UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

University Paris 1 Panthéon-Sorbonne

UFR 27 : Department of Mathematics and Computer Science

Major subject : Applied mathematics

Statistical learning and sensitivity analysis - Application to Oil and Gas production

Author:
Acharki Naoufal

Academic supervisor:
Prof. Garnier Josselin

Industrial supervisors:
Bertoncello Antoine
Agharzayeva Zinyat

A thesis submitted for the degree of

Master M2 : Mathematical Models and Methods of Economics and Finance

September 2019

Abstract

Predicting future's well production is a significant challenge. The ability to anticipate a well's performance helps operating companies plan long-term investments and optimize production strategies. Classical models of Machine Learning have been tested to analyze well production (e.g. Random Forest and Gradient Boosting). Although the results are promising, uncertainty was not accurately quantified, and influent variables were incorrectly identified.

Our main tool to understand the influence of inputs is the global sensitivity analysis, combining both variance-based measures : Sobol indices [I. Sobol, 1993] and Shapley values [Shapley, 1953], and dependence measures with Hilbert-Schmidt Independence Criterion [A. Gretton, 2005].

The Gaussian Processes (GP) are one of the most important Bayesian Machine Learning methods providing a probabilistic approach for supervised learning in kernel space functions Reproducing Kernel Hilbert Spaces (RKHS) [C.E. Rasmussen and C.K.I. Williams, 2006]. It has the advantage of interpolating, being interpretable in terms of predictions/uncertainty and estimating sensitivity indices faithfully.

However, the GP model building process remains challenging in the case of high-dimensional data. We deal with this problem by following a new methodology, proposed by B. Iooss and A. Marrel [2017], allowing to build a GP model with numerous inputs in an efficient manner by screening and performing joint modelling.

In this study, we apply the Gaussian Process (GP) model on unconventional fields with the goal of having the best predictions on "Production over 12 months" with accurate uncertainty. The GP hyperparameters were optimized to get approximately the same accuracy Q^2 as others Machine Learning (ML) models. It was also optimized so that Prediction Intervals achieve appropriate confidence levels, unlike Random Forest (overestimated) or Gradient Boosting (underestimated).

Keywords : • Uncertainty Quantification • Gaussian Processes • Sensitivity analysis • Machine Learning • Variables selection.

Résumé

La prévision de la production future de pétrole et de gaz des puits est un défi majeur. La capacité d'anticiper la performance d'un puits permet les compagnies pétroliéres à planifier leurs investissements à long terme et à optimiser leurs stratégies de production. Des modèles d'apprentissage machine ont été testés pour analyser la production (p. ex. Random Forest et GradientBoosting). Bien que les résultats soient prometteurs, les quantiles n'ont pas été quantifiée avec précision et les variables influentes ont été mal identifiées.

L'analyse de sensibilité globale constitue notre outil principal pour mesurer l'influence des variables d'entrée. elle combine les mesures basées sur la variance : Les indices de Sobol [I. Sobol, 1993] et les valeurs de Shapley [Shapley, 1953], en plus des mesures de dépendance, notamment le critère d'indépendance de Hilbert-Schmidt HSIC [A. Gretton, 2005].

Les Processus Gaussiens GP représentent l'une des méthodes Bayesiennes d'apprentissage plus importantes. ils se fondent sur une approche probabiliste pour l'apprentissage contrôlé dans l'espace des noyaux RKHS [C.E. Rasmussen et C.K.I. Williams, 2006]. Ils ont l'avantage d'interpoler, d'être interprétabled en termes de prédictions/incertitudes et d'estimer fidèlement les indices de sensibilité.

Toutefois, le modèle du krigeage par PG devient très couteux lorsqu'il s'agit des données de grande dimension. Pour contourner ce problème, nous suivons une méthodologie proposée par B. Iooss et A. Marrel [2017], consistant à construire un modèle GP avec de nombreuses variables entrées efficacement en criblant ces entrées et en effectuant une modélisation conjointe.

Dans cette étude, nous appliquons le modèle de krigeage par Processus Gaussians GP dans des champs non-conventionnels.Le but étant de prédire la production petrolière aprés 12 mois avec une incertitude précise. Les paramètres du modèle ont été optimisés de telle sorte à avoir approximativement la même précision Q^2 que les autres modèles d'apprentissage machine. Ces paramètres été optimisés également pour que les intervalles de prédiction atteignent des niveaux de confiance appropriés, contrairement à ForestRandom Forest (sur-estimés) ou Gradient Boosting (sous-estimés).

Mots clés : • Quantification des incertitudes • Processus Gaussiens • Analyse de sensibilité • Apprentissage Machine • Sélection de variables.

Acknowledgements

At the end of this work, I would like to express my gratitude and my thanks to all those who contributed to its realization:

Pr. Josselin Garnier, for his accompaniment, his exchanges and his availability all long these months;

Pr. Olivier Gueant, for his influence, his follow-up and his precious advices to seize this opportunity;

Pr. Olivier Roustant to whom I owe my passion and love for data science;

Antoine Bertoncello for guiding me to data analytics and Machine Learning research world;

Achraf Ourir and Zinyat Agharzayeva for their kidness, support and welcoming me inside "Geo-statistics and Uncertainty" team;

All trainees and other people of "Gisements" division that I met during this internship.

N. Acharki

Contents

List of Figures

List of Tables

Lists of acronyms

CV Cross Validation

FAST Fast Amplitude Sensitivity Test

GP Gaussian Process

GSA global sensitivity analysis

HS Hilbert-Schmidt

HSIC Hilbert-Schmidt Independence Criterion

LHS Hypercube Latin Sampling

LOO Leave-One-Out

MAE mean absolute error

ML Machine Learning

MLE Maximum Likelihood Estimator

MSE Mean Squared Error

PCA Principal Components Analysis

PCC Partial Correlation Coefficient

PRCC Partial Rank Correlation Coefficient

PRMS Petroleum Resource Management System

RKHS Reproducing Kernel Hilbert Spaces

RMSE Root Mean Squared Error

SA Sensitivity Analysis

SEC Securities and Exchange Commission

SFD space filling designs

SRC Standard Regression Coefficient

SRRC Standard Rank Regression Coefficient

Chapter 1

Introduction

1.1 General context

A fundamental challenge of oil and gas companies is to predict how much oil and gas they will produce in the future. It drives both their exploration and development strategy. It is also used as a critical metric (reserves) by investors to assess the company's value. Yet, forecasting a well's future production is challenging because subsurface reservoirs properties are never fully known. Consequently, a crucial task of reserve oil engineers is to estimate wells production with their associated uncertainty correctly. Modelling production profile is traditionally done by exponential *Decline Curve Analysis* [2] to fit a decline curve and estimate future oil production. Still, this standard approach does not take into account the specificities of each well.

TOTAL seeks to improve the evaluation and develop more precise and realistic models. To this end, TOTAL is investigating the use of a data-driven approach to forecasting a well's production. Several classical models of Machine Learning ML algorithms have been tested to analyze a well's production (e.g. Random Forest, Gradient Boosting). Though promising in terms of accuracy, the results did not give entire satisfaction. Uncertainty was not correctly quantified, and in particular, the extreme forecasts were poorly estimated when compared to field data. In addition, influent variables were incorrectly identified. One of the reasons might be that these models capture only correlations but not causes-effect links.

Developing ML methods that honour uncertainty is critical for Total. Indeed, reservoirs are heterogeneous and uncertain; all reserve estimates involve uncertainty which stems mainly from the lack of geological data and engineering data when the field is not explored and developed yet. Thus, uncertainty quantification is one of the most critical tasks. It allows companies to keep track of how much oil and gas is still left in their prospects and update their investors about their predictions of future gains/losses in both optimistic and pessimistic cases.

It is common in the gas and oil industry to describe the relative degree of uncertainty in terms of a low (P_{90})/high (P_{10}) range. This is consistent with both the **Petroleum Resource Management System (PRMS)** and the **Securities and Exchange Commission (SEC)**. Both define the reserves and resources estimates in terms of $P_{90}/P_{50}/P_{10}$ ranges :

The range of uncertainty of the recoverable and/or potentially recoverable volumes may be represented by either deterministic scenarios or by a probability distribution. When the range of uncertainty is represented by a probability distribution, a low, best, and high estimate shall be provided such that:

- *There should be at least a 90 percent probability (P_{90}) that the quantities actually recovered will equal or exceed the low estimate –proved, the highest figure–*
- *There should be at least a 50 percent probability (P_{50}) that the quantities actually recovered will equal or exceed the best estimate –median–*

- *There should be at least a 10 percent probability (P_{10}) that the quantities actually recovered will equal or exceed the high estimate –possible, the lowest figure–*

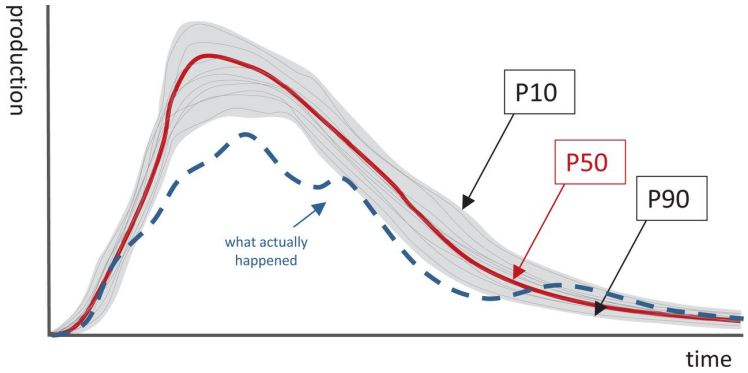


Figure 1.1 – Illustration of $P_{90}/P_{50}/P_{10}$ ranges for production decline curve

These agencies (**PRMS** and **SEC**) define rules for defining $P_{90}/P_{50}/P_{10}$ reserves estimates to be disclosed to security investors for publicly traded oil and gas companies. The primary objective is to provide investors with consistent information and associated value assessments obtained under the same assumptions to easily compare the financial performance of petroleum companies. That also allows companies to manage their oil and gas portfolios (for those listed on U.S. stock exchanges, they must also estimate proved reserves under **SEC** guidelines).

It is essential to know that reserves are the main assets of gas and oil companies. They are anticipated to be commercially recoverable from known accumulations after a given date. Their volumes and their associated monetary values are extremely important to the upstream petroleum industry. Better reserves estimates support better decisions in transactions, reservoir management, and asset portfolio management. Conversely, poor estimates can have very dangerous consequences on investors' confidence and may cause a dramatic drop in the stock market.

A good predictive model must be accurate and respect as much as possible the quantiles P_{10}/P_{90}. Still, targeting specific quantiles such as P_{10}, P_{50} and P_{90} realizations is a related challenge. The situation is more complex for a production forecast when the forecast is a timeline and not a scalar. There are many statistical methods to establish P_{10} and P_{90} quantiles, such as the Monte Carlo method or Bagging/Boostrap. However, they remain unpractical for some Machine Learning models (e.g. Gradient Boosting and Neural Networks), which require heavy computing resources for each simulation.

1.2 Objective and approaches

Production modelling depends on a large number of parameters and variables, including exploitation conditions and reservoir geological properties. The objective is to get an accurate predictive model of the oil/gas cumulative production at some given dates, or the max oil/gas rate, from the input variables describing the environment and well's characteristics. It must also quantify the uncertainty correctly and respect the constraint of P_{10}/P_{90}.

We propose to tackle the problem through the following approach :

First, we apply a sensitivity analysis to extract the influent input variables. In addition to variance-based measures, we explore a new class of sensitivity indices based on dependence measures, such as distance correlation and the Hilbert-Schmidt Independence Criterion. These two methods represent an alternative to screening in high dimension than Pearson's and Spearman's coefficients, which makes sorting the inputs more efficient.

Second, it seems that GP should be tractable in our setting. For this, we consider a new methodology proposed that makes it possible to build a Gaussian process model with a large number of inputs in a very efficient manner: It uses the screening results to sort the inputs so that the sorted inputs are successively included in the group of explanatory inputs for the Gaussian process model. In contrast, the other inputs are considered as global stochastic (i.e. unknown) inputs.

Finally, we optimize GP hyper-parameters by two different methods : Maximum Likelihood Estimator (MLE) and Cross Validation (CV). We get approximately the same accuracy R-squared as Random Forest or Gradient Boosting. We optimize also GP hyper-parameters by Cross-Validation to fit confidence level and respect consequently the definition of P_{10}/P_{90}.

Chapter 2

Global Sensitivity Analysis measures

Contents

In many engineering and research fields, using mathematical models or approximating physical models by surrogate models is crucial to describe phenomena (B. Sudret [3]). Nevertheless, the input values are often known only to some degree of uncertainty and are therefore described as random variables. The goal is then to understand the influence of inputs :

- Identify and prioritize the most influential inputs,
- Identify non-influential inputs in order to fix them to nominal values,
- calibrate some model inputs using some available information (real output observations, constraints, etc.).

Sensitivity Analysis (SA) methods are invaluable tools. They allow studying how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input (A. Saltelli *et al.* [4]). It may be used to determine the most contributing input variables to an output behaviour or to understand/interpret the dependencies and interaction structure of the model. Sensitivity Analysis can be helpful in many situations; one can mention model understanding by reducing the variance of the most influential inputs or simplifying by decreasing the number of variables in the model.

An easy and very intuitive way to examine the influence of input parameters is local sensitivity analysis, considered to be the first historical approach to sensitivity analysis SA. It consists of studying the impact of small input perturbations on the model output by calculating or estimating the partial derivatives of the model at specific points of interest. Unlike local sensitivity analysis, global sensitivity analysis (GSA) provides information about the influence due to variation over the whole system and offers a comprehensive approach to the model analysis.

The global sensitivity indices were suggested by (I. Sobol [5] [6]) in the early 1990s with variance-based methods, and then further developed by (A. Saltelli and T. Homma [7]). They have been considered as one of the most efficient and popular global SA techniques for a while. There's also Derivative based Global Sensitivity Measures (I. Sobol and S. Kucherenko, [8], I. M. Sobol and S. Kucherenko [9]) which have shown more efficiency and accuracy in the sensitivity analysis (B. Iooss and P. Lemaître [10]). Recently, a new class of sensitivity indices based on dependence measures is proposed (A. Gretton *et al.* [11], S. Da Veiga [12]) which overcomes some disadvantages of other sensitivity classes, especially in high dimension.

2.1 Importance measures (sampling-based global methods)

A naive way to answer the previous questions is the importance measures by fitting a model on the output Y given inputs $X = (X_1, \cdots, X_p)$, provided that the sample size n is sufficiently large, and studying this fitted model. For linear models, the main indices are:

- Pearson Correlation Coefficient :

$$\rho(X_k, Y) = \frac{\sum_{i=1}^{n} \left(X_k^{(i)} - \overline{X}_k\right)\left(Y^{(i)} - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_k^{(i)} - \overline{X}_k\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y^{(i)} - \overline{Y}\right)^2}} \tag{2.1}$$

  where $\overline{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_k^{(i)}$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y^{(i)}$ are the empirical mean of X_k and Y.
  Partial Correlation Coefficient (PCC) provides a linearity measure between variable X_k and output Y. If X_k and Y are independents, $\rho(X_k, Y)$ equals 0 but the reverse is not true.

- Standard Regression Coefficient (SRC) :

$$SRC_k = \beta_k \sqrt{\frac{\mathrm{Var}(X_k)}{\mathrm{Var}(Y)}} \tag{2.2}$$

  where β_k is the linear regression coefficient associated to X_k . It represents a share of variance if the linearity hypothesis is confirmed.

- PCC :

$$PCC_k = \rho(X_k - \hat{X}_{-k}, Y - \hat{Y}_k) \tag{2.3}$$

  where $\hat{X}_{-k}$ is the prediction of the linear model on X_k with respect to the other inputs X_{-k} and $\hat{Y}_k$ is the prediction of the linear model where X_k is missing. PCC_k measures the sensitivity of Y to X_k when the effects of the other inputs have been removed.

These linear and rank-based measures are part of the so-called sampling based global sensitivity analysis method. If the model is not linear but still monotonic, a rank transformation can be applied to the three measures by replacing the values by their ranks in each column of the matrix. Giving analogously, the Spearman's Correlation Coefficient ρ^S, the Standard Rank Regression Coefficient (SRRC), the Partial Rank Correlation Coefficient (PRCC) and Kendall rank correlation coefficient τ.

However, in practice the models are frequently non-linear and non-monotonic, the importance measures could fail to describe the influence of inputs.

2.2 Variance-based methods for sensitivity analysis

2.2.1 Sobol indices for independent variables

Consider the following model:

$$Y = f(X_1, .., X_p) \tag{2.4}$$

where the output Y is a scalar, f is a measurable function describing the model and the input factors $X_1, .., X_p$ are supposed at this part to be independent random variables described by known probability distributions.

An intuitive way to define the importance of input X_i is to analyze how the model output Y changes for different values of X_i. The idea is to compare the variance $\mathrm{Var}(Y)$ with the variance of Y knowing X_i (i.e the variance of the conditional expectation Y on X_i), these are the Sobol's indices.

The original idea behind the Sobol's indices is to represent the model as a sum of component functions with increasing dimensionality, and then decompose the output variance into the contribution associated with each input factor.

Definition 2.2.1 (Hoeffding decomposition of variance [13]) *Let us assume that the function f in (2.4) is L^2-integrable function over the unit hypercube $[0,1]^p$. It is possible to represent this function as a sum of elementary functions:*

$$f(\mathbf{X}) = f_0 + \sum_{i=1}^{p} f_i\left(X_i\right) + \sum_{i<j}^{p} f_{ij}\left(X_i, X_j\right) + \cdots + f_{12\ldots p}(\mathbf{X}) \tag{2.5}$$

This decomposition is unique under some conditions (Sobol [83]):

$$\int_0^1 f_{i_1\ldots i_s}\left(x_{i_1}, \ldots, x_{i_s}\right) dx_{i_k} = 0, 1 \leq k \leq s, \{i_1, \ldots, i_s\} \subseteq \{1, \ldots, p\} \tag{2.6}$$

In the Sensitivity Analysis framework, let us have the random vector $\mathbf{X} = (X_1, ..., X_p)$ where the variables are mutually independent, and the output $Y = f(\mathbf{X})$ of a deterministic model $f(\cdot)$. It can be shown that the variance of the output, $\mathrm{Var}(Y)$, can also be decomposed according to this functional decomposition, often referred to as functional ANOVA (B. Efron and C. Stein [14]):

$$\mathrm{Var}(Y) = \sum_{i=1}^{p} V_i(Y) + \sum_{1\leq i<j\leq p} V_{ij}(Y) + \cdots + V_{1,\ldots,p}(Y) \tag{2.7}$$

where $V_i(Y), V_{ij}(Y), ..., V_{1,2,\ldots,p}(Y)$ denote the variance of $f_i(Y), f_{ij}(Y), ..., f_{1,\ldots,p}(Y)$ respectively :

$$\begin{aligned}
V_i(Y) &= \mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right] \\
V_{ij}(Y) &= \mathrm{Var}\left[\mathbb{E}\left(Y|X_i, X_j\right)\right] - V_i(Y) - V_j(Y) \\
V_{ijk}(Y) &= \mathrm{Var}\left[\mathbb{E}\left(Y|X_i, X_j, X_k\right)\right] - V_{ij}(Y) - V_{ik}(Y) - V_{jk}(Y) - V_i(Y) - V_j(Y) - V_k(Y) \\
&\vdots \\
V_{1,\ldots,p}(Y) &= \mathrm{Var}(Y) - \sum_{i=1}^{p} V_i(Y) - \sum_{1\leq i<j\leq p} V_{ij}(Y) - \sum_{1\leq i_1<\ldots<\leq i_{p-1}\leq p} V_{i_1 i_{p-1}}(Y)
\end{aligned} \tag{2.8}$$

From this decomposition (2.7), The so-called "Sobol' indices" or "Variance-based sensitivity indices" (I. Sobol [5]) can be naturally obtained by dividing on $\mathrm{Var}(Y)$. These indices express the share of the variance of Y that is due to a given input or input combination.

Note that the first-order indexes S_i can be deduced from the first p terms of the decomposition (2.7) :

Definition 2.2.2 (Sobol first-order index) *The first-order sobol index of an input X_i is defined as the first term of the Hoeffding decomposition.*

$$S_i = \frac{V_i(Y)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right]}{\mathrm{Var}(Y)} \tag{2.9}$$

It measures how the expected value of Y varies for different values of X_i compared to the total variation of Y. The larger this quantity, the more important the contribution of X_i to the variance of Y.

The conditional expectation Y on X_i can be computed as the average of the model evaluations $f(.)$ from a sample of $X_{\sim i} = (X_1, ..., X_{i-1}, X_{i+1}, .., X_p)$ and a given $X_i = x$.

$$\mathbb{E}\left(Y|X_i\right) = g(X_i) \tag{2.10}$$

where $g(x) = \mathbb{E}\left(Y|X_i = x\right)$ for $x \in I$

The second-order sensitivity index, S_{ij}, expresses the amount of variance of Y explained by the interaction of the factors X_i and X_j subtracting order indices:

$$S_{ij} = \frac{V_{ij}(Y)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i, X_j\right)\right] - S_i - S_j}{\mathrm{Var}(Y)} \tag{2.11}$$

Consequently, for a subset u of k-indices it is possible to define the k-th order index:

$$S_u = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_u\right)\right]}{\mathrm{Var}(Y)} - \sum_{w \subseteq u} S_w \tag{2.12}$$

and so on until order p

A. Saltelli and T. Homma [7] define in particular the total sobol index of an input X_i is defined as :

$$S_{T_i} = S_i + \sum_{i<j} S_{ij} + \sum_{j \neq i, k \neq i, j<k} S_{ijk} + \ldots = \sum_{u \in \# i} S_u \tag{2.13}$$

To estimate Sobol' indices, many techniques have been developed including Fast Amplitude Sensitivity Test (FAST) due to H. Cukier *et al.* [15] and [16], and Monte Carlo sampling-based methods : I. Sobol [6] for first order and interaction indices and A. Saltelli [17] for first order and total indices.

However, FAST remains costly, unstable and biased when the number of inputs increases (larger than 10) (Tissot, J-Y. and Prieur, C. [18]). Unlike a Monte Carlo method which still feasible with high inputs data and provides less when performed with random repetition (B. Iooss *et al.* [19])

We recall the strong law of large numbers commonly used in Monte-Carlo method :

Theorem 1 (Strong Law of Large numbers) *Let X be a real-valued random variable. Let X_1,...,X_n be a sequence of i.i.d variables with the same law as X, and assume that $\mathbb{E}(X) = \mu$ exists and is finite.*

$$\textit{Then} \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{a.s.}}{\to} \mu = \mathbb{E}(X) \quad \textit{when } n \to \infty \tag{2.14}$$

This law makes it possible to estimate the expectation of any function of a random variable X by the estimator :

$$\hat{\mathbb{E}}[f(X)] = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \tag{2.15}$$

where $(x_i)_{i=1..n}$ is a n-sample realizations of the random variable X.

The rate of convergence of Monte Carlo method is in $O(\sqrt{n})$. Many alternative methods have been proposed to improve convergence's rate, in particular Quasi-Monte Carlo, with a rate of $O(n^{-\frac{3}{2}}(log(n))^{\frac{p-1}{2}})$ (A. Owen [20]).

Estimating Sobol's indices requires the estimation of the variance of the conditional expectation. An estimating technique called *Pick and Freeze* 2 due to I. Sobol [6] is useful in this case.

Lemma 2 (Pick and Freeze, I. Sobol [5]) *Let $\mathcal{J} \subseteq \{1,..,p\}$ and $\mathbf{X} = (X_1,..,X_p) = (X_{\mathcal{J}}, X_{\overline{\mathcal{J}}})$ be a random vector of independent variables and let $Y = f(X_{\mathcal{J}}, X_{\overline{\mathcal{J}}})$, then :*

$$\mathrm{Var}\left[\mathbb{E}\left(Y|X_{\mathcal{J}}\right)\right] = \mathrm{Cov}(Y, Y^{\mathcal{J}}) \tag{2.16}$$

where $Y^{\mathcal{J}} = f(X_{\mathcal{J}}, (X_{\overline{\mathcal{J}}})')$ and $(X_{\overline{\mathcal{J}}})'$ is an independent copy of $X_{\overline{\mathcal{J}}}$ (i.e random vector independent of $X_{\overline{\mathcal{J}}}$ with the same distribution).

