Statistical learning methods for nonlinear geochemical problems

Mary Edith Savino

université PARIS-SACLAY

Statistical learning methods for nonlinear geochemical problems

Méthodes d'apprentissage statistique pour l'étude de problèmes géochimiques non linéaires

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°574 mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques.
Référent : Faculté des sciences d'Orsay

Thèse préparée au sein de l' **UMR MIA Paris-Saclay (Université Paris-saclay, AgroParisTech, INRAE)**, sous la direction de **Céline LÉVY-LEDUC**, professeure de statistique à AgroParisTech, le co-encadrement de **Benoit COCHEPIN**, ingénieur en analyse de performance à l'Andra et le co-encadrement de **Marc LECONTE**, ingénieur en calcul scientifique à l'Andra.

Thèse soutenue à Paris-Saclay, le 30 Septembre 2024, par

Mary Edith SAVINO

Composition du jury

Membres du jury avec voix délibérative

Mathilde MOUGEOT Professeure, ENS Paris-Saclay et ENSIIE	Présidente
Olivier BOUAZIZ Professeur des universités, Laboratoire Paul Painlevé, Université de Lille	Rapporteur & Examinateur
Vincent LAGNEAU Professeur des universités et directeur du Centre de Géosciences, Mines Paris – PSL	Rapporteur & Examinateur
Marco DE LUCIA Chercheur associé, German Research Centre for Geosciences (GFZ), Potsdam	Examinateur
Nikolaos PRASIANAKIS Chercheur et responsable d'équipe, Paul Scherrer Institut (PSI), Villigen	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASM032

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Méthodes d'apprentissage statistique pour l'étude de problèmes géochimiques non linéaires
Mots clés: apprentissage statistique, estimation de fonction, sélection de variables, régression non-paramétrique, approches régularisées, simulations numériques

Résumé: Dans le cadre de simulations numériques de systèmes géochimiques s'intégrant dans un projet de stockage profond de déchets hautement radioactifs, nous proposons dans cette thèse deux méthodes d'estimation de fonction ainsi qu'une méthode de sélection de variables dans un modèle de régression non-paramétrique multivarié. Plus précisement, dans le Chapitre 2, nous présentons une procédure d'apprentissage actif utilisant les processus Gaussiens pour approcher des fonctions inconnues ayant plusieurs variables d'entrée. Cette méthode permet à chaque itération le calcul de l'incertitude globale sur l'estimation de la fonction et donc de choisir astucieusement les points en lesquels la fonction à estimer doit être évaluée. Ceci permet de réduire considérablement le nombre d'observations nécessaire à l'obtention d'une estimation satisfaisante de la fonction sous-jacente. De ce fait, cette méthode permet de limiter les appels à un logiciel dit "solveur" d'équations de réactions géochimiques, ce qui réduit les temps de calculs.

Dans le Chapitre 3, nous proposons une deuxième méthode d'estimation de fonctions non séquentielle consistant à approximer la fonction à estimer par une combinaison linéaire de B-splines et appelée GLOBER. Dans cette approche, les nœuds des B-splines pouvant être considérés comme des changements dans les dérivées de la fonction à estimer, ceux-ci sont choisis à l'aide du *generalized lasso*.

Dans le Chapitre 4, nous introduisons une nouvelle méthode de sélection de variables dans un modèle de régression non-paramérique multivarié, ABSORBER, pour identifier les variables dont dépend réellement la fonction inconnue considérée et réduire ainsi la complexité des systèmes géochimiques étudiés. Dans cette approche, nous considérons que la fonction à estimer peut être approximée par une combinaison linéaire de B-splines et de leurs termes d'interactions deux-à-deux. Les coefficients de chaque terme de la combinaison linéaire sont estimés en utilisant un critère des moindres carrés standard pénalisé par les normes ℓ_2 des dérivées partielles par rapport à chaque variable.

Les approches considérées ont été évaluées puis validées à l'aide de simulations numériques et ont toutes été appliquées à des systèmes géochimiques plus ou moins complexes. Des comparaisons à des méthodes de l'état de l'art ont également permis de montrer de meilleures performances obtenues par nos méthodes. Dans le Chapitre 5, les méthodes d'estimation de fonctions ainsi que la méthode de sélection de variables ont été appliquées dans le cadre d'un projet européen EURAD et comparées aux méthodes d'autres équipes impliquées dans le projet. Cette application a permis de montrer la performance de nos méthodes, notamment lorsque seules les variables pertinentes sélectionnées avec ABSORBER sont considérées. Les méthodes proposées ont été implémentées dans des packages R : `glober` et `absorber` qui sont disponibles sur le CRAN (Comprehensive R Archive Network).

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Title: Statistical learning methods for nonlinear geochemical problems
Keywords: statistical machine learning, function estimation, variable selection, nonparametric regression, regularized approaches, computer simulations

Abstract: In this thesis, we propose two function estimation methods and a variable selection method in a multivariate nonparametric model as part of numerical simulations of geochemical systems, for a deep geological disposal facility of highly radioactive waste. More specifically, in Chapter 2, we present an active learning procedure using Gaussian processes to approximate unknown functions having several input variables. This method allows for the computation of the global uncertainty of the function estimation at each iteration and thus, cunningly selects the most relevant observation points at which the function to estimate has to be evaluated. Consequently, the number of observations needed to obtain a satisfactory estimation of the underlying function is reduced, limiting calls to geochemical reaction equations solvers and reducing calculation times.

Additionally, in Chapter 3, we propose a non sequential function estimation method called GLOBER consisting in approximating the function to estimate by a linear combination of B-splines. In this approach, since the knots of the B-splines can be seen as changes in the derivatives of the function to estimate, they are selected using the *generalized lasso*.

In Chapter 4, we introduce a novel variable selection method in a multivariate nonparametric model, ABSORBER, to identify the variables the unknown function really depends on, thereby simplifying the geochemical system. In this approach, we assume that the function can be approximated by a linear combination of B-splines and their pairwise interaction terms. The coefficients of each term of the linear combination are estimated using the usual least squares criterion penalized by the ℓ_2-norms of the partial derivatives with respect to each variable.

The introduced approaches were evaluated and validated through numerical experiments and were all applied to geochemical systems of varying complexity. Comparisons with state-of-the-art methods demonstrated that our methods outperformed the others. In Chapter 5, the function estimation and variable selection methods were applied in the context of a European project, EURAD, and compared to methods devised by other scientific teams involved in the projet. This application highlighted the performance of our methods, particularly when only the relevant variables selected with ABSORBER were considered. The proposed methods have been implemented in R packages: `glober` and `absorber` which are available on the CRAN (Comprehensive R Archive Network).

Remerciements

Je voudrais avant tout remercier ma directrice de thèse, Céline Lévy-Leduc et mes co-encadrants de thèse, Marc Leconte et Benoit Cochepin qui m'ont encadrée pendant six mois de stage et qui m'ont finalement accompagnée pendant trois ans de thèse. J'aimerais vraiment vous remercier tous les trois pour cet encadrement car je m'estime extrêmement chanceuse d'avoir pu travailler avec vous et d'avoir eu un environnement de travail aussi riche et complet, tant d'un point de vue scientifique que social. J'ai pu bénéficier de cette multidisciplinarité, qui m'a permis d'aborder ce projet avec des points de vue très variés et enrichissants. Je n'aurais pas pu espérer mieux !

Merci, Céline, de m'avoir fait confiance dans ce projet et de m'avoir transmis les meilleurs outils pour réaliser mes travaux de recherche. Merci d'avoir été pour moi un exemple de force de travail et de rigueur à travers ta passion pour les mathématiques. Je te remercie également d'avoir toujours su me motiver et me remotiver lorsque j'en avais besoin, ainsi que pour ta grande disponibilité. Merci pour tous tes conseils, que ce soit pour le travail ou ma vie professionnelle. Il est difficile de quantifier avec des mots tout ce que j'ai appris grâce à toi et tout ce que tu m'as apporté.

Merci, Benoit et Marc, pour vos accompagnements respectifs. Vous m'avez fait découvrir un univers qui m'était peu familier et vous avez toujours été très patients et réactifs face à mes questions. Merci pour votre bonne humeur, votre gentillesse, vos conseils et pour m'avoir bien intégrée au sein de cette merveilleuse équipe, ainsi que pour tous nos échanges aux pauses café ou du midi. Vous êtes des encadrants formidables !

Je souhaiterais également remercier Mathilde Mougeot d'avoir accepté d'examiner et de présider mon jury de thèse. Je remercie Olivier Bouaziz et Vincent Lagneau d'avoir accepté de rapporter ma thèse et d'avoir consacré du temps à la lecture, ainsi qu'à la rédaction de commentaires constructifs et positifs sur mon manuscrit. Enfin, je souhaite remercier Marco De Lucia et Nikolaos Prasianakis, qui ont accepté d'examiner mon travail et d'avoir été présents à distance à ma soutenance de thèse.

Cette thèse n'aurait jamais pu avoir lieu sans l'aide précieuse de Guillaume Pépin et de Frédéric Plas. Je souhaite leur adresser ma gratitude pour tous nos échanges, l'intérêt qu'ils ont porté à mon stage de six mois et leur appui inestimable qui m'a permis de le poursuivre en trois années de thèse.

Je souhaite remercier tous les membres du service PSD-EPS de l'Andra, qui m'ont permis d'évoluer dans un environnement bienveillant et très enrichissant. Un merci plus spécifique à mon collègue et voisin de bureau de ces trois ans et demi, Antoine, pour nos nombreux échanges, mais aussi à Bernard pour tous ses précieux conseils et son aide indispensable quant à l'utilisation du cluster de calcul. Je remercie également Abdellah, Achim, Anaïs, Denis, Jacques, Jean, Matthieu, Renaud, Sandra, Timothé, Vincent et à toutes celles et ceux que j'ai côtoyé.e.s à l'Andra au cours de ces trois ans. Un grand merci à Marie pour son aide administrative, qui a su gérer, avec gentillesse et patience, les délais d'inscription (parfois très courts !) pour toutes les conférences auxquelles j'ai participé.

Je souhaite en parallèle remercier le laboratoire MIA-Paris-Saclay pour son accueil. Une mention particulière à *ma soeur de thèse* Marina, devenue une amie très chère, qui m'a apporté un grand soutien émotionnel et avec qui j'ai partagé des moments inoubliables. Merci à mon deuxième voisin de bureau, Louis, pour sa gentillesse et tous nos échanges ainsi que pour son aide en informatique, notamment pour m'avoir fait découvrir RustDesk, qui a changé le cours de ma thèse. Un grand merci également aux autres doctorants et doctorantes, post-docs et contractuels de l'équipe : Alizée, Annaïg, Armand, Barbara, Bastien, Caroline, Emré, Florian, François, Hayato, Jeanne, José, Jules, Marion, Mathilde, Saint-Clair, Tâm et Tanguy pour tous les moments partagés à l'agro ainsi qu'en dehors du travail et qui ont contribué à mon épanouissement personnel au sein de ce laboratoire. Je pourrais vous remercier toutes et tous individuellement tant chacun.e d'entre vous a joué un rôle important au cours de ma thèse, mais mes remerciements seraient alors

aussi longs que le manuscrit lui-même. Un grand remerciement à Camille, Christophe, David, Emmanuelle, Gabriel, Hugo, Isabelle, Jade, Jean-Benoit, Julie, Julien, Laure, Liliane, Lucia, Nicolas, Pierre, Sarah, Sophie, Stéphane et Tristan pour leur accueil et de m'avoir permis de m'intégrer au sein de cet environnement de travail productif et bienveillant. Un grand merci à Sébastien pour tous nos échanges et plus particulièrement sur le cinéma, ainsi qu'à Farida et Christelle pour leur précieux soutien administratif.

Merci à mes parents, Helen et Gil, d'avoir toujours été des modèles de réussite pour moi, merci pour les valeurs qu'ils m'ont inculquées, mais également pour leur soutien et leur amour inconditionnel qu'ils m'ont toujours porté. Un grand merci à ma petite sœur Anna pour son soutien et d'avoir toujours été présente pour moi, malgré la distance. Je sais que ces dernières années d'études ont été difficiles pour toi, mais je suis convaincue que tu réussiras tout ce que tu entreprendras.

Un grand merci à tous mes proches, mes ami.e.s de longue date, de collège, lycée, de prépa et d'AgroParisTech, sans qui ce parcours n'aurait jamais pu se concrétiser. Votre soutien m'a été essentiel. Merci à Thibault et Ben, qui malgré toutes ces années, ont toujours su être présents pour moi. Hâte de refaire d'autres sessions escalade avec vous. Je souhaite également remercier Adèle, Cléa, Hervé, Martin et Thomas, pour tous ces moments inoubliables, même après nos années à l'agro. Je souhaiterais également remercier Ellyn, Lucas et Ophélie, mes acolytes de stage à Sanofi qui sont toujours restés dans mon entourage et tout au long de ma thèse. Un remerciement très particulier à Camille et Célia, mes amies très chères à mon cœur, qui resteront toujours présentes pour moi, malgré tous les kilomètres qui nous séparent.

J'ai une pensée toute particulière pour mes professeures de mathématiques que j'ai eu la chance de rencontrer tout au long de mon parcours, notamment Mme Martine Vilatte, Mme Clémentine Portal et une fois de plus Céline, qui m'ont inspirée et ont été des figures féminines exemplaires, contribuant ainsi à mon désir de poursuivre une thèse en mathématiques appliquées.

Enfin, un immense merci à celui que cette thèse m'a permis de rencontrer et qui est aujourd'hui mon pilier et mon partenaire au quotidien. Merci Jérémy pour ton accompagnement, ton soutien et ton amour sans faille. Les mots me manquent pour exprimer à quel point je te suis reconnaissante, mais sache que je t'aime de tout mon cœur. J'espère pouvoir te rendre un jour tout le soutien inestimable que tu m'as apporté. Je te souhaite la plus grande des réussites, mais surtout d'être heureux dans tout ce que tu entreprendras.

Acknowledgements

I would like to express my gratitude to all the members of the EURAD project for their spirit, kindness and enriching exchanges we shared during my participation in the DONUT work package throughout this PhD journey. Your support and collaboration have been invaluable.

A special acknowledgment to Nikolaos Prasianakis and Marco De Lucia for their significant involvement in my PhD jury, as well as for their constructive advice and insightful feedback on my methods.

Finally, I would like to dedicate this work to my grandparents, who, I believe, would be proud of what I have achieved.

Table of contents

Chapter 1 - Introduction

1.1. Geochemical context

1.1.1. The Cigéo project

The research conducted in this thesis was financially supported by the French National Radioactive Waste Management Agency (Andra) which, according to their official website, aims at:

"*Identifying, implementing and guaranteeing safe management solutions for all French radioactive waste, in order to protect present and future generations from the risks inherent in such substances.*"

Radioactive waste is produced by various human activities but mainly by the nuclear power generation sector, as depicted in Figure 1.1A. According to the Environment Code of French Law, radioactive waste is defined as a substance containing radionuclides, whether natural or man-made, that is no longer used. Radionuclides are unstable nuclides with an excess number of neutrons or protons, leading to an excess of energy manifested as emission of high radiation. As a result, radioactive waste is characterized by its level of radioactivity and by its half-life, which corresponds to the time required to halve the number of radionuclides. Therefore, their activity and concentration necessitate radiological protection control. In France these radioactive materials account for a total volume of $1,760,000$ cubic meters[1], categorized into different waste groups, as illustrated in Figure 1.1B. This classification facilitates appropriate waste management through the development of specialized facilities to handle waste according to its radioactivity level. For instance, very low-level waste (VLLW) and low-to-intermediate-level short-lived waste (LILW-SL) referring to a half-life smaller than 30 years, are stored in surface disposal facilities, located at the Aube and Manche sites[1] for the latter. However, disposal facilities for low-level long-lived waste (LLW-LL), intermediate-level long-lived waste (ILW-LL) and high-level waste (HLW) are currently under study.

In particular, Andra has undertaken a significant project known as the Industrial Centre for Geological Disposal (Cigéo) aimed at providing a long-term solution for ILW-LL and HLW. These two waste categories account for only 3% of the total amount of radioactive waste but more than 99% of the total radioactivity. Therefore, this project poses a great challenge that has been the subject of intensive study by a huge number of scientists and engineers for more than three decades. The project involves the creation of a 500-meter deep geological disposal facility within a 250 square-kilometer area located between the Meuse and the Haute-Marne departments, as displayed in Figure 1.1C. This site was selected based on its geological characteristics which exhibits highly favorable properties for containing the radionuclides and preventing their dispersion into the environment. Furthermore, its depth provides protection against natural phenomena such as erosion and glaciation but also against anthropogenic aleas.

The Cigéo project has undergone various phases since its creation, ranging from conceptual and engineering designs, multiple safety, security, retrievability report submissions to multiple public consultations. Recently, a construction licence application was submitted in 2023 which is expected to culminate in a decree allowing the inital construction of the deep waste

[1] *National Inventory of radioactive materials and waste – 2023 Essentials*, available online at https://inventaire.andra.fr.

disposal facility. For further details on the progress of this project throughout the years, we refer the reader to the major milestones depicted in the Appendix 8.1.

As the Cigéo Project leads to the confinement of radioactive substances in the layer of a particular geological formation called Callovo-Oxfordian formation, both experimental studies and simulation models are necessary to validate the quality and safety of the project before undertaking any construction work. Specifically, an experimental laboratory at Bure allows for comprehensive assessments of the physical and geochemical characteristics. Furthermore, members of the international scientific community conducted significant experimental work to assess the suitability of rock materials for the disposal of radioactive waste. For instance, Moyce et al. (2014) investigated the alteration of surrounding rocks in alkaline cement waters over a 15-year period. Additionally, Fernández et al. (2018) examined the formation of magnesium silicates in a specific type of clay which can induce various chemical changes in rocks, including a reduction in porosity that can affect the transport properties of the host rock. The conducted experiments studied concrete and clay interactions for 13 and 10 years *in situ* and in a laboratory, respectively. In the context of the Cigéo project, an exhaustive list of safety requirements and measures over hundreds of thousands of years has to be respected, as the half-life of certain radioactive waste can extend to this time scale. These examples testify the extensive time required to obtain such results, making simulations essential supplementary tools for the examination of rock behavior under various scenarios and physical/chemical parameters. This facilitates the assessment of the performance of the different components of the disposal site, among other utilities.

1.1.2. Surrogate models in Reactive Transport Modelling (RTM)

1.1.2.1. Reactive Transport Modelling

A part of these simulations aims at assessing the geological media containing the encapsulated radioactive waste over the years and under physical and hydrological constraints. Effectively, the surrounding rocks can suffer from fractures which can lead to the infiltration of water and to the impairment of the geological structure. In this context, the chemical evolution of the different compounds of the geological materials is simulated to study for instance mineral precipitation/dissolution while considering a transport of groundwater. The comprehensive examination of fluid flow, geochemical reactions, heat transfer, and solute transport is commonly referred to as reactive transport modelling (RTM). Steefel et al. (2005) proposed a thorough introduction to RTM and its wide range of applications. Modelling the physics and chemistry in such scenarios entails coupling partial differential equations for transport with algebraic equations for chemical reactions. This presents a significant challenge, especially when the complexity of the system increases, making it difficult to solve (de Capitani and Brown (1987), Yeh and Tripathi (1989)).

To facilitate the implementation of such systems, a common approach is to consider chemical reactions and transport separately. This involves using an operator splitting approach where the discretized dispersion-advection equation is first solved independently, followed by the discretized chemical reaction equations. By employing this method, reactive transport modelling can be managed using two different solvers. While solutions obtained through operator splitting could have led to numerical errors in the past (Valocchi and Malmstead (1992); Barry et al. (1997); Simpson and Landman (2007)), this approach may offer great flexibility and simplicity (Steefel and McQuarrie, 1996). A substantial body of work has focused on this method for solving nonlinear reactive transport problems and compared different strategies for the

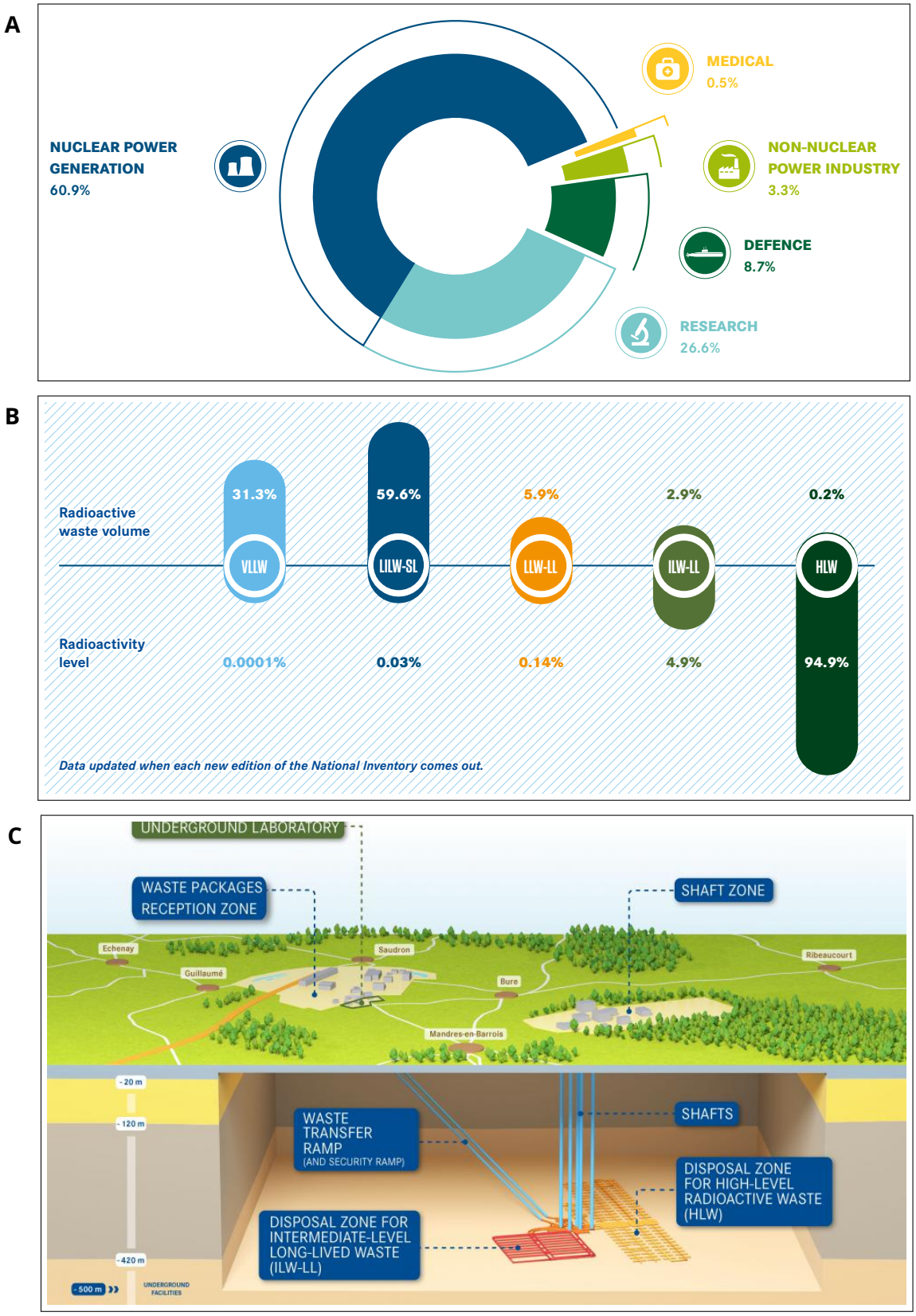

Figure 1.1: (A): Main producing sectors of radioactive waste in France. (B): Radioactive waste classification in France. VLLW: very low-level waste, LILW-SL: low-to-intermediate-level short-lived waste, LLW-LL: low-level long-lived waste, ILW-LL: intermediate-level long-lived waste, HLW: high-level waste. (C): Cigéo project, deep disposal facility for ILW-LL and HLW radioactive waste.

operator splitting approach, such as sequential iterative or sequential non-iterative schemes (Carrayrou et al., 2004; Marchuk, 1990). Notable contributions in this area have studied the convergence rate of these strategies (Kanney et al., 2003) and proposed improved algorithms to increase their computational efficiency (Lagneau and van Der Lee, 2010). Furthermore, Steefel et al. (2015) conducted a comprehensive benchmarking study on the available codes developed by the geochemical community notably for the use of operator splitting. We also refer the reader to this paper for further references and to Steefel (2019) for more recent advancements on RTMs. Despite recent efforts to enhance the accuracy of the operator splitting approach, it still faces computational challenges (Lu et al. (2022)).

In the context of radioactive waste disposal, several teams have studied RTM approaches applied to the surrounding rocks, the cementitious materials encapsulating radioactive wastes and their interactions with the environment. For instance, studies by Kosakowski and Berner (2013), Samper et al. (2016), and Wilson et al. (2018) explored reactive transport simulations in the context of Swiss, Spanish, and British radioactive waste management, respectively. Moreover, a collaborative European benchmark was established by Idiart et al. (2020) and proved that reactive transport simulation with chemical solvers and transport codes can effectively allow for the study of long-term interaction between hydrated concrete, in which radioactive waste are disposed, and the surrounding clay layer. Thus, this benchmark demonstrated that RTM is a powerful tool for safety evaluation.

Nethertheless, as modern simulations aim to represent real-world systems with increasing complexity, they tend to demand prohibitively large computational budgets. Chemical reactions are widely recognized as the limiting factor in computing such reactive-transport systems, especially considering the potentially numerous input and output variables (see Appelo and Postma (2004) for some examples). Despite recent improvements and a significant increase in computational resources and materials using highly performant computers, solving three-dimensional-large-scale models of complex RTM over many time steps still remain a major challenge.

1.1.2.2. Surrogate models

The emergence of what are commonly called *surrogate models* introduced a potential solution to circumvent the oversimplification of considered geochemical models required to make these simulations tractable. The idea is straightforward: explicitely solve the transport systems and then, based on a reduced set of solver-calculated solutions, use these models, also known as *proxy models*, to estimate the real solutions for simulation models suffering from a high computational cost. This approach allows for a drastic decrease in computational time while maintaining high accuracy of the calculated equilibrium state. The concept of surrogate models and their various applications have been extensively discussed in the literature to facilitate their use. We refer the reader to the book of Forrester et al. (2008) which provides a comprehensive overview of surrogate modelling for engineering design, covering polynomial models, radial basis function models, kriging and support vector regression. Within the scope of geosciences, a more recent review by Asher et al. (2015) outlines the primary categories of existing surrogate models applied to groundwater modelling. They note that no family of method universally outperforms others, as its efficiency strictly depends on its application.

Among all the existing methods, Machine learning (ML) approaches stand out as data-driven surrogate models that have garnered significant interest over the past few decades. Their capacity to accurately capture the nonlinear complexities inherent in real-world prob-

lems has established them as highly proficient estimation tools. A review by Razavi et al. (2012) delved into a broad range of statistical and ML-based surrogate models for water resources, demonstrating substantial computational efficiency and time savings, occasionally reaching up to 97% reduction of CPU time. More recent work by Jatnieks et al. (2016) compared a total of 32 statistical and ML methods to predict 7 output variables involved in RTMs. The authors determined that Bayesian Regularized Neural Networks emerged as the most efficient ML method in this context, exhibiting lower prediction errors. This method was then used to simulate a reactive transport system, yielding accurate predictions compared to traditional RTM simulations.

The use of surrogate models in reactive transport simulations for radioactive waste management represents a relatively novel approach. In recent works, Artificial Neural Networks (ANNs) have been employed as substitutes for traditional geochemical reactions solver. For example, Guérillot and Bruyelle (2020) considered a multilayer perceptron with one hidden layer to simulate reactive transport in a compositional reservoir flow simulation of dissolved CO_2. The obtained results demonstrated similar outcomes using the ANN approach and the traditional RTM simulations while significantly reducing computational cost. Additionally, Prasianakis et al. (2020) developed an ANN architecture capable of accurately describing both micro- and macroscopic interactions within geochemical systems, resulting in decreased simulation times. In a recent study by Laloy and Jacques (2022), a deep neural-network and a k-nearest neighbor regressor were compared as surrogate models for simulating a two-dimensional cementitious system under various transport conditions. Although both methods yielded satisfying results for the simplest geochemical systems, they struggled to accurately predict more complex systems involving multiple components. Nevertheless, they exhibited notable computational speed-ups, outperforming the traditional RTM solvers. More recently, Demirer et al. (2023) proposed a two-layered ANN designed to simulate a complex three-dimensional heterogeneous reactive-transport system under non-isothermal conditions. While this approach demonstrated promising accuracy and enhanced computational performance by one order of magnitude, it displayed error propagation accross multiple time steps for calcite amount. Furthermore, Collard et al. (2023) conducted a comparative study of five surrogate models – Random Forest, Gradient Boosting, AdaBoost, Support-Vector Machine and ANNs – to simulate a calcite-dolomite precipitation/dissolution geochemical system without transport. The ANN model outperformed the others in terms of prediction accuracy and achieved a computational gain of approximately three orders of magnitude compared to traditional solvers for geochemical equilibrium calculation. Finally, Laloy and Jacques (2019) demonstrated the accuracy of ANNs and Gaussian Processes (GP) in comparison to Polynomial Chaos Expansion, especially when training data is limited.

Other surrogate models have also been assessed in recent years, as presented by Leal et al. (2017). They proposed a novel strategy to markedly reduce the computation time of RTM simulations by up to two orders of magnitude. Their approach introduced an on-demand ML method based on sensitivity derivatives to determine the chemical state of an element from its previous state and from the infinitesimal changes of temperature, pressure and species amount. The resulting calculation is either accepted or rejected according to a specific criterion. In the latter case, the solution is computed using a geochemical solver. This approach enables the proxy model to be trained during the RTM simulation, thereby increasing its speed by a factor ranging from 60 to 125. This approach underwent further optimization in recent studies, such as Leal et al. (2020) and Kyas et al. (2022), which built upon the same philosophy to further

reduce execution time. Similarly, De Lucia and Kühn (2021) introduced a data-driven method using a mass balance criterion while the RTM is running to determine whether the surrogate model is accurate enough to be used or if the solution should be evaluated using the RTM solver.

These ANN-based methods have proven to be satisfactory surrogate models, enabling a significant decrease in computational time while maintaining high accuracy. However, they typically require a large number of observations, for example, Guérillot and Bruyelle (2020) needed at least $50,000$ training points to achieve their performance, while Collard et al. (2023) required $378,000$ samples. The high number of hyperparameters to tune can also be a limitation. Additionally, extending their applications to more complex real geochemical systems still encounters a bootleneck as it requires additional computational resources and/or larger training datasets to design more accurate surrogate models.

In this thesis, our goal is to address these issues by introducing two novel approaches for estimating multivariate nonlinear functions in a geochemical context. These developed methods are designed to demand a minimal number of observations while still providing a satisfying accuracy, thereby significantly accelerating the calculation process. Additionally, they are willing to have only a few parameters to tune, facilitating their implementation. Furthermore, along with these two estimation methods, we aim to devise a variable selection method tailored for nonlinear geochemical models to reduce their complexity, thus enabling more precise estimation.

1.2. Function estimation in multivariate regression models

1.2.1. State-of-the-art

Nonparametric regression is a well-established field which offers an alternative approach to parametric regression for function estimation when no prior information can be made regarding this function. More precisely, the nonparametric regression problem aims to estimate an unknown function f in the following model:

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = \left(x_i^{(1)}, \dots, x_i^{(p)}\right), \quad 1 \leq i \leq n, \tag{1.1}$$

where Y_i is a random variable modelling the ith observation of the *response variable* or *output variable*, $x_i^{(k)}$ denotes the ith value of the kth *predictor variable* or *input variable* and the ε_i are i.i.d centered random variables with variance σ^2.
The background presentation provided is mainly based on the books of Tsybakov (2009) and Hastie et al. (2009). We will focus on presenting the general insights of each method, as detailed materials are available in literature and the two presented books for further exploration.

1.2.1.1. One-dimensional case ($p = 1$)

Kernel regression: The fundamental concept behind kernel regression or kernel smoothing is to estimate the function f in (1.1) at each point by using its nearby observation points, resulting in a smooth estimation $\hat{f}$. It was initially defined as a probability density estimator of a random variable by Rosenblatt (1956) and Parzen (1962) before being extended to function estimation. A widely-used kernel regression model is the Nadaraya-Waston (N-W) estimator developed by Nadaraya (1964) and Watson (1964). By introducing a specific integrable function

$K : \mathbb{R} \to \mathbb{R}$, called kernel and satisfying $\int K(u)du = 1$, the N-W estimator is defined as:

$$\widehat{f}(x) = \begin{cases} \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}, & \text{if } \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \neq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{1.2}$$

where $h > 0$ is a bandwith parameter that has to be determined. A method for choosing h using cross-validation is proposed by Tsybakov (2009, p.27–31). As we can see in (1.2), this estimator can be viewed as a weighted average of the Y_i's, $i = 1, \ldots, n$, making it a linear estimator. Therefore, it provides a simple and intuitive approach for nonparametric regression while offering flexibility through adaptive bandwith selection. However, this method may suffer from a *boundary effect* leading to biased estimates as it assigns equal weight to all points within the bandwidth, regardless of their distance from the boundary (see Hastie et al. (2009, p.195) for visual representation of this effect). Furthermore, it may not be well-suited for multivariate regression models since it can suffer from the curse of dimensionality (Wasserman (2006, p.58)).

Local polynomial regression: This category of methods generalizes the N-W estimator by fitting local polynomials instead of constants. Consequently, the function f is approximated by fitting polynomials locally, thereby mitigating the *boundary effect* observed in the N-W estimator. This allows for more flexibility in capturing the underlying structure of the data. Thus, the local estimate at x, $x \in \mathbb{R}$, is defined as: $\widehat{f}(x) = \widehat{\beta}_0(x) + \sum_{j=1}^{q} \widehat{\beta}_j(x)$ where $\widehat{\beta}_0(x), \ldots, \widehat{\beta}_q(x)$, minimize the following weigthed least-square criterion :

$$\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \left(Y_i - \beta_0(x) - \sum_{j=1}^{q} \beta_j(x) x_i^j\right)^2. \tag{1.3}$$

From a historic point of view, local polynomial estimators were initially employed in the 1930s for the analysis of time series. They were later introduced to nonparametric regression by Stone et al. (1997). Building on these methods, the locally estimated scatterplot smoothing (loess) method was introduced as a robust locally weighted estimation model by Cleveland (1979). It fits a local polynomial model at each point using (1.3) but only considering its k nearest neighbors. Despite allowing for boundary correction even at higher dimensions, local polynomial regression methods suffer from high-dimensionality of the data when p exceeds 2 or 3 (Hastie et al. (2009, p.200), Cleveland and Devlin (1988)).

Projection methods: Another widely extented nonparametric approach is based on projection methods, also called linear basis expansion, which are approaches for global estimation. The main idea is to consider that f in (1.1) can be written as:

$$f(x) = \sum_{j=1}^{N} \beta_j \psi_j(x), \quad x \in \mathbb{R}, \tag{1.4}$$

where N is the total number of basis functions and ψ_j is the j^{th} function of the corresponding basis. Hence, the estimated parameters of β_j, denoted $\widehat{\beta}_j$ for $j = 1, \ldots, N$, are determined using a least-square criterion. They are then used to build the estimation of f by substituting β_j by $\widehat{\beta}_j$ in (1.4). This approach transforms the model from being written in terms of the original

variables x to being expressed in terms of additional transformed variables thereby facilitating the model interpretation. A broad range of basis functions is utilized in nonparametric regression such as the Fourier basis (Rice (1984)) and wavelets, which have been extensively used in signal processing and compression (Hastie et al. (2009, Chapter 5)). In particular, wavelets give information on both time and frequency localization, thus they have emerged as an alternative to Fourier basis since 1980s gaining popularity following the work of Meyer (1993). The strength of this method lies in their ability to modulate smoothness of the estimated signal. For further references, theoretical insights and applications on wavelets, we refer the reader to the books Hernández and Weiss (1996) and Härdle et al. (1998). More recent developments can be found in Mallat (1999).

Since the expansion of wavelets in the $1990s$, there has been a surge in the development of projection methods. As a result, other basis such as polynomials and splines have been gained interest to be used as basis functions ψ_j for $j = 1, \ldots, N$ in (1.4). Splines, in particular, are piecewise polynomial functions of degree k, continuous and depending on a list of fixed points called knots which need to be determined (De Boor, 1978; Wahba, 1990). Choosing the knot positions in regression splines can be tricky and may lead to overfitting and poor accuracy if done arbitrarily (Wood (2017, p.126)). This challenge led to the introduction of smoothing splines, a regularized method in which the least-square criterion has an additional term to penalize the second derivatives of f (Wang (2011); Green and Silverman (1994); Eubank (1999)). The second derivative of f can be seen as the roughness of the curve as it quantifies the rate of change of the slope of f. By penalizing it, smoothing splines automatically select knots from a predefined exhaustive list, resulting in a smooth estimation of f. For further details on this topic, we refer to Hastie et al. (2009, Section 5.2).

Another family of splines, known as B-splines, has gained popularity due to their computationally efficiency and stability. Introduced by De Boor (1978) in Chapter 9, B-splines are defined in such way that any spline function can be written as a linear combination of B-splines. This property coupled with their local support, makes them widely used for curve fitting (Piegl and Tiller (2012)) and function estimation using smoothing splines (Hastie et al. (2009, p.186–189)). Built upon the same idea, O'Sullivan (1986, Section 3) introduced a method based on penalized B-splines which became the most widely used class of penalized splines in statistical analyses thanks to its numerous properties and its fast R implementation (Wand and Ormerod (2008)). Inspired by this method, Eilers and Marx (1996) proposed another approach to penalized splines called P-splines. The idea is to used an exhaustive list of evenly-spaced knots in the B-splines basis and a difference penalty on the regression coefficients to shrink them close to 0, thus decreasing the model complexity. We refer to Eilers et al. (2015) for a recent reference on P-splines and its extensions and Eilers and Marx (2003) for an extension and applications to a two-dimensional regression model. More recently, Goepp et al. (2018) introduced a weighted adaptive ridge method to keep the most relevant knots of B-splines from an evenly-spaced list of points. It can be seen as a more interpretable procedure than P-splines, but showed similar results.

1.2.1.2. *Multivariate regression ($p > 1$)*

Generalized Additive Models (GAMs) and multivariate regression splines: In a nonlinear regression model, capturing the relationship between predictor and response variables can be challenging. Hastie and Tibshirani (1986) introduced GAMs to address this complexity. In

this case, we can express f in (1.1) as:

$$f(x) = \alpha + \sum_{j=1}^{p} f_j(x^{(j)}), \quad \alpha \in \mathbb{R}, \tag{1.5}$$

where f_j can be a cubic spline or smooth function. To fit this model, a backfitting procedure was introduced, iteratively computing $\widehat{f_j}$ by using the current estimates of the other functions $\widehat{f_k}$, for $k \neq j$ (Fox (2015, p.566-567); Wood (2017, p.209 section 4.11.1); Hastie et al. (2009, p.298 Algorithm 9.1)). However, if each function is modelled using an expansion of basis functions as in (1.4), a least-square criterion is sufficient to obtain the estimation of the resulting model. Supplementary materials on GAMs and applications can be found in (Wood, 2017, Chapters 4 and 5).

Stone and Koo (1985) provide an example and applications of additive splines for multivariate regression, along with discussions on knot selection. For further investigation on these topics, we recommand Wand (2000) and the references therein to the reader. The main advantage of these additive models is their simplicity of interpretation and generally straightforward implementation. However, they can be restrictive as they do not consider interaction terms. To address this limitation, approaches such as *tensor product* (Wasserman (2006, Chapter 8 p.193); Hastie et al. (2009, Chapter 5, p.162)) or *thin-plate splines* (Wahba (1990, Section 2.4); Green and Silverman (1994, Chapter 7); Wood (2017, p.150)), have been introduced to integrate interactions in regression splines. While these methods offer great flexibility, they may exhibit computational complexity, especially as the dimension p increases.

Tree-based method: Tree-based methods have gained increasing visibility, particularly since the introduction of Classification And Regression Trees (CART) by Breiman et al. (1984). The general idea of this method is to use binary split to iteratively define split points, resulting in M regions which leads to the following expression for f:

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{1}\{x \in R_m\}, \tag{1.6}$$

where c_m is a constant specific to each region R_m ($1 \leq m \leq M$) and $\mathbb{1}\{A\} = 1$ if A holds, 0 otherwise. Therefore, the estimate $\widehat{f}$ can be obtained by calculating $\widehat{c}_m = N_m^{-1} \sum_{x_i \in R_m} y_i$, where N_m is the cardinality of R_m for m belonging to $\{1, \ldots, M\}$. Regression trees are simple and easy to interpret and numerous algorithms for tree regression are now available, see Loh (2011) for a comparative work and applications of different regression tree algorithms. Nevertheless, these tree-based methods can exhibit some limitations (Hastie et al. (2009, Chapter 9, p.307)). For example, they do not handle missing data well, may present instability due to their hierarchical architecture and struggle to capture many additive effects with a high number of input variables p.

To tackle these issues, Bagging and its extension Random Forests were introduced by Breiman (1996) and Breiman (2001), respectively. These methods improve robustness of regression trees by averaging the resulting output values from large collection of decorrelated trees. Furthermore, they have become popular due to their interpretability, performance and optimized implementations in various programming languages (Hastie et al. (2009, Chapter 15, p 587)).

Another notable multivariate regression approach is Multivariate Adaptive Regression Splines (MARS) introduced by Friedman (1991). MARS follows a similar philosophy to the

CART model with a binary structure but uses a regression splines model with additive and interaction terms. By leveraging piecewise linear functions in a regression model such as (1.4), MARS produces a sparse model since the splines are locally nonzero. Additionally, through a forward and backward procedure to prune the number of knots and spline terms, MARS decreases the model's complexity, making it computationally advantageous.

Support Vector Regression (SVR): Support Vector Machines are widely known for its performance as a classifier but their extension, Support Vector Regression (SVR), has also demonstrated interesting features for regression tasks, as presented by Vapnik et al. (1996). By employing the kernel trick, SVR efficiently estimates nonlinear models as they are known to be insensitive to noisy observations. However, they can potentially result in poor predictions when the kernel function is not well chosen.

Artificial neural networks (ANNs): Since the introduction of the percepton concept in the 1950s by Rosenblatt (1958), ANNs have become immensely popular for both classification and regression problems (see LeCun et al. (2015)). They are nonlinear statistical models constructed by composing linear combinations of transformed input variables with a chosen activation function, which can be linear or nonlinear.

Despite their promising characteristics, ANNs can encounter several challenges. Firstly, their overparametrized architecture can lead to instability in performance due to initial configuration of the weights, which may result in local minima and suboptimal solutions (Aggarwal et al. (2018, Section 1.4.4)). Moreover, too many weights can cause overfitting of the data. To address this issue, regularization approaches can be employed, such as *weight decay* or alternatively, *bagging* or *ensemble methods* to average predictions over a collection of networks from randomly chosen training data (Goodfellow et al. (2016, Chapter 7); Aggarwal et al. (2018, Section 1.4.1)). Furthermore, training ANNs generally demands significant computational resources and their interpretability is often hindered as they are perceived as *'black boxes'*. Hyperparameter tuning, including determining the appropriate number of nodes and layers, can also be complex (Hastie et al. (2009, Chapter 11, p.389)). Finally, ANNs require a substantial number of observations to achieve a desired level of accuracy (Goodfellow et al. (2016, Part III)).

In the previously defined methods, the observation set consists of a predetermined number of observations, which may be obtained at a significant computational expense especially when it is large. This approach can be seen as a "passive" one. Various strategies can be employed to reduce this cost, for example by defining a small observation set and then sequentially and adaptively adding new observations according to a specific criterion. At each iteration, a new estimation can be obtained from the new observation set. In such cases, it is essential to define both a selection criterion and a stopping criterion for the sequential method, which is commonly referred as "active" learning.

1.2.1.3. Active learning

A commonly used approach for performing active learning consists in using a Bayesian framework. In this vein, Srinivas et al. (2012) proposed a strategy to optimize an unknown function by considering it as a sample of a zero-mean Gaussian Process, leveraging the mean and covariance of the posterior distribution. The method employs a multi-armed bandit problem, inspired by the sequential design defined by Robbins (1952), to optimize the target function

by sequentially selecting new observations. For further insights into Gaussian Processes and covariance functions, we refer the reader to Rasmussen and Williams (2006). More recently, inspired by Srinivas et al. (2012), Jala et al. (2016) presented four sequential approaches for estimating a quantile of a random variable which can be written as an unknown function of a random vector having a known distribution, by using as few observations as possible.

In the following section, we introduce a novel active learning approach using Gaussian Processes, inspired by Srinivas et al. (2012), for function estimation to sequentially select new observation points while reducing the approximation error. Additionally, we further present a complementary method based on multivariate regression B-splines with an adaptive knot selection to improve the estimation accuracy.

1.2.2. Contribution of Chapter 2

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

We present our function estimation approach with the previously introduced idea of active learning to use a sequentially well-chosen set of observation points. The function to estimate is a real-valued function f defined on a compact subset $\mathcal{A} \subset \mathbb{R}^d$, satisfying (1.1).

We adopt a Bayesian point of view which consists in considering f as a sample of a zero-mean Gaussian process (GP) having a covariance function k that we shall denote by GP(0,$k(\cdot,\cdot)$) in the following. The advantage of this approach is that, conditionally on a set of t observations $\mathbf{y}_t = (y_1, \ldots, y_t)'$ where $y_i = f(x_i)$, x_i belonging to $\mathcal{A}$, the posterior distribution is still a GP having a mean μ_t and a covariance function k_t given by

$$\mu_t(u) = \mathbf{k}_t(u)' \mathbf{K}_t^{-1} \mathbf{y}_t \ , \tag{1.7}$$

$$k_t(u,v) = k(u,v) - \mathbf{k}_t(u)' \mathbf{K}_t^{-1} \mathbf{k}_t(v) \ , \tag{1.8}$$

where $\mathbf{k}_t(u) = [k(x_1,u) \ldots k(x_t,u)]'$. Here $'$ denotes the matrix transposition, u and v are in $\mathcal{A}$ and $\mathbf{K}_t = [k(x_i,x_j)]_{1 \le i,j \le t}$, where the x_i's are in $\mathcal{A}$.

In our case, f models a physical quantity that is assumed to be smooth, so for our applications we shall consider two covariance functions that are commonly used in this case. The first one is the squared exponential (SE) covariance function

$$k_{\mathrm{SE}}(u,v) = \exp\left(-\frac{1}{2}(u-v)' M^{-1} (u-v)\right), \quad u,v \in \mathcal{A} \subset \mathbb{R}^d \ , \tag{1.9}$$

$$M = \mathrm{diag}\left(\ell_1^2, \ldots, \ell_d^2\right) \ , \ \ell_1 \ , \ \ell_2 \ , \ldots \ , \ell_d > 0 \ . \tag{1.10}$$

Here the $\boldsymbol{\ell} = (\ell_1 \ , \ \ell_2 \ , \ldots \ , \ell_d)$ hyperparameters are the characteristic length scales. Actually, these hyperparameters can be understood as how far you need to move along a particular axis in the input space so that the function values become uncorrelated. Note that Definition (1.9) allows us to model anisotropic response surfaces.

As explained in Rasmussen and Williams (2006), since this covariance function is infinitely differentiable, the GP with this covariance function has mean square derivatives of all orders.

As argued by Stein (1999) such strong smoothness assumptions may be unrealistic for modeling many physical processes, so we shall also consider another covariance function belonging to the Matérn class of covariance functions defined by

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} r\right)^{\nu} K_{\nu}\left(\sqrt{2\nu} r\right), \quad \nu > 0 ,$$

where K_ν is a modified Bessel function with Bessel order ν, see (Abramovitz and Stegun, 1965, Section 9.6), and r is defined by

$$r = \sqrt{(u-v)' M^{-1} (u-v)} , \quad u, v \in \mathcal{A} , \tag{1.11}$$

M being defined in (1.10). In this situation, as explained in Rasmussen and Williams (2006), the GP is q-times mean-square differentiable if and only if $\nu > q$. Here, we shall focus on the case where $\nu = 5/2$, for which $k_{\text{Matérn}}$ has a computationally advantageous expression. Indeed, for $\nu = p + \frac{1}{2}$, where p is in $\mathbb{N}$,

$$k_{\text{Matérn}}(r) = \exp\left(-\sqrt{2\nu} r\right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\sqrt{8\nu} r\right)^{p-i} , \tag{1.12}$$

with r defined in (1.11); see Abramovitz and Stegun (1965, Equation 10.2.15) for further details.

In the following, we shall denote by A a fine grid of $\mathcal{A}$:

$$\mathrm{A} = \{\mathrm{x}_1, \ldots, \mathrm{x}_m\} \subset \mathcal{A} . \tag{1.13}$$

This grid is either a regular grid of $\mathcal{A} \subset \mathbb{R}^d$ when d is small (usually 1 or 2) or a Latin Hypercube Sampling for larger values of d. Note that this grid contains the points at which the estimation of f is performed and that the points at which f is evaluated are chosen in this grid. Inspired by Srinivas et al. (2012) who proposed a sequential approach for maximizing a function by modeling it using a Gaussian process, we propose a strategy which consists in adding the new point x_{t+1} to the set of t observations at which f needs to be evaluated as follows:

$$x_{t+1} \in \operatorname*{Arg\,max}_{x \in \mathrm{A}} \sigma_t(x) ,$$

where

$$\sigma_t(x)^2 = k_t(x, x), \tag{1.14}$$

k_t being defined in (1.8) and $\operatorname*{Arg\,max}_{x \in \mathrm{A}} \sigma_t(x)$ being the set of $x \in \mathrm{A}$ where $\sigma_t(x)$ reaches its maximum. Note that the points $x_1, x_2, \ldots, x_t, x_{t+1}, \ldots$ at which f needs to be evaluated are chosen in the fine grid A of $\mathcal{A}$ defined in (1.13).

We propose using a maximum-likelihood strategy described in Rasmussen and Williams (2006) to estimate $\boldsymbol{\ell}$. This adds a step to the method previously described, as the ℓ_i's have to be estimated before evaluating the posterior distribution of the GP using (1.7) and (1.8). Hence, for the observation set $\{(x_1, y_1), \ldots, (x_t, y_t)\}$ with $y_i = f(x_i)$, $1 \leq i \leq t$, the posterior log-likelihood given by:

$$-\frac{1}{2} \mathbf{y}_t' \mathbf{K}_t^{-1} \mathbf{y}_t - \frac{1}{2} \log |\mathbf{K}_t| - \frac{t}{2} \log 2\pi ,$$

with $\mathbf{y}_t = (y_1, \ldots, y_t)'$ and $\mathbf{K}_t = [k(x_i, x_j)]_{1 \leq i,j \leq t}$, has to be maximized with respect to $\boldsymbol{\ell}$.

Different stopping criteria based on the following quantities were defined:

- **Ratio variance.** At each iteration t of our method, the following average is computed:

$$R_n(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\max_{x \in \mathrm{A}} \sigma_t^2(x)}{\max_{x \in \mathrm{A}} \sigma_{t-i}^2(x)}, \tag{1.15}$$

  where σ_t is defined in (1.14) and $n = 2$, 5 or 10.

- **Mobile average.** At each iteration t of our method, the following average is computed:

$$M_\ell(t) = \frac{1}{\ell} \sum_{j=0}^{\ell-1} \max_{x \in \mathrm{A}} \sigma_{t-j}^2(x) \tag{1.16}$$

  for $\ell = 5$ or 10 where σ_t is defined in (1.14).

- **Maximal variance.** At each iteration t of our method,

$$V(t) = \max_{x \in \mathrm{A}} \sigma_t^2(x) \tag{1.17}$$

  is computed where σ_t is defined in (1.14).

This active learning method for estimating unknown functions is initially assessed on one- and two-dimensional functions of geochemical reactions. Based on three statistical measures, we compared the performance of the two covariance functions defined in (1.9) and (1.12), using $\nu = \frac{5}{2}$, alongside the three stopping criteria introduced in (1.15), (1.16) and (1.17). The choice of the covariance function is found to be insignificant in these cases and criteria based on the ratio and mean average over at least 10 iterations exhibited the most satisfactory, significantly reducing the number of observations. However, the threshold for these stopping criteria must be specifically selected for the chosen covariance function. Further applications to higher dimensional geochemical cases yielded similar conclusions, emphasizing the low budget of observation points required to achieve interesting accuracy.

1.2.3. Contribution of Chapter 3

This section summarizes the article:
Savino, E. M., Lévy-Leduc, C. 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. Submitted and also available on arXiv preprint *(arXiv:2306.00686)*.

The proposed method is implemented in the `glober` R package which is available on the CRAN.

We propose estimating the function f appearing in (1.1) by approximating it with a linear combination of B-splines of order M ($M \geq 1$) introduced by De Boor (1978) in Chapter 9.

Let $\mathbf{t} = (t_1, \ldots, t_K)$ be a set of K points called knots which are crucial in the definition of the B-spline basis. We define the augmented knot sequence $\boldsymbol{\tau}$ such that:

$$\begin{aligned} \tau_1 = \ldots = \tau_M &= x_{min}, \\ \tau_{j+M} &= t_j, \quad j = 1, \ldots, K, \\ x_{max} = \tau_{K+M+1} &= \ldots = \tau_{K+2M}, \end{aligned}$$

$$\boldsymbol{\tau} = (\tau_1, \dots, \tau_{K+2M}) = (\underbrace{x_{min}, \dots, x_{min}}_{\text{M times}}, \underbrace{t_1, \dots, t_K}_{\mathbf{t}}, \underbrace{x_{max}, \dots, x_{max}}_{\text{M times}}),$$

where x_{min} and x_{max} are the lower and upper bounds of $\mathcal{S}$, a compact set of $\mathbb{R}$ on which f is defined.

B-splines are defined by De Boor (1978, p. 89-90) and Hastie et al. (2009, p. 160) as follows. Denoting by $B_{i,m}(x)$ the ith B-spline basis function of order m for the knot sequence $\boldsymbol{\tau}$ with $m \leq M$, they are defined by the following recursion:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } \; i = 1, \dots, K + 2M - 1, \tag{1.18}$$

and for $m \leq M$,

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x), \tag{1.19}$$

for $i = 1, \dots, (K + 2M - m)$.

The following introduced estimation method is called GLOBER. Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\mathbf{x} = (x_1, \dots, x_n)$ where Y_i and x_i are defined in (1.1). In the following, we shall assume that $x_1 < \dots < x_n$ and $M = q + 1$, with $q \geq 0$. Hence, when $q = 0$ (resp. $q = 1$, $q = 2$) f is approximated by piecewise constant (resp. linear, quadratic) functions.

Since the knots of a B-spline basis can be seen as changes in the $(q + 1)$th derivative of f, we propose finding them by using the generalized Lasso described in Tibshirani and Taylor (2011) and further studied in Tibshirani (2014). In the latter, they define the polynomial trend filtering which consists in approximating f by $\widehat{\boldsymbol{\beta}}(\lambda)$ defined as follows:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\operatorname{argmin}} \{ ||\mathbf{Y} - \boldsymbol{\beta}||_2^2 + \lambda ||D\,\boldsymbol{\beta}||_1 \}, \tag{1.20}$$

where $||y||_2^2 = \sum_{i=1}^n y_i^2$ for $y = (y_1, \dots, y_n)$ and $||u||_1 = \sum_{i=1}^m |u_i|$ for $u = (u_1, \dots, u_m)$, λ is a positive constant which has to be tuned and $D \in \mathbb{R}^{m \times n}$ is a specified penalty matrix, defined recursively as follows:

$$D = D_{tf,q+1} = D_0 \cdot D_{tf,q} \quad q \geq 0,$$

where "tf" is the abbreviation of "trend filtering", $(q+1)$ is the order of differentiation, $D_{tf,0} = \mathrm{Id}_{\mathbb{R}^n}$, the identity matrix of $\mathbb{R}^n$, and D_0 is the penalty matrix for the one-dimensional fused Lasso:

$$D_0 = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}.$$

The penalty matrix D is the discrete difference operator of order $(q + 1)$ and thus, $D\widehat{\boldsymbol{\beta}}$ estimates the $(q + 1)$st order derivative of f. Hence, observing the locations where $D\widehat{\boldsymbol{\beta}} \neq 0$ provides a way of finding the B-spline knots.

The matrix D is well-adapted when the observation points are evenly spaced. When it is not the case, it should be replaced by the following matrix $\Delta^{(q+1)}$ defined recursively as follows:

$$\Delta^{(q+1)} = \mathbf{W}_{(q+1)} \cdot D_0 \cdot \Delta^{(q)}, \quad q \geq 0,$$

where $\Delta^{(0)} = \mathrm{Id}_{\mathbb{R}^n}$ and $\mathbf{W}_{(q+1)}$ is the diagonal weight matrix defined by:

$$\mathbf{W}_{(q+1)} = \mathrm{diag}\left(\frac{1}{(x_{(q+1)+1} - x_{(q+1)})}, \frac{1}{(x_{(q+1)+2} - x_{(q+1)+1})}, \ldots, \frac{1}{(x_n - x_{n-1})}\right).$$

In both cases (evenly or unevenly-spaced observations), the number of rows of D and $\Delta^{(q+1)}$ equals $m = n - q - 1$.

Let us now more precisely explain how to choose the B-spline knots. Let $\Lambda = (\lambda_1, \ldots, \lambda_k)$ be a grid of penalization parameters λ_i. We define the resulting differentiated column vector $\mathbf{a}(\lambda)$ by:

$$\mathbf{a}(\lambda) = \Delta^{(q+1)} \cdot \widehat{\boldsymbol{\beta}}(\lambda),$$

where $\widehat{\boldsymbol{\beta}}(\lambda)$ is the solution of problem (1.20) when $D = \Delta^{(q+1)}$ and λ belongs to Λ.

The ordered vector of selected knots associated to λ is defined as follows:

$$\widehat{\mathbf{t}}_\lambda = \left(\widehat{t}_j\right)_{j=1,\ldots,K_\lambda} = \left(x_{p_j}\right)_{j=1,\ldots,K_\lambda}, \quad \text{with } p_j \in \mathcal{P}_\lambda, \tag{1.21}$$

where

$$\mathcal{P}_\lambda = \{\ell + 1,\ a_\ell(\lambda) \neq 0\ \} \quad \text{and} \quad K_\lambda = \sum_{\ell=1}^{m} \mathbb{1}\{a_\ell(\lambda) \neq 0\},$$

$a_\ell(\lambda)$ denoting the ℓth component of $\mathbf{a}(\lambda)$ and $\mathbb{1}\{A\} = 1$ if the event A holds and 0 if not.

