*, \text{Calcite}^*\}$, $\overline{\mathcal{V}_{f_6}} = \{\text{K}^*, \text{Cl}^*\}$ and $\overline{\mathcal{V}_{f_7}} = \{\text{K}^*, \text{Cl}^*, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}\}$.

The results for the application of our method, LassoNet and RF to f_6 and f_7 are displayed in Figure 14. Here, our method consistently selects the relevant variables belonging to $\mathcal{V}_{f_6}$ 62.5% of the time and the irrelevant ones belonging to $\overline{\mathcal{V}_{f_6}}$ are almost never selected. By increasing n, the percentage of selection for these irrelevant variables reaches 0%. In contrast, LassoNet selects all the variables and fails to discriminate the relevant from the irrelevant ones. Random Forests, on the other hand, tends to detect only one variable of $\mathcal{V}_{f_6}$ which is the calcite quantity. Furthermore, it tends to select K and Cl (5%) more than the relevant variables C and Ca (0%), even with $n = 2000$.

Adding fake variables does not deteriorate our method which continues to select only the relevant variables in $\mathcal{V}_{f_7}$. However, it does not improve the results for LassoNet which continues to select all ten variables from $\mathcal{V}_{f_7}$ and $\overline{\mathcal{V}_{f_7}}$. Interestingly, it helps the Random Forests approach in detecting the relevant variables C and Ca resulting in an increase in the percentage of selection up to 5%. Nevertheless, this emphasizes the efficiency of our method which outperforms the other two in this geochemical case.

Furthermore, we used the AIC to select the parameter $\boldsymbol{\lambda}$ and to automatically choose the relevant variables. The corresponding results are shown in Figure 15. This statistical criterion proves to be highly efficient as evidenced by $\mathrm{TPR}(\boldsymbol{\lambda}_{\mathrm{AIC}}) = 1$ and $\mathrm{FPR}(\boldsymbol{\lambda}_{\mathrm{AIC}}) = 0$, regardless of n and p.

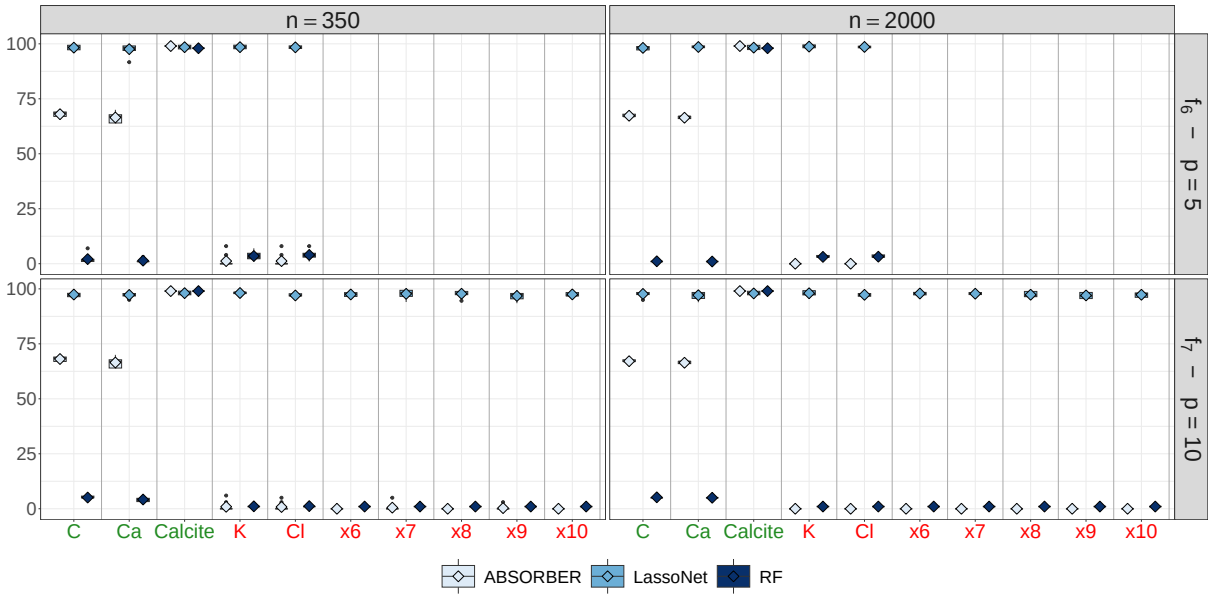

FIGURE 14. Percentage of selection of each variable of f_6 (top) and f_7 (bottom) with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