The first-order sensitivity indices (See Eq.2.9) can be expressed then as :

$$S_i = \frac{\mathrm{Cov}(Y, Y^{(i)})}{\mathrm{Var}(Y)} = \frac{\mathbb{E}(YY^{(i)}) - \mathbb{E}(Y)\mathbb{E}(Y^{(i)})}{\mathrm{Var}(Y)} = \frac{\mathbb{E}(YY^{(i)}) - \mathbb{E}(Y)^2}{\mathrm{Var}(Y)} \tag{2.17}$$

By considering an n-sample $\widehat{X}_n = (x_{k1},...,x_{kp})_{k=1..n}$ of realizations of input variables $(X_1,...,X_p)$ and $\widehat{Y}_n = (Y_k)_{k=1..n} = f\left(x_{k1},...,x_{kp}\right)_{k=1..n}$, a natural estimator of S_i consists in taking the empirical estimators of the mean $\mathbb{E}(Y) = \mu$ and of the variance $\mathrm{Var}(Y) = V$:

$$\hat{S}_i = \frac{\frac{1}{n}\sum_{k=1}^{n} Y_k Y_k^{(i)} - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right)^2}{\frac{1}{n}\sum_{k=1}^{n} Y_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right)^2} \tag{2.18}$$

where

$$\hat{\mu} = \tfrac{1}{n}\textstyle\sum_{k=1}^{n} Y_k = \tfrac{1}{n}\sum_{k=1}^{n} f(x_{k1},...,x_{kp})$$
$$\hat{V} = \tfrac{1}{n}\textstyle\sum_{k=1}^{n} Y_k^2 - \left(\sum_{k=1}^{n} Y_k\right)^2 = \tfrac{1}{n}\sum_{k=1}^{n} f^2(x_{k1},...,x_{kp}) - \hat{\mu}^2$$

Estimating the term $U_i = \sum_{k=1}^{n} Y_k Y_k^{(i)}$ can be done by using two samples of realizations of the input variables $\hat{X}_{n,i}^{(1)}$ and $\hat{X}_{n,i}^{(2)}$:

$$\hat{X}_{n,i}^{(1)} = (X_{k,i}^{(1)}, X_{k,\sim i}^{(1)})_{k=1..n} = (x_{k1}^{(1)},\ldots,x_{k(i-1)}^{(1)}, x_{ki}^{(1)}, x_{k(i+1)}^{(1)},\ldots,x_{kp}^{(1)})_{k=1..n}$$
$$\hat{X}_{n,i}^{(2)} = (X_{k,i}^{(1)}, X_{k,\sim i}^{(2)})_{k=1..n} = (x_{k1}^{(2)},\ldots,x_{k(i-1)}^{(2)}, x_{ki}^{(1)}, x_{k(i+1)}^{(2)}\cdots,x_{kp}^{(2)})_{k=1..n}$$

such that :

$$\hat{U}_i = \frac{1}{n}\sum_{k=1}^{n} Y_k Y_k^{(i)} = \frac{1}{n}\sum_{k=1}^{n} f\left(x_{k1}^{(1)},\ldots,x_{k(i-1)}^{(1)}, x_{ki}^{(1)}, x_{k(i+1)}^{(1)},\ldots,x_{kp}^{(1)}\right) f\left(x_{k1}^{(2)},\ldots,x_{k(i-1)}^{(2)}, x_{ki}^{(1)}, x_{k(i+1)}^{(2)}\cdots,x_{kp}^{(2)}\right) \tag{2.19}$$

Hence, The first-order sensitivity indices are obtained from $\hat{U}_i, \hat{\mu}^2$ and $\hat{V}$ by computing :

$$\hat{S}_i = \frac{\hat{U}_i - \hat{\mu}^2}{\hat{V}} \tag{2.20}$$

The number of indices growths in an exponential way with the number p of dimension: there are $2^p - 1$ indices. For computational time and interpretation reasons, the practitioner should not estimate indices of order higher than two and interest himself in the *total effect*.

Indeed, using the law of total variance below :

Theorem 3 (Law of total Variance, N. A. Weiss [21]) *If X and Y are two random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and the variance of Y is finite.*

$$\textit{Then} \quad \mathrm{Var}(Y) = \mathrm{Var}\left[\mathbb{E}\left(Y|X\right)\right] + \mathbb{E}\left[\mathrm{Var}\left(Y|X\right)\right] \tag{2.21}$$

When applying theorem 3 to $X_{\sim i}$ and normalizing by output variance $\mathrm{Var}(Y)$, we find the sum of two shares:

$$1 = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_{\sim i}\right)\right]}{\mathrm{Var}(Y)} + \frac{\mathbb{E}\left[\mathrm{Var}\left(Y|X_{\sim i}\right)\right]}{\mathrm{Var}(Y)} \tag{2.22}$$

Since the first term is the first order index of $X_{\sim i}$ (see Eq. (2.9)) measuring the contribution of all terms expect those where X_i appeared, the second term has to include all combined effects of the variables X_i. It is thus the total index S_{T_i} of X_i.

Definition 2.2.3 (Sobol total effect) *The total-order sensitivity index S_{T_i} accounting all the contributions to the output variation due to factor X_i, is defined as :*

$$S_{T_i} = \frac{\mathbb{E}\left[\mathrm{Var}\left(Y|X_{\sim i}\right)\right]}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_{\sim i}\right)\right]}{\mathrm{Var}(Y)} \tag{2.23}$$

such that

$$S_{\sim i} + S_{T_i} = 1 \tag{2.24}$$

S_{T_i} represents the expected variance of Y, when only X_i is varied. If it is small, it means that X_i can be fixed to a nominal value without impacting the output.

The total order indices (2.23) can be estimated analogously using Pick and Freeze rule 2 :

$$S_{T_i} = 1 - \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_{\sim i}\right)\right]}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Cov}(Y, Y^{(\sim i)})}{\mathrm{Var}(Y)} = 1 - \frac{\mathbb{E}(YY^{(\sim i)}) - \mathbb{E}(Y)^2}{\mathrm{Var}(Y)} \tag{2.25}$$

Using the same estimator of Sobol first-order as 2.18 except that this time it is applied to $X_{\sim i}$, we get :

$$\hat{S}_{T_i} = 1 - \frac{\frac{1}{n}\sum_{k=1}^{n} Y_k Y_k^{(\sim i)} - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right)^2}{\frac{1}{n}\sum_{k=1}^{n} Y_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right)^2} \tag{2.26}$$

Using also two samples of realizations of the input variables $\hat{X}^{(1)}_{n,\sim i}$ and $\hat{X}^{(2)}_{n,\sim i}$:

$$\hat{X}^{(1)}_{n,\sim i} = (X^{(1)}_{k,i}, X^{(1)}_{k,\sim i}) = (x^{(1)}_{k1}, \ldots, x^{(1)}_{k(i-1)}, x^{(1)}_{ki}, x^{(1)}_{k(i+1)}, \ldots, x^{(1)}_{kp})_{k=1..n}$$

$$\hat{X}^{(2)}_{n,\sim i} = (X^{(2)}_{k,i}, X^{(1)}_{k,\sim i}) = (x^{(1)}_{k1}, \ldots, x^{(1)}_{k(i-1)}, x^{(2)}_{ki}, x^{(1)}_{k(i+1)} \cdots, x^{(1)}_{kp})_{k=1..n}$$

such that :

$$\sum_{k=1}^{n} Y_k Y_k^{(\sim i)} = \frac{1}{n}\sum_{k=1}^{n} f\left(x_{k1}^{(1)},\ldots,x_{k(i-1)}^{(1)},x_{ki}^{(1)},x_{k(i+1)}^{(1)},\ldots,x_{kp}^{(1)}\right) \\ f\left(x_{k1}^{(1)},\ldots,x_{k(i-1)}^{(1)},x_{ki}^{(2)},x_{k(i+1)}^{(1)}\cdots,x_{kp}^{(1)}\right) \tag{2.27}$$

We finally get :

$$\hat{S}_{T_i} = 1 - \frac{\hat{U}_{\sim i} - \hat{\mu}^2}{\hat{V}} \tag{2.28}$$

Note that these indices correspond to the definition of Sobol first-order and total indices as defined from Hoeffding decomposition. Note also $S_{T_i} - S_i$ measures how much X_i is involved in interactions and that :

$$0 \leq S_i \leq S_{T_i} \leq 1. \tag{2.29}$$

2.2.2 Shapley values for dependent variables

The previous Sobol's indices - *first-order and total effect* - are often used together in the sensitivity analysis. Yet, they are founded on a gross assumption of independence of variables. They may fail to appropriately measure how sensitive the output is to uncertainty in the inputs when there is probabilistic dependence or correlation among the inputs (i.e. inputs are inter-correlated) and will consequently present a faulty model interpretation. The purpose now is to determine the output variance when there are some dependencies between model inputs or when they interact.

A very similar problem has been studied in the economics and game theory literature for a long time. The motivation is to find a fair way to attribute the value created in a team effort to the individual members of that team. Economists have studied the setting to measure the value that any subset of the team would have created. For that, they use an attribution method known as the Shapley value (L. Shapley [22]).

Definition and properties :

The Shapley value (L. Shapley [22]) is introduced in game theory to evaluate the *"fair share"* of the total gains to the players in a cooperative game. Mathematically, in Song *et al.* [23], a k-player game with the set of players $\mathcal{K} = \{1, 2, ..., k\}$ is defined as a real-valued function that maps a subset of $\mathcal{K}$ to its corresponding value (or cost), i.e., $v : 2^{\mathcal{K}} \to \mathbb{R}$ with $v(\emptyset) = 0$. Hence, $v(\mathcal{J})$ describes the cost that arises the members of subset/coalition $\mathcal{J}$ of $\mathcal{K}$ by cooperation in the game.

Definition 2.2.4 (Shapley value, L. Shapley [22]) *The Shapley value of player i with respect to $v(.)$ is defined as :*

$$Sh_i = \sum_{\mathcal{J}\subseteq\mathcal{K}\setminus\{i\}} \frac{(k - |\mathcal{J}| - 1)!|\mathcal{J}|!}{k!}(v(\mathcal{J}\cup\{i\}) - v(\mathcal{J})) \tag{2.30}$$

where $|\mathcal{J}|$ indicates the size of $\mathcal{J}$.

Formally, Shapley value Sh_i is the incremental cost of including player i in set $\mathcal{J}$ averaged over all sets $\mathcal{J} \subseteq \mathcal{K}\setminus\{i\}$. Concretely, it represents the average contribution of the player i when he decides to play the game and join a coalition $\mathcal{J}$ over all coalitions formed without player i.

The Shapley value (See Eq.2.30) has several appealing properties characterizing its uniqueness (Winter, [24]) :

1. Efficiency : $\sum_{i=1}^{k} Sh_i = v(|\mathcal{K}|)$
2. Symmetry : If $v(\mathcal{J}+i) = v(\mathcal{J}+j)$ for all $\mathcal{J} \subseteq \mathcal{K}\backslash\{i,j\}$ then $v(i) = v(j)$
3. Dummy : If $v(\mathcal{J}+i) = v(\mathcal{J})$ for all $\mathcal{J} \subseteq \mathcal{K}$ then $Sh_i = 0$
4. Additivity : If v and $\tilde{v}$ have Shapley values Sh and $\tilde{Sh}$ respectively then the game with value function $v(\mathcal{J}) + \tilde{v}(\mathcal{J})$ has Shapley value of $Sh_i + \tilde{Sh}_i$ for all $i \in \mathcal{K}$

Therefore, A. B. Owen [25] proposed an alternative sensitivity measure for dependent variables, based on the concept of the Shapley value in game theory and show that these values are interesting as they allocate the mutual contribution (due to correlation and interaction) of a group of inputs to each individual input within the group.

In the framework of global sensitivity analysis, the set of players $\mathcal{K}$ is the index set of inputs $\{X_1, .., X_k\}$, and the value function $v(.)$ for set of inputs (i.e coalition) $\mathcal{J} \subseteq \mathcal{K}$ is defined as its explanatory power of output variance:

$$v(\mathcal{J}) = \text{Var}\left[\mathbb{E}\left(Y|X_{\mathcal{J}}\right)\right] \tag{2.31}$$

$v(\mathcal{J})$ measures the variance of Y caused by the uncertainty of the inputs in $\mathcal{J}$. Obviously, we have :

$$v(\emptyset) = \text{Var}\left[\mathbb{E}\left(Y|X_{\emptyset}\right)\right] = \text{Var}\left[\mathbb{E}\left(Y\right)\right] = 0 \tag{2.32}$$
$$v(\mathcal{K}) = \text{Var}\left[\mathbb{E}\left(Y|X_{\mathcal{K}}\right)\right] = \text{Var}\left[\mathbb{E}\left(Y|\mathcal{F}\right)\right] = \text{Var}(Y) \tag{2.33}$$

E. Song *et al.* [23] proposed another value function $\tilde{v}$ such as :

$$\tilde{v}(\mathcal{J}) = \mathbb{E}\left[\text{Var}\left(Y|X_{-\mathcal{J}}\right)\right] \tag{2.34}$$

They also proved that Shapley values using both values functions v and $\tilde{v}$ are equivalent. B. Iooss and Cl. Prieur [26] use the normalized version of these two functions to estimate Shapley effects.

The incremental cost $v(\mathcal{J} \cup i) - v(\mathcal{J})$ then can be interpreted as the expected decrease in the variance of Y if we are given the input value of X_i out of all the unknown inputs in $\mathcal{J} \cup \{i\}$.

Finally, we define Shapley values for global sensitivity analysis as in 2.30 :

Definition 2.2.5 (Shapley value for sensitivity analysis) *Let $X = \{X_1, \cdots, X_k\}$ be a set of inputs and $\mathcal{K} = \{1, \cdots, k\}$, the Shapley value of an input X_i.*

$$Sh_i = \sum_{\mathcal{J} \subseteq \mathcal{K}\backslash\{i\}} \frac{(k - |\mathcal{J}| - 1)!|\mathcal{J}|!}{k!} (v(\mathcal{J} \cup \{i\}) - v(\mathcal{J})) \tag{2.35}$$

where $v(\mathcal{J}) = \text{Var}\left[\mathbb{E}\left(Y|X_{\mathcal{J}}\right)\right]$ and $|\mathcal{J}|$ indicates the size of $\mathcal{J}$

Comparison between Sobol' indices and Shapley values :

The first-order Sobol index S_i and the total index S_{T_i} multiplied by $\text{Var}(Y)$, can be defined as semi-values using the value function in 2.31 or in 2.34 :

$$\begin{aligned} S_i &= \text{Var}\left[\mathbb{E}\left(Y|X_i\right)\right] = v(\{i\}) - v(\emptyset) \\ &= \text{Var}(Y) - \mathbb{E}\left[\text{Var}\left(Y|X_{-\mathcal{K}\backslash\{i\}}\right)\right] = \tilde{v}(\mathcal{K}) - \tilde{v}(\mathcal{K}\backslash\{i\}) \end{aligned} \tag{2.36}$$

$$\begin{aligned} S_{T_i} &= \mathbb{E}\left[\text{Var}\left(Y|X_i\right)\right] = \tilde{v}(\{i\}) - \tilde{v}(\emptyset) \\ &= \text{Var}(Y) - \text{Var}\left[\mathbb{E}\left(Y|X_{-\mathcal{K}\backslash\{i\}}\right)\right] = v(\mathcal{K}) - v(\mathcal{K}\backslash\{i\}) \end{aligned} \tag{2.37}$$

Hence, the first-order and total Sobol effects are captured by shapley value when evaluating the incremental cost for subsets $\mathcal{K}\backslash\{i\}$ and $\emptyset$.

Under the assumptions of the four axioms in 2.2.2, and thanks to the propriety of efficiency 2.38, the Shapley value has an outstanding interest compared to Sobol first-order and total effect, it is the unique value such that the sum of all contributions for each input is equal to the total variance of the output.

$$\sum_{i=1}^{k} Sh_i = v(|\mathcal{K}|) = \mathrm{Var}(Y) \tag{2.38}$$

A. B. Owen [25] proved the inequality between Sobol's indices and Shapley value "sandwich effect" in the case of independent input variables :

$$S_i \leq Sh_i \leq S_{T_i} \quad \forall i \in \{1, 2, ...k\} \tag{2.39}$$

with equality if and only if the model is perfectly additive.

This inequality 2.39 doesn't hold anymore when inputs are correlated as shown by B. Iooss [26] for some joint distributions. However, Shapley value remains always between Sobol's indices :

$$S_i \leq Sh_i \leq S_{T_i} \text{ or } S_{T_i} \leq Sh_i \leq S_i \quad \forall i \in \{1, 2, ..., k\} \tag{2.40}$$

In their paper, B. Iooss [26] also showed some interesting results about Sobol's and Shapley values :

- If $S_{T_i} = 0$, then the model output can be written as a measurable function of $X_{\sim i}$ only.

  Proof : If $S_{T_i} = 0$, then $\mathbb{E}(\mathrm{Var}(Y|X_{\sim i})) = 0$ so $\mathrm{Var}(Y|X_{\sim i})) = \mathbb{E}([Y - \mathbb{E}(Y|X_{\sim i})]^2|X_{\sim i}) = 0$ almost surely. Consequently, by taking expectation $\mathbb{E}([Y - \mathbb{E}(Y|X_{\sim i})]^2) = 0$ so $Y - \mathbb{E}(Y|X_{\sim i}) = 0$ almost surely. Hence, $Y = \mathbb{E}(Y|X_{\sim i})$ which is a measurable function of $X_{\sim i} = (X_1, ...X_{i-1}, X_{i+1}, ...X_n)$.

- Assume $S_{T_i} = 0$, if $S_i > 0$, then X_i is correlated to $X_{\sim i} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_d)$.

  Proof : if $\mathbf{X} = (X_1, ..., X_d)$ are independent then $S_i = \mathrm{Var}[\mathbb{E}(Y|X_i)] = \mathrm{Var}[\mathbb{E}(f(X_{\sim i})|X_i)]$ By independence, $S_i = \mathrm{Var}[\mathbb{E}(f(X_1, ...X_{i-1}, X_{i+1}, ...X_n))] = \mathrm{Var}[\mathbb{E}(f(X_{\sim i})] = 0$

- If $Sh_i = 0$ then the input has no significant contribution to the variance of the output, neither by its interactions nor by its dependencies with other inputs.

  Proof : This is direct consequence of the equitable principle on which the allocation rule of Shapley value is based on.

Estimating Shapley values :

Computing Shapley effects could be expensive and unfeasible if the number of inputs is large. Indeed, they involve all subsets $\mathcal{J}$ of the inputs to evaluate $v(\mathcal{K})$ for all $\mathcal{J} \subseteq \mathcal{K}$, i.e., computing $2^k - 1$ variance components and $k!$.

Two algorithms have been proposed by J. Castro *et al.* [27] to estimate Shapley values given a cost function $v(.)$: "Exact permutation" and "Random permutation".

The "Exact permutation" algorithm traverses all possible permutations between the inputs in $\Pi(\mathcal{K})$ and estimate Sh_i in 2.30 by :

$$\hat{Sh_i} = \sum_{\pi \in \Pi(\mathcal{K})} \frac{1}{k!}(v(\pi \cup \{i\}) - v(\pi)) \tag{2.41}$$

while "Random permutation" algorithm consists of sampling some permutations of the inputs randomly $\pi_1, \cdots, \pi_m$ in $\Pi(\mathcal{K})$ and estimating Sh_i via Monte-Carlo simulation :

$$\hat{Sh}_i = \frac{1}{m}\sum_{k=1}^{m}(v(P_i(\pi_k) \cup \{i\}) - v(P_i(\pi_k)) \tag{2.42}$$

where $P_i(\pi)$ is the set the players that precede player i in π.

E. Song *et al.* [23] proposed an improvement to this algorithm that makes accurate approximation of Shapley effects in a efficient manner in terms of computation budget (See Algorithm 1 in 2.1).

Algorithm 1.

1. Choose m, N_V, N_O, and N_I; set $\widehat{Sh}_i = 0$ for $i = 1, 2, \ldots, k$ and $counter = 0$
2. For $q = 1, 2, \ldots, N_V$
   (a) Sample $\mathbf{X}_{\mathcal{K}}^{(q)}$ from $\mathbf{G}_{\mathcal{K}}$
   (b) Evaluate $Y^{(q)} = \eta(\mathbf{X}_{\mathcal{K}}^{(q)})$
3. Calculate $\bar{Y} = {N_V}^{-1}\sum_{q=1}^{N_V} Y^{(q)}$ and $\widehat{\mathrm{Var}}[Y] = (N_V - 1)^{-1}\sum_{q=1}^{N_V}(Y^{(q)} - \bar{Y})^2$
4. While $counter < m$
   (a) Generate $\pi \in \Pi(k)$
   (b) Set $prevC = 0$
   (c) For $j = 1, 2, \ldots, k$
      i. If $j = k$
      $\widehat{c}(P_{\pi(j)}(\pi) \cup \{\pi(j)\}) = \widehat{\mathrm{Var}}[Y]$
      ii. Else \\ comment: $0 < j < k$
         A. For $l = 1, 2, \ldots, N_O$
            Sample $\mathbf{X}^{(l)}_{-P_{\pi(j+1)}(\pi)}$ from $\mathbf{G}_{-P_{\pi(j+1)}(\pi)}$
            For $h = 1, 2, \ldots, N_I$
               Sample $\mathbf{X}^{(l,h)}_{P_{\pi(j+1)}(\pi)}$ from $\mathbf{G}_{P_{\pi(j+1)}(\pi)\left|\mathbf{X}^{(l)}_{-P_{\pi(j+1)}(\pi)}\right.}$
               Evaluate $Y^{(l,h)} = \eta\left(\mathbf{X}^{(l,h)}_{P_{\pi(j+1)}(\pi)}, \mathbf{X}^{(l)}_{-P_{\pi(j+1)}(\pi)}\right)$
            Calculate $\bar{Y}^{(l)} = {N_I}^{-1}\sum_{h=1}^{N_I} Y^{(l,h)}$
            Calculate $\widehat{\mathrm{Var}}\left[Y|\mathbf{X}^{(l)}_{-P_{\pi(j+1)}(\pi)}\right] = (N_I - 1)^{-1}\sum_{h=1}^{N_I}(Y^{(l,h)} - \bar{Y}^{(l)})^2$
         B. Calculate $\widehat{c}(P_{\pi(j+1)}(\pi)) = {N_O}^{-1}\sum_{l=1}^{N_O}\widehat{\mathrm{Var}}\left[Y|\mathbf{X}^{(l)}_{-P_{\pi(j+1)}(\pi)}\right]$
      iii. Calculate $\widehat{\Delta}_{\pi(j)}c(\pi) = \widehat{c}(P_{\pi(j+1)}(\pi)) - prevC$
      iv. Update $\widehat{Sh}_{\pi(j)} = \widehat{Sh}_{\pi(j)} + \widehat{\Delta}_{\pi(j)}c(\pi)$
      v. Set $prevC = \widehat{c}(P_{\pi(j+1)}(\pi))$
   (d) $counter = counter + 1$
5. $\widehat{Sh}_i = \widehat{Sh}_i/m$ for $i = 1, 2, \ldots, k$

Figure 2.1 – Algorithm 1 for estimating Shapley values

Instead of calculating the incremental cost for each i, Algorithm 1 below 2.1 presents a sequential procedure consisting of calculating the costs from the smallest subset of π to the largest and subtracts the previous set's cost $prevC$ to obtain the marginal cost, reducing thus the computing time of the original algorithm reduced by half.

However, the algorithm depends on various parameters: N_i (conditional variance estimation sample size), N_o (expectation estimation sample size), N_v (output variance estimation sample size) and m (random permutation number) and requires a total computing resources of $N_v + mN_iN_o(k-1)$ operations.

As $\hat{Sh}_i$ is unbiased estimator of Sh_i and $\mathrm{Var}(\hat{Sh}_i) \leq \frac{\mathrm{Var}(Y)^2}{m}$, a wise strategy of computational budget is to choose $N_i = 3$, $N_o = 1$, and m as large as possible (E. Song *et al.* [23]), until a sufficient precision (B. Iooss [26]). In addition, adding dummy inputs will increase the complexity of the algorithm without affecting the output Y nor the variance of the estimated Shapley effects as it would have the same bound as before.

2.3 Dependence measures for sensitivity analysis

The importance of the independence assumption for inference arises in many research and engineering fields (biomedical). The classical measures of importance, as introduced in [2.1] are mainly sensitive to a linear or monotonic relationship. Nevertheless, the situation is more challenging in reality, especially when Pearson's coefficient indicates a null value for dependent variables (See figure 2.2). Therefore, there's a need for a statistical measure that, on the one hand, can detect nonlinear relationships between variables. On the other hand, it generalizes the properties of the Pearson correlation coefficient and quantifies the independence when it is zero.

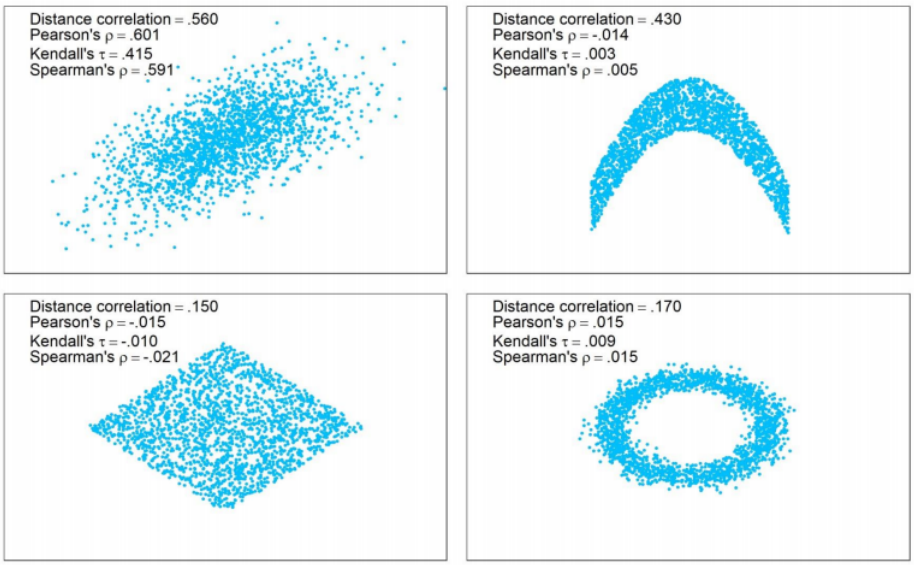


Figure 2.2 – Comparison between distance correlation and other linear/monotonic coefficients (L. Stasielowicz and R. Suck (2019) [1])

Formally, the main problem of measuring dependence is, given $(X, Y) = (x_1, y_1), \ldots, (x_n, y_n)$ with a joint distribution $\mathbb{P}_{X,Y}$, to determine whether $\mathbb{P}_{X,Y} = \mathbb{P}_X \mathbb{P}_Y$ (i.e X and Y are independent) [28].