The corresponding B-spline basis $B_{i,M}$ is defined by replacing the t_j in the augmented knot sequence $\boldsymbol{\tau}$ appearing in (1.18) and (1.19) by $\widehat{t}_j$ found in (1.21). Thus, we obtain the following estimator of f for each λ of Λ:

$$\widehat{f}_\lambda(x) = \sum_{i=1}^{q+K_\lambda+1} \widehat{\gamma}_i B_{i,M}(x), \tag{1.22}$$

where $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_i)_{1 \leq i \leq q+K_\lambda+1}$ is obtained using the following least-square criterion:

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{q+K_\lambda+1}}{\mathrm{argmin}} \|\mathbf{Y} - \mathbf{B}(\lambda)\,\boldsymbol{\gamma}\|_2^2, \tag{1.23}$$

where $\mathbf{B}(\lambda)$ is a $n \times (q + K_\lambda + 1)$ matrix having as ith column $(B_{i,M}(x_k))_{1 \leq k \leq n}$, i belonging to $\{1, \ldots, q + K_\lambda + 1\}$.

In order to choose the penalization parameter λ which leads to the best selection of knots, we use a criterion defined by Chen and Chen (2008) and recommended in Goepp et al. (2018), namely the extended Bayesian information criterion also called EBIC:

$$\mathrm{EBIC}(\lambda) = \mathrm{SS}(\lambda) + (q + K_\lambda + 1)\log n + 2\log\binom{q + K_{\max} + 1}{q + K_\lambda + 1}, \tag{1.24}$$

where $K_{\max}$ is the maximum number of knots that we can select (here $K_{\max} = n$) and $\mathrm{SS}(\lambda)$ is the sum of squares defined by:

$$\mathrm{SS}(\lambda) = \|\mathbf{Y} - \widehat{\mathbf{Y}}(\lambda)\|_2^2,$$

where

$$\widehat{\mathbf{Y}}(\lambda) = \mathbf{B}(\lambda)\widehat{\boldsymbol{\gamma}},$$

with $\widehat{\boldsymbol{\gamma}}$ and $\mathbf{B}(\lambda)$ being defined in (1.23). This criterion allows us to get a trade-off between a good approximation of the underlying function without using too many parameters. The final estimator of f is defined as follows:

$$\widehat{f}(x) = \widehat{f}_{\lambda_{\mathrm{EBIC}}}(x),$$

where $\widehat{f}_{\lambda}(x)$ is defined in (1.22) and

$$\lambda_{\text{EBIC}} = \underset{\lambda \in \Lambda}{\operatorname{argmin}}\{\text{EBIC}(\lambda)\}.$$

The extension to $d = 2$ is also proposed in this chapter, based on the following idea. Firstly, we consider that $\mathcal{S}$ is defined as the Cartesian product of two compact sets $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathbb{R}$. We seek at estimating f in (1.1) as follows:

$$\sum_{i=1}^{Q_1}\sum_{j=1}^{Q_2} \gamma_{ij}\, B_{1,i,M}(x_1) B_{2,j,M}(x_2), \quad (x_1, x_2) \in \mathcal{S}_1 \times \mathcal{S}_2, \tag{1.25}$$

where $B_{1,i,M}$ and $B_{2,j,M}$ are the B-spline basis of order M defined in (1.19) for the first and second dimension, respectively. In (1.25), $Q_1 = q + K_1 + 1$, $Q_2 = q + K_2 + 1$ with K_1 and K_2 the number of knots defined in the B-spline basis of the first and second variables, respectively and $M = q + 1$. So we can consider the two dimensions independently and thus, by fixing one dimension at a time, the problem can be rewritten as an estimation problem in the one-dimensional framework.

Therefore, for each dimension, we identify different candidate sets of knots. Then, for every combination of these sets, we compute the EBIC adapted from (1.24). The final estimator of f is obtained by building the B-spline basis for each dimension by using the knots selected by the EBIC. This chapter also introduces an extension of this method, called GLOBER-c, designed to accommodate higher dimensions and general sets of points that may not be generated by a Cartesian product. GLOBER-c achieves this by employing a clustering method to cluster the observations into groups and then, to identify the candidate set of knots for each dimension. Numerical experiments are conducted to assess the influence of the sampling of the observation set and noise levels. Their impact does not appear to be significant. Furthermore, the chapter explores some applications of the method to geochemical reactions. The results demonstrate that, in these specific contexts, our method outperforms several state-of-the-art alternatives, such as MARS, GP and a architecture of ANN.

1.3. Variable selection in nonlinear multivariate models

1.3.1. State-of-the-art

Another approach to reduce computational time and simplify geochemical simulations consists in reducing the number of input variables considered in the model, a strategy also known as *variable selection* or *feature selection*. Let us consider that we have n observations satisfying the nonparametric regression model defined in (1.1). We will 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 $\widetilde{x}$.

1.3.1.1. Kernel-based methods

The extension of linear lasso regression introduced by Tibshirani (1996) has paved the way for variable selection in nonlinear regression models. Inspired by Roth (2004), Lin and Zhang (2006) presented the Feature Vector Machine, a feature-wise nonlinear lasso regression for feature selection that serves as a sparse adaptation of SVRs. However, this method exhibits some constraints as it applies the same nonlinear kernel function to both the response and input

variables, limiting its flexibility in capturing nonlinear dependences. More recently, Yamada et al. (2014) introduced HSIC-Lasso and its alternative NOCCO-Lasso, two kernel-based methods that leverage a minimum redundancy maximum relevance criterion (mRMR), as defined by Peng et al. (2005), to overcome the limitations of the previous approach. These methods employ different kernel functions for the output and input variables for more flexibility. Based on kernel-based independence criteria such as HSIC or NOCCO, they select relevant variables while simultaneously discarding the most redundant ones.

Another approach proposed by Bertin and Lecué (2008) consists in employing a local polynomial regression model penalized by ℓ_1 regularization on the regression coefficients. The subset of selected variables is then used to estimate f through a local polynomial estimator, minimizing a weighted least-squares criterion as defined in (1.3). Strong theoretical insights are provided within this article. Based on the same regularization strategy, Allen (2013) introduced KNIFE, a weighted-kernel-based method for feature selection in kernel regression. This method incorporates weights within the kernel functions in a regularization approach, penalizing a loss function by the ℓ_1-norm of these weights to enforce stronger feature selection.

While these approaches aim at penalizing weights or regression coefficients, Rosasco et al. (2010) proposed a penalty based on the partial derivatives of the function with respect to each variable, selecting the most relevant variables for which the partial derivative is nonzero. A comprehensive investigation of this approach, including its selection properties, computational aspects and comparisons to other state-of-the-art methods can be found in Rosasco et al. (2013).

1.3.1.2. Additive models and regularization approaches

An extension of the Lasso approach to nonlinear regression models in the context of smoothing splines was presented by Lin and Zhang (2006) as COSSO. Comparisons with MARS (Friedman, 1991), defined in the previous section, as a variable selection due to its capability to prune irrelevant variable-associated terms, on a few examples revealed that COSSO exhibited superior performance in variable selection.

Similarly, Ravikumar et al. (2009) introduced a novel approach based on GAM previously defined in (1.5), called sparse additive models (SpAM). Regarded as a functional version of the Group Lasso (Yuan and Lin (2006)), SpAM treats smoothness and sparsity separately, enabling the expression of each f_j in equation (1.5) as a linear combination of basis functions, as illustrated in equation (1.4). B-splines are a possible choice for these basis functions. The use of P-splines, introduced in the previous section, can be extended to variable selection as discussed in Antoniadis et al. (2012). Their approach employs a regularization approach, the nonnegative garrot method defined by Breiman (1995), to simplify the model and estimate the regression coefficients. Additionally, Huang et al. (2010) proposed a two-step procedure using the adaptive group Lasso approach, an extension of the adaptive Lasso defined by Zou (2006) to group Lasso, for selecting relevant variables and reducing model complexity. Furthermore, Radchenko and James (2010) adapted the additive approach by incorporating pairwise interaction terms through the use of basis functions, thereby enhancing flexibility.

1.3.1.3. Tree-based methods

A popular variable selection method, proposed by Genuer et al. (2010), is based on Random Forests (RF). Inspired by the work of Díaz-Uriarte and Alvarez de Andrés (2006), the procedure involves two steps. Firstly, it calculates the Variable Importance (VI) of each feature by permuting them and measuring the difference in regression error using the original set of ob-

servations versus the permuted one. The variables are then ranked according to their VI and the top m variables are retained, m being a parameter to choose. In the second step, the pre-selected variables are further discriminated according to two distinct strategies, one dedicated to prediction improvements and the other to enhance interpretability. While this method efficiently identifies relevant variables, it may be sensitive to the choice of hyperparameters and correlated variables.

Similarly, Galelli and Castelletti (2013) introduced an iterative approach for variable selection using an ensemble method based on constructing multiple trees with the entire observation set. This method first sorts the variables according to their importance in terms of explained variance and then, at each iteration, selects the best variable among the top m variables based on regression error. The procedure continues until no further improvement is observed in the total regression model using all selected variables. This method is supposed to be more robust and avoid redundancy of correlated variables.

1.3.1.4. ANNs

ANN-based method have garnered interest for variable selection in recent years, proving highly efficient for nonlinear regression models and coping with high-dimensional data. A common approach involves using regularization methods on the weights to shrink those associated with less relevant variables. For example, Li et al. (2016) introduced a *deep feature selection* (DFS) method employing a multi-layer perceptron with an Elastic-Net penalty on the weights of the first layer, each corresponding to an input variable. They also included a penalty term on the hidden layers to reduce model complexity and prevent swelling of weights in the upper layers. Similarly, Feng and Simon (2017) introduced SPINN, a sparse input neural network featuring sparse group lasso regularization on the first-layer weights. This approach achieves sparsity by shrinking entire vectors of weights associated with irrelevant variables. Additionally, a ridge penalty is applied on the upper-layer weights to enhance connections of very few nodes. Ye and Sun (2018) extended this method by incorporating a drop-one-out approach, using a greedy algorithm to iteratively drop one weight at a time associated with a variable and observing changes in the loss function. Chen et al. (2021) proposed a regularized neural network architecture with a specific selection layer for DFS, providing strong theoretical insights and established selection consistency. The integration of skip-connections into the ANN architecture is another approach for simplifying the model, reducing computational training time and improving variable selection. For instance, Feng and Simon (2022) introduced SIER-net which achieves input sparsity and controls the number of active layers and nodes. Additionally, they proposed an ensemble extension using a Bayesian perspective to enhance the variable selection performance, especially in dealing with correlated or grouped variables. In the same vein, Lemhadri et al. (2021) proposed a residual feed-forward neural network, introduced by He et al. (2016). They employed a Lasso regression to shrink the weights of the residual layer associated with irrelevant variables. A strong constraint involves the direct elimination of the weights of the first hidden layer linked to these irrelevant variables. Another approach proposed by Liang et al. (2018) consists in using a Bayesian neural network architecture to select variables for which the marginal inclusion probability exceeds a predefined threshold.

One of the major drawbacks of these methods is the lack of interpretability and reproducibility. To address this issue, recent works based on knockoffs variables and False Discovery Rate control, as introduced by Candès et al. (2018), have been employed. These knockoff variables are used as controls for variable selection and are randomly generated to mimic the

arbitrary dependence structure among the original features remaining conditionally independent of the output variable, given the input variables. DeepPINK of Lu et al. (2018) is an ANN architecture leveraging this strategy by integrating knockoff variables in the model to define the variable importance of each feature. A group-feature selection version of this approach can be found in Zhu and Zhao (2021) for datasets with group structures.

A common limitation encountered with the majority of these methods is the computational time required for the training procedure, mostly due to hyperparameter tuning. This may raise questions about their efficiency, especially when very fast approaches such as SpAM or RF yield comparable selection and accuracy results.

1.3.2. Contribution of Chapter 4

This section summarizes the collaborative article:
Savino, E. M., Lévy-Leduc, C. A novel variable selection method in nonlinear multivariate models using B-splines with an application to geoscience. Submitted and also available on HAL preprint *(hal-04434820)*.

The proposed method is implemented in the `absorber` R package which is available on the CRAN.

Inspired by Radchenko and James (2010), we propose approximating the function $f(x^{(1)},\dots,x^{(p)})$ appearing in (1.1) by a linear combination of B-splines of each variable $x^{(1)},\dots,x^{(p)}$ and of pairwise interaction of them as follows:

$$F(x^{(1)},\dots,x^{(p)}) = \sum_{\ell=1}^{p}\sum_{k=1}^{K+M}\beta_k^{(\ell)}B_k^{(\ell)}(x^{(\ell)}) + \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p}\left(\sum_{k=1}^{K+M}\sum_{q=1}^{K+M}\beta_{k,q}^{(\ell,j)}B_k^{(\ell)}(x^{(\ell)})B_q^{(j)}(x^{(j)})\right), \tag{1.26}$$

where $B_k^{(\ell)} = B_{k,M}^{(\ell)}$ is defined in (1.18) and (1.19) and where $\beta_k^{(\ell)}$ and $\beta_{k,q}^{(\ell,j)}$ are unknown coefficients. Observe that the column vector $(F(x_i^{(1)},\dots,x_i^{(p)}))_{1\le i\le n}$ (1.26) can be rewritten as follows:

$$\sum_{\ell=1}^{p}\Psi_\ell\boldsymbol{\beta}_\ell + \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p}\Phi_{\ell j}\boldsymbol{\beta}_{\ell,j}. \tag{1.27}$$

where Ψ_ℓ is a $n\times(K+M)$ matrix such that its ith row is equal to $(B_1^{(\ell)}(x_i^{(\ell)}),\dots,B_{K+M}^{(\ell)}(x_i^{(\ell)}))$ and $\boldsymbol{\beta}_\ell = \left(\beta_1^{(\ell)}\ \dots\ \beta_{K+M}^{(\ell)}\right)^T$ for $1\le\ell\le p$, 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 $1\le\ell<j\le p$.

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 (1.27) through the minimization of the following regularized criterion:

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

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the Y_i's being defined in (1.1), $\partial_\ell F(x_i)$ denotes the ℓth partial derivative of F defined in (1.26) 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 (1.27), the criterion can be rewritten as follows:

$$\begin{aligned}
&\left(\widehat{\boldsymbol{\beta}}_1(\lambda), \ldots, \widehat{\boldsymbol{\beta}}_p(\lambda), \widehat{\boldsymbol{\beta}}_{1,2}(\lambda), \ldots, \widehat{\boldsymbol{\beta}}_{(p-1),p}(\lambda)\right) \\
&= \underset{\substack{(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p) \\ (\boldsymbol{\beta}_{12},\ldots,\boldsymbol{\beta}_{(p-1)p})}}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j} \right\|_2^2 \right. \\
&\left. + \lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell \boldsymbol{\beta}_\ell + \sum_{j=\ell+1}^{p} (\partial_\ell \Phi_{\ell j}) \boldsymbol{\beta}_{\ell,j} + \sum_{1 \le j < \ell} (\partial_\ell \Phi_{j\ell}) \boldsymbol{\beta}_{j,\ell} \right\|_2 \right),
\end{aligned} \tag{1.28}$$

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:

$$\lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell \boldsymbol{\beta}_\ell + (\partial_\ell \Phi_{\ell\bullet}) \boldsymbol{\beta}_{\ell\bullet} + (\partial_\ell \Phi_{\bullet\ell}) \boldsymbol{\beta}_{\bullet\ell} \right\|_2 =: \lambda \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell) \boldsymbol{\gamma}_\ell \right\|_2, \tag{1.29}$$

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:

$$\begin{aligned}
&\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j} \right\|_2^2 \\
&= \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \frac{1}{2} \left(\sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j} + \sum_{\ell=2}^{p} \sum_{j=1}^{\ell-1} \Phi_{j\ell} \boldsymbol{\beta}_{j,\ell} \right) \right\|_2^2 \\
&=: \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell \boldsymbol{\gamma}_\ell \right\|_2^2.
\end{aligned} \tag{1.30}$$

Equation (1.30) by defining $\Theta_\ell = \left(\Psi_\ell \ \ \frac{1}{2}\Phi_{\ell\bullet} \ \ \frac{1}{2}\Phi_{\bullet\ell}\right)$ and setting $\Theta_1 = \left(\Psi_1 \ \ \frac{1}{2}\Phi_{1\bullet}\right)$ and $\Theta_p = \left(\Psi_p \ \ \frac{1}{2}\Phi_{\bullet p}\right)$, where $\Phi_{\ell\bullet} = (\Phi_{\ell(\ell+1)} \ldots \Phi_{\ell p})$ and $\Phi_{\bullet\ell} = (\Phi_{1\ell} \ldots \Phi_{(\ell-1)\ell})$. Combining (1.29) and (1.30), (1.28) can be rewritten as:

$$\left(\widehat{\boldsymbol{\gamma}}_1(\lambda), \ldots, \widehat{\boldsymbol{\gamma}}_p(\lambda)\right) = \underset{(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_p)}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell \boldsymbol{\gamma}_\ell \right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell) \boldsymbol{\gamma}_\ell \right\|_2 \right). \tag{1.31}$$

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, (1.31) can be rewritten as:

$$\left(\widehat{\boldsymbol{\alpha}}_1(\lambda), \ldots, \widehat{\boldsymbol{\alpha}}_p(\lambda)\right) = \underset{(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_p)}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \widetilde{\mathbf{X}}_\ell \boldsymbol{\alpha}_\ell \right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\| \boldsymbol{\alpha}_\ell \right\|_2 \right). \tag{1.32}$$

The last formulation of our variable selection criterion (1.32), introduced as ABSORBER, 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 $\widehat{\boldsymbol{\gamma}}_\ell(\boldsymbol{\lambda})$ are thus obtained as follows:

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

Thus, we define the active variables for each $\boldsymbol{\lambda}$ belonging to a given set Λ as follows:

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

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 (1.1) actually depends that have to be selected among the p variables.

Our variable selection method was assessed through numerical experiments, by defining two strategies to detect relevant variables.The results were promising as ABSORBER exhibited good selection performance, even with noisy observations and an increased number of input variables p. Furthermore, our method was compared to two other state-of-the-art methods, namely Random Forests and LassoNet. These comparisons revealed that our method outperformed the two other methods in the context of their applications to nonlinear functions and a geochemical system, as only ABSORBER successfully identified the relevant variables.

1.4. Applications in the context of a EURAD work package

1.4.1. Context

The European Joint Programme on Radioactive Waste Management (EURAD) fosters European collaborations among organizations from 23 countries engaged in radioactive waste disposal projects, aiming to develop robust and sustained science, technology, and knowledge[2] for supporting safe radioactive waste management activities. One of the main goals of EURAD is to understand and quantify the evolution of interactions within the radioactive waste disposal facility. Thus, one of the undertaken studies is dedicated to investigating interfaces between different rock components involved in waste disposal to support the design and optimization of barrier systems. For this purpose, EURAD has proposed a work package (WP), entitled: "Development and Improvement of NUmerical methods and Tools for modeling coupled processes" (DONUT), which concentrates solely on numerical simulations.

Led by Francis Claret[3], this WP includes collaborators from different organizations and countries. Its objective is to address challenges encountered in RTMs, particularly CPU time constraints and the study of highly multi-scale coupled processes, by introducing cutting-edge methods for high-performance computing, as described by Claret et al. (2022). A significant aspect of the WP involves comparing innovative surrogate modeling approaches through a benchmark exercise, coordinated by Nikolaos Prasianakis[4], to assess the efficiency, robustness, accuracy, and computational time of the developed methods. Two geochemical applications are considered in the DONUT WP: a uranium study and the temporal evolution of the behavior of cementitious materials. The former pertains to radionuclide migration while the latter

[2] https://www.ejp-eurad.eu/publications/eurad-sra

[3] BRGM, 3 Avenue Claude Guillemin, 45060 Orléans, France

[4] Laboratory for Waste Management, Paul Scherrer Institute, CH, 5232, Villigen PSI, Switzerland

assesses one of the major construction materials existing in these radioactive waste disposal facilities. Several geochemical systems with an increasing level of complexity, corresponding to an increasing number of input variables, are defined to simulate these two applications.

1.4.2. Contribution of Chapter 5

A part of this section summarizes the article:
N.I. Prasianakis, E. Laloy, D. Jacques, J.C.L. Meeussen, C. Tournassat, G.D. Miron, D. A. Kulik, A. Idiart, E. Demirer, E. Coene, B. Cochepin, M. Leconte, M. E. Savino, J. Samper II, M. De Lucia, S. V. Churakov, O. Kolditz, C. Yang, J. Samper, F. Claret. Geochemistry and Machine Learning: review of methods and benchmarking. *To be submitted.*

In this chapter, our objective is to compare the developed estimation methods, with and without variable selection, against two other benchmark methods within the context of the DONUT WP, focusing specifically on the cementitious application case.

Input variables belonging to a p-dimensional space are created using a Latin Hypercube Sampling (LHS). Then, high-quality data are generated using powerful geochemical solvers such as ORCHESTRA, GEMS and Phreeqc (see Meeussen (2003); Kulik et al. (2013); Parkhurst and Appelo (2013), respectively, for more details on each solver). These datasets are used for training and validation purposes, allowing for the computation of statistical measures and facilitating comparisons between the different machine learning approaches. The geochemical context presented here concerns the hydration and evolution of cementitious systems under 25°C and takes as input variables the amounts of different oxides and the amount of water. Six cementitious systems are presented in the DONUT benchmark with an increasing number of input variables. Hereafter, we will only consider the most complex case involving 7 chemical elements as inputs. Our aim is to estimate 42 output chemical elements, related to aqueous species, solid phases and auxilliary variables.

Firstly, the variable selection ABSORBER introduced in Section 1.3.2 was applied to identify the most relevant variables for each output variable. The results revealed varying selected variables depending on the nature of the considered output element. For instance, solid phases tend to depend on only one or a few variables whereas the aqueous phases appear to rely on nearly every input variable. To simplify the study, only output variables depending on at most three input variables were considered for further analysis. Subsequently, we proposed to employ the active learning approach using Gaussian Processes (GP AL), as previously introduced in Section 1.2.2 with or without our variable selection approach. The obtained prediction error emphasized the benefits of using only the relevant variables. Then, we used GP AL, GLOBER and GLOBER-c, the latter two being presented in Section 1.2.3, to compare their efficiency with two other methods introduced by teams participating in the DONUT WP. Our methods incorporated only the relevant variables determined with ASORBER. The results showed superior performance in our favor, particularly for outputs depending on only one input variable, and once again demonstrated the value of using our estimation methods with only the variables selected with ABSORBER.

Chapter 2 - An active learning approach for improving the performance of equilibrium based chemical simulations

Publication

The content of this chapter is the subject of the article:
Savino, M., Lévy-Leduc, C., Leconte, M., Cochepin, B. (2022). An active learning approach for improving the performance of equilibrium based chemical simulations. *Computational Geosciences*, 26(2), 365–380.

Abstract

In this chapter, we propose a novel sequential data-driven method for dealing with equilibrium based chemical simulations, which can be seen as a specific machine learning approach called active learning. The underlying idea of our approach is to consider the function to estimate as a sample of a Gaussian process which allows us to compute the global uncertainty on the function estimation. Thanks to this estimation and with almost no parameter to tune, the proposed method sequentially chooses the most relevant input data at which the function to estimate has to be evaluated to build a surrogate model. Hence, the number of evaluations of the function to estimate is dramatically limited. Our active learning method is validated through numerical experiments and applied to a complex chemical system commonly used in geoscience.

Table of contents

2.1. Introduction

Computing the concentrations at equilibrium of reactive species is well known to be a challenging issue when the number of species is high and/or when the reaction involves the dissolution or the precipitation of minerals White et al. (1958); Smith (1980); de Capitani and Brown (1987). The numerical resolution of these non-linear problems can quickly become so time consuming that the coupling with other physical processes has to be simplified. For instance in the case of reactive transport, it means that the size of the geometric model has to be drastically limited leading typically to a one dimensional model or that the number of time steps has to be reduced. To overcome this issue, research efforts have been dedicated to the improvement of the numerical scheme aiming at speeding up the computations. A classical approach consists in using a splitting operator technique to solve separately the transport of the chemical species and the chemical reaction between those species Marchuk (1990); Sportisse (2000); Descombes (2001); Carrayrou et al. (2004); Simpson and Landman (2007). With this approach a specific optimization for each part of the resolution can be performed especially by taking advantage of the parallel architecture of computers Faragó and Geiser (2007); Geiser (2011); Geiser et al. (2020).

However, despite the significant improvements of the numerical solvers and preconitionners during the last decades, three dimensional large scale modelling of complex reactive transport over a long period of time, namely many time steps, remains almost impossible to solve with standard computers. Consequently, the recent success of machine learning (ML) in various fields have quickly drawn attention of geoscientists because ML seems to be able to solve very complex problems with a reasonable cost in terms of computational ressources.

The main idea behind the ML success is to provide an estimation of the solution of the full simulation model that can replace it. Two of the most popular approaches are model order reduction and data-driven models also called surrogate models. The first one requires to understand the underlying chemical processes to create a simplified model while preserving some physical principles Rao et al. (2013). In the second approach, the underlying chemical processes are not assumed to be known or understood and a model is solely built from a limited but potentially significant set of values of the solution of the full simulation model associated to some specific input values Guérillot and Bruyelle (2020). Since the number of required values is unknown a priori, choosing the optimal input values and parameters used for building the surrogate model is crucial and usually challenging.

In this chapter, we propose a novel sequential data-driven method for dealing with equilibrium based chemical simulation, which can thus be seen as an active learning approach inspired by the ideas contained in Srinivas et al. (2012); Jala et al. (2016). With such an approach, our goal is to minimize the number of evaluations of the function that has to be estimated to build a surrogate model. Our approach consists in modeling the function to estimate as a sample of a Gaussian Process (GP) which allows us to provide an error estimation to sequentially choose the most relevant input data until a given stopping criterion is fulfilled. The advantage of our approach is that the number of required evaluations of the function to estimate is very limited and that there are no parameter to tune.

The chapter is organized as follows. In Section 2.2, our approach is described. Some numerical experiments are provided in Section 2.3 to illustrate the statistical and numerical performance of our method. It is then applied in Section 2.4 to a multidimensional example coming from Kolditz et al. (2012) which includes several chemical elements and minerals.

2.2. Description of our approach

In this section, we describe our active learning approach for estimating a real-valued function f defined on a compact subset $\mathcal{A} \subset \mathbb{R}^d$ by using only a few number of sequentially well-chosen points at which f is evaluated.

We adopt a Bayesian point of view which consists in considering f as a sample of a zero-mean Gaussian process (GP) having a covariance function k that we shall denote by GP(0,$k(\cdot,\cdot)$) in the following. The advantage of this approach is that, conditionally on a set of t observations $\mathbf{y}_t = (y_1, \ldots, y_t)'$ where $y_i = f(x_i)$, x_i belonging to $\mathcal{A}$, the posterior distribution is still a GP having a mean μ_t and a covariance function k_t given by

$$\mu_t(u) = \mathbf{k}_t(u)' \mathbf{K}_t^{-1} \mathbf{y}_t \ , \tag{2.1}$$

$$k_t(u,v) = k(u,v) - \mathbf{k}_t(u)' \mathbf{K}_t^{-1} \mathbf{k}_t(v) \ , \tag{2.2}$$

where $\mathbf{k}_t(u) = [k(x_1,u) \ldots k(x_t,u)]'$. Here $'$ denotes the matrix transposition, u and v are in $\mathcal{A}$ and $\mathbf{K}_t = [k(x_i,x_j)]_{1 \leq i,j \leq t}$, where the x_i's are in $\mathcal{A}$. For further details on GP, we refer the reader to Rasmussen and Williams (2006) in which their properties are thoroughly presented.

In our case, f models a physical quantity that is assumed to be smooth, so for our applications we shall consider two covariance functions that are commonly used in this case. The first one is the squared exponential (SE) covariance function

$$k_{\text{SE}}(u,v) = \exp\left(-\frac{1}{2}(u-v)' M^{-1} (u-v)\right) \ , u, v \in \mathcal{A} \subset \mathbb{R}^d \ , \tag{2.3}$$

$$M = \text{diag}\left(\ell_1^2, \ldots, \ell_d^2\right) \ , \ \ell_1 \ , \ \ell_2 \ , \ldots \ , \ell_d > 0 \ . \tag{2.4}$$

Here the ℓ_1 , ℓ_2 , $\ldots$, ℓ_d hyperparameters are the characteristic length scales. Actually, these hyperparameters can be understood as how far you need to move along a particular axis in the input space so that the function values become uncorrelated. For further details, we refer the reader to Section 5.1 of Rasmussen and Williams (2006). Note that Definition (2.3) allows us to model anisotropic response surfaces.

As explained in Rasmussen and Williams (2006), since this covariance function is infinitely differentiable, the GP with this covariance function has mean square derivatives of all orders. As argued by Stein (1999) such strong smoothness assumptions may be unrealistic for modeling many physical processes, so we shall also consider another covariance function belonging to the Matérn class of covariance functions defined by

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} r\right)^{\nu} K_{\nu}\left(\sqrt{2\nu} r\right) \ , \nu > 0 \ , \tag{2.5}$$

where K_ν is a modified Bessel function with Bessel order ν, see (Abramovitz and Stegun, 1965, Section 9.6), and r is defined by

$$r = \sqrt{(u-v)' M^{-1} (u-v)} \ , u, v \in \mathcal{A} \ , \tag{2.6}$$

M being defined in (2.4). In this situation, as explained in Rasmussen and Williams (2006), the GP is q-times mean-square differentiable if and only if $\nu > q$. Here, we shall focus on the case where $\nu = 5/2$, for which $k_{\text{Matérn}}$ has a computationally advantageous expression. Indeed, for $\nu = p + \frac{1}{2}$, where p is in $\mathbb{N}$,

$$k_{\text{Matérn}}(r) = \exp\left(-\sqrt{2\nu} r\right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\sqrt{8\nu} r\right)^{p-i} \ , \tag{2.7}$$

with r defined in (2.6); see Abramovitz and Stegun (1965, Equation 10.2.15) for further details.

In the following, we shall denote by A a fine grid of $\mathcal{A}$:

$$\mathrm{A} = \{\mathrm{x}_1, \ldots, \mathrm{x}_m\} \subset \mathcal{A} \, . \tag{2.8}$$

This grid is either a regular grid of $\mathcal{A} \subset \mathbb{R}^d$ when d is small (usually 1 or 2) or a Latin Hypercube Sampling for larger values of d. Note that this grid contains the points at which the estimation of f is performed and that the points at which f is evaluated are chosen in this grid.

Inspired by Srinivas et al. (2012) who proposed a sequential approach for maximizing a function by modeling it using a Gaussian process, we propose a strategy which consists in adding the new point x_{t+1} to the set of t observations at which f needs to be evaluated as follows:

$$x_{t+1} \in \operatorname*{Arg\,max}_{x \in \mathrm{A}} \sigma_t(x) \, , \tag{2.9}$$

where

$$\sigma_t(x)^2 = k_t(x, x), \tag{2.10}$$

k_t being defined in (2.2) and $\operatorname*{Arg\,max}_{x \in \mathrm{A}} \sigma_t(x)$ being the set of $x \in \mathrm{A}$ where $\sigma_t(x)$ reaches its maximum. Note that the points $x_1, x_2, \ldots, x_t, x_{t+1}, \ldots$ at which f needs to be evaluated are chosen in the fine grid A of $\mathcal{A}$ defined in (2.8).

2.2.1. Estimating the characteristic length scales

Previously, we assumed that the characteristic length scales $\boldsymbol{\ell} = (\ell_i)_{\{1 \leq i \leq d\}}$ were known. However, this is obviously not the case in real-data applications. We propose using the maximum-likelihood strategy described in Rasmussen and Williams (2006) to estimate $\boldsymbol{\ell}$. This adds a step to the method previously described, as the ℓ_i's have to be estimated before evaluating the posterior distribution of the GP using (2.1) and (2.2). Hence, for the observation set $\{(x_1, y_1), \ldots, (x_t, y_t)\}$ with $y_i = f(x_i)$, $1 \leq i \leq t$, the posterior log-likelihood given by:

$$-\frac{1}{2}\mathbf{y}_t' \mathbf{K}_t^{-1} \mathbf{y}_t - \frac{1}{2} \log |\mathbf{K}_t| - \frac{t}{2} \log 2\pi \, , \tag{2.11}$$

with $\mathbf{y}_t = (y_1, \ldots, y_t)'$ and $\mathbf{K}_t = [k(x_i, x_j)]_{1 \leq i,j \leq t}$, has to be maximized with respect to $\boldsymbol{\ell}$.

2.2.2. Summary of our strategy

Our method was implemented by using the GaussianProcessRegressor class of the `scikit-learn 0.20.3` module of Python which only provides the computation of μ_t and σ_t defined in (2.1) and (2.10). Our sequential approach is summarized in Algorithm 1.

Further comments on the stopping criteria appearing in Algorithm 1 are given below.

2.2.3. Stopping criteria

Different stopping criteria based on the following quantities can be used.

- **Ratio variance.** At each iteration t of our method, the following average is computed:

$$R_n(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\max_{x \in \mathrm{A}} \sigma_t^2(x)}{\max_{x \in \mathrm{A}} \sigma_{t-i}^2(x)}, \tag{2.12}$$

  where σ_t is defined in (2.10) and $n = 2, 5$ or 10. This criterion will be then compared to a threshold to determine if the maximal variance reach a plateau. In some cases, σ_{t-i}^2 can

Algorithm 1

Input: $x_1, \ldots, x_{t_1}$ a small initial set of points of A where f has been evaluated
$t = t_1$; Choose a covariance function k among SE and Matérn.
While the stopping criterion is not fulfilled

- Estimate $\boldsymbol{\ell}$ by using (2.11)
- Evaluate the posterior distribution of the GP using (2.1) and (2.2), and the variance $\sigma_t(x)^2$ for all x in A
- Choose x_{t+1} in A using (2.9)
- Evaluate f at this point: $y_{t+1} = f(x_{t+1})$
- Add this new observation to the set of points at which f is evaluated which becomes $x_1, \ldots, x_t, x_{t+1}$
- $t \leftarrow t + 1$

The function f is estimated by μ_t defined in (2.1).

be less than σ_t^2 so in order to detect the smallest variations, we also have to make sure that the ratio does not exceed the inverse of the chosen threshold. Thus, the associated stopping criterion is: interrupt the algorithm when t is such that

$$0.9 < R_n(t) < \frac{1}{0.9}. \tag{2.13}$$

- **Mobile average.** At each iteration t of our method, the following average is computed:

$$M_\ell(t) = \frac{1}{\ell} \sum_{j=0}^{\ell-1} \max_{x \in \mathrm{A}} \sigma_{t-j}^2(x) \tag{2.14}$$

for $\ell = 5$ or 10 where σ_t is defined in (2.10). The associated stopping criterion is: interrupt the algorithm when t is such that

$$M_\ell(t) < 0.01. \tag{2.15}$$

- **Maximal variance.** At each iteration t of our method,

$$V(t) = \max_{x \in \mathrm{A}} \sigma_t^2(x) \tag{2.16}$$

is computed where σ_t is defined in (2.10). The associated stopping criterion is: interrupt the algorithm when t is such that

$$V(t) < s, \tag{2.17}$$

where $s = 0.01$ or 0.001 in the following.

The statistical performance of these different criteria are investigated in Section 2.3. Note that the values reported here for each criteria (0.9, 0.01 or 0.001) were chosen based on some numerical experiments since they appear to be relevant to detect a plateau in the maximal variance.

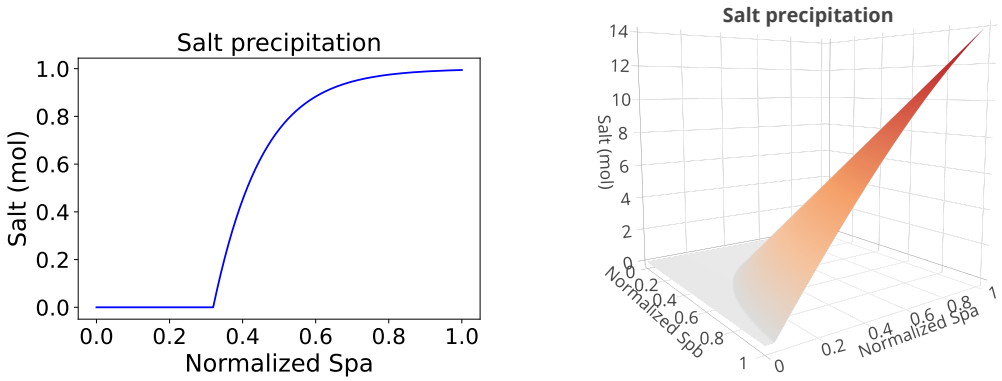

Figure 2.1: Functions f to estimate when $d = 1$ (left) and $d = 2$ (right).

2.3. Numerical experiments

To illustrate our method we consider hereafter the estimation of the amount of a "Salt" mineral as a function of the concentrations of its constituents Sp_a^+ and Sp_b^-. For this example, the thermodynamic constants of the halite salt (NaCl) were considered because there are only two constitutive elements and because they do not depend on the pH of the solution. From our point of view, there is no theoretical limitation in the application of our method to more complex salts or minerals.

Following the law of mass action, the dissolution reaction of this mineral writes:

$$\mathrm{Salt} \rightleftharpoons \mathrm{Sp}_\mathrm{a}^+ + \mathrm{Sp}_\mathrm{b}^-.$$

At equilibrium, the activity of these elements $\mathrm{a}_{Sp_a^+}$ and $\mathrm{a}_{Sp_b^-}$ obey the solubility product

$$\mathrm{K}_{\mathrm{Salt}} = \mathrm{a}_{\mathrm{Sp}_\mathrm{a}^+}\mathrm{a}_{\mathrm{Sp}_\mathrm{b}^-} = 10^{1.570}.$$

The amount of Salt was first calculated with PHREEQC Parkhurst and Appelo (2013) as a function of the concentrations of Sp_a^+, which is normalized so that $\mathcal{A} = [0, 1]$. It corresponds to the case $d = 1$ below. The corresponding function f is displayed in the left part of Figure 2.1 where A is a regular grid of $\mathcal{A}$ with $m = 1140$ points. Then, the amount of Salt was computed with PHREEQC as a function of the concentrations of Sp_a^+ and Sp_b^-, which are also normalized so that $\mathcal{A} = [0, 1]^2$. It corresponds to the case $d = 2$ below. The corresponding function f is displayed in the right part of Figure 2.1 where A is a regular grid of $\mathcal{A}$ with $m = 40\,000$ points.

2.3.1. Case $d = 1$

The different steps of our approach summarized in Algorithm 1 are illustrated in Figure 2.2 where our procedure was arbitrarily stopped after 40 evaluations. Here, we used the SE covariance function defined in (2.3).

The approach starts with $t_1 = 3$ points randomly chosen in A. Then, a new point in green is added to the set of points at which an evaluation of f is required. This point corresponds to the position on the x-axis where the uncertainty σ_t^2 associated to the estimation of f is maximized. We can see from this figure which displays the true function f, the estimation of f and the points at which f has been evaluated that 35 evaluation points are enough to obtain a very accurate estimation of f. To further investigate the statistical performance of

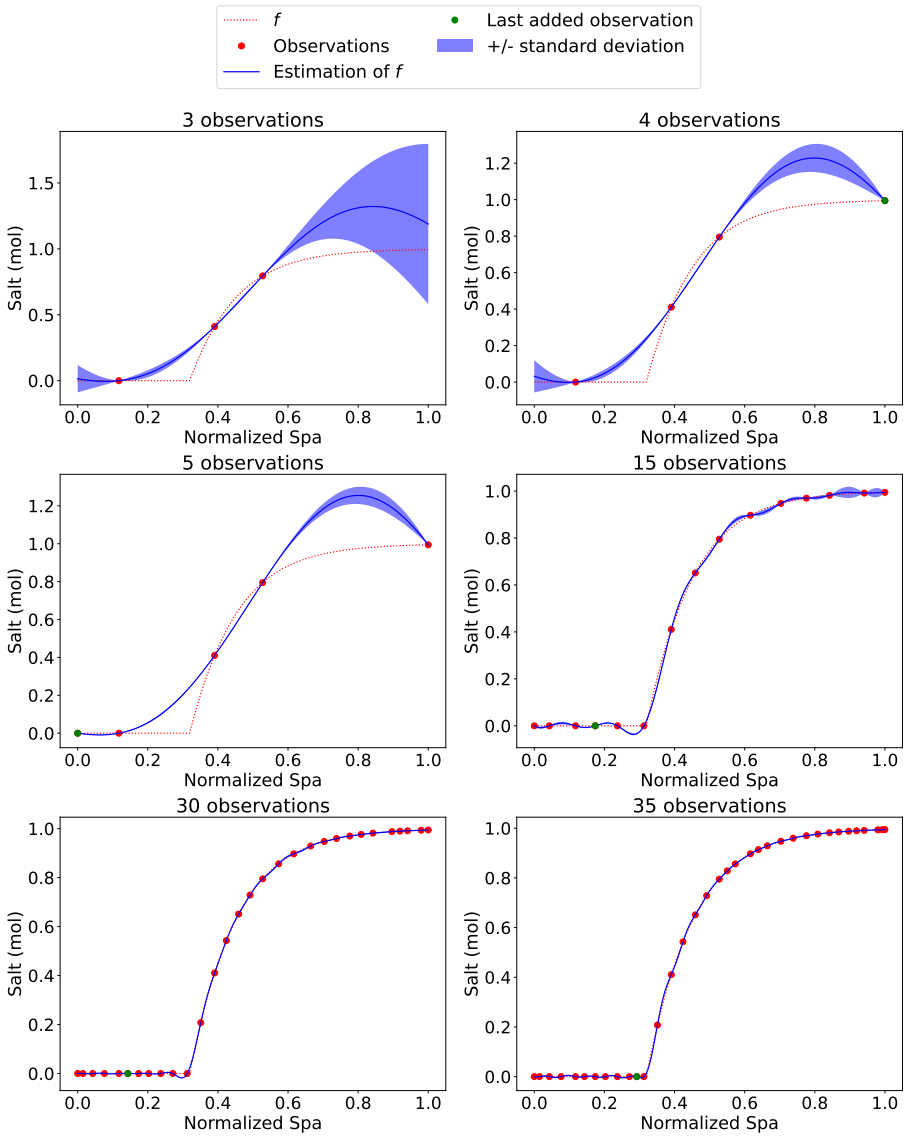

Figure 2.2: Illustration of our active learning approach for estimating the function displayed in the left part of Figure 2.1 by starting from $t_1 = 3$ observations randomly chosen in A with the squared exponential covariance function.

our approach, we used the following measures:

$$\text{Normalized MAE}(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{|y_i - \mu_t(x_i)|}{y_{max} - y_{min}}, \tag{2.18}$$

where μ_t is the estimation of f obtained at iteration t, m is the number of elements in the grid A and y_{min} and y_{max} are the minimum and maximum values, respectively, found for the evaluation of f on the initial grid ;

$$\text{Normalized sup norm}(t) = \max_{1\leq i\leq m}\frac{|y_i - \mu_t(x_i)|}{y_{max} - y_{min}}. \tag{2.19}$$

$$V(t) = \max_{x\in\mathrm{A}}\sigma_t^2(x), \tag{2.20}$$

where σ_t is defined in (2.10).

The average and the standard deviation of these measures obtained from 10 replications of the initial set of points are displayed in Figure 2.3 for the covariance functions defined in (2.3) and (2.7) and $3 \leq t \leq 40$. Note that the average and the standard deviation are computed by using 10 different initial sets of points.

We can see from this figure that the performance of our approach is slightly better for the Matérn covariance function than for the squared exponential function. It can indeed reach a normalized MAE (resp. normalized sup norm) of 10^{-3} (resp. $10^{-1.5}$) by using only 40 evaluations of the function to estimate. This might come from the discontinuity of the first derivative of the function to estimate where the salt starts to precipitate.

In the left part of Figure 2.4 the statistical performance of our approach including the stopping criteria are further investigated thanks to the computation of the previous performance measures defined in (2.18), (2.19) and (2.16): $\text{Normalized MAE}(t^\star)$, $\text{Normalized Sup norm}(t^\star)$ and $V(t^\star)$ where $t^\star$ is the stopping iteration which may be different for each stopping criterion.

We can see from the left part of Figure 2.4 that among all of the stopping criteria, "ratio variance 5" (R_5), "ratio variance 10" (R_{10}) and "mobile average 10" (M_{10}) are those providing the best estimations of the function f. Moreover, we can observe from the right part of this figure that our active learning approach only requires between 15 and 40 evaluations of the function to estimate instead of the 1140 points of the initial grid to provide a very accurate estimation of the function f. With such an approach, we can thus expect a significant reduction of the computational time especially in situations where the computational load associated to the evaluation of f is high. Figure 2.4 also shows that, in this case, the impact of the covariance function is not significant even though the first derivative of the function to approximate is not continuous, namely where the salt precipitates.

2.3.2. Case $d = 2$

In order to further assess the performance of our approach we now consider the estimation of the amount of Salt as a function of the concentrations of Sp_a^+ and Sp_b^-.

The different steps of our approach summarized in Algorithm 1 are illustrated in Figure 2.5. Here, we used the SE covariance function defined in (2.3).

The approach starts with $t_1 = 3$ points randomly chosen in $\mathrm{A} \subset [0,1]^2$ obtained thanks to a regular grid of 200×200 points. Then, new points (orange bullets) are added one by one to the set of points at which an evaluation of f is required. These points correspond at each iteration to the position in $\mathrm{A} \subset [0,1]^2$ where the uncertainty σ_t^2 associated to the estimation of

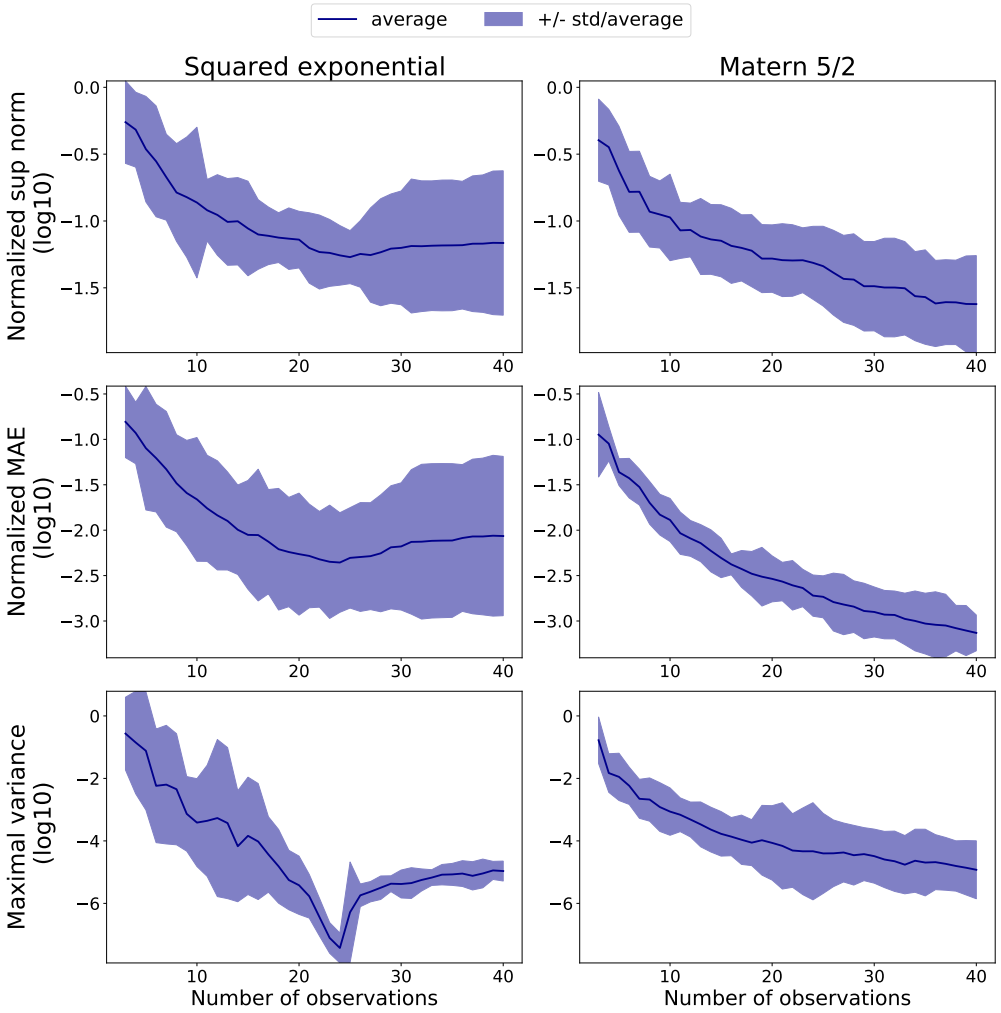

Figure 2.3: Average and standard deviation of different statistical measures for the squared exponential covariance function defined in (2.3) (left) and for the Matern covariance function defined in (2.7) (right) in the case $d = 1$.

f is maximized. We can see from this figure which displays the true function f, the estimation of f and the points at which f has been evaluated that 35 evaluation points are enough to obtain a very accurate estimation of f.

In the $d = 2$ case, the average and the standard deviation of the statistical measures defined in (2.16), (2.18) and (2.19) obtained from 10 replications of the initial set of points are displayed in Figure 2.6 for the squared exponential and the Matérn covariance function defined in (2.3) and (2.7)) for $3 \leq t \leq 100$. We can see that for both choices of covariance function the performance of our approach are similar: it can reach a normalized sup norm (resp. normalized MAE) of $10^{-1.5}$ (resp. $10^{-2.5}$) by using only 100 evaluations of the function to estimate instead of the $40\,000$ points of the grid A. We also observe a smoother behavior of the maximal variance with the Matérn covariance function even though the final values are close.

We can see from the left part of Figure 2.7 that most of the stopping criteria provide an accurate estimation of the function except "ratio variance 2" (R_2). As for the $d = 1$ case, the stopping criteria R_{10} and M_{10} provide very satisfactory results. Moreover, we can observe

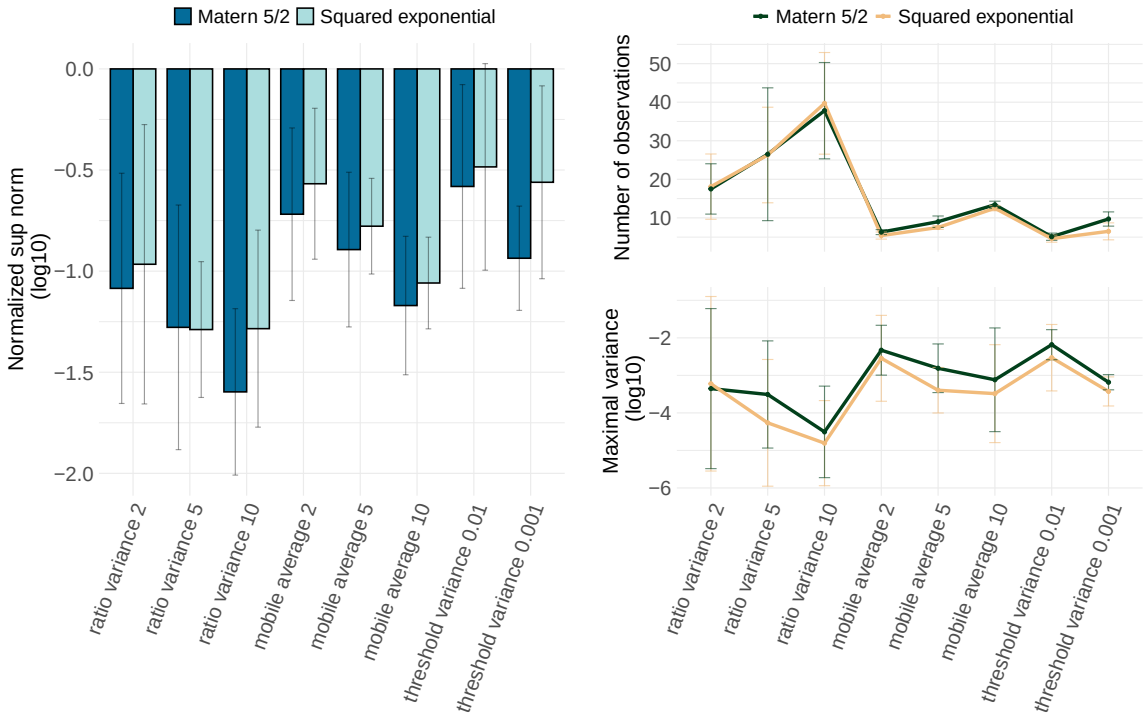

Figure 2.4: Left: Statistical assessment of the error estimation of f displayed in the left part of Figure 2.1 for the stopping criteria defined in (2.13), (2.15) and (2.17) for the squared exponential and the Matérn covariance functions. Top right: Number of evaluations required for the considered stopping criteria. Bottom right: Values of $V(t^\star)$ where V is defined in (2.16) and $t^\star$ is the stopping iteration which changes from one stopping criterion to another.

from the right part of this figure that thanks to our active learning approach, $30-50$ evaluations of the function to estimate are required instead of the $40\,000$ points of the initial grid to provide a very accurate estimation of the function f. Once again, with our approach, we can thus expect a significant reduction of the computational burden especially in situations where the computational load associated to the evaluation of f is high.

In this case, the choice of the covariance function might result from a trade-off between accuracy and number of evaluation points. However, the accuracy and the number of evaluation points do not change drastically suggesting that the choice of the covariance function is still not significant.

2.4. Application to a multidimensional geochemical system

The chemical problem solved in this section derives from Kolditz et al. (2012). The chemical setup is based on the thermodynamic data for aqueous species and minerals available in the Phreeqc.dat database distributed with PHREEQC Parkhurst and Appelo (2013). The compositional system actually solved consists of 14 species in solution, 2 mineral components, 8 geochemical reactions and 2 mineral dissolution-precipitation reactions:

$$\mathrm{H_2O} \rightleftharpoons \mathrm{H^+} + \mathrm{OH^-}, \log \mathrm{K_1} = -13.987$$

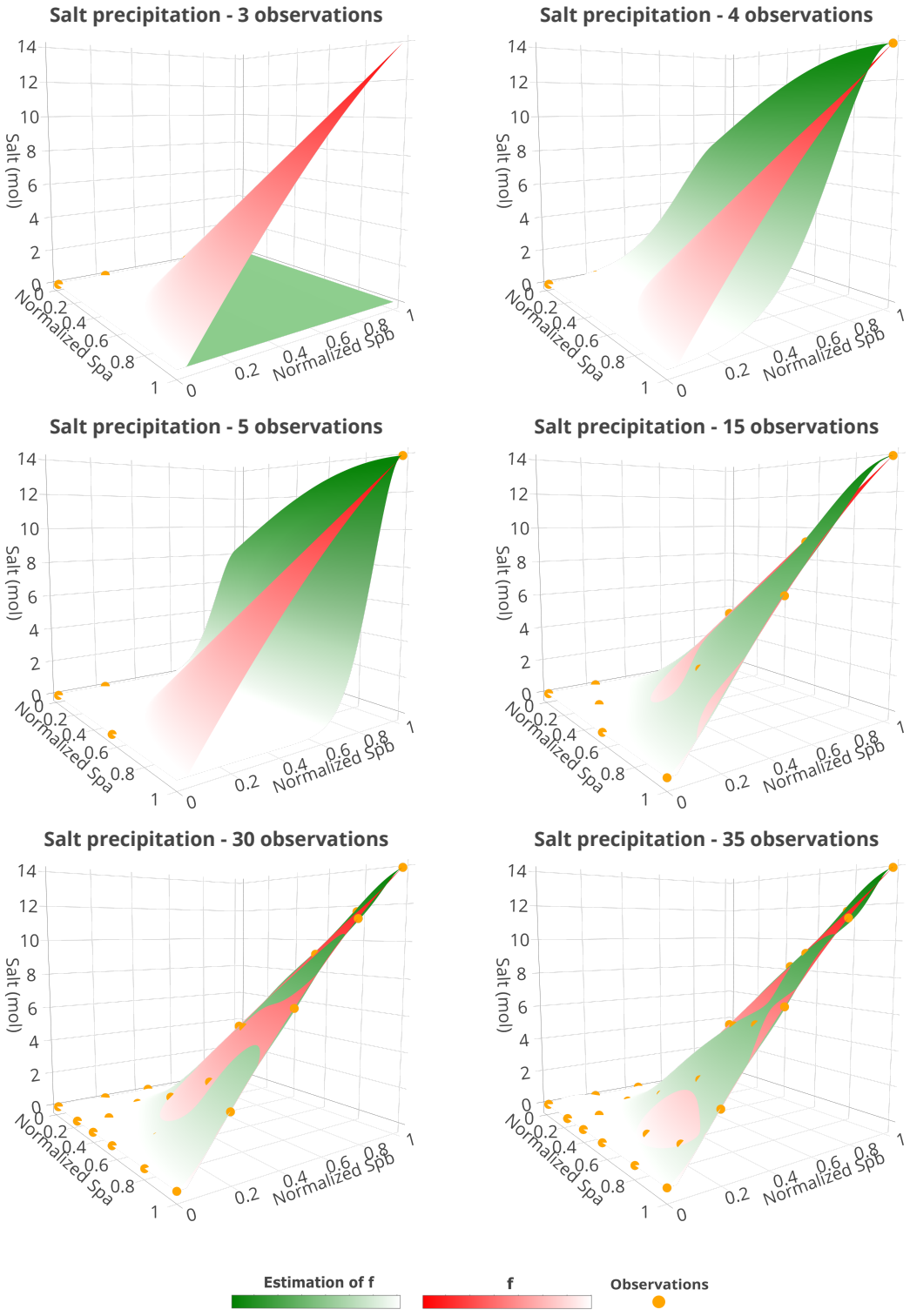

Figure 2.5: Illustration of our active learning approach for estimating the function displayed in the right part of Figure 2.1 by starting from $t_1 = 3$ observations randomly chosen in $\mathrm{A} \subset [0, 1]^2$ for the squared exponential covariance function.

