



HAL
open science

Nonparametric adaptive estimation of order 1 Sobol indices in stochastic models, with an application to Epidemiology

Gwenaëlle Castellan, Anthony Cousien, Viet Chi Tran

► **To cite this version:**

Gwenaëlle Castellan, Anthony Cousien, Viet Chi Tran. Nonparametric adaptive estimation of order 1 Sobol indices in stochastic models, with an application to Epidemiology. 2015. hal-01249333v1

HAL Id: hal-01249333

<https://hal.science/hal-01249333v1>

Preprint submitted on 31 Dec 2015 (v1), last revised 15 Jan 2020 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nonparametric adaptive estimation of order 1 Sobol indices in stochastic models, with an application to Epidemiology

Gwenaëlle Castellan,^{*} Anthony Cousien,[†] Viet Chi Tran[‡]

December 31, 2015

Abstract

The global sensitivity analysis is a set of methods aiming at quantifying the influence of the uncertainty about the inputs parameters of a model on the variability of the responses. In a deterministic framework, i.e. when the same inputs values give always the same outputs values, the estimation of the Sobol indices is a commonly-used method. This method is based on the variance decomposition aiming at estimating the contribution of each parameter (or combination of parameters) on the variance of the response. We consider here the estimation of the Sobol indices of order 1, which are usually estimated by replicated simulations of the model. In the case of a stochastic framework, i.e. when the model response is not unique for a same input parameter set due to random numbers generation in the model, metamodels are often used to approximate the mean and the dispersion of the response by deterministic functions thus allowing to recover the classical deterministic framework. We propose a new non-parametric estimator without the need of defining a metamodel to estimate the Sobol indices of order 1. The estimator is based on warped wavelets and it is adaptative in the regularity of the model. The convergence of the mean square error to zero, when the number of simulations of the model tend to infinity, is computed and an elbow effect is shown, depending on the regularity of the model.

Keywords: Sensitivity analysis in a stochastic framework; Sobol indices of order 1; adaptive non-parametric inference; warped wavelets; Nadaraya-Watson estimator; model selection; applications to epidemiology; SIR model; spread of the Hepatitis Virus C among drug users.

MSC2010: 49Q12; 62G08; 62P10.

1 Sobol indices

In a mathematical model where the output $y \in \mathbb{R}$ depends on a set of $p \in \mathbb{N}$ input parameters $x = (x_1, \dots, x_p) \in \mathbb{R}^p$ through the relation $y = f(x)$, there are various ways to measure the influence of the input x_ℓ , for $\ell \in \{1, \dots, p\}$, on y . In this article, we are interested in Sobol indices [21], which are based on an ANOVA decomposition. These indices have been proposed to take into account the uncertainty on the input parameters that are here considered as a realisation of a set of independent random variables $X = (X_1, \dots, X_p)$, with a known distribution.

^{*}Laboratoire Paul Painlevé, UMR CNRS 8524, UFR de Mathématiques, Université de Lille, Cité Scientifique, 59655 Villeneuve d'Ascq Cedex, France.

[†]INSERM, IAME, UMR 1137, Université Paris Diderot, Sorbonne Paris Cité, 75018 Paris, France

[‡]Laboratoire Paul Painlevé, UMR CNRS 8524, UFR de Mathématiques, Université de Lille, Cité Scientifique, 59655 Villeneuve d'Ascq Cedex, France.

Authors are listed in alphabetical order.

Denoting by $Y = f(X)$ the random response, the first order Sobol indices can be defined for $\ell \in \{1, \dots, p\}$ by

$$S_\ell = \frac{\text{Var}(\mathbb{E}[Y | X_\ell])}{\text{Var}(Y)}. \quad (1.1)$$

This first order index S_ℓ correspond to the sensitivity of the model to X_ℓ alone. Higher order indices can also be defined using ANOVA decomposition : considering $(\ell, \ell') \in \{1, \dots, p\}$, we can define the second order sensitivity, corresponding to the sensitivity of the model to the interaction between X_ℓ and $X_{\ell'}$ index by

$$S_{\ell\ell'} = \frac{\text{Var}(\mathbb{E}[Y | X_\ell, X_{\ell'}])}{\text{Var}(Y)} - S_\ell - S_{\ell'} \quad (1.2)$$

We can also define the total sensitivity indices by

$$S_{T_\ell} = \sum_{L \subset \{1, \dots, p\} | \ell \in L} S_L. \quad (1.3)$$

As the estimation of the Sobol indices can be computer time consuming, a usual practice consists in estimating the first order and total indices, to assess 1) the sensitivity of the model to each parameter taking alone and 2) the possible interactions, which are quantified by the difference between the total order and the first order index for each parameter. Several numerical procedures to estimate the Sobol indices have been proposed, in particular by Jansen [13] (see also [19, 20]). These estimators, that we recall in the sequel, are based on Monte-Carlo simulations of $(Y, X_1 \dots X_p)$.

The literature focuses on deterministic relations between the input and output parameters. In a stochastic framework where the model response Y is not unique for given input parameters, few works have been done, randomness being usually limited to input variables. Assume that:

$$Y = f(X, \varepsilon), \quad (1.4)$$

where $X = (X_1, \dots X_p)$ still denotes the random variables modelling the uncertainty of the input parameters and where ε is a noise variable. When noise is added in the model, the classical estimators do not always work: Y can be chaotic regarding the value of ε . Moreover, this variable is not always controllable by the user.

When the function f is linear, we can refer to [8]. For general cases, it is possible to add the seed of the noise as an additional input parameter and compute classical estimators, or to propose a meta-model, i.e. a deterministic function approximating the mean and the dispersion of the response by deterministic functions allows to come back in the classical deterministic framework (e.g. Janon et al. [12], Marrel et al. [17]). We study here another point of view, which is based on the non-parametric statistical estimation of the term $\text{Var}(\mathbb{E}[Y | X_\ell])$ appearing in the numerator of (1.1). We propose here a new approach based on the Nadaraya-Watson kernel regression or on wavelet decompositions. The kernel estimator presented in the sequel has been introduced independently from us by Solís [22]. An advantage of these non-parametric estimators is that their computations requires less simulations of the model. For Jansen estimators, the number of calls of f required to compute the sensitivity indices is $n(p + 2)$, where n is the number of independent random vectors $(Y^i, X_1^i, \dots X_p^i)$ ($i \in \{1, \dots n\}$) that are sampled for the Monte-Carlo, making the estimation of the sensitivity indices time-consuming for sophisticated models with many parameters. In addition, for the non-parametric estimators, the convergence of the mean square error to zero may be faster than for Monte-Carlo estimators, depending on the regularity of the model.

In a first section, we present the non-parametric estimators of the Sobol indices of order 1 in the case of the stochastic model (1.4) and study their convergence rates. These estimators are then computed and compared for toy examples introduced by Ishigami [11]. We then address models from Epidemiology. First, the stochastic continuous-time SIR model is considered, in which the population of size N is divided into three compartments: the susceptibles, infectious and removed individuals (see e.g. [1] for an introduction). Infections and removals occur at random times whose laws depend on the composition of the population and on the infection and removal parameters λ and μ as input variables. The output variable Y can be the prevalence or the incidence at a given time T for instance. Y naturally depends on λ , μ and on the randomness underlying the occurrence of random times. Then, we consider a stochastic multi-level epidemic model for the transmission of Hepatitis C virus (HCV) among people who inject drugs (PWID) that has been introduced by Cousien et al. [5, 6]. This model describes an individual-based population of PWID that is structured by compartments showing the state of individuals in the health-care system and by a contact-graph indicating who inject with whom. Additionally the advance of HCV in each patient is also taken into account. The input variables are the different parameters of the model. Outputs depend on these inputs, on the randomness of event occurrences and on the randomness of the social graph. We compare the sensitivity analysis performed by estimating the Sobol indices of order 1 with the naive sensitivity analysis performed in [5, 6] by letting the parameters vary in an *a priori* chosen windows.

In the sequel, C denotes a constant that can vary from line to line.

2 A non-parametric estimator of the Sobol indices of order 1

Denoting by $V_\ell = \mathbb{E}(\mathbb{E}^2(Y | X_\ell))$, we have:

$$S_\ell = \frac{V_\ell - \mathbb{E}(Y)^2}{\text{Var}(Y)}, \quad (2.1)$$

which can be approximated by

$$\hat{S}_\ell = \frac{\hat{V}_\ell - \bar{Y}^2}{\hat{\sigma}_Y^2} \quad (2.2)$$

where

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j \text{ and } \hat{\sigma}_Y^2 = \frac{1}{n} \sum_{j=1}^n (Y_j - \bar{Y})^2$$

are the empirical mean and variance of Y . In this article, we propose 2 approximations \hat{V}_ℓ of V_ℓ , based on Nadaraya-Watson and on warped wavelet estimators. At an advanced stage of this work, we learned that the Nadaraya-Watson-based estimator of Sobol indices of order 1 had also been proposed and studied in the PhD of Solís [22]. Using a result on estimation of covariances by Loubes et al. [16], they obtain an elbow effect. However their estimation is not adaptive. For the warped wavelet estimator, we propose a model selection procedure based on a work by Laurent and Massart [15] to make the estimator adaptive.