2.3.1 Distance correlation measure

G. J. Szekely *et al.* ([29] and [30]) introduced a new measure, called the distance correlation coefficient, to address the shortcomings of the Pearson correlation coefficient. They were also providing a new approach to the problem of testing the joint independence of random vectors.

The distance correlation coefficient has now been applied in many contexts, In particular in Astrophysical data [31] where it exhibited higher statistical power (i.e., fewer false positives) than the Pearson coefficient to find nonlinear associations and identified smaller sets of variables that provide equivalent statistical information.

Definition 2.3.1 (Characteristic function) *Let p and q be positive integers. Let $X = (X_1, ..., X_p) \in \mathbb{R}^p$ and $Y = (Y_1, ..., Y_q) \in \mathbb{R}^q$ be random vectors. Let denote ϕ_X and ϕ_Y the characteristic function of X and Y , respectively, and $\phi_{X,Y}$ their joint characteristic function :*

$$\begin{aligned} \phi_X(t) &= \mathbb{E}[\exp(i\langle t, X\rangle)] \quad \forall t \in \mathbb{R}^p \\ \phi_Y(s) &= \mathbb{E}[\exp(i\langle s, Y\rangle)] \quad \forall s \in \mathbb{R}^q \\ \phi_{X,Y}(t,s) &= \mathbb{E}[\exp(i\langle t, X\rangle + i\langle s, Y\rangle)] \quad \forall (t,s) \in \mathbb{R}^p \times \mathbb{R}^q \end{aligned} \tag{2.43}$$

Theorem 4 (Independence characterization) *Let X and Y be two random variables. Let Φ_X and Φ_Y be the characteristic function of X and Y.*

$$\textit{Then} \quad X \textit{ and } Y \textit{ are mutually independent} \quad \textit{iif} \quad \Phi_{X,Y}(s,t) = \Phi_X(s)\Phi_Y(t) \tag{2.44}$$

Definition 2.3.2 (Distance covariance) *The distance covariance between the random vectors X and Y as the non-negative number $\mathcal{V}(X,Y)$ defined by :*

$$\mathcal{V}(X,Y) = \left(\frac{1}{c_p c_q} \int \frac{|\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2}{\|s\|^{p+1}\|t\|^{q+1}} \mathrm{d}s\mathrm{d}t \right)^{1/2} \tag{2.45}$$

where $|.|$ denotes the modulus of complex numbers and

$$c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)} \quad \textit{where} \quad \Gamma \textit{ is the complete gamma function} \tag{2.46}$$

$c_d = C(d, \alpha = 1)$ verifies the lemma below :

Lemma 5 (J. Szekeley *et al.* [29]) *If $0 < \alpha < 2$, then for all x in $\mathbb{R}^d$:*

$$\int_{\mathbb{R}^d} \frac{1 - \cos\langle t, x\rangle}{|t|_d^{d+\alpha}} dt = C(d,\alpha)|x|^\alpha \tag{2.47}$$

The integral in 2.3.2 is well defined by using lemma 5, in fact :

$$\begin{aligned} &\int_{\mathbb{R}^{p+q}} \frac{|\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2}{\|s\|^{p+1}\|t\|^{q+1}} \mathrm{d}s\mathrm{d}t \\ &\leq \int_{\mathbb{R}^{p+q}} \frac{\left(1 - |\phi_X(s)|^2\right)\left(1 - |\phi_Y(t)|^2\right)}{\|s\|^{p+1}\|t\|^{q+1}} \mathrm{d}s\mathrm{d}t \\ &= \int_{\mathbb{R}^p} \frac{1 - |\phi_X(s)|^2}{\|s\|^{p+1}} \mathrm{d}s \int_{\mathbb{R}^q} \frac{1 - |\phi_Y(t)|^2}{\|t\|^{q+1}} \mathrm{d}t \\ &= \int_{\mathbb{R}^p} \frac{1 - \phi_X(s)\overline{\phi_{X'}}(s)}{\|s\|^{p+1}} \mathrm{d}s \int_{\mathbb{R}^q} \frac{1 - \phi_Y(t)\overline{\phi_{Y'}}(t)}{\|t\|^{q+1}} \mathrm{d}t \\ &= \int_{\mathbb{R}^p} \frac{1 - \phi_{X-X'}(s)}{\|s\|^{p+1}} \mathrm{d}s \int_{\mathbb{R}^q} \frac{1 - \phi_{Y-Y'}(t)}{\|t\|^{q+1}} \mathrm{d}t \\ &= \int_{\mathbb{R}^p} \frac{1 - \mathbb{E}\left[\exp\left(i\langle t, X - X'\rangle\right)\right]}{\|s\|^{p+1}} \mathrm{d}s \int_{\mathbb{R}^q} \frac{1 - \mathbb{E}\left[\exp\left(i\langle t, Y - Y'\rangle\right)\right]}{\|t\|^{p+1}} \mathrm{d}s \\ &= \mathbb{E}\left[\int_{\mathbb{R}^p} \frac{1 - \cos\langle t, X - X'\rangle}{|t|_p^{1+p}} dt\right] \cdot \mathbb{E}[\left[\int_{\mathbb{R}^q} \frac{1 - \cos\langle s, Y - Y'\rangle}{|s|_q^{1+q}} ds\right] \\ &= c_p c_q \, \mathbb{E}\left|X - X'\right|_p \mathbb{E}\left|Y - Y'\right|_q < \infty \end{aligned} \tag{2.48}$$

The distance correlation coefficient between X and Y is defined as :

$$\mathcal{R}(X,Y) = \frac{\mathcal{V}(X,Y)}{\sqrt{\mathcal{V}(X,X)\mathcal{V}(Y,Y)}} \tag{2.49}$$

if $\mathcal{V}(X,X), \mathcal{V}(Y,Y) > 0$; otherwise, $\mathcal{R}(X,Y)$ is defined to be 0. It has the following proprieties :

- $0 \leq \mathcal{R}(X,Y) \leq 1$
- $\mathcal{R}(X,Y) = 0$ if and only if X and Y are mutually independent (direct result of theorem 4)

The second propriety give a powerful advantage of distance correlation $\mathcal{R}(X,Y)$ over the Pearson coefficient and other classical measures of correlation

The distance covariance in 2.45 can be expanded and computed in terms of expectations of pairwise Euclidean distances :

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{X,X',Y,Y'} \|X - X'\|_2 \|Y - Y'\|_2 \\ &+ \mathbb{E}_{X,X'} \|X - X'\|_2 \mathbb{E}_{Y,Y'} \|Y - Y'\|_2 \\ &- 2\mathbb{E}_{X,Y} \left[\mathbb{E}_{X'} \|X - X'\|_2 \mathbb{E}_{Y'} \|Y - Y'\|_2\right]\end{aligned} \tag{2.50}$$

where (X', Y') is an i.i.d copy of (X, Y), so a natural estimator of $\mathcal{V}^2(X,Y)$ consists of estimating each expectation $\mathbb{E}(.)$ by $\frac{1}{n}\sum_{i=1}^{n}$ and is given by :

$$\begin{aligned}\mathcal{V}_n^2(X,Y) &= \frac{1}{n^2}\sum_{i,j=1}^{n} \|X_i - X_j\|_2 \|Y_i - Y_j\|_2 \\ &+ \frac{1}{n^2}\sum_{i,j=1}^{n} \|X_i - X_j\|_2 \frac{1}{n^2}\sum_{i,j=1}^{n} \|Y_i - Y_j\|_2 \\ &- \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n} \|X_i - X_j\|_2 \frac{1}{n}\sum_{j=1}^{n} \|Y_i - Y_j\|_2\right]\end{aligned} \tag{2.51}$$

The empirical distance variance is defined in a simplified form from this estimator has been simplified by A. Feuerverger [32] and G. J. Szekely [29] :

$$\mathcal{V}_n(X,Y) = \frac{1}{n}\sqrt{\sum_{k,l=1}^{n} A_{kl}B_{kl}} \tag{2.52}$$

where, for $1 \leq k, l \leq n$:

$$a_{kl} = |X_k - X_l\|_p; \quad \overline{a}_{k.} = \frac{1}{n}\sum_{j=1}^{n} a_{kj} \tag{2.53}$$

$$\overline{a}_{.l} = \frac{1}{n}\sum_{i=1}^{n} a_{il}; \quad \overline{a} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij} \tag{2.54}$$

$$\text{and} \quad \overline{A}_{kl} = a_{kl} - \overline{a}_{k.} - \overline{a}_{.l} + \overline{a} \tag{2.55}$$

and similarly

$$b_{kl} = |Y_k - Y_l\|_q; \quad \overline{b}_{k.} = \frac{1}{n}\sum_{j=1}^{n} B_{kj} \tag{2.56}$$

$$\overline{b}_{.l} = \frac{1}{n}\sum_{i=1}^{n} B_{il}; \quad \overline{b} = \frac{1}{n^2}\sum_{i,j=1}^{n} B_{ij} \tag{2.57}$$

$$\text{and} \quad \overline{B}_{kl} = a_{kl} - \overline{B}_{k.} - \overline{b}_{.l} + \overline{b} \tag{2.58}$$

Although $\mathcal{V}_n(X,Y)$ is a consistent estimator of $\mathcal{V}(X,Y)$, it is easy to see that it is biased. A correction is proposed by G. J.Szekely and M. L. Rizzo [33] for an unbiased version of $\mathcal{V}_n(X,Y)$

The empirical distance correlation $\mathcal{R}_n(X,Y)$ is then :

$$\mathcal{R}_n(X,Y) = \frac{\mathcal{V}_n(X,Y)}{\sqrt{\mathcal{V}_n(X,X)\mathcal{V}_n(Y,Y)}} \tag{2.59}$$

if $\mathcal{V}_n(X,X)$ and $\mathcal{V}_n(Y,Y) > 0$; otherwise, $\mathcal{R}_n(X,Y)$ is defined to be 0.

Remark 1 *$\mathcal{V}_n(X,X) = 0$ if and only if every sample observation is identical i.e $X_1 = ... = X_n$*

Coming back to sensitivity analysis, we define a new index based on distance correlation :

$$S_{X_k}^{dCor} = \mathcal{R}(X_k, Y) \tag{2.60}$$

$S_{X_k}^{dCor}$ will measure the dependence between an input variable X_k and the output Y. It is expected to detect nonlinear relationships and quantify effectively the impact of X_k on Y.

2.3.2 The Hilbert-Schmidt Independence Criterion

In this subsection, we describe how to measure the independence via The Hilbert-Schmidt Independence Criterion, the following theorems and the RKHS theory represent the root idea behind this criterion :

Theorem 6 (Independence by covariance) *Let X and Y be two random variables. Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be two continuous bounded functions.*

$$\textit{Then} \quad X \textit{ and } Y \textit{ are mutually independent} \quad \textit{iif} \quad \forall f,g \ \ \mathrm{Cov}(f(X), g(Y)) = 0 \tag{2.61}$$

Proof : $\Longrightarrow$) Let X and Y be two independent random variables and let f, g be two bounded continuous functions.
By König-Huygens theorem $\mathrm{Cov}(f(X), g(Y)) = \mathbb{E}(f(X)g(Y)) - \mathbb{E}(f(X))\mathbb{E}(g(Y)) = 0$ since :

$$\begin{aligned}
\mathbb{E}(f(X)g(Y)) &= \int_{\mathbb{R}^2} f(x)g(y)f_{X,Y}(x,y)dxdy && (2.62)\\
&= \int_{\mathbb{R}^2} f(x)g(y)f_X(x)f_Y(y)dxdy \quad \text{by independence of } X \text{ and } Y && (2.63)\\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f(x)g(y)f_X(x)f_Y(y)dx \right] dx && (2.64)\\
&= \int_{\mathbb{R}} f(y)f_X(x) \left[\int_{\mathbb{R}} g(y)f_Y(y)dy \right] dx && (2.65)\\
&= \left(\int_{\mathbb{R}} f(x)f_X(x)dx \right) \left(\int_{\mathbb{R}} g(y)f_Y(y)dy \right) && (2.66)\\
&= \mathbb{E}(f(X))\mathbb{E}(g(Y)) && (2.67)
\end{aligned}$$

$\Longleftarrow$) Take $f(x) = \exp(i\langle t, x\rangle)$ and $g(y) = \exp(i\langle s, y\rangle)$ for all $t \in \mathbb{R}^p$ and $s \in \mathbb{R}^q$.
f and g are continuous bounded function such that $\mathrm{Cov}(f(X), g(Y)) = 0$ i.e $\mathbb{E}(f(X)g(Y)) = \phi_{X,Y} = \mathbb{E}(f(X))\mathbb{E}(g(Y)) = \phi_X\phi_Y$, this implies the independence of X and Y by theorem 2.43.

Theorem 7 (Cross-Covariance Operator) *Let X and Y be two random variables. We denote by $\mathcal{C}_b(\mathbb{R})$ the space of continuous and bounded functions in $\mathbb{R}$.*

$$\exists!\ C_{X,Y} \in \mathcal{L}(\mathcal{C}_b(\mathbb{R}), \forall (f,g) \in \mathcal{C}_b(\mathbb{R}) \times \mathcal{C}_b(\mathbb{R}) : \quad \langle f(X), C_{X,Y}g(Y)\rangle = \mathrm{Cov}(f(X), g(Y)) \tag{2.68}$$

where $\langle ., . \rangle$ is inner product $\langle X, Y\rangle = X^\top Y$. C_{XY} is called the cross-covariance operator

Proof : The application $B : \mathcal{C}_b(\mathbb{R}) \times \mathcal{C}_b(\mathbb{R}) \to \mathbb{R} : B(f,g) = \mathrm{Cov}(f(X), g(Y))$ is a bilinear bounded form on a Hilbert space, the corollary of Riesz's representation theorem for bilinear forms guarantee the existence and uniqueness of C_{XY}

RKHS Theory : Reproducing Kernel Hilbert Spaces

We begin by introducing Hilbert spaces and kernels which form the building block of reproducing kernel Hilbert spaces as introduced by A. Berlinet and C. Thomas-Agnan [34].

Definition 2.3.3 (Hilbert space) *A Hilbert Space is an inner product space that is complete and separable with respect to the norm defined by the inner product (i.e Cauchy sequence limits)*

Definition 2.3.4 (Kernel) *Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that :*

$$\forall x, x' \in \mathcal{X} \quad k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

The feature map ϕ of every point $x \in \mathcal{X}$ is a function such that : $\phi(x) = k(\cdot, x)$. In particular, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle k(x, \cdot), k(\cdot, y) \rangle_{\mathcal{H}} = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$

It has the following proprieties :

1. k is symmetric : $\forall (x, y) \in \mathcal{X} \times \mathcal{X} : k(x, y) = k(y, x)$.
2. k is positive semi-definite : $\forall x_1, x_2, ..., x_n \in \mathcal{X}$, the "**Gram Matrix**" $\mathbf{K}$ defined by $K_{ij} = k(x_i, x_j)$ is positive semi-definite i.e $\forall a \in \mathbb{R}^n \; a^\top \mathbf{K} a \geq 0$

The following list of functions are common examples of positive semi-definite kernels on $\mathcal{X} = \mathbb{R}^n$

- Linear kernel : $k(x, x') = \langle x, x' \rangle$
- Gaussian kernel with length-scale $\sigma > 0$: $k(x, x') = \exp\left(-\frac{\|x - x'\|_{\mathrm{L}^n}^2}{2\sigma^2}\right)$
- Matérn kernel with parameters σ and ν : $k_\nu(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|x - x'\|}{\rho}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{\|x - x'\|}{\rho}\right)$ where Γ is the complete Gamma function, K_ν is the modified Bessel function of the second kind.
- Polynomial kernel of degree $d \in \mathbb{N}$: $k(x, x') = (\langle x, x' \rangle + 1)^d$

Theorem 8 (Sum of kernels are kernels) *Given $\alpha > 0$ and k, k_1 and k_2 all kernels on $\mathcal{X}$, then αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.*

Theorem 9 (Product of kernels are kernels) *Given k_1 on $\mathcal{X}_1$ and k_2 on $\mathcal{X}_2$, then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$ then $k := k_1 \times k_2$ is a kernel on $\mathcal{X}$.*

Definition 2.3.5 (Space of real-valued functions on $\mathcal{X}$) *Let $\mathcal{X}$ be a set. Then the space*

$$\mathcal{F}(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \mid f \text{ `is a function}\}$$

together with the standard scalar multiplication and summation defined for all $\lambda \in \mathbb{R}$, and for all $f, g \in \mathcal{F}(\mathcal{X})$ by :

$$(\lambda f)(x) := \lambda f(x) \qquad \forall x \in \mathcal{X}$$

$$(f + g)(x) := f(x) + g(x) \qquad \forall x \in \mathcal{X}$$

forms a linear space over $\mathbb{R}$. We call $\mathcal{F}(\mathcal{X})$ the space of real-valued functions on $\mathcal{X}$.

The Reproducing kernel Hilbert spaces on $\mathcal{X}$ are well-behaved sub-spaces of $\mathcal{F}(\mathcal{X})$. This is made precise in the following definition

Definition 2.3.6 (Reproducing kernel Hilbert spaces) *Let $\mathcal{X}$ be a compact set. Let $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X})$ be a Hilbert space. Then $\mathcal{H}$ is called a RKHS if there exists a kernel k on $\mathcal{X}$ satisfying :*

- $\forall x \in \mathcal{X} : k(x, \cdot) \in \mathcal{H}$
- $\forall f \in \mathcal{H}, \forall x \in \mathcal{X} : \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$

The second property is called "the reproducing property". We call k a reproducing kernel of $\mathcal{H}$

Theorem 10 (Uniqueness of the kernel) *Let $\mathcal{X}$ be a set and let $\mathcal{H}$ be an RKHS on $\mathcal{X}$. Assume both k and $\tilde{k}$ are reproducing kernels of $\mathcal{H}$. Then $k = \tilde{k}$.*

Theorem 11 (Moore-Aronszajn) *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive definite kernel. There is a **unique** RKHS $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ with reproducing kernel k .*

Remark 2 *The feature map ϕ is not unique, only kernel k is unique.*

To summarize up the RKHS theory, if $\mathcal{H}$ is a RKHS and $\mathcal{X}$ is non-empty set of points, then for each $x \in \mathcal{X}$ there exists, by the Riesz's representation theorem a function (i.e feature map ϕ) $\phi(x) = k(x, \cdot)$ in $\mathcal{H}$ (called representer) with the reproducing property $\mathcal{F}_x(f) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$ where $\mathcal{F}_x(f)$ design the evaluation application of $f \in \mathcal{H}$ on x.

The Cross-Covariance Operator and HSIC :

In the framework of RKHS (A. Gretton *et al.* [35]); Let X be a $\mathcal{X}-$valued random vector with a distribution $\mathbb{P}_X$ and consider a RKHS space $\mathcal{F}$ of functions$f : \mathcal{X} \to \mathbb{R}$ with kernel k_X and dot product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. Similarly, we can also define a second random vector $Y \in \mathcal{Y}$ with distribution $\mathbb{P}_Y$ and a RKHS space $\mathcal{G}$ of functions $g : \mathcal{Y} \to \mathbb{R}$ with kernel k_Y and dot product $\langle \cdot, \cdot \rangle_{\mathcal{G}}$.

When $\mathcal{F}$ and $\mathcal{G}$ are RKHS with universal kernels k_X and k_Y (i.e $\mathcal{F}$ and $\mathcal{G}$ are dense in the space of bounded continuous functions 2.3.5) on the compact domains $\mathcal{X}$ and $\mathcal{Y}$, all functions f and g are continuous and bounded. We are now in a position to define the cross-covariance operator from theorem 7 for every $f \in \mathcal{F}$ and $g \in \mathcal{G}$:

$$\langle f(X), C_{X,Y} g(Y)\rangle = \mathrm{Cov}(f(X), g(Y)) = \mathbb{E}_{XY}\left([f(X) - \mathbb{E}_X(f(X))]\,[g(Y) - \mathbb{E}_Y(g(Y))]\right) \quad (2.69)$$

Definition 2.3.7 (The cross-covariance operator) *The cross-covariance operator C_{XY} associated to the joint distribution $\mathbb{P}_{XY}$ of (X, Y) is the unique linear operator $C_{XY} : \mathcal{G} \to \mathcal{F}$:*

$$C_{XY} := \mathbb{E}_{XY}[(\phi(x) - \mu_X) \otimes (\psi(Y) - \mu_Y)]$$

where $\otimes$ denotes the tensor product, ϕ (resp. ψ) is a feature map of k_X (resp. k_Y), μ_X and μ_Y are such that :

$$\langle \mu_X, f\rangle_{\mathcal{F}} := \mathbb{E}_X\left[\langle \phi(X), f\rangle_{\mathcal{F}}\right] = \mathbb{E}_X[f(X)]$$
$$\langle \mu_Y, f\rangle_{\mathcal{F}} := \mathbb{E}_Y\left[\langle \psi(Y), f\rangle_{\mathcal{F}}\right] = \mathbb{E}_Y[g(Y)]$$

The Cross-Covariance C_{XY} operator was introduced by CR. Baker [36] for general Hilbert spaces and by K. Fukumizu *et al.* [37] without investigating it in measuring dependencies

In the same framework of RKHS, and thanks to theorem 6 which characterizes the independence with cross-covariance :

Lemma 12 (C_{XY} and Independence) *Let (X, Y) be two random variables with the joint distribution $\mathbb{P}_{XY}$ and let C_{XY} cross-covariance operator associated to (X, Y)*

$$\textit{The largest singular value of } C_{XY} \textit{ is null} \quad \textit{iif} \quad X \textit{ and } Y \textit{ are independent} \quad (2.70)$$

The cross-covariance operator therefore induces an independence criterion, The Hilbert-Schmidt independence criterion, proposed by A. Gretton *et al.* [28] and build upon on cross covariance operators in RKHS.

Definition 2.3.8 (The Hilbert-Schmidt independence criterion) *Given separable RKHSs $\mathcal{F}$, $\mathcal{G}$ and a joint measure $\mathbb{P}_{XY}$, the Hilbert-Schmidt Independence Criterion (HSIC) is defined as the Hilbert-Schmidt (HS) norm of the associated cross-covariance operator C_{XY} :*

$$HSIC(X,Y)_{\mathcal{F},\mathcal{G}} := \|C_{XY}\|^2_{\text{HS}}$$

where the Hilbert-Schmidt HS norm of a linear operator C is defined as :

$$\|C\|^2_{\text{HS}} := \sum_{ij} \langle C v_i \, u_i \rangle^2_{\mathcal{F}} \tag{2.71}$$

where v_i and u_j are orthonormal basis of $\mathcal{G}$ and $\mathcal{F}$, respectively. This is simply the generalization of the Frobenius norm on matrices.

The HSIC criterion can also be expanded according to the following form (A. Gretton *et al.* [28]) :

$$\begin{aligned} HSIC(X,Y)_{\mathcal{F},\mathcal{G}} &= \|C_{XY}\|^2_{HS} \\ &= \mathbb{E}_{X,X',Y,Y'} k_{\mathcal{X}}(X,X') \, k_{\mathcal{Y}}(Y,Y') \\ &+ \mathbb{E}_{X,X'} k_{\mathcal{X}}(X,X') \, \mathbb{E}_{Y,Y'} k_{\mathcal{Y}}(Y,Y') \\ &- 2\mathbb{E}_{X,Y} \left[\mathbb{E}_{X'} k_{\mathcal{X}}(X,X') \, \mathbb{E}_{Y'} k_{\mathcal{Y}}(Y,Y') \right] \end{aligned} \tag{2.72}$$

where (X', Y') is an independent copy of (X, Y) drawn from $\mathbb{P}_{XY}$.