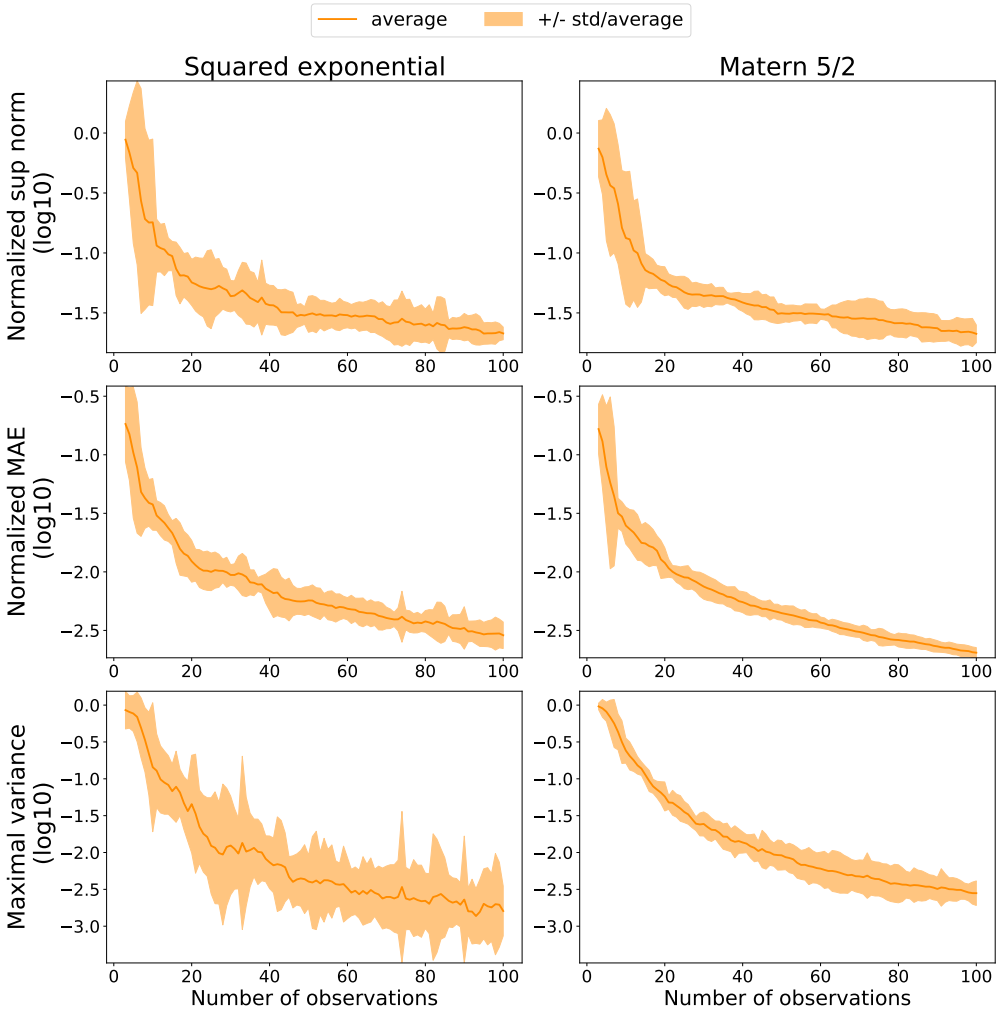

Figure 2.6: Average and standard deviation of different statistical measures for the squared exponential covariance function defined in (2.3) (left) and for the Matern covariance function defined in (2.7) (right) in the case $d = 2$.

$$\mathrm{HCO_3^- \rightleftharpoons CO_3^{2-} + H^+, logK_2 = -10.329}$$
$$\mathrm{CO_2 + H_2O \rightleftharpoons CO_3^{2-} + 2H^+, logK_3 = -16.681}$$
$$\mathrm{CaHCO_3^+ \rightleftharpoons Ca^{2+} + CO_3^{2-} + H^+, logK_4 = -11.435}$$
$$\mathrm{MgHCO_3^+ \rightleftharpoons Mg^{2+} + H^+ + CO_3^{2-}, logK_5 = -11.399}$$
$$\mathrm{CaCO_{3(aq)} \rightleftharpoons Ca^{2+} + CO_3^{2-}, logK_6 = -3.224}$$
$$\mathrm{MgCO_3 \rightleftharpoons Mg^{2+} + CO_3^{2-}, logK_7 = -2.98}$$
$$\mathrm{MgOH^+ + H^+ \rightleftharpoons Mg^{2+} + H_2O, logK_8 = 11.44}$$
$$\mathrm{Calcite \rightleftharpoons CO_3^{2-} + Ca^{2+}, logK_9 = -8.48}$$
$$\mathrm{Dolomite \rightleftharpoons Ca^{2+} + Mg^{2+} + 2CO_3^{2-}, logK_{10} = -17.09}$$

Then, each amount of mineral (calcite or dolomite, respectively) is computed with PHREEQC Parkhurst and Appelo (2013) as a function of the total elemental concentrations (C,

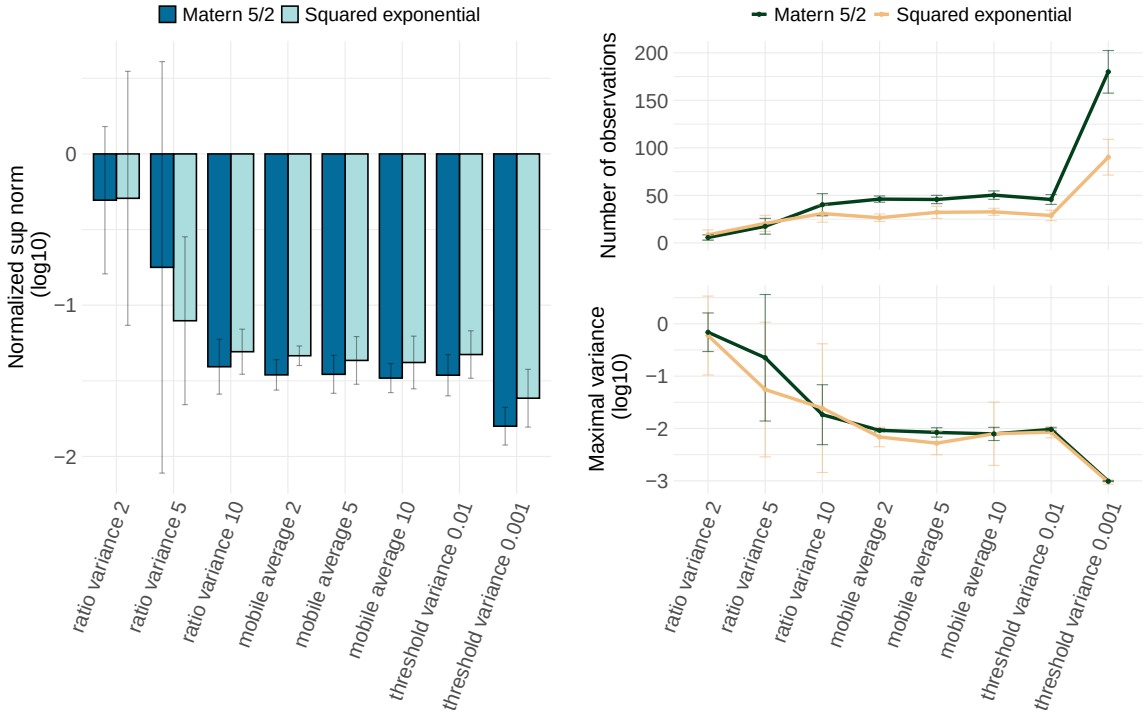

Figure 2.7: Left: Statistical assessment of the error estimation of f displayed in the right part of Figure 2.1 for the stopping criteria defined in (2.13), (2.15) and (2.17) for the squared exponential and the Matérn covariance functions. Top right: Number of evaluations required for the different considered stopping criteria. Bottom right: Values of $V(t^\star)$ where V is defined in (2.16) and $t^\star$ is the stopping iteration which changes from one stopping criterion to another.

Ca, Cl, Mg), the pH (as $-\log(\mathrm{H}^+)$) and the mineral amount (dolomite or calcite, respectively), which are normalized so that $\mathcal{A} = [0,1]^6$. Here, our goal is to estimate the functions f_1 and f_2 defined as follows:

$$\text{calcite} = f_1(\text{C, Ca, Cl, Mg, pH, dolomite}) \text{ and } \text{dolomite} = f_2(\text{C, Ca, Cl, Mg, pH, calcite}), \tag{2.21}$$

by using the minimal number of evaluations of these functions. For this, we shall use a grid A built thanks to a Latin Hypercube Sampling (LHS) of $\mathcal{A}$ with $m = 100\,000$ points.

In the left part of Figure 2.8 the amount of calcite is displayed as a function of C and Ca for Cl=2×10^{-3} mol/kgw, Mg=10^{-5} mol/kgw, pH=10, dolomite=0 mol which corresponds to $f_1(\mathrm{C}, \mathrm{Ca}, 2 \times 10^{-3}, 10^{-5}, 10, 0)$. In the right part of Figure 2.8 the amount of dolomite is displayed as a function of Ca and Mg for C=5×10^{-4} mol/kgw, Cl=2×10^{-3} mol/kgw, pH=10, calcite=0 mol which corresponds to $f_2(5 \times 10^{-4}, \mathrm{Ca}, 2 \times 10^{-3}, \mathrm{Mg}, 10, 0)$.

Illustrations of our active learning approach for estimating these functions are shown in Figures 2.13 and 2.14 of the Appendix.

2.4.1. Calcite precipitation

The average and the standard deviation of the different statistical measures obtained from 10 replications of the initial set of points are shown in Figure 2.9 for the squared exponential and the Matérn covariance functions defined in (2.3) and (2.7) for $3 \leq t \leq 500$. We can see

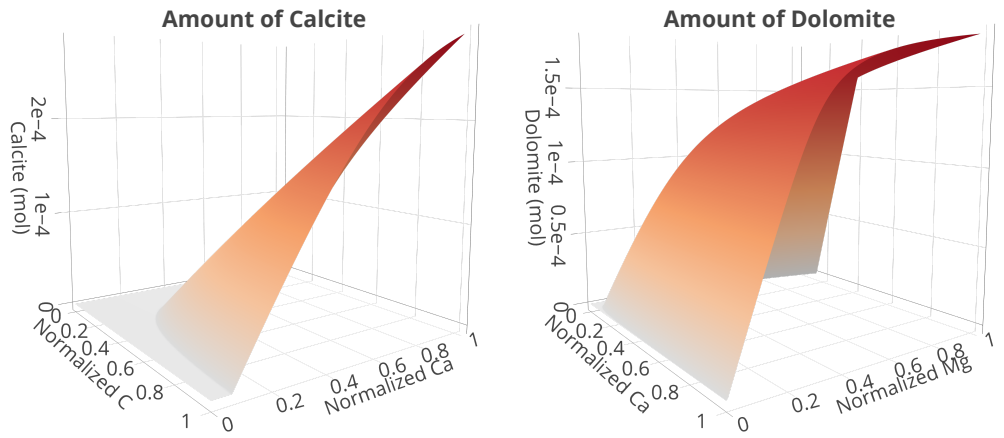

Figure 2.8: Left : Amount of calcite as a function of C and Ca for Cl=2×10^{-3} mol/kgw, Mg=10^{-5} mol/kgw, pH=10, dolomite=0 mol: $f_1(\mathrm{C}, \mathrm{Ca}, 2 \times 10^{-3}, 10^{-5}, 10, 0)$ where f_1 is defined in (2.21). Right : Amount of dolomite as a function of Ca and Mg for C=5×10^{-4} mol/kgw, Cl=2×10^{-3} mol/kgw, pH=10, calcite=0 mol: $f_2(5 \times 10^{-4}, \mathrm{Ca}, 2 \times 10^{-3}, \mathrm{Mg}, 10, 0)$ where f_2 is defined in (2.21).

that for both choices of covariance functions, the maximal variance and the statistical precision measures keep decreasing as the number of evaluations increases. For instance, our method allows us to have a normalized sup norm (resp. normalized MAE) of $10^{-0.5}$ (resp. $10^{-1.4}$) with only 500 evaluations instead of the 100000 points of the grid A for both covariance functions. However, the maximal variance is around $10^{-3.5}$ (resp. $10^{-1.5}$) for the squared exponential (resp. Matérn) covariance function.

Moreover, we can see from Figure 2.10 that when the mobile average M_ℓ criteria and the squared exponential covariance function are used the final estimation of f_1 is obtained with around 100 evaluations of f_1 instead of 10^5. To obtain similar statistical performance with the Matérn covariance more than 750 observations are required. The difference between the two covariance functions probably comes from the behavior of the maximal variance. It is still strongly decreasing after 500 observations for the squared exponential covariance function which is not the case for the Matérn covariance function.

2.4.2. Dolomite precipitation

Similarly to the previous case, the average and the standard deviation of the different statistical measures obtained from 10 replications of the initial set of points are shown in Figure 2.11 for the squared exponential and the Matérn covariance functions defined in (2.3) and (2.7) for $3 \leq t \leq 500$. We obtained similar conclusions as for the calcite precipitation case, see Figure 2.12.

2.5. Conclusion

We have shown that our method has two main features which make it very attractive. Firstly, it is very efficient from a practical point of view thanks to the Gaussian Process modeling which enables us to sequentially build the surrogate model with a low number of points

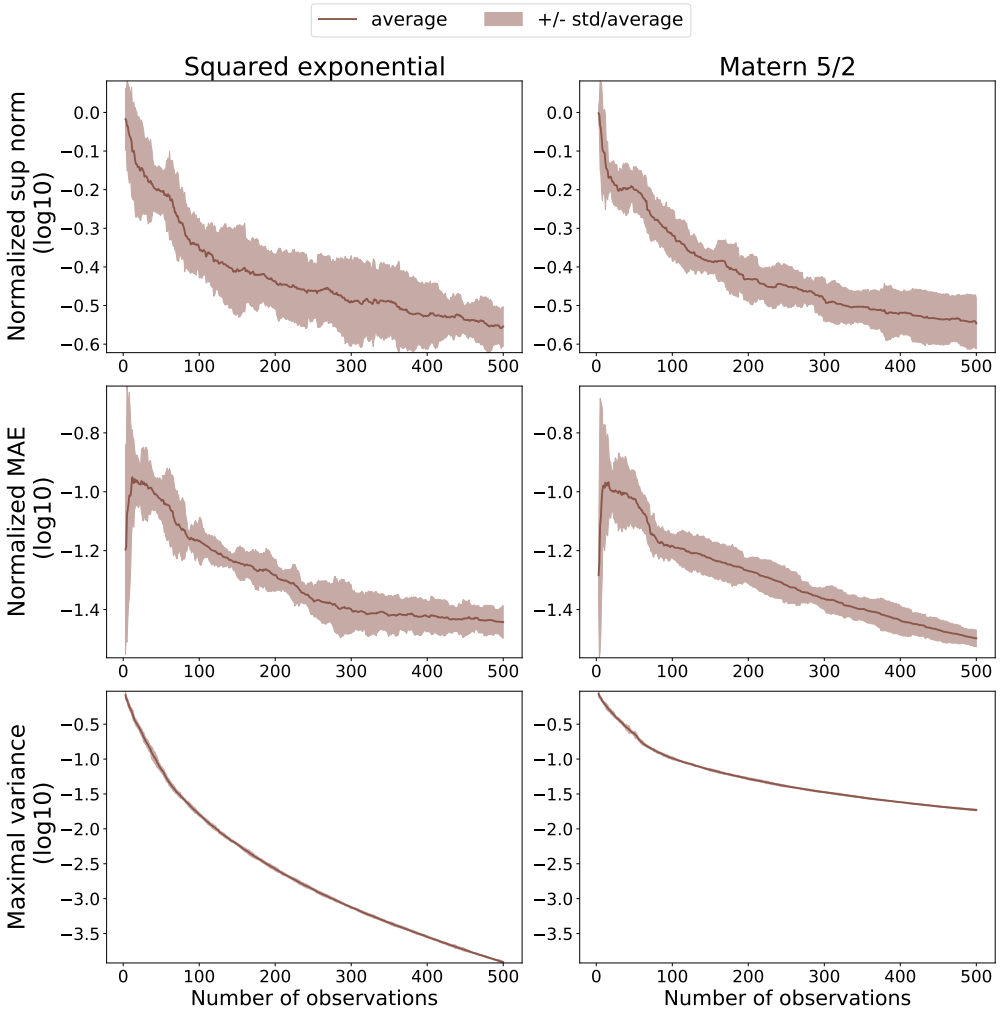

Figure 2.9: Average and standard deviation of different statistical measures for the squared exponential and the Matérn covariance functions defined in (2.3) and (2.7) for the calcite precipitation problem with $d = 6$.

and almost no parameters to tune. Secondly, its very low computational burden makes its use possible on complex chemical reactions involving singular behaviors like precipitation and dissolution of minerals. Our method could also be applied to more complex geochemical systems like surface complexation or ion exchange that can be described with laws of mass action. Effectively, these two features have further potential applications on much larger sets of reactive species or with coupled physical processes namely in reactive transport modeling. This will be the subject of a future work.

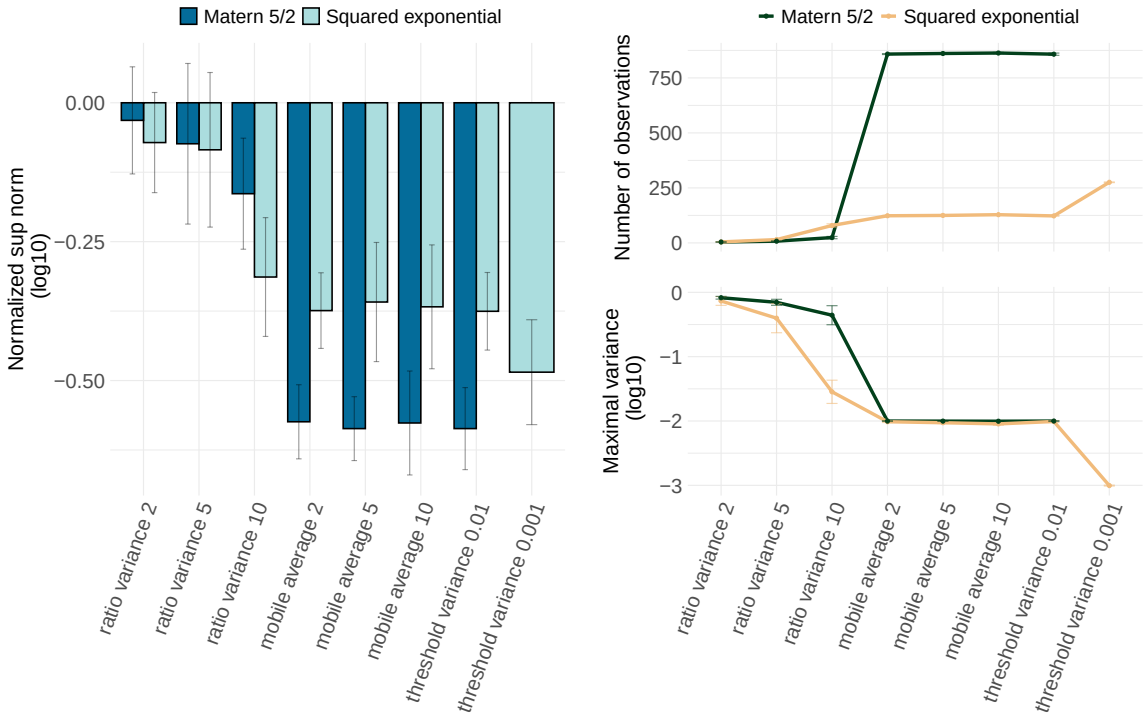

Figure 2.10: Left: Statistical assessment of the error estimation of f_1 defined in (2.21) for the stopping criteria defined in (2.13), (2.15) and (2.17) for the squared exponential and the Matérn covariance function defined in (2.3) and (2.7). Top right: Number of evaluations required for the different considered stopping criteria. Bottom right: Values of $V(t^\star)$ where V is defined in (2.16) and $t^\star$ is the stopping iteration which changes from one stopping criterion to another.

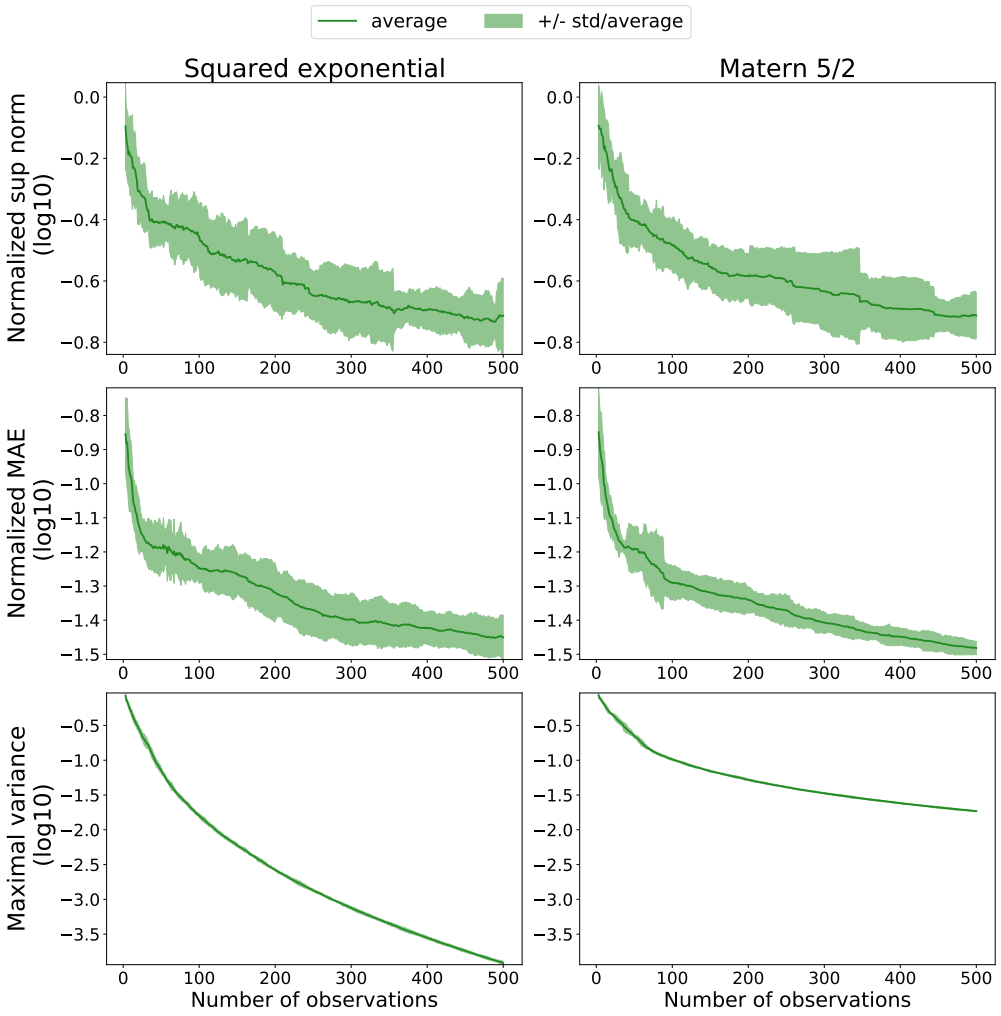

Figure 2.11: Average and standard deviation of different statistical measures for the squared exponential and the Matérn covariance functions defined in (2.3) and (2.7) for the dolomite precipitation problem with $d = 6$.

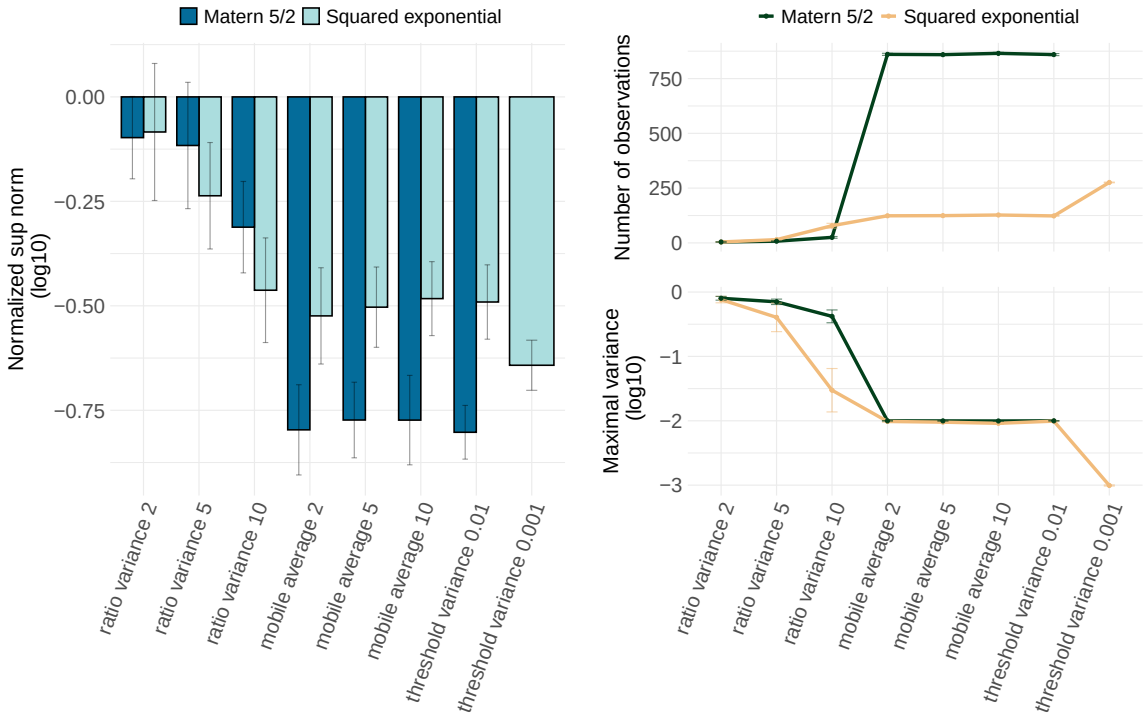

Figure 2.12: Left: Statistical assessment of the error estimation of f_2 defined in (2.21) for the stopping criteria defined in (2.13), (2.15) and (2.17) for the squared exponential and the Matérn covariance functions defined in (2.3) and (2.7). Top right: Number of evaluations required for the different considered stopping criteria. Bottom right: Values of $V(t^\star)$ where V is defined in (2.16) and $t^\star$ is the stopping iteration which changes from one stopping criterion to another.

2.6. Appendix: Additional plots

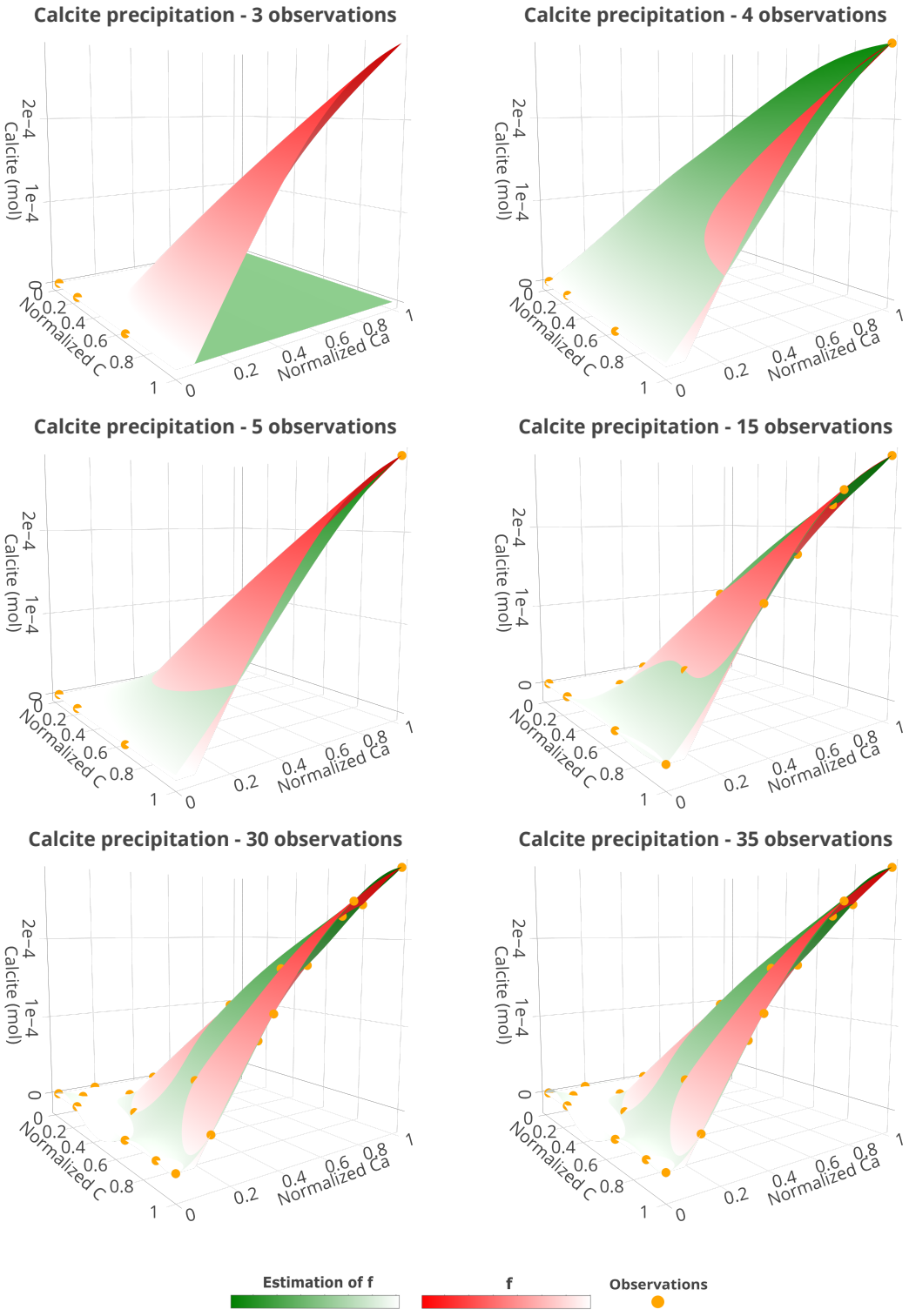

Figure 2.13: Illustration of our active learning approach for estimating the function $f_1(\mathrm{C}, \mathrm{Ca}, 2\times 10^{-3}, 10^{-5}, 10, 0)$ displayed in the left part of Figure 2.8 by starting from $t_1 = 3$ observations randomly chosen in $\mathrm{A} \subset [0,1]^2$. Here, the squared exponential covariance function was used.

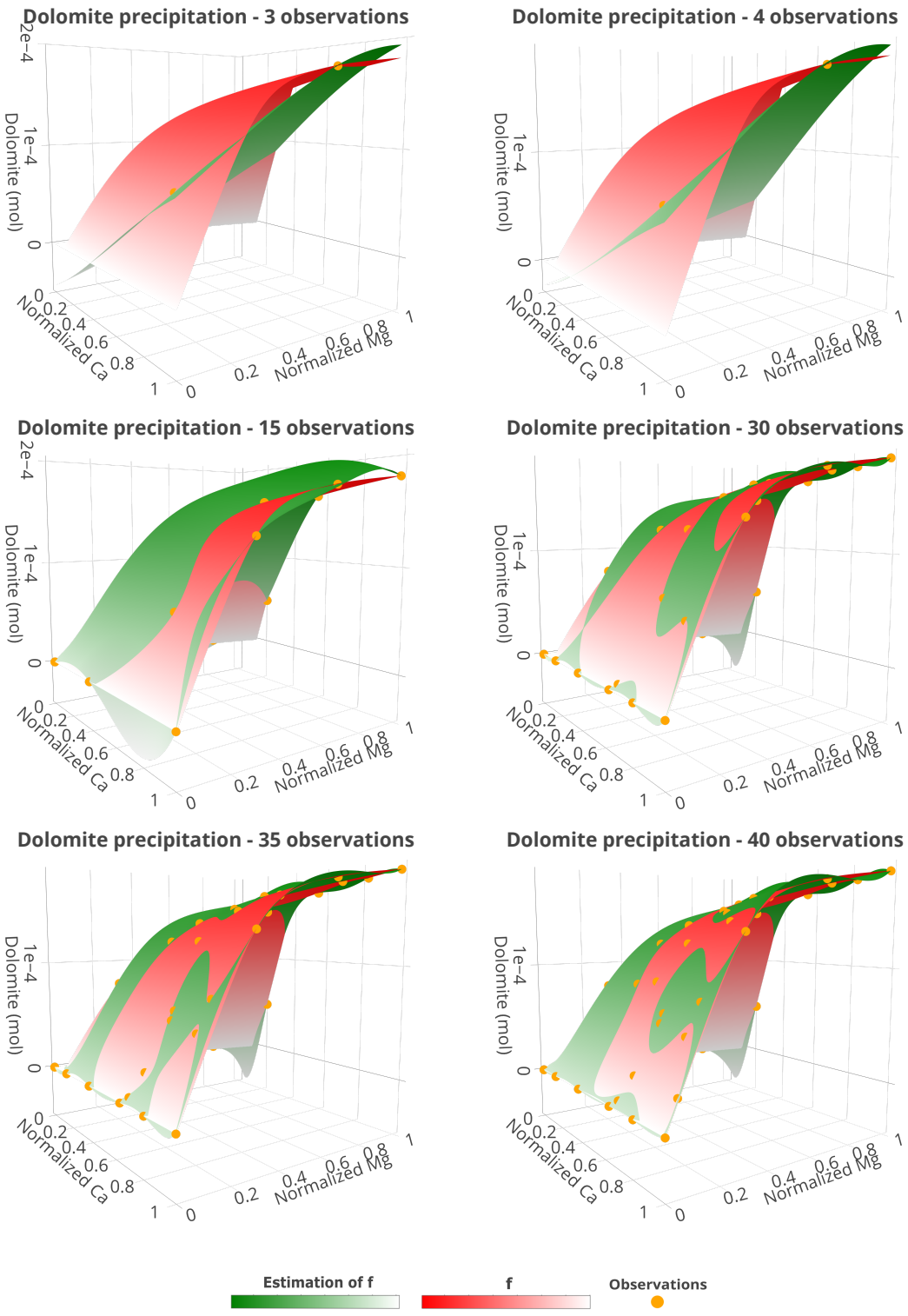

Figure 2.14: Illustration of our active learning approach for estimating the function $f_2(5 \times 10^{-4}, \mathrm{Ca}, 2 \times 10^{-3}, \mathrm{Mg}, 10, 0)$ displayed in the right part of Figure 2.8 by starting from $t_1 = 3$ observations randomly chosen in $\mathrm{A} \subset [0, 1]^2$. Here, the squared exponential covariance function was used.

Chapter 3 - A novel approach for estimating functions based on an adaptive knot selection for B-splines with an application to geoscience

Scientific contribution

The content of this chapter is the subject of the article:
Savino, E. M., Lévy-Leduc, C. 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. Submitted and also available on arXiv preprint *(arXiv:2306.00686)*.

The proposed method is implemented in the `glober` R package available from the CRAN.

Abstract

In this chapter, we will outline a novel data-driven method for estimating functions in a multivariate nonparametric regression model based on an adaptive knot selection for B-splines. The underlying idea of our approach for selecting knots is to apply the generalized lasso, since the knots of the B-spline basis can be seen as changes in the derivatives of the function to be estimated. This method was then extended to functions depending on several variables by processing each dimension independently, thus reducing the problem to a univariate setting. The regularization parameters were chosen by means of a criterion based on EBIC. The nonparametric estimator was obtained using a multivariate B-spline regression with the corresponding selected knots. Our procedure was validated through numerical experiments by varying the number of observations and the level of noise to investigate its robustness. The influence of observation sampling was also assessed and our method was applied to a chemical system commonly used in geoscience. For each different framework considered in this chapter, our approach performed better than state-of-the-art methods. Our completely data-driven method is implemented in the `glober` R package which is available on the Comprehensive R Archive Network (CRAN).

Table of contents

3.1. Introduction

In geochemical models, computing the concentrations of reactive species at equilibrium is well-known to be a challenging task especially when the number of species is large and/or when the reactions involve the dissolution or the precipitation of minerals, see White et al. (1958), Smith (1980) and de Capitani and Brown (1987) for further details. The numerical resolution of these non-linear problems can become so time consuming that coupling them with other physical processes may require to be simplified. For instance in the case of reactive transport, the size of the geometric model has to be drastically limited. To overcome this issue, researchers have been focusing their work on improving the numerical scheme to speed up computations.

However, despite the significant improvements of the numerical solvers and preconditioners over the past few decades, solving three dimensional large scale modelling of complex reactive transport over many time steps is still nearly impossible using standard computers. Consequently, geoscientists are more and more interested in devising approaches which can provide an estimation of the solution of the full simulation model (sometimes also called surrogate model) from a limited set of observations obtained with the full simulation model from specific input values that can thus replace it. Hence, the problem can be reformulated as the estimation of an unknown function f in the following regression model:

$$Y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{3.1}$$

where the ε_i are i.i.d centered random variables of variance σ^2 and the x_i are observation points which belong to a compact set $\mathcal{S}$ of $\mathbb{R}^d$, $d \geq 1$. In the reactive transport modelling field (RTM), several surrogate models have been proposed, we refer the reader to Asher et al. (2015) and Jatnieks et al. (2016) for a comparison of the different approaches. Artificial neural networks have recently gained a huge interest in RTM (Guérillot and Bruyelle (2020)), more especially through Deep Neural Networks (DNNs), since their approximations have a high accuracy compared to other estimators, see Laloy and Jacques (2019). Nevertheless, despite all the effort for improving the efficiency of DNNs via the conjunction of computational advancements for training ever-larger networks and improvements of backpropagation algorithms, DNNs still remain difficult to exploit when the quantity of training data is not sufficient, see Karpatne et al. (2018), especially when a high number of parameters needs to be calibrated. Recently, Savino et al. (2022) proposed an active learning approach to drastically decrease the number of training observations to use by modeling the function to estimate as a sample of Gaussian Processes. This method has given promising results but is not necessarily the most suitable approach for noisy observation sets. In order to circumvent this limitation, nonparametric estimation approaches based on splines are known to be an efficient tool, see Wahba (1990) for further details on this kind of methods.

Nonparametric estimation approaches based on splines consist in approximating the function to estimate by a linear combination of splines which are functions defined by pre-selecting a well chosen set of knots. In this framework, Friedman (1991) proposed an efficient approach called Multivariate Adaptive Regression Splines (MARS) which can be used when the function to estimate has several input variables. However, MARS has not shown better performance than other state-of-the-art methods on a concrete RTM application displayed in Jatnieks et al. (2016). A theorical and experimental comparison has been undertaken by Eckle and Schmidt-Hieber (2019) and demonstrated that DNNs can outperform MARS but with a specific number

of parameters and they do not necessarily give better results for every numerical application. This conclusion was also drawn by Zhang and Goh (2013) and Zhang and Goh (2016) in which the authors demonstrated an equivalent accuracy and performance between a back-propagation neural network architecture and MARS on geotechnical applications but a better interpretability and a higher computational efficiency was demonstrated for the latter.

Other articles proposed approximating the function to estimate by a linear combination of B-splines defined in De Boor (1978) since they display an attractive stability and a computationnal efficiency. Their ability to approximate complicated functions while being unsensitive to noisy observation sets have made them very interesting in the past few decades. Since their definition depends on a pre-defined sequence of knot locations, many strategies have been developed to optimize the selection of these points in order to avoid overfitting and so to ensure the best approximation of the underlying function. O'Sullivan (1986) described an innovative method, introduced as O-splines by Wand and Ormerod (2008), to estimate a function f by selecting simultaneously the number and the locations of knots from an arbitrary set of values. Its main goal was to penalize an integrated square of the second order derivative also called roughness in order to determine the coefficients of the linear combination of B-splines. However, the computation of this method was tedious with higher order derivatives. To circumvent this issue, Eilers and Marx (1996) proposed a discrete version of this method called P-splines which uses a discrete penalty matrix and a ℓ_2-norm penalized least-square criterion (ridge approach) to determine the coefficients of the B-splines defined from evenly-spaced knots (see Wand and Ormerod (2008) for a detailed comparison between O-splines and P-splines). These P-splines have been used in an impressive list of work of curve fitting (see Eilers et al. (2015) for a review) and have been extended to the multivariate setting, see Eilers and Marx (2003) for an application to smoothing two-dimensional signals. Li and Cao (2022) have adapted these P-splines to apply them to unevenly-spaced knots by defining a general weighted difference penalty matrix adapted to regular and irregular knot spacing. To drastically limit the number of knots, Goepp et al. (2018) have proposed a weigthed adaptive ridge method called A-splines which aims at discarding the less relevant knots and by defining new B-splines from the selected knots. This method appeared to be more interpretable than the P-splines method but their statistical performance is equivalent.

Another approach was introduced for B-spline curve fitting by Yuan et al. (2013). Their idea is to first select the most pertinent B-splines from a multi-resolution basis by applying the Lasso criterion to get the locations of the knots. Then, after a pruning step to reduce once again the number of knots, the final B-splines are built from these small sets of knots. The estimation of the function is obtained by fitting a linear combination of these B-splines to the observations by a least-square approach. This method seems to have promising results but is not available for two-dimensional functions yet.

In this chapter, we propose a novel data-driven approach for estimating the function f in the multivariate nonparametric regression model (3.1) based on an adaptive knot selection for B-splines. Since the knots of a B-spline basis can be seen as changes in the derivatives of f, we propose finding the most relevant ones, based on the work of Denis et al. (2020), by using the generalized lasso described in Tibshirani and Taylor (2011) and further studied in Tibshirani (2014). A B-spline basis is then defined from these selected knots and a least-square approach is undertaken to determine the coefficients of the linear combination of B-splines. Sadhanala et al. (2021) have proposed a multivariate version of trend filtering (a specific generalized lasso form) called Kronecker trend filtering (KTF) to extend it to smoothing functions with multiple

input variables. It implies the use of a huge penalty matrix defined as the Kronecker product of univariate trend filtering penalty operators and of a unique regularization parameter common to every input variables. In order to drastically reduce the dimensions of the difference penalty matrix and to allow a better flexibility in the regularization step, the extension of our method to functions with two input variables is presented by simply considering each dimension separately to reduce the problem to the one-dimensional setting. We also propose a way to extend our approach to higher dimensional settings where the observation points do not necessarily come from a Cartesian product of the sets in each dimension.

This chapter is organized as follows. Section 3.2 describes the methodology that we propose for our adaptive knot selection method for the one and two-dimensional settings. Section 3.3 investigates the performance of our approach through numerical experiments. In Section 3.4, we apply our method to the data that motivated this study. Finally, in Section 3.5.3, we extend our approach to more general observation point settings.

3.2. Methodology

In this section, we describe our innovative nonparametric method to estimate the function f defined in (3.1). We will introduce our method first for one-dimensional functions ($d = 1$), then in a second section we will extend it to the two-dimensional case ($d = 2$).

3.2.1. Description of our method in the one-dimensional case

We propose estimating the function f appearing in (3.1) by approximating it with a linear combination of B-splines of order M ($M \geq 1$) introduced by De Boor (1978) in Chapter 9.

Let $\mathbf{t} = (t_1, \ldots, t_K)$ be a set of K points called knots which are crucial in the definition of the B-spline basis. We define the augmented knot sequence $\boldsymbol{\tau}$ such that:

$$\tau_1 = \ldots = \tau_M = x_{min},$$

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

$$x_{max} = \tau_{K+M+1} = \ldots = \tau_{K+2M},$$

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{K+2M}) = (\underbrace{x_{min}, \ldots, x_{min}}_{\text{M times}}, \underbrace{t_1, \ldots, t_K}_{\mathbf{t}}, \underbrace{x_{max}, \ldots, x_{max}}_{\text{M times}}),$$

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

B-splines are defined by (De Boor, 1978, p. 89-90) and (Hastie et al., 2009, p. 160) as follows. Denoting by $B_{i,m}(x)$ the ith B-spline basis function of order m for the knot sequence $\boldsymbol{\tau}$ with $m \leq M$, they are defined by the following recursion:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, \ldots, K + 2M - 1, \tag{3.2}$$

and for $m \leq M$,

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x), \tag{3.3}$$

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

In the next section we will describe how to choose the set of knots $\mathbf{t}$ to estimate f.

3.2.1.1. Creation of a candidate set of knots.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$ where Y_i and x_i are defined in (3.1). In the following, we shall assume that $x_1 < \ldots < x_n$ and $M = q + 1$, with $q \geq 0$. Hence, when $q = 0$ (resp. $q = 1$, $q = 2$) f is approximated with piecewise constant (resp. linear, quadratic) functions.

Since the knots of a B-spline basis can be seen as changes in the $(q+1)$th derivative of f, we propose finding them by using the generalized Lasso described in Tibshirani and Taylor (2011) and further studied in Tibshirani (2014). In the latter, they define the polynomial trend filtering which consists in approximating f by $\widehat{\boldsymbol{\beta}}(\lambda)$ defined as follows:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\operatorname{argmin}} \{ ||\mathbf{Y} - \boldsymbol{\beta}||_2^2 + \lambda ||D\,\boldsymbol{\beta}||_1 \}, \tag{3.4}$$

where $||y||_2^2 = \sum_{i=1}^n y_i^2$ for $y = (y_1, \ldots, y_n)$ and $||u||_1 = \sum_{i=1}^m |u_i|$ for $u = (u_1, \ldots, u_m)$, λ is a positive constant which has to be tuned and $D \in \mathbb{R}^{m \times n}$ is a specified penalty matrix, defined recursively as follows:

$$D = D_{tf,q+1} = D_0 \cdot D_{tf,q} \quad q \geq 0, \tag{3.5}$$

where "tf" is the abbreviation of "trend filtering", $(q+1)$ is the order of differentiation, $D_{tf,0} = \mathrm{Id}_{\mathbb{R}^n}$, the identity matrix of $\mathbb{R}^n$, and D_0 is the penalty matrix for the one-dimensional fused Lasso:

$$D_0 = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

The penalty matrix D is the discrete difference operator of order $(q+1)$ and thus, $D\widehat{\boldsymbol{\beta}}$ estimates the $(q+1)$st order derivative of f. Hence, observing the locations where $D\widehat{\boldsymbol{\beta}} \neq 0$ provides a way of finding the B-spline knots.

The matrix D is well-adapted when the observation points are evenly spaced. When it is not the case, it should be replaced by the following matrix $\Delta^{(q+1)}$ defined recursively as follows:

$$\Delta^{(q+1)} = \mathbf{W}_{(q+1)} \cdot D_0 \cdot \Delta^{(q)}, \quad q \geq 0,$$

where $\Delta^{(0)} = \mathrm{Id}_{\mathbb{R}^n}$ and $\mathbf{W}_{(q+1)}$ is the diagonal weight matrix defined by:

$$\mathbf{W}_{(q+1)} = \operatorname{diag}\left(\frac{1}{(x_{(q+1)+1} - x_{(q+1)})}, \frac{1}{(x_{(q+1)+2} - x_{(q+1)+1})}, \ldots, \frac{1}{(x_n - x_{n-1})} \right).$$

In both cases (evenly or unevenly-spaced observations), the number of rows of D and $\Delta^{(q+1)}$ equals $m = n - q - 1$.

Let us now more precisely explain how to choose the B-spline knots. Let $\Lambda = (\lambda_1, \ldots, \lambda_k)$ be a grid of penalization parameters λ_i. We define the resulting differentiated column vector $\mathbf{a}(\lambda)$ by:

$$\mathbf{a}(\lambda) = \Delta^{(q+1)} \cdot \widehat{\boldsymbol{\beta}}(\lambda), \tag{3.6}$$

where $\widehat{\boldsymbol{\beta}}(\lambda)$ is the solution of problem (3.4) when $D = \Delta^{(q+1)}$ and λ belongs to Λ.

The ordered vector of selected knots associated to λ is defined as follows:

$$\widehat{\mathbf{t}}_\lambda = \left(\widehat{t}_j\right)_{j=1,\ldots,K_\lambda} = \left(x_{p_j}\right)_{j=1,\ldots,K_\lambda}, \quad \text{with } p_j \in \mathcal{P}_\lambda, \tag{3.7}$$

where

$$\mathcal{P}_\lambda = \{\ell + 1,\ a_\ell(\lambda) \neq 0\ \} \quad \text{and} \quad K_\lambda = \sum_{\ell=1}^{m} \mathbb{1}\{a_\ell(\lambda) \neq 0\}, \tag{3.8}$$

$a_\ell(\lambda)$ denoting the ℓth component of $\mathbf{a}(\lambda)$ and $\mathbb{1}\{A\} = 1$ if the event A holds and 0 if not.

The corresponding B-spline basis $B_{i,M}$ is defined by replacing the t_j in the augmented knot sequence $\boldsymbol{\tau}$ appearing in (3.2) and (3.3) by $\widehat{t_j}$ found in (3.7). Thus, we obtain the following estimator of f for each λ of Λ:

$$\widehat{f_\lambda}(x) = \sum_{i=1}^{q+K_\lambda+1} \widehat{\gamma_i} B_{i,M}(x), \tag{3.9}$$

where $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma_i})_{1\leq i\leq q+K_\lambda+1}$ is obtained using the following least-square criterion:

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{q+K_\lambda+1}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{B}(\lambda)\,\boldsymbol{\gamma}\,\|_2^2\,, \tag{3.10}$$

where $\mathbf{B}(\lambda)$ is a $n \times (q + K_\lambda + 1)$ matrix having as ith column $(B_{i,M}(x_k))_{1\leq k\leq n}$, i belonging to $\{1, \ldots, q + K_\lambda + 1\}$.

3.2.1.2. Choice of the penalization parameter of the regularized method.

In order to choose the penalization parameter λ which leads to the best selection of knots, we use a criterion defined by Chen and Chen (2008) and recommended in Goepp et al. (2018), namely the extended Bayesian information criterion also called EBIC:

$$\mathsf{EBIC}(\lambda) = \mathsf{SS}(\lambda) + (q + K_\lambda + 1)\log n + 2\log\binom{q + K_{\mathsf{max}} + 1}{q + K_\lambda + 1}, \tag{3.11}$$

where K_{max} is the maximum number of knots that we can select (here $K_{\mathsf{max}} = n$) and $\mathsf{SS}(\lambda)$ is the sum of squares defined by:

$$\mathsf{SS}(\lambda) = \|\mathbf{Y} - \widehat{\mathbf{Y}}(\lambda)\|_2^2\,, \tag{3.12}$$

where

$$\widehat{\mathbf{Y}}(\lambda) = \mathbf{B}(\lambda)\widehat{\boldsymbol{\gamma}},$$

with $\widehat{\boldsymbol{\gamma}}$ and $\mathbf{B}(\lambda)$ being defined in (3.10). This criterion allows us to get a trade-off between a good approximation of the underlying function without using too many parameters. The final estimator of f is defined as follows:

$$\widehat{f}(x) = \widehat{f_{\lambda_{\mathsf{EBIC}}}}(x), \tag{3.13}$$

where $\widehat{f_\lambda}(x)$ is defined in (3.9) and

$$\lambda_{\mathsf{EBIC}} = \underset{\lambda\in\Lambda}{\operatorname{argmin}}\{\mathsf{EBIC}(\lambda)\}. \tag{3.14}$$

3.2.1.3. Illustration of our method on a simple case.

In order to illustrate our method we apply it to a noisy set of observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$ where the Y_i are defined in (3.1) and $f = f_1$ is a linear combination of quadratic B-splines ($M = 3$) with $\mathbf{t} = (0.1, 0.27, 0.745)$ defined as follows:

$$f_1(x) = -2.5B_{2,3}(x) + 4.3B_{5,3}(x), \quad x \in [0, 1]. \tag{3.15}$$

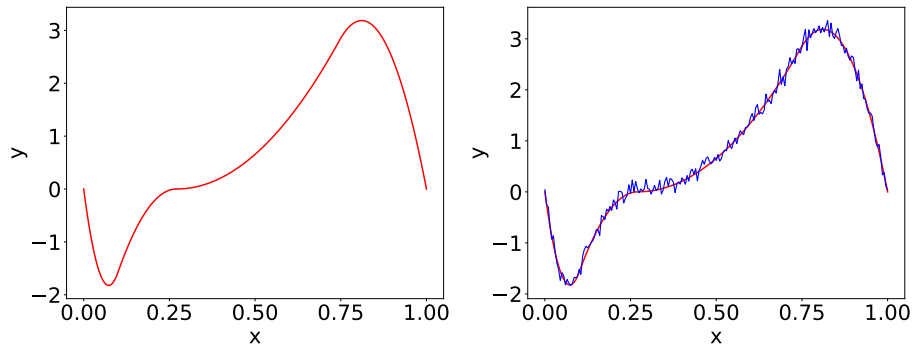

Figure 3.1: Function f_1 to estimate (left) and a noisy set of observations $Y_1, \ldots, Y_{201}$ with $\sigma = 0.1$ (right).

In (3.1), the ε_i are i.i.d Gaussian centered random variables with $\sigma = 0.1$. The set of knots $\mathbf{t}$ belongs to the observation set $\{x_1, \ldots, x_n\}$. The corresponding $(f_1(x_i))_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$ are displayed in Figure 3.1 for $n = 201$. Since we want to approximate quadratic B-splines, we must choose q such that the method can detect the changes in the third derivative so here $q + 1 = 3$. In order to assess the performance of our knot selection procedure, we compute the Hausdorff distance defined as follows:

$$d(\mathbf{t}, \widehat{\mathbf{t}}_\lambda) = \max\left(d_1(\mathbf{t}, \widehat{\mathbf{t}}_\lambda),\ d_2(\mathbf{t}, \widehat{\mathbf{t}}_\lambda)\right), \tag{3.16}$$

where

$$d_1(\mathbf{u}, \mathbf{v}) = \sup_{v\in\mathbf{v}} \inf_{u\in\mathbf{u}} |u - v|,$$

$$d_2(\mathbf{u}, \mathbf{v}) = d_1(\mathbf{v}, \mathbf{u}).$$

Figure 3.2 displays the boxplots of the first and second part of the Hausdorff distance and of the number of selected knots K_λ for $\lambda = \lambda_{\text{EBIC}}$ obtained from 10 different samplings of $x_1, \ldots, x_n$. The first boxplots are obtained for $n = 15$ then new observation points are randomly added to the current observation sets in order to have an increasing number of observations such that $n \leq 100$. We can see from this figure that from $n = 70$ the second part of the Hausdorff distance is close to 0 which means that the estimated knots are near from the real ones. These results are obtained with an almost constant number of selected knots $K_{\lambda_{\text{EBIC}}} = 6$.

For a comparison purpose, we displayed in Figure 3.3 the results obtained for $\lambda = \lambda_{\text{opt}}$, where λ_{opt} is defined by:

$$\lambda_{\text{opt}} = \underset{\lambda\in\Lambda}{\text{argmin}}\left(\text{Normalized sup norm}(\lambda)\right),$$

with

$$\text{Normalized sup norm}(\lambda) = \max_{1\leq k\leq N} \frac{\left|f(x_k) - \widehat{f}_\lambda(x_k)\right|}{f_{max} - f_{min}}, \tag{3.17}$$

where $\widehat{f}_\lambda$ is defined in (3.9). In (3.17), N ($N > n$) is the cardinality of the set of evenly-spaced points $\{x_1, \ldots, x_N\}$ of $[0, 1]$ which contains the observation points $x_1, \ldots, x_n$ as well as additional points where f has not been observed. Moreover, f_{min} and f_{max} denote the minimum

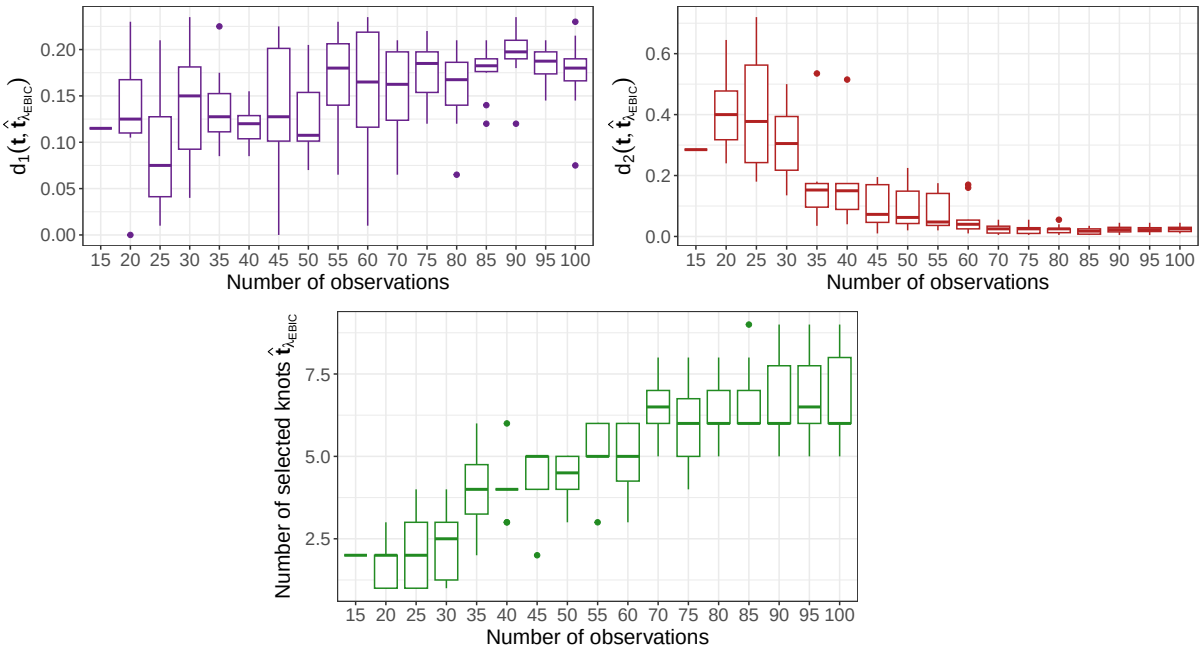

Figure 3.2: Top left: Boxplots for the first part of the Hausdorff distance as a function of n. Top right: boxplots for the second part of the Hausdorff distance as a function of n. Bottom: number of estimated knots as a function of n with $\lambda = \lambda_{\text{EBIC}}$ for estimating f_1.

and maximum values of f evaluated on $\{x_1, \ldots, x_N\}$, respectively. We can see from this figure that the performance obtained when λ is optimally chosen is on a par with that of λ_{EBIC} which means that our procedure for choosing λ is almost optimal.

The corresponding performance is shown on the right part of Figure 3.4 for $N = 201$. This figure displays the average of the most stringent metric (Normalized Sup Norm) obtained from 10 different samplings of $x_1, \ldots, x_n$ for each n. We observed from this figure that the Normalized Sup Norm reaches $10^{-1.75}$ (resp. $10^{-1.5}$) for λ_{opt} (resp. λ_{EBIC}) which represents a normalized maximum absolute error of 2% (resp. 3%). Once again, these results show that the choice of λ does not alter the performance of our approach.

3.2.2. Extension to the two-dimensional case.

In this section, we will extend the previous method for estimating a two-dimensional function f from the observations $(Y_i)_{1 \leq i \leq n}$ defined in Model (3.1) when $d = 2$.