2.1 Definitions

Assume that we have n independent couples $(Y^i, X_1^i, \dots, X_p^i)$ in $\mathbb{R} \times \mathbb{R}^p$, for $i \in \{1, \dots, n\}$, generated by (1.4). Let us start with the kernel-based estimator:

Definition 2.1. Let $K : \mathbb{R} \mapsto \mathbb{R}$ be a kernel such that $\int_{\mathbb{R}} K(u)du = 1$ and Let $h > 0$ be a window and let us denote $K_h(x) = K(x/h)/h$. An estimator of S_ℓ for $\ell \in \{1, \dots, p\}$ is:

$$\widehat{S}_\ell^{(NW)} = \frac{\frac{1}{n} \sum_{i=1}^n \left(\frac{\sum_{j=1}^n Y_j K_h(X_\ell^j - X_\ell^i)}{\sum_{j=1}^n K_h(X_\ell^j - X_\ell^i)} \right)^2 - \bar{Y}^2}{\widehat{\sigma}_Y^2}. \quad (2.3)$$

This estimator is based on the Nadaraya-Watson estimator of $\mathbb{E}(Y | X_\ell = x)$ given by (e.g. [25])

$$\frac{\sum_{j=1}^n Y_j K_h(X_\ell^j - x)}{\sum_{j=1}^n K_h(X_\ell^j - x)}.$$

Replacing this expression in (2.2) provides $\widehat{S}_\ell^{(NW)}$. As mentioned before, this estimator has also been proposed by Solís [22].

Our second estimator is based on a warped wavelet decomposition of $\mathbb{E}(Y | X_\ell = x)$. For introduction to such decomposition, refer to [4, 14]. Let us denote by G_ℓ the cumulative distribution function of X_ℓ .

Let $(\psi_{jk})_{j \geq -1, k \in \mathbb{Z}}$ be a Hilbert wavelet basis of L^2 . In the sequel, we denote by $\langle f, g \rangle = \int_{\mathbb{R}} f(u)g(u)du$, for $f, g \in L^2$, the usual scalar product of L^2 . The wavelet ψ_{-10} is the father wavelet, and for $k \in \mathbb{Z}$, $\psi_{-1k}(x) = \psi_{-10}(x - k)$. The wavelet ψ_{00} is the mother wavelet, and for $j \geq 0$, $k \in \mathbb{Z}$, $\psi_{jk}(x) = 2^{j/2} \psi_{00}(2^j x - k)$.

Definition 2.2. Let us define for $j \geq -1$, $k \in \mathbb{Z}$,

$$\widehat{\beta}_{jk}^\ell = \frac{1}{n} \sum_{i=1}^n Y_i \psi_{jk}(G_\ell(X_\ell^i)). \quad (2.4)$$

Then, we define the (block thresholding) estimator of \widehat{S}_ℓ , for $J_n := \lceil \log_2 \left(\frac{\sqrt{n}}{\log(n)} \right) \rceil$, as:

$$\widehat{S}_\ell^{(WW)} = \frac{\widehat{\theta}_\ell - \bar{Y}^2}{\widehat{\sigma}_Y^2}, \quad (2.5)$$

$$\text{where } \widehat{\theta}_\ell = \sum_{j=-1}^{J_n} \left[\sum_{k \in \mathbb{N}} (\widehat{\beta}_{jk}^\ell)^2 - w(j) \right] \mathbb{1}_{\sum_{k \in \mathbb{N}} (\widehat{\beta}_{jk}^\ell)^2 \geq w(j)} \quad (2.6)$$

$$\text{with } w(j) = K' \left(\frac{2^j + \log 2}{n} \right) \text{ so that } \text{pen}(\mathcal{J}) = \sum_{j \in \mathcal{J}} w(j) = K \left(\frac{2^{J_{\max}}}{n} + \frac{x_{\mathcal{J}}}{n} \right) \quad (2.7)$$

where K and K' are positive constants, where $J_{\max} := \max \mathcal{J}$ and where

$$x_{\mathcal{J}} = J_{\max} \log(2). \quad (2.8)$$

Notice for the proofs that for $x_{\mathcal{J}}$ as in (2.8),

$$\limsup_{n \rightarrow +\infty} \frac{1}{n} \sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} e^{-x_{\mathcal{J}}} 2^{2J_{\max}} < +\infty. \quad (2.9)$$

Indeed, for a given $J_{\max} \leq J_n$ there are $2^{J_{\max}+1}$ subsets $\mathcal{J} \subset \{-1, \dots, J_n\}$ such that $\max \mathcal{J} = J_{\max}$. Thus:

$$\sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} e^{-x_{\mathcal{J}}} 2^{2J_{\max}} = C + C' \sum_{J_{\max}=0}^{J_n} 2^{J_{\max}} 2^{-J_{\max}} 2^{2J_{\max}} = C 2^{2J_n} = C \frac{n}{\log^2(n)},$$

where C in the first equality corresponds to the case $\mathcal{J} = \{-1\}$.

An expression of the constant K appears in the proofs of Section 4 (where the mean square error is studied). However this constant is hard to compute in practice and depends on inequalities that are maybe not optimal. Indeed, the proof is concentrated on the orders in n and in the dimension of the model corresponding to \mathcal{J} , not on obtaining the best constants. For applications, the constant K appearing in the penalty $\text{pen}(\mathcal{J})$ can be chosen by a slope heuristic approach (see e.g. [2]) explained at the end of the section.

Let us present the idea explaining the estimator proposed in Definition 2.2. Let us introduce centered random variables η_ℓ such that

$$Y = f(X, \varepsilon) = \mathbb{E}(Y | X_\ell) + \eta_\ell. \quad (2.10)$$

Let $g_\ell(x) = \mathbb{E}(Y | X_\ell = x)$ and $h_\ell(u) = g_\ell \circ G_\ell^{-1}(u)$. h_ℓ is a function from $[0, 1] \mapsto \mathbb{R}$ that belong to L^2 since $Y \in L^2$. Then

$$h_\ell(u) = \sum_{j \geq -1} \sum_{k \in \mathbb{Z}} \beta_{jk}^\ell \psi_{jk}(u), \quad \text{with} \quad \beta_{jk}^\ell = \int_0^1 h_\ell(u) \psi_{jk}(u) du = \int_{\mathbb{R}} g_\ell(x) \psi_{jk}(G_\ell(x)) G_\ell(dx). \quad (2.11)$$

Notice that the sum in k is finite because the function h_ℓ has compact support in $[0, 1]$. It is then natural to estimate $h_\ell(u)$ by

$$\widehat{h}_\ell = \sum_{j \geq -1} \sum_{k \in \mathbb{Z}} \widehat{\beta}_{jk}^\ell \psi_{jk}(u), \quad (2.12)$$

and we then have:

$$\begin{aligned} V_\ell &= \mathbb{E}(\mathbb{E}^2(Y | X_\ell)) = \int_{\mathbb{R}} G_\ell(dx) \left(\sum_{j \geq -1} \sum_{k \in \mathbb{Z}} \beta_{jk}^\ell \psi_{jk}(G_\ell(x)) \right)^2 = \int_0^1 \left(\sum_{j \geq -1} \sum_{k \in \mathbb{Z}} \beta_{jk}^\ell \psi_{jk}(u) \right)^2 du \\ &= \sum_{j \geq -1} \sum_{k \in \mathbb{Z}} (\beta_{jk}^\ell)^2 = \|h_\ell\|_2^2. \end{aligned} \quad (2.13)$$

Adaptive estimation of $\|h_\ell\|_2^2$ has been studied in [15], which provides the block thresholding estimator $\widehat{\theta}_\ell$ in the Definition 2.2. The idea is: 1) to sum the terms $(\beta_{jk}^\ell)^2$, for $j \geq 0$, by blocks $\{(j, k), k \in \mathbb{Z}\}$ for $j \in \{-1, \dots, J_n\}$ with a penalty $w(j)$ for each block to avoid choosing too large j s, 2) to cut the blocks that do not sufficiently contribute to the sum, in order to obtain statistical adaptation.