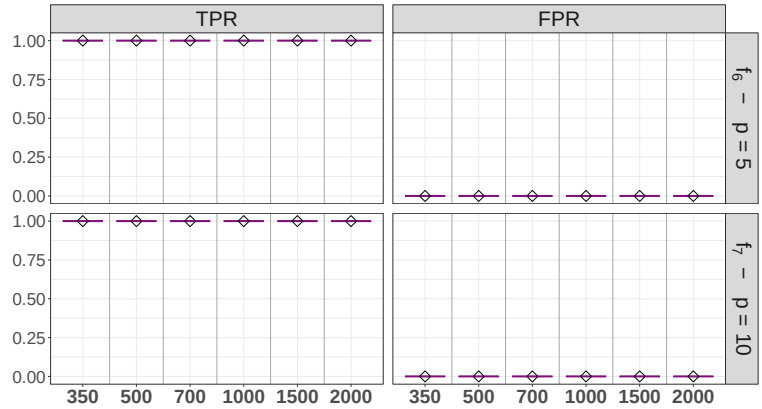

FIGURE 15. TPR($\boldsymbol{\lambda}$) and FPR($\boldsymbol{\lambda}$) values by choosing $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\text{AIC}}$ for f_6 (top) and f_7 (bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory, pp. 267–281. In B.N.Petrov, F.Csaki (Eds.).

Antoniadis, A., I. Gijbels, and A. Verhasselt (2012). Variable selection in additive models using P-splines. Technometrics 54(4), 425–438.

Asher, M. J., B. F. Croke, A. J. Jakeman, and L. J. Peeters (2015). A review of surrogate models and their application to groundwater modeling. Water Resources Research 51(8), 5957–5973.

Breiman, L. (2001). Random forests. Machine learning 45, 5–32.

Chen, Y., Q. Gao, F. Liang, and X. Wang (2021). Nonlinear variable selection via deep neural networks. Journal of Computational and Graphical Statistics 30(2), 484–492.

De Boor, C. (1978). A practical guide to splines, Volume 27. Springer-Verlag New York.

Demirer, E., E. Coene, A. Iraola, A. Nardi, E. Abarca, A. Idiart, G. de Paola, and N. Rodríguez-Morillas (2023). Improving the performance of reactive transport simulations using artificial neural networks. Transport in Porous Media 149(1), 271–297.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. Statistical science 11(2), 89–121.

Feng, J. and N. Simon (2017). Sparse-input neural networks for high-dimensional nonparametric regression and classification. arXiv preprint arXiv:1711.07592.

Feng, J. and N. Simon (2022). Ensembled sparse-input hierarchical networks for high-dimensional datasets. Statistical Analysis and Data Mining: The ASA Data Science Journal 15(6), 736–750.

Friedman, J. H. (1991). Multivariate adaptive regression splines. The annals of statistics 19(1), 1–67.

Genuer, R., J.-M. Poggi, and C. Tuleau-Malot (2010). Variable selection using random forests. Pattern recognition letters 31(14), 2225–2236.

Gijbels, I., A. Verhasselt, and I. Vrinssen (2015). Variable selection using P-splines. Wiley Interdisciplinary Reviews: Computational Statistics 7(1), 1–20.

Guérillot, D. and J. Bruyelle (2020). Geochemical equilibrium determination using an artificial neural network in compositional reservoir flow simulation. Computational Geosciences 24(2), 697–707.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The Elements of Statistical Learning: Data mining, inference, and prediction. New York, NY, USA: Springer New York Inc.

He, K., X. Zhang, S. Ren, and J. Sun (2016, June). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. Annals of statistics 38(4), 2282.

