A variable selection method in a multivariate nonparametric regression model: Application to geoscience

Mary E. Savino [1] & Céline Lévy-Leduc [2]

[1] *Andra, 92290 Châtenay-Malabry, France and Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120 Palaiseau, France*
mary.savino@agroparistech.fr
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120 Palaiseau, France*
celine.levy-leduc@agroparistech.fr

Résumé. Nous présentons ici une nouvelle méthode de sélection de variables dans un modèle de régression non-paramétrique multivarié et reposant sur des données afin d'identifier les variables dont dépend réellement la fonction de régression. Cette méthode consiste à approcher la fonction sous-jacente par une combinaison linéaire de B-splines d'ordre M ainsi que par la combinaison de leurs interactions deux-à-deux. Les coefficients de cette combinaison linéaire sont estimés en minimisant le critère des moindres carrés pénalisé par la somme des normes ℓ_2 des dérivées partielles par rapport à chaque variable dont dépend la fonction. Nous montrons que la méthode proposée peut être reformulée sous la forme d'un critère de type Group Lasso. Nous validons notre approche à travers différentes expériences numériques en faisant notamment varier le nombre d'observations, le niveau de bruit et le nombre total de variables. Nous la comparons également à deux autres méthodes de l'état de l'art et une application à un système géochimique réel est présentée. A travers ces différentes applications, notre approche démontre de meilleures performances statistiques que les autres méthodes auxquelles nous l'avons comparée. Notre méthode est implémentée dans le package R `absorber` qui sera bientôt disponible sur le "Comprehensive R Archive Network" (CRAN).

Mots-clés. Sélection de variables, régression non paramétrique, B-splines, Group Lasso

Summary. In this presentation, 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 underlying idea of our method consists in approximating the 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 penalized least-squares criterion. The penalization consists of the sum of the ℓ_2-norms of the partial derivatives with respect to the different variables on which the function depends. We show that our proposed method can be reformulated as a Group Lasso problem. We investigate the statistical performance of our approach through numerical experiments varying the number of observations, the noise level and the total number of variables. We also compare it to two other state-of-the-art methods. An application to a geochemical system is also proposed. In these different frameworks, our approach exhibits better performance than the other methods. Our completely data-driven method is implemented in the `absorber` R package which will be soon available on the Comprehensive R Archive Network (CRAN).

Keywords. Variable selection, nonparametric regression, B-splines, Group Lasso

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. 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). 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 $\widetilde{x}$.

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) depends. Our approach, presented in Section 2, consists in 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 show that our proposed method can be reformulated as a Group Lasso problem defined by Yuan and Lin (2006). Two different approaches to choose the penalization parameter are presented. The statistical performance of our approach are investigated in Section 3 and a geochemical application is given in Section 4.

2 Method

2.1 Approximation of f using B-splines

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:

$$\tau_{\ell,1} = \ldots = \tau_{\ell,M} = x_{min}^{(\ell)},$$

$$\tau_{\ell,j+M} = t_{\ell,j}, \quad j = 1, \ldots, K,$$

$$\tau_{\ell,K+M+1} = \ldots = \tau_{\ell,K+2M} = x_{max}^{(\ell)},$$

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

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 $\boldsymbol{\beta}_{\ell,j} = \left(\beta_{1,1}^{(\ell,j)} \ \beta_{1,2}^{(\ell,j)} \ \ldots \ \beta_{K+M,K+M}^{(\ell,j)}\right)^T$ 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) through the minimization of the following regularized criterion:

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

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$, 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 $x_i = (x_i^{(1)}, \ldots, x_i^{(p)})$ 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:

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

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 $(\partial_\ell \Phi_{\ell\bullet}) = \big((\partial_\ell \Phi_{\ell(\ell+1)}) \ldots (\partial_\ell \Phi_{\ell p})\big)$, $(\partial_\ell \Phi_{\bullet\ell}) = \big((\partial_\ell \Phi_{1\ell}) \ldots (\partial_\ell \Phi_{(\ell-1)\ell})\big)$, $\boldsymbol{\beta}_{\ell\bullet} = \left(\boldsymbol{\beta}_{\ell,(\ell+1)}^T \ldots \boldsymbol{\beta}_{\ell,p}^T\right)^T$ and $\boldsymbol{\beta}_{\bullet\ell} = \left(\boldsymbol{\beta}_{1,\ell}^T \ldots \boldsymbol{\beta}_{(\ell-1),\ell}^T\right)^T$, the penalty term can be written as:

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

where $\boldsymbol{\gamma}_\ell = \left(\boldsymbol{\beta}_\ell^T \ \boldsymbol{\beta}_{\ell\bullet}^T \ \boldsymbol{\beta}_{\bullet\ell}^T\right)^T$. The least-squares term can be rewritten as follows:

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

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:

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

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:

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

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. The coefficients $\widehat{\boldsymbol{\gamma}}_\ell(\lambda)$ are thus obtained as follows:

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

Thus, we define the active variables for each λ belonging to a given set Λ as follows:

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

where $\widehat{\gamma}_{\ell,k}(\lambda)$ is the kth coefficient of $\widehat{\boldsymbol{\gamma}}_\ell(\lambda)$. We also introduce the set $\mathcal{V}_f$ of the indices of the d relevant variables on which f in (1) actually depends that have to be selected among the p variables. We also denote the set $\overline{\mathcal{V}_f}$ of the indices of the irrelevant variables on which f does not depend.

3 Numerical experiments

In this section, we will investigate the statistical performance of our method called ABSORBER and 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 the efficiency of our method, 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), using their default parameters.

3.1 Metrics and selection criteria

We first introduce metrics to assess the efficiency of our method:

- **True Positive Rate (TPR)** and the **False Positive Rate (FPR)**, for each λ:

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

  where $d < p$, $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$, $\mathrm{TP}(\lambda)$ and $\mathrm{FP}(\lambda)$ are the number of true selected variables and the number of false selected variables for λ, respectively. $\mathcal{V}_\lambda$, $\mathcal{V}_f$ and $\overline{\mathcal{V}_f}$ and are introduced in (12) and in the text following this equation.

We also propose two criteria to choose λ. One allows the user to choose a threshold of percentage of selection and the other leverages the Akaike Information Criterion (AIC), introduced in Akaike (1973) to automatically choose λ. Both are defined as follows:

- **Percentage of variable selection**, for each variable ℓ belonging to $\{1, \ldots, p\}$:

$$P_\ell = \frac{100}{|\Lambda|} \sum_{\lambda \in \Lambda} \mathbb{1}\{\ell \in \mathcal{V}_\lambda\}\,, \tag{13}$$

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

- **AIC**:

$$\mathrm{AIC}(\lambda) = n \ln\left(\frac{\mathrm{RSS}(\lambda)}{n}\right) + 2T_\lambda\,, \tag{14}$$

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

$$\mathrm{RSS}(\lambda) = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}(\lambda)\right\|_2^2 \quad \text{with} \quad \widehat{\mathbf{Y}}(\lambda) = \sum_{\ell=1}^{p} \Theta_\ell \widehat{\boldsymbol{\gamma}}_\ell(\lambda), \tag{15}$$

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

$$\lambda_{\mathrm{AIC}} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \left(\mathrm{AIC}(\lambda)\right). \tag{16}$$

3.2 Results

We apply ABSORBER, LassoNet and RF on noisy observations satisfying (1) with $f = f_1$ defined as:

$$f_1(x^{(1)}, \ldots, x^{(10)}) = 1.8 \cos\left(x^{(1)}\right) \sin\left(x^{(7)} + 1\right) - 5 \ln\left(x^{(3)} + 1\right) - \frac{0.9}{{x^{(10)}}^2 + 1}, \qquad (17)$$
$$(x^{(1)}, \ldots, x^{(10)}) \in [0, 1]^{10}.$$

and we calculate the percentage of selection defined in the previous section. Here, $\mathcal{V}_{f_1} = \{1, 3, 7, 10\}$ are the relevant variables to be selected. Note that we use $K = 1$ evenly spaced knots in the B-splines bases involved in ABSORBER. We refer the reader to Savino and Lévy-Leduc (2024) for a numerical investigation justifying this choice.

Similarly to ABSORBER, the percentage of selection can be computed for each penalization parameter of LassoNet. Concerning the RF method, we convert the percentage of increased mean square error for each variable as the model excludes them one-by-one into a percentage of selection for each of them.

The results obtained for $n = 700$ and $n = 2000$ observations with ten random samplings of the set are displayed in Figure 1. Despite being sensitive to the noise level σ in the obser-