Here $\mathcal{S}$ is defined as the Cartesian product of two compact sets $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathbb{R}$. More precisely, we will consider $(x_{11}, \ldots, x_{1n_1})$ belonging to $\mathcal{S}_1$ and similarly $(x_{21}, \ldots, x_{2n_2})$ belonging to $\mathcal{S}_2$ the n_1 and n_2 observation values for the first and second variables of f at which f is evaluated, respectively. Thus, the set of observations belonging to $\mathcal{S}$ will be defined as:

$$\big((x_{11}, x_{21}), (x_{11}, x_{22}), \ldots, (x_{11}, x_{2n_2}), (x_{12}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{12}, x_{2n_2}), \ldots, (x_{1n_1}, x_{2n_2})\big)$$
$$= \big((x_{1k}, x_{2\ell})\big)_{1 \leq k \leq n_1, 1 \leq \ell \leq n_2}.$$

For estimating f, we shall approximate it by a linear combination of multidimensional B-splines defined in Hastie et al. (2009, p. 162-163) as the tensor product of the B-splines of order

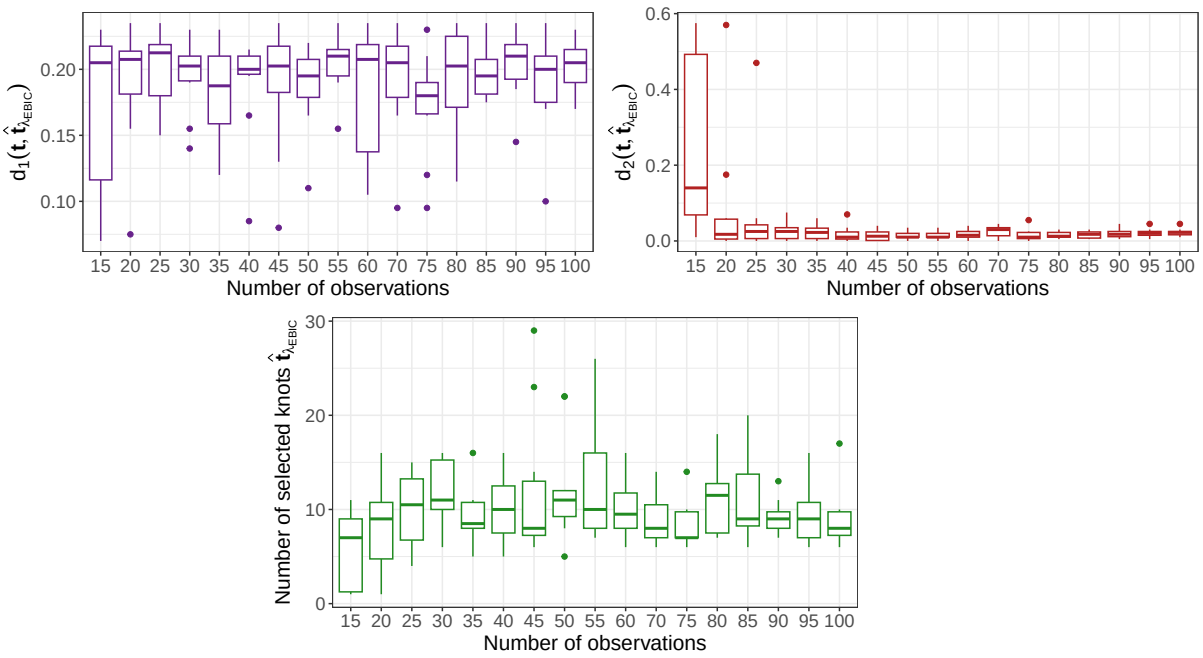

Figure 3.3: Similar to Figure 3.2 by choosing $\lambda = \lambda_{\text{opt}}$ for estimating f_1.

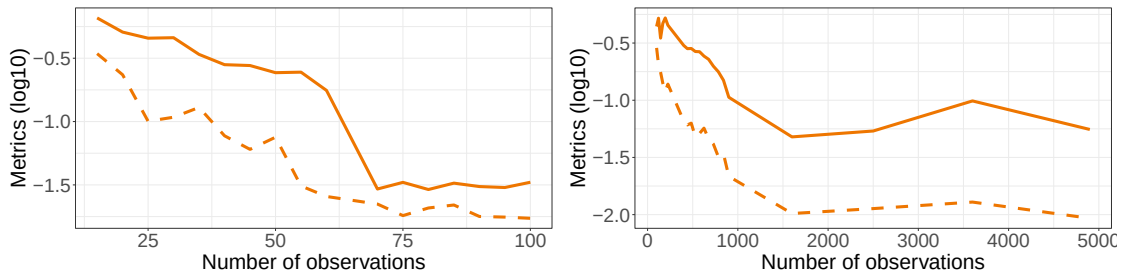

Figure 3.4: Statistical performance (Normalized Sup Norm) of the method using λ_{opt} (dashed) and λ_{EBIC} (solid) for the estimation of f_1 (right) and f_2 (left) obtained from 10 replications.

M introduced in Section 3.2.1. More precisely, $f(x) = f(x_1, x_2)$ will be approximated by:

$$\sum_{i=1}^{Q_1} \sum_{j=1}^{Q_2} \gamma_{ij} \, B_{1,i,M}(x_1) B_{2,j,M}(x_2), \tag{3.18}$$

where $B_{1,i,M}$ and $B_{2,j,M}$ are the B-spline basis of order M defined in (3.3) for the first and second dimension, respectively. In (3.18), $Q_1 = q + K_1 + 1$, $Q_2 = q + K_2 + 1$ with K_1 and K_2 the number of knots defined in the B-spline basis of the first and second variables, respectively and $M = q + 1$.

3.2.2.1. Creation of a candidate set of knots.

The idea is to consider the two dimensions independently and thus, by fixing one dimension at a time, the problem can be rewritten as an estimation problem in the one-dimensional framework.

First, we shall consider the knot selection of the B-spline basis of the first dimension by fixing the second dimension to a certain value of x_2 belonging to $\{x_{21}, \ldots, x_{2n_2}\}$. Thus, we can apply the polynomial trend filtering method described in Section 3.2.1 and get the grid of penalization parameters $\left(\boldsymbol{\lambda}_{(1,i),k}\right)_{1\leq k\leq s_i}$, with i belonging to $\{1, \ldots, n_2\}$ and s_i corresponds to the number of penalization parameters. The index 1 in $(1, i)$ denotes the first dimension and i indexes the ith value of $\{x_{21}, \ldots, x_{2n_2}\}$. For each value $\boldsymbol{\lambda}_{(1,i),k}$ with k belonging to $\{1, \ldots, s_i\}$, we can get the corresponding selected knots by following the procedure described in Section 3.2.1.2: after calculating $\mathbf{a}(\boldsymbol{\lambda}_{(1,i),k})$ as in (3.6), we can determine the set of knots $\widehat{\mathbf{t}}_{1,\boldsymbol{\lambda}_{(1,i),k}}$ as in (3.7). In order to take into account all the information obtained for each value of x_2, we gather the selected knots into a single vector depending on the value of the penalization parameters. Nevertheless, because not all the vectors $\left(\boldsymbol{\lambda}_{(1,i),k}\right)_{1\leq k\leq s_i}$ have exactly the same values $\boldsymbol{\lambda}_{(1,i),k}$ and the same number of penalization parameters s_i when i varies, we shall define the set of *equivalent regularization parameters* $\widetilde{\Lambda}_1$ and the minimal number of penalization parameters s_{min_1}:

$$\widetilde{\Lambda}_1 = \left\{\widetilde{\boldsymbol{\lambda}}_{1,1}, \ldots, \widetilde{\boldsymbol{\lambda}}_{1,s_{min_1}}\right\} \quad \text{and} \quad s_{min_1} = \min_{1\leq i\leq n_2} s_i \,, \tag{3.19}$$

where

$$\widetilde{\boldsymbol{\lambda}}_{1,k} = \left(\boldsymbol{\lambda}_{(1,i),k}\right)_{1\leq i\leq n_2}, \quad 1 \leq k \leq s_{min_1}. \tag{3.20}$$

In (3.20), $\widetilde{\boldsymbol{\lambda}}_{1,k}$ can be seen as the vector of parameters which penalize (3.4) at an equivalent strength for each fixed value of x_2. We can therefore get the vector of selected knots $\widehat{\mathbf{t}}_{1,\widetilde{\boldsymbol{\lambda}}_{1,k}}$ for the first dimension by grouping together and ordering all the corresponding selected knots of $\widetilde{\boldsymbol{\lambda}}_{1,k}$.

We proceed the same way to get the set of equivalent parameters $\widetilde{\Lambda}_2$ for the second dimension by fixing this time the value of x_1 and with s_{min_2} defined similarly as in (3.19) for i belonging to $\{1, \ldots, n_1\}$. Analogously as in (3.19) and (3.20), we have:

$$\widetilde{\Lambda}_2 = \left\{\widetilde{\boldsymbol{\lambda}}_{2,1}, \ldots, \widetilde{\boldsymbol{\lambda}}_{2,s_{min_2}}\right\} \quad \text{and} \quad \widetilde{\boldsymbol{\lambda}}_{2,\ell} = \left(\boldsymbol{\lambda}_{(2,i),\ell}\right)_{1\leq i\leq n_1}, \quad 1 \leq \ell \leq s_{min_2}.$$

Moreover, as well as for the first dimension, the vector of selected knots for the second dimension for each $\widetilde{\boldsymbol{\lambda}}_{2,\ell}$ is defined as $\widehat{\mathbf{t}}_{2,\widetilde{\boldsymbol{\lambda}}_{2,\ell}}$, ℓ belonging to $\{1, \ldots, s_{min_2}\}$.

In the following, let us consider two generic penalization parameters $\widetilde{\boldsymbol{\lambda}}_1$ belonging to $\widetilde{\Lambda}_1$ and $\widetilde{\boldsymbol{\lambda}}_2$ belonging to $\widetilde{\Lambda}_2$. Thus, we can define the candidate sets of knots for both dimensions $\widehat{\mathbf{t}}_{1,\widetilde{\boldsymbol{\lambda}}_1}$ and $\widehat{\mathbf{t}}_{2,\widetilde{\boldsymbol{\lambda}}_2}$. We must now determine which combination of penalization parameters $\widetilde{\boldsymbol{\lambda}}_1$ and $\widetilde{\boldsymbol{\lambda}}_2$ and hence, which combination of selected knots for the first and second dimension, allows us to get an optimal estimator of f.

3.2.2.2. Choice of the penalization parameters of the regularized method.

In order to choose the penalization parameters leading to the best selection of knots, we consider the following EBIC criterion which can be seen as the adaptation to the two-dimensional case of the one defined in (3.11):

$$\text{EBIC}\left(\widetilde{\boldsymbol{\lambda}}_1, \widetilde{\boldsymbol{\lambda}}_2\right) = \text{SS}\left(\widetilde{\boldsymbol{\lambda}}_1, \widetilde{\boldsymbol{\lambda}}_2\right) + \widetilde{Q}_1\widetilde{Q}_2 \log n + 2\log\binom{(q+n_1+1)(q+n_2+1)}{\widetilde{Q}_1\widetilde{Q}_2}. \tag{3.21}$$

where $\widetilde{Q}_1 = q + K_{\widetilde{\boldsymbol{\lambda}}_1} + 1$ and $\widetilde{Q}_2 = q + K_{\widetilde{\boldsymbol{\lambda}}_2} + 1$, $K_{\widetilde{\boldsymbol{\lambda}}_1}$ and $K_{\widetilde{\boldsymbol{\lambda}}_2}$ being the number of selected knots with the parameters $\widetilde{\boldsymbol{\lambda}}_1$ and $\widetilde{\boldsymbol{\lambda}}_2$ for the first and second dimension, respectively.

In (3.21), SS $\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right)$ is defined as:

$$\text{SS}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right) = \left\| \mathbf{Y} - \widehat{\mathbf{Y}}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right) \right\|_2^2,$$

where

$$\widehat{\mathbf{Y}}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right) = \mathbf{B}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right)\widehat{\gamma}, \tag{3.22}$$

and $\mathbf{B}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right)$ is defined as:

$$\mathbf{B}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right) = \mathbf{B}\left(\widetilde{\lambda}_1\right) \otimes \mathbf{B}\left(\widetilde{\lambda}_2\right), \tag{3.23}$$

$E \otimes F$ denoting the Kronecker product of the matrices E and F. In (3.23), $\mathbf{B}\left(\widetilde{\lambda}_1\right)$ is a $n_1 \times \widetilde{Q}_1$ matrix having as ith column $(B_{1,i,M}(x_{1k}))_{1\leq k\leq n_1}$, i belonging to $\{1, \ldots, \widetilde{Q}_1\}$ and $\mathbf{B}\left(\widetilde{\lambda}_2\right)$ is a $n_2 \times \widetilde{Q}_2$ matrix having as jth column $(B_{2,j,M}(x_{2\ell}))_{1\leq \ell\leq n_2}$, j belonging to $\{1, \ldots, \widetilde{Q}_2\}$.

In (3.22), $\widehat{\gamma} = (\widehat{\gamma}_{ij})_{1\leq i\leq \widetilde{Q}_1, 1\leq j\leq \widetilde{Q}_2}$ is obtained using the following least-square criterion:

$$\widehat{\gamma} = \underset{\gamma\in\mathbb{R}^{\widetilde{Q}_1\widetilde{Q}_2}}{\operatorname{argmin}} \left\| \mathbf{Y} - \mathbf{B}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right)\gamma \right\|_2^2. \tag{3.24}$$

As in Equation (3.14), we shall define $\widetilde{\lambda}_{1,\text{EBIC}}$ and $\widetilde{\lambda}_{2,\text{EBIC}}$ the penalization parameters which verify:

$$\left(\widetilde{\lambda}_{1,\text{EBIC}}, \widetilde{\lambda}_{2,\text{EBIC}}\right) = \underset{\widetilde{\lambda}_1\in\widetilde{\Lambda}_1, \widetilde{\lambda}_2\in\widetilde{\Lambda}_2}{\operatorname{argmin}} \left\{\text{EBIC}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right)\right\}.$$

Hence, the final estimator of f is defined as:

$$\widehat{f}(x_1, x_2) = \widehat{f}_{\widetilde{\lambda}_{1,\text{EBIC}}, \widetilde{\lambda}_{2,\text{EBIC}}}(x_1, x_2),$$

with $\widehat{f}_{\widetilde{\lambda}_1, \widetilde{\lambda}_2}$ defined as:

$$\widehat{f}_{\widetilde{\lambda}_1, \widetilde{\lambda}_2}(x) = \widehat{f}_{\widetilde{\lambda}_1, \widetilde{\lambda}_2}(x_1, x_2) = \sum_{i=1}^{\widetilde{Q}_1}\sum_{j=1}^{\widetilde{Q}_2} \widehat{\gamma}_{ij}\, B_{1,i,M}(x_1)B_{2,j,M}(x_2). \tag{3.25}$$

In (3.25), $\widehat{\gamma} = (\widehat{\gamma}_{ij})_{1\leq i\leq \widetilde{Q}_1, 1\leq j\leq \widetilde{Q}_2}$ is obtained as in (3.24).

3.2.2.3. Illustration of our method on a simple case.

To illustrate the extension of our method to the two-dimensional case, we propose estimating a function $f = f_2$ which is a linear combination of tensor product of quadratic B-splines ($M = 3$) with $\mathbf{t}_1 = (0.24, 0.545)$ and $\mathbf{t}_2 = (0.395, 0.645)$:

$$f_2(x_1, x_2) = 2.3B_{1,3,3}(x_1)B_{2,3,3}(x_2) - 1.5B_{1,4,3}(x_1)B_{2,5,3}(x_2), \quad (x_1, x_2) \in [0, 1]^2, \tag{3.26}$$

where $B_{i,j,M}$ is defined in (3.18) with $\mathbf{t}_1$ and $\mathbf{t}_2$ the knots involved in the definition of $B_{1,j,M}$ and $B_{2,j,M}$, respectively. We shall apply our method to a noisy set of observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$ where the Y_i are defined in (3.1) and the ε_i are i.i.d Gaussian centered random variables with $\sigma = 0.01$. The set of knots $\mathbf{t}_1$ and $\mathbf{t}_2$ are a part of the observation set $\{x_{11}, \ldots, x_{1n_1}\}$ and $\{x_{21}, \ldots, x_{2n_2}\}$, respectively. The corresponding $(Y_i)_{1\leq i\leq n}$ (resp.

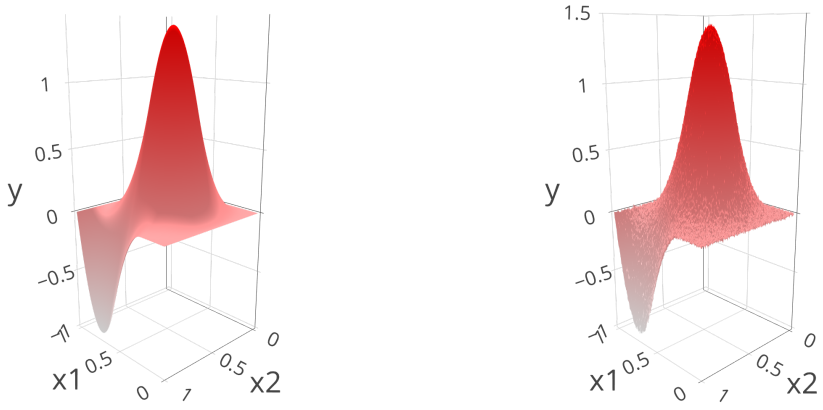

Figure 3.5: Function f_2 to estimate (left) and a noisy set of observations $Y_1, \ldots, Y_{40401}$ with $\sigma = 0.01$ (right).

$(f_2(x_{1,k}, x_{2,\ell}))_{1\leq k\leq n_1, 1\leq \ell\leq n_2}$) are displayed in the right (resp. left) part of Figure 3.5 for $n = n_1 n_2 = 201^2 = 40401$.

In order to assess the performance of our knot selection procedure, we shall use the Hausdorff distance defined in (3.16) for each dimension independently. The results for the first and second part of the Hausdorff distance and the number of selected knots for both dimensions are displayed in the boxplots of Figure 3.18 of the Appendix for $\widetilde{\lambda}_1 = \widetilde{\lambda}_{1,\text{EBIC}}$ and $\widetilde{\lambda}_2 = \widetilde{\lambda}_{2,\text{EBIC}}$ and from 10 different samplings of $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$. New observation points are then randomly added to the current observation sets in order to have an increasing number of observations. We can see from this figure that from $n = 1600$ and so from 40 observation points by dimension, the second part of the Hausdorff distance is close to 0 which means that the estimated knots are near from the real ones. The numbers of selected knots required to get these results are between 5 and 10. Similarly as for the one-dimensional case, we compare these results with those obtained for $\widetilde{\lambda}_1 = \widetilde{\lambda}_{1,\text{opt}}$ and $\widetilde{\lambda}_2 = \widetilde{\lambda}_{2,\text{opt}}$, two optimal parameters defined as:

$$\left(\widetilde{\lambda}_{1,\text{opt}}, \widetilde{\lambda}_{2,\text{opt}}\right) = \underset{\widetilde{\lambda}_1\in\widetilde{\Lambda}_1, \widetilde{\lambda}_2\in\widetilde{\Lambda}_2}{\operatorname{argmin}} \left\{\text{Normalized sup norm } \left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right)\right\}$$

with Normalized sup norm being defined in (3.17) depending here on the values of $\widetilde{\lambda}_1$ and $\widetilde{\lambda}_2$, with $\widehat{f}_{\lambda}$ becoming $\widehat{f}_{\widetilde{\lambda}_1,\widetilde{\lambda}_2}$ defined in (3.25) and x_k belonging to the set:

$$\{x_1, \ldots, x_N\} = \left\{(x_{11}, x_{21}), (x_{11}, x_{22}), \ldots, (x_{11}, x_{2N_2}), (x_{12}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{12}, x_{2N_2}), \ldots, (x_{1N_1}, x_{2N_2})\right\}. \tag{3.27}$$

N is the cardinality of $\{x_1, \ldots, x_N\}$ and is such that $N = N_1 N_2$. We can see from Figure 3.19 of the Appendix, where the results are displayed, that they are comparable to those found for $\widetilde{\lambda}_{1,\text{EBIC}}$ and $\widetilde{\lambda}_{2,\text{EBIC}}$. This means that our choice of the penalization parameters does not alter the performance of our approach. The corresponding performance is shown on the right part of Figure 3.4 for $N = 40401$ and from 10 different samplings of $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$.

The most stringent metric (Normalized Sup Norm) reaches 10^{-2} (resp. $10^{-1.4}$) for $\lambda_{1,\text{opt}}$ and $\lambda_{2,\text{opt}}$ (resp. $\lambda_{1,\text{EBIC}}$ and $\lambda_{2,\text{EBIC}}$) which represents a normalized maximum absolute error of 1% (resp. 4%). Once again, these results show that the choice of $\widetilde{\lambda}_1$ and $\widetilde{\lambda}_2$ does not alter the performance of our approach. We can see for both penalization parameters, the optimal and the ones from the EBIC criterion, that the performance reaches a plateau from $n = 1600$ for $K_{\widetilde{\lambda}_{1,\text{opt}}} = K_{\widetilde{\lambda}_{1,\text{EBIC}}} = 6$ and $K_{\widetilde{\lambda}_{2,\text{opt}}} = K_{\widetilde{\lambda}_{2,\text{EBIC}}} = 9$, which is on a par with what has been found for the number of selected knots for the one-dimensional case in Figures 3.2 and 3.3 for $n \geq 40$.

3.3. Numerical experiments

In this section, we will study the behavior of our method using the EBIC criterion called GLOBER for Generalized LassO for knot selection in multivariate B-splinE Regression and implemented in the `glober` R package when the variance of the noise σ^2 increases and when the observation set changes.

To assess the efficiency of our method, we will compare it to state-of-the-art approaches: Gaussian Processes (GP) described in Rasmussen and Williams (2006) and implemented in the Python package `scikit-learn`, Multivariate Adaptive Spline Regression (MARS) introduced in Friedman (1991) and implemented in the R package `earth` and Deep Neural Networks (DNN) implemented in the R package `keras`.

For the GP, we chose the squared exponential covariance function as defined in Savino et al. (2022). For the MARS approach, we used the default settings proposed in the `earth` package. It has to be noticed that for the two-dimensional case the interaction terms are included in the model in order not to penalize it. The architecture of the DNN was chosen arbitrarily since our goal is not to optimize it in this chapter. More precisely, we used a 2-hidden-layered structure composed of 10 neurons per layer. The activation function of the hidden layers was the RELU function since it is one of the most used functions. In order to train this DNN, we used the stochastic gradient descent method Adam as the optimizer and the Mean Squared Error (MSE) as the loss function. According to the analysis of loss function curves during a preprocessing step, we trained our DNN over 300 epochs for functions of $d = 1$ and 50 epochs for functions of $d = 2$ to avoid overfitting.

3.3.1. Influence of σ on the statistical performance of the method

We first investigate the influence of the level of noise on the performance of GLOBER. To do so, we applied our method to observations corrupted with two different levels of noise and we computed the average Normalized Sup Norm in (3.17), for 10 different samplings of the observations. In both cases ($d = 1$ or 2), the set of knots used to define the underlying function to estimate belongs to the observation set.

3.3.1.1. One-dimensional case ($d = 1$)

We first study the estimation of the function f_1 defined in (3.15) from a noisy set of observations. The corresponding $(Y_i)_{1\leq i\leq n}$ for $\sigma = 0.05$ (resp. $\sigma = 0.25$) and $n = 201$ are displayed in the left (resp. right) part of Figure 3.6.

The two parts of the Hausdorff distance between the real knots and the estimated ones as well as the number of selected knots obtained for the noisiest observation set ($\sigma = 0.25$) are displayed in Figure 3.7. We can see from these results that even the highest value of σ ($\sigma = 0.25$) does not alter the second part of the Hausdorff distance d_2 and that the number of

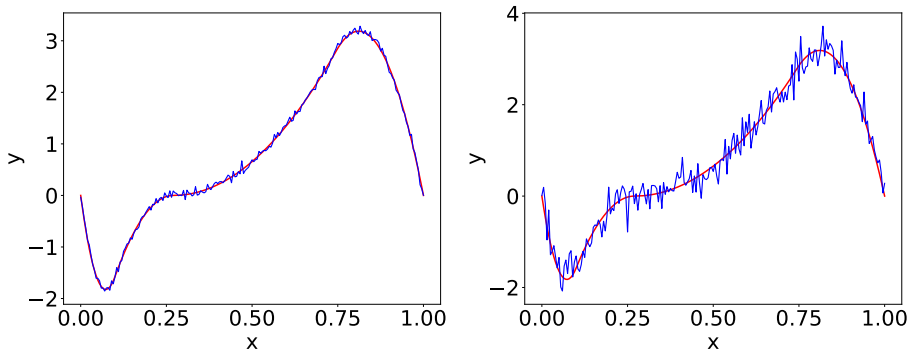

Figure 3.6: Function f_1 to estimate with a noisy set of observations $Y_1, \dots, Y_{201}$ of $\sigma = 0.05$ (left) and $\sigma = 0.25$ (right).

selected knots remains the same as the one previously found for $\sigma = 0.1$.

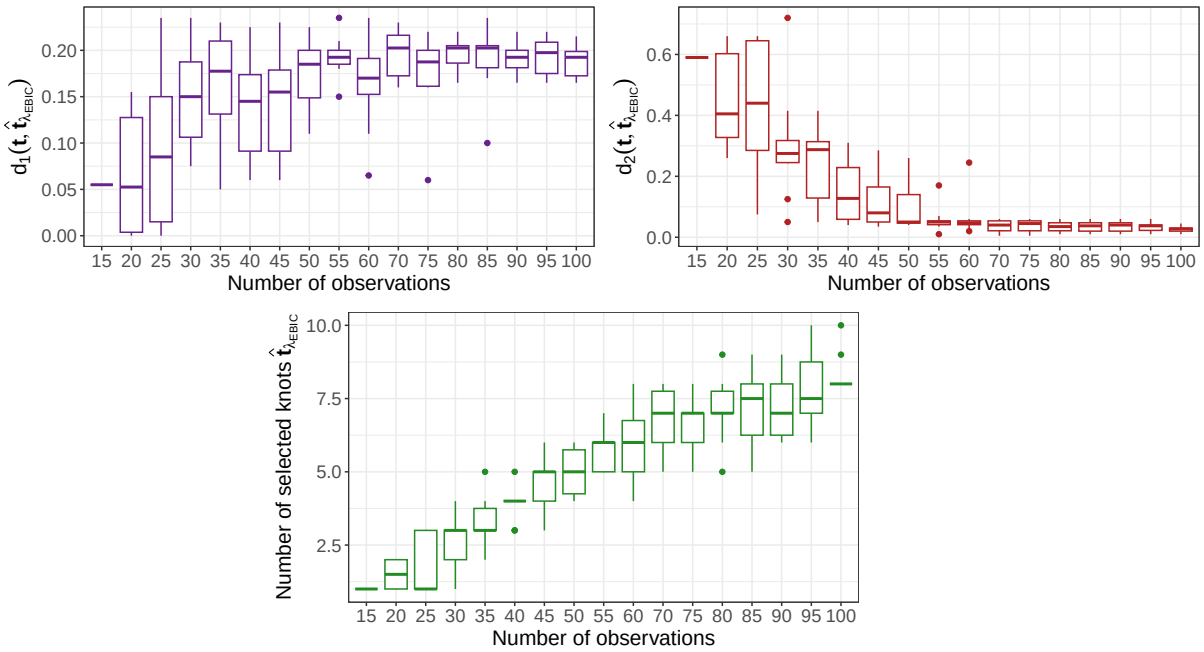

Figure 3.7: Similar to Figure 3.2 with $\lambda = \lambda_{\text{EBIC}}$ for estimating f_1 when $\sigma = 0.25$.

The corresponding results for the statistical performance defined in (3.17) for n varying from 15 to 100 are displayed in Figure 3.8 for $N = 201$. We can see from this figure that the level of noise deteriorates the performance of every method. However, our approach still has high levels of precision since the Normalized Sup Norm varies from $10^{-1.75}$ to $10^{-1.25}$ for $n = 100$ which allows it to outperform the other methods. The surprising results obtained by the Gaussian Processes in the right part of Figure 3.8 may either come from the fact that the noise is too high or from the fact that the selected points may be too close thus leading to a ill-conditioned covariance matrix.

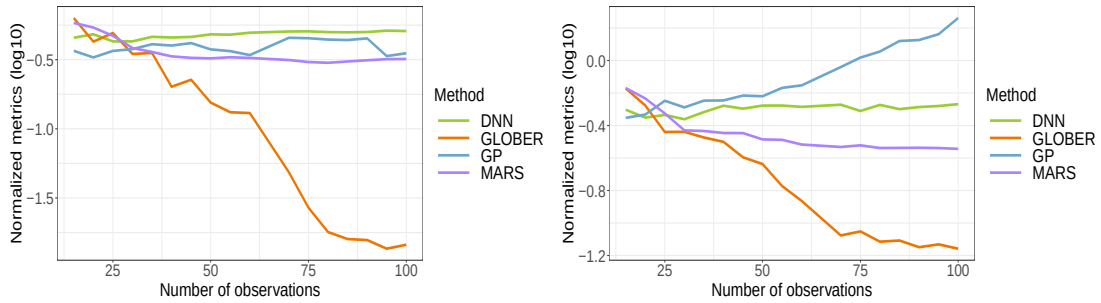

Figure 3.8: Statistical performance (Normalized Sup Norm) of GLOBER for estimating f_1 when $\sigma = 0.05$ (left) and $\sigma = 0.25$ (right) and of the state-of-the-art methods obtained from 10 replications.

3.3.1.2. Two-dimensional case ($d = 2$)

In this part, we focus on the estimation of the function f_2 for $d = 2$ from a noisy set of observations $(Y_i)_{1\leq i\leq n}$ obtained with $\sigma = 0.005$ (resp. $\sigma = 0.05$) and $n = 40401$. The corresponding Y_is are displayed in the left (resp. right) part of Figure 3.9.

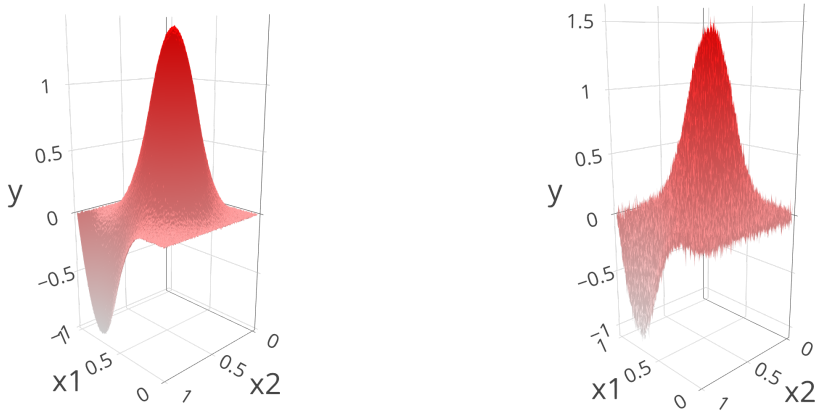

Figure 3.9: Function f_2 to estimate with a noisy set of observations $Y_1, \ldots, Y_{40401}$ with $\sigma = 0.005$ (left) and $\sigma = 0.05$ (right).

The two parts of the Hausdorff distance between the estimated and the real knots as well as the number of selected knots obtained for the noisiest observation set ($\sigma = 0.05$) are shown in Figure 3.20 of the Appendix. As for the one-dimensional case, we can see that the results are similar to those previously obtained with $\sigma = 0.01$ in Figure 3.19 of the Appendix: the number of selected knots is the same and the second part of the Hausdorff distance tends to 0.

Figure 3.21 of the Appendix displays the average of the statistical performance obtained from 10 random samplings of the set of observations for $N = 40401$ where the statistical measure is defined in (3.17). We can see that even though an alteration of the performance is visible, our approach still outperforms the other methods since the normalized metric keeps decreasing, contrary to the DNNs and the MARS approaches which seem to reach a plateau and the Gaussian Processes which led to very poor accuracy on noisy observations. Once again,

our method remains robust with highly noisy observation sets both in the one-dimensional and in the two-dimensional case.

3.3.2. Influence of the sampling of the observation set

We now assess the influence of the sampling of the observation set on the performance of our approach. To do so, we apply our method on randomly chosen observations and we calculate the average Normalized Sup Norm defined in (3.17) on 10 different samplings of the observations. In such situations, the knots used to define the function to estimate are not necessarily included in the set of observations and in this case, cannot thus be chosen as knots of the B-spline basis. Then, we compare it to the case where the set of observations necessarily contains the set of knots used to define the underlying function f to estimate.

3.3.2.1. One-dimensional case ($d = 1$)

We first assess the estimation of the function f_1 defined in (3.15) from a noisy set of observations obtained with $\sigma = 0.05$. Figure 3.10 shows the Hausdorff distance between the set of knots $\mathbf{t}$ of the function f_1 and the observation set $\mathbf{x}$ and the Hausdorff distance between the set of knots $\mathbf{t}$ and its estimation with our method in the case where the knots are not necessarily included in the observation set. We can see from this figure that for each element of the observation set, there exists at least one knot at a distance smaller than 0.25. Moreover, for each knot there exists at least one point of the observation set at a distance very close to 0 for large enough n. In addition, we can see from the plot on the bottom right that for large enough n, there exists for each knot an estimated one which is very close.

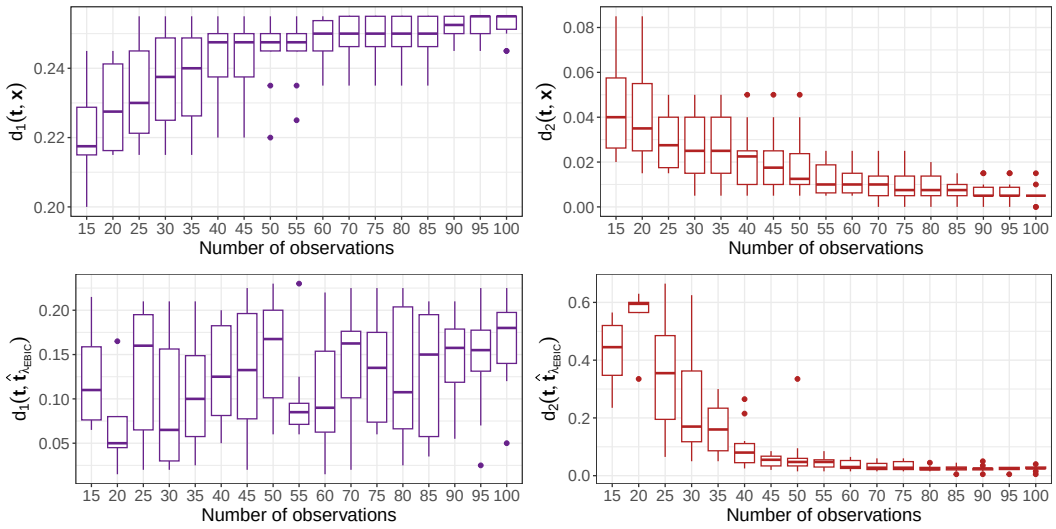

Figure 3.10: Boxplots of the first part of the Hausdorff distance as a function of n: $d_1(\mathbf{t}, \mathbf{x})$ (top left) and $d_1(\mathbf{t}, \widehat{\mathbf{t}}_{\lambda_{\text{EBIC}}})$ (bottom left) and for the second part of the Hausdorff distance as a function of n: $d_2(\mathbf{t}, \mathbf{x})$ (top right) and $d_2(\mathbf{t}, \widehat{\mathbf{t}}_{\lambda_{\text{EBIC}}})$ (bottom right) for estimating f_1 when the observation set is randomly chosen and $\sigma = 0.05$.

In the case where $\mathbf{t}$ belongs to $\mathbf{x}$, the results are displayed in Figure 3.22 of the Appendix. The boxplot on the top left shows the same results for the distance $d_1(\mathbf{t}, \mathbf{x})$ as for the random sampling and the boxplot on the top right confirms that $\mathbf{t}$ belongs to $\mathbf{x}$ since $d_2(\mathbf{t}, \mathbf{x}) = 0$ at every value of n. Furthermore, we can see similar results on the bottom left and bottom right

boxplots since distances $d_1(\mathbf{t}, \widehat{\mathbf{t}}_{\lambda_{\text{EBIC}}})$ and $d_2(\mathbf{t}, \widehat{\mathbf{t}}_{\lambda_{\text{EBIC}}})$ have exactly the same behavior as for the random sampling case.

Finally, the number of selected knots displayed in Figure 3.11 shows comparable results between the random sampling of the observations and when $\mathbf{t}$ belongs to $\mathbf{x}$. Therefore, the random sampling of the observations does not seem to affect the knot selection of our method.

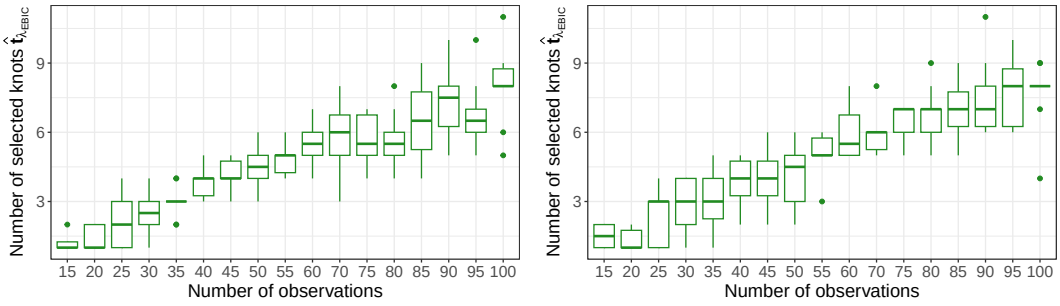

Figure 3.11: Number of estimated knots as a function of n for the estimation of f_1 with GLOBER from a random sampling of observations (left) and when $\mathbf{t}$ belongs to $\mathbf{x}$ (right) with $\sigma = 0.05$.

The results of the statistical performance of our method for the estimation of the function f_1 are displayed in Figure 3.12 for $N = 201$. We can clearly see that the random sampling of the observation set does not deteriorate the performance of our method in comparison to the case where the observation set contains all the knots: the value of the Normalized Sup Norm reaches in both cases $10^{-1.75}$ for $n = 100$ and our method still outperforms the other ones.

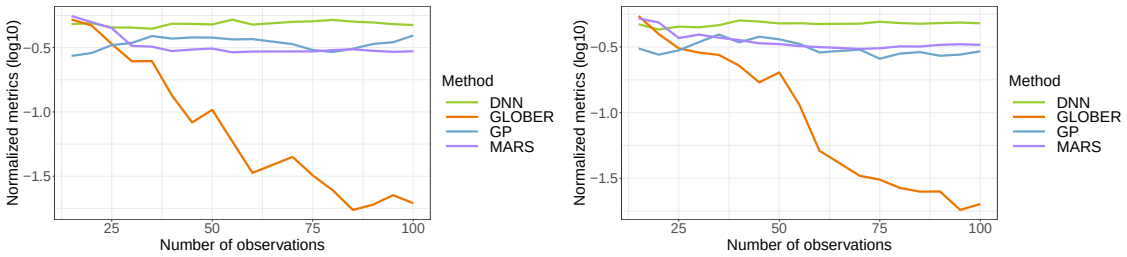

Figure 3.12: Statistical performance (Normalized Sup Norm) of GLOBER for estimating f_1 from a random sampling of the observation set (left) and with $\mathbf{t}$ belonging to the observation set (right) with $\sigma = 0.05$. Comparison to the performance of the state-of-the-art methods obtained from 10 replications.

3.3.2.2. Two-dimensional case ($d = 2$)

Similarly to the previous part, we study the estimation of the function f_2 from a noisy set of observations obtained with $\sigma = 0.01$. Figure 3.23 of the Appendix shows the Hausdorff distance between the set of knots $\mathbf{t}_1$ (resp. $\mathbf{t}_2$) of the function f_2 and the observation set $\mathbf{x}_1 = \{x_{11}, \ldots, x_{1n_1}\}$ (resp. $\mathbf{x}_2 = \{x_{21}, \ldots, x_{2n_2}\}$) for the first (resp. second) dimension. It also displays the Hausdorff distance between the set of knots $\mathbf{t}_1$ (resp. $\mathbf{t}_2$) and its estimation with our method $\widehat{\mathbf{t}}_{1,\widetilde{\lambda}_1}$ (resp. $\widehat{\mathbf{t}}_{2,\widetilde{\lambda}_2}$), with $\widetilde{\lambda}_1 = \widetilde{\lambda}_{1,\text{EBIC}}$ (resp. $\widetilde{\lambda}_2 = \widetilde{\lambda}_{2,\text{EBIC}}$) for the first (resp. second) dimension in the case where the knots do not necessarily belong to the observation set. We can see from this figure that for each element of the observation set, there exists at

least one knot at a distance smaller than 0.45 (resp. 0.38) for the points of the first dimension (resp. second dimension). Furthermore, for each knot of each dimension there exists at least one point of the observation set of the corresponding dimension at a distance very close to 0 when n is large enough (at least 20 points per dimension). For the case where $\mathbf{t}_1$ and $\mathbf{t}_2$ belong to $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, we can see from Figure 3.24 of the Appendix that the distances $d_1(\mathbf{t}_1, \mathbf{x}_1)$ and $d_1(\mathbf{t}_2, \mathbf{x}_2)$ have the same behavior as for the random sampling. Moreover, the boxplots on the top right shows that $d_2(\mathbf{t}_1, \mathbf{x}_1) = 0$ and $d_2(\mathbf{t}_2, \mathbf{x}_2) = 0$ at every n value which confirms that the sets of knots belong to the observation set. Similar results can be observed for the evolution of the distances $d_1(\mathbf{t}_1, \widehat{\mathbf{t}}_{1,\widetilde{\lambda}_1})$ (resp. $d_1(\mathbf{t}_2, \widehat{\mathbf{t}}_{2,\widetilde{\lambda}_2})$) and $d_2(\mathbf{t}_1, \widehat{\mathbf{t}}_{1,\widetilde{\lambda}_1})$ (resp. $d_2(\mathbf{t}_2, \widehat{\mathbf{t}}_{2,\widetilde{\lambda}_2})$) for the random sampling and when the knots belong to the observation set. Indeed, in both cases we can see from the plot on the bottom right of Figures 3.23 and 3.24 of the Appendix that for large enough n, there exists for each knot of each dimension an estimated one which is very close. Moreover, the number of selected knots displayed in Figure 3.25 of the Appendix shows comparable results between the random sampling of the observations and when $\mathbf{t}$ belongs to $\mathbf{x}$. Therefore, the random sampling of the observations does not seem to affect the knot selection of our method for $d = 2$.

As it is the case for $d = 1$, we can see from Figure 3.26 of the Appendix where the statistical performance of our method for estimating f_2 for $N = 40401$ are displayed that the performance of our approach is not altered by the sampling of the observation set. More precisely, the value of the Normalized Sup Norm reaches in both cases $10^{-1.5}$ for $n = 1600$ and our method still outperforms the other approaches which seem to reach a constant value.

3.3.3. Numerical performance

The goal of this section is to investigate the computational times of our approach GLOBER implemented in the `glober` R package available on the CRAN as a function of the number of observation points. The timings were obtained on a workstation with 31.2GB of RAM and Intel Core i7 (3.8GHz) CPU. The average computational times and their standard deviation obtained from 30 independent executions are displayed in Figure 3.13. We can see from this figure that it only takes 600ms to estimate the underlying function with our approach in the one-dimensional. In the two-dimensional case, the computational time is larger but even in this case the computational burden of our approach is low since it only takes 110 seconds for processing 1600 observations.

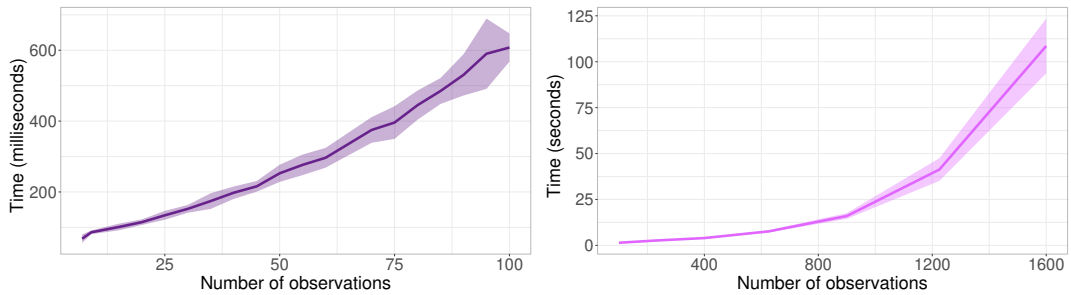

Figure 3.13: Average and standard deviation of the execution time measured in milliseconds for $d = 1$ (left) and in seconds for $d = 2$ (right) on 30 independent executions.

3.4. Application to geochemical systems

In this section, we apply our method to simple chemical problems with one or two input concentrations which correspond to the estimation of a function f from observations satisfying (3.1) for $d = 1$ or $d = 2$, respectively. In geochemical simulations, high-quality data can be generated using powerful solvers such as PHREEQC (see Parkhurst and Appelo (2013) for more details). These solvers are optimized to describe real chemical equations with extreme precision. This is the reason why hereafter the noise is very small.

In the following, the increasing number of observations is obtained by randomly adding new points from a given grid of N points to the pre-existing observation sets. Since the functions to estimate come from chemical processes and are known to be smooth, we propose taking $q \geq 2$. Nevertheless, we observed that using $q = 3$ does not yield better results compared to $q = 2$. This is the reason why we will present in the sequel results obtained for $q = 2$.

3.4.1. One-dimensional application ($d = 1$)

We consider hereafter the estimation of the amount of a "Salt" mineral as a function of the concentrations of its constituents Sp_a^+ and Sp_b^- as in Savino et al. (2022). For this example, the thermodynamic constants of the halite Salt (NaCl) were considered because there are only two constitutive elements and they do not depend on the pH of the solution. Following the law of mass action, the dissolution reaction of this mineral writes:

$$\mathrm{Salt} \rightleftharpoons \mathrm{Sp_a^+} + \mathrm{Sp_b^-}.$$

At equilibrium, the activity of these elements $\mathrm{a}_{Sp_a^+}$ and $\mathrm{a}_{Sp_b^-}$ obey the solubility product

$$\mathrm{K_{Salt}} = \mathrm{a_{Sp_a^+}} \mathrm{a_{Sp_b^-}} = 10^{1.570}.$$

To reduce the problem to a one dimensional case, we consider the amount of Salt as a function of the normalized concentration of Sp_a^+ with an additional numerical noise, while fixing the value of Sp_b^-. Hence, the normalized concentration of Sp_a^+ belongs to $[0, 1]$. The corresponding data are displayed on the left part of Figure 3.14.

Our method is used to estimate this function f_3 by using a set of a varying number of observation points ($7 \leq n \leq 100$) and is then compared to the state-of-the-art methods described in Section 3.3. Figure 3.15 displays the illustration of the estimation of f_3 for a varying number of n. We can see by adding new points to the observation set (blue crosses) that GLOBER has better chance to choose knots (blue bullets) among observation points which depict more precisely changes in the underlying curve. However, it seems that 40 observations are enough to have a good approximation of our function f_3 since the estimation (black curve) fits quite perfectly to the observation points (red curve).

The corresponding statistical performance is computed thanks to (2.19) where $f(x_k)$ is replaced by Y_k and f_{min} (resp. f_{max}) is replaced by the minimal (resp. maximal) value of $(Y_k)_{1 \leq k \leq 1140}$. The results are displayed on the right side of Figure 3.14 for $N = 1140$. We can see from this figure that our method outperforms the other approaches and allows us to get a very good accuracy of the amount of Salt since the average Normalized Sup Norm reaches $10^{-1.5}$ for only $n = 100$.

3.4.2. Two-dimensional application ($d = 2$)

Our method is then applied to a more complex chemical problem which derives from the calcite dissolution and precipitation study described in Kolditz et al. (2012). The corresponding

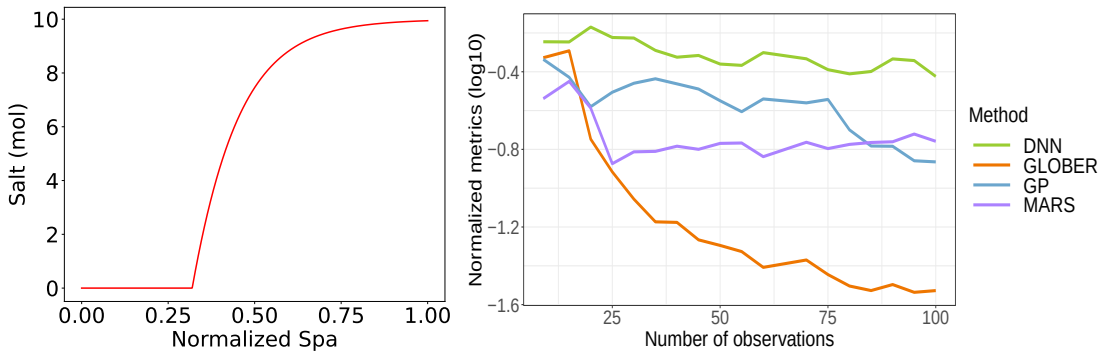

Figure 3.14: Left: Values of Salt obtained with PHREEQC to estimate f_3 when $d = 1$. Right: Statistical performance (Normalized Sup Norm) of GLOBER and of the state-of-the-art methods for estimating f_3. The average values are obtained from 10 replications.

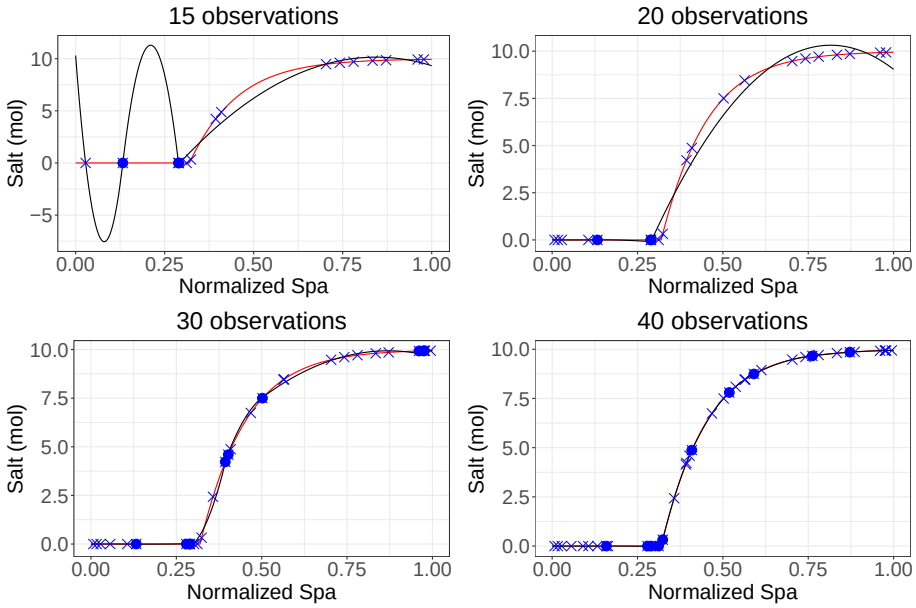

Figure 3.15: Illustration of the estimation of the amount of Salt depending on the normalized concentration of Sp_a^+ for 15 (top left), 20 (top right), 30 (bottom left) and 40 observations (bottom right). The red curve displays the values of the observations, the black curve corresponds to the estimation of f_3, the blue crosses are the observation points used for estimating f_3 and the blue bullets are the observation points chosen as estimated knots.

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. However, in this chapter, we will only consider the dolomite precipitation:

$$\mathrm{Dolomite} \rightleftharpoons \mathrm{Ca}^{2+} + \mathrm{Mg}^{2+} + 2\mathrm{CO}_3^{2-}, \quad \log\mathrm{K}_{10} = -17.09.$$

The amount of dolomite is computed with PHREEQC as a function of the total elemental concentrations (C, Ca, Cl, Mg), the pH (as $-\log(\mathrm{H}^+)$) and the amount of calcite. In this chapter,

we will only consider the dolomite precipitation for C=5$\times 10^{-4}$ mol/kgw, Cl=2×10^{-3} mol/kgw, pH=10, calcite=0 mol in order to reduce the problem to a two-dimensional case. Thus, the function f_4 to estimate is defined by considering the normalized concentrations of Ca and Mg as input variables. The data produced by PHREEQC in this context and from which f_4 is estimated are displayed in the left part of Figure 3.16.

We seek to estimate f_4 by applying our method to an increasing number of observations ($100 \leq n \leq 1600$) which corresponds to an increasing number of points per dimension ($10 \leq n_1, n_2 \leq 40$ with $n = n_1 n_2$). The resulting illustration is displayed in Figure 3.27 of the Appendix and shows an improvement of the fitting of GLOBER (green curve) to the observation data (red curve) by adding new points to the observation set (orange bullets) in order to get a perfect overlapping for $n \geq 900$.

The corresponding statistical performance is computed thanks to (2.19) where $f(x_k)$ is replaced by Y_k and f_{min} (resp. f_{max}) is replaced by the minimal (resp. maximal) value of $(Y_k)_{1 \leq k \leq 40000}$. The results are displayed on the right side of Figure 3.16 for $N = 40000$. Similarly to what has been shown for the precipitation of Salt, our method gives satisfactory results as the Normalized Sup Norm reaches 10^{-1}. Moreover, our method still outperforms the other ones for which the statistical metric seems to rapidly reach a constant value.

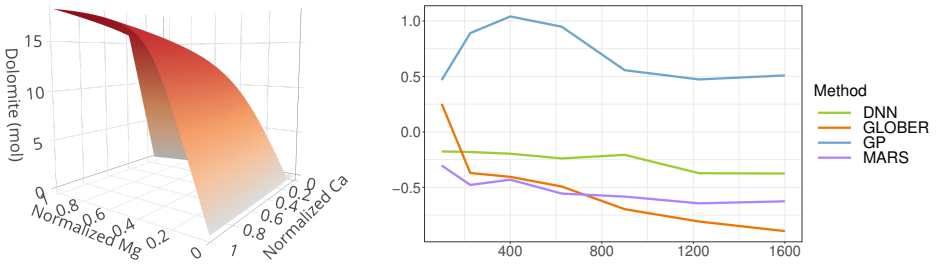

Figure 3.16: Left: Amount of dolomite obtained with PHREEQC to estimate f_4 when $d = 2$. Right: Statistical performance (Normalized Sup Norm) of GLOBER and of the state-of-the-art methods for estimating f_4. The average values are obtained from 10 replications.

3.5. Extension to higher dimensional and more general observation settings

Here, let us consider the case where $x_i \in \mathbb{R}^d$ with $d \geq 2$ and f is still a function to estimate from (3.1). In this section, the grid of N points and the observation sets are no longer constrained to result from a Cartesian product of d compact sets, such as in Section 3.2.2.

3.5.1. Adaptation of the knot selection method by using clustering

In order to avoid finding the knots associated to a given dimension for each fixed value of the others, we propose in this section to cluster the n observation points by using a k-means approach. More precisely, for each dimension j, the n observation points are gathered into clusters based on their multidimensional characteristics excluding dimension j. Then, the one-dimensional approach described in Section 3.2.1 is applied to the Y's associated to the x's belonging to this cluster to find the knots in dimension j. The number k of clusters is chosen as

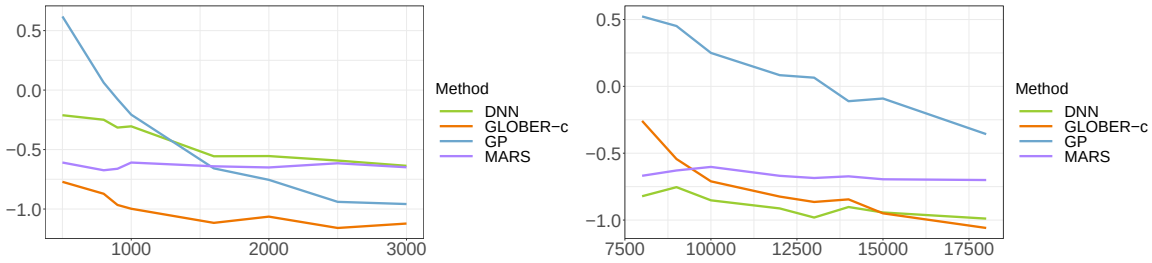

Figure 3.17: Statistical performance (Normalized Sup Norm) of GLOBER-c and of the state-of-the-art methods for estimating f_4 (left) and f_5 (right). The average values are obtained from 10 replications.

the integer part of n divided by 25 to guarantee a large enough amount of points per cluster. Thus, since sets of knots are found for each $\widetilde{\boldsymbol{\lambda}}_j$ of $\widetilde{\Lambda_j}$ we gather the knots $\widehat{\mathbf{t}}_{\widetilde{\boldsymbol{\lambda}}_j}$ into clusters. The final knot sets are built by keeping the median value of each cluster of knots. The number of knot clusters is chosen as the value where there is a change in the slope of the within-sum of squares. Finally, the EBIC criterion defined in (3.21) is used to find the final combination of knot sets.

3.5.2. Case where $d = 2$

This method was first validated on the two-dimensional framework application by comparing the results obtained for the clusterized version called GLOBER-c on the Dolomite precipitation case described in Section 3.4.2. The obtained results are displayed in the left part of Figure 3.17. We can see from this figure that compared to the results obtained in Section 3.4.2 with the non-clusterized version, the maximal absolute error is smaller than 10^{-1} for $n = 1600$ and that our method still outperforms the three others.

3.5.3. Case where $d = 3$

In this section, we apply our method to an extension of the geochemical system described in Section 3.4.2. Our goal is to estimate f_5 which describes the relationship between the normalized input concentrations of C, Ca and Mg and the amount of Dolomite. The obtained results are displayed in the right part of Figure 3.17. We can see from this figure that our method is the only one that displays a maximal absolute error below 10^{-1}. Moreover, it outperforms the three others when the number of observations is larger than 15000 observations.

3.6. Conclusion

In this chapter, we propose a novel approach for estimating functions in a multivariate nonparametric regression setting using an adaptive knot selection method for B-splines. Our procedure is implemented in the `glober` R package which is available on the Comprehensive R Archive Network (CRAN). In the course of this study, we showed that our approach is very attractive since, compared to other alternative approaches, it is very efficient from a statistical point of view and can benefit from a low computational burden when dealing with a high number of observations.