Jatnieks, J., M. De Lucia, D. Dransch, and M. Sips (2016). Data-driven surrogate model approach for improving the performance of reactive transport simulations. Energy Procedia 97, 447–453.

Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar (2018). Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering 31(8), 1544–1554.

Kolditz, O., U.-J. Görke, H. Shao, and W. Wang (2012). Thermo-hydro-mechanical-chemical processes in porous media: benchmarks and examples, Volume 86. Springer Science & Business Media.

Laloy, E. and D. Jacques (2019). Emulation of cpu-demanding reactive transport models: a comparison of gaussian processes, polynomial chaos expansion, and deep neural networks. Computational Geosciences 23, 1193–1215.

Laloy, E. and D. Jacques (2022). Speeding up reactive transport simulations in cement systems by surrogate geochemical modeling: deep neural networks and k-nearest neighbors. Transport in Porous Media 143(2), 433–462.

Lary, D. J., A. H. Alavi, A. H. Gandomi, and A. L. Walker (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers 7(1), 3–10.

Leal, A. M., D. A. Kulik, and M. O. Saar (2017). Ultra-fast reactive transport simulations when chemical reactions meet machine learning: chemical equilibrium. arXiv preprint arXiv:1708.04825.

Lemhadri, I., F. Ruan, L. Abraham, and R. Tibshirani (2021). Lassonet: A neural network with feature sparsity. The Journal of Machine Learning Research 22(1), 5633–5661.

Li, Y., C.-Y. Chen, and W. W. Wasserman (2016). Deep feature selection: theory and application to identify enhancers and promoters. Journal of Computational Biology 23(5), 322–336.

Liang, F., Q. Li, and L. Zhou (2018). Bayesian neural networks for selection of drug sensitive genes. Journal of the American Statistical Association 113(523), 955–972.

Liang, X., A. Cohen, A. S. Heinsfeld, F. Pestilli, and D. J. McDonald (2022). sparsegl: An r package for estimating sparse group lasso. arXiv preprint arXiv:2208.02942.

Liaw, A., M. Wiener, et al. (2002). Classification and regression by randomforest. R news 2(3), 18–22.

Lin, Y. and H. H. Zhang (2006). Component selection and smoothing in multivariate nonparametric regression. The annals of statistics 34(5), 2272–2297.

Lu, Y., Y. Fan, J. Lv, and W. Stafford Noble (2018). Deeppink: reproducible feature selection in deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Advances in Neural Information Processing Systems, Volume 31. Curran Associates, Inc.

Parkhurst, D. L. and C. Appelo (2013). Description of input and examples for PHREEQC version 3: a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. U.S.G.S. Techniques and Methods.

Prasianakis, N. I., R. Haller, M. Mahrous, J. Poonoosamy, W. Pfingsten, and S. V. Churakov (2020). Neural network based process coupling and parameter upscaling in reactive transport simulations. Geochimica et Cosmochimica Acta 291, 126–143.

Radchenko, P. and G. M. James (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. Journal of the American Statistical Association 105(492), 1541–1553.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009). Sparse additive models. Journal of the Royal Statistical Society Series B: Statistical Methodology 71(5), 1009–1030.

Razavi, S., B. A. Tolson, and D. H. Burn (2012). Review of surrogate modeling in water resources. Water Resources Research 48(7).

Rosasco, L., M. Santoro, S. Mosci, A. Verri, and S. Villa (2010). A regularization approach to nonlinear variable selection. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 653–660. JMLR Workshop and Conference Proceedings.

Savino, M., C. Lévy-Leduc, M. Leconte, and B. Cochepin (2022). An active learning approach for improving the performance of equilibrium based chemical simulations. Computational Geosciences 26(2), 365–380.

Savino, M. E. and C. Lévy-Leduc (2023). A novel approach for estimating functions in the multivariate setting based on an adaptive knot selection for b-splines with an application to a chemical system used in geoscience. arXiv preprint arXiv:2306.00686.

Steefel, C. I. (2019). Reactive transport at the crossroads. Reviews in Mineralogy and Geochemistry 85(1), 1–26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58(1), 267–288.