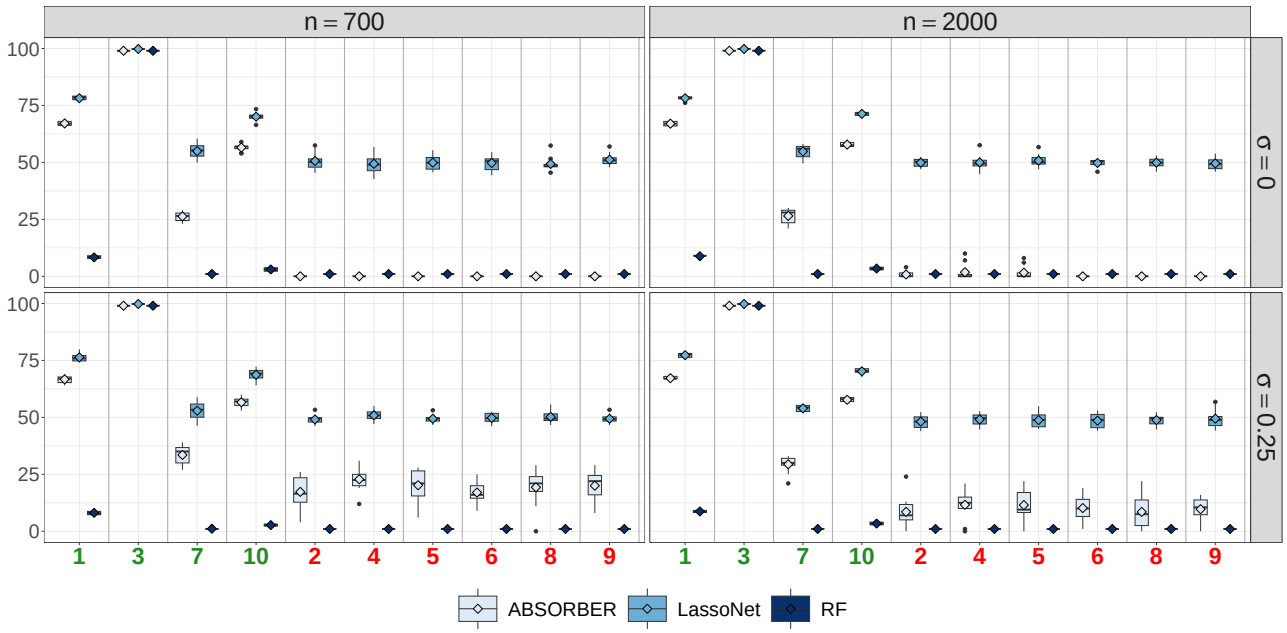

Figure 1: Percentage of selection of each variable of f_1 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

vations, ABSORBER succeeds in selecting the relevant variables with a satisfying percentage (between 25% and 40%) and allows us to choose a threshold at the visible gap (25 %) to

select the correct variables. In contrast, LassoNet and Random Forests either select irrelevant variables at a high percentage (50%) or fail to select the relevant variables, respectively. Increasing the number of observations allows us to reduce the percentage of selection for irrelevant variables with ABSORBER to 12% while maintaining the minimum percentage of relevant variables up to 30%. The two other methods appear to be unaffected by changes in n. These conclusions emphasize that our method outperforms those two methods for variable selection while requiring only a few parameters to choose.

We then apply our variable selection method using λ_{AIC} as described in (16) and we calculate the value of $\text{TPR}(\lambda_{\text{AIC}})$ and $\text{FPR}(\lambda_{\text{AIC}})$. We can observe that increasing σ slightly alters the performance of our method. However, for smaller noise levels ($\sigma < 0.25$) and with $n \geq 700$ our method enables $\text{TPR}(\lambda_{\text{AIC}}) = 1$ while maintaining $\text{FPR}(\lambda_{\text{AIC}})$ to a value smaller than 0.05. This means that almost no irrelevant variables are chosen demonstrating the efficiency of our variable selection procedure.

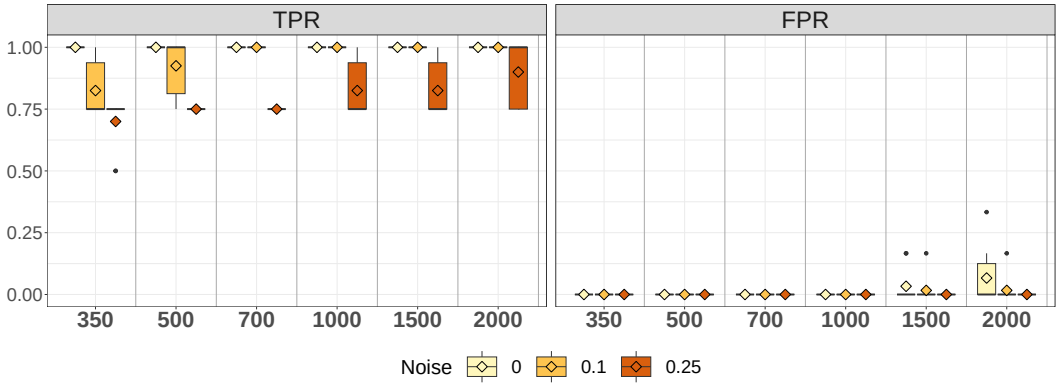

Figure 2: $\text{TPR}(\lambda)$ and $\text{FPR}(\lambda)$ values by choosing $\lambda = \lambda_{\text{AIC}}$ for f_1 with three noise levels in the observation sets. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