As a special case, HSIC includes the distance covariance, one can notice the similarity between the generalized distance covariance (see Eq.2.50) and the HSIC criterion (see Eq.2.72). Indeed, D. Sejdinovic *et al.* [38] studied the linked between both measures and showed that :

$$\mathcal{V}(X,Y) = 4HSIC(X,Y)_{\mathcal{F},\mathcal{G}} \tag{2.73}$$

As mentioned before, an important property of $HSIC(X,Y)_{\mathcal{F},\mathcal{G}}$ is the theorem below :

Theorem 13 (C_{XY} and Independence, A. Gretton *et al.* [28]) *Denote by $\mathcal{F}$ and $\mathcal{G}$ two RKHSs with universal kernels k and l on the compact domains $\mathcal{X}$ and $\mathcal{Y}$ respectively. We assume without loss of generality that $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$ for all $f \in \mathcal{F}$ and $g \in \mathcal{F}$.*

Then $\|C_{XY}\|_{\text{HS}} = 0$ *if and only if* X *and* Y *are independent.*

Moreover, B. K. Sriperumbudur *et al.* [39] show that the following kernels are universal :

- Gaussian kernel with length-scale $\sigma > 0$: $k(x,x') = \exp\left(-\frac{\|x-x'\|^2_{\mathbb{L}^n}}{2\sigma^2}\right)$
- Exponential/Laplace kernel with length-scale $\sigma > 0$: $k(x,x') = \exp\left(-\frac{\|x-x'\|}{\sigma}\right)$
- Matérn kernel with parameters σ and ν : $k_\nu(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|x-x'\|}{\rho}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{\|x-x'\|}{\rho}\right)$

The kernel-based distance correlation (i.e HSIC coefficient) between X and Y is defined as :

$$R(X,Y)_{\mathcal{F},\mathcal{G}} = \frac{HSIC(X,Y)_{\mathcal{F},\mathcal{G}}}{\sqrt{HSIC(X,X)_{\mathcal{F},\mathcal{G}} HSIC(Y,Y)_{\mathcal{G},\mathcal{G}}}} \tag{2.74}$$

Where the kernels inducing $\mathcal{F}$ and $\mathcal{G}$ have to be chosen within the class of universal kernels (Back to sensitivity analysis, we can finally propose a sensitivity index based on HSIC measure :

$$S^{HSIC_{\mathcal{F},\mathcal{G}}}_{X_k} = R(X_k,Y)_{\mathcal{F},\mathcal{G}} \tag{2.75}$$

Estimating HSIC criterion :

The main idea of estimating HSIC criterion is similar to empirical distance covariance in 2.50. Assume that $(X_i, Y_i)_{\{} i = 1, \cdots, n\}$ is a sample of the random vector (X, Y) =, then the natural empirical HSIC from Eq. 2.72 is :

$$\begin{aligned} HSIC_n(X,Y)_{\mathcal{F},\mathcal{G}} = & \frac{1}{n^2}\sum_{i,j=1} k_{\mathcal{X}}\left(X_i, X_j\right) k_{\mathcal{Y}}\left(Y_i, Y_j\right) \\ & + \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\mathcal{X}}\left(X_i, X_j\right) \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\mathcal{Y}}\left(Y_i, Y_j\right) \\ & - \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n} k_{\mathcal{X}}\left(X_i, X_j\right) \frac{1}{n}\sum_{j=1}^{n} k_{\mathcal{Y}}\left(Y_i, Y_j\right)\right] \end{aligned} \tag{2.76}$$

Let denote $\mathbf{K}_X$, $\mathbf{K}_Y$ the Gram matrices of k_X, k_Y with entries $\mathbf{K}_X(i,j) = k_X(X_i, X_j)$ and $\mathbf{K}_Y(i,j) = k_Y(Y_i, Y_j)$. and let H a $n \times n$ matrix with $H_{ij} = \delta_{ij} - \frac{1}{n}$. A. Gretton *et al.* [28] propose the following consistent estimator for $HSIC(X,Y)_{\mathcal{F},\mathcal{G}}$:

$$HSIC_n(X,Y)_{\mathcal{F},\mathcal{G}} = \frac{1}{(n-1)^2}\operatorname{Tr}(\mathbf{K}_X H \mathbf{K}_Y H) \tag{2.77}$$

They show that this estimator is biased by $O(1/n)$ but the convergence is uniform. An unbiased estimator is also introduced by L. Song *et al.* [40].

With this estimator, one can test whether the dependence is statistically significant or not for two random variables. An advantage of $HSIC(X,Y)_{\mathcal{F},\mathcal{G}}$ is that it can be computed in $O(n^2)$ time, whereas distance covariance $\mathcal{V}(X,Y)$ and other kernel methods cost at least $O(n^3)$ before approximations are made.

Despite being in $O(n^2)$, using HSIC as an independence criterion still remains difficult to compute when n is too large. A low rank decomposition of the Gram matrices via an incomplete Cholesky decomposition could be useful to get an accurate approximation to HSIC and save computing cost.

Lemma 14 (Efficient approximation by Cholesky decomposition, A. Gretton *et al.* [28]) *Let $K \approx AA^T$ and $L \approx BB$, where $A \in \mathbb{R}^{n \times d_f}$ and $B \in \mathbb{R}^{n \times d_g}$. Then we may approximate $\operatorname{Tr}(HKHL)$ in $O(n(d_f^2 + d_g^2))$ time.*

We also propose another method to save computing time while estimating $HSIC_(X,X)_{\mathcal{F},\mathcal{F}}$ and $HSIC_(Y,Y)_{\mathcal{G},\mathcal{G}}$ by diagonalizing HK and HL by spectral theorem : $HK = PD_KP^T$ and $HL = QD_LQ^T$ where $P, Q \in \mathcal{O}(\mathbb{R}^n$ and $D_K, D_L \in \mathcal{D}(\mathbb{R}^n)$

Lemma 15 (Efficient approximation by eigen values) $\operatorname{Tr}[(HK)^2] = \sum_{\lambda_K \in \mathbb{S}_p(HK)} \lambda_K^2$ *and* $\operatorname{Tr}[(HL)^2] = \sum_{\lambda_L \in \mathbb{S}_p(HL)} \lambda_L^2$. *Then we may approximate* $\operatorname{Tr}(HKHK)$, $\operatorname{Tr}(HLHL)$ *in* $O(\frac{n^2}{2})$ *time*

2.4 Sensitivity measures : limits and discussion

The most important issues of sensitivity analysis methods are computational more than conceptual. Indeed, we can present these issues for both classes of sensitivity measures :

- variance-based measures: the numerical computation of indices requires an evaluation function or numerical model $f(\cdot)$ to condition on an input X_i and estimate the conditional expectation. However, on the one hand, one doesn't know the physical or mathematical

model behind it but has only observational data. In this case, it's recommended to have a good Machine Learning model to estimate these indices correctly. On the other hand, many computer models require a lot of computational time to perform one run. It is impossible (or at least not practical) to perform the number of runs needed to estimate Sobol's indices or Shapley values with the required precision.

- Dependence measures: Although they can be estimated directly from observational data without any model, these kernel-based methods are still heavy to compute when n is too large (n larger than 1000 for most common cases) as we deal with matrix product of many matrices $n \times n$. Moreover, the impact of choice kernels associated with HSIC should be studied.

In the next chapter, and to ease the computational burden, we suggest Gaussian Processes metamodel in which the computer model $f(\cdot)$ is costly to run or unknown. It is a usual engineering practice for estimating variance-based sensitivity indices.

Chapter 3

Gaussian Processes modeling

Contents

The kriging method (G. Matheron [41], N.A.C. Cressie [42]) has been developed for spatial interpolation problems; it takes into account spatial statistical structure of the estimated variable. Sacks *et al.* [43] have extended the kriging principles to computer experiments by considering the correlation between two responses of a computer code depending on the distance between input variables.

The kriging model (also called Gaussian Process model) has been set up with its basis in probability theory (C. E. Rasmussen and C. K. I. Williams [44]), it presents several advantages, especially the interpolation and interpretability properties. Moreover, numerous authors (for example, C. Currin *et al.* [45], T. Santner *et al.* [46] and E. Vazquez *et al.* [47]) show that this model can provide a statistical framework to compute an efficient predictor of code response with an associated uncertainty.

3.1 Gaussian Process and covariance functions

In this section, we define several notions of random process and covariance functions that will be use in GP models in section 3.2). In all manuscripts, we consider a domain of interest $\mathcal{D} \subseteq \mathbb{R}^d$,

Definition 3.1.1 (Stochastic process) *A real-valued random process (or random function) on $\mathcal{D}$ is an application Y, that associates a random variable $Y(x)$ to each $x \in \mathcal{D}$. All the random variables $Y(x)$, for $x \in \mathcal{D}$, are defined respectively to a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

Alternatively a stochastic process is a function on $\mathbb{R}^d$ that is unknown, or that depends of underlying random phenomena. It is characterized by :

- Mean function $\mathcal{M} : x \to \mathcal{M}(x) = \mathbb{E}(Z(x))$
- Covariance function $\mathbf{C} : (x_1, x_2) \to \mathbf{C}(x_1, x_2) = \mathrm{cov}(Z(x_1), Z(x_2))$

Definition 3.1.2 (Trajectory of a random process) *For each fixed $\omega \in \Omega$, the real-valued function $\mathcal{D} : x \to Y(\omega, x)$ is called a trajectory (or a realization or a sample function) of the random process Y.*

Definition 3.1.3 (Gaussian variables and vectors) *A random variable X is a Gaussian variable with mean μ and variance $\sigma^2 > 0$ (i.e $X \sim \mathcal{N}(\mu, \sigma^2)$) when its probability density function is:*

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3.1}$$

A n-dimensional random vector $Y = (Y_1, \cdots, Y_n)$ is a Gaussian vector with mean vector $\mu = \mathbb{E}(Y)$ and invertible covariance matrix $\mathbf{K} = \mathrm{cov}(Y)$, (i.e $Y \sim \mathcal{N}(m, \mathbf{K})$) when either :

- *Any linear combination of its components is a Gaussian random variable.*
- *Its characteristic function has the form :*

$$\Phi_Y(x) = \exp\left(i\langle x, \mu\rangle - \frac{1}{2}x^\top \mathbf{K} x\right) \quad \forall x \in \mathbb{R}^n \tag{3.2}$$

Definition 3.1.4 (Gaussian Process, C. E. Rasmussen and C. K. I. Williams [44]) *A stochastic process Z on $\mathbb{R}^d$ is a Gaussian Process GP when for all $(x_1, ..., x_n)$, the random vector $(Z(x_1), ..., Z(x_n))$ is Gaussian.*

Gaussian Processes presents some advantages : they are simple to define and simulate from their mean and covariance functions. and the Gaussian distribution is reasonable for modeling a large variety of random variables.

To indicate that a random function $f(x)$ follows a Gaussian process, we write :

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, x_0))$$

where x and x_0 are arbitrary input variables. the mean and covariance functions associated are:

$$\mu(x) = \mathbb{E}[f(x)] \quad k(x, x_0) = \mathbb{E}[(f(x) - \mu(x))(f(x_0) - \mu(x_0))^\top]$$

The covariance function $k(x, x_0)$ is a symmetric positive semi-definite (i.e. kernel), usually stationary ($k(x, x_0) = C(|x - x_0|)$. We can also write $k(x, x_0) = \sigma^2 \mathbf{R}(x, x_0)$, where $\mathbf{R}(x, x_0)$ is the auto-correlation function and σ^2 is the process variance.

We recall commonly used kernels in $\mathbb{R}$:

- Matérn kernel : $k_\nu(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|x-x'\|}{\theta}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{\|x-x'\|}{\theta}\right)$ where $\sigma > 0$ is the amplitude, $\theta > 0$ is the length-scale, Γ is the complete Gamma function and K_ν is the modified Bessel function of the second kind.
  - $\sigma^2 > 0$ is the variance amplitude, the larger σ^2 is, the larger the scale of the trajectories.
  - $\theta > 0$ is the characteristic length-scale, it controls how fast the functions sampled from your GP oscillate.
  - ν is the smoothness hyper-parameter that controls the degree of regularity (differentiability) of the resultant GP.

  Some particular cases of Matérn kernel are when $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$ and $\nu \to \infty$.
- Exponential kernel ($\nu = \frac{1}{2}$) : $k_{Exp}(x, x') = \sigma^2 \exp\left(-\frac{|x-x'|}{\theta}\right)$ corresponding to the known Ornstein-Uhlenbeck process

- Matérn 3/2 kernel $(\nu = \frac{3}{2})$: $k_{\mathrm{M}_{3/2}}(x, x') = \sigma^2 \left(1 + \sqrt{3}\frac{|x-x'|}{\theta}\right) \exp\left(-\sqrt{3}\frac{|x-x'|}{\theta}\right)$
- Matérn 5/2 kernel $(\nu = \frac{5}{2})$: $k_{\mathrm{M}_{5/2}}(x, x') = \sigma^2 \left(1 + \sqrt{5}\frac{|x-x'|}{\theta} + \sqrt{5}\frac{(x-x')^2}{3\theta^2}\right) \cdot \exp\left(-\sqrt{5}\frac{|x-x'|}{\theta}\right)$
- Gaussian kernel $(\nu \to \infty)$: $k_{Gauss}(x, x') = \sigma^2 \exp\left(-\frac{\|x-x'\|_2^2}{2\theta^2}\right)$.

The choice of the covariance function is important as it enables to synthesize the information from the Gaussian Process, see figures 3.1, 3.2 and 3.3.

Figure 3.1 – Trajectories of Gaussian processes for different covariance functions from left to right

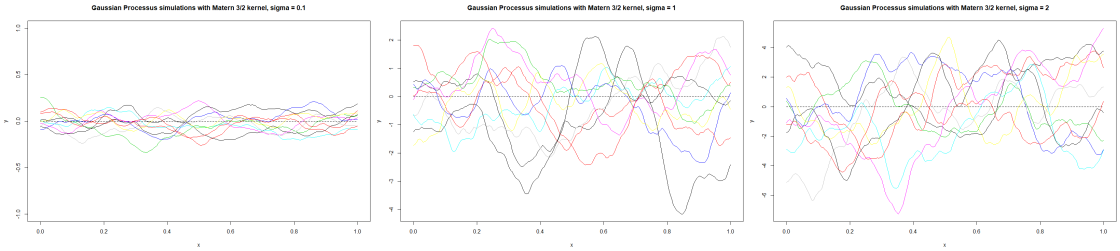


Figure 3.2 – Influence of the variance amplitude σ^2. trajectories of Gaussian processes : Matérn 3/2 with $\sigma^2 = 0.1, 1, 2$ from left to right

In the case of a Gaussian process on $\mathbb{R}^d$, the amplitude σ^2 and smoothness ν are still defined as one value, but the length-scale $\theta = (\theta_1, \cdots, \theta_d) \in \mathbb{R}_+^d$ is now defined as a vector. when θ_i is particularly small, then the variable X_i is particularly important, this allows us to get a rank/hierarchy of the input variables $X_1, \cdots, X_d$ according to their correlation lengths '$(\theta_1, \cdots, \theta_d)$

As mentioned in the chapter 2, it is possible to combine the sum and the product of kernels (See theorems 8 and 9), We can obtain a more complex covariance model based on classical kernels in $\mathbb{R}$:

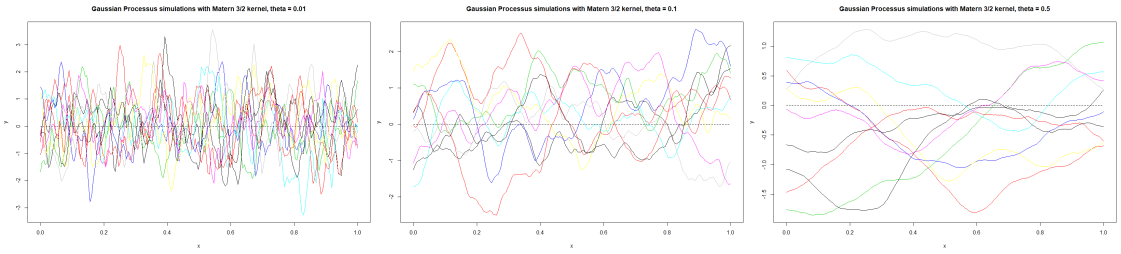


Figure 3.3 – Influence of the length-correlation θ. trajectories of Gaussian processes : Matérn 3/2 with $\theta = 0.01, 0.1, 0.5$ from left to right

- The radial model (isotropic model) defined by :

$$\mathbf{K}_{radial}(x, x') = \sigma^2 k\left(\sqrt{\sum_{i=1}^{d} \frac{|x_i - x_i'|}{\theta_i^2}}\right) \tag{3.3}$$

- The tensorized product model defined by :

$$\mathbf{K}_{TensorProduct}(x, x') = \sigma^2 \bigotimes_{i=1}^{d} k_i(x_i, x_i', \theta_i) \tag{3.4}$$

- The tensorized additive model defined by :

$$\mathbf{K}_{TensorSum}(x, x') = \bigoplus_{i=1}^{d} \sigma_i^2 k_i(x_i, x_i', \theta_i) \tag{3.5}$$

Other classical covariance functions can be build, such as the Power-exponential by tensorizing the exponential kernel k_{Exp} parameterized also by $0 < p \leq 2$:

$$\mathbf{K}_{PowExp}(x, x') = \sigma^2 \prod_{i=1}^{d} \exp\left(-\left(\frac{|x_i - x_i'|}{\theta_i}\right)^p\right) \tag{3.6}$$

or the quasi-periodic GP (H. Tolba *et al.* [48]) by multiplying a periodic kernel by a non periodic kernel.

3.2 Gaussian Process regressor (kriging)

Let us consider n realizations of a physical model or computer code. Each realization $Y(x)$ of the output corresponds to a d-dimensional input vector $x = (x_1, \cdots, x_d) \in \mathcal{D}$. The n points corresponding to the model/code runs are called an experimental design and are denoted as $\mathbf{X} = \left(x^{(1)}, \cdots, x^{(n)}\right)$ where $x^{(i)} = (x_1^{(i)}, \cdots, x_d^{(i)}) \in \mathcal{D}$. The outputs will be denoted as $Y = \left(y^{(1)}, \cdots, y^{(n)}\right)$ with $y^{(i)} = Y(x^{(i)})$. Gaussian Process GP modeling treats the deterministic response $y(x)$ as a realization of a random function $Y(x)$, including a regression part and a centered stochastic process.

Definition 3.2.1 (Gaussian Process modeling) *Gaussian process modelling (i.e Kriging) assumes that the map $\mathcal{D} : x \to Y = f(x)$ is a realization of a Gaussian process:*

$$Y(x, \omega) = f_{trend}(x) + Z(x, \omega)$$

where :

- $x \in \mathcal{D} \subseteq \mathbb{R}^d$ *is the input vector*
- $f_{trend}(x) = \sum_{i=1}^{p} \beta_i f_i(x) = \beta^\top \boldsymbol{F}(x)$ *is the trend part,* $f_i, i = 1, \cdots, p$ *are predefined (e.g. polynomial) functions and* $\beta = \{\beta_1, \cdots, \beta_p\}$ *are the regression coefficients.*
- $Z(x, \omega)$ *is a stationary, zero-mean, with variance* σ^2 *Gaussian Process :*

$$\mathbb{E}[Z(x)] = 0 \quad \textit{and} \quad \mathrm{Var}[Z(x)] = \sigma^2 \quad \forall x \in \mathcal{D}$$

Hence, Gaussian Process regression is a Bayesian non-parametric regression which assumes a GP prior over the regression functions (C. E. Rasmussen and C. K. I. Williams [44]), which can be converted into a posterior over functions once some data has been observed. It consists in approximating $f(x) \sim GP(\mu(x), k(x, x_0))$ using a training set of n observations $\mathcal{D}_{train} = \{(x^{(i)}, y^{(i)}),\ i \in \{1, \cdots, n\}\}$ in order to predict $y^* = y(x^*)$ at a new point $x^* \notin \mathcal{D}_{train}$.

There are three sub-cases of Kriging model, depending on the assumption made on the existing knowledge of the model Y :

- The Simple Kriging : the mean function is assumed to be known i.e $p = 1, f_1 = 1$ and known constant β_1. Equivalently, when working in the simple Kriging framework, we will consider a centered Gaussian process Y.
- The Ordinary Kriging : the mean function is assumed to be constant but unknown i.e $p = 1, f_1 = 1$ and unknown constant β_1
- The Universal Kriging : the mean function at $x \in \mathcal{D}$ is assumed to be of the form $\sum_{i=1}^{p} \beta_i f_i(x)$, where f_i is in set of arbitrary functions $\{f_i(x) = x^{i-1}, j = 1, ..., p\}$ and unknown scalar coefficients β_i.

The parameters $\beta = (\beta_1, ..., \beta_p)$ are subject to an estimation problem, they can be estimated by Generalized Least Squares.

3.3 Joint and conditional distribution : Kriging prediction

Under the hypothesis of a GP model (3.2.1), for each point $x^{(i)} \in \mathcal{D}$, $Y^{(i)} := Y(x^{(i)})$ can be written :

$$Y^{(i)} = \sum_{j=1}^{p} \beta_j f_j\left(x^{(i)}\right) + \sigma \bar{Z} = {f^{(i)}}^\top \beta + \sigma \bar{Z} \tag{3.7}$$

where $\bar{Z} \sim \mathcal{N}(0, 1)$ and $f^{(i)} = \left(f_j\left(x^{(i)}\right)\right)_{j=1,..,p}$.

$Y^{(i)}$ is then a Gaussian variable :

$$Y^{(i)} \sim \mathcal{N}({f^{(i)}}^\top \beta, \sigma^2) \quad \text{with } \mathrm{Cov}[Y^{(i)}, Y^{(j)}] = \sigma^2 \mathbf{R}_\theta(x^{(i)}, x^{(j)}) = \mathbf{K}_{ij} \tag{3.8}$$

where $\mathbf{R}_\theta$ the auto-correlation function.

As a result, the joint distribution of the learning sample $\boldsymbol{Y}$ is Gaussian :

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{F}\beta, \sigma^2 \mathbf{R}_\theta) = \mathcal{N}(\mathbf{F}\beta, \mathbf{K}) \tag{3.9}$$

where $\mathbf{F} \in \mathbb{R}^{n \times p}$ is the regression matrix such that $F_{ij} = f_j(x^{(i)}), i = 1, .., n, j = 1, ..., p$ and $\mathbf{K} = \sigma^2 \mathbf{R}_\theta \in \mathbb{R}^{n \times n}$ is the covariance matrix.

Proof : Let $\boldsymbol{Y} = \left(Y^{(1)}, ..., Y^{(n)}\right)$ be a random vector with mean $\mathbb{E}(\boldsymbol{Y}) = \left({f^{(1)}}^\top \beta, ..., {f^{(n)}}^\top \beta\right) = \mathbf{F}\beta$ and covariance matrix $\mathbf{K} = (\mathbf{K}_{i,j})_{1 \le i,j \le n}$. $\mathbf{K}$ is symmetric definite positive, by Cholesky decomposition there exists a matrix $A \in \mathbb{R}^n$ such that : $\mathbf{K} = AA^\top$. We can write then $Y = \mathbf{F}\beta + AZ$

where $Z \sim \mathcal{N}(0_{n,1}, I_n)$.

$$\begin{aligned}
\phi_{\boldsymbol{Y}}(x) &= \exp\left[i\langle x, \mathbf{F}\beta\rangle\right] \mathbb{E}\left[\exp(i\langle x, AZ\rangle)\right] \\
&= \exp\left[i\langle x, \mathbf{F}\beta\rangle\right] \mathbb{E}\left[\exp(i\langle A^\top x, Z\rangle)\right] \\
&= \exp\left[i\langle x, \mathbf{F}\beta\rangle\right] \Phi_Z\left(A^\top x\right) \\
&= \exp\left[i\langle t, \mathbf{F}\beta\rangle\right] \exp\left[-\frac{1}{2} x^\top A A^\top x\right] \\
&= \exp\left[i\langle x, \mathbf{F}\beta\rangle - \frac{1}{2} x^\top \mathbf{K} x\right]
\end{aligned} \tag{3.10}$$

for $x \in \mathbb{R}^{(n,1)}$, which proves the Gaussianity of $\boldsymbol{Y}$.

Using this result and in the same setting, we want to predict $\boldsymbol{y^*}$ the value of f at a new fixed point $\boldsymbol{x^*} = \{x_1^*, \cdots, x_d^*\}$. The joint probability distribution of $(\boldsymbol{Y}, \boldsymbol{y}^*)$ of the observed data Y and $\boldsymbol{y^*} = f(\boldsymbol{x^*})$ is given by:

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{y}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{F}^\top \beta \\ \boldsymbol{f^*}^\top \beta \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \boldsymbol{k^*} \\ \boldsymbol{k^*}^\top & \sigma^2 \end{bmatrix}\right) \tag{3.11}$$

where $\boldsymbol{k^*} = k(\mathbf{X}, \boldsymbol{x^*}; \theta) \in \mathbb{R}^n$ the cross-covariance vector and $\boldsymbol{f^*} = \{f_1(\boldsymbol{x^*}), \cdots, f_p(\boldsymbol{x^*})\}$ the vector of regressors in $\boldsymbol{x^*}$.