Notice that

$$\widehat{\theta}_\ell = \sup_{\mathcal{J} \subset \{-1, 0, \dots, J_n\}} \sum_{j \in \mathcal{J}} \left[\sum_{k \in \mathbb{N}} (\widehat{\beta}_{jk}^\ell)^2 - w(j) \right] = \sup_{\mathcal{J} \subset \{-1, 0, \dots, J_n\}} \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{N}} (\widehat{\beta}_{jk}^\ell)^2 - \text{pen}(\mathcal{J}). \quad (2.14)$$

In view of this identity, $\widehat{\theta}_\ell$ can be seen as an estimator of V_ℓ resulting from a model selection on the choice of the blocks $\{(j, k), k \in \mathbb{Z}\}$, $j \in \{-1, \dots, J_n\}$ that are kept, with the penalty function $\text{pen}(\mathcal{J}) = \sum_{j \in \mathcal{J}} w(j)$, for $\mathcal{J} \subset \{-1, \dots, J_n\}$.

For a given K appearing in the definition of the penalty function pen (2.7), let us denote by \mathcal{J}_K the subset of indices j of $\{-1, \dots, J_n\}$ achieving the supremum in the r.h.s. of (2.14). Plotting $\text{Card}(\mathcal{J}_K)$ as a function of K , the slope heuristic tells us to choose K as value where the curve has a sudden decrease.

2.2 Statistical properties

In this Section, we are interested in the rate of convergence to zero of the mean square error (MSE) $\mathbb{E}((S_\ell - \widehat{S}_\ell)^2)$. Let us consider the generic estimator \widehat{S}_ℓ defined in (2.2), where \widehat{V}_ℓ is an estimator of $V_\ell = \mathbb{E}(\mathbb{E}^2(Y | X_\ell))$. We first start with a Lemma stating that the MSE can be obtained from the rate of convergence of \widehat{V}_ℓ to V_ℓ . Then, we recall the result of Solís [22], where an elbow effect for the MSE is shown when the regularity of the density of (X_ℓ, Y) varies. The case of the warped wavelet estimator is studied at the end of the section.

Lemma 2.3. *Consider the generic estimator \widehat{S}_ℓ defined in (2.2). Then there is a constant C such that:*

$$\mathbb{E}((S_\ell - \widehat{S}_\ell)^2) \leq \frac{C}{n} + \frac{4}{\text{Var}(Y)^2} \mathbb{E}[(\widehat{V}_\ell - V_\ell)^2]. \quad (2.15)$$

Proof. From (2.1) and (2.2),

$$\begin{aligned} \mathbb{E}((S_\ell - \widehat{S}_\ell)^2) &= \mathbb{E}\left[\left(\frac{V_\ell - \mathbb{E}(Y)^2}{\text{Var}(Y)} - \frac{\widehat{V}_\ell - \bar{Y}^2}{\widehat{\sigma}_Y^2}\right)^2\right] \\ &\leq 2\mathbb{E}\left[\left(\frac{\mathbb{E}(Y)^2}{\text{Var}(Y)} - \frac{\bar{Y}^2}{\widehat{\sigma}_Y^2}\right)^2\right] + 2\mathbb{E}\left[\left(\frac{V_\ell}{\text{Var}(Y)} - \frac{\widehat{V}_\ell}{\widehat{\sigma}_Y^2}\right)^2\right]. \end{aligned} \quad (2.16)$$

The first term in the right hand side (r.h.s.) is in C/n . For the second term in the right hand side of (2.16):

$$\mathbb{E}\left[\left(\frac{V_\ell}{\text{Var}(Y)} - \frac{\widehat{V}_\ell}{\widehat{\sigma}_Y^2}\right)^2\right] \leq 2\mathbb{E}\left[\widehat{V}_\ell^2 \left(\frac{1}{\text{Var}(Y)} - \frac{1}{\widehat{\sigma}_Y^2}\right)^2\right] + \frac{2}{\text{Var}(Y)^2} \mathbb{E}[(\widehat{V}_\ell - V_\ell)^2]. \quad (2.17)$$

The first term in the r.h.s. is also in C/n , which concludes the proof. ■

2.2.1 MSE for the Nadaraya-Watson estimator

Using the preceding Lemma, Loubes Marteau and Solís prove an elbow effect for the estimator $\widehat{S}_\ell^{(NW)}$. Let us introduce $\mathcal{H}(\alpha, L)$, for $\alpha, L > 0$, the set of functions ϕ of class $[\alpha]$, whose derivative $\phi^{([\alpha])}$ is $\alpha - [\alpha]$ Hölder continuous with constant L .

Proposition 2.4 (Loubes Marteau and Solís [22, 16]). *Assume that $\mathbb{E}(X_\ell^4) < +\infty$, that the joint density $\phi(x, y)$ of (X_ℓ, Y) belongs to $\mathcal{H}(\alpha, L)$, for $\alpha, L > 0$ and that the marginal density of X_ℓ , ϕ_ℓ belongs to $\mathcal{H}(\alpha', L')$ for $\alpha' > \alpha$ and $L' > 0$. Then:*

If $\alpha \geq 2$, there exists a constant $C > 0$ such that

$$\mathbb{E}((S_\ell - \widehat{S}_\ell)^2) \leq \frac{C}{n}.$$

If $\alpha < 2$, there exists a constant $C > 0$ such that

$$\mathbb{E}((S_\ell - \widehat{S}_\ell)^2) \leq C \left(\frac{\log^2 n}{n}\right)^{\frac{2\alpha}{\alpha+2}}.$$

For smooth functions ($\alpha \geq 2$), Loubes et al. recover a parametric rate, while they still have a nonparametric one when $\alpha < 2$. Their result is based on (2.15) and a bound for $\mathbb{E}[(\widehat{V}_\ell - V_\ell)^2]$ given by [16, Th. 1], whose proof is technical. Since their result is not adaptive, they require the knowledge of the window h for numerical implementation. Our purpose is to provide a similar result for the warped wavelet adaptive estimator, with a shorter proof.

2.2.2 MSE for the warped wavelet estimator

Let us introduce first some additional notation. We define, for $\mathcal{J} \subset \{-1, \dots, J_n\}$, the projection $h_{\mathcal{J},\ell}$ of h on the subspace spanned by $\{\psi_{jk}$, with $j \in \mathcal{J}$, $k \in \mathbb{Z}\}$ and its estimator $\widehat{h}_{\mathcal{J},\ell}$:

$$h_{\mathcal{J},\ell}(u) = \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \beta_{jk}^\ell \psi_{jk}(u) \quad (2.18)$$

$$\widehat{h}_{\mathcal{J},\ell}(u) = \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \widehat{\beta}_{jk}^\ell \psi_{jk}(u). \quad (2.19)$$

We also introduce the estimator of V_ℓ for a fixed subset of resolutions \mathcal{J} :

$$\widehat{\theta}_{\mathcal{J},\ell} = \|\widehat{h}_{\mathcal{J},\ell}\|_2^2 = \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} (\widehat{\beta}_{jk}^\ell)^2. \quad (2.20)$$

Note that $\widehat{\theta}_{\mathcal{J},\ell}$ is one possible estimator \widehat{V}_ℓ in Lemma 2.3.

The estimators $\widehat{\beta}_{jk}^\ell$ and $\widehat{\theta}_{\mathcal{J},\ell}$ have natural expressions in term of the empirical process $\gamma_n(dx)$ defined as follows:

Definition 2.5. *The empirical measure associated with our problem is:*

$$\gamma_n(dx) = \frac{1}{n} \sum_{i=1}^n Y_i \delta_{G_\ell(X_\ell^i)}(dx) \quad (2.21)$$

where $\delta_a(dx)$ denotes the Dirac mass in a .