4 A geochemical application

We propose applying our method ABSORBER to a real geochemical system derived from the calcite dissolution and precipitation study in Kolditz et al. (2012). Our observation sets are here generated using the geochemical solver PHREEQC as in Parkhurst and Appelo (2013). For the purposes of this study, we specifically focus on the calcite precipitation/dissolution such that we can define a function f_2 depending on normalized concentrations and quantities of elements:

$$\begin{aligned}\text{Calcite} &= f_2(\text{C}^*, \text{Ca}^*, \text{K}^*, \text{Cl}^*, \text{Calcite}^*, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}) \\ &= \widetilde{f}_2(\text{C}^*, \text{Ca}^*, \text{Calcite}^*),\end{aligned}$$

where Calcite denotes the amount of produced calcite and 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 $[0, 1]$. $x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}$ are synthetic irrelevant variables obtained through uniform sampling between 0 and 1. Hereafter, $\mathcal{V}_{f_2} = \{\text{C}^*, \text{Ca}^*, \text{Calcite}^*\}$ and $\overline{\mathcal{V}_{f_2}} = \{\text{K}^*, \text{Cl}^*\, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}\}$. We aim at retrieving $\mathcal{V}_{f_2}$ by applying our method.

The results for the application of ABSORBER, LassoNet and RF to f_2 are displayed in the left part of Figure 3. Our method consistently succeeds in selecting the relevant variables belonging to $\mathcal{V}_{f_2}$ (62.5% of selection) while discarding the irrelevant ones (almost 0% of selection) even with a small dataset size $n = 350$. 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 $\mathcal{V}_{f_2}$. This shows once again that our method outperforms the other two in this geochemical case.

Furthermore, we used the AIC to select the parameter λ and to automatically choose the relevant variables. The corresponding results are shown in right part of Figure 3. This statistical criterion proves to be highly efficient as evidenced by $\text{TPR}(\lambda_{\text{AIC}}) = 1$ and $\text{FPR}(\lambda_{\text{AIC}}) = 0$, regardless of n.

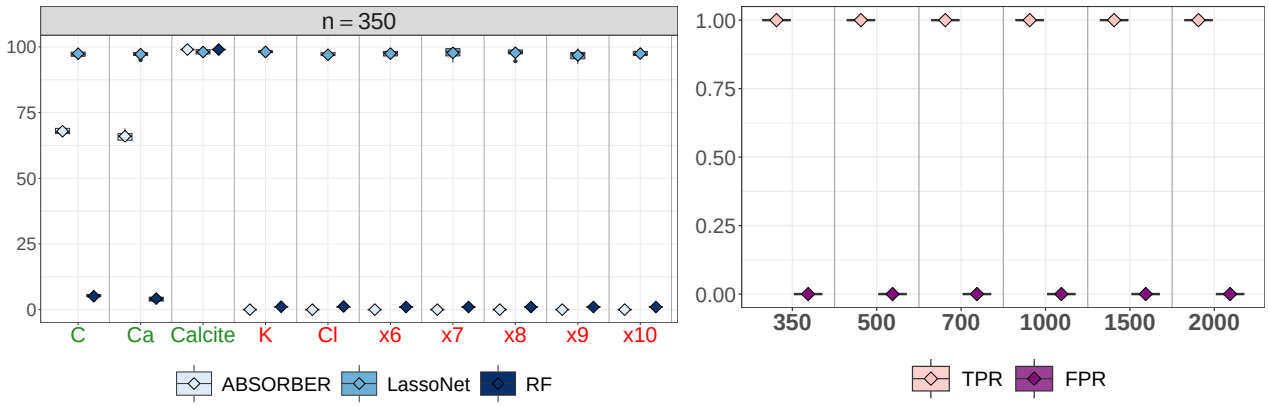

Figure 3: (Left) Percentage of selection of each variable of f_2 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with $n = 350$ observations. (Right) $\text{TPR}(\lambda)$ and $\text{FPR}(\lambda)$ values by choosing $\lambda = \lambda_{\text{AIC}}$ for f_2 with an increasing number of observations n. 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.).

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.

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

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.

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.

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.

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.

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.

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.

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. E. and C. Lévy-Leduc (2024). A novel variable selection method in a nonlinear multivariate model using B-splines with an application to geoscience. hal-04434820.

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.