The theorem below is useful to deduce the distribution of the posterior (F. Bachoc [49])

Theorem 16 (Gaussian conditioning theorem) *Let (Y_1, Y_2) be a Gaussian vector such as :*

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{1,1} & \mathbf{K}_{1,2} \\ \mathbf{K}_{2,1} & \mathbf{K}_{2,2} \end{pmatrix}\right) \tag{3.12}$$

Then, $Y_2|Y_1 = y_1$ (i.e Y_2 conditionally on $Y_1 = y_1$) follows a Gaussian distribution

$$Y_2|Y_1 = y_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_2 + \mathbf{K}_{2,1}\mathbf{K}_{1,1}^{-1}\left(\boldsymbol{y}_1 - \boldsymbol{\mu}_1\right), \mathbf{K}_{2,2} - \mathbf{K}_{2,1}\mathbf{K}_{1,1}^{-1}\mathbf{K}_{1,2}\right) \tag{3.13}$$

By conditioning this joint distribution on the learning sample Y in 3.11, It can be shown that the conditional distribution of y^* is also Gaussian :

$$\boldsymbol{y^*}|X, \boldsymbol{Y}, \boldsymbol{x^*} \sim \mathcal{N}\left(\mu(\boldsymbol{x^*}), \sigma^2(\boldsymbol{x^*})\right) \tag{3.14}$$

where:

$$\mu(\boldsymbol{x^*}) = \boldsymbol{f^*}^\top \beta + \boldsymbol{k^*}^\top \mathbf{K}(\boldsymbol{Y} - \mathbf{F}\beta) \tag{3.15}$$

$$\sigma^2(\boldsymbol{x^*}) = \sigma^2 - \boldsymbol{k^*}^\top \mathbf{K}^{-1}\ \boldsymbol{k^*} + \left(\boldsymbol{f^*} - \mathbf{F}\ \mathbf{K}^{-1}\boldsymbol{k^*}\right)^\top \left(\mathbf{F}^\top \mathbf{K}^{-1}\mathbf{F}\right)^\top \left(\boldsymbol{f^*} - \mathbf{F}\ \mathbf{K}^{-1}\boldsymbol{k^*}\right) \tag{3.16}$$

The conditional mean, known as kriging mean and denoted now by $\tilde{y}$ (Eq. 3.18), is used as a predictor. It has a regression part $\boldsymbol{f^*}^\top \beta = \sum_{j=1}^p \beta_i f_j(\boldsymbol{x^*})$ and a local correction. It can be applied to any other new point x_{new} :

$$\tilde{y}(x_{new}) := \mathbb{E}(Y(x_{new})|\boldsymbol{Y}) = {f_{new}}^\top \beta + k(x_{new})^\top \mathbf{K}(\boldsymbol{Y} - \mathbf{F}\beta) \tag{3.17}$$

Thus $\tilde{y}$, the mean prediction for $Y(x_{new})$, can be written as a linear combination of kernel functions, each one centered on a training point:

$$\tilde{y}(x_{new}) = f(x_{new})^\top \beta + k(x_{new})^\top \mathbf{K}(\boldsymbol{Y} - \mathbf{F}\beta) \tag{3.18}$$

$$= \sum_{j=1}^p \beta_j f_j(x_{new}) + \sum_{i=1}^n \alpha_i k(x^{(i)}, x_{new}) \tag{3.19}$$

where $\alpha = \mathbf{K}(\boldsymbol{Y} - \mathbf{F}\beta)$.

These coefficients α_i will be referred to as *parameters;* they are updated each time a new observation is made (as opposed to the parameters of the kernel, referred to as *hyperparameters*, which are not updated once training is over (see subsection)

Remark 3 *The kernel part of prediction function* $x_{new} \rightarrow \sum_{i=1}^{n} \alpha_i k(x^{(i)}, x_{new})$ *vanishes when* x_{new} *is far from the observation points* $\{x^{(1)}, ..., x^{(n)}\}$. *Hence the prediction of 3.19 is essentially meant for interpolation.*

The variance formula corresponds (Eq. 3.20) to the Mean Squared Error (MSE) of this predictor and is also known as the kriging variance $\tilde{\sigma}^2$. It gives a local indicator of the prediction accuracy.

$$\tilde{\sigma}^2(x_{new}) := \mathrm{Var}(Y(x_{new})|\boldsymbol{Y}) = \mathrm{Var}(Y(x_{new}) - k(x_{new})^\top \mathbf{K}^{-1}\ k(x_{new}) + \tag{3.20}$$

$$\left(f_{new} - \mathbf{F}\ \mathbf{K}^{-1} k(x_{new})\right)^\top \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \left(f_{new} - \mathbf{F}\ \mathbf{K}^{-1} k(x_{new})\right) \tag{3.21}$$

Due to the Gaussianity of the predictor $Y(x_{new}) \sim \mathcal{N}\left(\tilde{y}(x_{new}), \tilde{\sigma}^2(x_{new})\right)$ one can derive confidence intervals on the prediction with confidence level $(1 - \alpha)$, e.g. 95%, one gets:

$$\tilde{y}(x_{new}) - q_{\alpha/2} \times \tilde{\sigma}^2(x_{new}) \leq Y(x_{new}) \leq \tilde{y}(x_{new}) + q_{\alpha/2} \times \tilde{\sigma}^2(x_{new}) \tag{3.22}$$

In particular, for $(1 - \alpha) = 95\%$:

$$\tilde{y}(x_{new}) - 1.96 \times \tilde{\sigma}^2(x_{new}) \leq Y(x_{new}) \leq \tilde{y}(x_{new}) + 1.96 \times \tilde{\sigma}^2(x_{new}) \tag{3.23}$$

The most outstanding advantage of GP model compared to other models comes from the previous equations. In fact, kriging model provides an mathematical formula for the distribution of the output variable at an arbitrary new point x_{new} (Eq. 3.18, Eq. 3.20 and 3.22). This distribution formula can be used for sensitivity analysis and uncertainty quantification, as well as for quantile evaluation (J. Oakley *et al.* [50]) instead of costly methods based for example on a Monte Carlo algorithm. All these considerations and possible extensions of GP modeling of GP represent significant advantages (C. Currin *et al.*[45], C. E. Rasmussen and C. K. I. Williams [44]).

Remark 4 *When* $x_{new} = x^{(i)}$ *for a particular* i, *it results from 3.18 and 3.20 that :* $\tilde{y}(x^{(i)}) = y_i$ *and* $\tilde{\sigma}^2(x^{(i)}) = 0$. *We say that the GP predictor interpolates the experimental design : the prediction is the value itself and the associated uncertainty is zero.*

3.4 Estimating GP model parameters and hyper-parameters

The GP model 3.7 is characterized by the regression parameter vector β and the covariance parameters (σ, θ) (in addition p for Power-Exponential kernel). In practice, we know neither the parameters vector β of nor the hyperparameters vector (σ, θ).

Constructing a GP model and computing the kriging mean and variance as shown in 3.18 and 3.20 implies estimating these parameters. A choice can be made *a prior* based on prior knowledge of the distribution and data, then a estimation is performed from the experimental design $\mathcal{D} = \{(X^{(1)}, Y^{(1)}), ..., (X^{(n)}, Y^{(n)})\}$. This is commonly done by Maximum Likelihood method (ML) or CV, which makes the GP model be an Bayesian estimator of the maximum *a posterior*.

3.4.1 Maximum Likelihood Estimator

Given a GP model and Assuming that data follows a joint Gaussian distribution $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{F}\beta, \sigma^2\mathbf{R}(\theta))$ the negative log-likelihood of $\boldsymbol{Y}$ can be written as:

$$\begin{aligned} -\log \mathcal{L}\left(\beta, \sigma^2, \theta | \boldsymbol{Y}\right) =& \frac{1}{2\sigma^2}(\boldsymbol{Y} - \mathbf{F}\beta)^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{Y} - \mathbf{F}\beta) + \frac{n}{2}\log(2\pi) \\ &+ \frac{n}{2}\log\left(\sigma^2\right) + \frac{1}{2}\log(\det \mathbf{R}_{\boldsymbol{\theta}}) \end{aligned} \tag{3.24}$$

The solution $(\widehat{\beta}, \widehat{\sigma}^2)$ is obtained by solving :

$$\frac{\partial(-\log \mathcal{L})}{\partial \beta} = \mathbf{F}^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{Y} - \mathbf{F}\boldsymbol{\beta}) = 0 \quad ; \quad \frac{\partial(-\log \mathcal{L})}{\partial \sigma^2} = 0 \tag{3.25}$$

Given the covariance parameters $(\sigma, \theta; p)$, the maximum likelihood estimator of $\widehat{\beta}_{ML}$ is the generalized least squares estimator:

$$\widehat{\beta}_{ML}(\boldsymbol{\theta}) = \left(\mathbf{F}^{\mathrm{T}} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1} \boldsymbol{Y} \tag{3.26}$$

and the maximum likelihood estimator of $\widehat{\sigma}_{ML}$ is:

$$\widehat{\sigma}^2_{ML}(\boldsymbol{\theta}) = \frac{1}{n}(\boldsymbol{Y} - \mathbf{F} \cdot \widehat{\beta}_{ML})^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1} \cdot (\boldsymbol{Y} - \mathbf{F}\widehat{\beta}_{ML}) \tag{3.27}$$

As $\widehat{\sigma}^2_{ML}$ and $\widehat{\beta}_{ML}$ depend on hyper-parameters of covariance θ. We Substitute them into the log-likelihood $-\log \mathcal{L}$ to obtain the optimal choice θ minimizing equivalently the reduced likelihood function $-\log \tilde{\mathcal{L}}$ such that :

$$-\log \tilde{\mathcal{L}}(\boldsymbol{\theta}) = \ln\left(\hat{\sigma}^2_{ML}(\boldsymbol{\theta})\right) + \frac{1}{n}\ln\left(\det \mathbf{R}_{\boldsymbol{\theta}}\right) \tag{3.28}$$

Thus, maximum likelihood estimation of $\widehat{\theta}_{ML}$ consists in numerical optimization of the function defined as follows:

$$\widehat{\theta}_{ML} \in \operatorname{argmin}_{\theta \in \Theta} -\log \tilde{\mathcal{L}}(\boldsymbol{\theta}) \tag{3.29}$$

Minimizing function $-\log \tilde{\mathcal{L}}$ in 3.29 is an optimization problem that is numerically costly and difficult to solve with $O(n^3)$ computational cost. Several difficulties make this optimization problem computationally heavy, mainly, the large number of parameters which imposes the use of a sequential method of resolution, where different parameters are introduced step by step, and the large parameter's domain due to and the lack of prior bounds requires an exploratory algorithm (stochastic gradient, multistart ...) able to explore the domain in an optimal way.

3.4.2 Bayesian full approach estimation

In Maximum Likelihood, the estimator MLE we look for point $(\beta, \sigma^2, \theta)$ that maximizes the likelihood $\mathcal{L}\left(\beta, \sigma^2, \theta | \boldsymbol{Y}\right)$ as in 3.24. The optimal value, $(\beta_{ML}, \sigma^2_{ML}, \theta_{ML})$ is not a random variable but a point estimate (i.e a Dirac distribution centred on $(\beta_{ML}, \sigma^2_{ML}, \theta_{ML})$)

Bayesian estimation integrates the uncertainty about the unknown parameter and treats $(\beta, \sigma^2, \theta)$ as a random variable. In this method, the estimated hyper-parameters are probability density functions, rather than estimating a single point as in Maximum Likelihood MLE.

We recall the Bayes's rule in theorem 17 below:

Theorem 17 (Bayes' Rule for parameters distribution) *Let θ be a set of probability distribution parameters that best explains the observations $\boldsymbol{Y}$, the Bayes' Rule assumes that:*

$$f_{\theta|\boldsymbol{Y}} = \frac{f_{\boldsymbol{Y}|\theta}\ f_{\theta}}{f_{\boldsymbol{Y}}} \quad i.e \quad Posterior = \frac{Likelihood \times Prior}{Evidence}$$

The posterior hyper-parameters $\boldsymbol{\theta}_{posterior}$ for the model are given by :

$$f_{\boldsymbol{\theta}|\boldsymbol{Y},X} = \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta},X}\ f_{\boldsymbol{\theta}}}{f_{\boldsymbol{Y}|X}} \tag{3.30}$$

where $f_{\boldsymbol{Y}|X}$ is the joint distribution of $\boldsymbol{Y}$ given by Eq 3.9.

3.4.3 Cross-Validation Estimator

For this subsection on the Cross Validation estimation which represent an alternative to estimate the covariance hyper-parameters (σ^2, θ). The regression coefficients β will not be studied with this estimation method.

Definition 3.4.1 (Leave-One-Out Mean Square Error criterion) *The Leave-One-Out (LOO) Mean Square Error (MSE) criterion is defined by :*

$$LOO(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,\boldsymbol{\theta}})^2 \tag{3.31}$$

where, for $1 \leq i \leq n$, $\hat{y}_{i,\boldsymbol{\theta}}$ is the prediction given in 3.18 of y_i by a GP model when trained on $\{y_1, ..., y_{i-1}, y_{i+1}, ..., y_n\}$, given the covariance function $\mathbf{K}_{\sigma,\theta} = \sigma^2\mathbf{R}_\theta$

F. Bachoc [49] showed that $LOO(\boldsymbol{\theta})$ can be written with explicit quadratic forms :

$$LOO(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{y}^\top\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\ \mathrm{Diag}\left(\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\right)^{-2}\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\boldsymbol{y} \tag{3.32}$$

where $\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-} := \mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{F}\left(\mathbf{F}^\top\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{R}_{\boldsymbol{\theta}}^{-1}$ and $\mathbf{F}$ is the regression matrix as defined in 3.9.

The Cross-Validation estimator CV aims to estimate θ by minimizing the LOO MSE criterion

$$\hat{\theta}_{MSE} \in \mathrm{argmin}_{\theta\in\Theta}\ \frac{1}{n}\boldsymbol{y}^\top\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\ \mathrm{Diag}\left(\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\right)^{-2}\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\boldsymbol{y}. \tag{3.33}$$

This criterion reflects only the quality of the point wise prediction of 3.18. It doesn't estimate The variance parameter σ^2. We define another Cross-Validation estimator for this purpose :

Definition 3.4.2 (Leave-One-Out variance estimator) *The LOO Cross-Validation estimator of σ^2 is defined as :*

$$\hat{\sigma}^2_{MSE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(y_i - \hat{y}_{i,\hat{\theta}_{MSE}}\right)^2}{\hat{c}^2_{i,\hat{\theta}_{MSE}}} \tag{3.34}$$

where $\hat{\theta}_{MSE}$ is obtain from 3.33

F. Bachoc [49] also showed that $\hat{\sigma}^2_{LOO}$ can be written with an explicit quadratic forms:

$$\hat{\sigma}^2_{MSE} = \frac{1}{n}\boldsymbol{y}^\top\tilde{\mathbf{R}}_{\tilde{\boldsymbol{\theta}}_{MSE}}^{-}\ \mathrm{Diag}\left(\tilde{\mathbf{R}}_{\tilde{\boldsymbol{\theta}}_{MSE}}^{-}\right)^{-1}\tilde{\mathbf{R}}_{\tilde{\boldsymbol{\theta}}_{MSE}}^{-}\boldsymbol{y} \tag{3.35}$$

The expressions 3.45 and 3.35 allow to estimate θ by Cross-Validation by minimizing a criterion that has the same computational complexity of $O(n^3)$ as Maximum Likelihood, but it has the advantage of being more efficient when the covariance function is mis-specified.

In the following part, Let interest us with the confidence intervals of the GP model 3.22, we define a new metric, called the $(1-\alpha)$-probability score, denoted $\mathbb{P}_{1-\alpha}^{score}$

Definition 3.4.3 ($(1-\alpha)$-probability score) *Let $\boldsymbol{y} = \{y^{(1)}, ..., y^{(n)}\}$ and $\tilde{\boldsymbol{y}} = \{\tilde{y}^{(1)}, ..., \tilde{y}^{(n)}\}$ (resp $\tilde{\boldsymbol{\sigma}} = \{\tilde{\sigma}^{(1)}, ..., \tilde{\sigma}^{(n)}\}$) be the mean (resp standard-deviation) predictions of $\boldsymbol{y}$ given by a GP model with covariance matrix $\mathbf{K}_{\sigma,\theta} = \sigma^2 \mathbf{R}_\theta$*

$(1-\alpha)$-probability score score describes the probability of getting $y^{(i)}$ inside in predictions intervals with confidence level $(1-\alpha)$

$$\mathbb{P}_{1-\alpha}^{score}(\theta, \sigma) = \mathbb{P}(y \in PI_{1-\alpha}(\tilde{y})) = \mathbb{P}_{1-\alpha/2}\left(\frac{y-\tilde{y}}{\tilde{\sigma}}\right) - \mathbb{P}_{\alpha/2}\left(\frac{y-\tilde{y}}{\tilde{\sigma}}\right) \tag{3.36}$$

Empirically, $\mathbb{P}_{1-\alpha}^{score}$ is the percentage of points $y^{(i)}$ belonging to prediction intervals $PI_{1-\alpha}(\tilde{y}_\theta))$. Ideally, it should be close to $1-\alpha$ (e.g. if we define $1-\alpha = 95\%$ (i.e $\alpha = 5\%$), we would like to have 95% of points inside these intervals, see 3.4)

This criterion is very useful in our case, it will allow us to measure if a GP model is reliable in terms of predictions, otherwise, optimize it to fit and respect the P_{90}/P_{10} rules.

Definition 3.4.4 (Leave-One-Out $(1-\alpha)$-probability score) *Let $\boldsymbol{y} = \{y_1, ..., y_n\}$ and $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{\sigma}}$) be the mean, standard-deviation predictions of $\boldsymbol{y}$ by a given GP model.*

$$LOO_{\mathbb{P}_{1-\alpha}^{score}}(\theta, \sigma) = \left(\mathbb{P}_{1-\alpha}^{score} - (1-\alpha)\right)^2 \tag{3.37}$$

The previous $LOO_{\mathbb{P}_{1-\alpha}^{score}}$ can be simplified. One can write $\mathbb{P}(\frac{y-\tilde{y}}{\tilde{\sigma}})$ as an expectation :

$$\mathbb{P}_{1-\alpha/2}\left(\frac{y-\tilde{y}_\theta}{\tilde{\sigma}_\theta}\right) = \mathbb{E}\left(\mathbb{1}_{\frac{y-\tilde{y}_\theta}{\tilde{\sigma}_\theta} \leq q_{1-\alpha/2}}\right) \simeq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} \leq q_{1-\alpha/2}} \tag{3.38}$$

$$\mathbb{P}_{\alpha/2}\left(\frac{y-\tilde{y}_\theta}{\tilde{\sigma}_\theta}\right) = \mathbb{E}\left(\mathbb{1}_{\frac{y-\tilde{y}_\theta}{\tilde{\sigma}_\theta} \leq q_{\alpha/2}}\right) \simeq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} \leq q_{\alpha/2}} \tag{3.39}$$

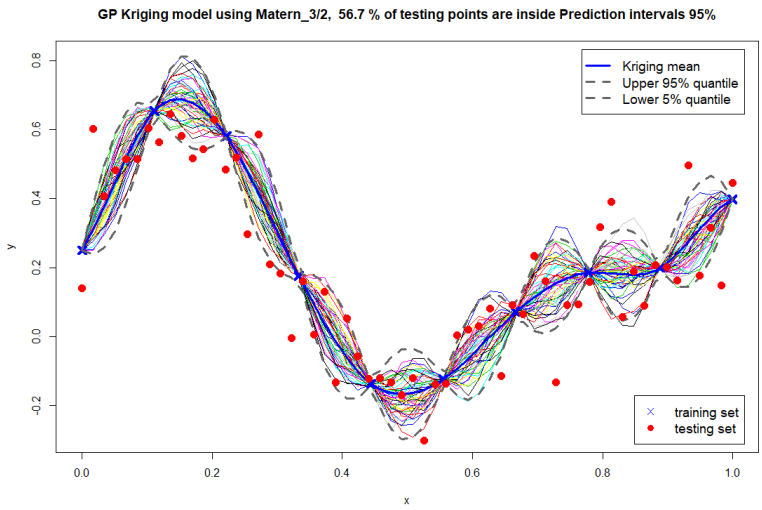
Such that :

$$\mathbb{1}_{\frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} \leq q_{1-\alpha/2}} = \mathbb{1}_{q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} \geq 0} = \frac{\left(q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right)^+}{\left|q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right|} = \frac{1}{2}\left(1 + \frac{q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}}{\sqrt{\left(q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right)^2}}\right) \tag{3.40}$$

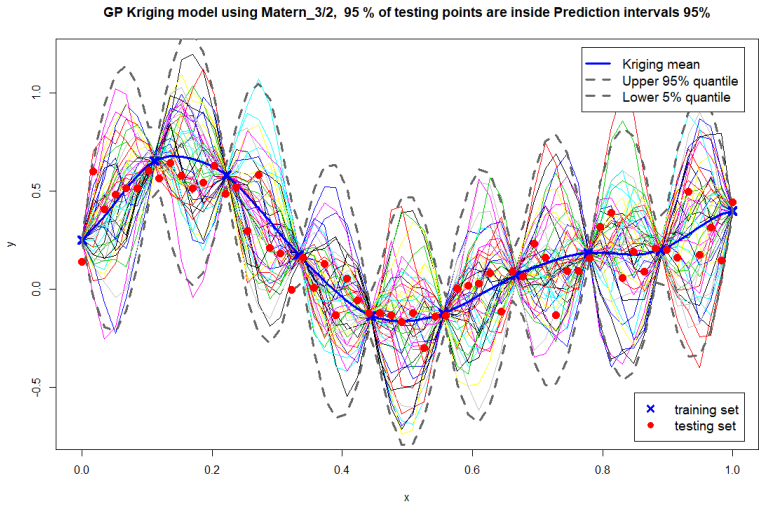
$$\mathbb{1}_{\frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} < q_{\alpha/2}} = \mathbb{1}_{q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} > 0} = \frac{\left(q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right)^+}{\left|q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right|} = \frac{1}{2}\left(1 + \frac{q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}}{\sqrt{\left(q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right)^2}}\right) \tag{3.41}$$

By Plugging the obtained expressions in 3.40 and 3.41 in $LOO_{\mathbb{P}_{1-\alpha}^{score}}$ 3.37 :

$$\begin{aligned} LOO_{\mathbb{P}_{1-\alpha}^{score}} &= \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} \leq q_{1-\alpha/2}} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i} < q_{\alpha/2}} - (1-\alpha)\right)^2 \\ &= \left(\frac{1}{2n}\sum_{i=1}^{n}\left(\frac{q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}}{\sqrt{\left(q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right)^2}} - \frac{q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}}{\sqrt{\left(q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right)^2}}\right) - (1-\alpha)\right)^2 \\ &= \left(\frac{1}{2n}\sum_{i=1}^{n}\left(h\left(q_{1-\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right) - h\left(q_{\alpha/2} - \frac{y_i-\tilde{y}_i}{\tilde{\sigma}_i}\right)\right) - (1-\alpha)\right)^2 \end{aligned} \tag{3.42}$$



(a)



(b)

Figure 3.4 – Case of $\alpha = 5\%$; a) Hyper-parameters are optimized by Maximum Likelihood 3.4.1, only $55,7\%$ of points are inside Predictions Intervals. b) Hyper-parameters are optimized in such way to have exactly $1 - \alpha = 95\%$ of points

where $h(x) = \frac{x}{\sqrt{x^2}}$ when $x > 0$ and $h(0) = 1$

In LOO-CV, the expressions of $\tilde{y}_i$ and $\tilde{\sigma}_i$ are given by (C. E. Rassmussen & C. K. I. Williams [44]) :

$$\tilde{y}_i = y_i - \frac{[\mathbf{K}^{-1}\mathbf{y}]_i}{[\mathbf{K}^{-1}]_{ii}}, \quad \text{and} \quad \sigma_i^2 = \frac{1}{[\mathbf{K}^{-1}]_{ii}} \tag{3.43}$$

Replacing these values in the equation 3.42 leads to :

$$LOO_{\mathbb{P}_{1-\alpha}^{score}} = \left(\frac{1}{2n} \sum_{i=1}^{n} \left(h\left(q_{1-\alpha/2} + \frac{[\mathbf{K}^{-1}\mathbf{y}]_i}{\sqrt{[\mathbf{K}^{-1}]_{ii}}} \right) - h\left(q_{\alpha/2} + \frac{[\mathbf{K}^{-1}\mathbf{y}]_i}{\sqrt{[\mathbf{K}^{-1}]_{ii}}} \right) \right) - (1-\alpha) \right)^2 \tag{3.44}$$

Hence, a Cross-Validation estimator of σ, θ for fitting PI with confidence level $(1 - \alpha)$ is by minimizing the $LOO_{\mathbb{P}_{1-\alpha}^{score}}$ above :

$$(\hat{\theta}_{score}, \hat{\sigma}_{score}) \in \text{argmin}_{\theta \in \Theta, \sigma \in \Sigma} \; LOO_{\mathbb{P}_{1-\alpha}^{score}}(\theta, \sigma) \tag{3.45}$$

Remark 5 *We tried to minimize* $\text{argmin}_{\theta\in\Theta,\sigma\in\Sigma}\ \left(\mathbb{P}_{\alpha}^{score}-(1-\alpha)\right)^2$ *by solving* $\nabla\left(\mathbb{P}_{\alpha}^{score}-(1-\alpha)\right)^2=0$ *but this method seems to be very heavy to solve even with optimisation algorithms.*

Remark 6 *Some packages (e.g. "kergp") provide directly LOO-CV predictive mean* $\tilde{y}_i$ *and variance* $\tilde{\sigma}_i$ *while building Gaussian Process model. In this case, minimizing LOO* $\left(\mathbb{P}_{1-\alpha}^{score}\right)$ *as in 3.42 becomes much easier and faster than 3.44 which requires inverting* $\mathbf{K}$.