Yamada, M., W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama (2014). High-dimensional feature selection by feature-wise kernelized lasso. Neural computation 26(1), 185–207.

Ye, M. and Y. Sun (2018). Variable selection via penalized neural network: a drop-out-one loss approach. In International Conference on Machine Learning, pp. 5620–5629. PMLR.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B: Statistical Methodology 68(1), 49–67.

Zhu, G. and T. Zhao (2021). Deep-gknock: nonlinear group-feature selection with deep neural networks. Neural Networks 135, 139–147.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology 67(2), 301–320.

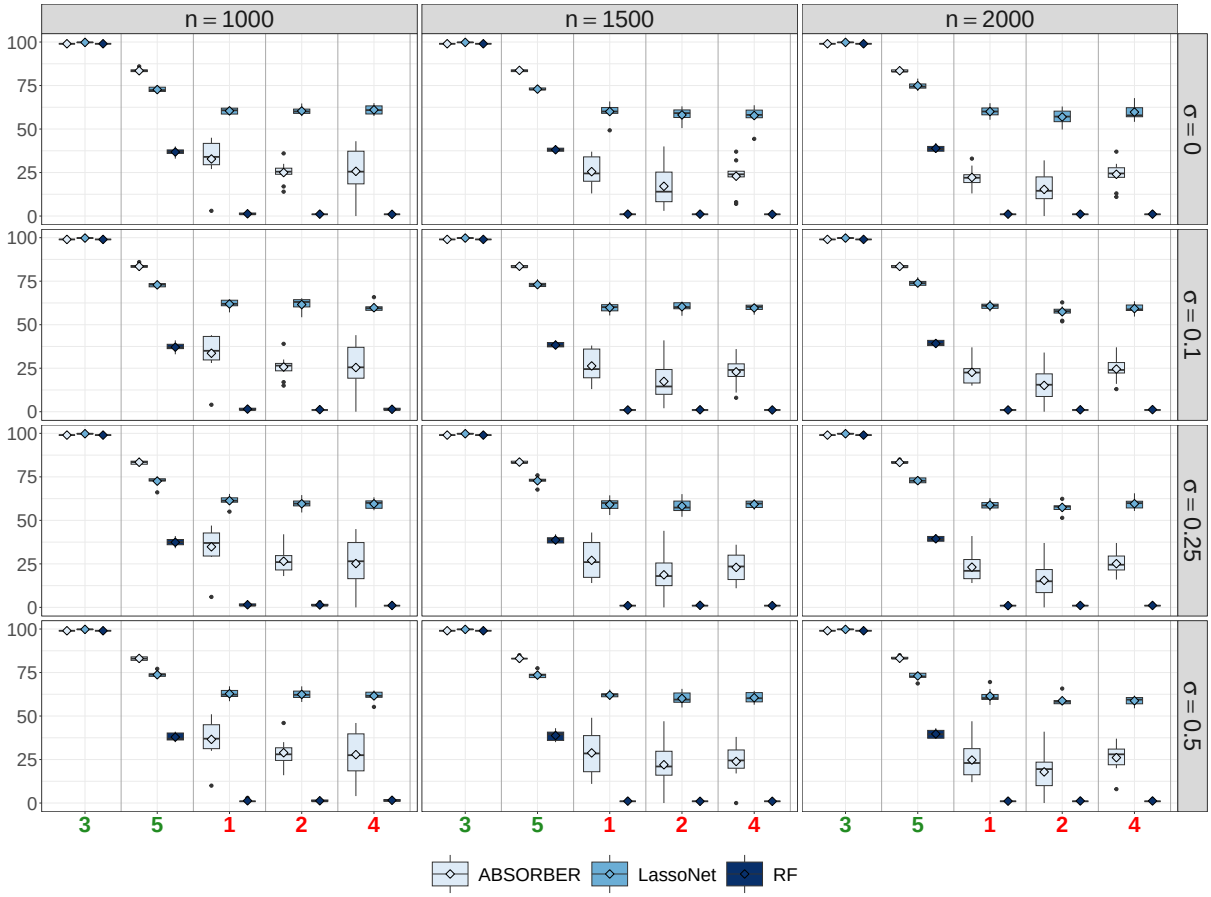

FIGURE 16. Percentage of selection of each variable of f_1 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