3.7. Appendix : Additional plots

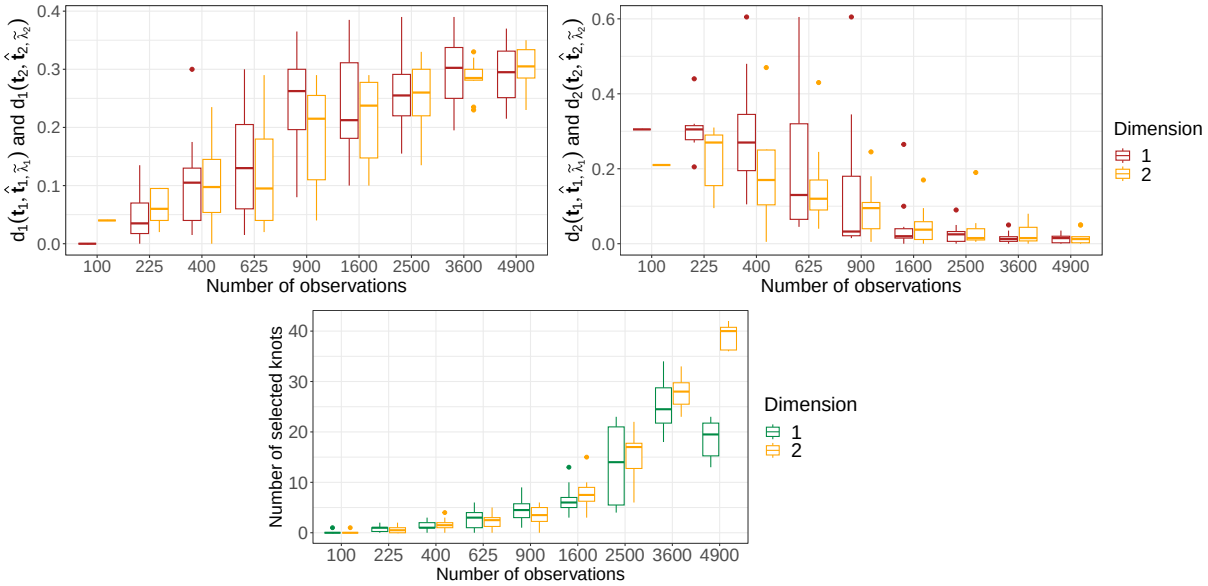

Figure 3.18: Top left: Boxplots for the first part of the Hausdorff distance (d_1) as a function of $n = n_1 n_2$ and top right: boxplots for the second part of the Hausdorff distance (d_2) as a function of $n = n_1 n_2$ between the two sets of knots $\mathbf{t}_1$ and $\widehat{\mathbf{t}}_{1,\widetilde{\lambda}_1}$ and $\mathbf{t}_2$ and $\widehat{\mathbf{t}}_{2,\widetilde{\lambda}_2}$ for the first and second dimension, respectively. Bottom: number of estimated knots as a function of $n = n_1 n_2$ by choosing $\widetilde{\lambda}_1 = \widetilde{\lambda}_{1,\text{EBIC}}$ and $\widetilde{\lambda}_2 = \widetilde{\lambda}_{2,\text{EBIC}}$ for the estimation of f_2 with $\sigma = 0.01$.

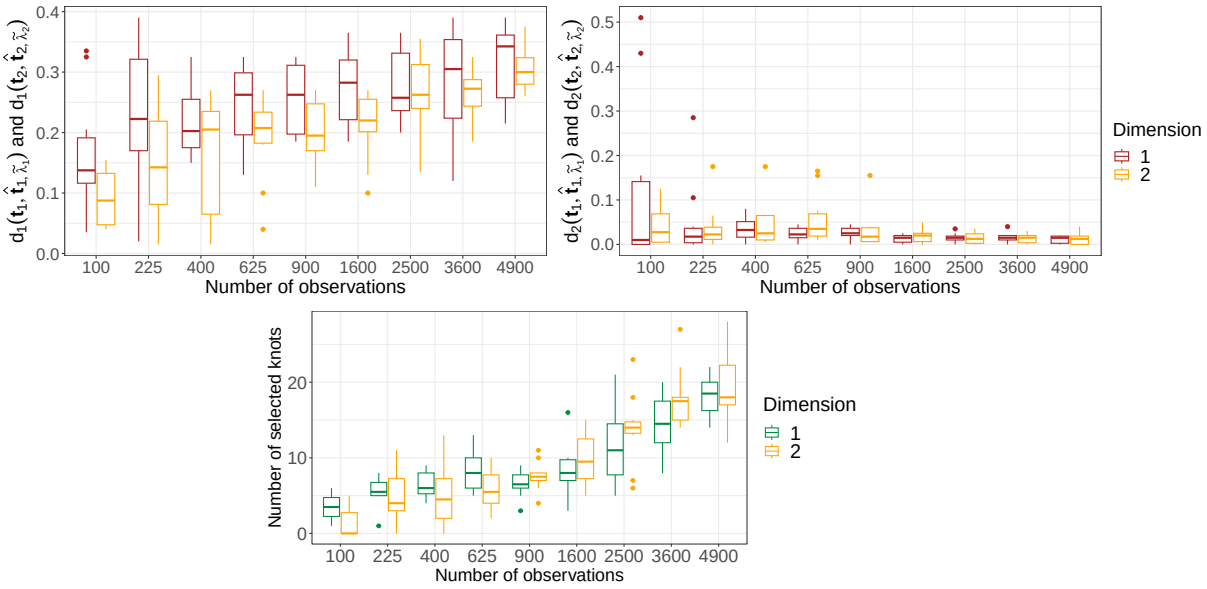

Figure 3.19: Similar to Figure 3.18 by choosing $\widetilde{\lambda}_1 = \widetilde{\lambda}_{1,\text{opt}}$ and $\widetilde{\lambda}_2 = \widetilde{\lambda}_{2,\text{opt}}$ for the estimation of f_2 with $\sigma = 0.01$.

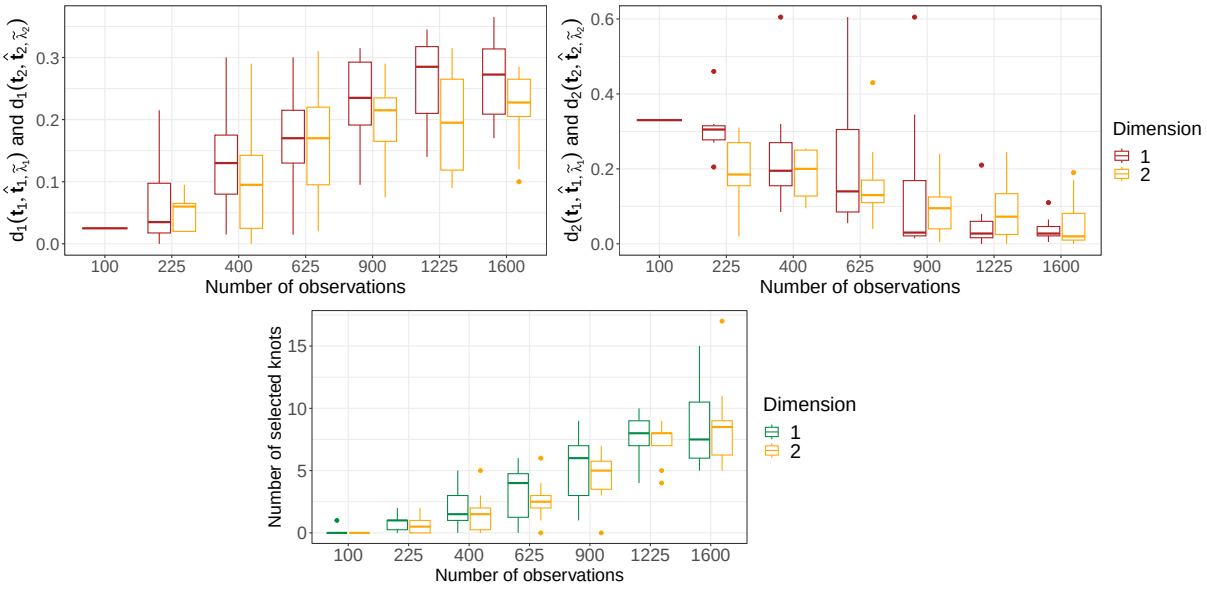

Figure 3.20: Similar to Figure 3.18 by choosing $\widetilde{\lambda}_1 = \widetilde{\lambda}_{1,\text{EBIC}}$ and $\widetilde{\lambda}_2 = \widetilde{\lambda}_{2,\text{EBIC}}$ for the estimation of f_2 with $\sigma = 0.05$.

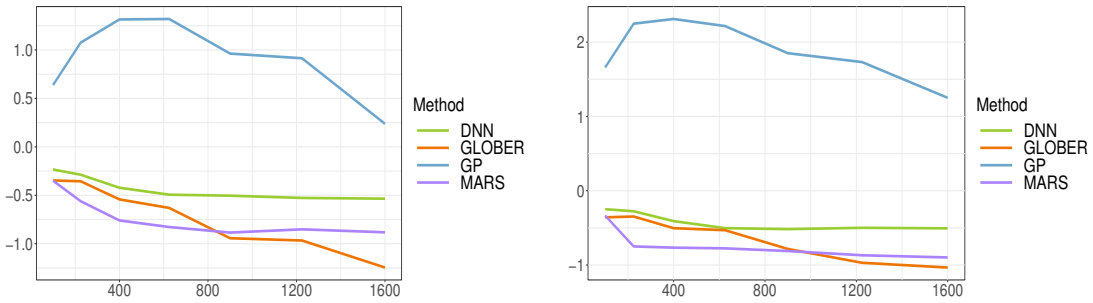

Figure 3.21: Statistical performance (Normalized Sup Norm) of our method using the EBIC criterion for $\sigma = 0.005$ (left) and $\sigma = 0.05$ (right) and of the state-of-the-art methods obtained from 10 replications.

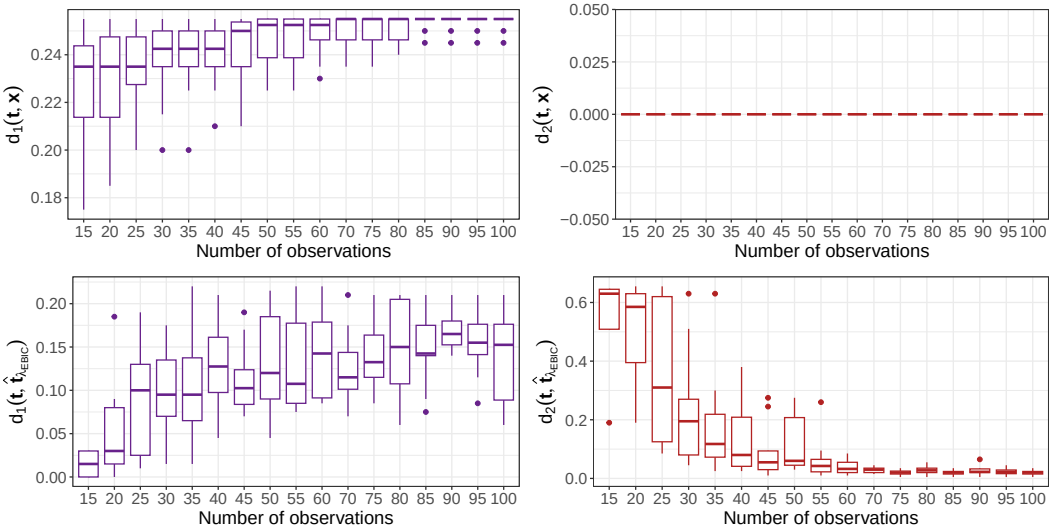

Figure 3.22: Similar to Figure 3.10 for the estimation of f_1 with $\mathbf{t}$ belonging to $\mathbf{x}$ and $\sigma = 0.05$.

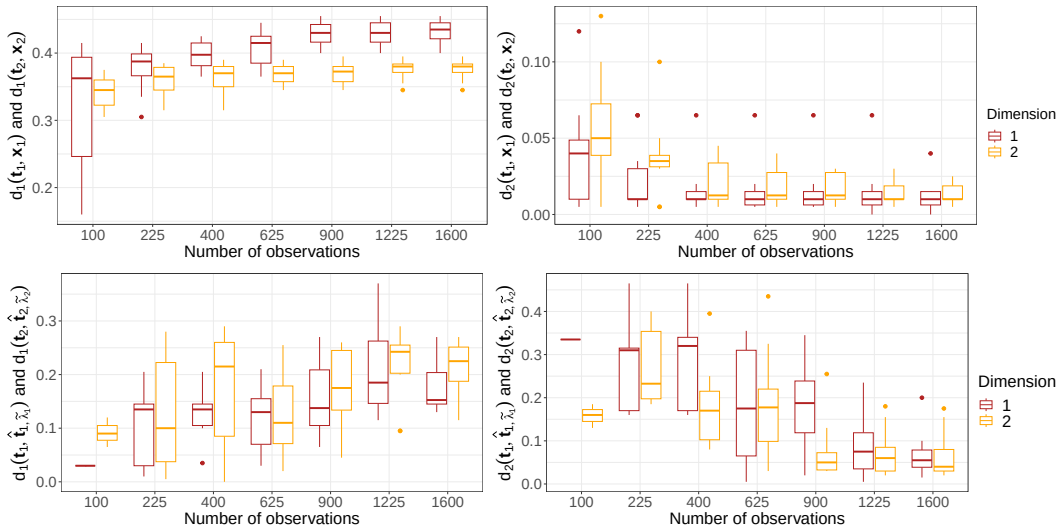

Figure 3.23: Boxplots of the first part of the Hausdorff distance as a function of n: $d_1(\mathbf{t}_1, \mathbf{x}_1)$ (resp. $d_1(\mathbf{t}_2, \mathbf{x}_2)$) (top left) and $d_1(\mathbf{t}_1, \widehat{\mathbf{t}}_{1,\widetilde{\lambda}_1})$ (resp. $d_1(\mathbf{t}_2, \widehat{\mathbf{t}}_{2,\widetilde{\lambda}_2})$) (bottom left) and for the second part of the Hausdorff distance as a function of n: $d_2(\mathbf{t}_1, \mathbf{x}_1)$ (resp. $d_2(\mathbf{t}_2, \mathbf{x}_2)$) (top right) and $d_2(\mathbf{t}_1, \widehat{\mathbf{t}}_{1,\widetilde{\lambda}_1})$ (resp. $d_2(\mathbf{t}_2, \widehat{\mathbf{t}}_{2,\widetilde{\lambda}_2})$) (bottom right) for the first (resp. second) dimension, for the estimation of f_2 by choosing $\widetilde{\lambda}_1 = \widetilde{\lambda}_{1,\text{EBIC}}$ and $\widetilde{\lambda}_2 = \widetilde{\lambda}_{2,\text{EBIC}}$ with a random sampling of the observation set with $\sigma = 0.01$.

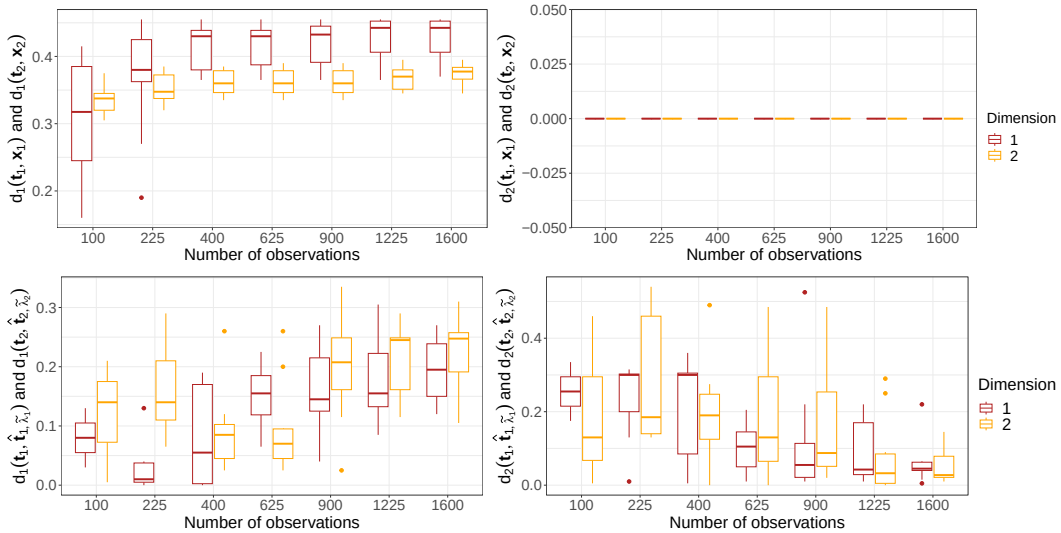

Figure 3.24: Similar to Figure 3.23 for the estimation of f_2 with $\mathbf{t}_1$ belonging to $\mathbf{x}_1$ (resp. $\mathbf{t}_2$ belonging to $\mathbf{x}_2$) and $\sigma = 0.01$.

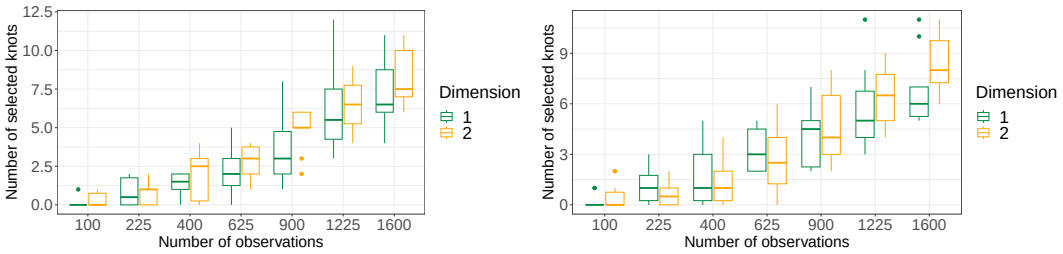

Figure 3.25: Left: number of estimated knots as a function of n for the estimation of f_2 with GLOBER from a random sampling of observations (left) and when $\mathbf{t}_1$ and $\mathbf{t}_2$ belong to $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively (right) with $\sigma = 0.01$.

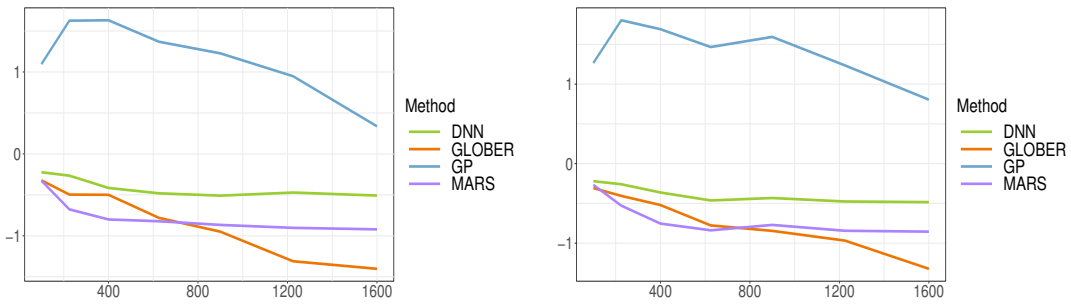

Figure 3.26: Statistical performance (Normalized Sup Norm) of GLOBER from a random sampling of the noisy observation set (left) and with $\mathbf{t}_1$ and $\mathbf{t}_2$ belonging to the observation set (right) with $\sigma = 0.01$. Comparison to the performance of state-of-the-art methods obtained from 10 replications.

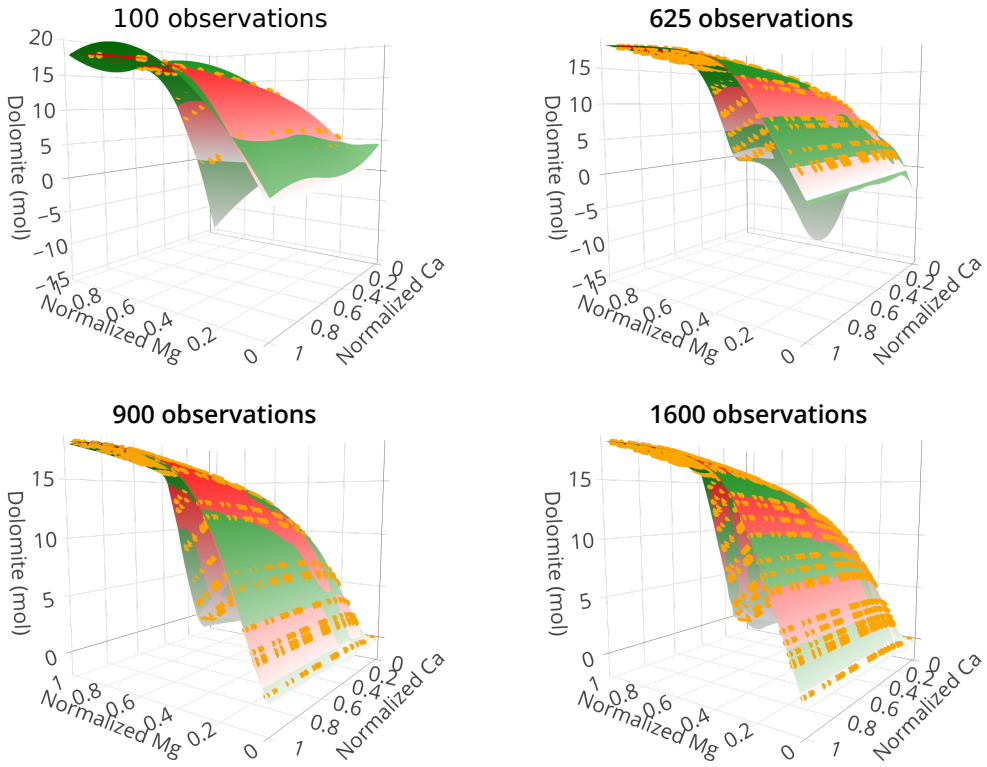

Figure 3.27: Illustration of the estimation of the amount of dolomite depending on the normalized concentration of Ca and Mg for 100 (top left), 625 (top right), 900 (bottom left) and 1600 observations (bottom right). The red surface describes the values of the observations, the green surface corresponds to the estimation with GLOBER and the orange bullets are the observation points used for the estimation.

Chapter 4 - A novel variable selection method in nonlinear multivariate models using B-splines with an application to geoscience

Scientific contribution

The content of this chapter is the subject of the article:
Savino, E. M., Lévy-Leduc, C. A novel variable selection method in nonlinear multivariate models using B-splines with an application to geoscience. Submitted and also available on HAL preprint *(hal-04434820)*.

The proposed method is implemented in the `absorber` R package available from the CRAN.

Abstract

In this chapter, 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 compare 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).

Table of contents

4.1. 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 such as in Laloy and Jacques (2019). However, the computational efficiency of ANN in reducing simulation time comes at the cost of requiring a large dataset and often demands extensive CPU times for training and tuning the hyperparameters, see Karpatne et al. (2018). Furthermore, an approach based on an on-demand training algorithm presented in Leal et al. (2017) allows to train the model at runtime to iteratively build the training dataset. Analogously, an active learning approach has also been introduced to RTM by 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:

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = (x_i^{(1)}, \dots, x_i^{(p)}) \in \mathbb{R}^p, \quad 1 \leq i \leq n \tag{4.1}$$

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 (4.1) 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 (4.1) 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 (4.1). 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 (4.1) 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 chapter, 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 (4.1) 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 chapter is organized as follows. Section 4.2 presents the methodology that we propose for variable selection in nonlinear models. Section 4.3 investigates the performance of our approach through numerical experiments. Finally, in Section 4.4, we apply our method to a real geochemical application that motivated this study.

4.2. Methodology

4.2.1. 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:

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

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

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

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

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:

$$B_{k,1}^{(\ell)}(x^{(\ell)}) = \begin{cases} 1 & \text{if } \tau_{\ell,k} \leq x^{(\ell)} < \tau_{\ell,k+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k = 1, \ldots, K + 2M - 1, \tag{4.2}$$

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

$$B_{k,m}^{(\ell)}(x^{(\ell)}) = \frac{x^{(\ell)} - \tau_{\ell,k}}{\tau_{\ell,k+m-1} - \tau_{\ell,k}} B_{k,m-1}^{(\ell)}(x^{(\ell)}) + \frac{\tau_{\ell,k+m} - x^{(\ell)}}{\tau_{\ell,k+m} - \tau_{\ell,k+1}} B_{k+1,m-1}^{(\ell)}(x^{(\ell)}), \tag{4.3}$$

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 (4.1) by a linear combination of B-splines of each variable $x^{(1)}, \ldots, x^{(p)}$ and of pairwise interaction of them as follows:

$$F(x^{(1)}, \ldots, x^{(p)}) = \sum_{\ell=1}^{p} \sum_{k=1}^{K+M} \beta_k^{(\ell)} B_k^{(\ell)}(x^{(\ell)}) + \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \left(\sum_{k=1}^{K+M} \sum_{q=1}^{K+M} \beta_{k,q}^{(\ell,j)} B_k^{(\ell)}(x^{(\ell)}) B_q^{(j)}(x^{(j)}) \right), \tag{4.4}$$

where $B_k^{(\ell)} = B_{k,M}^{(\ell)}$ is defined in (4.2) and (4.3) and where $\beta_k^{(\ell)}$ and $\beta_{k,q}^{(\ell,j)}$ are unknown coefficients.

Observe that the column vector $(F(x_i^{(1)}, \ldots, x_i^{(p)}))_{1 \leq i \leq n}$ (4.4) can be rewritten as follows:

$$\sum_{\ell=1}^{p} \Psi_\ell \, \boldsymbol{\beta}_\ell + \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \, \boldsymbol{\beta}_{\ell,j} \, . \tag{4.5}$$

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 $\boldsymbol{\beta}_\ell = \begin{pmatrix} \beta_1^{(\ell)} & \ldots & \beta_{K+M}^{(\ell)} \end{pmatrix}^T$ 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 $\boldsymbol{\beta}_{\ell,j} = \begin{pmatrix} \beta_{1,1}^{(\ell,j)} & \beta_{1,2}^{(\ell,j)} & \ldots & \beta_{K+M,K+M}^{(\ell,j)} \end{pmatrix}^T$ for $1 \leq \ell < j \leq p$.

4.2.2. 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 (4.5) by minimizing the following regularized criterion:

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

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the Y_i's being defined in (4.1), $\partial_\ell F(x_i)$ denotes the ℓth partial derivative of F defined in (4.4) 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 (4.5) the criterion can be rewritten as follows:

$$\begin{aligned}
&\left(\widehat{\boldsymbol{\beta}}_1(\lambda),\dots,\widehat{\boldsymbol{\beta}}_p(\lambda),\widehat{\boldsymbol{\beta}}_{1,2}(\lambda),\dots,\widehat{\boldsymbol{\beta}}_{(p-1),p}(\lambda)\right)\\
&= \underset{\substack{(\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_p)\\(\boldsymbol{\beta}_{12},\dots,\boldsymbol{\beta}_{(p-1)p})}}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell\, \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p} \Phi_{\ell j}\, \boldsymbol{\beta}_{\ell,j} \right\|_2^2 \right.\\
&\left. +\lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell\, \boldsymbol{\beta}_\ell + \sum_{j=\ell+1}^{p} (\partial_\ell \Phi_{\ell j})\, \boldsymbol{\beta}_{\ell,j} + \sum_{1\le j<\ell} (\partial_\ell \Phi_{j\ell})\, \boldsymbol{\beta}_{j,\ell} \right\|_2 \right),
\end{aligned} \tag{4.6}$$

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

$$\lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell\, \boldsymbol{\beta}_\ell + (\partial_\ell \Phi_{\ell\bullet})\, \boldsymbol{\beta}_{\ell\bullet} + (\partial_\ell \Phi_{\bullet\ell})\, \boldsymbol{\beta}_{\bullet\ell} \right\|_2 =: \lambda \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell)\, \boldsymbol{\gamma}_\ell \right\|_2, \tag{4.7}$$

where $\boldsymbol{\gamma}_\ell = \left(\boldsymbol{\beta}_\ell^T \;\; \boldsymbol{\beta}_{\ell\bullet}^T \;\; \boldsymbol{\beta}_{\bullet\ell}^T\right)^T$. 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:

$$\begin{aligned}
&\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell\, \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p} \Phi_{\ell j}\, \boldsymbol{\beta}_{\ell,j} \right\|_2^2\\
&= \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell\, \boldsymbol{\beta}_\ell - \frac{1}{2} \left(\sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p} \Phi_{\ell j}\, \boldsymbol{\beta}_{\ell,j} + \sum_{\ell=2}^{p}\sum_{j=1}^{\ell-1} \Phi_{j\ell}\, \boldsymbol{\beta}_{j,\ell} \right) \right\|_2^2\\
&=: \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell\, \boldsymbol{\gamma}_\ell \right\|_2^2.
\end{aligned} \tag{4.8}$$

Equation (4.8) comes by defining $\Theta_\ell = \left(\Psi_\ell \quad \frac{1}{2}\Phi_{\ell\bullet} \quad \frac{1}{2}\Phi_{\bullet\ell}\right)$ and setting $\Theta_1 = \left(\Psi_1 \quad \frac{1}{2}\Phi_{1\bullet}\right)$ and $\Theta_p = \left(\Psi_p \quad \frac{1}{2}\Phi_{\bullet p}\right)$, 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 (4.7) and (4.8), (4.6) can be rewritten as:

$$\left(\widehat{\boldsymbol{\gamma}}_1(\lambda),\dots,\widehat{\boldsymbol{\gamma}}_p(\lambda)\right) = \underset{(\boldsymbol{\gamma}_1,\dots,\boldsymbol{\gamma}_p)}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell\, \boldsymbol{\gamma}_\ell \right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell)\, \boldsymbol{\gamma}_\ell \right\|_2 \right). \tag{4.9}$$

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, (4.9) can be rewritten as:

$$\left(\widehat{\boldsymbol{\alpha}}_1(\lambda),\dots,\widehat{\boldsymbol{\alpha}}_p(\lambda)\right) = \underset{(\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_p)}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \widetilde{\mathbf{X}}_\ell\, \boldsymbol{\alpha}_\ell \right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\| \boldsymbol{\alpha}_\ell \right\|_2 \right). \tag{4.10}$$

The last formulation of our variable selection criterion (4.10) 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 λ, this package provides a set of penalization parameters Λ and the coefficients $\widehat{\boldsymbol{\alpha}}_\ell(\lambda)$ for λ belonging to Λ. The coefficients $\widehat{\gamma}_\ell(\lambda)$ are thus obtained as follows

$$\widehat{\gamma}_\ell(\lambda) = (\partial_\ell \Theta_\ell)^+ \widehat{\boldsymbol{\alpha}}_\ell(\lambda). \tag{4.11}$$

We then define the active variables for each λ in Λ as follows:

$$\mathcal{V}_\lambda = \left\{ \ell,\ \sum_{k\geq 1} |\widehat{\gamma}_{\ell,k}(\lambda)| \neq 0 \right\}, \tag{4.12}$$

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

We also introduce the set $\mathcal{V}_f$ of the indices of the d relevant variables on which f in (4.1) 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.

4.2.3. 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 λ belonging to Λ, we computed the True Positive Rate (TPR) and the False Positive Rate (FPR), defined as:

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

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. $\mathcal{V}_f$, $\overline{\mathcal{V}_f}$ and $\mathcal{V}_\lambda$ are introduced in the previous section. We can then draw the ROC curve where each point has as coordinates $(\mathrm{FPR}(\lambda), \mathrm{TPR}(\lambda))$ for λ 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:

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

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

$$\begin{gathered} f_1(x^{(1)},x^{(2)},x^{(3)},x^{(4)},x^{(5)}) = 2B_2^{(3)}(x^{(3)})B_4^{(5)}(x^{(5)}) + 2B_4^{(5)}(x^{(5)}) - 5B_2^{(3)}(x^{(3)}), \\ (x^{(1)},\dots,x^{(5)}) \in [0,1]^5, \end{gathered} \tag{4.13}$$

$$\begin{gathered} f_2(x^{(1)},\dots,x^{(10)}) = 1.8\cos(x^{(1)})\sin(x^{(7)}+1) - 5\ln(x^{(3)}+1) - \frac{0.9}{{x^{(10)}}^2+1}, \\ (x^{(1)},\dots,x^{(10)}) \in [0,1]^{10}. \end{gathered} \tag{4.14}$$

In (4.13), 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 $\{1, \dots, 5\}$. 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 $\sigma = 0.25$ in Figure 4.1 and in Figure 4.3 for $n = 700$ and $n = 2000$, respectively. Firstly, we can clearly see for $n = 700$ in Figure 4.1 that for f_1, all the values of K are satisfying as they allow us to get max(TPR $-$ FPR) $= 1$ and AUC $= 1$.

However, for $f = f_2$, we do not have necessarily AUC $= 1$ when max(TPR $-$ FPR) $= 1$ for instance for $K = 1$, which does not imply that this method does not select properly the relevant variables. To illustrate this idea, one can relate to Figure 4.2 in which the ROC curves for $f = f_2$ are drawn for each value of K belonging to $\{1, \dots, 10\}$. Here, we can observe that a good variable selection method will not necessarily lead to AUC $= 1$ since 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 4.3 for f_2. The results for $n = 700$ in Figure 4.1 allow us to discriminate a value of K which gives good selection for both $\sigma = 0$ and $\sigma = 0.25$ and for both functions f_1 and f_2. Firstly, for f_2 without any noise in the observation set ($\sigma = 0$) we can see that only the true variables are selected for $K = 1$ and $K = 2$ since AUC $= 0$ and max(TPR $-$ FPR) $= 1$. Moreover, these two values of K are the only cases where max(TPR $-$ FPR) $= 1$ for all the different samplings of the observation set when $\sigma = 0.25$. Higher values of K drastically deteriorate the AUC and the max(TPR−FPR), especially for noisy observation sets. Choosing a small value for K offers the advantage of simplifying our model and speeding up computational executions. For all these reasons, we decide to only focus on using our method with $K = 1$ knot in the B-spline basis.

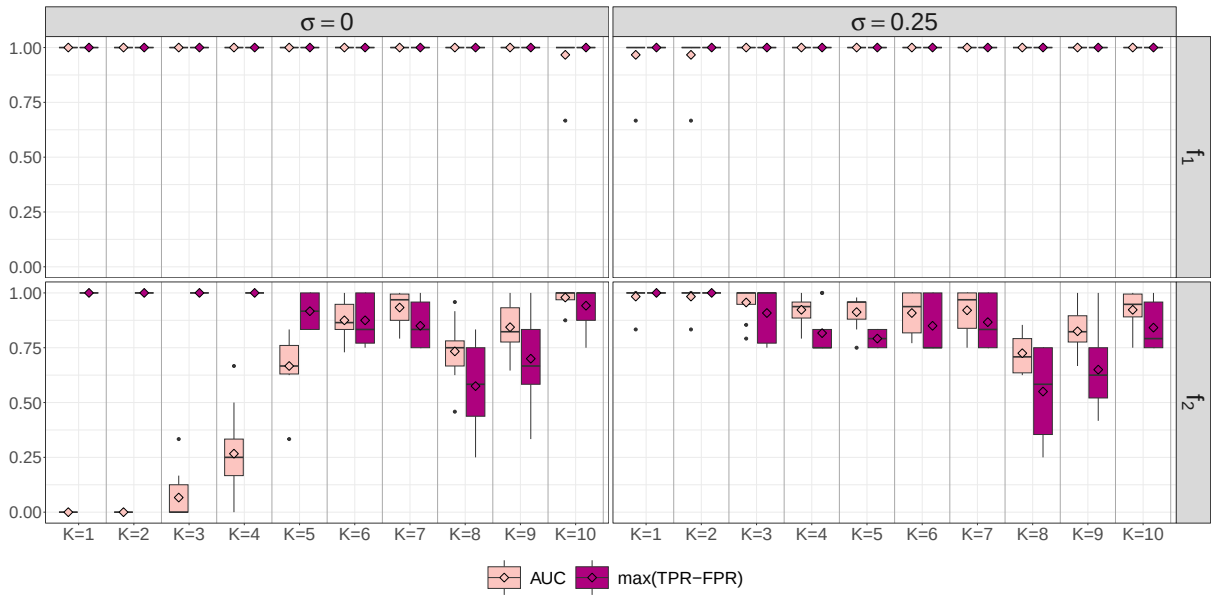

Figure 4.1: AUC and max$(\text{TPR} - \text{FPR})$ calculated for an increasing value of K for f_1 (top) and f_2 (bottom) with noise (right) or without noise (left) in the observation set $\mathbf{Y} = (Y_1, \dots, Y_{700})$. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty bullets inside the boxplots represent the mean value and the plain bullets outside the boxplots are the extreme values.

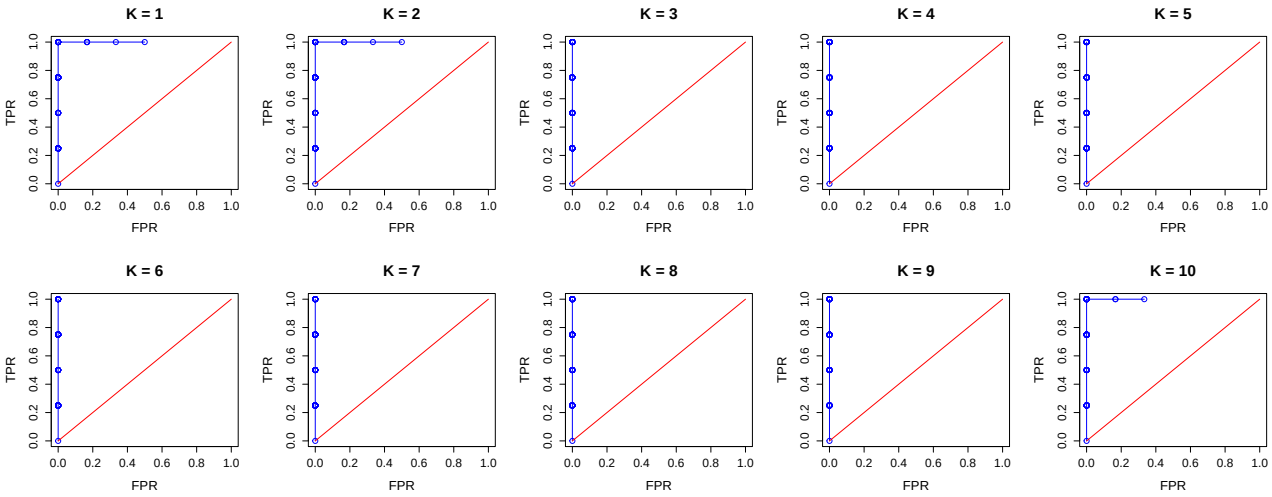

Figure 4.2: ROC curves for an increasing value of K for f_2 with no noisy observations $\mathbf{Y} = (Y_1, \ldots, Y_{2000})$ ($\sigma = 0$ in (4.1)) and for one sampling of the observation set (blue line). The red line corresponds to the identity function TPR $=$ FPR.

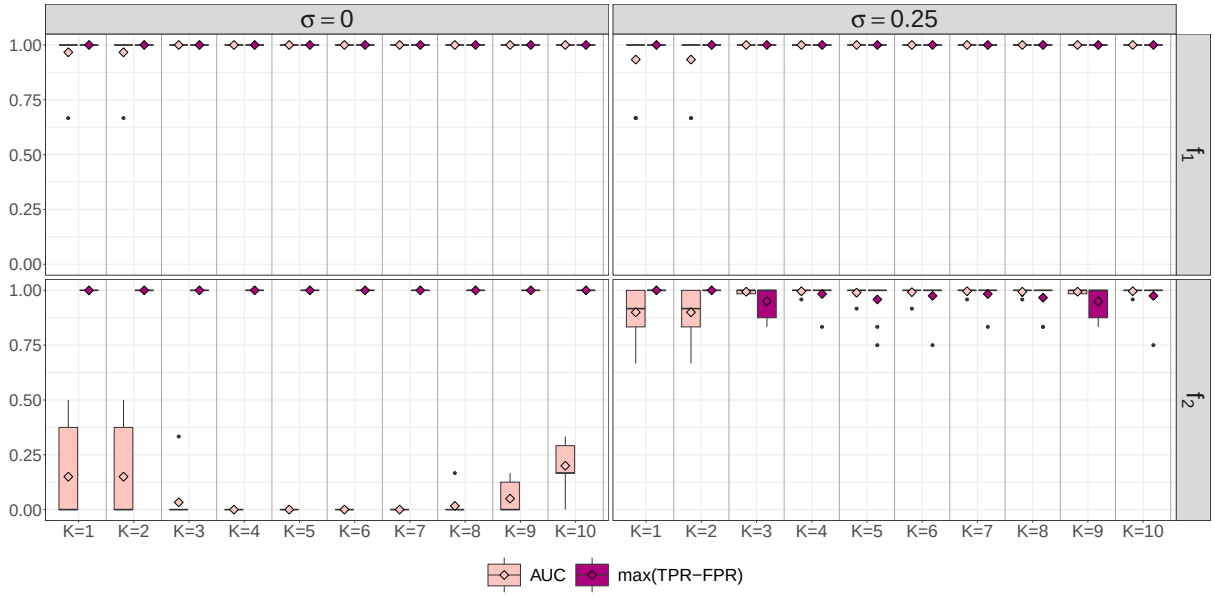

Figure 4.3: AUC and $\max(\text{TPR} - \text{FPR})$ calculated for an increasing value of K for f_1 (top) and f_2 (bottom) with noise (right) or without noise (left) in the observation set $\mathbf{Y} = (Y_1, \ldots, Y_{2000})$. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty bullets inside the boxplots represent the mean value and the plain bullets outside the boxplots are the extreme values.

4.2.4. Choice of λ

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 $\{1, \dots, p\}$ by:

$$P_\ell = \frac{100}{|\Lambda|} \sum_{\lambda \in \Lambda} \mathbb{1}\{\ell \in \mathcal{V}_\lambda\}\,, \tag{4.15}$$

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 (4.12).

Results for the percentage of selection of variables are displayed in Figure 4.4 (resp. Figure 4.5) 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 62.5% of the λs belonging to Λ. Moreover, we can observe a noticeable 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 $n = 700$ 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 (25% for variable 7 and $\sigma = 0$ and 35% for $\sigma = 0.25$). However, we can see smaller percentages 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 $n = 700$ 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 $\sigma = 0.25$ 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 4.4 and 4.5 which are visualizations of the output of our method for one sample of the observation set for f_1 and f_2, respectively.

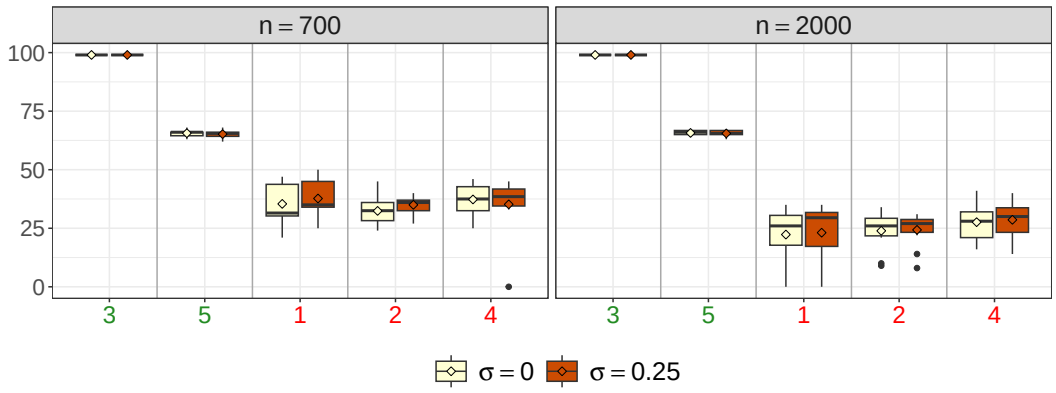

Figure 4.4: Percentage of selection of each variable for f_1 with $n = 700$ (left) and $n = 2000$ (right) and for $\sigma = 0$ or $\sigma = 0.25$. The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_1. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty bullets inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

We also propose another method to automatically choose λ. 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:

$$\mathsf{AIC}(\lambda) = n \ln\left(\frac{\mathsf{RSS}(\lambda)}{n}\right) + 2T_\lambda\,, \tag{4.16}$$

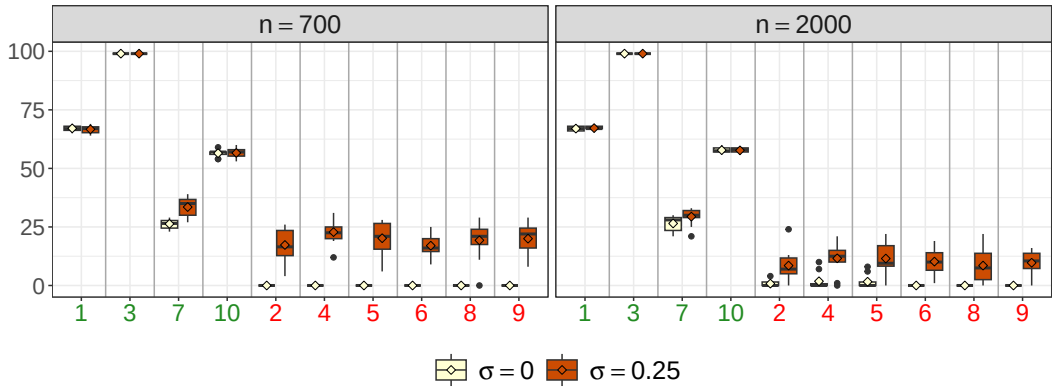

Figure 4.5: Percentage of selection of each variable for f_2 with $n = 700$ (left) and $n = 2000$ (right) and for $\sigma = 0$ or $\sigma = 0.25$. The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_2. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

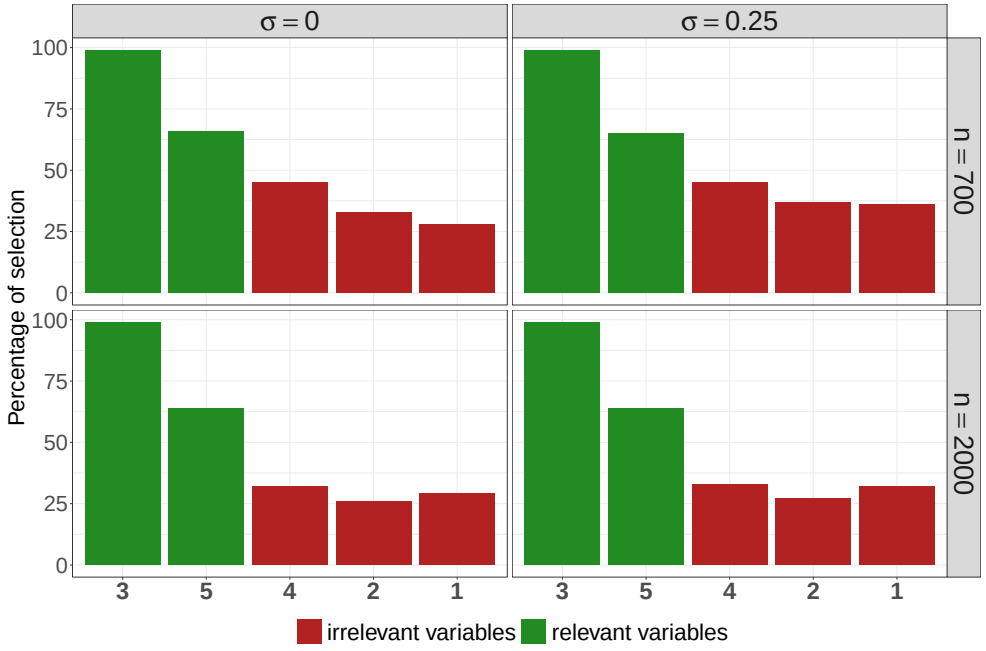

Figure 4.6: Percentage of selection of each variable for f_1 with $n = 700$ (top) and $n = 2000$ (bottom) and for $\sigma = 0$ (left) or $\sigma = 0.25$ (right). The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_1. Only one sampling of $\mathbf{Y}$ is used to obtain the results displayed.

where T_λ is the number of terms appearing in (4.5) by keeping only the variables selected with λ and $\text{RSS}(\lambda)$ is the residual sum of squares defined as follows:

$$\begin{aligned} \text{RSS}(\lambda) &= \left\| \mathbf{Y} - \widehat{\mathbf{Y}}(\lambda) \right\|_2^2, \\ \text{with} \quad \widehat{\mathbf{Y}}(\lambda) &= \sum_{\ell=1}^{p} \Theta_\ell \widehat{\gamma}_\ell(\lambda), \end{aligned} \tag{4.17}$$

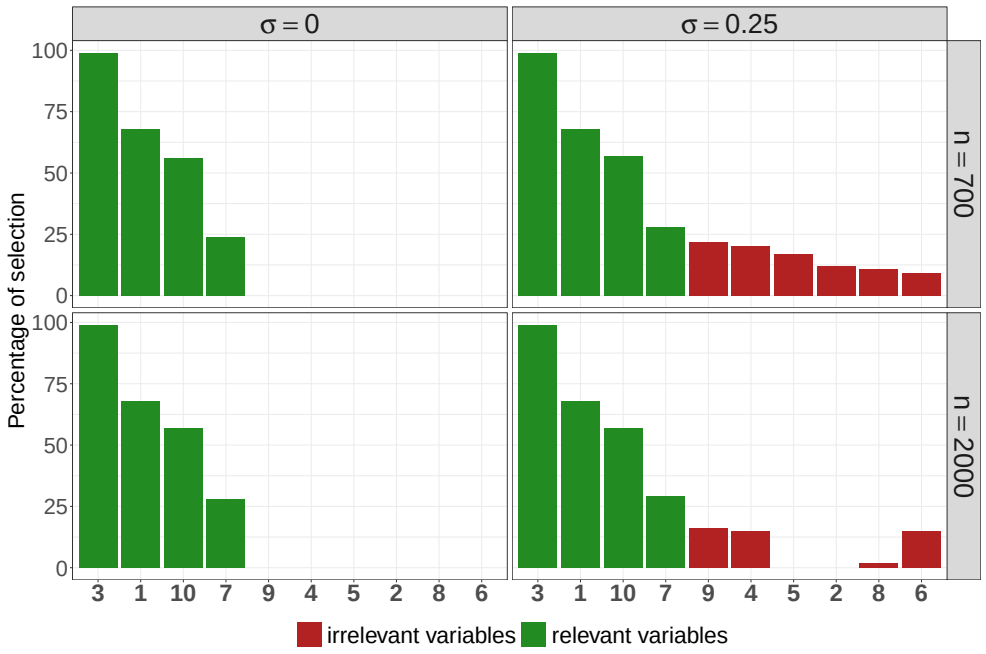

Figure 4.7: Percentage of selection of each variable for f_2 with $n = 700$ (top) and $n = 2000$ (bottom) and for $\sigma = 0$ (left) or $\sigma = 0.25$ (right). The green (resp. red) variables indicate the true relevant (resp. irrelevant) variables for f_2. Only one sampling of $\mathbf{Y}$ is used to obtain the displayed results.

where $\widehat{\gamma}_\ell(\lambda)$ is defined in (4.11). Then, the chosen $\lambda = \lambda_{\mathsf{AIC}}$ is such that:

$$\lambda_{\mathsf{AIC}} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \left(\mathsf{AIC}(\lambda)\right). \tag{4.18}$$

The $\mathsf{TPR}(\lambda)$ and $\mathsf{FPR}(\lambda)$ obtained with $\lambda = \lambda_{\mathsf{AIC}}$ for both functions f_1 and f_2 are displayed in Figure 4.8 for $n = 700$ and $n = 2000$ and for $\sigma = 0$ or $\sigma = 0.25$. We can observe for f_1 that even with a small number of observations ($n = 700$) this criterion allows us to get $\mathsf{TPR}(\lambda_{\mathsf{AIC}}) = 1$ while having $\mathsf{FPR}(\lambda_{\mathsf{AIC}}) = 0$ 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 $\sigma = 0.25$, we indeed have $\mathsf{TPR}(\lambda_{\mathsf{AIC}}) < 1$ and $\mathsf{FPR}(\lambda_{\mathsf{AIC}}) = 0$. By increasing the value of n from 700 to 2000, the value of $\mathsf{TPR}(\lambda_{\mathsf{AIC}})$ is increased and thus the number of relevant selected variables. Moreover, for unnoisy set of observations the relevant variables are recovered from $n = 700$. 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.

4.3. 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

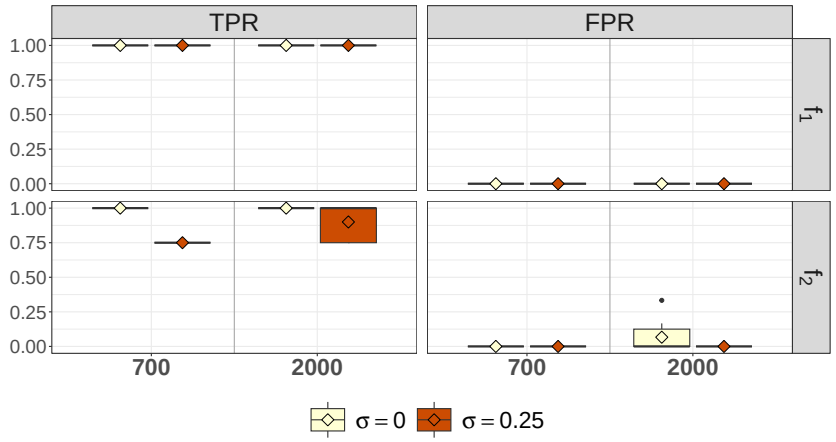

Figure 4.8: TPR(λ) and FPR(λ) values by choosing $\lambda = \lambda_{\text{AIC}}$ for f_1 (top) and f_2 (bottom) with an unnoisy ($\sigma = 0$) or noisy ($\sigma = 0.25$) set of observations. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

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 (4.15). 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 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.

4.3.1. 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 (4.15) for 10 samplings of the observation set and the results are shown in Figure 4.9 (resp. Figure 4.10) for $f = f_1$ (resp. $f = f_2$) with n belonging to $\{700, 1000, 2000\}$ (resp. $\{700, 1000\}$). The relevant variables are displayed in green and the irrelevant variables are displayed in red. Additional results for f_2 are presented in the Appendix in Figure 4.16. As we can see in Figure 4.9, 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 $\sigma = 0.25$, the percentage of irrelevant variables selected

with our method drops from nearly 37.5% for $n = 700$ to 25% with $n = 2000$. 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 close to that of variable 5 belonging to $\mathcal{V}_{f_1}$, with a high selection rate of 62.5%. It has to be noticed that Random Forests selects variable 5 with only 37.5% of selection whereas our method (resp. LassoNet) selects it with a percentage of 65% (resp. 73%).

Let us now study the application of these methods to f_2. The corresponding results are displayed in Figure 4.10 and in Figure 4.16 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 18.5% of selection when $\sigma = 0.25$ 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 to 32%. Here, using a threshold at 25% allows us to discriminate the relevant variables from the irrelevant ones.

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 as variables 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.

4.3.2. 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:

$$f_3\left(x^{(1)},\dots,x^{(5)}\right) = 1.8\cos\left(x^{(1)}\right)\sin\left(x^{(3)}+1\right) - 5\ln\left(x^{(3)}+1\right) - \frac{0.9}{(x^{(4)})^2+1} \\ \left(x^{(1)},\dots,x^{(5)}\right) \in [0,1]^5, \tag{4.19}$$

$$f_4\left(x^{(1)},\dots,x^{(25)}\right) = 1.8\cos\left(x^{(1)}\right)\sin\left(x^{(7)}+1\right) - 5\ln\left(x^{(3)}+1\right) - \frac{0.9}{\left(x^{(10)}\right)^2+1}, \\ \left(x^{(1)},\dots,x^{(25)}\right) \in [0,1]^{25}. \tag{4.20}$$

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 4.2.3. The results for these comparisons are displayed in Figure 4.11 for $n = 700$ and $n = 2000$.

We can see from this figure that our method is affected by p but only when the observation set is corrupted with significant noise ($\sigma = 0.25$) and has a reduced size ($n = 700$). Specifically, the value of max(TPR $-$ FPR) equals 1 for $p \leq 10$ and is less than 1 for $p = 25$. In contrast, the efficiency of the two other methods is impacted by the value of p even for $\sigma = 0$ as the

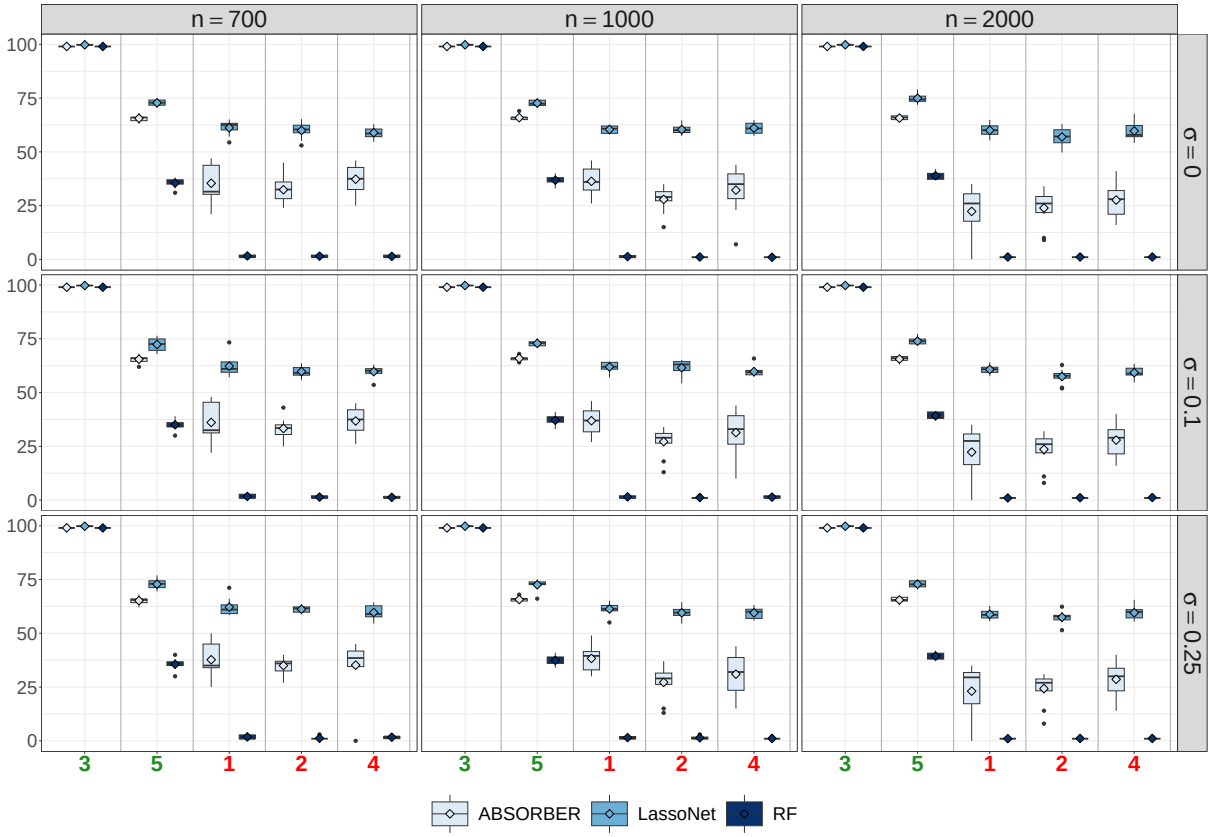

Figure 4.9: Percentage of selection of each variable of f_1 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

metrics presented in Figure 4.11A are deteriorated for $p \geq 10$. For instance, with LassoNet, max(TPR $-$ FPR) $= 1$ and AUC $= 1$ with $p = 5$, regardless of σ. However, max(TPR $-$ FPR) becomes less that 1 for $p = 10$ and it keeps decreasing as the values of p and σ increase. Even in the easiest case where $p = 5$ and $\sigma = 0$, Random Forests is less competitive than the other approaches and its performance remain unchanged regardless the values of n, σ and p. For both of these two methods and for $\sigma = 0$, the AUC being close to 1 indicates a high FPR, suggesting that these methods select multiple irrelevant variables while omitting one or more relevant ones. Increasing the number of observations up to $n = 2000$ improves the results in Figure 4.11B for our method ABSORBER and Lassonet. Nevertheless, Random Forests continue to display less satisfactory results. Our method is the only one showing satisfactory results as max(TPR $-$ 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.