To summarize, the Cross-Validation procedure is applied in two separate cases according to two different criterion : for Mean Square Error criterion we give priority first to the point-wise prediction at a new point to estimate θ, and second estimate the global variance σ^2 adapted to the Leave-One-Out prediction errors. For $(1-\alpha)$-probability score, θ,σ are optimized so prediction intervals length will be fitted to respect the confidence level $(1-\alpha)$. Combining both criteria Mean Square Error MSE and $(1-\alpha)$-probability score will be subject of a future research.

Chapter 4

Modelling Methodology with Gaussian Process

Contents

Although being a robust statistical model for supervised learning and uncertainty quantification, the theoretical efficiency of the Gaussian Process is limited when modelling a function in a high-dimension domain. Unfortunately, the Gaussian Process model is computationally expensive and not well adapted to high-dimensional problems, principally due to inversion problems of the covariance matrix in the kriging mean 3.18 and variance 3.20, and to hyper-parameters estimation, while solving the minimization problems by Maximum likelihood and Cross-Validation for methods, which has a computational cost of $O(n^3 + n^2 d)$. Thus, the kriging model becomes unfeasible with many input variables d and requires a large number of training points $n \gg 10d$ in addition to computing sensitivity indices whom the complexity increases exponentially with the number of inputs.

In this kind of situation, variable selection techniques must be applied to reduce the complexity of the model. We deal with this problem by following a new methodology, proposed by B. Iooss and A. Marrel ([51], 2017), allowing to build a GP model with a large number of inputs efficiently by screening and joint modelling. It could also be applied to other types of Machine Learning models. The idea behind the methodology is simple, include the first influent variables and keep the non-selected variables to quantify the uncertainty caused by the dimensionality reduction. Our objective is twofold: Build a highly predictive GP model with a few variables with a proper uncertainty, and analyze the sensitivity of parameters that control production.

4.1 Step 1 - Standardization of numerical input variables

The suitable procedure to construct a GP model requires optimal space filling designs (SFD) with orthogonality properties (K-T. Fang *et al.* [52]) (e.g. Hypercube Latin Sampling (LHS) with Maximin criterion or $\mathcal{L}_2-$discrepancy). Such designs provide a full coverage of the input space an allows investigating the whole variation domain of the uncertain parameters.

However, in Gas and Oil industry, the production data have been generated a long time ago according to a specific or unknown distribution, building a space filling designs is probably not

possible anymore. In this case, all not available data are removed and a standardization is applied to the original data according to the $Z-$score :

$$Z = \frac{X - \mu}{\sigma} \tag{4.1}$$

where X is a given input, $\mu = \mathbb{E}(X)$ is the mean of X and $\sigma^2 = \text{Var}(X)$ is the standard deviation. The standardization has the objective of increasing the robustness of hyper-parameter estimating algorithm, and simplifying the choices of bounds and starting points.

Remark 7 *We do not change the input probability distributions after standardization, they are used later in the sensitivity analysis.*

Remark 8 *The categorical inputs are not taken into account in this stage. This task will be accomplished in the future by building group kernels (O. Roustant et al. [53])*

4.2 Step 2 - Screening initial input variables by decreasing influence

The sensitivity measures, as introduced in chapter 2, are very useful in prioritizing inputs and can be quantitatively interpreted. They represent our main tool to identify the most influent variables from a set $\mathbf{X} = \{X_1, ..., X_d\}$ and sort them by decreasing order of influence. This step is called : screening inputs.

In our case, we don't have any computer code behind well Oil and Gas production, so, as a sort criterion, we choose the kernel-based distance correlation $R(X_k, Y)_{\mathcal{F}_\parallel, \mathcal{G}}$ (HSIC coefficient 2.74) between the input X_k variable and the response Y. Thanks to the non-linear kernels which remove hypotheses such as linearity or monotony, the HSIC measure considers dependencies that are more complex will measure of the influence of inputs $\mathbf{X}$ on the output Y efficiently. Furthermore, the $R(\mathbf{X}, Y)_{\mathcal{F},\mathcal{G}}$ can help to divide input variables $\mathbf{X}$ into two sub-groups : *"the significant ones"* and *"the non-significant ones"*, based on the independence statistical test and depending on the significance level of these tests. The *"significant ones"* will be taken into account sequentially while building GP model. At the end of this step, the "significant inputs" are ordered in a set $X_{sort} = \{X_{\Pi(1)}, ..., X_{\Pi(d)}\}$ where $\Pi(j)$ design the index of j^{th} most influent variable

Remark 9 *Another screening method, commonly used in ensemble learning algorithms, can be applied by exploring variables sequentially : At iteration j^{th}, look for variable that maximizes the predictivity coefficient Q^2 (see section 4.4) :*

$$X_{\Pi(j+1)} = \text{argmax}_{i \notin \Pi\{1,...,j\}} \, Q^2(\mathbf{X}_{\Pi\{1,...,j\}} \cup X_i) \tag{4.2}$$

4.3 Step 3: GP joint modeling with sequential building process

To estimate GP model hyper-parameters faster and efficiently, we precede in a progressive procedure (loop on all input variables in $\mathbf{X}$) that combines sorted inputs from screening step and joint modeling.

At each iteration j, we consider only the j first sorted inputs in explanatory inputs variables $\mathbf{X}_{\text{exp}} = \{X_{\Pi(1)}, ..., X_{\Pi(j)}\}$ while the remaining inputs are considered as a stochastic parameter $\mathbf{X}_\epsilon = \{X_{\Pi(j+1)}, ..., X_{\Pi(d)}\}$, then we perform a joint GP modeling.

The approach, as described in [54] and in [51], consists on building two GP models with $\mathbf{X}_{\text{exp}}$ to fit mean and dispersion components such as :

$$\begin{aligned} Y_m\left(\mathbf{X}_{\text{exp}}\right) &= \mathbb{E}\left(Y|\mathbf{X}_{\text{exp}}\right) \\ Y_d\left(\mathbf{X}_{\text{exp}}\right) &= \text{Var}\left(Y|\mathbf{X}_{\text{exp}}\right) = \mathbb{E}\left[\left(Y - Y_m\left(\mathbf{X}_{\text{exp}}\right)\right)^2 |\mathbf{X}_{\text{exp}}\right] \end{aligned} \tag{4.3}$$

The first GP model is built from a defined covariance model (e.g. Matérn $3/2$) with homoscedastic nugget effect, $GP^j_{m,1}$ for the mean component to fit Y. Then a second GP, denoted $GP^j_{d,1}$, is built for the dispersion component with the same covariance model to fit the squared residuals $(Y - \tilde{Y}_{m,1})^2$ from the predictor of $GP^j_{m,1}$.

$GP^j_{d,1}$ estimates the dispersion error at each point, it can be considered as the value of the heteroscedastic nugget effect. The nugget effect is thus updated in the covariance matrix $\mathbf{K}_\theta$. We repeat the same step by building more GPs $GP^j_{m,i}$ and $GP^j_{d,i}$ on the mean and dispersion component and updating the estimated nugget, but it seems that one $GP^j_{m,2}$ and $GP^j_{d,2}$ is enough to remove the homoscedastic hypothesis.

The final GP GP^j model is built with the updated heteroscedastic nugget effect. Its hyper-parameters are optimized by taking hyper-parameters obtained at the $(j-1)^{th}$ iteration as starting points for the optimization.

This procedure is summarized up in the following algorithm :

Algorithm 1 Sequential procedure of joint modeling

Ensure: $X_{sort} = \{X_{\Pi(1)}, ..., X_{\Pi(d)}\}$
for $j = 1...d$ **do**
(0) Set $\mathbf{X}_{exp} == \{X_{\Pi(1)}, ..., X_{\Pi(j)}\}$
(1) Build a GP model $GP^j_{m,1}$ with $\mathbf{X}_{exp}$ to fit Y and estimates $\tilde{Y}_{m,1} = \mathbb{E}\left(Y|\mathbf{X}_{\text{exp}}\right)$
(2) Build a GP model $GP^j_{d,1}$ with $\mathbf{X}_{exp}$ to fit Y and estimates $\tilde{Y}_{d,1} = \mathbb{E}\left[\left(Y - \tilde{Y}_{m,1}\right)^2 |\mathbf{X}_{\text{exp}}\right]$
(3) Update the covariance matrix by the estimated nugget effect $\epsilon = \tilde{Y}_{d,1}$
$\mathbf{K}_{\theta,\sigma} \leftarrow \mathbf{K}_{\theta,\sigma} + \epsilon\mathbf{I}_n$
(4) Build a GP model $GP^j_{m,2}$ to fit Y with the new covariance matrix $\mathbf{K}_\theta$ and estimates $\tilde{Y}_{m,2} = \mathbb{E}\left(Y|\mathbf{X}_{\text{exp}}\right)$
(5) Repeat (2) and (3) using a GP model $GP^j_{d,2}$
(6) Build a GP model GP^j with $\mathbf{X}_{exp}$ to fit $\tilde{Y} = \mathbb{E}\left(Y|\mathbf{X}_{\text{exp}}\right)$
(7) Estimates the new hyper-parameters $(\sigma, \theta, \beta)_j$ by taking $(\sigma, \theta, \beta)_{j-1}$ as starting point
(8) Computing the model accuracy ${Q_j}^2$

$$Q^2_j = 1 - \frac{\sum_{i=1}^{n_{test}} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_{\text{test}}} \left(y_i - \overline{y}\right)^2} \tag{4.4}$$

end for

4.4 Step 4: Assessment of GP model predictivity

The accuracy coefficient Q^2, corresponding to the classical coefficient of determination R^2, is computed for a test sample or by Cross-Validation, to evaluate the accuracy of the model:

$$Q^2_{\mathbf{X}} = 1 - \frac{\sum_{i=1}^{n_{test}} \left(Y_i - \tilde{Y}_i\right)^2}{\sum_{i=1}^{n_{\text{test}}} \left(Y_i - \overline{Y}\right)^2} \tag{4.5}$$

where $Y = \{Y_1, ..., Y_{n_{test}}\}$ denotes the n_{test} observations of the test set, $\overline{Y} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} Y_i$ is their empirical mean and $\tilde{Y} = \{\tilde{Y}_1, ..., \tilde{Y}_{n_{test}}\}$ represents the GP model predicted values obtain

from Eq. 3.18, built on $\mathbf{X}$ with estimated parameters (σ, θ, β). The closer to one the Q^2, the better the accuracy of the model

Other simple validation criteria can be used: the mean absolute error (MAE), the mean square error MSE (see, for example, J. P. C. Kleijnen and R. G. Sargent [55]). Some statistical and graphical analyses of residuals (e.g. QQ-plot) can also provide detailed descriptions and diagnostics.

Definition 4.4.1 (The Mean Square Error MSE) *Given a predictor f^*, Y the vector of n observed values and $\tilde{Y}$ the vector of predictions of Y by f^*. The Mean Square Error MSE of f^* is defined as the variance of predictions errors (residuals):*

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \tilde{Y}_i\right)^2 \tag{4.6}$$

MSE provides information about the goodness of predictor fitting. The smaller the MSE value, the better the predictions are.

Definition 4.4.2 (The Root Mean Square Error RMSE) *The Root Mean Square Error RMSE is defined as the standard deviation of the prediction errors (residuals):*

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \tilde{Y}_i\right)^2} = \sqrt{MSE} \tag{4.7}$$

Definition 4.4.3 (The Mean Absolute Error MAE) *The Mean Absolute Error MAE is defined similarly to The Mean Square Error MSE expect that it takes the absolute value of the prediction errors (residuals) :*

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \tilde{Y}_i\right| \tag{4.8}$$

Remark 10 *Q^2 is a biased criterion because one can obtain a different score for the same predictions when the observations Y are dispersed from $\overline{Y}$. It is important to consider always Mean Square Errors MSE*

Remark 11 *MSE is a biased criterion because it penalizes large errors (e.g. outliers) than small errors. MAE is useful in this case as it penalizes errors in the same way whether they are small or large.*

At this step, we explore all families of covariance functions described in section 3.1 and different covariance structures until find the appropriate model that builds the optimal kriging model. At the end, we draw a graph presenting the evolution of model's accuracy Q^2 with features or number of variables added into the model.

4.5 Step 5: Optimizing the final GPs for special criterion's

Once step 4.3 and 4.4 are done, we investigate the *"accuracy vs features"* graph to see at which variable the accuracy Q^2 stagnate or fall, so we can build two GP models : GP_{pred} that will be used for the predictive part and uncertainty quantification by considering only the first variables increasing accuracy, and one other GP_{global} for the sensitivity analysis of all input variables (variance-based sensitivity measures vs dependence measures).

- The first model GP_{pred} is built from $p \leq d$ variables which makes him faster in estimating procedure. GP_{pred} is used to predict production at given time or the oil/gas max rate, its hyper-parameters $(\sigma_{ML}, \theta_{ML}, \beta_{ML})$ are estimated by Maximum likelihood for an accurate faithful model in a firstly, then we estimate $(\theta_{CV}^{score}, \sigma_{CV}^{score})$ by Cross-Validation method with the same β_{ML} to fit the $(1-\alpha)$-probability score criterion. In the end, we get a predictive model respecting the P_{10}/P_{90} with a reasonable Q^2 .
- The second GP_{global} is built as explained in step 4.3 using all inputs variables. The main criterion is MSE for point-wise prediction, it's preferable to have the higher Q^2 with both estimators (Maximum Likelihood and Cross-Validation). GP_{global} will be used mainly to estimate variance-based sensitivity measures (Sobol and Shapley).

4.6 Step 6: Sensitivity analysis of the GP model

The last step is dedicated to sensitivity analysis of production data using GP_{global}, we study the influence of parameters and the uncertainty propagation by computing Sobol's first-order, total effect and Shapley value for each variable (Although we still work on high dimensional with GP_{global}). We compare in particular these indices with HSIC measure.

Chapter 5

Application and results

Contents

5.1 Application to Analytical functions

In this section, dedicated to the first experiments, we focus on GP kriging for different analytical functions, considered as test case, where the dimensionality varies from $d = 2$ to $d = 20$. The objective is to compare GP joint model with sequential approach against simple GP model, Maximum Likelihood and Cross-Validation estimators and sensitivity indices.

Let us consider the following analytical functions, defined on the hypercube $\mathbf{X} \in [0,1]^d$:

- O. Roustant *et al.* (2018) [53] Additive function :

$$f_{addfun6d}(\mathbf{X}) = X_1^3 + \cos(\pi X_2) + |X_3|\sin(X_3^2) + 3X_4^3 + 3\cos(\pi X_5) + 3|X_6|\sin(X_6^2) \quad (5.1)$$

- H. Moon (2010) [56] Low-Dimensionality function :

$$f_{MoonLD}(\mathbf{X}) = X_1 + X_2 + 3X_1X_3 \quad (5.2)$$

- H. Moon *et al.* (2012) [57] High-Dimensionality function :

$$f_{MoonHD}(\mathbf{X}) = -19.71X_1X_{18} + 23.72X_1X_{19} - 13.34X_{19}^2 + 28.99X_7X_{12} \quad (5.3)$$

- W. J. Morokoff and R. E. Caflisch (1995) [58] function :

$$f_{Morokoff}(\mathbf{X}) = \left(1 + \frac{1}{d}\right)^d \prod_{i=1}^{d} X_i^{1/d} \quad (5.4)$$

- T. Crestaux *et al.* (2009) [59] Sobol G-function :

$$f_{Sobol}(\mathbf{X}) = \prod_{i=1}^{d=8} \frac{|4X_i - 2 + a_i|}{1 + a_i} \tag{5.5}$$

where a_i are parameters, such that $a_i \geq 0$.

- M. D. Morris *et al.* (2006) [60] function :

$$f_{Morris}(\mathbf{X}) = \alpha \sum_{i=1}^{k} \left(X_i + \beta \sum_{i<j=2}^{k} X_i X_j \right) \tag{5.6}$$

where $\alpha = \sqrt{12} - 6\sqrt{0.1\,(k-1)}$, $\beta = 12\sqrt{0.1\,(k-1)}$, $i, j = \{1, ..., 20\}$ and $k = 10$ is an integer controlling the number of influential inputs.

- C. Linkletter *et al.* (2006) [61] decreasing function :

$$f_{Linkletter}(\mathbf{X}) = 0.2X_1 + \frac{0.2}{2}X_2 + \frac{0.2}{4}X_3 + \frac{0.2}{8}X_4 + \frac{0.2}{16}X_5 + \frac{0.2}{32}X_6 + \frac{0.2}{64}X_7 + \frac{0.2}{128}X_8 \tag{5.7}$$

- T. Ishigami and T. Homma (1991) [62] function :

$$f_{Ishigami}(\mathbf{X}) = \sin(\pi X_1) + a\sin(\pi X_2)^2 + bX_3^4 \sin(\pi X_1) \tag{5.8}$$

where $a = 7$ and $b = 0.1$ (Marrel et al.)

- Testing function :

$$f_{Test}(\mathbf{X}) = \sin(\pi X_1) + \cos\left(\pi \frac{X_2}{4}\right) + \sqrt{X_1 X_2} + 0.X_3 \tag{5.9}$$

$\mathbf{X} = \{X_1, ..., X_d\}$ are assumed to be $i.i.d$ variables, except in some cases in the subsection 5.1.3) where we deal with dependent variables. The "kergp" package, available on R-Cran is used while building GP kriging model following (See 4.4 in 4). Finally, the training *Design of Experiments* (DoE) is a Latin Hypercube Sample LHS with $n_{Fit} = 50$ points whereas the testing DoE is also a LHS built from $n_{Val} = 200$ points.

5.1.1 Maximum Likelihood vs Cross-Validation

Firstly, for each function, two GP models are built using Maximum Likelihood estimator and Cross-Validation estimator. The metrics used to compare both models are RMSE (See 4.4.2) and prediction accuracy Q^2 (See 4.5), the table 5.1 summarizes the results :

	RMSE MLE	RMSE Cross-Validation	Q^2 MLE	Q^2 Cross-Validation
Additive fun ($d = 6$)	0,325	0,191	0,983	0,994
Moon fun HD ($d = 20$)	7,353	3,042	0,149	0,759
Moon fun SD ($d = 3$)	$8{,}104.10^{-4}$	0,101	0,999	0,988
Morokoff.fun ($d = 3$)	0,072	0,074	0,974	0,962
Morokoff.fun ($d = 10$)	0,127	0,134	0,855	0,837
Test fun ($d = 3$)	0,011	0,013	0,999	0,999
Sobol fun ($d = 8$)	0,181	0,214	0,931	0,904
Morris fun ($d = 20$)	37,16	33,16	-0,008	0,221
Linkletter fun ($d = 8$)	$1{,}76.10^{-6}$	$2{,}27.10^{-6}$	0,999	0,999
Ishigami fun ($d = 3$)	$33{,}4.10^{-4}$	$5{,}29.10^{-4}$	0,999	0,999

Table 5.1 – Accuracy Q^2 and the RMSE error for Maximum Likelihood and Cross-Validation.

One can notices that Cross-Validation method is more efficient in the case of Moon HD function 5.3 or Morris function 5.6. More generally, the Cross-Validation method is well adapted to model mis-specifications i.e when the covariance of Y cannot be computed exactly with a covariance function (See F. Bachoc [49]).

5.1.2 GP Building process : Sequential approach vs classical approach

Secondly, for high-dimensional functions which has a good Q^2, we perform a HSIC-based screening on inputs and build a joint GP model according to the sequential approach in 4.3. On the other hand, we build a simple GP model including all inputs :

	RMSE Joint GP	RMSE Simple GP	Q^2 Joint GP	Q^2 Simple GP
Additive fun ($d = 6$)	0,467	0,465	0,969	0,966
Morokoff.fun ($d = 10$)	0,116	0,119	0,806	0,797
Sobol fun ($d = 8$)	0,181	0,429	0,917	0,534
Linkletter fun ($d = 8$)	1,76 10^{-6}	2,27.10^{-6}	0,999	0,999

Table 5.2 – Accuracy Q^2 and RMSE obtained for Maximum Likelihood estimated GP model by joint modeling (See step 4.3) vs simple GP model.

	RMSE Joint GP	RMSE Simple GP	Q^2 Joint GP	Q^2 Simple GP
Additive fun ($d = 6$)	0,406	0,406	0,975	0,975
Moon fun HD ($d = 20$)	3,314	3,212	0,804	0,791
Morokoff.fun ($d = 10$)	0,116	0,119	0,806	0,797

Table 5.3 – Accuracy Q^2 and RMSE obtained for Cross-Validation GP model by joint modeling (See step 4.3) vs simple GP model.

Obviously, the accuracy Q^2 is improved between 0.3% to 38%. the higher is Q^2, the less significant the difference between the two models. Yet, the proposed methodology is still an efficient and robust method to build GP models with a high-dimensional data.

Remark 12 *Joint modeling by Cross-Validation estimator suffers from some computation problems (due to matrix inversion). reason why we don't present the comparison's results for Sobol and Morris function.*

5.1.3 Sensitivity analysis indices

In this subsection, we illustrate the sensitivity indices for some of the previous functions : Additive function 5.1, Linkletter 5.7, G-sobol 5.7 and Testing function 5.9. We do not compute these indices from the function itself, we use the optimal surrogate GP models, as shown in the previous section, such that Q^2 is close to one.

- Case of independent variables :

  $\mathbf{X} = \{X_1, ..., X_d\}$ are assumed to be independent variables sampled in a LHS Design of Experiment. Sensitivity indices are computed and presented in figure 5.1 :

  1. Some indices are estimated negatively, in particular first-order indices, in this case one can conclude that these indices are equal to zero.
  2. The most influent variables are those with low coefficient a_i or expressed by a function with high variations.

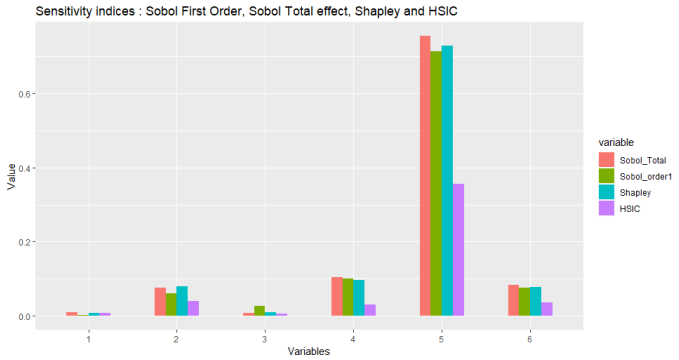
3. Variance-based measures and HSIC rank influent variables similarly regardless to their value for each input X_i (i.e. input's influence ranking does not change from a measure to another).
4. When the function is perfectly additive (e.g. Additive func 5.1 and Linkletter 5.7), the equality of sensitivity indices $S_i = Sh_i = S_{T_i}$ holds.

- Case of dependent variables :

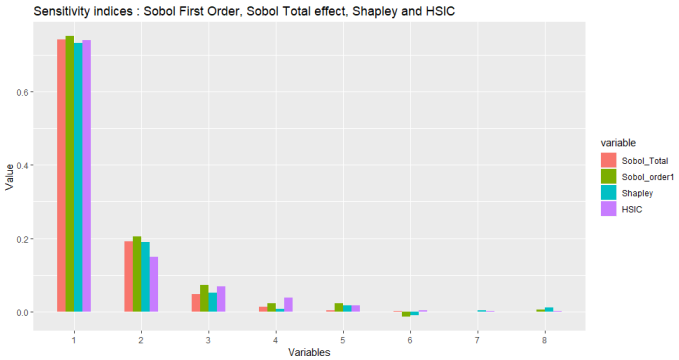
  $\mathbf{X} = \{X_1, ..., X_d\}$ are no more assumed to be independent and we consider only the Testing function.

  1. S_{T_3}=0 in all cases, this result is consistent with claim 1) in 2.2.2 : $f(.)$ is a measurable function of X_1 and X_2 but not X_3.
  2. - When $X_3, X_2 \sim \mathcal{U}[0,1], X_1 = \frac{1+X_3^2}{2} + \epsilon$, $S_3 > 0$ which is also consistent with claim 2 in 2.2.2 : The contribution of X_3 to the variance of Y is contained in the contribution of X_1.
  3. - When $X_3 \sim \mathcal{U}[0,1], X_1 = \frac{\sqrt{X_3}+X_3^2}{2} + \epsilon$ and $X_2 = \frac{X_1^2+X_3}{2} + \epsilon$, HSIC gives a high importance to X_2 due to its strong dependence to X_1 and X_3.
  4. - When $X_1, X_2 \sim \mathcal{U}[0,1]$ and $X_3 = \sin(X_1) + \epsilon$, $HSIC_{\mathcal{F},\mathcal{G}}$ captures the dependence between X_3 and Y through X_1 while Variance-based indices (Sh_3 and S_3) do not capture any effect on Y.