5. Appendix: Additional plots

ANDRA, 1/7 RUE JEAN MONNET, 92290 CHÂTENAY-MALABRY, FRANCE AND UNIVERSITÉ PARIS-SACLAY, AGROPARISTECH, INRAE, UMR MIA PARIS-SACLAY, 91120, PALAISEAU, FRANCE

UNIVERSITÉ PARIS-SACLAY, AGROPARISTECH, INRAE, UMR MIA PARIS-SACLAY, 91120, PALAISEAU, FRANCE

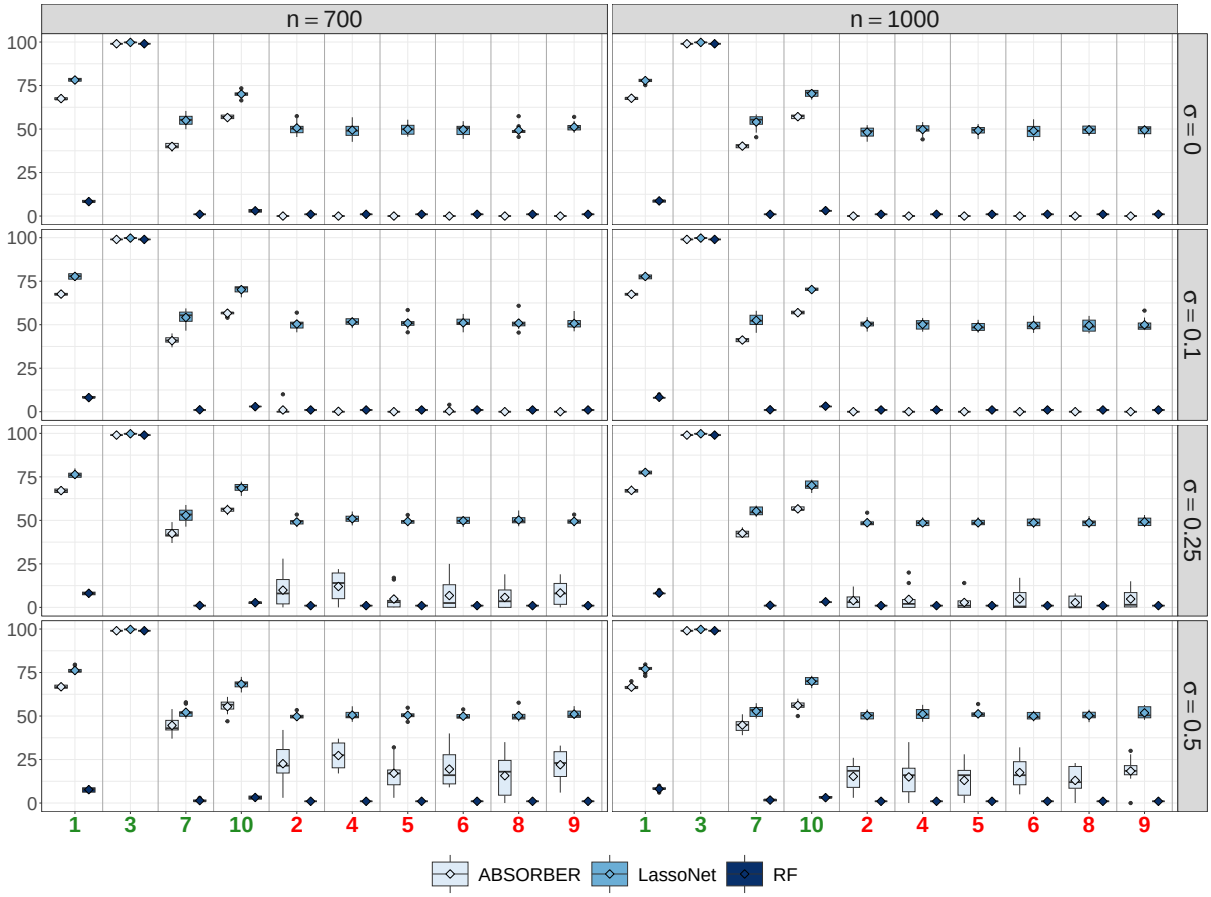

FIGURE 17. Percentage of selection of each variable of f_2 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