For a measurable function f , $\gamma_n(f) = \frac{1}{n} \sum_{i=1}^n Y_i f(G_\ell(X_\ell^i))$. We also define the centered integral of f with respect to $\gamma_n(dx)$ as:

$$\bar{\gamma}_n(f) = \gamma_n(f) - \mathbb{E}(\gamma_n(f)) \quad (2.22)$$

$$= \frac{1}{n} \sum_{i=1}^n \left(Y_i f(G_\ell(X_\ell^i)) - \mathbb{E}[Y_i f(G_\ell(X_\ell^i))] \right). \quad (2.23)$$

Using the empirical measure $\gamma_n(dx)$, we have:

$$\widehat{\beta}_{jk}^\ell = \gamma_n(\psi_{jk}) = \beta_{jk}^\ell + \bar{\gamma}_n(\psi_{jk}).$$

Let us introduce the correction term

$$\zeta_n = 2\bar{\gamma}_n(h_\ell) \quad (2.24)$$

$$\begin{aligned} &= 2 \left[\frac{1}{n} \sum_{i=1}^n Y_i h_\ell(G_\ell(X_\ell^i)) - \mathbb{E} \left(Y_1 h_\ell(G_\ell(X_\ell^1)) \right) \right] \\ &= 2 \left[\frac{1}{n} \sum_{i=1}^n h_\ell^2(G_\ell(X_\ell^i)) - \|h_\ell\|_2^2 \right] + \frac{2}{n} \sum_{i=1}^n \eta_\ell^i h_\ell(G_\ell(X_\ell^i)). \end{aligned} \quad (2.25)$$

Theorem 2.6. *Let us assume that the random variables Y are bounded by a constant M , and let us choose a father and a mother wavelets ψ_{-10} and ψ_{00} that are continuous with compact support (and thus bounded). The estimator $\widehat{\theta}_\ell$ defined in (2.6) is almost surely finite, and:*

$$\mathbb{E} \left[(\widehat{\theta}_\ell - V_\ell - \zeta_n)^2 \right] \leq C \inf_{\mathcal{J} \subset \{-1, \dots, J_n\}} \left(\|h_\ell - h_{\mathcal{J},\ell}\|_2^4 + \frac{2^{J_{max}}}{n^2} \right) + \frac{C'}{n \log^2(n)}, \quad (2.26)$$

for constants C and $C' > 0$.

We deduce the following corollary from the estimate obtained above. Let us consider the Besov space $\mathcal{B}(\alpha, 2, \infty)$ of functions $h = \sum_{j \geq -1} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}$ of L^2 such that

$$|h|_{\alpha, 2, \infty} := \sum_{j \geq 0} 2^{j\alpha} \sqrt{\sup_{0 < v \leq 2^{-j}} \int_0^{1-v} |h(u+v) - h(u)|^2 du} < +\infty.$$

For a $h \in \mathcal{B}(\alpha, 2, \infty)$ and $h_{\mathcal{J}}$ its projection on $\text{Vect}\{\psi_{jk}, j \in \mathcal{J} = \{-1, \dots, J_{\max}\}, k \in \mathbb{Z}\}$, we have the following approximation result from [10, Th. 9.4].

Proposition 2.7 (Härdle Kerkyacharian Picard and Tsybakov). *Assume that the wavelet function ψ_{-10} has compact support and is of class \mathcal{C}^N for an integer $N > 0$. Then, if $h \in \mathcal{B}(\alpha, 2, \infty)$ with $\alpha < N + 1$,*

$$\sup_{\mathcal{J} \subset \mathbb{N} \cup \{-1\}} 2^{\alpha J_{\max}} \|h - h_{\mathcal{J}}\|_2 = \sup_{\mathcal{J} \subset \mathbb{N} \cup \{-1\}} 2^{\alpha J_{\max}} \left(\sum_{j \geq J_{\max}} \sum_{k \in \mathbb{Z}} \beta_{jk}^2 \right)^{1/2} < +\infty. \quad (2.27)$$

Notice that Theorem 9.4 of [10] requires assumptions that are fulfilled when ψ_{-10} has compact support and is smooth enough (see comment after the Corol. 8.2 of [10]).

Corollary 2.8. *If ψ_{-10} has compact support and is of class \mathcal{C}^N for an integer $N > 0$ and if h_{ℓ} belongs to a ball of radius $R > 0$ of $\mathcal{B}(\alpha, 2, \infty)$ for $0 < \alpha < N + 1$, then*

$$\sup_{h \in \mathcal{B}(\alpha, 2, \infty)} \mathbb{E} \left[(\widehat{\theta}_{\ell} - V_{\ell})^2 \right] \leq C \left(n^{-\frac{8\alpha}{4\alpha+1}} + \frac{1}{n} \right). \quad (2.28)$$

As a consequence, we obtain the following elbow effect:

If $\alpha \geq \frac{1}{4}$, there exists a constant $C > 0$ such that

$$\mathbb{E}((S_{\ell} - \widehat{S}_{\ell})^2) \leq \frac{C}{n}.$$

If $\alpha < \frac{1}{4}$, there exists a constant $C > 0$ such that

$$\mathbb{E}((S_{\ell} - \widehat{S}_{\ell})^2) \leq C n^{-\frac{8\alpha}{4\alpha+1}}.$$

The proof of Theorem 2.6 is postponed to Section 4. Let us remark that in comparison with the result of Loubes et al. [16], the regularity assumption is on the function h_{ℓ} rather than on the joint density $\phi(x, y)$ of (X_{ℓ}, Y) . The adaptivity of our estimator is then welcomed since the function h_{ℓ} is *a priori* unknown. Remark that in application, the joint density $\phi(x, y)$ also has to be estimated and hence has an unknown regularity. For very regular functions $\alpha \rightarrow +\infty$, we recover a rate of convergence in n^{-2} in both cases.

Notice that in the case when $\alpha > 1/4$, we can show from the estimate of Th. 2.6 that:

$$\lim_{n \rightarrow +\infty} n \mathbb{E} \left[(\widehat{\theta}_{\ell} - V_{\ell} - \zeta_n)^2 \right] = 0, \quad (2.29)$$

which yields that $\sqrt{n}(\widehat{\theta}_{\ell} - V_{\ell} - \zeta_n)$ converges to 0 in L^2 . Since $\sqrt{n}\zeta_n$ converges in distribution to $\mathcal{N}\left(0, 4\text{Var}(Y_1 h_{\ell}(G_{\ell}(X_{\ell}^1)))\right)$ by the central limit theorem, we obtain that:

$$\lim_{n \rightarrow +\infty} \sqrt{n}(\widehat{\theta}_{\ell} - V_{\ell}) = \mathcal{N}\left(0, 4\text{Var}(Y_1 h_{\ell}(G_{\ell}(X_{\ell}^1)))\right), \quad (2.30)$$

in distribution.

The result of Corollary 2.8 is stated for functions h_{ℓ} belonging to $\mathcal{B}(\alpha, 2, \infty)$, but the generalization to other Besov space might be possible.

2.3 Numerical tests on toy models

We start with considering a toy model called the Ishigami function and defined as:

$$Y = f(X_1, X_2, X_3) = \sin(X_1) + 7 \sin(X_2)^2 + 0.1 X_3^4 \sin(X_1) \quad (2.31)$$

where X_i are independent uniform random variables in $[-\pi, \pi]$ (see e.g. [11, 19]).

Firstly, we consider this model with (X_1, X_2, X_3) as input parameters and compute the associated Sobol indices. For the Ishigami function, all the Sobol sensitivity indices are known.

$$S_1 = 0.3139, \quad S_2 = 0.4424, \quad S_3 = 0.$$

Secondly, following Marrel et al. [17], we consider the case where (X_1, X_2) are the input parameters and X_3 a nuisance random parameter. However, the Sobol indices have the same values as in the standard case.

In both cases, we compare the Nadaraya-Watson estimator of the Sobol indices of order 1 with the Jansen estimator [13] that is one of the classical estimator found in the literature (for the case of outputs that are deterministic functions of the inputs). The numerical implementation of the wavelet estimator is a work in progress. The Jansen estimator is based on the mixing of two replications of the sample (Y, X_1, \dots, X_p) , as described below.

Let us consider two samples $(X_1^{(1),i}, \dots, X_p^{(1),i}, i \in \{1, \dots, n\})$ and $(X_1^{(2),i}, \dots, X_p^{(2),i}, i \in \{1, \dots, n\})$ of i.i.d. p -uplets distributed as (X_1, \dots, X_p) . The Jansen estimators for the first order Sobol indices are, $\forall \ell \in 1, \dots, p$:

$$\widehat{S}_\ell = \widehat{\sigma}_Y^2 - \frac{1}{2n} \sum_{i=1}^n (f(X_1^{(2),i}, \dots, X_p^{(2),i}) - f(X_1^{(1),i}, \dots, X_{\ell-1}^{(1),i}, X_\ell^{(2),i}, X_{\ell+1}^{(1),i}, \dots, X_p^{(1),i}))^2 \quad (2.32)$$

The total order sensitivity indices are estimated by:

$$\widehat{S}_{T_\ell} = \frac{1}{2n} \sum_{i=1}^n (f(X_1^{(1),i}, \dots, X_p^{(1),i}) - f(X_1^{(1),i}, \dots, X_{\ell-1}^{(1),i}, X_\ell^{(2),i}, X_{\ell+1}^{(1),i}, \dots, X_p^{(1),i}))^2 \quad (2.33)$$

Notice that the estimation of the Sobol indices using Jansen estimators requires $N(p+2)$ simulations of the model. We computed the non-parametric estimators first from a sample of size n , then from a sample of size $(p+2)n$ to have a similar number of simulations of the model. We used $n = 10,000$ and we performed 1,000 replications to estimate the bias and MSE for each estimator. For the Nadaraya-Watson estimator, we choose $h = 0.01$. For the warped wavelet estimator, simulations are due very soon!