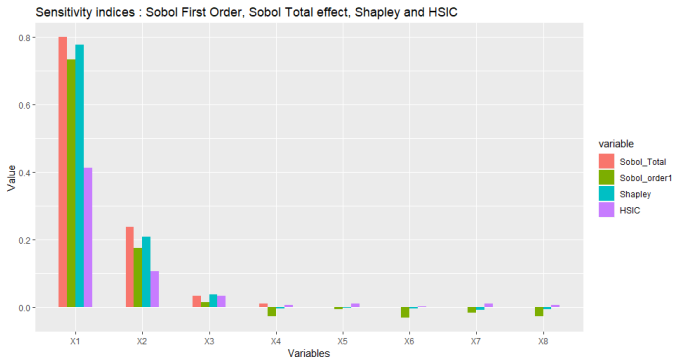
  The previous result is very interesting in sensitivity analyse / causal inference with dependent variables : **When $Sh_i = 0$ but $S_{X_i}^{HSIC_{\mathcal{F},\mathcal{G}}} \neq 0$ then X_i and Y are dependent mutually of a variable controlling both of them (confounder)**.



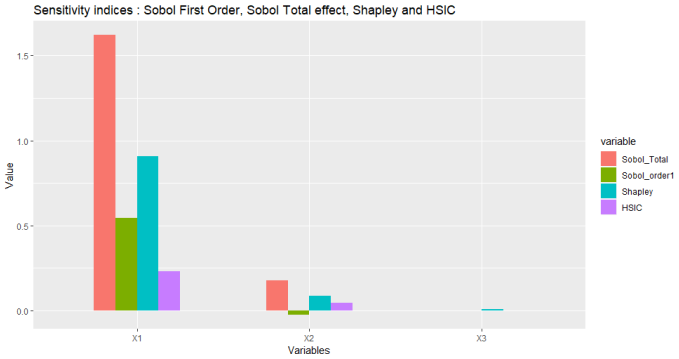
(a) Additive function



(b) Linkletter function

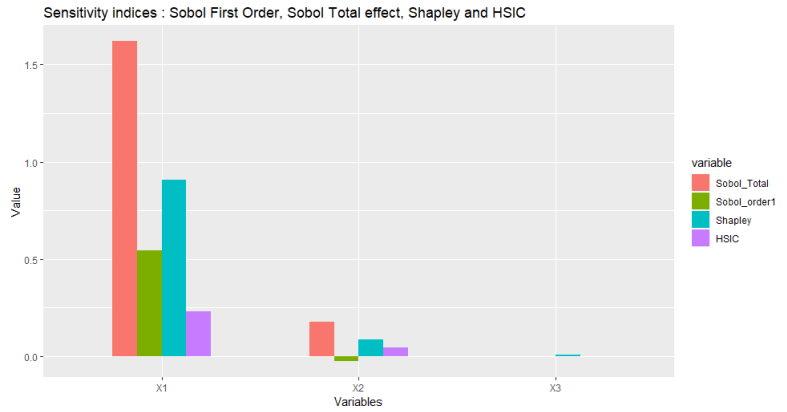


(c) Sobol-G function

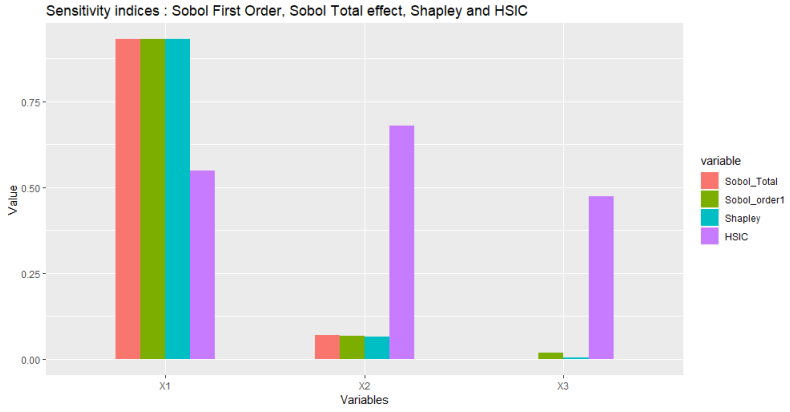


(d) Testing function

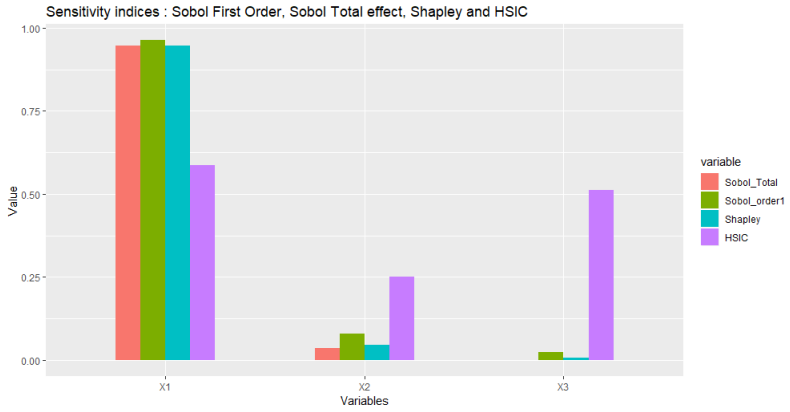
Figure 5.1 – Sensitivity analysis with independent variables



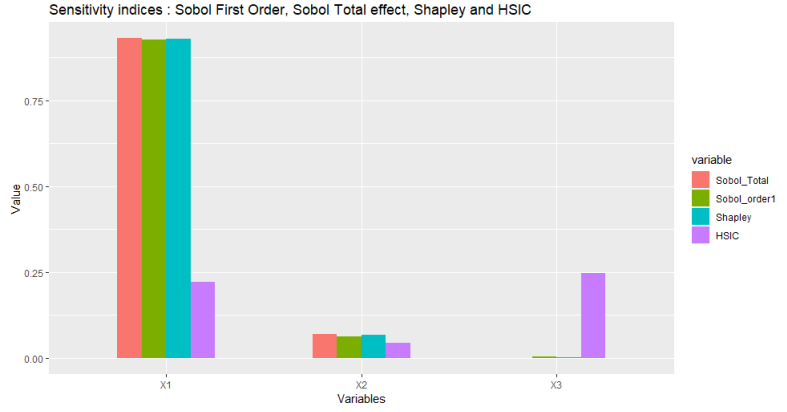
(a)



(b)



(c)



(d)

Figure 5.2 – Sensitivity analysis with dependant variable for Testing function

5.2 Application to production data : UTICA Shale

5.2.1 Presentation

Reservoir engineers and experts seek to predict gas and oil well production of new/undiscovered wells, not only from historical data of previous well but also from other well characteristics and operational parameters. This will allow them to anticipate the economic performance of the well, and also to better manage the supply chain of which it is a part.

5.2.2 Data description and exploratory analysis

Our data-set, *field data*, is derived from unconventional wells. It contains approximately 2700 wells with more than 120 variables, including localization, production time-series, exploitation conditions and the associated geological data. However, Only a few variables are interesting to us in this study: Production, localization, well characteristics and "Fracturing design" (see figure 5.3), the scatter plot of data-set is 5.12.

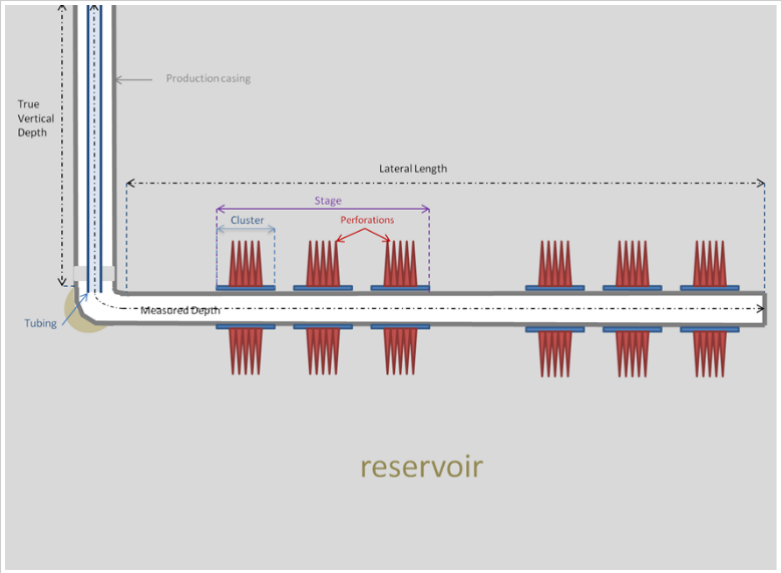


Figure 5.3 – Illustration of a well in reservoir : True Vertical Depth, Lateral Length and bottom hole point

- **Longitude_BH, Latitude_BH** : Longitude and Latitude of bottom hole point (decimal degrees).
- **TrueVerticalDepth_FT** : Total vertical depth to bottom of wellbore (feet).
- **LateralLength_FT** : Length of the horizontal drain (feet).
- **ProppantIntensity_LBSPerFT, ProppantLoading_LBSPerGAL** and **Proppant_LBS** : Characteristics of Proppant injected during completion (LBS)
- **WaterIntensity_GALPerFT** and **TotalWaterPumped_GAL**: Characteristics of Water injected during completion (GAL)
- **FluidIntensity_GALPerFT** and **TotalFluidPumped_GAL**: Characteristics of Fluid injected during completion (GAL)
- **First12MonthProd_BOE** : Production of Gas and Oil over 12 months (BOE) - variable of interest to predict -

After removing Not-Available rows and extracting only interest variables, our data-set has become composed of 12 variables and 1580 rows and standardized as described in 4.1.

The Principal Components Analysis (PCA) results shows that the first factorial plan, composed of production and some "Fracturing design" as 1^{st} axis, and localization variables as 2^{nd} axis, represents approximately 60% of initial data, which not enough to represent it faithfully because 40% of information are lost. Moreover, according to the correlation circle, *First12MonthProd_BOE* is likely to be more correlated to Fracturing design, whereas *TrueVerticalDepth_FT* and *LateralLength_FT* can be taken separately in the 3^{rd} and 4^{th} axis

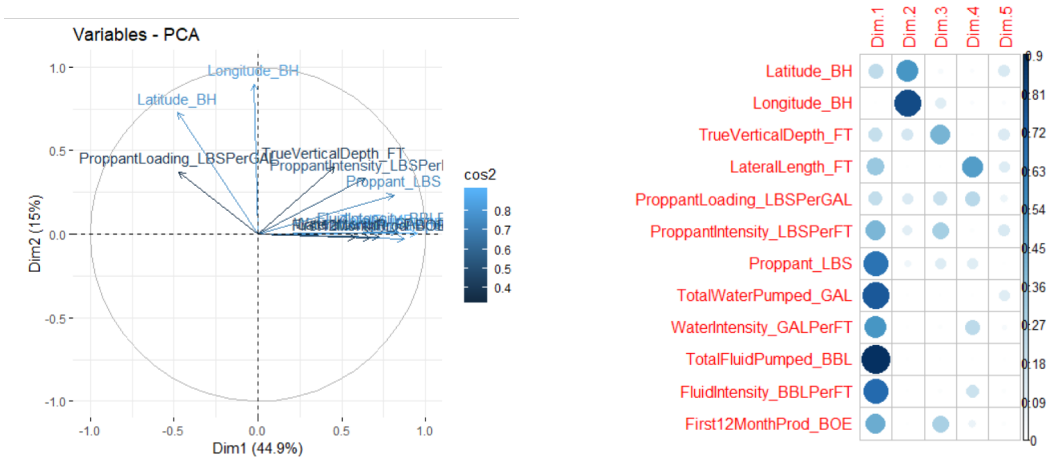


Figure 5.4 – On the left : Correlation Circle of 1^{st} factorial plan. On the right : Contribution of each variable to factorial axis

Regarding screening inputs (Step 2 4.2) with HSIC, We identify *TrueVerticalDepth_FT* and *Latitude_BH* as most influent variables while *Longitude_BH* and *ProppantIntensity_LBSPerFT* are less influent.

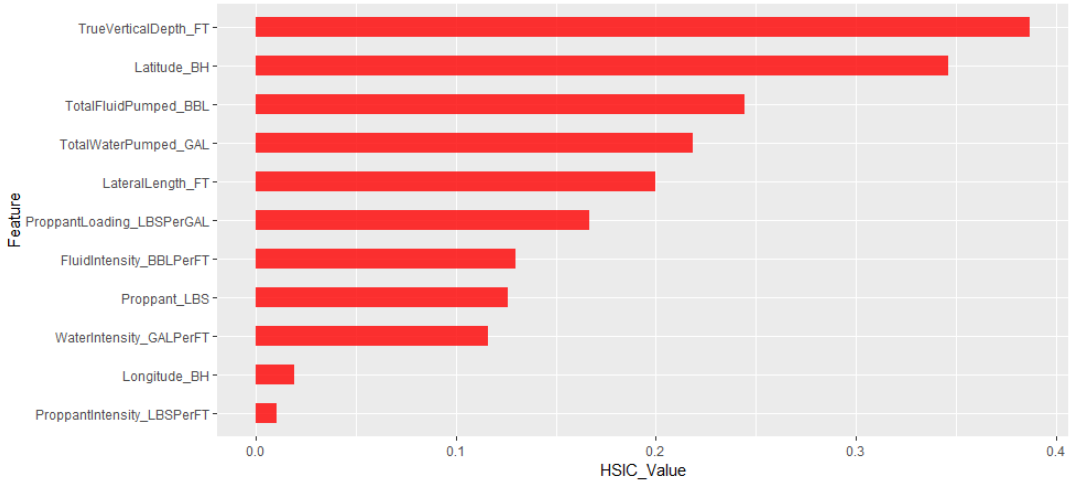


Figure 5.5 – HSIC indices $S_{X_k}^{HSIC_{\mathcal{F},\mathcal{G}}}$ for Production inputs

$HSIC_{\mathcal{F},\mathcal{G}}$ exhibits also a high dependence between *First12MonthProd_BOE* and Fracturing parameters as PCA correlation circle (see 5.4) who fails to identify the influence *TrueVerticalDepth_FT* and *Latitude_BH*.

5.2.3 Modeling Production with GP

Predicting the Production over 12 months

The data-set has been split to a training set of $n_{train} = 1100$ observations and testing set of $n_{test} = 480$ (corresponding to 75%-25% split).

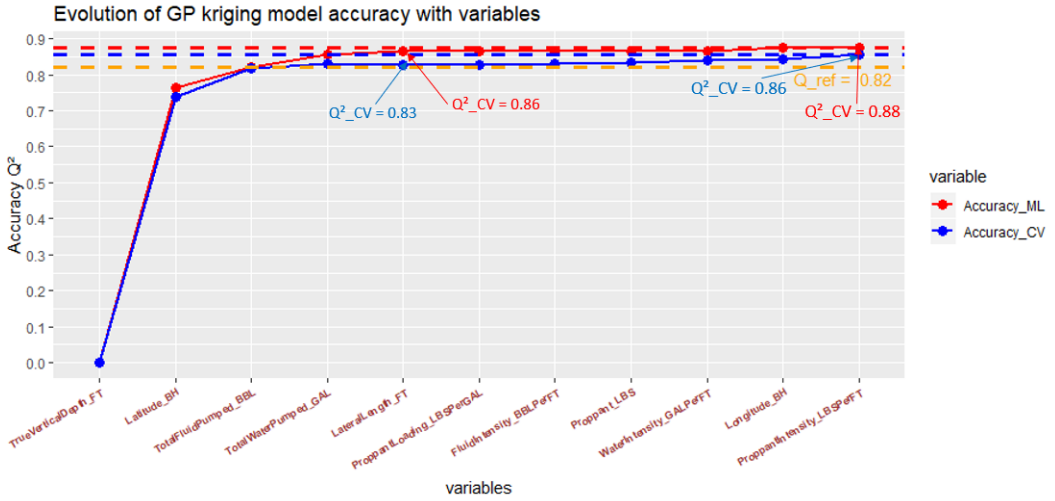


Figure 5.6 – Evolution of model's accuracy Q^2 for each feature included at iteration j^{th}

From the HSIC-Based screening (5.5) and build a joint GP model according to the sequential approach in 4.3 to predict production $Y = First12MonthProd\_BOE$ for testing set, considered as undiscovered wells, we build also a simple GP model including all inputs, for comparison purposes. The covariance model chosen is an Radial model with a Matérn 3/2 kernel.

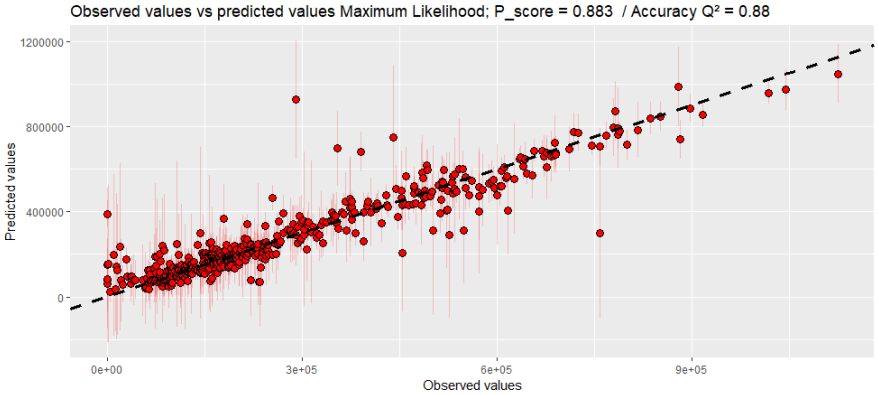
By examining the graph above (5.6) , we infer that :

- Firstly, the sequential joint modeling improve Q^2 by 5% as long as the stochastic part due to nugget effect is more important in our data-set.
- Secondly, the accuracy is increasing until the 5^{th} iteration corresponding to $LateralLength\_FT$ where it stagnates before increasing a little bit at $Longitude\_BH$. The predictive model GP_{pred} is built using only 5 most influent variables ($TrueVerticalDepth\_FT$ to $LateralLength\_FT$), the global model GP_{global} is built using all inputs.
- Finally, Maximum Likelihood method seems to predict better Cross-Validation especially in the first iterations.

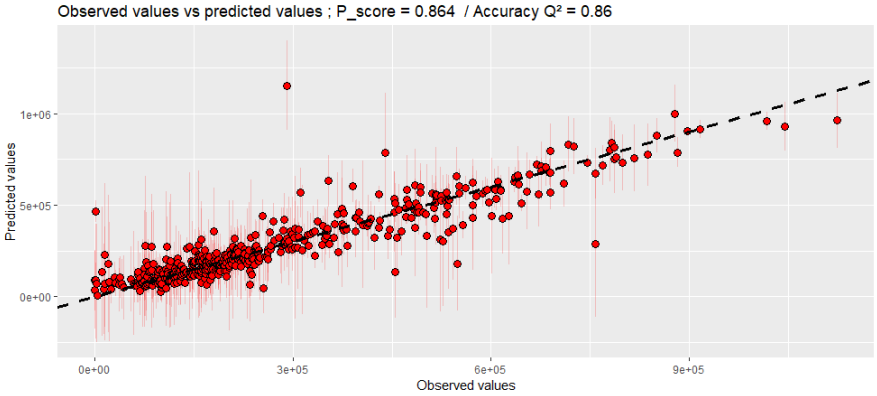
Remark 13 *The GP parameter's estimator do not converge for $TrueVerticalDepth\_FT$. We set $Q^2 = 0$ as a default accuracy.*

Remark 14 *In GP_{pred} we lose 0.02 in accuracy Q^2 but we reduce the complexity of them model for hyper-parameters estimation by more than half.*

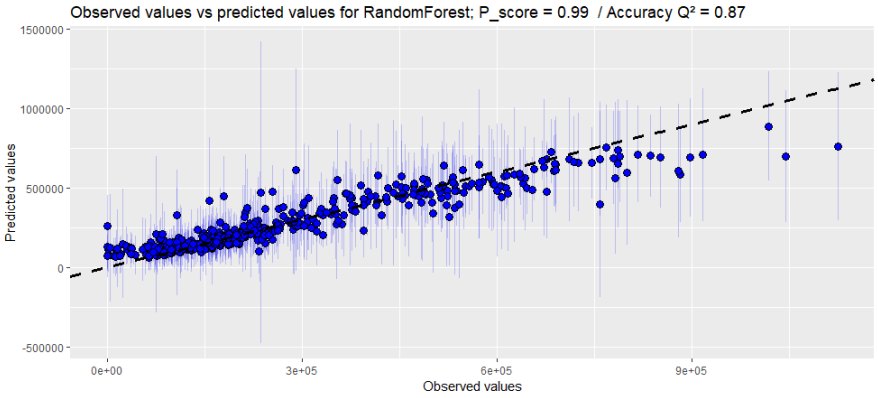
In the following, we choose the default confidence level $1-\alpha = 95\%$ (i.e. $\alpha = 5\%$) and compare the obtained kriging model with other ML algorithms, in particular Random Forest and Gradient Boosting.



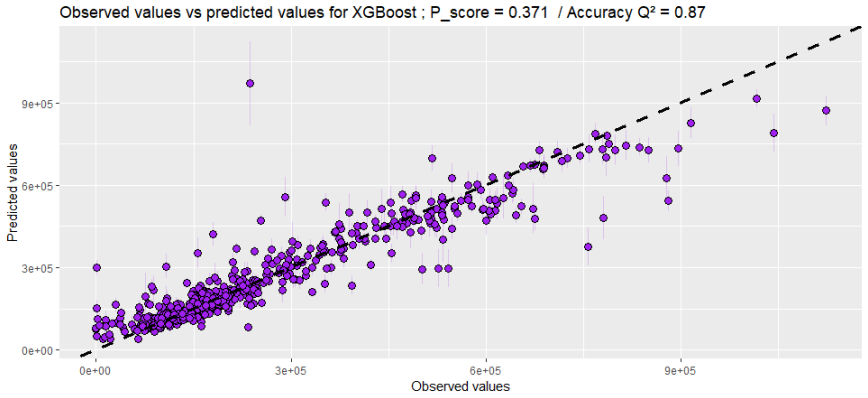
(a) Maximum Likelihood GP model



(b) Cross-Validation GP model



(c) Random Forest



(d) XGBoost

Figure 5.7 – Comparison of different ML models accuracy and score

	Maximum Likelihood	Cross-Validation	Random Forest	XGBoost
Computing time (s)	207.5	507.2	2.33	96.3

Table 5.4 – Computing time for different ML models

Clearly, the GP Models have approximately the same predictivity compared to XGBoost and RandomForest. Furthermore, we quite underestimate *First12MonthProd_BOE* unlike XGBoost, which underestimate real production heavily (less than 38% of predictions are inside $IC_{95\%}$) Random Forest, which overestimates it (more than 99%). Still, in terms of complexity and computing resources, the GP model is computationally expensive, and one would instead prefer using Random Forest for building reliable models as it is much faster and more robust.

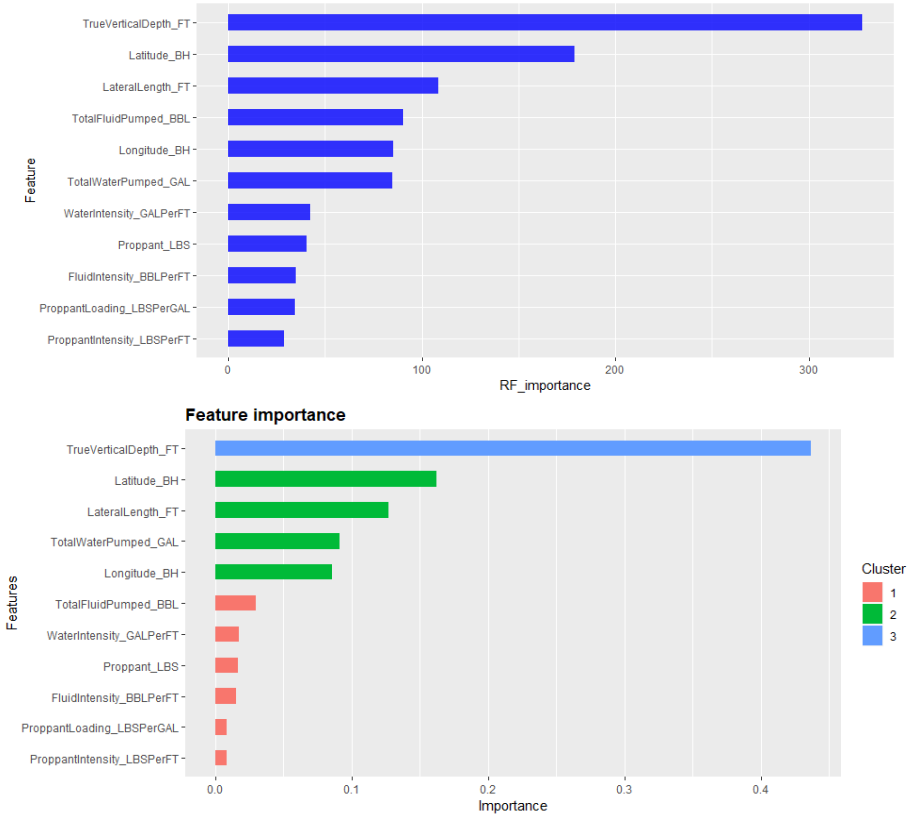


Figure 5.8 – Importance feature selection for XGBoost and Random Forest

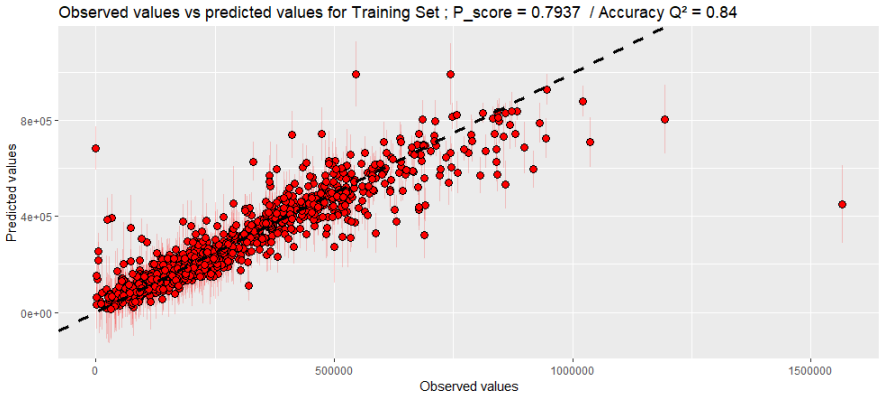
From the importance feature ranking 5.8, it is interesting to see that both XGBoost and Random Forest are exactly the same and that some features were ranked more important than those with HSIC, in particular, *Longitude_BH* is ranked as 5^{th} important variable, which also can explain the slight increase in accuracy Q^2 at step 10^{th} (See 5.6).