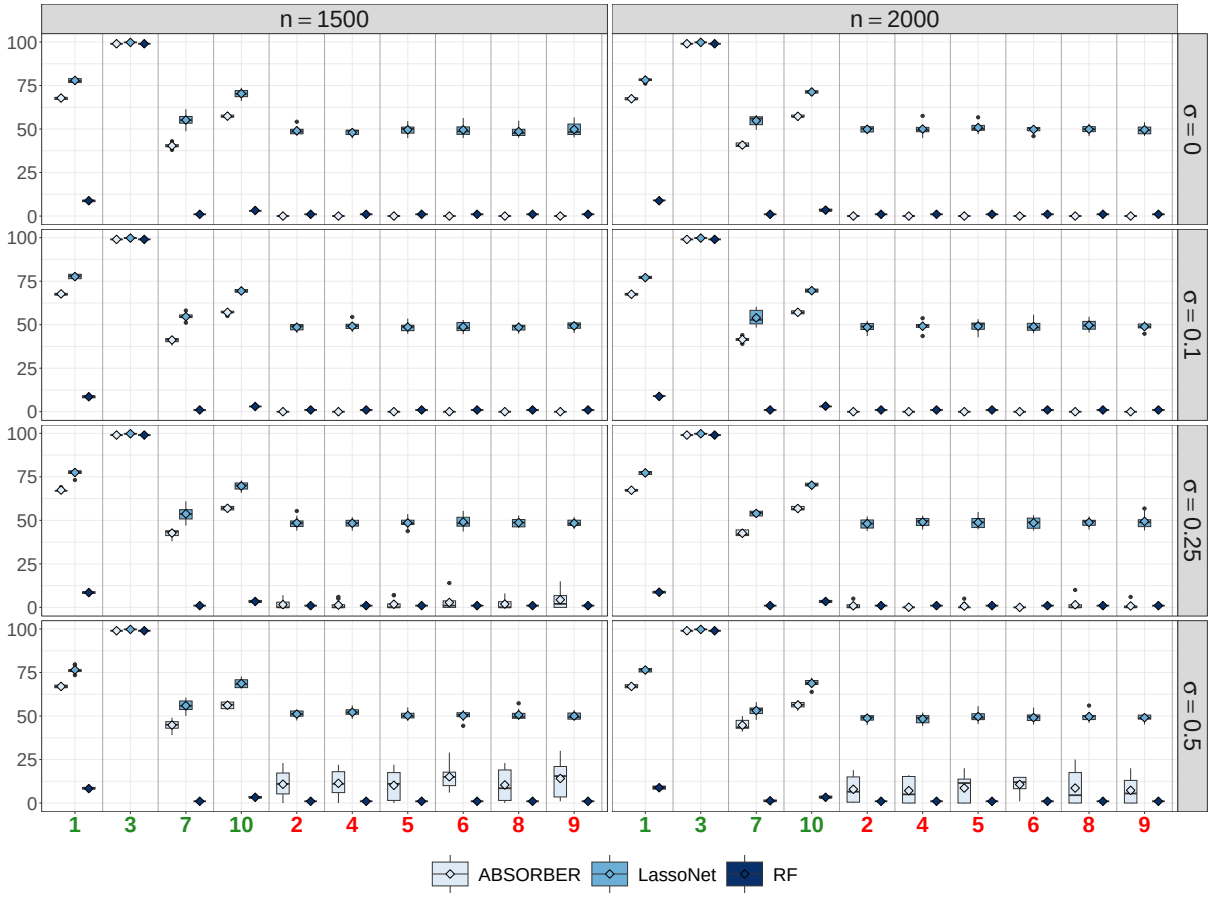

FIGURE 18. Percentage of selection of each variable of f_2 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.