Table 1: *Estimates of the bias and MSE for the parameters X_1, X_2 and X_3 in the Ishigami function, for 1,000 replications and $n = 10,000$*

| Method | $\mathbb{E}[\widehat{S}_1 - S_1]$ | $\mathbb{E}[(\widehat{S}_1 - S_1)^2]$ | $\mathbb{E}[\widehat{S}_2 - S_2]$ | $\mathbb{E}[(\widehat{S}_2 - S_2)^2]$ | $\mathbb{E}[\widehat{S}_3 - S_3]$ | $\mathbb{E}[(\widehat{S}_3 - S_3)^2]$ |
|---------------------------|-----------------------------------|---------------------------------------|-----------------------------------|---------------------------------------|-----------------------------------|---------------------------------------|
| Jansen, $n(p+1)$ | 9.9e-4 | 1.8e-4 | 3.2e-5 | 1.0e-4 | 8.6e-4 | 5.6e-4 |
| Nadaraya-Watson, n | 6.6e-3 | 8.8e-5 | 4.4e-3 | 8.1e-5 | 9.5e-3 | 9.3e-5 |
| Nadaraya-Watson, $n(p+1)$ | 1.5e-3 | 1.1e-5 | 3.4e-4 | 1.6e-5 | 2.0e-3 | 4.3e-6 |

We can see that in the deterministic framework results Table 1 that for the 3 indices, the mean bias is higher for the Nadaraya-Watson estimator, even when the same number of simulation were computed $n(p+1)$. However, with the non-parametric estimator, the MSE was lower with the Nadaraya-Watson, even with five times less simulations of the model available. The results are similar in the stochastic framework Table 2, with the exception of the estimation of S_2 for which, with the same number of model simulations, the bias is comparable for Jansen and Nadaraya Watson (-7.8e-4 vs. 6.8e-4).

Table 2: Estimates of the bias and MSE for the parameters X_1 and X_2 in the Ishigami function, when X_3 is considered as a perturbation parameter, for 1,000 replications and $n = 10,000$

| Method | $\mathbb{E}[\hat{S}_1 - S_1]$ | $\mathbb{E}[(\hat{S}_1 - S_1)^2]$ | $\mathbb{E}[\hat{S}_2 - S_2]$ | $\mathbb{E}[(\hat{S}_2 - S_2)^2]$ |
|---------------------------|-------------------------------|-----------------------------------|-------------------------------|-----------------------------------|
| Jansen, $n(p+1)$ | -5.6e-4 | 2.0e-4 | -7.8e-4 | 1.8e-4 |
| Nadaraya-Watson, n | 5.8e-3 | 7.8e-5 | 4.7e-3 | 8.4e-5 |
| Nadaraya-Watson, $n(p+1)$ | 1.6e-3 | 1.2e-5 | 6.8e-4 | 1.5e-5 |

3 Sobol indices for epidemiological problems

We now consider two stochastic individual-based models of epidemiology in continuous time. In both cases, the population is of size N and divided into compartments. Input parameters are the rates describing the times that individuals stay in each compartment. These rates are usually estimated from epidemiological studies or clinical trials, but there can be uncertainty on their values. The restricted size of the sample in these studies brings uncertainty on the estimates, which are given with uncertainty intervals (classically, a 95% confidence interval). Different studies can provide different estimates for the same parameters. The study populations can be subject to selection biases. In the case of clinical trials where the efficacy of a treatment is estimated, the estimates can be optimistic compared with what will be the effectiveness in real-life, due to the protocol of the trials. It is important to quantify how these uncertainty on parameters estimations could impact the results and the conclusion of a modelling study.

3.1 SIR model and ODE metamodels

In the first model, we consider the usual SIR model, with three compartments: susceptibles, infectious and removed (e.g. [1]). We denote by S_t^N , I_t^N and R_t^N the respective sizes of the corresponding sub-populations at time $t \geq 0$, with $S_t^N + I_t^N + R_t^N = N$. At the population level, infections occur at the rate $\frac{\lambda}{N} S_t^N I_t^N$ and removals at the rate μI_t^N . The idea is that to each pair of susceptible-infectious individuals a random independent clock with parameter λ/N is attached and to each infectious individual an independent clock with parameter μ is attached. The input parameters are the rates λ and μ . The output parameter is the final size of the epidemic, i.e. at a time $T > 0$ where $I_T^N = 0$, $Y = (I_T^N + R_T^N)/N$.

It is possible to describe the evolution of $(S_t^N/N, I_t^N/N, R_t^N/N)_{t \geq 0}$ by a stochastic differential equation driven by Poisson point measures and it is known that when $N \rightarrow +\infty$, this stochastic process converges in $\mathbb{D}(\mathbb{R}_+, \mathbb{R}^3)$ to the unique solution $(s_t, i_t, r_t)_{t \geq 0}$ of the following system of ordinary differential equations (e.g. [1, 24]):

$$\begin{cases} \frac{ds}{dt} = -\lambda s_t i_t \\ \frac{di}{dt} = \lambda s_t i_t - \mu i_t \\ \frac{dr}{dt} = \mu i_t. \end{cases}$$

The fluctuations associated with this convergence have also been established. The limiting equations provide a natural deterministic approximating meta-model (recall [17]) for which sensitivity indices can be computed.

We applied the Jansen estimator and the Nadaraya-Watson estimator to the estimation of the first order Sobol indices of S_λ and S_μ . We applied these estimators to the SIR stochastic process and to SIR deterministic model (as a metamodel approximating the stochastic process) described above.

In a first example, we simulated a close population of 1200 individuals, with $S_0^{1200} = 1190$, $I_0^{1200} = 10$ and $R_0^{1200} = 0$. We choose this high population size to ensure the convergence of the simulations in the stochastic process to the solution of the ordinary differential equation system. The parameters distribution were Beta(2,2) distributions renormalized to have $\lambda/N \in [1/15000, 3/15000]$ and $\mu \in [1/15, 3/15]$.

In a second example, to increase the influence of ε on the output Y , we decreased the population size to 120 individuals, with $S_0^{120} = 119$, $I_0^{120} = 1$ and $R_0^{120} = 0$ and parameters drawn from Beta(2,2) distributions renormalized to have $\lambda/N \in [1/1500, 3/1500]$ and $\mu \in [1/15, 3/15]$.

The smoothed distributions of the first order Sobol indices are presented below, for 1,000 replications of the estimators and $n = 10,000$ are presented Figures 1 and 2 .

For the deterministic framework, the results are similar. The Jansen estimator for S_λ is $\hat{S}_\lambda = 0.45$ (standard deviation=8.1e-3) and $\hat{S}_\lambda^{(NW)} = 0.45$ (6.9e-3); for S_μ , we have $\hat{S}_\mu = 0.50$ (7.7e-3) and $\hat{S}_\mu^{(NW)} = 0.50$ (7.2e-3). We can however underline that the estimation of the Sobol indice with Jansen estimators required 40,000 simulations of the model vs. 10,000 for the Nadaraya-Watson estimator.

For the stochastic model with $N = 1,200$, we obtained similar results, but the Sobol indices were lower, due to the contribution of the randomness of the model to the variance of Y : $\hat{S}_\lambda = 0.38$ (9.6e-3) and $\hat{S}_\lambda^{(NW)} = 0.38$ (7.6e-3); $\hat{S}_\mu = 0.46$ (8.6e-3) and $\hat{S}_\mu^{(NW)} = 0.45$ (7.8e-3).

However, when we increased the contribution of the randomness to the variance of Y by considering a smaller population (120 individuals), the results were different: for Jansen $\hat{S}_\lambda = 0.063$ (0.022) and $\hat{S}_\lambda^{(NW)} = 0.087$ (5.4e-3); $\hat{S}_\mu = 0.12$ (0.021) and $\hat{S}_\mu^{(NW)} = 0.13$ (7.3e-3). We can not compare the two estimators in absence of theoretical values for the Sobol indices, but we can see that for small values, the two estimators provide different values, with the Nadaraya-Watson giving higher mean estimates but tighter distributions than the Jansen estimator.