In Figure 4.12, the impact of the value p on our variable selection procedure using AIC is assessed. We can observe that p has little effect on the efficiency of our method, in contrast to σ which has a more pronounced negative impact on the metrics. Specifically, the TPR values are affected by σ as it necessitates an increasing value of n as σ grows to achieve TPR$(\lambda_{\text{AIC}}) = 1$.

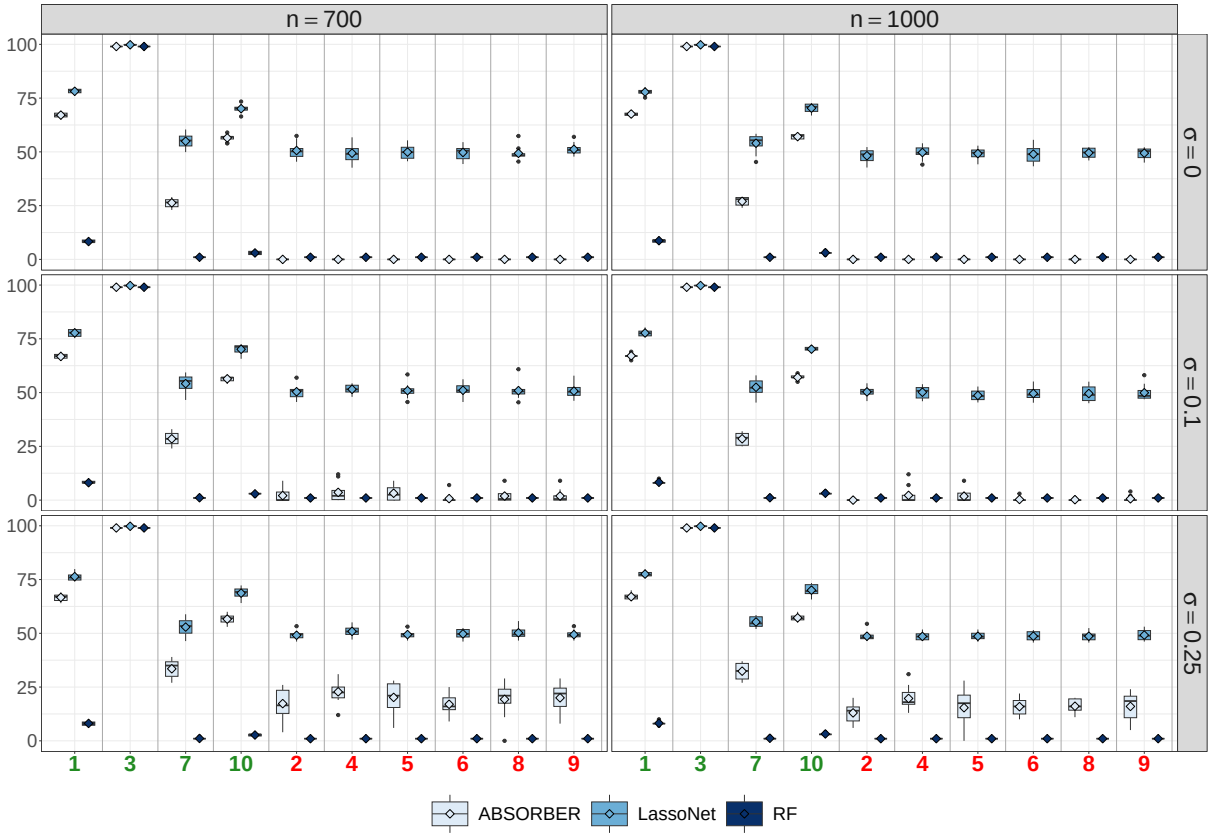

Figure 4.10: Percentage of selection of each variable of f_2 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

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. Since the number of relevant variables is $d = 4$ even for $p = 5$, the selection of just one irrelevant variable leads to $\text{FPR}(\lambda_{\text{AIC}}) = 0.25$, thereby explaining the observed results for $n = 2000$. However, such trends are not visible for high value of p, demonstrating the efficiency of our variable selection procedure.

4.3.3. Numerical performance

The goal of this section is to investigate the computational times of our variable selection approach implemented in the `absorber` R package. Our variable selection method is applied to f_2, f_3 and f_4 (defined in (4.14), (4.19) and (4.20), respectively) for $p = 5$, $p = 10$ and $p = 25$, 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 4.13. We can see from this figure that it only takes 57 seconds to perform variable selection on a function with $p = 25$ variables from $n = 2000$ observations. It has to be noticed that the execution times reported are mainly due to the use of the `sparsegl` package.

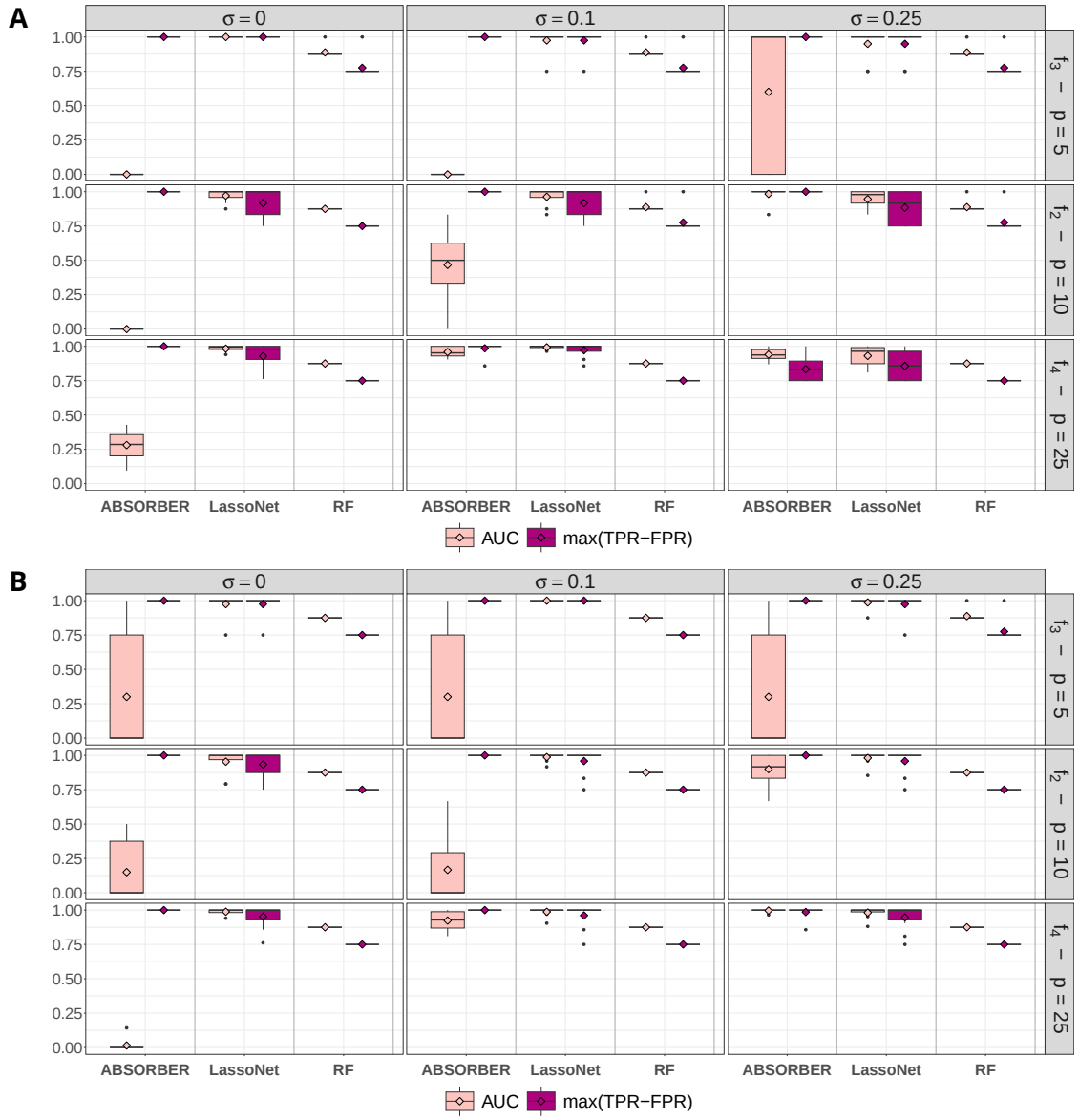

Figure 4.11: AUC and max$(\text{TPR} - \text{FPR})$ calculated for three different variable selection methods: ABSORBER, LassoNet and RandomForests (RF) applied to three functions (top to bottom) f_3, f_2 and f_4 with $n = 700$ (A) and $n = 2000$ (B) and an increasing value of σ (left to right). 10 random 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.

4.4. Application to a 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 min-

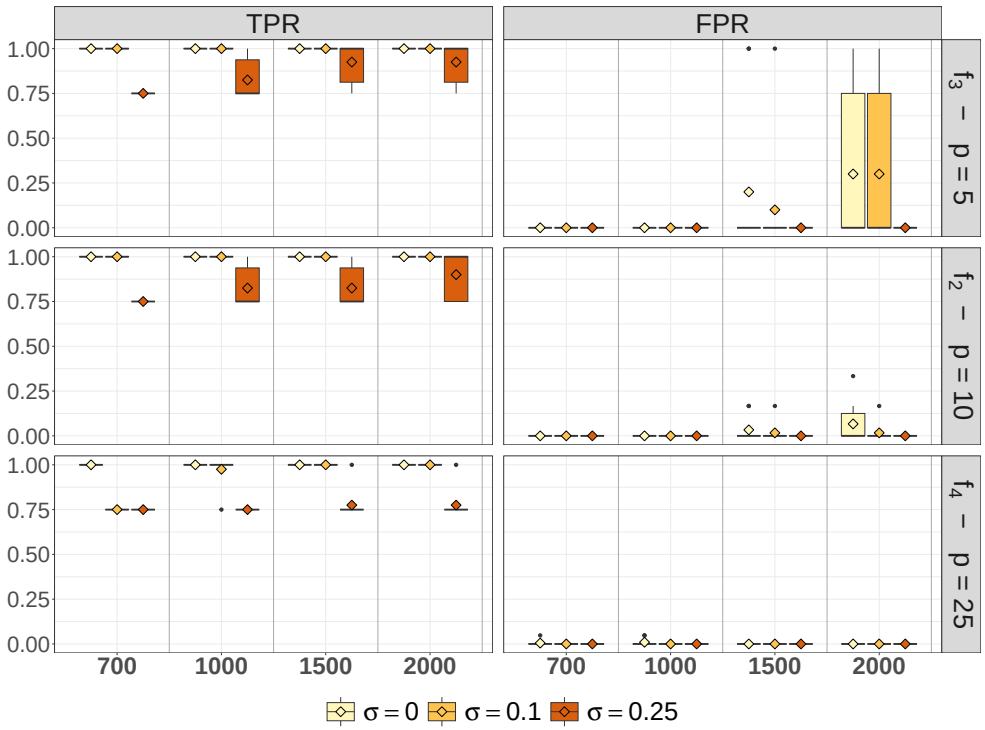

Figure 4.12: TPR(λ) and FPR(λ) values by choosing $\lambda = \lambda_{\text{AIC}}$ for f_3, f_2 and f_4 with three noise levels in the observation sets. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

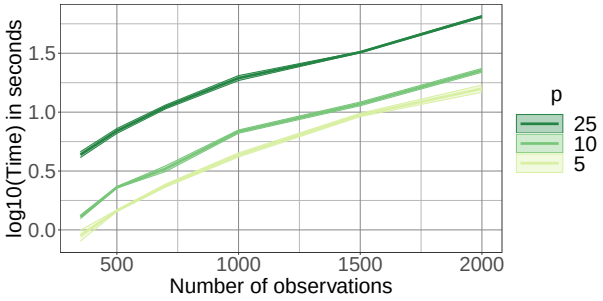

Figure 4.13: Execution times for an increasing number of observations n and values of p. The tested functions are here f_2, f_3 and f_4 with unnoisy observation sets ($\sigma = 0$).

eral components, 8 geochemical reactions and 2 mineral dissolution-precipitation reactions. For the purposes of this chapter, 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:

$$\begin{aligned}\text{Calcite} &= f_5(\text{C}^*, \text{Ca}^*, \text{K}^*, \text{Cl}^*, \text{Calcite}^*)\\ &= \widetilde{f}_5(\text{C}^*, \text{Ca}^*, \text{Calcite}^*),\end{aligned}$$

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_6 which takes into account known fake variables to study the behavior of all three methods:

$$\begin{aligned}\text{Calcite} &= f_6(\text{C}^*, \text{Ca}^*, \text{K}^*, \text{Cl}^*, \text{Calcite}^*, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)})\\ &= \widetilde{f}_5(\text{C}^*, \text{Ca}^*, \text{Calcite}^*),\end{aligned}$$

where $x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}$ are synthetic irrelevant variables obtained through uniform sampling between 0 and 1. Hereafter, $\mathcal{V}_{f_5} = \mathcal{V}_{f_6} = \{\text{C}^*, \text{Ca}^*, \text{Calcite}^*\}$, $\overline{\mathcal{V}_{f_5}} = \{\text{K}^*, \text{Cl}^*\}$ and $\overline{\mathcal{V}_{f_6}} = \{\text{K}^*, \text{Cl}^*, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}\}$.

The results for the application of our method, LassoNet and RF to f_5 and f_6 are displayed in Figure 4.14. Here, our method consistently selects the relevant variables belonging to $\mathcal{V}_{f_5}$ 65% of the time and the irrelevant ones belonging to $\overline{\mathcal{V}_{f_5}}$ 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 $\mathcal{V}_{f_5}$ 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 $\mathcal{V}_{f_6}$. However, it does not improve the results for LassoNet which continues to select all ten variables from $\mathcal{V}_{f_6}$ and $\overline{\mathcal{V}_{f_6}}$. 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 λ and to automatically choose the relevant variables. The corresponding results are shown in Figure 4.15. 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 and p.

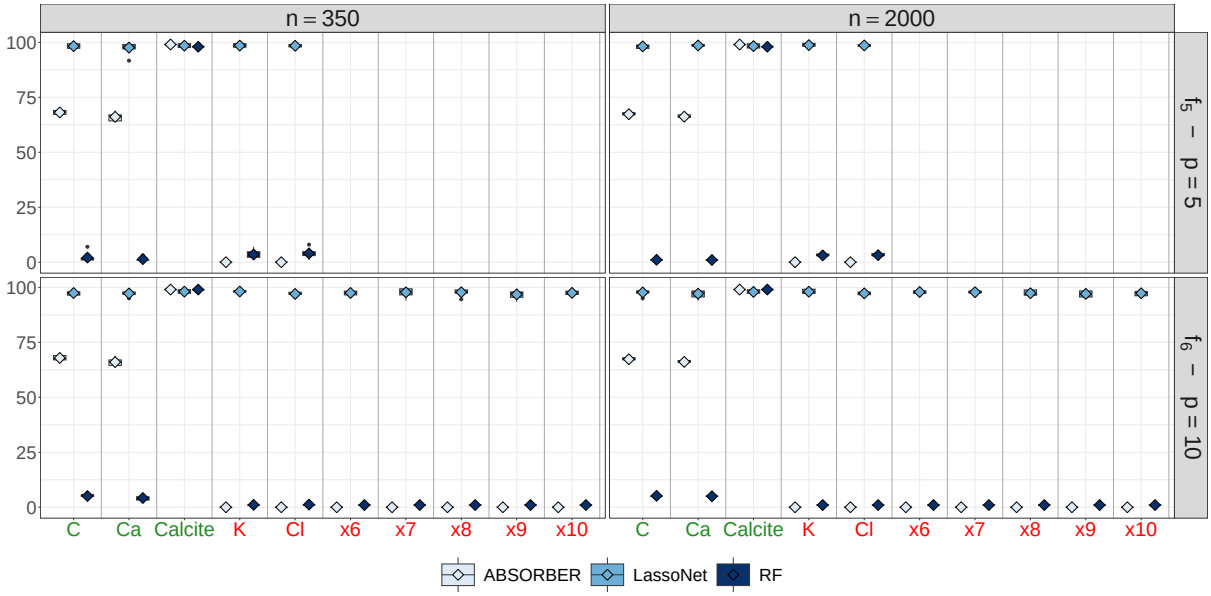

Figure 4.14: Percentage of selection of each variable of f_5 (top) and f_6 (bottom) with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

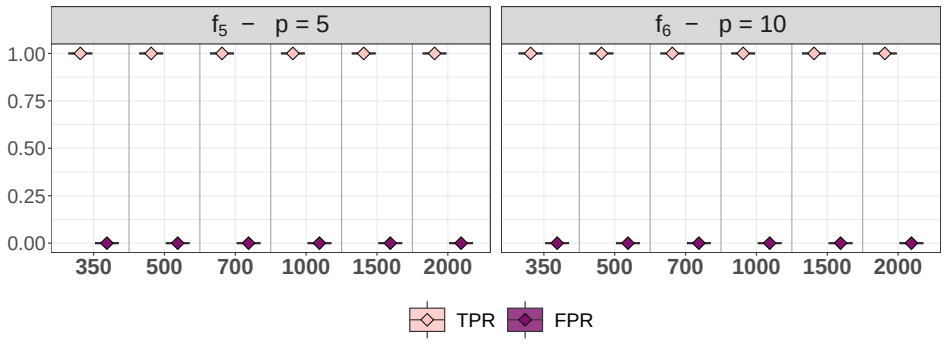

Figure 4.15: TPR(λ) and FPR(λ) values by choosing $\lambda = \lambda_{\text{AIC}}$ for f_5 (left) and f_6 (right). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

4.5. Appendix: Additional plots

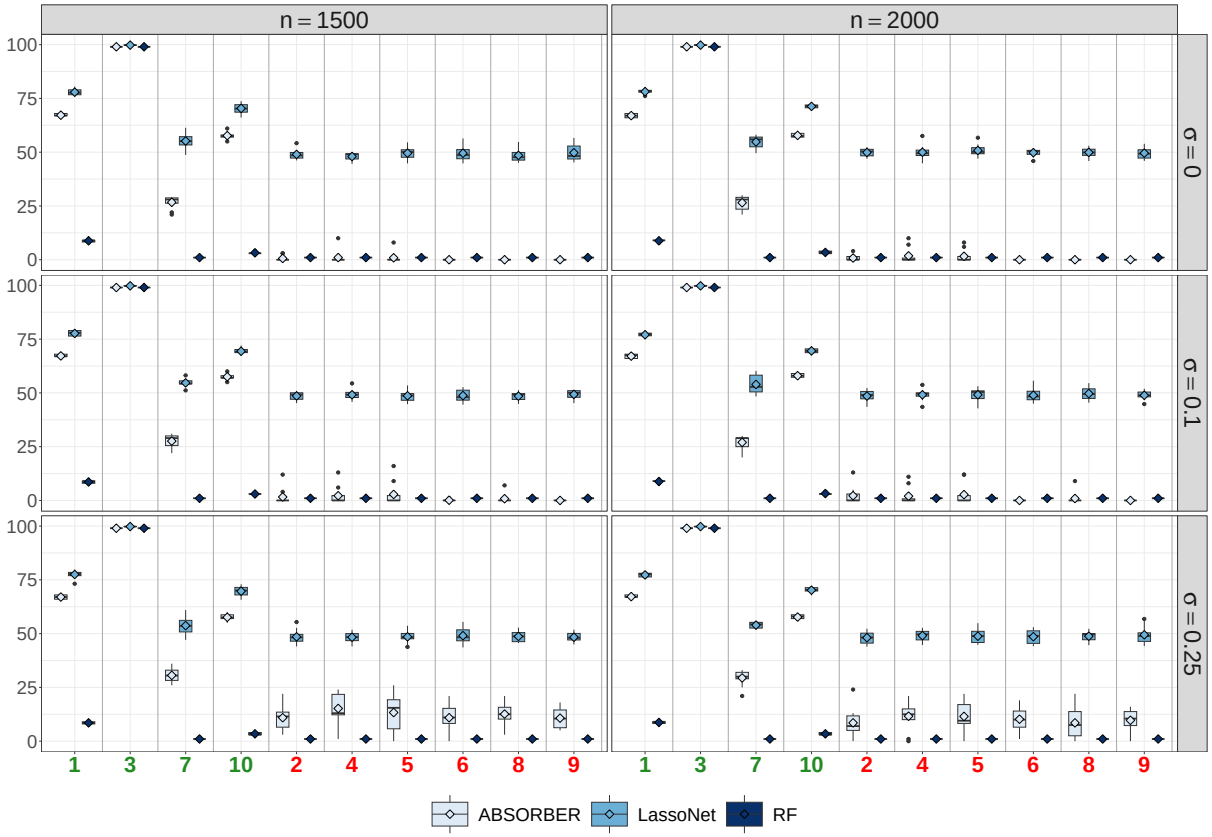

Figure 4.16: Percentage of selection of each variable of f_2 with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations n (left to right) and of the value of σ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

Chapter 5 - Applications of the developed methods on geochemical systems in the context of a EURAD work package

Publication

A part of the content of this chapter is the subject of the collaborative article:
N.I. Prasianakis, E. Laloy, D. Jacques, J.C.L. Meeussen, C. Tournassat, G.D. Miron, D. A. Kulik, A. Idiart, E. Demirer, E. Coene, B. Cochepin, M. Leconte, M. E. Savino, J. Samper II, M. De Lucia, S. V. Churakov, O. Kolditz, C. Yang, J. Samper, F. Claret. Geochemistry and Machine Learning: review of methods and benchmarking.
To be submitted.

Abstract

In this chapter, we present additional results using our variable selection method and the estimation methods presented in Chapter 2, 3 and 4 through my participation to a work package of the European Joint Programme on Radioactive Waste Management (EURAD). This collaborative work will be the subject of a scientific article to be submitted in the forthcoming months.

Table of contents

5.1. Context and geochemical systems

This section focuses on our involvement to a machine learning work package (WP) called Development and Improvement Of Numerical methods and Tools for modelling coupled processes (DONUT) which is a part of EURAD. This project consists in the collaboration of European teams focused on various machine learning approaches applied to geochemistry and reactive transport simulations in the context of radioactive waste management.
This WP is led by Francis Claret[1] and the benchmark exercise is coordinated by Nikolaos Prasianakis[2], which includes several collaborators contributing to its advancement. Among them, we can particularly cite: Diederik Jacques[3], J.C.L. Meeussen[4], Dmitrii Kulik[4], Eric Laloy[3], Andrès Idiart[5], Ersan Demirer[5], Emilie Coene[5], Javier Samper[6], Javier Samper II[6], Marco De Lucia[7], Benoit Cochepin[8], Marc Leconte[8] and Mary Savino[8,9].

The main objective of this benchmark is to assess the performance and accuracy of each proposed method, aiming to highlight their respective advantages as well as limitations. To do so, pre-defined geochemical systems with an increasing level of complexity are considered. Input variables belonging to a p-dimensional space were generated using a Latin Hypercube Sampling (LHS). Then, high-quality data were generated using powerful geochemical solvers such as ORCHESTRA, GEMS and Phreeqc (see Meeussen (2003); Kulik et al. (2013); Parkhurst and Appelo (2013), respectively, for more details on each solver). These datasets were used for training and validation purposes, allowing for the computation of statistical measures and facilitating comparisons between the different machine learning approaches. The geochemical context presented here concerns the hydration and evolution of cementitious systems under 25°C and takes as input variables the amounts of different oxides and the amount of water. Six cementitious systems are presented in the DONUT benchmark with an increasing number of input variables. In this work package, we proposed to use the active learning approach using Gaussian Processes (GP AL), as described in Chapter 2 of this thesis, comparing it to the non active learning approach (GP noAL) and to the methods employed by other teams. Hereafter, we will only display results for the most complex case involving 7 chemical elements as inputs. We seek at estimating a total of 44 output chemical elements which are related to aqueous species, solid phases and auxilliary variables. Since two of them consistently yielded only zero values during the dataset generation process, we will not include them in our subsequent applications. These elements are summarized in Table 5.1.

5.2. Application of the methods introduced in the previous chapters

In the following, we will mainly employ GP AL using the most relevant input variables selected with ABSORBER. We will then compare these results to those obtained by considering all

[1]BRGM, 3 Avenue Claude Guillemin, 45060 Orléans, France
[2]Laboratory for Waste Management, Paul Scherrer Institute, CH, 5232, Villigen PSI, Switzerland
[3]Engineered and Geosystems Analysis, Belgian Nuclear Research Centre, Belgium
[4]Nuclear Research and Consultancy Group (NRG), Petten, The Netherlands.
[5]AMPHOS 21 Consulting, S.L., Calle Venezuela, 103, 08019, Barcelona, Spain
[6]Centro de Investigaciones Científicas, ETS Ingenieros de Caminos, Universidade da Coruña, A Coruña, Spain
[7]Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany
[8]Andra, 1/7 Rue Jean Monnet, 92290, Chatenay-Malabry, France
[9]Universite Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France

7 input chemical elements	42 output chemical elements
CaO, SiO_2, Al_2O_3, SO_2, K_2O, CO_2, H_2O	• *total amount of elements in the dissolved form:* Ca_{aq}, Si_{aq}, Al_{aq}, S_{aq}, K_{aq}, C_{aq}, O_{aq}, H_{aq}, pH, • *total amount of elements in the solid form:* Ca_s, Si_s, Al_s, S_s, K_s, C_s, O_s, H_s, • *total amount of elements in the solid solution form:* Ca_{ss}, Si_{ss}, K_{ss}, H_2O_{ss}, V_{ss}, mET_{ss}, Al_ET_{ss}, Ca_ET_{ss}, S_ET_{ss}, C_ET_{ss}, $H_2O\_ET_{ss}$, • *total amount of minerals:* Portlandite, AmorfSi, Gibbsite, Katoite, Monosulfate, Hemicarbonate, Monocarbonate, Straetlingite, Chabazite, Calcite, Thaumasite, • *auxillary variables:* MassWater, mCSHQ, GelWater

Table 5.1: Geochemical system considered in this application.

input variables, presented in the collaborative article. Additionally, we propose an application of GLOBER-c with the selected variables and compare it to the results obtained for GP noAL and displayed in the article.

5.2.1. Variable selection with ABSORBER

Firstly, three sets of $3,500$ observations were sampled from the $50,000$ observation training table to apply ABSORBER. The corresponding results are depicted in Figure 5.1 for the percentage of selection of each variable and in Figure 5.2 which illustrates the selection using the AIC for the three different samplings. The first figure highlights that not all input variables hold the same importance in estimating the output. As evidenced by this figure, the solid phases (s) tend to depend on only one or a few variables whereas the aqueous phases (aq) appear to rely on nearly every input variable, as observed, for instance, for H_{aq}. Additionally, the solid solution (ss) output elements exhibit a similar dependence on input elements. Similarly, the "ET_{ss}" elements share the same relevant variables. The results in Figure 5.2 allow for a more straightforward discrimination between relevant and irrelevant variables, facilitating the selection process for estimation. To highlight the effect of variable selection, we will concentrate solely on output elements that depend on a maximum of three input variables, according to the results in Figure 5.2, which concerns 10 elements among the total 42. These 10 considered output elements and their corresponding selected variables are displayed in Table 5.2.

5.2.2. Validation of our selection method through a first application of GP AL

To demonstrate the efficiency of our variable selection method, we apply our active learning method to estimate the considered outputs. In this case, we initiate the process with 15 points randomly chosen among the training table and set the stopping criterion as a threshold on the number of observations, fixed at $1,000$. This estimation method is applied considering only the d relevant variables found with ABSORBER, and displayed in Table 5.2, or all of the $p = 7$ input variables of Table 5.1. Then, we compare the statistical measures defined in (2.18) and (2.19). We define a supplementary measure, the normalized Root Mean Square Error (RMSE) at the tth iteration of our active learning approach as:

$$\text{Normalized RMSE}(t) = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \mu_t(x_i)\right)^2}}{y_{max} - y_{min}}, \tag{5.1}$$

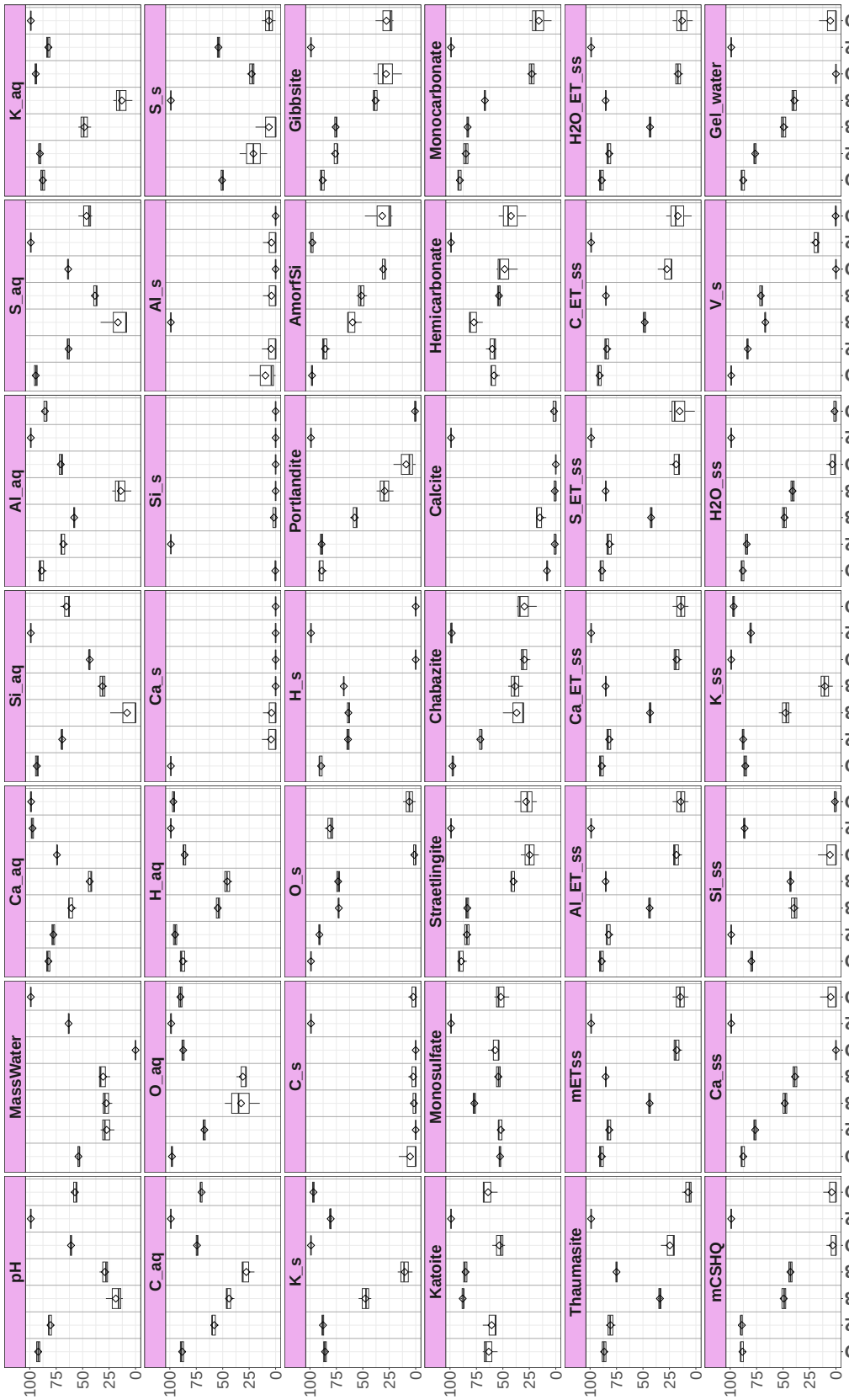

Figure 5.1: Percentage of selection using ABSORBER of each variable of the most complex geochemical system studied by the DONUT benchmark obtained with three independent samplings of the observation set.

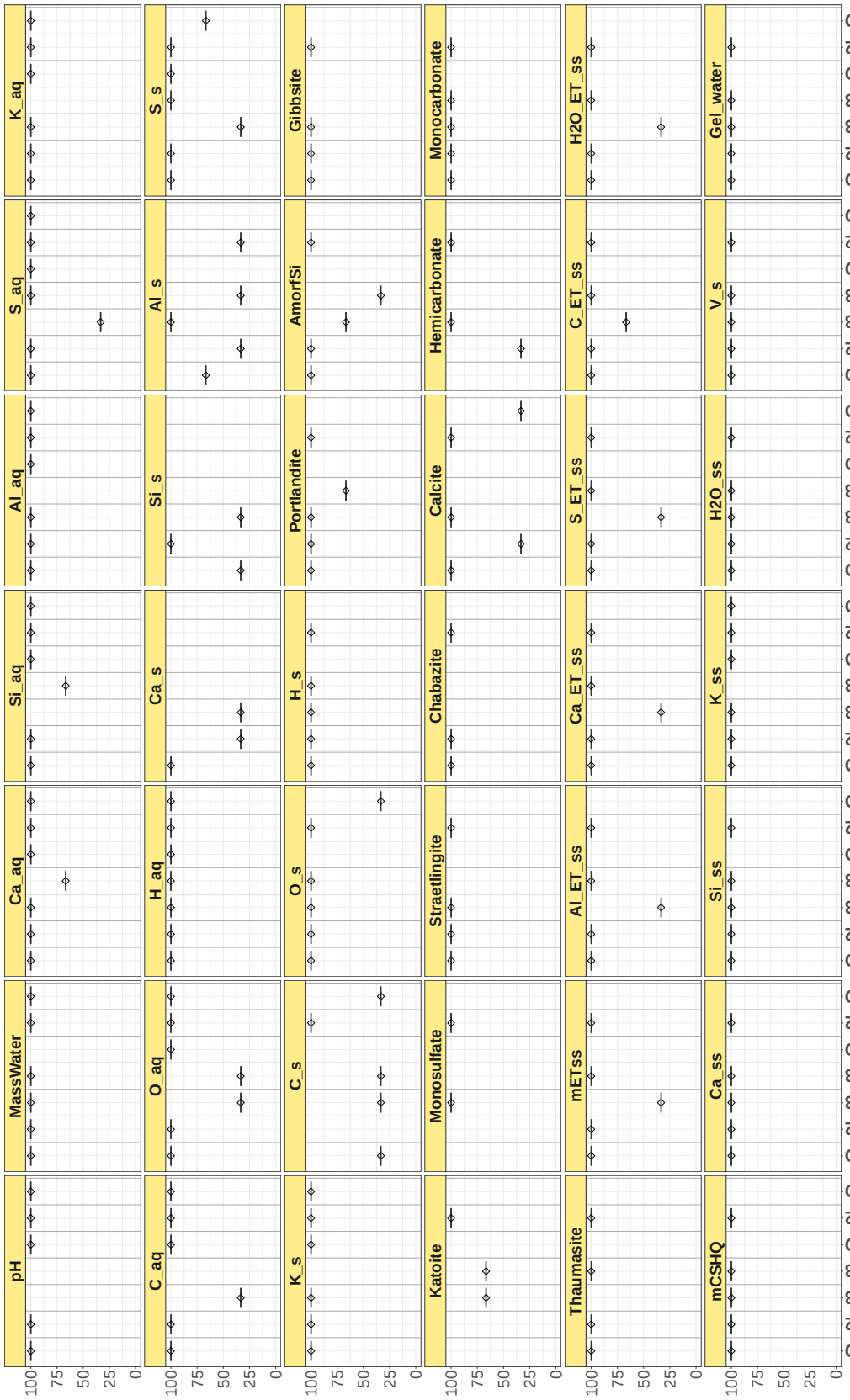

Figure 5.2: Percentage of selection of each variable using ABSORBER with AIC. The boxplots were obtained with three independent samplings of the observation set. Note that an absence of bar means that the corresponding variable is not selected.

Considered output	Selected variables with ABSORBER
Ca_s	CaO (1)
C_s	CO_2 (1)
Si_s	SiO_2 (1)
Al_s	Al_2O_3 (1)
Katoite	CO_2 (1)
Monosulfate	Al_2O_3, CO_2 (2)
Hemicarbonate	Al_2O_3, CO_2 (2)
AmorfSi	CaO, SiO_2, CO_2 (3)
Chabazite	CaO, SiO_2, CO_2 (3)
Calcite	CaO, Al_2O_3, CO_2 (3)

Table 5.2: Selected variables identified with ABSORBER for estimating the 10 output variables, based on the results in Figure 5.2. The number in brackets indicates the number of selected variables.

where y_i, μ_t, y_{max} and y_{min} are introduced in Chapter 2 after (2.18) and (2.19). The summarized results for seven different steps of the active learning process are presented in boxplots displayed in Figure 5.3 for the 10 outputs and the three different samplings of the observation set.

This figure demonstrates that integrating our variable selection method with the active learning approach, limited to a maximum of $1,000$ observations, significantly reduces the estimation mean error as evidenced by both the average and the extreme values of the normalized MAE and RMSE. Furthermore, it either maintains or decreases the maximum normalized error while also reducing the number of input dimensions, regardless the number of observations n. We can also conclude that increasing the number of observations beyond $1,000$ will not necessarily lead to different results, as the average values seem to reach a plateau for each measure and method. To illustrate this improvement throughout each step of active learning, the results for four output elements are shown in Figure 5.4. The top two plots display a significant enhancement in estimation error when using only one relevant variable, as suggested by our variable selection method. Moreover, as observed in the bottom two plots depicting results for Monosulfate and Chabazite, the normalized MAE and RMSE exhibit improvement when the input dimension is reduced to 2 and 3 variables selected by ABSORBER, respectively.

5.2.3. Application of GP AL and GLOBER and comparison to the other methods of the benchmark

Finally, we apply our GLOBER and GLOBER-c methods, defined in Chapter 3 and denoted in the figures as Glober $d \leq 3$, to estimate the 10 outputs studied in the previous section, using only the relevant variables suggested by ABSORBER and displayed in Table 5.2, and the three samplings of $3,500$ observations used for the variable selection step in Section 5.2.1. We compare the corresponding results to those obtained for the non-active learning approach of Gaussian Processes using the same observation sets, denoted GP noAL $d \leq 3$. It should be noted that since the solid outputs presented here depend only on one variable according to ABSORBER, these two approaches need very few observations to run correctly. Hence, we use only 100 observations for the total amount of elements in the solid phases (Ca_s, C_s, Si_s and

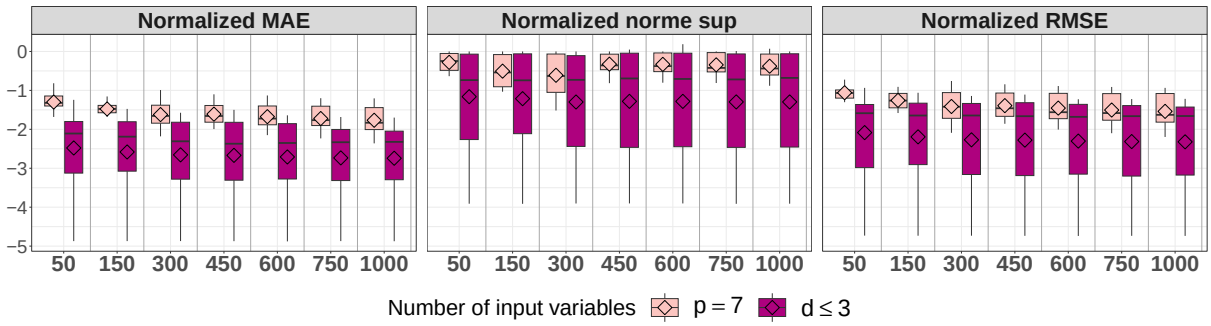

Figure 5.3: Global statistical measures obtained with ($d \leq 3$) or without ($p = 7$) a variable selection for the estimation of 10 outputs with 3 random samplings of the observation set. The method displayed here is the active learning approach of Gaussian Processes at 7 different steps of the process (7 values of n), reaching $1,000$ points at the end. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

Al_s).

In parallel, we apply the active learning approach initially selecting 15 randomly chosen points from the training dataset of $50,000$ points. To reduce the execution time, the stopping criterion was chosen to be $1,000$ observation points. This method is applied using all of the 7 variables or only the most relevant ones selected with ABSORBER. We can refer in the results to GP AL $p = 7$ and GP AL $d \leq 3$, respectively. We compare all these approaches to two other teams of the DONUT benchmark, namely A21 and SCK. These two teams have developed two distinct Deep Neural Network architectures, both trained with the same tables. Hereafter, we will present the results using the training table containing $5,000$ points for these two methods (A21 $p = 7$ and SCK $p = 7$). Hereafter, we considered the normalized MAE and the normalized RMSE as statistical measures. All of these outcomes are depicted in Figure 5.5.

From this figure, we can see that employing the active learning approach while only selecting the relevant variables, results in a global reduction in estimation error, as evidenced by the comparison between the yellow and pink bars. Moreover, Glober $d \leq 3$ (red bars) exhibits improved results compared to GP noAL $d \leq 3$ (orange bars) for 7 outputs out of the total of 10 using the same observation set, highlighting the advantage of employing this method. Although SCK seems to outperform the other methods on half of the outputs with $5,000$ training points, A21 demonstrates corrupted results, especially when dealing with a large number of zero values in the training tables (Katoite, Monosulfate, AmorfSi). Furthermore, Glober $d \leq 3$ and GP AL $d \leq 3$ show improved results for the estimation of solid outputs compared to the other methods with solely $n = 100$ and $n = 1000$ observations, respectively. Overall, opting for $d \leq 3$ leads to enhanced precision in estimating geochemical outputs.

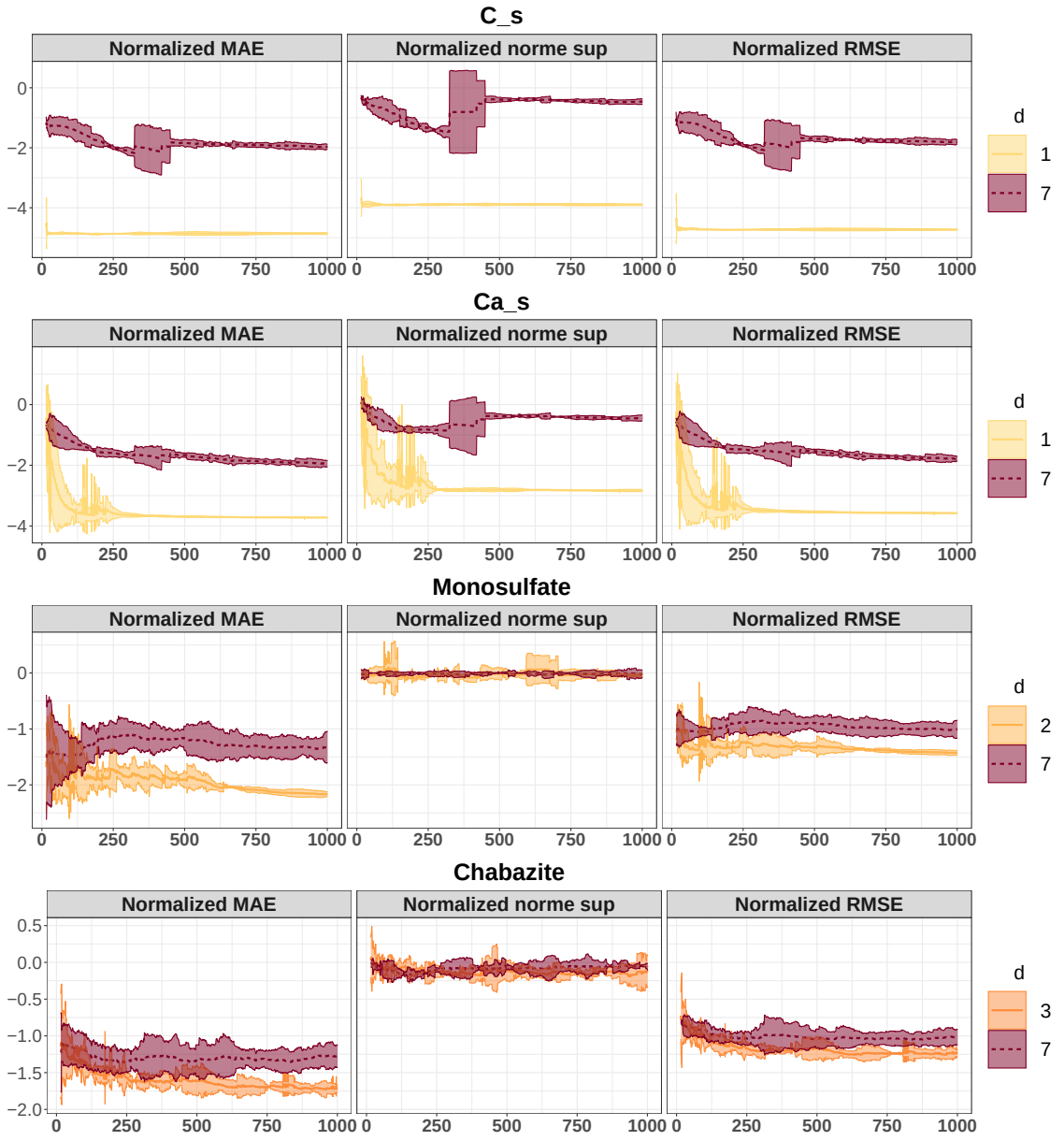

Figure 5.4: Average and standard deviation of different statistical measures for the active learning approach to estimate the considered outputs with the d input variables selected with ABSORBER or all of the $p = 7$ variables. Values are $\log_{10}$-transformed for each measure.

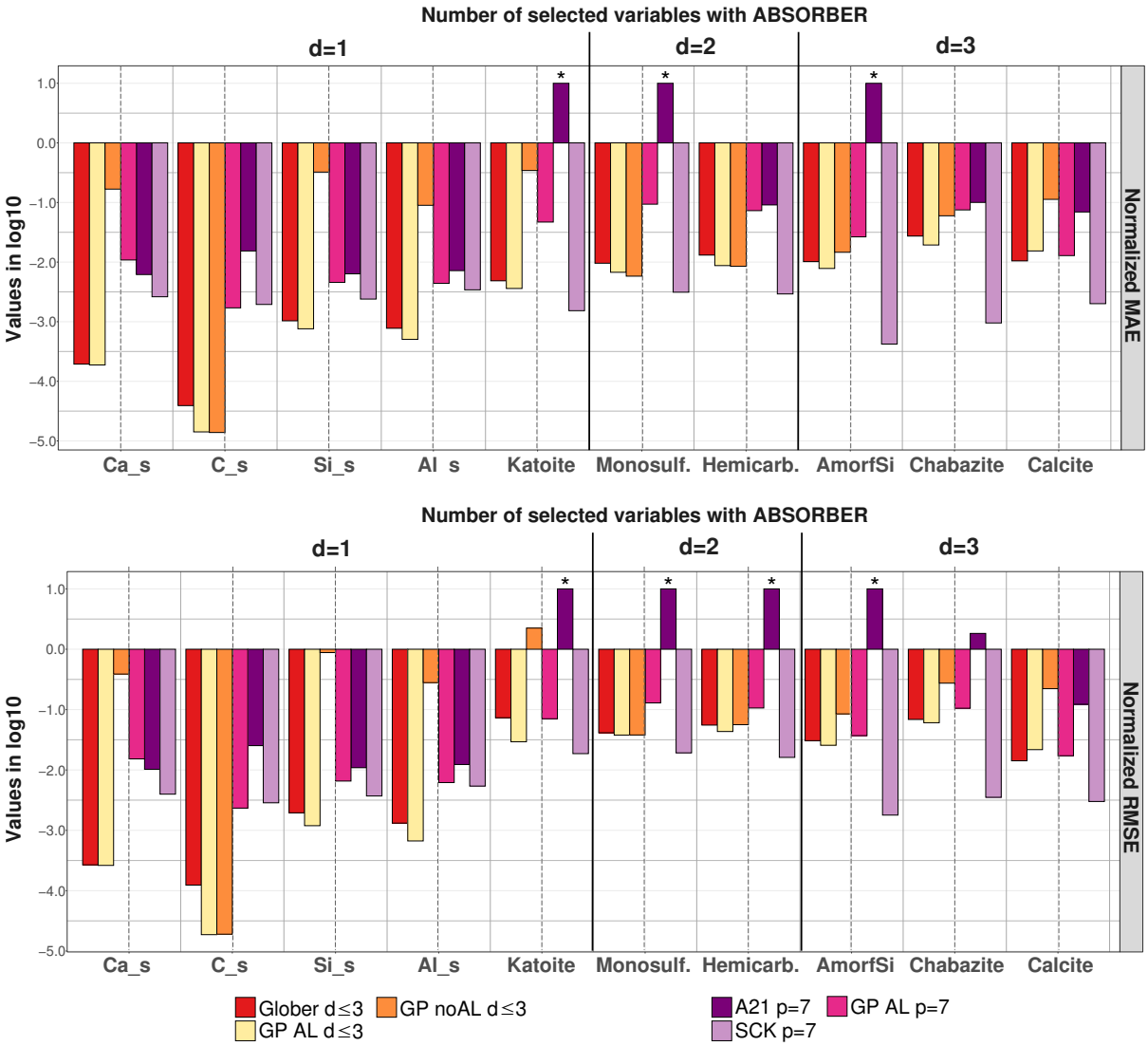

Figure 5.5: Statistical measures obtained for six distinct methods with ($d \leq 3$) or without ($p = 7$) the variable selection step prior to the estimation of 10 output variables. Values are $\log_{10}$-transformed for each measure and are obtained using a varying number of observations. The $*$ label indicates a $\log_{10}$ value greater than 1. Monosulf. and Hemicarb. stand for Monosulfate and Hemicarbonate, respectively.

Chapter 6 - Conclusion and Perspectives

6.1. Summary of the developed methods

In this PhD thesis, we developed function estimation methods as well as a variable selection approach in nonparametric multivariate settings applied to geochemical systems.

Firstly, we proposed a novel sequential data-driven method involving an active learning approach for estimating functions in a nonparametric and multivariate framework. Using a Bayesian perspective, the function to estimate is considered to be a sample of a Gaussian process which allows for the computation of the global uncertainty on the function estimation. Thanks to this property, the proposed method sequentially selects the most relevant observation points at which the function to estimate has to be evaluated. This method showed satisfactory estimation accuracy while using a limited number of observations for the estimation of functions depicting geochemical reactions.

Subsequently, we proposed another approach called GLOBER to estimate functions in multivariate nonparametric regression settings based on an adaptive knot selection for B-splines. The underlying idea for selecting the knots is to apply the generalized lasso, since the knots can be seen as changes in the derivatives of the function to be estimated. Numerical experiments were conducted to assess the sensitivity of our approach to noise levels and sampling. Additionally, comparisons to state-of-the art approaches through their applications on geochemical reactions demonstrated superior results for our method.

Furthermore, we introduced ABSORBER, a novel variable selection method for multivariate nonparametric regression models developed to identify only the variables on which the regression function truly depends. This approach consists in considering the function to estimate as a linear combination of B-splines and their pairwise interaction terms. The coefficients of this linear combination are estimated by minimizing a least-squares criterion penalized by the ℓ_2-norms of the partial derivatives with respect to each variable on which the function depends. We showed that this criterion can be rewritten as a Group Lasso problem, allowing the selection of relevant variables for which the corresponding estimated coefficients are nonzero. This approach was validated through numerical experiments and compared to state-of-the-art methods on a geochemical reaction.

Finally, we proposed to combine ABSORBER and the two previously described function estimation methods in the context of the DONUT work package of the EURAD European project. We showed that our method outperformed the approaches of two other teams contributing to the work package, especially when the functions appeared to depend solely on a single variable according to ABSORBER.

In the forthcoming section, we will discuss some perspectives for further continuation of this research.

6.2. Future work

6.2.1. Improving function estimation involved in geochemical applications with physics-informed approaches

In the presented geochemical systems, we considered equilibrium-based reactions without employing physics knowledge. One possible future direction could be the development of physics-informed aproaches to enhance the performance of our function estimation methods. Recently, interesting advances have been achieved using physics-informed machine learning methods. Specifically, Raissi et al. (2019) introduced Physics-Informed Neural Networks (PINNs) which have gained interest due to their ability to integrate information from both data and physics models, for instance by introducing a constraint on the function to estimate which has to be a solution of a given partial differential equation. Despite offering more interpretable solutions for DNNs while using shallower and simpler architectures, PINNs can still exhibit limitations. These include being computationally expensive, particularly when dealing with complex loss functions, which can result in highly non-convex optimization problems (Karniadakis et al., 2021).

To circumvent these issues, we can consider extending the physics-informed approach to other nonparametric estimation methods in nonlinear regression models, such as the k-nearest neighbors (kNN) algorithm. For instance, in the context of deep geological facilities for high-level waste management, Hu et al. (2024) used a physics-informed kNN approach for thermal-hydraulic modelling, leading to improved results compared to other data-driven state-of-the-art machine learning methods, including simple kNN.

In the geochemical context, De Lucia and Kühn (2021) proposed a physics-informed approach by using a mass balance tolerance threshold to either accept or reject the estimation of the output variables obtained with a gradient boosting-based surrogate model. They also introduced a decision-tree method that employs new input variables integrating equilibrium constants information.

Inspired by this idea, we could consider embedding a physics constraint in the presented nonparametric regression methods such as GLOBER developed in Chapter 3. For instance, we could integrate a mass conservation constraint in the procedure to estimate the coefficients of the B-spline terms. More generally, we could imagine to include this kind of physics-constraint in any statistical learning approach that could be used to estimate the unknown function in a multivariate nonparametric regression model.

6.2.2. Reactive Transport Modelling (RTM) applications leveraging the developed approaches

In this thesis, we presented geochemical applications to solve chemical equilibrium reactions. The next step would be to apply the developed methods to a reactive transport model depicting water-rock interactions, using an operator splitting approach to handle the transport phase and the chemical reactions separately. This approach would estimate the amounts of elements involved in chemical reactions at each time iteration using our function estimation methods as surrogate models.

A first application in the context of the radioactive waste storage would be a 1D reactive transport geochemical system based on Kolditz et al. (2012, p.313–314). In this case, we aim at simulating a calcite precipitation system involving multiple reactions, including precipitation-dissolution of minerals, as described in Kolditz et al. (2012, Figure 15.2). Consequently, a major

challenge is to precisely estimate the amounts of elements C, Ca, Cl, Mg as well as the minerals calcite and dolomite, at each time step of the reaction transport and at each position of the considered domain.

In this direction, we attempted to use the active learning approach based on Gaussian Processes (GP AL) proposed in Chapter 2. The method exhibited an estimation accuracy which should be further improved to provide satisfactory results in the context of RTM. As a potential way of improvement, we propose applying our variable selection approach ABSORBER to take into account only the relevant variables in the estimation of the amount of each element. As we can see from Figure 6.1 which displays the results obtained with ABSORBER, some variables such as the amounts of Cl, Mg and calcite indeed depend only on at most half the input variables, thus drastically decreasing the model's complexity.

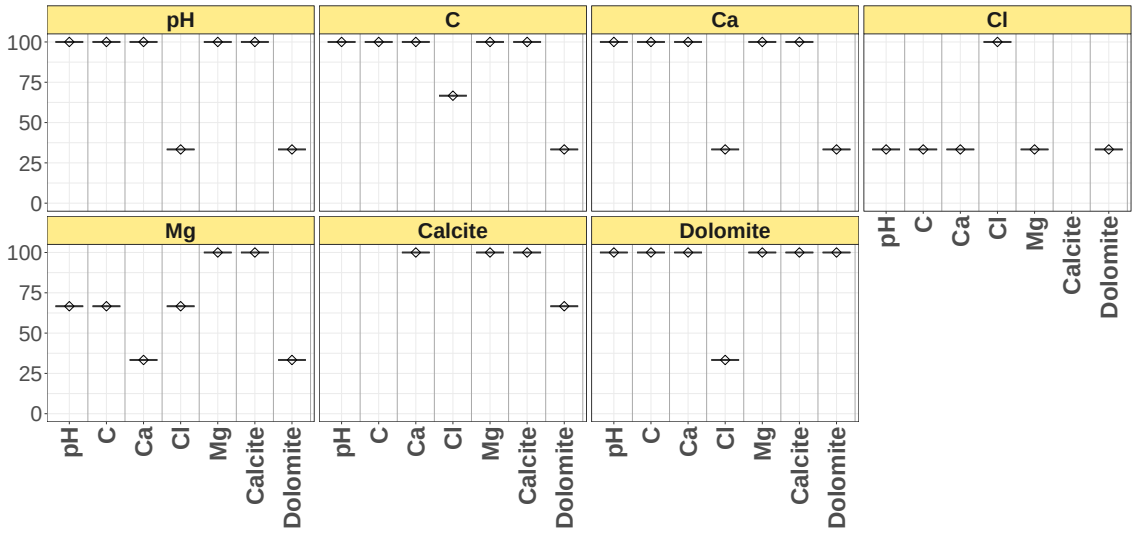

Figure 6.1: Percentage of selection using ABSORBER of each variable of the amount of elements to estimate in the studied RTM by employing the AIC. The boxplots were obtained with three independent samplings of $3,500$ points in the observation set. Note that an absence of bar means that the corresponding variable is not selected.