The importance feature selection, as explained in remark 9, is useful for predictive modelling purposes (but not for model explainability or sensitivity analysis). Unfortunately, it is inapplicable in the case of GP models as it requires $d!$ built models (iterations) to explore all possibilities and pick the most important variables for Q^2. However, we can serve us for importance feature selection of the previous ML algorithms to build GP_{pred}. Indeed, taking the five first important variables by Random Forest or XGBoost and using MLE, we get $Q^2 = 0.88$ for the new GP_{pred}.

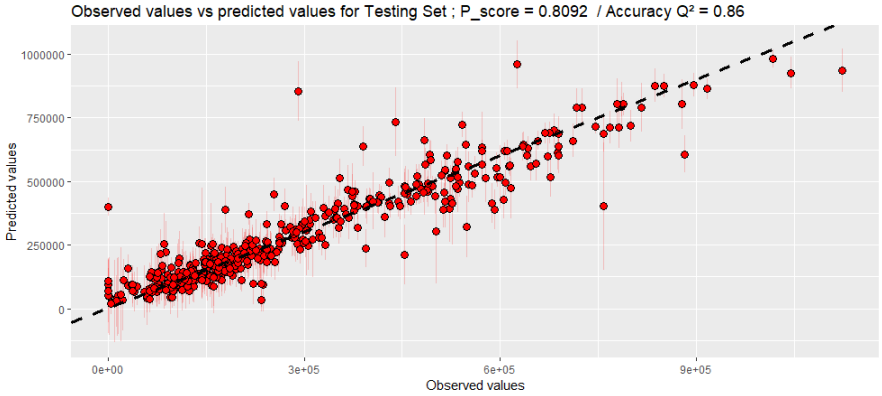
Uncertainty quantification : Estimating percentiles probability P_{90}/P_{10}

The definitions of rules P_{90}/P_{10} of PRMS and SEC correspond in fact to confidence level $1-\alpha = 80\%$. Since the purpose now is the estimate properly this two percentile, the GP_{pred} and Cross-

Validation Estimator are the main tool for. The predictions and associated Prediction's intervals of the GP_{pred}^{score} are presented below in 5.9 :



(a) Training set



(b) Testing set

Figure 5.9 – $\mathbb{P}_{1-0.20}^{score}$ obtained by LOO Cross-Validation

In both cases, the $\mathbb{P}_{1-0.20}^{score}$ criterion is respected as much as possible even if the accuracy Q^2 decreases a little bit. In particular, we obtain the empirical percentiles $\tilde{P}_{10} = \frac{1}{2}\left(1 - \mathbb{P}_{1-0.20}^{score}\right) \simeq 9.54\%$ and $\tilde{P}_{90} = \frac{1}{2}\left(1 + \mathbb{P}_{1-0.20}^{score}\right) \simeq 90.46\%$ for the testing set

5.2.4 Sensitivity indices

In the sensitivity analysis of the model, GP_{global} is used despite its high dimension which makes estimating Shapley values more difficult :

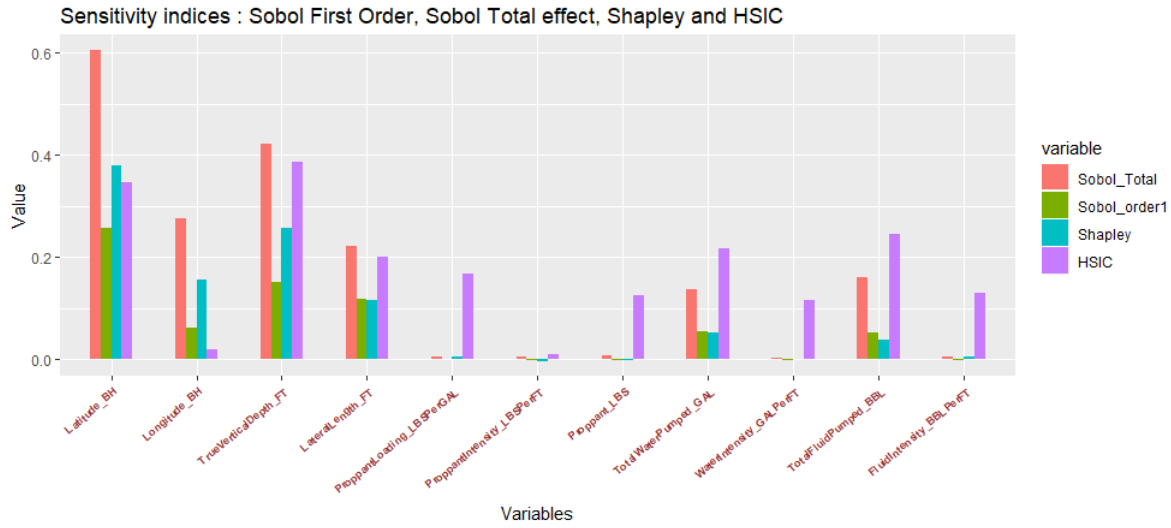


Figure 5.10 – Sensitivity analysis indices for UTICA Production data.

Concerning Sobol first-order and Shapley values, although *Latitude_BH* was ranked 2^{nd}, its contribution to the variance of the production is more important than other variables. In addition, $S_{Longitude\_BH}$ is small which means that it that its contribution alone to the production is also small but when it is combined with other variables (e.g. *Latitude_BH*), the contribution becomes more important.

Remark 15 *The "sandwich effect" of Sobol's indices and Shapley values is also valid in our case regardless of estimation's error.*

For Sobol total effects, all variables with $S_{T_i} \approx 0$ can be considered as not intervening in the model $Y = f(.)$. The model can be written then as :

$$\begin{aligned} First12MonthProd\_BOE = f(&TrueVerticalDepth\_FT, Longitude\_BH, Latitude\_BH, \\ &LateralLength\_FT, TotalFluidPumped\_BBL, TotalWaterPumped\_GAL) \end{aligned} \tag{5.10}$$

Note that these variables correspond to those ranked by Random Forest and XGBoost as most important. They are variables representing the predictive part of the model $Y = f(.)$

Remark 16 *When looking at scatter plot of Longitude_BH and First12MonthPro_BOE in 5.12, we can observe that these two variables cannot be fitted* ***alone*** *to predict production, it could be a reason why $HSIC_{\mathcal{F},\mathcal{G}}$ fails to rank them as influent variables.*

When a GP model GP^* is built (Matén 3/2 kernel) by taking the less important variables : *WaterIntensity_GALPerFT* to *ProppantIntensity_LBSPerFT*, the obtain accuracy is $Q^{*2} = 0.56$. That means the less important variables are still useful to gather information about production and able to predict 56% of real production.

From 5.11, One can deduce that :

- *ProppantLoading_LBSPerGAL* do not affect production neither directly nor by its dependencies, *HSIC* could be explained by a certain hidden variable affecting both of *ProppantLoading_LBSPerGAL* and *First12MonthProd_BOE*.

- $S_{ProppantIntensity\_LBSPerFT} = S_{WaterIntensity\_GALPerFT} = 0$ means that both of *Proppant-Loading_ LBSPerGAL* are affecting production only by their interactions/dependencies and that their contribution to the variance of the model is null.
- In 5.10, HSIC measured some dependencies between *WaterIntensity_ GALPerFT*, *WaterIntensity_ GALPerFT* and *Proppant_ LBS* and *First12MonthProd_ BOE* although other variance-based indices have given them a very small value. These dependencies are detected now by Sobol's and Shapley indices when most important variables are taken out. However, we are unable to interpret exactly the difference between HSIC and variance-based indices for *ProppantIntensity_ LBSPerFT* and *ProppantLoading_ LBSPeGAL* (could be caused by the imprecision of GP^* ? presence of a confounder?)

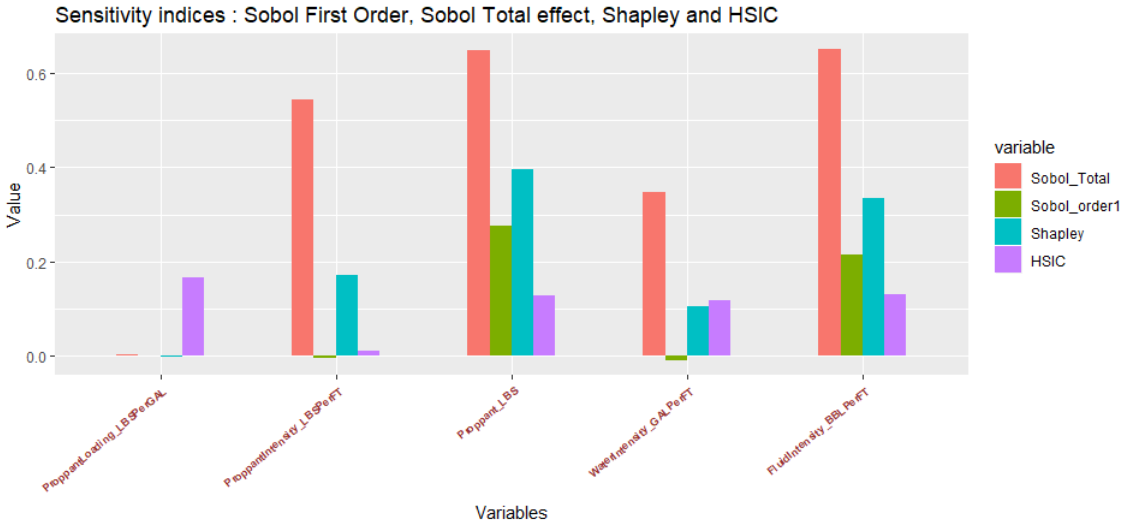


Figure 5.11 – Sensitivity analysis indices for remaining variables.

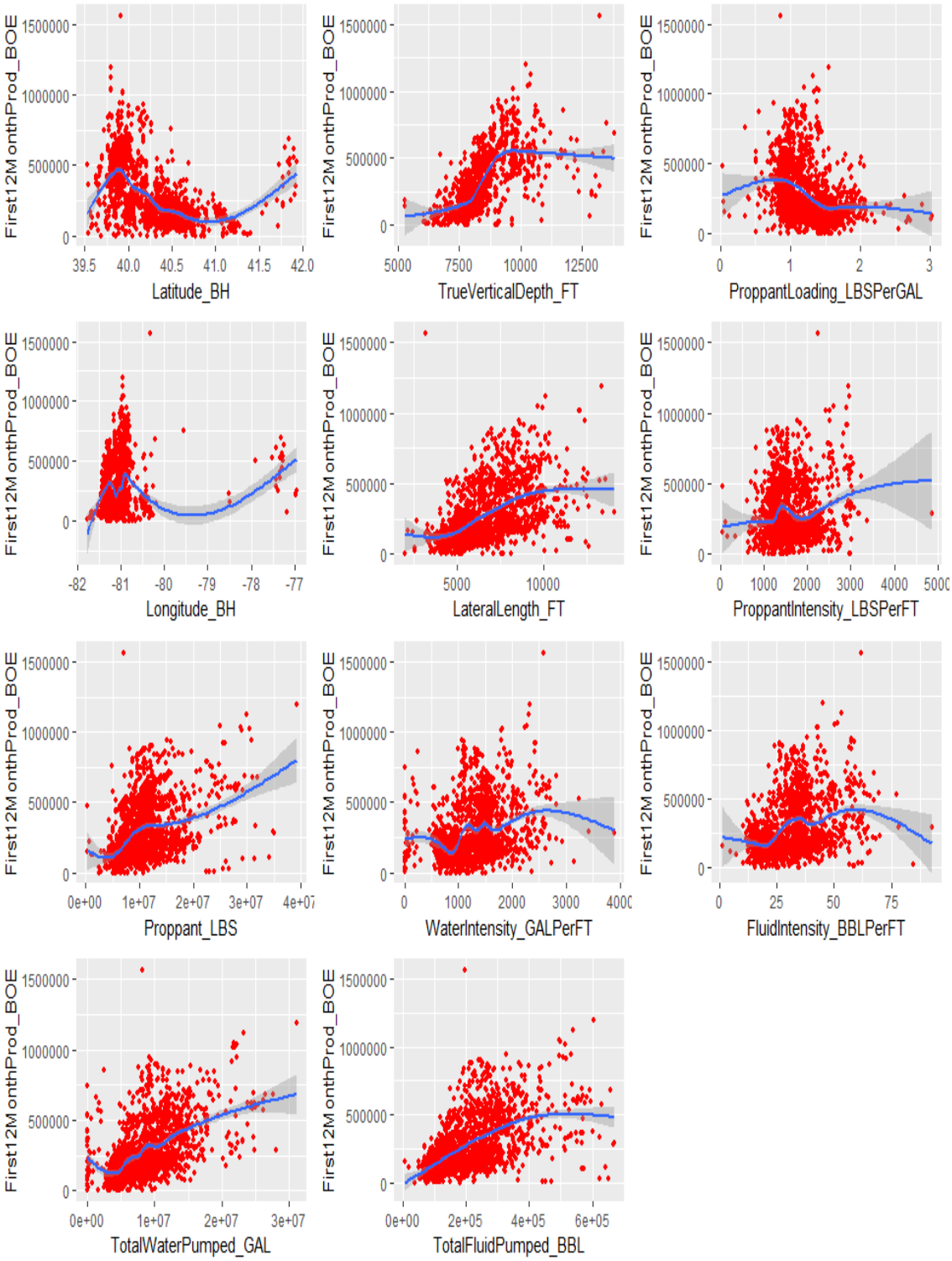


Figure 5.12 – Scatter plot of Y =First12MonthProd_BOE for UTICA inputs; the blue lines design the smooth mean of observed data

Chapter 6

Conclusion

The general field of this master thesis was uncertainty quantification and sensitivity analysis. In this conclusion, we give a brief summary of our work followed by a number of directions for future research.

This thesis aimed to investigate the issue of uncertainty quantification for Oil & Gas production. The purpose of this thesis is twofold, on the one hand, to present appropriate statistical methods and procedures for analyzing inputs contribution and dealing with parameter uncertainty. On the other hand, develop a predictive model in terms of accuracy and/or with respect to the SEC and PRMS rules of percentiles P_{90}/P_{10}.

The two different concepts of sensitivity analysis measures have been presented in chapter 2, both of them share the ability to study the contribution of inputs to the model uncertainty and provide deeper insight into the output behaviour. However, in some cases, they remain inapplicable as they suffer from many problems due mainly to the computational cost of these indices.

The Gaussian Processes GP, based on RKHS theory, have also shown their efficiency in modelling data. Although being computationally expensive, they allow building high predictive models with a well-founded framework for learning and model selection. In addition, the choice of kernels and the robustness of the Gaussian Process make fitting the percentile P_{90}/P_{10} possible by hyper-parameters Cross-Validation estimating.

Finally, it would be necessary to explore and investigate new research topics further to expand the findings of this thesis, in particular:

- Estimating Sobol's indices and Shapley values when the model's or surrogate model's accuracy is low: As mentioned in section 2.4, estimating these indices requires conditioning on a variable X_i, when the surrogate model $\hat{f}(.)$ is poor, they are estimated imprecisely. In addition, the hardness of conditional testing (peters may make computing Shapley values from data more challenging.

- Interpreting HSIC measure between X_k and the output : In some cases (See 5.1.3 and 5.2.4), $S_{X_k}^{HSIC_{\mathcal{F},\mathcal{G}}}$ indicates a non-null value meaning that X_k and Y are dependent somehow, this is insufficient to decide whether X_k appears in the model $Y = f(.)$ or not even through its dependencies with other inputs.

- Multivariate HSIC measure: It could be interesting to exhibit the HSIC not only between two vectors but in the multivariate case between Y and set of vectors $X_{\mathcal{J}}$ where $\mathcal{J} \subseteq \mathbb{N}$ to understand better dependencies between $X_k \in X_{\mathcal{J}}$ and Y

- Choice of kernels in HSIC: In the original HSIC papers [11], and [35], we show that HSIC depends on the universal kernels $\mathcal{F}, \mathcal{G}$ and the Gaussian kernel was chosen. However, there is no theoretical justification for this choice. The impact of kernels must be studied to see if HSIC is stable or not.

Bibliography

[1] L. Stasielowicz and R. Suck. Distance correlation: Discovering meta-analytic relationships between variables when other correlation coefficients fail. In *Research Synthesis, Dubrovnik: Methods in meta-analysis*. ZPID (Leibniz Institute for Psychology Information), May 2019.

[2] C. R. Clarkson. Production data analysis of unconventional gas wells: Review of theory and best practices. *Production data analysis of unconventional gas wells: Review of theory and best practices*, pages 109–110, 101—-146, 2013.

[3] B. Sudret. *Uncertainty propagation and sensitivity analysis in mechanical models – Contributions to structural reliability and stochastic spectral methods.* PhD thesis, Université Blaise Pascal, Clermont-Ferrand, France, 2007.

[4] A. Saltelli, K. Chan, and E.M. Scott. *Sensitivity Analysis*. Wiley Series in Probability and Statistics., 2000.

[5] I. Sobol. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, pages 112–118, 1990.

[6] I. Sobol. On sensitivity estimation for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, pages 407–414, 1993.

[7] A. Saltelli and T. Homma. Importance measures in global sensitivity analysis of non linear models. *Reliability Engineering and System Safety*, pages 1—17, 1996.

[8] I. Sobol and S. Kucherenko. Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, 79:3009–3017, 2009.

[9] I.M. Sobol and S. Kucherenko. Derivative based global sensitivity measures. *Procedia - Social and Behavioral Sciences*, 2(6):7745 – 7746, 2010. Sixth International Conference on Sensitivity Analysis of Model Output.

[10] B. Iooss and P. Lemaître. *A review on global sensitivity analysis methods.* Springer., 2014.

[11] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. *International conference on algorithmic learning theory*, 2005.

[12] S. Da Veiga. Global sensitivity analysis with dependence measures. *ArXiv*, abs/1311.2483, 2013.

[13] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325, 09 1948.

[14] B. Efron and C. Stein. The jackknife estimate of variance. *Ann. Statist.*, 9(3):586–596, 05 1981.

[15] H. Cukier, R.I. Levine, and K. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 1978.

[16] H. Cukier, R.I. Levine, and K. Shuler. A quantitative, model-independent method for global sensitivity analysis of model output. *Technometrics*, pages 39—51, 1999.

[17] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communication*, page 280–297, 2002.

[18] J-Y. Tissot and C. Prieur. A bias correction method for the estimation of sensitivity indices based on random balance designs. *Reliability Engineering and System Safety*, pages 205—-213, 2012.

[19] B. Iooss, F. Van Dorpe, and N. Devictor. Response surfaces and sensitivity analyses for an environmental model of dose calculations. *Reliability Engineering and System Safety*, page 1241–1251, 2006.

[20] A. Owen. Randomly permuted (t,m,s)-nets and (t,s)-sequences. In H. Niederreiter and P. J. Shieu, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*. Springer-Verlag, New York, USA, 1994.

[21] N. A. Weiss. *A Course in Probability*. Pearson Addison Wesley, 2005.

[22] L. Shapley. *A value for N-person games*. Defense Technical Information Center, 1953.

[23] E. Song, B. L. Nelson, and J. Staum. Shapley effects for global sensitivity analysis: Theory and computation. *Society for Industrial and Applied Mathematics*, 2016.

[24] E. Winter. *The Shapley value*. Handbook of game theory with economic applications, 2002.

[25] A. B. Owen. Sobol' indices and shapley value. *Society of Industrial and Applied Mathematics*, 2014.

[26] B. Iooss and C. Prieur. Shapley effects for sensitivity analysis with correlated inputs: comparisons with Sobol' indices, numerical estimation and applications. working paper or preprint, March 2019.

[27] J. Castroa, D. Gomez, and J. Tejada. Polynomial calculation of the shapley value based on sampling. *Computers and Operations Research*, pages 1726 – 1730, 2009.

[28] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[29] G. J. Szekely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing independence by correlation of distance. *Annals of Statistics*, pages 2769—2794, 2007.

[30] G. J. Szekely and M. L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, pages 1236–1265, 2009.

[31] E. Martinez-Gomez, M. T. Richards, and D. St. P. Richards. Distance correlation methods for discovering associations in large astrophysical database. *Astrophysical Journal*, page 39 (11 pp), 2014.

[32] A. Feuerverg. A consistent test for bivariate dependence. *International Statistical Review*, pages 419–433, 1993.

[33] G. J. Szekely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, pages 1249—1272, 2013.

[34] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[35] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Scholkopf, and N. Logothetis. Kernel constrained covariance for dependence measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 1–8, January 2005.

[36] C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[37] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, 2004.

[38] D. Sejdinovic, B. K. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *ArXiv*, abs/1207.6076, 2012.

[39] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *NIPS*, 2009.

[40] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(1):1393–1434, May 2012.

[41] G. Matheron. *La Théorie des variables régionalisées, et ses applications.* Les Cahiers du Centre de morphologie mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines de Paris, 1970.

[42] N.A.C. Cressie. *Statistics for spatial data.* Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1993.

[43] J. Sacks, W .J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–423, 11 1989.

[44] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, 2005.

[45] C. Currin, T. J. Mitchell, M. D. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991.

[46] T. Santner, B. Williams, and W. Notz. *The Design and Analysis Computer Experiments.* Springer Series in Statistics, 01 2003.

[47] E. Vazquez, E. Walter, and G. Fleury. Intrinsic kriging and prior information. *Applied Stochastic Models in Business and Industry*, 21:215 – 226, 03 2005.

[48] H. Tolba, N. Dkhili, J. Nou, J. Eynard, S. Thil, and S. Grieu. GHI forecasting using Gaussian process regression. In *IFAC Workshop on Control of Smart Grid and Renewable Energy Systems*, Jeju, South Korea, June 2019.

[49] F. Bachoc. *Estimation paramétrique de la fonction de covariance dans le modèle de Krigeage par processus Gaussiens : application à la quantification des incertitues en simulation numérique.* PhD thesis, University Paris 7, 2013. 2013PA077111.

[50] J. Oakley, A. Hagan, and A. O'Hagan. Probabilistic sensitivity analysis of complex models: A bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:751 – 769, 08 2004.

[51] B. Iooss and A. Marrel. An efficient methodology for the analysis and modeling of computer experiments with large number of inputs. In *UNCECOMP 2017 2nd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering*, pages 187–197, Rhodes Island, Greece, June 2017.

[52] K-T. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments (Computer Science and Data Analysis).* Chapman and Hall/CRC, 2005.

[53] O. Roustant, E. Padonou, Y. Deville, A. Clément, G. Perrin, J. Giorla, and H.P. Wynn. Group kernels for Gaussian process metamodels with categorical inputs. working paper or preprint, July 2018.

[54] A. Marrel, B. Iooss, S. Da Veiga, and M. Ribatet. Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing*, 22:833–847, 2012.

[55] J. P. C. Kleijnen and R. G. Sargent. A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, 120:14–29, 2000.

[56] H. Moon. *Design and Analysis of Computer Experiments for Screening Input Variables*. PhD thesis, The Ohio State University, Columbus, OH, USA, 2010. AAI3425387.

[57] H. Moon, A. M. Dean, and T. J. Santner. Two-stage sensitivity-based group screening in computer experiments. *Technometrics*, 54(4):376–387, 2012.

[58] W. J. Morokoff and R. E. Caflisch. Quasi-monte carlo integration. *JOURNAL OF COMPUTATIONAL PHYSICS*, 122:218–230, 1995.

[59] T. Crestaux, O. P. Le Maître, and J-M Martinez. Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety*, 94:1161–1172, 05 2009.

[60] M. D. Morris, L. Moore, and M. D. McKay. Sampling plans based on balanced incomplete block designs for evaluating the importance of computer model inputs. *Journal of Statistical Planning and Inference*, 136:3203–3220, 09 2006.

[61] C. Linkletter, D. Bingham, N. Hengartner, D. Higdon, and K. Ye. Variable selection for gaussian process models in computer experiments. *Technometrics*, 48:478–490, 11 2006.

[62] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In *[1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis*, pages 398 – 403, 01 1991.