3.2 Application to the spread of HVC among drug users

Chronic hepatitis C is a major cause of liver failure in the world, responsible of approximately 500,000 deaths annually [26]. Hepatitis C is a bloodborne disease, and the transmission remains high in people who inject drugs (PWID) due to injecting equipment sharing [23]. Until recently, the main approaches to decrease HCV transmission among PWID in high income countries relied on injection prevention and on risk reduction measures (access to sterile equipment, opioid substitution therapies, etc.). The arrival of highly effective antiviral treatments offers the opportunity to use the treatment as a mean to prevent HCV transmission, by treating infected PWID before they have transmitted the infection [9].

In this context, a stochastic, individual-based dynamic model was used to assess the impact of the treatment on HCV transmission in PWID in Paris area [7]. This model included HCV transmission on a random graph modelling PWID social network, the cascade of care of chronic hepatitis C and the progression of the liver disease. A brief description of the model for HCV infection and cascade of care is available in 3, for a detailed description and the details of the parameters values with their uncertainty intervals, the reader can refer to [7].

The parameter values used in this analysis were mainly provided by epidemiological studies and were subject to uncertainty. This kind of model requires high computing time, and thus the sensitivity analysis using Monte-Carlo estimators of Sobol indices is difficult, due to the number of simulations needed. We estimated Sobol indices using the Nadaraya-Watson non-parametric

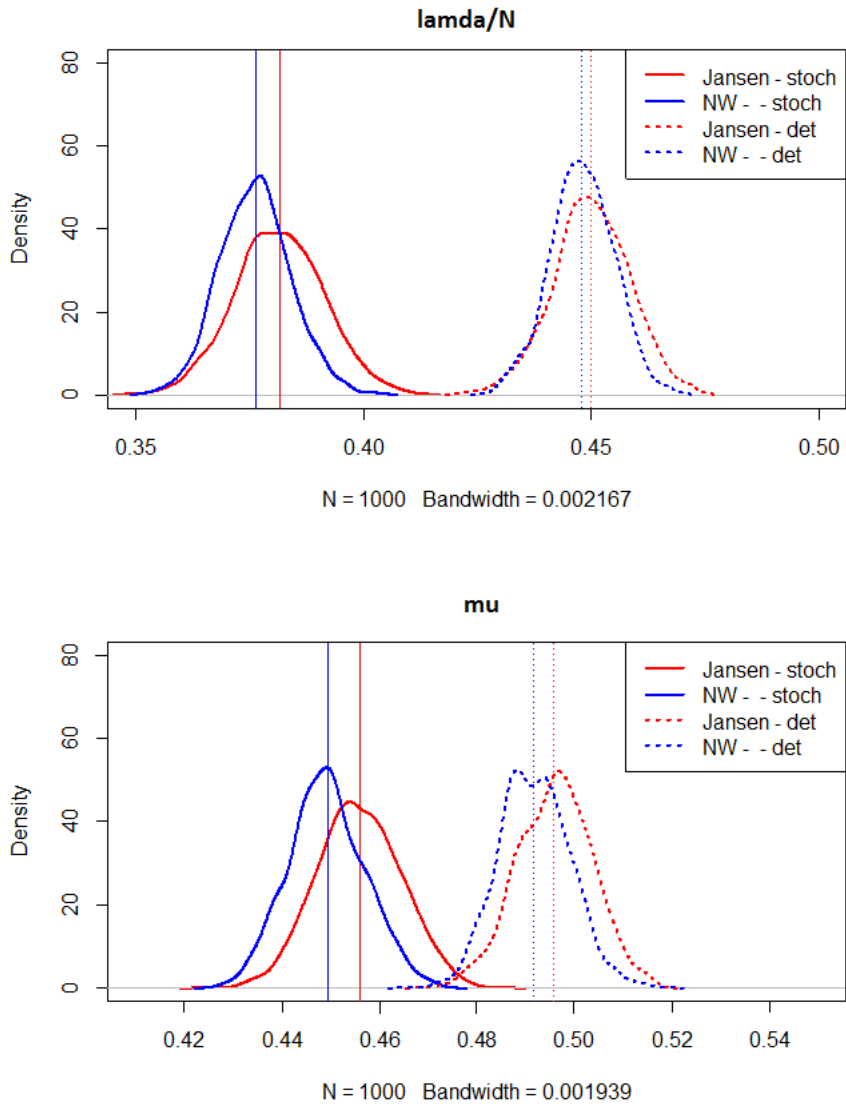


Figure 1: Smoothed densities of the first order sensitivity indices estimates for λ/N and μ in a deterministic (dotted lines) and in a stochastic (plain lines) SIR model, with $N = 1,200$ individuals.

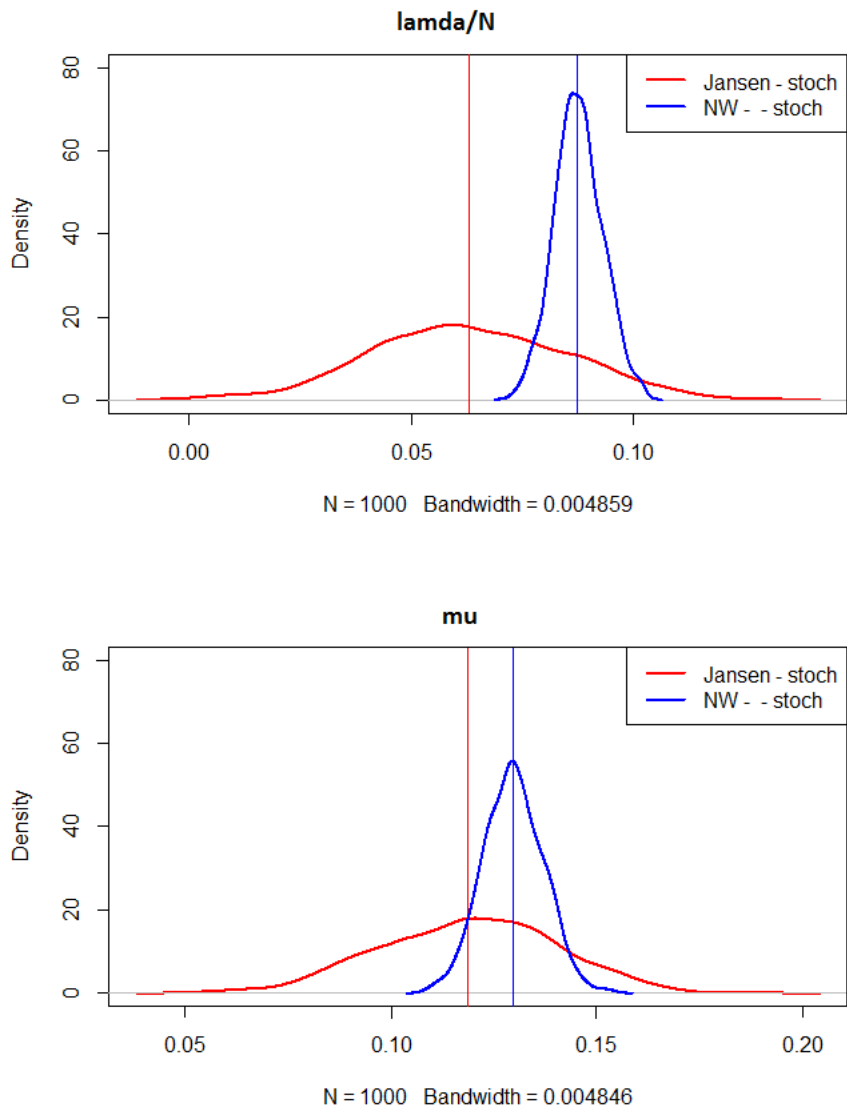


Figure 2: Smoothed densities of the first order sensitivity indices estimates for λ/N and μ in a stochastic (plain lines) SIR model, with $N = 120$ individuals.