Thus, the resulting approach that could be considered for performing RTMs would be the one described in Figure 6.2.

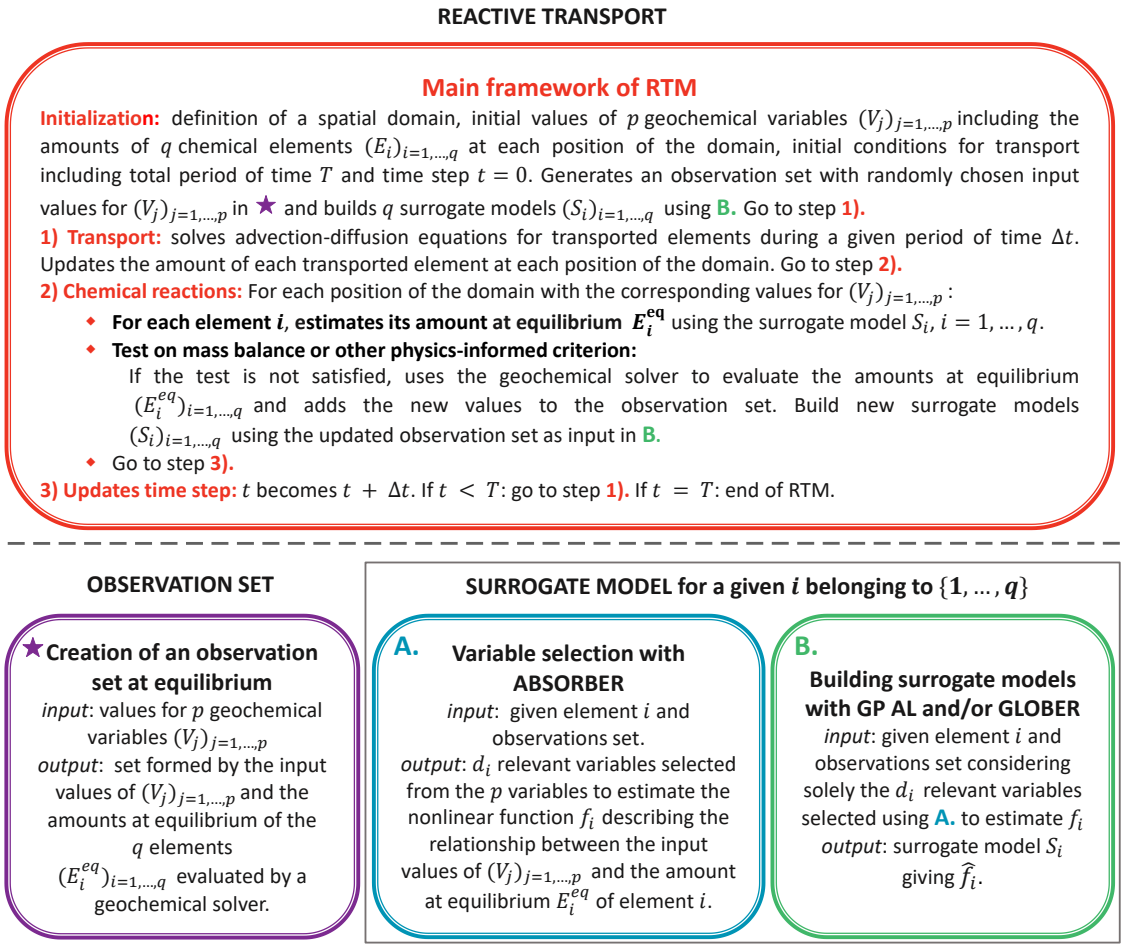

Figure 6.2: Integration our developed methods in the context of RTM to estimate the amount at equilibrium of q elements after each step of reactive transport, including the transport phase and chemical reactions. The RTM is performed over a total period of time of T.

Chapter 7 - En bref

7.1. Contexte géochimique

7.1.1. Le project Cigéo

La thèse ici présentée a été financée par l'Agence nationale pour la gestion des déchets radioactifs (Andra) dont l'objectif principal est d'après leur site officiel de :

> « Trouver, mettre en œuvre et garantir des solutions de gestion sûres pour l'ensemble des déchets radioactifs français afin de protéger les générations présentes et futures du risque que présentent ces déchets. »

Ces déchets sont produits par le biais de multiples activités humaines dont principalement par les groupes électro-nucléaires, comme indiqué dans la Figure 7.1A. D'après le code de l'environnement inscrit dans la legislation française, les déchets radioactifs sont définis comme étant des substances sans aucune utilité contenant des radionucléides, qu'ils soient d'origine anthropique ou non. Les radionucléides sont des noyaux instables présentant un excédent en neutrons ou en protons, ce qui leur confère un surplus d'énergie se manifestant par l'émission de rayonnements ionisants. De ce fait, les déchets radioactifs sont caractérisés par leur niveau de radioactivité ainsi que par leur temps de demi-vie, qui correspond au temps nécessaire afin de diviser par deux le nombre de radionucléides. Par conséquent, leur activité ainsi que leur concentration nécessitent un contrôle de radioprotection. En France, ces matériaux radioactifs comptent pour un volume total de $1.760.000$[1] mètres cubes et sont catégorisés dans différents groupes selon leur niveau de radioactivité et le temps de demi-vie, comme illustré dans la Figure 7.1B. Cette classification permet une meilleure gestion de ces déchets à travers le développement d'infrastructures spécialisées. Par exemple, les déchets de très faible activité (TFA) et de faible et moyenne activité à durée de vie courte (FMA-VC) associés à des temps de demi-vie de moins de 30 ans, sont stockés dans des installations de stockage en surface, situées pour ces derniers dans l'Aube et dans la Manche[1]. Cependant, en ce qui concerne les déchets de faible activité à vie longue (FA-VL), de moyenne activité à vie longue (MA-VL) et de haute activité (HA), des projets d'installations de stockage appropriées sont actuellement en cours d'étude.

Plus particulièrement, l'Andra a entrepris un projet considérable connu sous le nom de Centre industriel de stockage géologique (Cigéo) dont le principal objectif est de fournir une solution à long-terme pour le stockage des déchets MA-VL et HA. En effet, ces catégories représentent seulement 3% de la quantité totale de déchets radioactifs, mais plus de 99% de la radioactivité totale. C'est pourquoi ce projet constitue un défi majeur et fait l'objet d'études intensives mobilisant ainsi un grand nombre de scientifiques et d'ingénieurs depuis plus de 30 ans. Ce projet prévoit la création d'une infrastructure géologique de stockage à 500 mètres de profondeur dans une zone de 250 kilomètres carrés localisée entre les départements de la Meuse et la Haute-Marne, comme le montre la Figure 7.1C. Ce site a été sélectionné en raison de ses caractéristiques géologiques qui lui confèrent des propriétés très favorables au confinement des radionucléides et à la prévention de leur dispersion dans l'environnement.

[1] *Inventaire national des matières et déchets radioactifs – Les essentiels 2023*, disponible en ligne à https://inventaire.andra.fr.

De plus, sa profondeur offre une protection contre les phénomènes naturels de surface tels que l'érosion et la glaciation, mais aussi contre d'éventuels aléas anthropiques.

Le projet Cigéo a connu plusieurs phases depuis sa création, allant des études conceptuelles et techniques aux multiples débats publics en passant par les nombreux rapports techniques sur la sûreté, la sécurité et la récupérabilité. Récemment, une demande d'autorisation de création a été déposée en 2023 dans le but d'aboutir à un décret permettant le début des travaux de construction des installations de Cigéo. Pour plus de détails sur la progression de ce projet, nous renvoyons le lecteur aux principales dates clés présentées dans l'Appendice 8.2.

Puisque le projet implique le confinement de substances radioactives dans une couche d'une formation géologique particulière appelée Cavollo-Oxfordien, des études expérimentales ainsi que des simulations numériques sont indispensables afin de démontrer la qualité et la sûreté du projet avant d'entreprendre toute construction. Pour cela, un laboratoire expérimental situé sur le site-même du projet à Bure permet d'étudier de manière exhaustive les caractéristiques physiques et géochimiques des composants géologiques. De plus, d'autres travaux expérimentaux réalisés par des membres de la communauté scientifique ont évalué l'aptitude de certains matériaux rocheux à être utilisés dans ce cadre-ci. Par exemple, Moyce et al. (2014) ont étudié l'altération des roches environnantes dans des eaux de ciment alcalines sur une période de 15 ans. En outre, Fernández et al. (2018) ont analysé la formation de silicates de magnésium dans un type spécifique d'argile qui induit potentiellement divers changements chimiques dans les roches, notamment une réduction de leur porosité ce qui peut affecter le transport des fluides au sein de celle-ci. Dans le cadre de cette étude-ci, les expériences ont été réalisées pour étudier les interactions entre du béton et une roche argileuse pendant 13 et 10 ans, respectivement *in situ* et en laboratoire. Dans le contexte de Cigéo, puisque la demi-vie de certains déchets radioactifs peut s'étendre à des échelles de temps de l'ordre des centaines de milliers d'années, une liste exhaustive d'exigences et de mesures de sécurité doit être respectée sur ces périodes de temps données. Ces exemples témoignent du temps considérable nécessaire à l'obtention de tels résultats et rend indispensable l'emploi des simulations numériques comme outils complémentaires pour l'examen du comportement des roches sous divers scénarios et paramètres physico-chimiques. Ainsi, ces outils facilitent l'évaluation des performances des différents composants du site de stockage, parmi d'autres utilités.

7.1.2. Introduction de modèles de substitution pour la modélisation de transports réactifs

7.1.2.1. Modélisation de transports réactifs

Une partie des simulations réalisées vise à étudier l'évolution des milieux géologiques contenant les déchets radioactifs encapsulés, sur plusieurs années et sous des contraintes physiques et hydrologiques. En effet, les roches environnantes peuvent subir des microfractures, ce qui peut conduire à l'infiltration d'eau et à la dégradation de la structure géologique. Un exemple de cas d'étude dans ce contexte est la simulation de l'évolution chimique des différents composants des matériaux géologiques soumis à la précipitation et la dissolution minéralogique au cours d'un transport d'une eau souterraine. L'analyse complète de l'écoulement des fluides, des réactions géochimiques, du transfert de chaleur et du transport des solutés est communément appelé modélisation de transports réactifs ou *reactive transport modelling* (RTM). Steefel et al. (2015) ont proposé une introduction approfondie à la RTM et à son large éventail d'applications. Dans de tels scénarios, la modélisation de la physique et de la chimie implique le couplage d'équations différentielles partielles pour le transport avec des

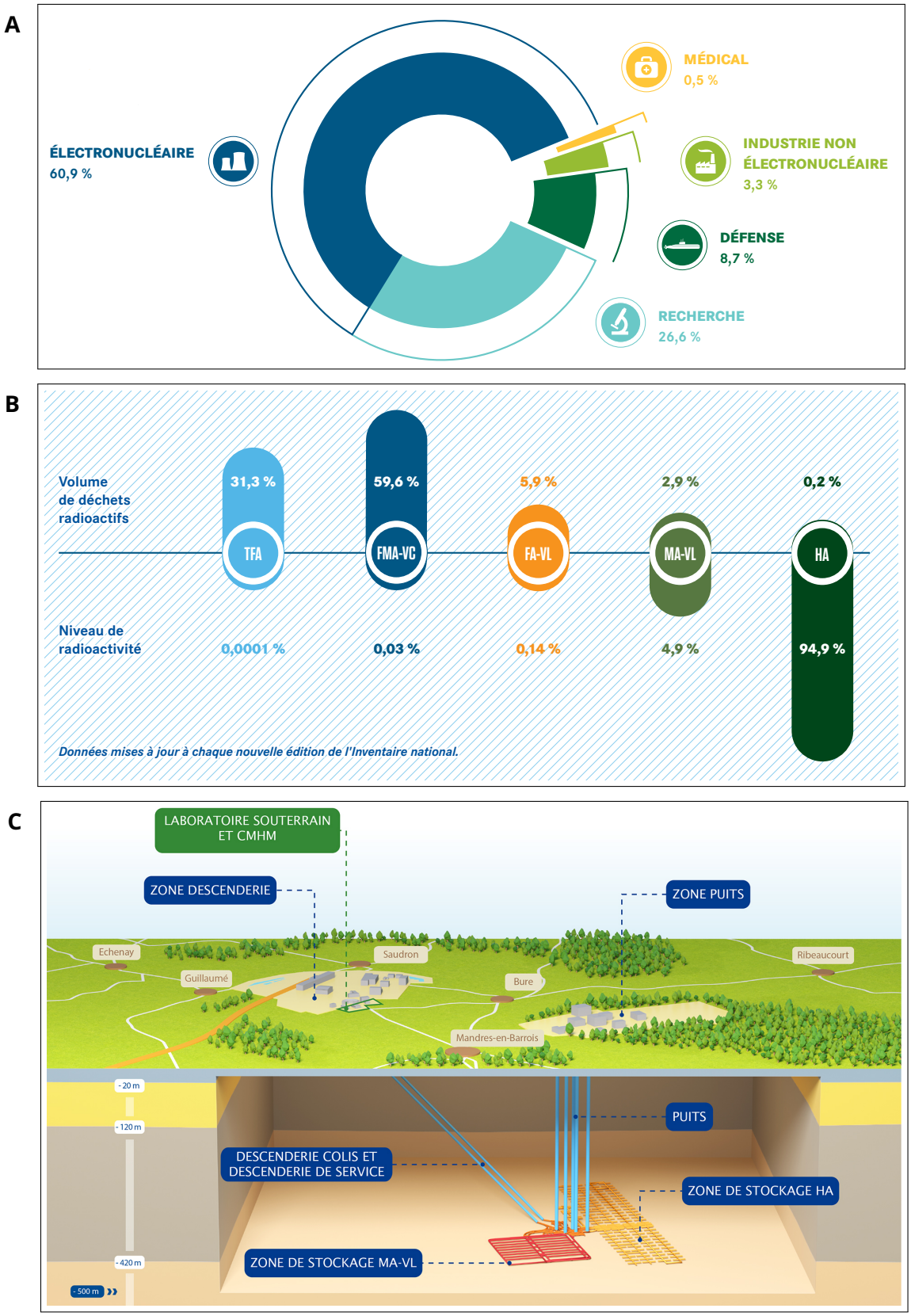

Figure 7.1: (A) : Principaux secteurs producteurs de déchets radioactifs en France. (B) : Classification des déchets radioactifs en français. Déchets TFA : de très faible activité, FMA-VC : de faible et moyenne activité à vie courte, FA-VL : de faible activité à vie longue, MA-VL : de moyenne activité à vie longue, HA : de haute activité. (C) : Cigéo, projet de stockage des déchets MA-VL et HA.

équations algébriques pour les réactions chimiques. Ce couplage représente un défi majeur lorsque la résolution des équations devient difficile, notamment due à une augmentation de la complexité du système (de Capitani and Brown (1987); Yeh and Tripathi (1989)).

Afin de faciliter l'implémentation de tels systèmes, une approche courante consiste à considérer séparément les réactions chimiques et le transport. Cela implique d'utiliser une méthode de décomposition par opérateurs ou *operator splitting*, où la résolution de l'équation de dispersion-advection discrétisée est d'abord réalisée, suivie par celle des équations de réaction chimique discrétisées. Par le biais de cette méthode, la modélisation du transport réactif peut être gérée à l'aide de deux solveurs différents. Bien que les solutions obtenues par cette décomposition par opérateurs aient pu entraîner des erreurs numériques dans le passé (Valocchi and Malmstead (1992); Barry et al. (1997); Simpson and Landman (2007)), cette méthode peut offrir une grande flexibilité et simplicité d'usage (Steefel and McQuarrie, 1996). Un ensemble important de travaux s'est concentré sur cette méthode dans le but de résoudre des problèmes non-linéaires de transport réactif et a comparé différentes stratégies considérées pour l'approche de décomposition par opérateurs, telles que les schémas séquentiels itératifs ou non-itératifs (Carrayrou et al., 2004). D'importantes contributions dans ce domaine ont étudié le taux de convergence de ces stratégies (Kanney et al., 2003) et ont proposé des algorithmes améliorés pour augmenter leur efficacité computationnelle (Lagneau and van Der Lee, 2010). De plus, Steefel et al. (2015) ont mené une étude comparative exhaustive sur des codes disponibles développés par la communauté géochimique en utilisant notamment l'approche de décomposition par opérateurs. Nous recommandons également au lecteur de se référer à cet article pour d'autres références et à Steefel (2019) pour les avancées plus récentes sur les RTMs. Malgré les efforts récents pour améliorer la précision de l'approche de décomposition par opérateurs, celle-ci reste confrontée à des défis computationnels (Lu et al., 2022).

Dans le contexte du stockage des déchets radioactifs, plusieurs équipes ont étudié les approches RTM appliquées aux roches environnantes, aux matériaux cimentaires encapsulant les déchets radioactifs et à leurs interactions avec l'environnement. Par exemple, des études menées par Kosakowski and Berner (2013), Samper et al. (2016) et Wilson et al. (2018) ont exploré les simulations de transports réactifs dans le cadre des programmes de gestion de déchets radioactifs respectivement en Suisse, Espagne et au Royaume-Uni. De plus, une étude comparative réalisée dans le contexte d'un projet européen collaboratif, établie par Idiart et al. (2020), a prouvé que la simulation du transport réactif avec des solveurs chimiques et des codes de transport peut effectivement permettre l'étude de l'interaction à long terme entre le béton hydraté, dans lequel les déchets radioactifs sont stockés, et la couche d'argile environnante. Ainsi, ce benchmark a démontré que la RTM est un outil puissant pour l'évaluation de la sûreté d'un projet de stockage de déchets radioactifs en milieu géologique, tel que Cigéo.

Néanmoins, puisque les simulations actuelles visent à représenter des systèmes réels de plus en plus complexes, elles ont tendance à exiger des coûts computationnels non-négligeables. La résolution des réactions chimiques est reconnue comme étant le facteur limitant dans l'implémentation de tels systèmes de transport réactif, en particulier si l'on considère un nombre important de variables d'entrée et de sortie potentielles (voir Appelo and Postma (2004) pour quelques exemples). Malgré de récentes améliorations et une augmentation significative des ressources de calcul et matérielles par l'utilisation d'ordinateurs très puissants, la résolution de modèles tri-dimensionnels à grande échelle de RTM complexes considérant de nombreux pas de temps reste à ce jour un défi important.

7.1.2.2. Modèles de substitution

L'émergence de ce que l'on appelle communément les modèles de substitution a introduit une solution alternative à la simplification excessive des modèles géochimiques considérés qui est aujourd'hui nécessaire pour rendre ces simulations réalisables. L'idée est simple : résoudre explicitement les systèmes de transport puis, sur la base d'un ensemble réduit de solutions calculées par le solveur, utiliser ces modèles de substitution, également connus sous le nom de *proxy models*, pour estimer les solutions réelles de modèles souffrant d'un temps de calcul élevé. Cette approche permet de réduire considérablement le temps de calcul tout en assurant une grande précision de l'état d'équilibre calculé. Le concept de modèles de substitution et leurs diverses applications ont été largement discutés dans la littérature afin de faciliter leur utilisation. Nous renvoyons le lecteur à l'ouvrage de Forrester et al. (2008) qui fournit un aperçu complet des modèles de substitution appliqués à de l'ingénierie et couvre notamment les modèles polynomiaux, les modèles de fonction de base radiale, le krigeage et les machines à vecteur de support. Dans le domaine des géosciences, une revue plus récente réalisée par Asher et al. (2015) décrit les principales catégories de modèles de substitution appliqués à la modélisation des eaux souterraines. Ils notent qu'aucune famille de méthode ne surpasse universellement les autres, car son efficacité dépend strictement de son application.

Parmi toutes les méthodes existantes, les approches d'apprentissage automatique ou dites de *Machine Learning* (ML) se distinguent comme étant des méthodes basées sur des données qui ont suscité un intérêt considérable au cours des dernières décennies. Leur capacité à capturer avec précision les complexités non-linéaires inhérentes aux problèmes du monde réel en a fait des outils d'estimation très performants. Une revue de Razavi et al. (2012) dans le contexte de la simulation des ressources en eau s'est intéressée à un large éventail de modèles de substitution statistiques et de ML, démontrant une performance de calcul et des économies de temps substantielles, atteignant parfois jusqu'à 97% de réduction de temps CPU. Des travaux plus récents entrepris par Jatnieks et al. (2016) ont comparé un total de 32 méthodes statistiques et de ML pour prédire 7 variables de sortie impliquées dans les RTMs. Les auteurs ont déterminé que les réseaux de neurones régularisés bayésiens étaient la méthode de ML la plus efficace dans le contexte considéré, présentant des erreurs de prédiction plus faibles. Cette méthode a ensuite été utilisée pour simuler un système de transport réactif, produisant des estimations précises des solutions obtenues par les méthodes traditionnelles de simulations de RTM.

L'utilisation de modèles de substitution dans les simulations de transport réactif pour la gestion des déchets radioactifs est une approche relativement nouvelle. Dans le cadre de travaux récents, les réseaux de neurones artificiels (ANN) ont été employés comme substituts aux solveurs traditionnels de réactions géochimiques. Par exemple, Guérillot and Bruyelle (2020) ont utilisé un perceptron multicouche composé d'une couche cachée pour simuler le transport réactif dans le cadre de la modélisation d'une injection de CO_2 dissous dans une formation géologique sédimentaire. Les résultats obtenus ont montré des sorties similaires en utilisant l'approche ANN et les simulations RTM traditionnelles, tout en réduisant considérablement le coût computationnel. De plus, Prasianakis et al. (2020) ont développé une architecture d'ANN capable de décrire avec précision les interactions à la fois micro et macroscopiques au sein des systèmes géochimiques, ce qui a conduit à une diminution des temps de simulation. Dans une étude récente par Laloy and Jacques (2022), un ANN et une méthode des k plus proches voisins ont été comparés en tant que modèles de substitution pour simuler un système cimentaire bidimensionnel sous diverses conditions de transport. Bien que les deux méthodes aient

donné des résultats satisfaisants pour les systèmes géochimiques les plus simples, elles ont eu du mal à estimer avec précision des systèmes plus complexes impliquant de nombreux composants. Néanmoins, ces méthodes ont montré une augmentation de la rapidité des calculs, surpassant les solveurs RTM traditionnels. Plus récemment, Demirer et al. (2023) ont proposé un ANN à deux couches conçu pour simuler un système hétérogène complexe tridimensionnel de transport réactif dans des conditions non isothermes. Bien que cette approche ait démontré une précision satisfaisante et amélioré la performance de calcul d'un ordre de grandeur, ils ont pu observer une propagation d'erreur au cours des multiples pas de temps pour la quantité de calcite. De plus, Collard et al. (2023) ont mené une étude comparative de cinq modèles de substitution - Forêt Aléatoire, Gradient Boosting, AdaBoost, *support vector machine* (SVM) et ANN - pour simuler un système géochimique de précipitation/dissolution de calcite-dolomite sans phénomène de transport. Le modèle ANN a surpassé les autres en termes de précision de prédiction et a permis un gain de calcul d'environ trois ordres de grandeur par rapport aux solveurs traditionnels pour le calcul de l'état d'équilibre géochimique. Enfin, Laloy and Jacques (2019) ont démontré la précision des réseaux de neurones profonds et des processus gaussiens (GP) en comparaison à une méthode de projection du chaos polynomial, en particulier lorsque les données d'entraînement sont limitées.

D'autres modèles de substitution ont également été développées ces dernières années, comme celle présentée par Leal et al. (2017). Dans cet article, ils ont proposé une nouvelle stratégie pour réduire considérablement le temps de calcul des simulations RTM, jusqu'à deux ordres de grandeur. Leur approche consiste en une méthode de ML dite à la demande qui repose sur les dérivées dites « de sensibilité » pour déterminer l'état chimique d'un élément à partir de son état précédent et des changements infinitésimaux de température, de pression et de quantité d'espèces. Le résultat ainsi obtenu est soit accepté soit rejeté selon un critère spécifique. S'il est rejeté, la solution est calculée à l'aide d'un solveur géochimique traditionnel. Cette approche permet d'entraîner le modèle de substitution pendant la simulation RTM, augmentant ainsi sa vitesse par un facteur allant de 60 à 125. Cette approche a fait l'objet d'optimisations supplémentaires dans des études plus récentes, telles que celles de Leal et al. (2020) et de Kyas et al. (2022), qui se sont inspirées de la même méthodologie pour réduire davantage le temps d'exécution. De même, De Lucia and Kühn (2021) ont présenté une méthode qui repose sur les données et qui utilise un critère de bilan de masse pendant la simulation du RTM pour déterminer si le modèle de substitution est suffisamment précis pour être utilisé ou si la solution doit être évaluée à l'aide du solveur géochimique.

Ces méthodes utilisant des ANN se sont avérées être des modèles de substitution satisfaisants, permettant une réduction significative du temps de calcul tout en conservant une précision importante. Cependant, elles nécessitent généralement un grand nombre d'observations, par exemple, Guérillot and Bruyelle (2020) ont eu besoin d'au moins 50.000 points d'entraînement pour atteindre leur performance, tandis que Collard et al. (2023) ont eu besoin de 378.000 points. De plus, le nombre élevé d'hyperparamètres à optimiser peut également représenter une contrainte non négligeable. De ce fait, l'extension de leurs applications à des systèmes géochimiques réels plus complexes se heurte encore à un obstacle, car elle nécessite des ressources informatiques supplémentaires et/ou des ensembles de données d'entraînement plus importants pour concevoir des modèles de substitution suffisamment précis.

Dans cette thèse, notre objectif est de résoudre ces problèmes en proposant deux nouvelles approches pour l'estimation de fonctions non-linéaires multivariées dans un contexte géochimique. Ces méthodes sont définies pour exiger un nombre minimal d'observations tout

en offrant une précision satisfaisante, accélérant ainsi considérablement le processus de calcul. En outre, elles sont conçues pour n'avoir que quelques paramètres à déterminer, ce qui facilite leur mise en œuvre. Par ailleurs, en plus de ces deux méthodes d'estimation, nous présenterons une méthode de sélection des variables adaptée aux modèles géochimiques non-linéaires afin de réduire leur complexité et de permettre ainsi une estimation plus précise des solutions des réactions chimiques.

7.2. Estimation de fonction dans des modèles de régression multivariée

7.2.1. État de l'art

La régression non-paramétrique est un domaine bien établi qui offre une approche alternative à la régression paramétrique pour l'estimation d'une fonction lorsqu'aucune information préalable ne peut être fournie concernant cette fonction. Plus précisément, le problème de la régression non-paramétrique vise à estimer une fonction inconnue f dans le modèle suivant :

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = \left(x_i^{(1)}, \ldots, x_i^{(p)}\right), \quad 1 \leq i \leq n, \tag{7.1}$$

où Y_i est une variable aléatoire modélisant la $i^{\text{ème}}$ observation de la *variable réponse* ou *variable de sortie*, $x_i^{(k)}$ désigne la $i^{\text{ème}}$ valeur de la $k^{\text{ème}}$ *variable explicative* ou *variable d'entrée* et les ε_i sont des variables aléatoires centrées i.i.d de variance σ^2.
La présentation du contexte repose principalement sur les livres Tsybakov (2009) et Hastie et al. (2009). Nous nous concentrerons sur l'introduction des idées générales de chaque méthode, étant donné que leur détail est disponible dans la littérature et dans les deux livres précédents pour une compréhension plus approfondie.

7.2.1.1. *Cas unidimensionnel ($p = 1$)*

Méthode de régression à noyau : Le concept fondamental de la régression à noyau ou du lissage à noyau consiste à estimer la fonction f dans (7.1) en chaque point en utilisant ses points d'observation proches, ce qui permet d'obtenir une estimation $\hat{f}$. Il a été initialement présenté comme un estimateur de densité de probabilité d'une variable aléatoire par Rosenblatt (1956) et Parzen (1962) avant d'être étendu à l'estimation de fonction.

Un modèle de régression à noyau largement utilisé est l'estimateur de Nadaraya-Waston (N-W) développé par Nadaraya (1964) et Watson (1964). En introduisant une fonction intégrable spécifique $K : \mathbb{R} \to \mathbb{R}$, appelée noyau et satisfaisant $\int K(u)du = 1$, l'estimateur N-W est défini comme suit :

$$\widehat{f}(x) = \begin{cases} \dfrac{\sum_{i=1}^n Y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}, & \text{si } \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \neq 0, \\ 0 & \text{sinon,} \end{cases} \tag{7.2}$$

où $h > 0$ est la fenêtre, un paramètre à déterminer. Une méthode pour choisir h employant la validation croisée a été proposée par Tsybakov (2009, p.27–31). Comme nous pouvons le voir dans (7.2), cet estimateur peut être considéré comme une moyenne pondérée des Y_i, $i = 1, \ldots, n$, ce qui en fait un estimateur linéaire. Il fournit donc une approche simple et intuitive pour la régression non-paramétrique tout en offrant une certaine flexibilité grâce à la sélection adaptative de la fenêtre. Toutefois, cette méthode peut souffrir d'un *effet de bord* conduisant à des estimations biaisées car elle attribue le même poids à tous

les points à l'intérieur d'une même fenêtre, quelle que soit leur distance par rapport aux bornes du domaine considéré (voir Hastie et al. (2009, p.195) pour une représentation visuelle de cet effet). En outre, il se peut qu'elle ne soit pas bien adaptée aux modèles de régression multivariée, car elle peut souffrir du fléau de la dimension Wasserman (2006, p.58)).

Régression par polynômes locaux : Cette catégorie de méthodes généralise l'estimateur de N-W en ajustant des polynômes locaux au lieu de constantes. Par conséquent, la fonction f est approximée par l'ajustement de polynômes locaux, ce qui permet d'atténuer l'*effet de bord* observé dans l'estimateur de N-W. Cela permet une plus grande flexibilité dans la capture de la structure sous-jacente des données. Ainsi, l'estimation locale en x, $x \in \mathbb{R}$, est définie comme suit : $\widehat{f}(x) = \widehat{\beta}_0(x) + \sum_{j=1}^{q} \widehat{\beta}_j(x)$ où $\widehat{\beta}_0(x), \ldots, \widehat{\beta}_q(x)$, minimisant le critère des moindres carrés pondérés suivant :

$$\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\left(Y_i - \beta_0(x) - \sum_{j=1}^{q} \beta_j(x) x_i^j\right)^2 . \tag{7.3}$$

D'un point de vue historique, les estimateurs par polynômes locaux ont été initialement utilisés dans les années 1930 pour l'analyse des séries temporelles. Ils ont ensuite été employés pour la régression non-paramétrique par Stone et al. (1997). En se basant sur ces méthodes, la régression *loess* a été introduite en tant que modèle d'estimation robuste pondéré localement par Cleveland (1979). Elle ajuste un modèle par polynômes locaux à chaque point en utilisant (7.3) mais en ne considérant que ses k plus proches voisins. Bien qu'elle permette la correction de l'*effet de bord* même à des dimensions plus élevées, cette méthode de régression souffre de l'augmentation de la dimensionnalité des données, notamment lorsque p dépasse 2 ou 3 (Hastie et al. (2009, p.200), Cleveland and Devlin (1988)).

Méthodes par projection : Une autre approche non-paramétrique largement répandue repose sur les méthodes par projection. L'idée fondamentale est de considérer que f dans (7.1) peut être écrite comme :

$$f(x) = \sum_{j=1}^{N} \beta_j \psi_j(x), \quad x \in \mathbb{R}, \tag{7.4}$$

où N est le nombre total de fonctions de base et ψ_j est la $j^{\text{ème}}$ fonction de la base correspondante. Par conséquent, les estimateurs des paramètres β_j, notés $\widehat{\beta}_j$ pour $j = 1, \ldots, N$, sont déterminés à l'aide d'un critère des moindres carrés. Ils sont ensuite utilisés pour construire un estimateur de f en remplaçant β_j par $\widehat{\beta}_j$ dans (7.4). Cette approche transforme le modèle, qui n'est plus écrit en termes de variables initiales x, mais exprimé comme la combinaison linéaire de variables transformées, ce qui facilite l'interprétation du modèle. Un large éventail de fonctions de base peut être utilisé dans le cadre de la régression non-paramétrique, comme la base de Fourier (Rice (1984)) et les ondelettes, qui ont été largement utilisées dans le traitement et la compression des signaux (Hastie et al. (2009, Chapitre 5)). En particulier, les ondelettes donnent des informations sur la localisation temporelle et fréquentielle de la fonction, et sont donc apparues comme une alternative à la base de Fourier depuis les années 1980, gagnant en popularité à la suite des travaux de Meyer (1993). La force de cette méthode réside dans leur capacité à moduler la régularité du signal estimé. Pour des références supplémentaires sur la théorie et les applications des ondelettes, nous renvoyons le lecteur aux ouvrages de

Hernández and Weiss (1996) et de Härdle et al. (1998). Des travaux plus récents sont présentés dans Mallat (1999).

Depuis que l'utilisation des ondelettes s'est répandue dans les années 1990, le développement des méthodes par projection a connu un essor considérable. En conséquence, d'autres bases telles que les polynômes et les splines ont suscité un intérêt croissant pour être utilisées comme fonctions de base ψ_j pour $j = 1, \ldots, N$ dans (7.4). Les splines, en particulier, sont des fonctions polynomiales par morceaux de degré k, continues et dépendant d'une liste de points fixes appelés nœuds à déterminer (De Boor, 1978; Wahba, 1990). Le choix des positions des nœuds dans les splines de régression peut être délicat et peut conduire à un surajustement et à une faible précision s'il est effectué de manière arbitraire (Wood (2017, p.126)). Ce défi a conduit à l'introduction des splines de lissage, une méthode régularisée dans laquelle le critère des moindres carrés comporte un terme supplémentaire qui pénalise les dérivées secondes de f (Wang (2011); Green and Silverman (1994); Eubank (1999)). La dérivée seconde de f représente la régularité de la courbe car elle quantifie le taux de changement de la pente de f. En la pénalisant, les splines de lissage sélectionnent automatiquement les nœuds à partir d'une liste exhaustive prédéfinie, ce qui permet d'obtenir une estimation lisse de f. Pour plus de détails sur ce sujet, nous référons le lecteur à Hastie et al. (2009, Section 5.2).

Une autre famille de splines, connue sous le nom de B-splines, a gagné en popularité en raison de son efficacité en termes de calcul et de sa stabilité. Introduites dans le Chapitre 9 de De Boor (1978), les B-splines sont définies de telle sorte que toute fonction spline peut être écrite comme une combinaison linéaire de B-splines. Cette propriété, associée à un support local, font d'elles des fonctions très utilisées pour l'ajustement de courbes (Piegl and Tiller (2012)) et l'estimation de fonctions à l'aide de splines de lissage (Hastie et al. (2009, p.186–189)). Suivant cette approche, O'Sullivan (1986, Section 3) a introduit une méthode qui repose sur les B-splines pénalisées, devenant ainsi la classe de splines pénalisées la plus utilisée dans les analyses statistiques grâce à ses nombreuses propriétés et à son implémentation en R (Wand and Ormerod (2008)). Inspirés par cette méthode, Eilers and Marx (1996) ont proposé une autre approche des B-splines pénalisées appelée P-splines. L'idée est d'utiliser une liste exhaustive de nœuds régulièrement espacés dans la définition de la base des B-splines ainsi qu'une pénalité sur la différence des coefficients de régression pour inciter le modèle à les réduire à 0, diminuant ainsi la complexité de celui-ci. Nous renvoyons à Eilers et al. (2015) pour une référence récente sur les P-splines et ses extensions et à Eilers and Marx (2003) pour une adaptation et des applications à un modèle de régression à deux variables d'entrée. Plus récemment, Goepp et al. (2018) a présenté une méthode de régression Ridge adaptative pondérée pour conserver les nœuds les plus pertinents des B-splines à partir d'une liste de points régulièrement espacés. Cette approche peut être considérée comme une procédure plus interprétable que les P-splines, mais a donné des résultats similaires à travers diverses simulations numériques.

7.2.1.2. Régression multivariée ($p > 1$)

Modèle additif généralisé (GAMs) et régression multivariée par splines : Dans un modèle de régression non-linéaire, il peut être difficile de saisir la relation entre les variables explicative et la variable réponse. Hastie and Tibshirani (1986) a introduit les GAM pour répondre à cette complexité. Dans ce cas, nous pouvons exprimer f dans (7.1) comme suit :

$$f(x) = \alpha + \sum_{j=1}^{p} f_j(x^{(j)}), \quad \alpha \in \mathbb{R}, \tag{7.5}$$

où f_j peut être une spline cubique ou une autre fonction régulière. Pour ajuster ce modèle, une procédure de *backfitting* a été proposée, calculant de manière itérative $\widehat{f_j}$ en utilisant les estimations actuelles des autres fonctions $\widehat{f_k}$, pour $k \neq j$ (Fox (2015, p.566-567); Wood (2017, p.209 section 4.11.1); Hastie et al. (2009, p.298 Algorithme 9.1)). Toutefois, si chaque fonction est modélisée à l'aide d'une combinaison linéaire de fonctions de base comme dans (7.4), un critère des moindres carrés suffit pour obtenir l'estimation du modèle obtenu. Des compléments sur les GAMs et leurs applications sont disponibles dans (Wood, 2017, Chapitres 4 et 5).

Stone and Koo (1985) fournissent un exemple et des applications de splines additives pour la régression multivariée, ainsi que des discussions sur la sélection des nœuds. Pour une étude plus approfondie de ces sujets, nous recommandons au lecteur l'article Wand (2000) et les références qu'il contient. Le principal avantage de ces modèles additifs est leur facilité d'interprétation et leur implémentation généralement simple. Cependant, ils peuvent être assez restrictifs car ils ne prennent pas en compte les termes d'interaction. Pour remédier à cette limitation, des approches telles que le produit tensoriel de splines (Wasserman (2006, Chapitre 8 p.193); Hastie et al. (2009, Chapitre 5, p.162)) ou les *thin-plate splines* (Wahba (1990, Section 2.4); Green and Silverman (1994, Chapitre 7); Wood (2017, p.150)), ont été introduites pour intégrer les interactions dans les splines de régression. Bien que ces méthodes offrent une grande flexibilité, elles peuvent présenter une complexité en termes de calcul, en particulier lorsque la dimension p augmente.

Méthodes par arbres de décision : Les méthodes reposant sur les arbres ont gagné en visibilité, en particulier depuis l'introduction des arbres de classification et de régression (CART) par Breiman et al. (1984). L'idée principale de cette méthode est d'utiliser la division binaire pour définir itérativement des nœuds, ce qui résulte en M régions qui conduisent à l'expression suivante pour f :

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{1}\{x \in R_m\}, \tag{7.6}$$

où c_m est une constante spécifique à chaque région R_m ($1 \leq m \leq M$) et $\mathbb{1}\{A\} = 1$ si l'événement A se réalise et vaut 0 sinon. Par conséquent, l'estimateur $\widehat{f}$ peut être obtenu en calculant $\widehat{c}_m = N_m^{-1} \sum_{x_i \in R_m} y_i$, où N_m est la cardinalité de R_m pour m appartenant à $\{1, \ldots, M\}$. Les arbres de régression sont simples et faciles à interpréter et de nombreux algorithmes sont désormais disponibles ; voir Loh (2011) pour un travail comparatif et des applications de différents algorithmes d'arbres de régression. Néanmoins, ces méthodes peuvent présenter certaines limitations (Hastie et al. (2009, Chapitre 9, p.307)). Par exemple, la gestion des données manquantes n'est pas optimale, ces méthodes peuvent présenter une instabilité en raison de leur architecture hiérarchique et peinent à capturer de nombreux effets additifs avec un nombre élevé de variables d'entrée p.

Pour résoudre ces problèmes, la méthode du *bagging* et son extension, les forêts aléatoires, ont été introduites respectivement par Breiman (1996) et Breiman (2001). Ces méthodes permettent d'améliorer la robustesse des arbres de régression en calculant la moyenne des valeurs de sortie obtenues à partir d'une large collection d'arbres décorrélés. En outre, elles sont devenues populaires en raison de leur interprétabilité, de leurs performances et de leurs implémentations optimisées dans divers langages de programmation (Hastie et al. (2009, Chapitre 15, p 587)).

Une autre approche notable de la régression multivariée est celle de la régression multivariée par spline adaptative (MARS) présentée par Friedman (1991). MARS repose sur

une méthodologie similaire au modèle CART avec une structure binaire, mais utilise un modèle de régression par spline avec des termes additifs et d'interactions. En tirant parti de fonctions linéaires par morceaux dans un modèle de régression tel que (7.4), MARS produit un modèle parcimonieux puisque les splines sont localement non nulles. De plus, grâce à une procédure *forward* et *backward* visant à réduire le nombre de nœuds et de splines, la complexité du modèle est drastiquement réduite, ce qui rend MARS avantageuse d'un point de vue computationnel.

Machines par vecteurs de support : Les machines à vecteurs de support sont largement connues pour leurs performances en tant que classifieurs, mais leur adaptation à de la régression (SVR), a également démontré des propriétés intéressantes, comme le présente Vapnik et al. (1996). En tirant profit de l'astuce du noyau, les SVRs estiment efficacement les modèles non-linéaires car ils sont connus pour être insensibles aux observations bruitées. Cependant, ils peuvent potentiellement donner lieu à de mauvaises prédictions lorsque la fonction de noyau n'est pas bien choisie.

Réseaux de neurones artificiels (ANNs) : Depuis l'introduction du percepton dans les années 1950 par Rosenblatt (1958), les ANNs sont devenus très populaires pour les problèmes de classification et de régression (voir LeCun et al. (2015)). Il s'agit de modèles statistiques non-linéaires construits par la composition en chaîne de combinaisons linéaires de variables d'entrée transformées par une fonction d'activation choisie, qui peut être linéaire ou non.

Malgré leurs caractéristiques intéressantes, les ANNs peuvent rencontrer plusieurs difficultés. Tout d'abord, leur architecture extrêmement paramétrée peut entraîner une instabilité des performances en raison de la configuration initiale des poids, qui peut aboutir à des minima locaux et à des solutions sous-optimales (Aggarwal et al. (2018, Section 1.4.4)). Ainsi, un trop grand nombre de poids peut entraîner un surajustement des données. Pour résoudre ce problème, des approches de régularisation peuvent être employées, telles que le *weight decay* ou alternativement, le *bagging* ou les méthodes d'ensemble permettant de faire la moyenne des prédictions sur une collection de réseaux à partir de données d'entraînement choisies au hasard (Goodfellow et al. (2016, Chapter 7); Aggarwal et al. (2018, Section 1.4.1)). De plus, la construction des ANNs nécessite généralement des ressources informatiques importantes et leur interprétabilité est souvent entravée car ils sont perçus comme des « *boîtes noires* ». La détermination des hyperparamètres, y compris le choix du nombre optimisé de nœuds et de couches, peut également être complexe (Hastie et al. (2009, Chapitre 11, p.389)). Enfin, les ANNs nécessitent souvent un nombre important d'observations pour atteindre le niveau de précision souhaité (Goodfellow et al. (2016, Part III)).

Dans les méthodes présentées précédemment, l'ensemble d'observations consiste en un nombre prédéterminé de points, obtenus à un coût de calcul significatif, en particulier lorsque ce nombre est important. Cette approche peut être vue comme « passive ». Diverses stratégies peuvent être employées pour réduire ce coût, par exemple en définissant un ensemble réduit d'observations, puis en y ajoutant séquentiellement et de manière adaptative de nouvelles observations en fonction d'un critère spécifique. À chaque itération, une nouvelle estimation peut être obtenue à partir du nouvel ensemble d'observations. Dans ce cas, il est essentiel de définir à la fois un critère de sélection et un critère d'arrêt pour la méthode séquentielle, qui est communément appelée « apprentissage actif » ou *active learning*.

7.2.1.3. Apprentissage actif

Une approche couramment utilisée pour l'apprentissage actif consiste à utiliser un cadre bayésien. Dans cette veine, Srinivas et al. (2012) a proposé une stratégie pour optimiser une fonction inconnue en la considérant comme une réalisation d'un processus gaussien de moyenne nulle, en tirant parti de la moyenne et de la covariance de la distribution a posteriori. La méthode utilise le problème du bandit manchot, inspiré par la procédure séquentielle définie par Robbins (1952), pour optimiser la fonction cible en sélectionnant séquentiellement de nouvelles observations. Pour davantage de précisions sur les processus gaussiens et les fonctions de covariance associées, nous renvoyons le lecteur à Rasmussen and Williams (2006).

Dans la section suivante, nous introduisons une nouvelle approche d'apprentissage actif utilisant des processus gaussiens, inspirée par Srinivas et al. (2012), pour l'estimation de fonction afin de sélectionner séquentiellement de nouveaux points d'observations tout en réduisant l'erreur d'approximation. De plus, nous présentons une méthode complémentaire qui repose sur un modèle de régression multivariée employant les B-splines avec une sélection adaptative de leurs nœuds pour améliorer la précision de l'estimation.

7.2.2. Contribution du Chapitre 2

Cette section résume l'article suivant :
Savino, M., Lévy-Leduc, C., Leconte, M., Cochepin, B. (2022). An active learning approach for improving the performance of equilibrium based chemical simulations. *Computational Geosciences*, 26(2), 365–380.

Nous présentons ici notre approche de l'estimation de fonctions avec l'idée précédemment introduite de l'apprentissage actif pour utiliser un ensemble de points d'observation séquentiellement bien choisis. La fonction à estimer est une fonction f à valeurs réelles définie sur un sous-ensemble compact $\mathcal{A} \subset \mathbb{R}^d$, satisfaisant (7.1).

Nous adoptons un point de vue bayésien qui consiste à considérer f comme une trajectoire d'un processus gaussien (GP) à moyenne nulle ayant une fonction de covariance k que nous désignerons par la suite par GP$(0,k(\cdot,\cdot))$. L'avantage de cette approche est que, conditionnellement à un ensemble de t observations $\mathbf{y}_t = (y_1, \ldots, y_t)'$ où $y_i = f(x_i)$, x_i appartenant à $\mathcal{A}$, la distribution a posteriori est toujours un GP de moyenne μ_t et de fonction de covariance k_t donnée par :

$$\mu_t(u) = \mathbf{k}_t(u)' \mathbf{K}_t^{-1} \mathbf{y}_t \ , \tag{7.7}$$

$$k_t(u,v) = k(u,v) - \mathbf{k}_t(u)' \mathbf{K}_t^{-1} \mathbf{k}_t(v) \ , \tag{7.8}$$

où $\mathbf{k}_t(u) = [k(x_1,u) \ldots k(x_t,u)]'$. Ici, $'$ désigne la transposition matricielle, u et v sont dans $\mathcal{A}$ et $\mathbf{K}_t = [k(x_i,x_j)]_{1 \le i,j \le t}$, où les x_i appartiennent à $\mathcal{A}$.

Dans notre cas, f modélise une quantité physique qui est supposée être régulière. De ce fait, nous considèrerons deux fonctions de covariance pour nos applications qui sont couramment utilisées. La première est la fonction de covariance exponentielle carrée (SE) :

$$k_{\mathrm{SE}}(u,v) = \exp\left(-\frac{1}{2}(u-v)' M^{-1} (u-v)\right), \quad u,v \in \mathcal{A} \subset \mathbb{R}^d \ , \tag{7.9}$$

$$M = \mathrm{diag}\left(\ell_1^2, \ldots, \ell_d^2\right), \ \ell_1 \ , \ \ell_2 \ , \ldots \ , \ell_d > 0 \ . \tag{7.10}$$

Ici, $\ell_1, \ell_2, \ldots, \ell_d$ sont des hyperparamètres appelés longueurs de corrélation. Ces hyperparamètres peuvent être compris comme la distance à parcourir le long d'un axe donné dans l'espace d'entrée pour que les valeurs de la fonction ne soient plus corrélées. Il faut noter que la définition (7.9) nous permet de modéliser des surfaces de réponse anisotropes.

Comme l'explique Rasmussen and Williams (2006), puisque cette fonction de covariance est infiniment dérivable, définir un GP avec cette fonction de covariance permet d'obtenir des dérivées en moyenne quadratique de tout ordre. Selon Stein (1999), de telles hypothèses de régularité peuvent être irréalistes pour modéliser de nombreux processus physiques. Nous allons donc également considérer une autre fonction de covariance appartenant à la classe des fonctions de covariance de Matérn définie par :

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} r\right)^{\nu} K_{\nu}\left(\sqrt{2\nu} r\right), \quad \nu > 0 \,,$$

où K_ν est une fonction de Bessel modifiée d'ordre ν, voir Abramovitz and Stegun (1965, Section 9.6), et r est défini par :

$$r = \sqrt{(u-v)' M^{-1} (u-v)}, \quad u, v \in \mathcal{A} \,, \tag{7.11}$$

M étant défini dans (7.10). Dans cette situation, comme expliqué dans Rasmussen and Williams (2006), le GP est q fois différentiable en moyenne quadratique, si et seulement si $\nu > q$. Nous nous concentrerons ici sur le cas où $\nu = 5/2$, pour lequel $k_{\text{Matérn}}$ a une expression avantageuse d'un point de vue computationnel. En effet, pour $\nu = p + \frac{1}{2}$, où p appartient à $\mathbb{N}$,

$$k_{\text{Matérn}}(r) = \exp\left(-\sqrt{2\nu} r\right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\sqrt{8\nu} r\right)^{p-i} \,, \tag{7.12}$$

où r est introduit dans (7.11) ; voir Abramovitz and Stegun (1965, Equation 10.2.15) pour plus de détails.

Dans la suite, nous désignerons par A une grille fine de $\mathcal{A}$:

$$\mathrm{A} = \{\mathrm{x}_1, \ldots, \mathrm{x}_m\} \subset \mathcal{A} \,. \tag{7.13}$$

Cette grille est soit une grille régulière de $\mathcal{A} \subset \mathbb{R}^d$ lorsque d est petit (généralement 1 ou 2), soit obtenue grâce à un *Latin Hypercube Sampling* (LHS) pour des valeurs plus importantes de d. Il est à noter que cette grille contient à la fois les points en lesquels l'estimation de f est effectuée et les points choisis pour lesquels f est évaluée avec le solveur. Inspirés par Srinivas et al. (2012) qui a proposé une approche séquentielle pour maximiser une fonction en la considérant comme une trajectoire de processus gaussien, nous proposons une stratégie qui consiste à ajouter le nouveau point x_{t+1} à l'ensemble des t observations pour lesquelles f doit être évaluée comme suit :

$$x_{t+1} \in \operatorname*{Arg\,max}_{x \in \mathrm{A}} \sigma_t(x) \,,$$

où

$$\sigma_t(x)^2 = k_t(x, x), \tag{7.14}$$

k_t est définie dans (7.8) et $\operatorname*{Arg\,max}_{x \in \mathrm{A}} \sigma_t(x)$ est l'ensemble des $x \in \mathrm{A}$ pour lesquels $\sigma_t(x)$ vaut la valeur maximale. Notons que les points $x_1, x_2, \ldots, x_t, x_{t+1}, \ldots$ pour lesquels f doit être évaluée sont choisis dans la grille fine A de $\mathcal{A}$ définie dans (7.13).

Nous proposons d'utiliser une stratégie de maximum de vraisemblance décrite dans Rasmussen and Williams (2006) pour estimer $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_d)$. Cela ajoute une étape à la méthode décrite précédemment, car les ℓ_i doivent être estimés avant d'évaluer la distribution a posteriori du GP à l'aide de (7.7) et de (7.8). Par conséquent, pour l'ensemble d'observations $\{(x_1, y_1), \ldots, (x_t, y_t)\}$ avec $y_i = f(x_i),\ 1 \leq i \leq t$, la log-vraisemblance a posteriori est donnée par :

$$-\frac{1}{2}\mathbf{y}_t'\mathbf{K}_t^{-1}\mathbf{y}_t - \frac{1}{2}\log|\mathbf{K}_t| - \frac{t}{2}\log 2\pi\ ,$$

avec $\mathbf{y}_t = (y_1, \ldots, y_t)'$ et $\mathbf{K}_t = [k(x_i, x_j)]_{1\leq i,j\leq t}$. Elle doit être maximisée par rapport à $\boldsymbol{\ell}$.

Différents critères d'arrêt reposant sur les éléments suivants ont été définis :

- **Ratio des variances.** A chaque itération t de notre méthode, la moyenne suivante est calculée :

$$R_n(t) = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\max_{x\in\mathrm{A}}\sigma_t^2(x)}{\max_{x\in\mathrm{A}}\sigma_{t-i}^2(x)}, \tag{7.15}$$

  où σ_t est définie dans (7.14) et $n = 2$, 5 ou 10.

- **Moyenne mobile.** A chaque itération t de notre méthode, la moyenne suivante est calculée :

$$M_\ell(t) = \frac{1}{\ell}\sum_{j=0}^{\ell-1}\max_{x\in\mathrm{A}}\sigma_{t-j}^2(x) \tag{7.16}$$

  pour $\ell = 5$ ou 10 où σ_t est définie dans (7.14).

- **Variance maximale.** A chaque itération t de notre méthode,

$$V(t) = \max_{x\in\mathrm{A}}\sigma_t^2(x) \tag{7.17}$$

  est calculé, avec σ_t définie dans (7.14).

Cette méthode d'apprentissage actif pour l'estimation de fonctions inconnues est d'abord évaluée sur des fonctions unidimensionnelles et bidimensionnelles de réactions géochimiques. Sur la base de trois mesures statistiques, nous avons comparé les performances des deux fonctions de covariance définies dans (7.9) et (7.12), en utilisant $\nu = \frac{5}{2}$, ainsi que les trois critères d'arrêt introduits dans (7.15), (7.16) et (7.17). Le choix de la fonction de covariance s'avère insignifiant dans ces cas d'étude et les critères reposant sur le ratio et la moyenne sur au moins 10 itérations se sont révélés les plus satisfaisants, réduisant de manière significative le nombre d'observations. Cependant, le seuil de ces critères d'arrêt doit être spécifiquement sélectionné pour la fonction de covariance choisie. D'autres applications à des cas géochimiques de plus grandes dimensions ont abouti à des conclusions similaires, soulignant le faible nombre de points d'observations requis pour obtenir une précision intéressante.

7.2.3. Contribution du Chapitre 3

Cette section résume l'article suivant :
Savino, E. M., Lévy-Leduc, C. 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. Soumis et disponible en preprint sur arXiv : *(arXiv :2306.00686)*.

La méthode proposée est implémentée dans le package R `glober` disponible sur le CRAN.

Nous proposons d'estimer la fonction f apparaissant dans (7.1) en l'approximant avec une combinaison linéaire de B-splines d'ordre M ($M \geq 1$) introduites par De Boor (1978) au Chapitre 9.

Soit $\mathbf{t} = (t_1, \ldots, t_K)$ un ensemble de K points appelés nœuds qui sont cruciaux dans la définition de la base de B-splines. Nous définissons la séquence augmentée de nœuds $\boldsymbol{\tau}$ telle que :

$$\tau_1 = \ldots = \tau_M = x_{min},$$
$$\tau_{j+M} = t_j, \quad j = 1, \ldots, K,$$
$$x_{max} = \tau_{K+M+1} = \ldots = \tau_{K+2M},$$
$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{K+2M}) = (\underbrace{x_{min}, \ldots, x_{min}}_{\text{M fois}}, \underbrace{t_1, \ldots, t_K}_{\mathbf{t}}, \underbrace{x_{max}, \ldots, x_{max}}_{\text{M fois}}),$$

où x_{min} et x_{max} sont respectivement les bornes minimale et maximale de $\mathcal{S}$, un ensemble compact de $\mathbb{R}$ sur lequel f est définie.

Les B-splines sont définies par De Boor (1978, p. 89-90) et Hastie et al. (2009, p. 160) comme suit. En désignant par $B_{i,m}(x)$ la $i^{\text{ème}}$ B-spline d'ordre m pour la séquence de nœuds $\boldsymbol{\tau}$ avec $m \leq M$, nous définissons la récurrence suivante :

$$B_{i,1}(x) = \begin{cases} 1 & \text{si } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{sinon} \end{cases} \quad \text{pour } \; i = 1, \ldots, K + 2M - 1, \tag{7.18}$$

et pour $m \leq M$,

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x), \tag{7.19}$$

pour $i = 1, \ldots, (K + 2M - m)$.

La méthode d'estimation introduite ci-dessous est appelée GLOBER. Soit $\mathbf{Y} = (Y_1, \ldots, Y_n)$ et $\mathbf{x} = (x_1, \ldots, x_n)$ où Y_i et x_i sont définis dans (7.1). Dans la suite, nous supposerons que $x_1 < \ldots < x_n$ et $M = q + 1$, avec $q \geq 0$. Par conséquent, lorsque $q = 0$ (resp. $q = 1$, $q = 2$) f est approximée par des fonctions constantes par morceaux (resp. linéaires, quadratiques).

Puisque les nœuds d'une base de B-splines peuvent être considérés comme des changements dans la dérivée $(q+1)^{\text{ème}}$ de f, nous proposons de les trouver en utilisant le *generalized lasso* décrit par Tibshirani and Taylor (2011) et étudié plus en détail par Tibshirani (2014). Dans ce dernier, ils définissent le *trend filtering* polynomial qui consiste à approcher f par $\widehat{\boldsymbol{\beta}}(\lambda)$ définie comme suit :

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\operatorname{argmin}} \{ ||\mathbf{Y} - \boldsymbol{\beta}||_2^2 + \lambda ||D\,\boldsymbol{\beta}||_1 \}, \tag{7.20}$$

où $||y||_2^2 = \sum_{i=1}^n y_i^2$ pour $y = (y_1, \ldots, y_n)$ et $||u||_1 = \sum_{i=1}^m |u_i|$ pour $u = (u_1, \ldots, u_m)$, λ est une constante strictement positive à déterminer et $D \in \mathbb{R}^{m \times n}$ est une matrice de pénalité spécifique, définie par récurrence :

$$D = D_{tf,q+1} = D_0 \cdot D_{tf,q} \quad q \geq 0,$$

où "tf" est l'abréviation pour "trend filtering", $(q+1)$ est l'ordre de pénalité, $D_{tf,0} = \mathrm{Id}_{\mathbb{R}^n}$, et la matrice identité de $\mathbb{R}^n$, et D_0 est la matrice de pénalité pour le *fused lasso* unidimensionnel :

$$D_0 = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

La matrice de pénalité D est l'opérateur de différence discrète d'ordre $(q+1)$ et donc, $D\widehat{\boldsymbol{\beta}}$ estime la dérivée $(q+1)^{\text{ème}}$ de f. Ainsi, observer les positions où $D\widehat{\boldsymbol{\beta}} \neq 0$ permet de trouver les nœuds des B-splines.

La matrice D est bien adaptée lorsque les points d'observation sont régulièrement espacés. Lorsque ce n'est pas le cas, elle doit être remplacée par la matrice suivante $\Delta^{(q+1)}$ définie récursivement comme suit :

$$\Delta^{(q+1)} = \mathbf{W}_{(q+1)} \cdot D_0 \cdot \Delta^{(q)}, \quad q \geq 0,$$

où $\Delta^{(0)} = \mathrm{Id}_{\mathbb{R}^n}$ et $\mathbf{W}_{(q+1)}$ est la matrice diagonale de poids définie par :

$$\mathbf{W}_{(q+1)} = \mathrm{diag}\left(\frac{1}{(x_{(q+1)+1} - x_{(q+1)})}, \frac{1}{(x_{(q+1)+2} - x_{(q+1)+1})}, \ldots, \frac{1}{(x_n - x_{n-1})} \right).$$

Dans les deux cas (observations régulièrement ou irrégulièrement espacées), le nombre de lignes de D et $\Delta^{(q+1)}$ est égal à $m = n - q - 1$.

Nous allons expliquer maintenant plus précisément comment choisir les nœuds de la base de B-splines. Soit $\Lambda = (\lambda_1, \ldots, \lambda_k)$ une grille de paramètres de pénalisation candidats λ_i. Nous définissons le vecteur colonne différencié résultant $\mathbf{a}(\lambda)$ comme étant :

$$\mathbf{a}(\lambda) = \Delta^{(q+1)} \cdot \widehat{\boldsymbol{\beta}}(\lambda),$$

où $\widehat{\boldsymbol{\beta}}(\lambda)$ est solution du problème (7.20) pour $D = \Delta^{(q+1)}$ et λ appartenant à Λ.

Le vecteur ordonné des nœuds sélectionnés associés à λ est défini comme suit :

$$\widehat{\mathbf{t}}_\lambda = \left(\widehat{t}_j\right)_{j=1,\ldots,K_\lambda} = \left(x_{p_j}\right)_{j=1,\ldots,K_\lambda}, \quad \text{with } p_j \in \mathcal{P}_\lambda, \tag{7.21}$$

où

$$\mathcal{P}_\lambda = \{\ell + 1,\ a_\ell(\lambda) \neq 0\} \quad \text{and} \quad K_\lambda = \sum_{\ell=1}^{m} \mathbb{1}\{a_\ell(\lambda) \neq 0\},$$

$a_\ell(\lambda)$ désigne la $\ell^{\text{ème}}$ composante de $\mathbf{a}(\lambda)$ et $\mathbb{1}\{A\} = 1$ si l'événement A se réalise et vaut 0 sinon.

La base de B-splines correspondante $B_{i,M}$ est définie en remplaçant les t_j de la séquence augmentée de nœuds $\boldsymbol{\tau}$ apparaissant dans (7.18) et (7.19) par $\widehat{t}_j$ trouvés dans (7.21). Nous obtenons ainsi l'estimateur de f pour chaque λ de Λ suivant :

$$\widehat{f}_\lambda(x) = \sum_{i=1}^{q+K_\lambda+1} \widehat{\gamma}_i B_{i,M}(x), \tag{7.22}$$

où $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_i)_{1\leq i\leq q+K_\lambda+1}$ est obtenu en utilisant le critère des moindres carrés suivant :

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{q+K_\lambda+1}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{B}(\boldsymbol{\lambda})\,\boldsymbol{\gamma}\|_2^2, \tag{7.23}$$

où $\mathbf{B}(\boldsymbol{\lambda})$ est une matrice de taille $n \times (q + K_\lambda + 1)$ qui a pour $i^{\text{ème}}$ colonne $(B_{i,M}(x_k))_{1\leq k\leq n}$, i appartenant à $\{1, \ldots, q + K_\lambda + 1\}$.

Afin de choisir le paramètre de pénalisation $\boldsymbol{\lambda}$ qui conduit à la meilleure sélection de nœuds, nous utilisons un critère défini par Chen and Chen (2008) et recommandé par Goepp et al. (2018), à savoir le critère d'information bayésien étendu également appelé EBIC :

$$\mathsf{EBIC}(\boldsymbol{\lambda}) = \mathsf{SS}(\boldsymbol{\lambda}) + (q + K_\lambda + 1)\log n + 2\log\binom{q + K_{\mathsf{max}} + 1}{q + K_\lambda + 1}, \tag{7.24}$$

où K_{max} est le nombre maximum de nœuds qui peuvent être sélectionnés (ici $K_{\mathsf{max}} = n$) et $\mathsf{SS}(\boldsymbol{\lambda})$ est la somme des carrés définie par :

$$\mathsf{SS}(\boldsymbol{\lambda}) = \|\mathbf{Y} - \widehat{\mathbf{Y}}(\boldsymbol{\lambda})\|_2^2,$$

où

$$\widehat{\mathbf{Y}}(\boldsymbol{\lambda}) = \mathbf{B}(\boldsymbol{\lambda})\widehat{\boldsymbol{\gamma}},$$

avec $\widehat{\boldsymbol{\gamma}}$ et $\mathbf{B}(\boldsymbol{\lambda})$ défini dans (7.23). Ce critère nous permet d'obtenir un compromis entre une bonne approximation de la fonction sous-jacente et un nombre raisonnable de paramètres à estimer. L'estimateur final de f est défini comme étant :

$$\widehat{f}(x) = \widehat{f}_{\boldsymbol{\lambda}_{\mathsf{EBIC}}}(x),$$

où $\widehat{f}_{\boldsymbol{\lambda}}(x)$ est définie dans (7.22) et

$$\boldsymbol{\lambda}_{\mathsf{EBIC}} = \underset{\boldsymbol{\lambda}\in\Lambda}{\operatorname{argmin}}\{\mathsf{EBIC}(\boldsymbol{\lambda})\}.$$

L'extension à $d = 2$ est également proposée dans ce chapitre, selon la méthodologie suivante. Tout d'abord, nous considérons que $\mathcal{S}$ est défini comme le produit cartésien de deux ensembles compacts $\mathcal{S}_1$ et $\mathcal{S}_2$ de $\mathbb{R}$. Nous cherchons à estimer f dans (7.1) comme suit :

$$\sum_{i=1}^{Q_1}\sum_{j=1}^{Q_2}\gamma_{ij}\,B_{1,i,M}(x_1)B_{2,j,M}(x_2), \quad (x_1, x_2) \in \mathcal{S}_1 \times \mathcal{S}_2, \tag{7.25}$$

où $B_{1,i,M}$ et $B_{2,j,M}$ sont les bases de B-splines d'ordre M définies par (7.19) pour respectivement la première et la deuxième dimension. Dans (7.25), $Q_1 = q + K_1 + 1$, $Q_2 = q + K_2 + 1$ avec K_1 et K_2 le nombre de nœuds définis dans la base de B-splines de la première et de la deuxième variable, respectivement et $M = q + 1$. Nous pouvons donc considérer les deux dimensions de manière indépendante et ainsi, en fixant une dimension à la fois, le problème peut être réécrit comme un problème d'estimation dans le cadre unidimensionnel.

Par conséquent, pour chaque dimension, nous identifions différents ensembles de nœuds candidats. Ensuite, pour chaque combinaison de ces ensembles, nous calculons l'EBIC défini à partir de (7.24). L'estimateur final de f est obtenu en construisant la base de B-splines pour chaque dimension en utilisant la combinaison de nœuds sélectionnée par l'EBIC. Ce chapitre présente également une extension de cette méthode, appelée GLOBER-c, conçue pour prendre en compte des dimensions plus élevées et des ensembles généraux de points qui ne peuvent pas être générés par un produit cartésien. GLOBER-c y parvient en employant une méthode de *clustering* pour regrouper les observations dans différents groupes, puis pour identifier l'ensemble de nœuds candidats pour chaque dimension. Des expériences numériques sont menées pour évaluer l'influence de l'échantillonnage de l'ensemble d'observations et du niveau de bruit. Leur impact ne semble pas significatif. En outre, le chapitre explore certaines applications de la méthode aux réactions géochimiques. Les résultats démontrent que, dans ces contextes spécifiques, notre méthode surpasse plusieurs méthodes de l'état de l'art, telles que MARS, GP et une architecture de ANN.

7.3. Sélection de variables dans les modèles non-linéaires multivariés

7.3.1. État de l'art

Une autre approche pour réduire le temps de calcul et simplifier les simulations géochimiques consiste à réduire le nombre de variables d'entrée prises en compte dans le modèle, une stratégie également connue sous le nom de *sélection de variables* ou *sélection de caractéristiques*. Considérons que nous avons n observations satisfaisant le modèle de régression non-paramétrique défini dans (7.1). Nous supposerons que f ne dépend en réalité que de d variables au lieu de p, avec $d < p$, ce qui signifie qu'il existe une fonction à valeurs réelles $\widetilde{f}$ telle que $f(x) = \widetilde{f}(\widetilde{x})$, où $x \in \mathbb{R}^p$ et $\widetilde{x} \in \mathbb{R}^d$. La sélection des variables consiste à identifier les composantes de $\widetilde{x}$.

7.3.1.1. Méthodes à noyau

L'extension de la régression lasso linéaire introduite par Tibshirani (1996) a ouvert la voie à la sélection de variables dans les modèles de régression non-linéaires. Inspirés par Roth (2004), Lin and Zhang (2006) ont présenté la *Feature Vector Machine*, une régression lasso non-linéaire pour la sélection de caractéristiques qui sert d'adaptation parcimonieuse des SVR. Toutefois, cette méthode présente certaines contraintes car elle applique la même fonction de noyau non-linéaire à la fois à la réponse et aux variables d'entrée, ce qui limite sa flexibilité dans la capture des dépendances non-linéaires. Plus récemment, Yamada et al. (2014) ont présenté HSIC-Lasso et sa version alternative NOCCO-Lasso, deux méthodes à noyau qui exploitent un critère de redondance minimale et de pertinence maximale (mRMR), tel que défini par Peng et al. (2005), pour contourner les limitations de l'approche précédente. Ces méthodes utilisent différentes fonctions de noyau pour les variables de sortie et d'entrée pour plus de flexibilité. Sur la base de critères d'indépendance propres aux méthodes à noyau, tels que HSIC ou NOCCO, elles sélectionnent les variables pertinentes tout en éliminant celles qui sont les plus redondantes.

Une autre approche proposée par Bertin and Lecué (2008) consiste à employer un modèle de régression par polynômes locaux pénalisé par une régularisation ℓ_1 sur les coefficients de régression. Le sous-ensemble de variables sélectionnées est ensuite utilisé pour estimer f à l'aide d'un estimateur par polynômes locaux, en minimisant un critère des moindres carrés pondérés tel que défini dans (7.3). Cette approche est théoriquement justifiée. En se basant sur la même stratégie de régularisation, Allen (2013) ont introduit KNIFE, une méthode à noyau pondéré pour la sélection de caractéristiques. Cette méthode incorpore des poids dans les fonctions de noyau et propose de minimiser une fonction de perte avec une pénalité sur la norme ℓ_1 de ces poids pour renforcer la discrimination des caractéristiques les plus pertinentes.

Alors que ces approches visent à pénaliser les poids ou les coefficients de régression, Rosasco et al. (2010) ont proposé une pénalité qui repose sur les dérivées partielles de la fonction par rapport à chaque variable, sélectionnant les variables les plus pertinentes pour lesquelles la dérivée partielle est non nulle. Une étude approfondie de cette approche, incluant ses propriétés de sélection, ses aspects computationnels et des comparaisons avec d'autres méthodes, peut être trouvée dans l'article Rosasco et al. (2013).

7.3.1.2. Modèles additifs et régression régularisée

Une extension de l'approche lasso aux modèles de régression non-linéaire dans le contexte des splines de lissage a été présentée par Lin and Zhang (2006) sous le nom de COSSO. Des comparaisons avec MARS (Friedman, 1991), défini dans la section précédente, en tant que méthode de sélection de variables en raison de sa capacité à élaguer les termes associés à des variables non pertinentes, sur quelques exemples ont révélé que COSSO présentait une performance supérieure en matière de sélection de variables.

De manière analogue, Ravikumar et al. (2009) ont introduit une nouvelle approche fondée sur les GAM précédemment définis dans (7.5), appelée *sparse additive models* (SpAM). Considérée comme une version fonctionnelle du *group lasso* (Yuan and Lin (2006)), SpAM traite la régularité et la parcimonie séparément, permettant l'expression de chaque f_j dans l'équation (7.5) comme une combinaison linéaire de fonctions de base (voir (7.4)). Les B-splines sont un choix possible pour ces fonctions de base. L'utilisation des P-splines, introduites dans la section précédente, peut être étendue à la sélection de variables comme discuté dans Antoniadis et al. (2012). Leur approche utilise une méthode de régularisation, le *nonnegative garrot* introduit par Breiman (1995), pour simplifier le modèle et estimer les coefficients de régression. De plus, Huang et al. (2010) ont proposé une procédure en deux étapes utilisant une approche adaptative du *group lasso* qui est une extension du lasso adaptatif défini par Zou (2006) au *group lasso*, pour sélectionner les variables pertinentes et réduire la complexité du modèle. Enfin, Radchenko and James (2010) ont adapté l'approche additive SpAM en incorporant des termes d'interaction deux à deux à travers l'utilisation de fonctions de base, améliorant ainsi la flexibilité du modèle.

7.3.1.3. Méthodes par arbres de décision

Une méthode populaire de sélection de variables, proposée par Genuer et al. (2010), est fondée sur les Forêts Aléatoires ou *random forests* (RF). Inspirée par le travail de Díaz-Uriarte and Alvarez de Andrés (2006), la procédure comprend deux étapes. Premièrement, elle calcule l'importance de chaque variable (VI) en les permutant et en mesurant la différence entre l'erreur de régression en utilisant l'ensemble original d'observations et celui obtenu avec l'ensemble permuté. Les variables sont ensuite classées en fonction de leur VI et les m premières variables sont conservées, m étant un paramètre à choisir. Dans la deuxième étape, les variables présélectionnées sont ensuite discriminées selon deux stratégies distinctes, l'une dédiée à l'amélioration des prédictions et l'autre au renforcement de l'interprétabilité. Bien que cette méthode identifie efficacement les variables pertinentes, elle peut être sensible au choix des hyperparamètres et aux variables corrélées.

De manière similaire, Galelli and Castelletti (2013) ont introduit une approche itérative de sélection de variables en utilisant une méthode d'ensemble reposant sur la construction de plusieurs arbres avec l'ensemble complet des observations. Cette méthode classe d'abord les variables en fonction de leur importance en termes de variance expliquée, puis, à chaque itération, sélectionne la meilleure variable parmi les m premières variables en fonction de l'erreur de régression. La procédure se poursuit jusqu'à ce qu'aucune amélioration supplémentaire ne soit observée dans le modèle de régression total en utilisant toutes les variables sélectionnées. Cette méthode est censée être plus robuste et éviter la redondance des variables corrélées.

7.3.1.4. ANNs

Les méthodes fondées sur les ANN ont suscité un intérêt important pour la sélection de variables ces dernières années, se révélant très efficaces pour les modèles de régression non-

linéaires et pour faire face aux données de grande dimension. Une approche courante consiste à utiliser des méthodes de régularisation sur les poids pour réduire ceux associés aux variables les moins pertinentes. Par exemple, Li et al. (2016) ont introduit une méthode de sélection de caractéristiques appelée *deep feature selection* (DFS) utilisant un perceptron multicouche avec une pénalité Elastic-Net sur les poids de la première couche, chacun correspondant à une variable d'entrée. Ils ont également inclus un terme de pénalité sur les couches cachées pour réduire la complexité du modèle et éviter l'augmentation des poids dans les couches supérieures. De manière analogue, Feng and Simon (2017) ont introduit SPINN, un réseau de neurones présentant une régularisation *sparse group lasso* sur les poids de la première couche. Cette approche permet d'atteindre un modèle parcimonieux en réduisant l'ensemble des vecteurs de poids associés aux variables non pertinentes. De plus, une pénalité Ridge est appliquée sur les poids des couches supérieures pour privilégier les connexions entre très peu de nœuds. Ye and Sun (2018) ont étendu cette méthode en incorporant une approche *drop-one-out*, en utilisant un algorithme glouton pour supprimer itérativement un poids à la fois associé à une variable et en observant les effets dans la fonction de perte. Chen et al. (2021) ont proposé une architecture de réseau de neurones régularisé avec une couche de sélection spécifique à la DFS, pour laquelle ils ont démontré la consistance de sélection sous certaines contraintes. L'intégration de *skip-connections* dans l'architecture des ANN est une autre approche pour simplifier le modèle, réduire le temps d'entraînement et améliorer la sélection de variables. Par exemple, Feng and Simon (2022) ont présenté SIER-net qui permet d'obtenir une architecture simplifiée et parcimonieuse en contrôlant notamment le nombre de couches et de nœuds actifs. De plus, ils ont proposé une extension de cette approche à une méthode d'ensemble en utilisant une perspective bayésienne pour améliorer la performance de sélection de variables, plus particulièrement pour mieux gérer les variables corrélées ou groupées. Sur ce même principe, Lemhadri et al. (2021) ont proposé un réseau de neurones résiduel à action directe, introduit par He et al. (2016). Ils ont utilisé une régression lasso pour réduire les poids de la couche résiduelle associée aux variables non pertinentes. Une contrainte forte implique l'élimination directe des poids de la première couche cachée liée à ces variables non pertinentes. Une autre approche proposée par Liang et al. (2018) consiste à utiliser une architecture de réseau neuronal bayésien pour sélectionner les variables pour lesquelles la probabilité d'inclusion marginale dépasse un seuil prédéfini.

L'un des principaux inconvénients de ces méthodes est le manque d'interprétabilité et de reproductibilité. Pour résoudre ce problème, des travaux récents fondés sur les variables dites *knockoff* et le contrôle du *false discovery rate*, introduits par Candès et al. (2018), ont été utilisés. Ces variables *knockoff* sont utilisées comme contrôles pour la sélection de variables et sont générées de manière aléatoire pour imiter la structure de dépendance arbitraire qui existe au sein des variables d'entrée tout en restant indépendantes de la variable de sortie, conditionnellement aux variables d'entrée. DeepPINK de Lu et al. (2018) est une architecture de ANNs qui exploite cette stratégie en intégrant des variables *knockoff* dans le modèle pour déterminer l'importance de chaque caractéristique. Une version de sélection par groupe de caractéristiques de cette approche peut être trouvée dans Zhu and Zhao (2021) pour les ensembles de données présentant des structures groupées.

Une limitation couramment rencontrée avec la majorité de ces méthodes est le temps de calcul nécessaire pour la procédure d'entraînement, principalement en raison du réglage des hyperparamètres. Cela peut soulever des questions sur leur efficacité, surtout lorsque des approches très rapides telles que SpAM ou les RF donnent des résultats de sélection et de

précision comparables.

7.3.2. Contribution du Chapitre 4

Cette section résume l'article suivant :
Savino, E. M., Lévy-Leduc, C. A novel variable selection method in nonlinear multivariate models using B-splines with an application to geoscience. Soumis et disponible en pre-print sur HAL : *(hal-04434820)*.

La méthode proposée est implémentée dans le package R `absorber` disponible sur le CRAN.

Inspirés par Radchenko and James (2010), nous proposons d'approcher la fonction $f(x^{(1)},\dots,x^{(p)})$ apparaissant dans (7.1) par une combinaison linéaire de B-splines de chaque variable $x^{(1)},\dots,x^{(p)}$ et de l'interaction deux par deux de celles-ci comme suit :

$$F(x^{(1)},\dots,x^{(p)}) = \sum_{\ell=1}^{p}\sum_{k=1}^{K+M}\beta_k^{(\ell)}B_k^{(\ell)}(x^{(\ell)}) + \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p}\left(\sum_{k=1}^{K+M}\sum_{q=1}^{K+M}\beta_{k,q}^{(\ell,j)}B_k^{(\ell)}(x^{(\ell)})B_q^{(j)}(x^{(j)})\right), \tag{7.26}$$

où $B_k^{(\ell)} = B_{k,M}^{(\ell)}$ est définie dans (7.18) et (7.19) et où $\beta_k^{(\ell)}$ et $\beta_{k,q}^{(\ell,j)}$ sont des coefficients inconnus. Nous pouvons observer que le vecteur colonne $(F(x_i^{(1)},\dots,x_i^{(p)}))_{1\leq i\leq n}$ dans (7.26) peut se réécrire comme étant :

$$\sum_{\ell=1}^{p}\Psi_\ell\boldsymbol{\beta}_\ell + \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p}\Phi_{\ell j}\boldsymbol{\beta}_{\ell,j}. \tag{7.27}$$

où Ψ_ℓ est une matrice de taille $n\times(K+M)$ telle que sa $i^{\text{ème}}$ ligne est égale à $(B_1^{(\ell)}(x_i^{(\ell)}),\dots,B_{K+M}^{(\ell)}(x_i^{(\ell)}))$ et $\boldsymbol{\beta}_\ell = \begin{pmatrix}\beta_1^{(\ell)} & \dots & \beta_{K+M}^{(\ell)}\end{pmatrix}^T$ pour $1\leq\ell\leq p$, A^T désignant la transposée de A. De plus, $\Phi_{\ell j}$ est une matrice de taille $n\times(K+M)^2$ telle que sa $i^{\text{ème}}$ ligne satisfait $(\Phi_{\ell j})_{i,\bullet} = ((\Psi_\ell)_{i,\bullet}\otimes(\Psi_j)_{i,\bullet})$, $\otimes$ désignant ici le produit de Kronecker, $(\Psi_\ell)_{i,\bullet}$ correspond à la $i^{\text{ème}}$ ligne de Ψ_ℓ et $\boldsymbol{\beta}_{\ell,j} = \begin{pmatrix}\beta_{1,1}^{(\ell,j)} & \beta_{1,2}^{(\ell,j)} & \dots & \beta_{K+M,K+M}^{(\ell,j)}\end{pmatrix}^T$ pour $1\leq\ell<j\leq p$.

Guidés par la méthodologie de Rosasco et al. (2010), nous proposons de sélectionner les variables dont f dépend en estimant les coefficients $\boldsymbol{\beta}_\ell$ et $\boldsymbol{\beta}_{\ell,j}$ apparaissant dans (7.27) par la minimisation du critère régularisé suivant :

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

où $\mathbf{Y} = (Y_1,\dots,Y_n)$, les Y_i's sont introduits dans (7.1), $\partial_\ell F(x_i)$ désigne la $\ell^{\text{ème}}$ dérivée partielle de F définie dans (7.26) à un certain point d'observation $x_i = (x_i^{(1)},\dots,x_i^{(p)})$ et $||y||_2^2 = \sum_{i=1}^n y_i^2$. Il est à noter que l'idée sous-jacente à ce critère est que lorsqu'une fonction ne dépend pas d'une variable, sa dérivée partielle par rapport à cette variable est égale à zéro.

En utilisant la définition de F donnée dans (1.27), le critère peut être réécrit comme suit :

$$\begin{aligned}
&\left(\widehat{\boldsymbol{\beta}}_1(\lambda),\ldots,\widehat{\boldsymbol{\beta}}_p(\lambda),\widehat{\boldsymbol{\beta}}_{1,2}(\lambda),\ldots,\widehat{\boldsymbol{\beta}}_{(p-1),p}(\lambda)\right)\\
&= \underset{\substack{(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p)\\(\boldsymbol{\beta}_{12},\ldots,\boldsymbol{\beta}_{(p-1)p})}}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p} \Phi_{\ell j}\boldsymbol{\beta}_{\ell,j} \right\|_2^2 \right.\\
&\left. +\lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell \boldsymbol{\beta}_\ell + \sum_{j=\ell+1}^{p} (\partial_\ell \Phi_{\ell j})\boldsymbol{\beta}_{\ell,j} + \sum_{1\le j<\ell} (\partial_\ell \Phi_{j\ell})\boldsymbol{\beta}_{j,\ell} \right\|_2 \right),
\end{aligned} \tag{7.28}$$

où Ψ'_ℓ est une matrice de taille $n \times (K+M)$ telle que $(\Psi'_\ell)_{i,k} = B_k^{(\ell)'}(x_i^{(\ell)})$, $B_k^{(\ell)'}$ désigne la dérivée première de $B_k^{(\ell)}$. La $i^{\text{ème}}$ ligne de $(\partial_\ell \Phi_{\ell j})$ (resp. $(\partial_\ell \Phi_{j\ell})$) est définie comme étant $(\partial_\ell \Phi_{\ell j})_{i,\bullet} = ((\Psi'_\ell)_{i,\bullet} \otimes (\Psi_j)_{i,\bullet})$ (resp. $(\partial_\ell \Phi_{j\ell})_{i,\bullet} = ((\Psi_j)_{i,\bullet} \otimes (\Psi'_\ell)_{i,\bullet})$). En posant $(\partial_\ell \Phi_{\ell\bullet}) = \big((\partial_\ell \Phi_{\ell(\ell+1)}) \ldots (\partial_\ell \Phi_{\ell p})\big)$, $(\partial_\ell \Phi_{\bullet\ell}) = \big((\partial_\ell \Phi_{1\ell}) \ldots (\partial_\ell \Phi_{(\ell-1)\ell})\big)$, $\boldsymbol{\beta}_{\ell\bullet} = \left(\boldsymbol{\beta}_{\ell,(\ell+1)}^T \ldots \boldsymbol{\beta}_{\ell,p}^T\right)^T$ et $\boldsymbol{\beta}_{\bullet\ell} = \left(\boldsymbol{\beta}_{1,\ell}^T \ldots \boldsymbol{\beta}_{(\ell-1),\ell}^T\right)^T$, le terme de la pénalité peut s'écrire comme suit :

$$\lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell \boldsymbol{\beta}_\ell + (\partial_\ell \Phi_{\ell\bullet})\boldsymbol{\beta}_{\ell\bullet} + (\partial_\ell \Phi_{\bullet\ell})\boldsymbol{\beta}_{\bullet\ell} \right\|_2 =: \lambda \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell)\boldsymbol{\gamma}_\ell \right\|_2, \tag{7.29}$$

où $\boldsymbol{\gamma}_\ell = \left(\boldsymbol{\beta}_\ell^T \ \boldsymbol{\beta}_{\ell\bullet}^T \ \boldsymbol{\beta}_{\bullet\ell}^T\right)^T$. Le critère des moindres carrés peut être réécrit comme suit :

$$\begin{aligned}
&\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p} \Phi_{\ell j}\boldsymbol{\beta}_{\ell,j} \right\|_2^2\\
&= \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \frac{1}{2}\left(\sum_{\ell=1}^{p-1}\sum_{j=\ell+1}^{p} \Phi_{\ell j}\boldsymbol{\beta}_{\ell,j} + \sum_{\ell=2}^{p}\sum_{j=1}^{\ell-1} \Phi_{j\ell}\boldsymbol{\beta}_{j,\ell} \right) \right\|_2^2\\
&=: \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell \boldsymbol{\gamma}_\ell \right\|_2^2.
\end{aligned} \tag{7.30}$$

L'équation (7.30) est obtenue en définissant $\Theta_\ell = \left(\Psi_\ell \quad \frac{1}{2}\Phi_{\ell\bullet} \quad \frac{1}{2}\Phi_{\bullet\ell}\right)$ et en posant $\Theta_1 = \left(\Psi_1 \quad \frac{1}{2}\Phi_{1\bullet}\right)$ et $\Theta_p = \left(\Psi_p \quad \frac{1}{2}\Phi_{\bullet p}\right)$, où $\Phi_{\ell\bullet} = (\Phi_{\ell(\ell+1)} \ldots \Phi_{\ell p})$ et $\Phi_{\bullet\ell} = (\Phi_{1\ell} \ldots \Phi_{(\ell-1)\ell})$. En combinant (7.29) et (7.30), (7.28) peut être reformulée comme suit :

$$\left(\widehat{\boldsymbol{\gamma}}_1(\lambda),\ldots,\widehat{\boldsymbol{\gamma}}_p(\lambda)\right) = \underset{(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_p)}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell \boldsymbol{\gamma}_\ell \right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell)\boldsymbol{\gamma}_\ell \right\|_2 \right). \tag{7.31}$$

En définissant $\boldsymbol{\alpha}_\ell = (\partial_\ell \Theta_\ell)\boldsymbol{\gamma}_\ell$ et $\widetilde{\mathbf{X}}_\ell = \Theta_\ell (\partial_\ell \Theta_\ell)^+$, A^+ étant la matrice inverse de Moore-Penrose de A, (7.31) peut se réécrire comme étant :

$$\left(\widehat{\boldsymbol{\alpha}}_1(\lambda),\ldots,\widehat{\boldsymbol{\alpha}}_p(\lambda)\right) = \underset{(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_p)}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \sum_{\ell=1}^{p} \widetilde{\mathbf{X}}_\ell \boldsymbol{\alpha}_\ell \right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\| \boldsymbol{\alpha}_\ell \right\|_2 \right). \tag{7.32}$$

La dernière formulation de notre critère de sélection de variable (7.32), présenté comme ABSORBER, peut être considérée comme un problème de *group lasso* introduit par Yuan and Lin

(2006), où ici la taille p_ℓ de chaque groupe ℓ appartenant à $\{1, \ldots, p\}$ est égale à n. Les coefficients $\widehat{\boldsymbol{\gamma}}_\ell(\boldsymbol{\lambda})$ sont donc obtenus comme suit :

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

Ainsi, nous pouvons définir les variables actives pour chaque $\boldsymbol{\lambda}$ appartenant à un ensemble donné Λ comme suit :

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

où $\widehat{\gamma}_{\ell,k}(\boldsymbol{\lambda})$ est le $k^{\text{ème}}$ coefficient de $\widehat{\boldsymbol{\gamma}}_\ell(\boldsymbol{\lambda})$. On introduit également l'ensemble $\mathcal{V}_f$ des indices des d variables pertinentes dont f dans (7.1) dépend réellement et que nous cherchons à sélectionner parmi les p variables.

Notre méthode de sélection de variables a été évaluée par le biais d'expériences numériques, en définissant deux stratégies pour détecter les variables pertinentes. Les résultats sont prometteurs car ABSORBER a montré de bonnes performances de sélection, même avec des observations bruitées et un nombre accru de variables d'entrée p. De plus, notre méthode a été comparée à d'autres approches récentes à savoir les forêts aléatoires et LassoNet. Ces comparaisons ont révélé que notre méthode était plus performante que les deux autres méthodes dans le contexte d'applications à des fonctions non-linéaires et à un système géochimique, car seul ABSORBER a réussi à identifier avec succès les variables pertinentes.

7.4. Applications dans le cadre d'un groupe de travail d'EURAD

7.4.1. Contexte

Le *European Joint Programme on Radioactive Waste Management* (EURAD) est un programme qui encourage les collaborations européennes entre des organisations de 23 pays engagés dans des projets de stockage des déchets radioactifs, dans le but de développer une technologie et des connaissances robustes et durables[2] pour soutenir les activités de gestion sûre des déchets radioactifs. L'un des principaux objectifs d'EURAD est de comprendre et de quantifier l'évolution des interactions au sein de l'installation du stockage des déchets radioactifs. Ainsi, l'une des études proposées est consacrée à l'analyse des interfaces entre différents composants rocheux impliqués dans le stockage des déchets afin de soutenir la conception et l'optimisation des systèmes dite « de barrière ». À cette fin, EURAD a proposé un groupe de travail ou *work package* (WP), intitulé : « Development and Improvement of NUmerical methods and Tools for modeling coupled processes » (DONUT), qui se concentre uniquement sur les simulations numériques.

Mené par Francis Claret[3], ce groupe de travail comprend des collaborateurs de différents organisations et pays. Son objectif est de relever les défis rencontrés dans les modèles de transport réactif (RTMs), en particulier les contraintes de temps CPU et l'étude des processus couplés multi-échelles, en introduisant des méthodes de pointe pour le calcul de haute performance, comme décrit par Claret et al. (2022). Un aspect significatif de ce groupe de travail implique la comparaison des approches innovantes de modèles de substitution à travers une analyse comparative, ou *benchmark*, pour évaluer l'efficacité, la robustesse, la précision et le

[2] https://www.ejp-eurad.eu/publications/eurad-sra

[3] BRGM, 3 Avenue Claude Guillemin, 45060 Orléans, France

temps de calcul des méthodes développées. Ce *benchmark* est coordonné par Nikolaos Prasianakis[4]. Deux applications géochimiques sont considérées dans le WP DONUT : une étude sur l'uranium et l'évolution temporelle du comportement de matériaux cimentaires. La première concerne la migration des radionucléides tandis que la seconde évalue l'un des principaux matériaux existant dans les installations de stockage des déchets radioactifs. Plusieurs systèmes géochimiques avec un niveau croissant de complexité, correspondant à un nombre accru de variables d'entrée, sont définis pour simuler ces deux applications.

7.4.2. Contribution du Chapitre 5

Une partie de cette section résume l'article collaboratif :
N.I. Prasianakis, E. Laloy, D. Jacques, J.C.L. Meeussen, C. Tournassat, G.D. Miron, D. A. Kulik, A. Idiart, E. Demirer, E. Coene, B. Cochepin, M. Leconte, M. E. Savino, J. Samper II, M. De Lucia, S. V. Churakov, O. Kolditz, C. Yang, J. Samper, F. Claret. Geochemistry and Machine Learning : review of methods and benchmarking.
Bientôt soumis.

Dans ce chapitre, notre objectif est de comparer les méthodes d'estimation développées au cours de cette thèse, avec et sans sélection de variables, par rapport à deux autres méthodes de référence dans le contexte du WP DONUT, en se concentrant spécifiquement sur le cas d'application cimentaire.

Des variables d'entrée appartenant à un espace p-dimensionnel sont créées en utilisant un *Latin Hypercube Sampling* (LHS). Puis, des données de haute qualité sont générées en utilisant des solveurs géochimiques puissants tels que ORCHESTRA, GEMS et Phreeqc (voir respectivement Meeussen (2003); Kulik et al. (2013); Parkhurst and Appelo (2013) pour plus de détails sur chaque solveur). Ces ensembles de données sont utilisés à des fins d'entraînement et de validation, permettant le calcul de mesures statistiques et facilitant les comparaisons entre les différentes approches d'apprentissage automatique. Le contexte géochimique présenté ici concerne l'hydratation et l'évolution des systèmes cimentaires à 25°C et prend comme variables d'entrée les quantités de différents oxydes et la quantité d'eau. Six systèmes cimentaires sont présentés au sein de DONUT avec un nombre croissant de variables d'entrée. Dans cette section, nous ne considérons que le cas le plus complexe impliquant 7 éléments chimiques en tant que variable d'entrées. Notre objectif est d'estimer 42 éléments chimiques de sortie, liés aux espèces aqueuses, aux phases solides et à des variables auxiliaires.

Tout d'abord, la méthode de sélection de variables ABSORBER introduite dans la Section 7.3.2 a été appliquée pour identifier les variables les plus pertinentes pour chaque variable de sortie. Les résultats ont révélé des variables sélectionnées différentes selon la nature la variable de sortie considérée. Par exemple, les phases solides ont tendance à dépendre d'une ou de quelques variables seulement tandis que les phases aqueuses semblent dépendre de presque toutes les variables d'entrée. Pour simplifier l'étude, seules les variables de sortie dépendant d'au plus trois variables d'entrée ont été prises en compte pour une analyse plus approfondie. Puis, nous avons proposé d'utiliser l'approche d'apprentissage actif de la méthode reposant sur les processus gaussiens (GP AL), précédemment introduite dans la Section 7.2.2 avec ou sans prendre en compte les résultats de l'approche de sélection de variable. L'erreur de prédiction obtenue a mis en évidence les avantages de n'utiliser que les variables pertinentes sélectionnées par ABSORBER. Ensuite, nous avons utilisé GP AL, GLOBER et GLOBER-c, ces deux

[4]Laboratory for Waste Management, Paul Scherrer Institute, CH, 5232, Villigen PSI, Switzerland

derniers étant introduits dans la Section 7.2.3, pour comparer leur efficacité avec deux autres méthodes présentées par des équipes participant au WP DONUT. Les résultats ont montré une performance supérieure en notre faveur, notamment pour les sorties qui semble dépendre d'une seule variable d'entrée ce qui a mis en évidence une fois de plus l'intérêt d'utiliser nos méthodes d'estimation conjointement avec les variables sélectionnées par ABSORBER.

Chapter 8 - Appendix

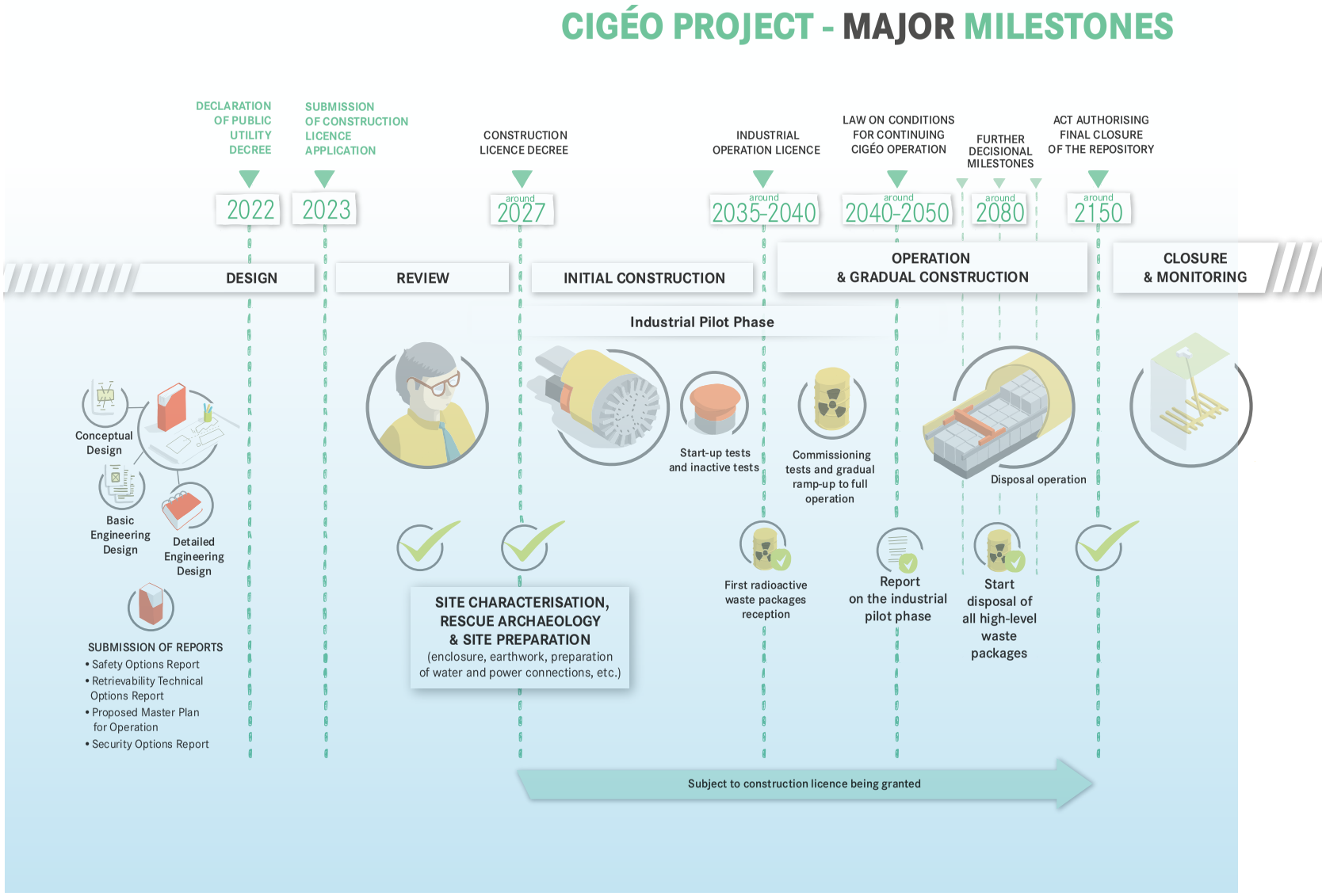

Figure 8.1: Cigéo major milestones.

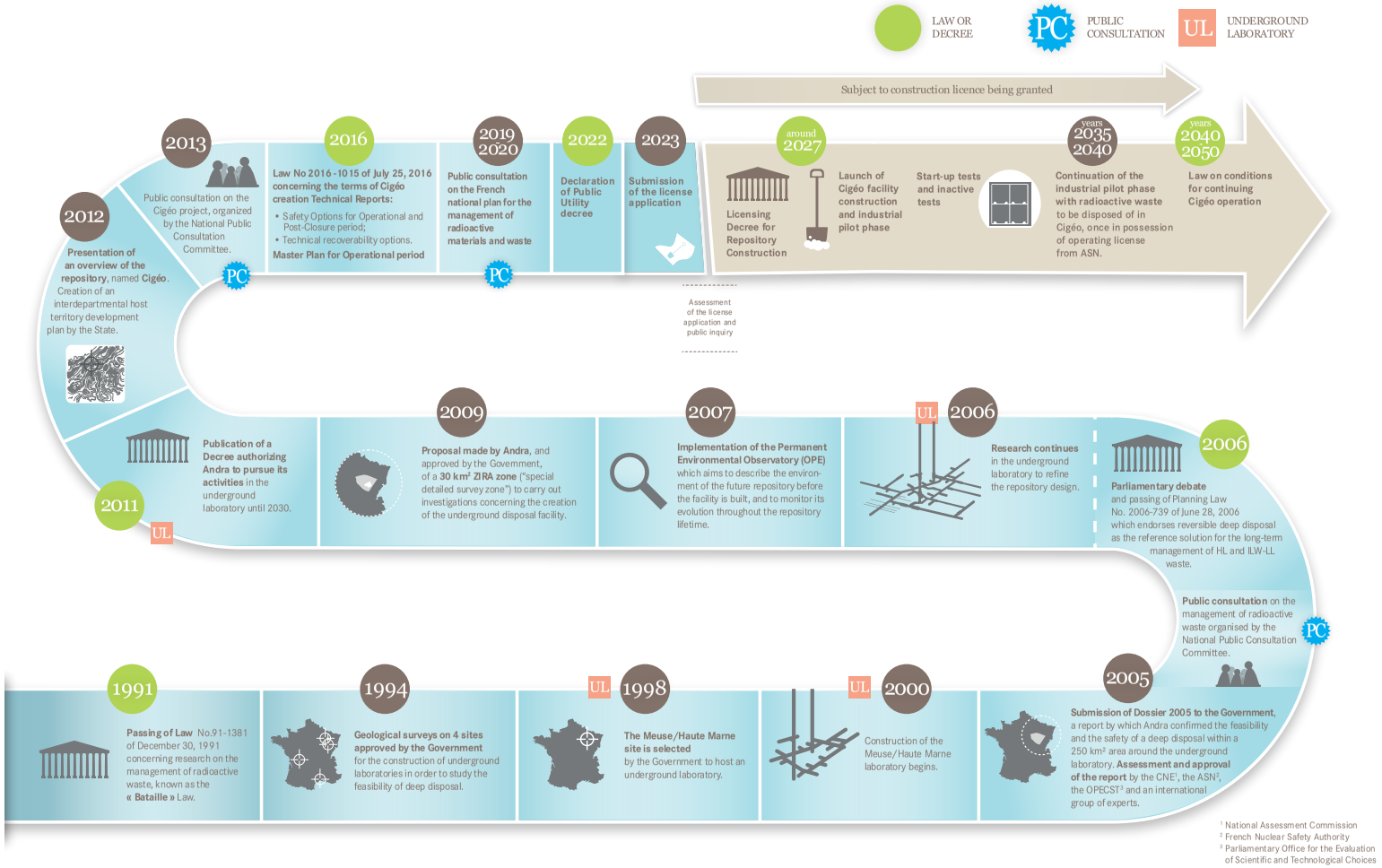

Figure 8.2: Principales dates clés du projet Cigéo.

Bibliography

Abramovitz, M. and I. Stegun (1965). Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. Dover books on mathematics. Dover Publications.

Aggarwal, C. C. et al. (2018). Neural networks and deep learning. Springer.

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.).

Allen, G. I. (2013). Automatic feature selection via weighted kernels and regularization. Journal of Computational and Graphical Statistics 22(2), 284–299.

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

Appelo, C. A. J. and D. Postma (2004). Geochemistry, groundwater and pollution. CRC press.

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

Barry, D., C. Miller, P. Culligan, and K. Bajracharya (1997). Analysis of split operator methods for nonlinear and multispecies groundwater chemical transport models. Mathematics and Computers in Simulation 43(3-6), 331–341.

Bertin, K. and G. Lecué (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. Electronic Journal of Statistics 2(none), 1224 – 1241.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. Technometrics 37(4), 373–384.

Breiman, L. (1996). Bagging predictors. Machine learning 24, 123–140.

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

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen (1984). Classification and Regression Trees. Chapman and Hall/CRC.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society Series B: Statistical Methodology 80(3), 551–577.

Carrayrou, J., R. Mosé, and P. Behra (2004). Operator-splitting procedures for reactive transport and comparison of mass balance errors. Journal of Contaminant Hydrology 68(3-4), 239–268.

Chen, J. and Z. Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95(3), 759–771.

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.

Claret, F., A. Dauzeres, D. Jacques, P. Sellin, B. Cochepin, L. De Windt, J. Garibay-Rodriguez, J. Govaerts, O. Leupin, A. Monlopez, et al. (2022). Modelling of the long-term evolution and performance of engineered barrier system. EPJ N-Nuclear Sciences & Technologies 8(41), 15.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74(368), 829–836.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association 83(403), 596–610.

Collard, N., T. Faney, P. Teboul, P. Bachaud, M. Cacas-Stentz, and C. Gout (2023). Machine learning model predicting hydrothermal dolomitisation for future coupling of basin modelling and geochemical simulations. Chemical Geology 637, 121676.

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

de Capitani, C. and T. H. Brown (1987). The computation of chemical equilibrium in complex systems containing non-ideal solutions. Geochimica et Cosmochimica Acta 51(10), 2639–2652.

De Lucia, M. and M. Kühn (2021). DecTree v1.0 – chemistry speedup in reactive transport simulations: purely data-driven and physics-based surrogates. Geoscientific Model Development Discussions 2021, 1–26.

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.

Denis, C., E. Lebarbier, C. Lévy-Leduc, O. Martin, and L. Sansonnet (2020). A novel regularized approach for functional data clustering: An application to milking kinetics in dairy goats. Journal of the Royal Statistical Society Series C: Applied Statistics 69(3), 623–640.

Descombes, S. (2001). Convergence of a splitting method of high order for reaction-diffusion systems. Mathematics of Computation 70(236), 1481–1501.

Díaz-Uriarte, R. and S. Alvarez de Andrés (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics 7, 1–13.

Eckle, K. and J. Schmidt-Hieber (2019). A comparison of deep networks with ReLU activation function and linear spline-type methods. Neural Networks 110, 232–242.

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

Eilers, P. H. and B. D. Marx (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemometrics and intelligent laboratory systems 66(2), 159–174.

Eilers, P. H., B. D. Marx, and M. Durbán (2015). Twenty years of P-splines. SORT: statistics and operations research transactions 39(2), 0149–186.

Eubank, R. L. (1999). Nonparametric regression and spline smoothing. CRC press.

Faragó, I. and J. Geiser (2007). Iterative operator-splitting methods for linear problems. International Journal of Computational Science and Engineering 3(4), 255–263.

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.

Fernández, R., D. González-Santamaría, M. Angulo, E. Torres, A. I. Ruiz, M. J. Turrero, and J. Cuevas (2018). Geochemical conditions for the formation of Mg silicates phases in bentonite and implications for radioactive waste disposal. Applied geochemistry 93, 1–9.

Forrester, A., A. Sobester, and A. Keane (2008). Engineering design via surrogate modelling: a practical guide. John Wiley & Sons.

Fox, J. (2015). Applied regression analysis and generalized linear models. Sage publications.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics 19(1), 1–67.

Galelli, S. and A. Castelletti (2013). Tree-based iterative input variable selection for hydrological modeling. Water Resources Research 49(7), 4295–4310.

Geiser, J. (2011). Iterative splitting methods for differential equations. Taylor & Francis Group: Boca Raton, FL, USA; London, UK; New York, NY, USA.

Geiser, J., J. L. Hueso, and E. Martínez (2020). Parallel iterative splitting methods: Algorithms and applications. In AIP Conference Proceedings, Volume 2293, pp. 420081. AIP Publishing LLC.

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.

Goepp, V., O. Bouaziz, and G. Nuel (2018). Spline regression with automatic knot selection. arXiv preprint arXiv:1808.01770.

Goodfellow, I., Y. Bengio, and A. Courville (2016). Deep Learning. MIT Press.

Green, P. J. and B. W. Silverman (1994). Nonparametric regression and generalized linear models: a roughness penalty approach. Monographs on statistics and applied probability 58. Boca Raton London New York etc: Chapman & Hall , 1994.

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.

Härdle, W., G. Kerkyacharian, D. Picard, and A. Tsybakov (1998). Wavelets and Approximation, pp. 71–100. New York, NY: Springer New York.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. Statistical Science 1(3), 297–310.

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).

Hernández, E. and G. Weiss (1996). A first course on wavelets. CRC press.

Hu, G., N. Prasianakis, S. V. Churakov, and W. Pfingsten (2024). Performance analysis of data-driven and physics-informed machine learning methods for thermal-hydraulic processes in full-scale emplacement experiment. Applied Thermal Engineering 245, 122836.

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

Idiart, A., M. Laviña, G. Kosakowski, B. Cochepin, J. C. Meeussen, J. Samper, A. Mon, V. Montoya, I. Munier, J. Poonoosamy, et al. (2020). Reactive transport modelling of a low-pH concrete/clay interface. Applied geochemistry 115, 104562.

Jala, M., C. Levy-Leduc, Éric Moulines, E. Conil, and J. Wiart (2016). Sequential design of computer experiments for the assessment of fetus exposure to electromagnetic fields. Technometrics 58(1), 30–42.

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.

Kanney, J. F., C. T. Miller, and C. T. Kelley (2003). Convergence of iterative split-operator approaches for approximating nonlinear reactive transport problems. Advances in Water Resources 26(3), 247–261.

Karniadakis, G. E., I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang (2021). Physics-informed machine learning. Nature Reviews Physics 3(6), 422–440.

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.

Kosakowski, G. and U. Berner (2013). The evolution of clay rock/cement interfaces in a cementitious repository for low-and intermediate level radioactive waste. Physics and Chemistry of the Earth, Parts A/B/C 64, 65–86.

Kulik, D. A., T. Wagner, S. V. Dmytrieva, G. Kosakowski, F. F. Hingerl, K. V. Chudnenko, and U. R. Berner (2013). GEM-Selektor geochemical modeling package: revised algorithm and GEMS3K numerical kernel for coupled simulation codes. Computational Geosciences 17, 1–24.

Kyas, S., D. Volpatto, M. O. Saar, and A. M. Leal (2022). Accelerated reactive transport simulations in heterogeneous porous media using Reaktoro and Firedrake. Computational Geosciences 26(2), 295–327.

Lagneau, V. and J. van Der Lee (2010). Operator-splitting-based reactive transport models in strong feedback of porosity change: The contribution of analytical solutions for accuracy validation and estimator improvement. Journal of contaminant hydrology 112(1-4), 118–129.

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.

Leal, A. M., S. Kyas, D. A. Kulik, and M. O. Saar (2020). Accelerating reactive transport modeling: on-demand machine learning algorithm for chemical equilibrium calculations. Transport in Porous Media 133(2), 161–204.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. Nature 521(7553), 436–444.

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.

Li, Z. and J. Cao (2022). General P-splines for non-uniform B-splines. arXiv preprint arXiv:2201.06808.

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.

Loh, W.-Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery 1(1), 14–23.

Lu, R., T. Nagel, J. Poonoosamy, D. Naumov, T. Fischer, V. Montoya, O. Kolditz, and H. Shao (2022). A new operator-splitting finite element scheme for reactive transport modeling in saturated porous media. Computers & Geosciences 163, 105106.

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.

Mallat, S. (1999). A wavelet tour of signal processing. Elsevier.

Marchuk, G. (1990). Splitting and alternating direction methods. In Handbook of Numerical Analysis, Volume 1, pp. 197–462. Elsevier.

Meeussen, J. C. (2003). ORCHESTRA: An object-oriented framework for implementing chemical equilibrium models. Environmental science & technology 37(6), 1175–1182.

Meyer, Y. (1993). Wavelets and Operators. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Moyce, E. B., C. Rochelle, K. Morris, A. E. Milodowski, X. Chen, S. Thornton, J. S. Small, and S. Shaw (2014). Rock alteration in alkaline cement waters over 15 years and its relevance to the geological disposal of nuclear waste. Applied Geochemistry 50, 91–105.

Nadaraya, E. A. (1964). Some new estimates for distribution functions. Theory of Probability & Its Applications 9(3), 497–500.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. Statistical Science 1(4), 502–518.

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.

Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3), 1065–1076.

Peng, H., F. Long, and C. Ding (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence 27(8), 1226–1238.

Piegl, L. and W. Tiller (2012). The NURBS book. Springer Science & Business Media.

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.

Raissi, M., P. Perdikaris, and G. E. Karniadakis (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational physics 378, 686–707.

Rao, S., A. van der Schaft, K. van Eunen, B. M. Bakker, and B. Jayawardhana (2013). Model-order reduction of biochemical reaction networks. In 2013 European Control Conference (ECC), pp. 4502–4507. IEEE.

Rasmussen, C. E. and C. K. I. Williams (2006). Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

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), W07401.

Rice, J. (1984). Bandwidth choice for nonparametric regression. The Annals of Statistics 12, 1215–1230.

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society 58(5), 527 – 535.

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.

Rosasco, L. A., S. Villa, S. Mosci, M. Santoro, and A. Verri (2013). Nonparametric sparsity and regularization. Journal of Machine Learning Research 14(1), 1665–1714.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65(6), 386.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27(3), 832 – 837.

Roth, V. (2004). The generalized LASSO. IEEE transactions on neural networks 15(1), 16–28.

Sadhanala, V., Y.-X. Wang, A. J. Hu, and R. J. Tibshirani (2021). Multivariate trend filtering for lattice data. arXiv preprint arXiv:2112.14758.

Samper, J., A. Naves, L. Montenegro, and A. Mon (2016). Reactive transport modelling of the long-term interactions of corrosion products and compacted bentonite in a HLW repository in granite: Uncertainties and relevance for performance assessment. Applied Geochemistry 67, 42–51.

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.

Simpson, M. J. and K. A. Landman (2007). Analysis of split operator methods applied to reactive transport with monod kinetics. Advances in Water Resources 30(9), 2026–2033.

Smith, W. R. (1980). The computation of chemical equilibria in complex systems. Industrial & Engineering Chemistry Fundamentals 19(1), 1–10.

Sportisse, B. (2000). An analysis of operator splitting techniques in the stiff case. Journal of computational physics 161(1), 140–168.

Srinivas, N., A. Krause, S. Kakade, and M. Seeger (2012). Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. IEEE Information Theory 58, 3258–3265.

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

Steefel, C. I., C. A. J. Appelo, B. Arora, D. Jacques, T. Kalbacher, O. Kolditz, V. Lagneau, P. Lichtner, K. U. Mayer, J. Meeussen, et al. (2015). Reactive transport codes for subsurface environmental simulation. Computational Geosciences 19, 445–478.

Steefel, C. I., D. J. DePaolo, and P. C. Lichtner (2005). Reactive transport modeling: An essential tool and a new research approach for the earth sciences. Earth and Planetary Science Letters 240(3-4), 539–558.

Steefel, C. I. and K. McQuarrie (1996). Approaches to modeling of reactive transport in porous media. Reviews in mineralogy 34, 83–130.

Stein, M. L. (1999). Interpolation of spatial data. Springer Series in Statistics. Springer-Verlag.

Stone, C. J., M. H. Hansen, C. Kooperberg, and Y. K. Truong (1997). Polynomial splines and their tensor products in extended linear modeling. The Annals of Statistics 25(4), 1371–1425.

Stone, C. J. and C.-Y. Koo (1985). Additive splines in statistics. Proceedings of the American Statistical Association Original 45, 48.

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

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. The Annals of Statistics 42(1), 285–323.

Tibshirani, R. J. and J. Taylor (2011). The solution path of the generalized lasso. The Annals of Statistics 39(3), 1335 – 1371.

Tsybakov, A. B. (2009). Introduction to Nonparametric Regression, Volume 41. Springer, New York.

Valocchi, A. J. and M. Malmstead (1992). Accuracy of operator splitting for advection-dispersion-reaction problems. Water Resources Research 28(5), 1471–1476.

Vapnik, V., S. Golowich, and A. Smola (1996). Support vector method for function approximation, regression estimation and signal processing. In Advances in Neural Information Processing Systems, Volume 9. MIT Press.

Wahba, G. (1990). Spline Models for Observational Data. Society for Industrial and Applied Mathematics.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. Computational Statistics 15, 443–462.

Wand, M. P. and J. T. Ormerod (2008). On semiparametric regression with O'Sullivan penalized splines. Australian & New Zealand Journal of Statistics 50(2), 179–198.

Wang, Y. (2011). Smoothing Splines: Methods and Applications. CRC press.

Wasserman, L. (2006). All of Nonparametric Statistics. Springer Science & Business Media.

Watson, G. S. (1964). Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A 26(4), 359–372.

White, W. B., S. M. Johnson, and G. B. Dantzig (1958). Chemical equilibrium in complex mixtures. The Journal of Chemical Physics 28(5), 751–755.

Wilson, J. C., S. Benbow, and R. Metcalfe (2018). Reactive transport modelling of a cement backfill for radioactive waste disposal. Cement and Concrete Research 111, 81–93.

Wood, S. N. (2017). Generalized Additive Models: an introduction with R. Chapman and Hall/CRC.

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.

Yeh, G. and V. Tripathi (1989). A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components. Water resources research 25(1), 93–108.

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.

Yuan, Y., N. Chen, and S. Zhou (2013). Adaptive B-spline knot selection using multi-resolution basis set. IIE Transactions 45(12), 1263–1277.

Zhang, W. and A. T. Goh (2013). Multivariate adaptive regression splines for analysis of geotechnical engineering systems. Computers and Geotechnics 48, 82–95.

Zhang, W. and A. T. Goh (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geoscience Frontiers 7(1), 45–52.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476), 1418–1429.

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.