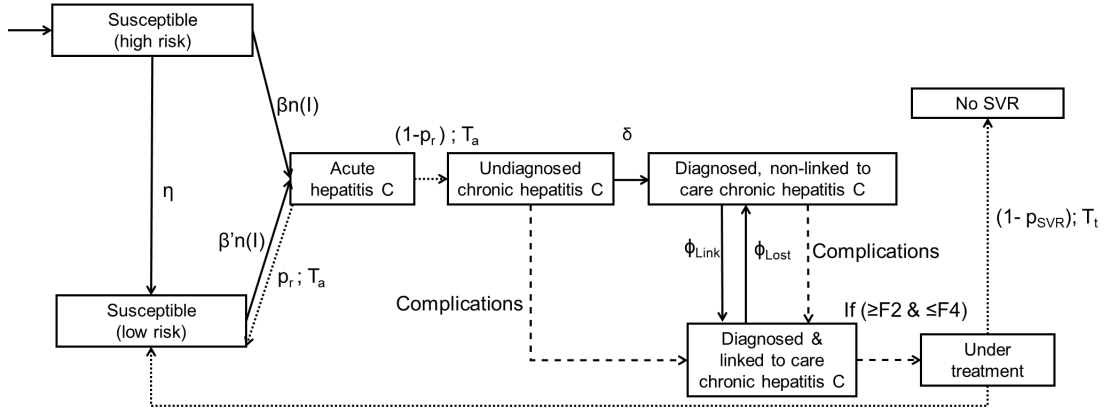


Figure 3: *Diagram flow of infection and cascade of care modelling for HCV infection among PWID. Greek letters refer to rates, p_r and p_{SVR} to probabilities and T_a and T_t to (deterministic) time before leaving the compartment. β depends on the status of the PWID with respect to the risk reduction measures (access to sterile injecting equipment, access to substitution therapies). n_i denotes the number of infected injecting partners of the PWID. δ depends on the status of the PWID with respect to injection: active or inactive injector (i.e. before or after the cessation of injection). The liver disease progression is quantified by a score (score Metavir for the fibrosis progression) between $F0$ and $F4$ (cirrhosis). “Complications” refers to the two cirrhosis complications: decompensated cirrhosis and hepatocellular carcinoma*

estimator, and with Y the prevalence after 10 years of simulation, and with uniform distributions on the uncertainty interval for each parameter. We used $n = 10,000$ simulations of the model. For comparison, we also represented the sensitivity using a Tornado diagram, classically used in epidemiology. A Tornado diagram is built using the extremal values of the uncertainty interval for each parameter. The model is simulated Y by changing the mean parameter value by the extremal values, one at a time. The other parameters values are let at their mean value (i.e. the value from the main analysis). Then, the parameters are sorted by decreasing width of the interval of Y values, and the deviation from the main analysis result is represented in a bar plot.

Results are presented Figure 4 With the Tornado diagram, the most sensitive parameters are the infection rate per infected injecting partner, the transition rate from a fibrosis score of $F0/F1$ to a score of $F2/F3$ and the combination of the linkage to care/loss to follow-up rate (which were varied together to estimate the impact of the uncertainty about the linkage to care of PWID). With the Sobol indices, we obtained consistent results. However, as the Sobol indices can be interpreted as the contribution of each parameter to this variance. We can thus see that a large part of the variance of Y is explained by the infection rate per infected partner alone, with a Sobol index of 0.59, and by the transition rate from a fibrosis score of $F0/F1$ to a score of $F2/F3$, with a Sobol index of 0.31. Other parameters contribute only marginally, and particularly linkage to care/loss to follow-up rate, which represent only 4% of the whole variance, according to these results. However, the sum of all the sensitivity indices estimated was 1.20, which is > 1 .

4 Proofs

4.1 Proof of Theorem 2.6

We follow the scheme of the proof of Theorem 1 in [15]. The main difficulty here is that we are not in a Gaussian framework and that we use the empirical process $\bar{\gamma}_n$, which introduces much

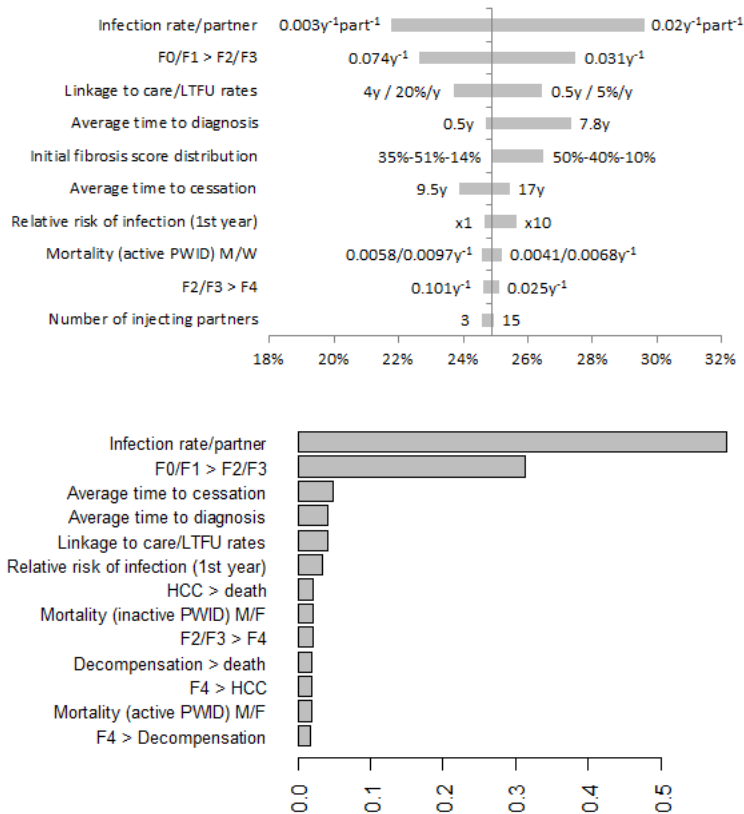


Figure 4: Tornado diagram (upper figure), and Sobol indices estimated using the Nadaraya-Watson estimator and sorted by decreasing value (lower figure). The values represented on the Tornado diagram are the extremal values of the uncertainty interval for each parameter. LTFU=loss to follow-up, HCC=Hepatocellular carcinoma, M=Male, F=Female. “Cessation” refers to the cessation of the injections. “F0/F1 > F2/F3” refers to the transition rate from a fibrosis score F0 or F1 to a fibrosis score F2 or F3 (and similarly for other rates).

technical difficulties.

In the sequel, C denotes a constant that can vary from line to line.

Using Lemma 2.3, we concentrate on the MSE $\mathbb{E}((\widehat{\theta}_\ell - V_\ell)^2)$. First, we will prove that:

$$\mathbb{E}\left[\left(\widehat{\theta}_\ell - V_\ell - \zeta_n\right)^2\right] \leq \inf_{\mathcal{J} \subset \{-1, \dots, J_n\}} \mathbb{E}\left[\left(-\widehat{\theta}_{\mathcal{J}, \ell} + V_\ell + \zeta_n\right)_+^2\right] + \frac{C}{n \log^2(n)}, \quad (4.1)$$

where $\widehat{\theta}_{\mathcal{J}, \ell}$ has been defined in (2.14). Then, considering the first term in the r.h.s. of (4.1), we prove:

$$\mathbb{E}\left[\left(-\widehat{\theta}_{\mathcal{J}, \ell} + V_\ell + \zeta_n\right)_+^2\right] \leq C\left(\|h_\ell - h_{\mathcal{J}, \ell}\|_2^4 + \frac{\log^2(n) + 2^{J_{\max}}}{n^2}\right) \quad (4.2)$$

Step 1:

From (2.14), and letting $A_{\mathcal{J}} = \widehat{\theta}_{\mathcal{J}, \ell} - V_\ell - \zeta_n$, we have:

$$\widehat{\theta}_\ell - V_\ell - \zeta_n = \sup_{\mathcal{J} \subset \{-1, \dots, J_n\}} A_{\mathcal{J}}.$$

Since

$$\left| \sup_{\mathcal{J}} A_{\mathcal{J}} \right| \leq \max \left[\sup_{\mathcal{J}} (A_{\mathcal{J}})_+, \inf_{\mathcal{J}} (A_{\mathcal{J}})_- \right],$$

we have that

$$\begin{aligned} \mathbb{E}\left(\sup_{\mathcal{J}} A_{\mathcal{J}}^2\right) &\leq \sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} \mathbb{E}\left((A_{\mathcal{J}})_+^2\right) + \inf_{\mathcal{J} \subset \{-1, \dots, J_n\}} \mathbb{E}\left((A_{\mathcal{J}})_-^2\right) \\ &\leq \sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} \mathbb{E}\left((A_{\mathcal{J}})_+^2\right) + \inf_{\mathcal{J} \subset \{-1, \dots, J_n\}} \mathbb{E}\left((V_\ell - \widehat{\theta}_{\mathcal{J}, \ell} + \zeta_n)_+^2\right). \end{aligned} \quad (4.3)$$

The second term correspond to what appears in (4.1) and will be treated in Step 4. Let us consider the first term of the r.h.s. We start by rewriting

$$\begin{aligned} A_{\mathcal{J}} &= \widehat{\theta}_{\mathcal{J}, \ell} - V_\ell - \zeta_n \\ &= \|\widehat{h}_{\mathcal{J}, \ell}\|_2^2 - \text{pen}(\mathcal{J}) - \|h_\ell\|_2^2 - \zeta_n \\ &= (\|\widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}\|_2^2 + \|h_{\mathcal{J}, \ell}\|_2^2 + 2\langle \widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}, h_{\mathcal{J}, \ell} \rangle) \\ &\quad - (\|h_\ell - h_{\mathcal{J}, \ell}\|_2^2 + \|h_{\mathcal{J}, \ell}\|_2^2 + 2\langle h_\ell - h_{\mathcal{J}, \ell}, h_{\mathcal{J}, \ell} \rangle) - \zeta_n - \text{pen}(\mathcal{J}) \\ &= \|\widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}\|_2^2 + 2\langle \widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}, h_{\mathcal{J}, \ell} \rangle - \|h_\ell - h_{\mathcal{J}, \ell}\|_2^2 - \zeta_n - \text{pen}(\mathcal{J}), \end{aligned} \quad (4.4)$$

since $\langle h_\ell - h_{\mathcal{J}, \ell}, h_{\mathcal{J}, \ell} \rangle = 0$ by definition of $h_{\mathcal{J}, \ell}$ as projection of h_ℓ on the subspace generated by $\{\psi_{jk}, j \in \mathcal{J}, k \in \mathbb{Z}\}$.

Thus:

$$\begin{aligned} \mathbb{E}\left((A_{\mathcal{J}})_+^2\right) &\leq 2\mathbb{E}\left(\left(\|\widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}\|_2^2 - \text{pen}_1(\mathcal{J})\right)^2\right) \\ &\quad + 2\mathbb{E}\left(\left(2\langle \widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}, h_{\mathcal{J}, \ell} \rangle - \|h_\ell - h_{\mathcal{J}, \ell}\|_2^2 - \zeta_n - \text{pen}_2(\mathcal{J})\right)^2\right). \end{aligned} \quad (4.5)$$

where

$$\text{pen}_1(\mathcal{J}) = \frac{K2^{J_{\max}}}{n}, \text{ and } \text{pen}_2(\mathcal{J}) = \frac{4M^2 x_{\mathcal{J}}}{n} + \frac{4M^2 \|\varphi_{\mathcal{J}}\|_\infty^2 x_{\mathcal{J}}^2}{n^2}. \quad (4.6)$$

Step 2: Upper bound of the first term in the r.h.s. of (4.5)

Reformulation of $\|\widehat{h}_{\mathcal{J},\ell} - h_{\mathcal{J},\ell}\|_2^2$

The first term in the r.h.s. of (4.4) is the approximation error of $h_{\mathcal{J}}$ by $\widehat{h}_{\mathcal{J},\ell}$ and equals

$$\|\widehat{h}_{\mathcal{J},\ell} - h_{\mathcal{J},\ell}\|_2^2 = \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} (\widehat{\beta}_{jk} - \beta_{jk})^2 = \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \bar{\gamma}_n(\psi_{jk})^2.$$

To control it, let us introduce, for coefficients $a = (a_{jk}, -1 \leq j \leq J_n, k \in \mathbb{Z})$, the set $\mathcal{F}_{1,\mathcal{J}} = \{\sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} a_{jk} \psi_{jk}, a_{jk} \in \mathbb{Q}, \|a\|_2 \leq 1\}$, which is countable and dense in the unit ball of $L^2([0, 1])$:

$$\begin{aligned} \left(\sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \bar{\gamma}_n(\psi_{jk})^2 \right)^{1/2} &= \sup_{\|a\|_2 \leq 1} \left| \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} a_{jk} \bar{\gamma}_n(\psi_{jk}) \right| \\ &= \sup_{\|a\|_2 \leq 1} \left| \bar{\gamma}_n \left(\sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} a_{jk} \psi_{jk} \right) \right| \\ &= \sup_{f \in \mathcal{F}_{1,\mathcal{J}}} |\bar{\gamma}_n(f)| := \chi_n(\mathcal{J}). \end{aligned} \quad (4.7)$$

Let us introduce, for $\rho > 0$,

$$\Omega_{\mathcal{J}}(\rho) = \left\{ \forall j \in \mathcal{J}, \sum_{k \in \mathbb{Z}} |\bar{\gamma}_n(\psi_{jk})| \leq \rho 2^{-j/2} \right\}. \quad (4.8)$$

Then, to upper bound the first term in (4.5), we can write:

$$\mathbb{E} \left(\left(\|\widehat{h}_{\mathcal{J},\ell} - h_{\mathcal{J},\ell}\|_2^2 - \text{pen}_1(\mathcal{J}) \right)^2 \right) \leq 2A_1(\mathcal{J}) + 2A_2(\mathcal{J}) \quad (4.9)$$

where

$$A_1(\mathcal{J}) = \mathbb{E} \left(\left(\chi_n^2(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}(\rho)} - \text{pen}_1(\mathcal{J}) \right)^2 \right), \quad \text{and} \quad A_2(\mathcal{J}) = \mathbb{E} \left(\chi_n^4(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}^c(\rho)} \right). \quad (4.10)$$

The upper bounds of $A_1(\mathcal{J})$ and $A_2(\mathcal{J})$ make the object of the remainder of Step 2. We use ideas developed in [3]. To upper bound $A_1(\mathcal{J})$, we use the identity

$$A_1(\mathcal{J}) = \int_0^{+\infty} t \mathbb{P}(\chi_n^2(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}(\rho)} - \text{pen}_1(\mathcal{J}) > t) dt, \quad (4.11)$$

and look for deviation inequalities of $\chi_n^2(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}(\rho)}$. Then, estimates of the probability of $\Omega_{\mathcal{J}}^c(\rho)$ are studied to control $A_2(\mathcal{J})$.

Deviation inequality for $\sup_{a \in \Lambda_{\mathcal{J}}} |\bar{\gamma}_n(f)|$

The supremum in (4.7) is obtained for

$$\bar{a}_{jk} = \frac{\bar{\gamma}_n(\psi_{jk})}{\chi_n(\mathcal{J})}. \quad (4.12)$$

On the set $\Omega_{\mathcal{J}}(\rho) \cap \{\chi_n(\mathcal{J}) > z\}$, for a constant $z > 0$ that shall be fixed in the sequel, we have for all $j \in \mathcal{J}$,

$$\sum_{k \in \mathbb{Z}} |\bar{a}_{jk}| = \frac{\sum_{k \in \mathbb{Z}} |\bar{\gamma}_n(\psi_{jk})|}{\chi_n(\mathcal{J})} \leq \frac{\rho 2^{-j/2}}{z}.$$

As a consequence, on the set $\Omega_{\mathcal{J}}(\rho) \cap \{\chi_n(\mathcal{J}) > z\}$, we can restrict the research of the optima to the set

$$\Lambda_{\mathcal{J}} = \left\{ a = (a_{jk})_{j \geq -1, k \in \mathbb{Z}} \in \mathbb{Q}^{\{-1, \dots\} \times \mathbb{Z}}, a_{jk} = 0 \text{ if } j \notin \mathcal{J}, \sum_{k \in \mathbb{Z}} |a_{jk}| \leq \frac{\rho 2^{-j/2}}{z} \text{ if } j \in \mathcal{J} \right\},$$

which is countable.

We can then use Talagrand inequality (see [18, p.170]) to obtain that for all $\eta > 0$ and $x > 0$,

$$\mathbb{P} \left(\sup_{a \in \Lambda_{\mathcal{J}}} |\bar{\gamma}_n(f)| \geq (1 + \eta) \mathbb{E} \left(\sup_{a \in \Lambda_{\mathcal{J}}} |\bar{\gamma}_n(f)| \right) + \sqrt{2\nu_n x} + \left(\frac{1}{3} + \frac{1}{\eta} \right) b_n x \right) \leq e^{-x}, \quad (4.13)$$

where $\mathbb{E} \left(\sup_{a \in \Lambda_{\mathcal{J}}} |\bar{\gamma}_n(f)| \right)$ and where ν_n and b_n can be chosen respectively as $\nu_n = M^2/n$ and $b_n = 2M \|\psi\|_{\infty} \rho \text{Card}(\mathcal{J})/nz$. Indeed, ν_n is an upper bound of:

$$\frac{1}{n} \sup_{a \in \Lambda_{\mathcal{J}}} \text{Var} \left(Y_1 \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} a_{jk} \psi_{jk}(G_{\ell}(X_{\ell}^1)) \right) \leq \frac{M^2}{n} \sup_{a \in \Lambda_{\mathcal{J}}} \left\| \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} a_{jk} \psi_{jk} \right\|_2^2 \leq \frac{M^2}{n}, \quad (4.14)$$

from the definition of $\Lambda_{\mathcal{J}}$. As for the term b_n , it can be obtained from:

$$\begin{aligned} & \frac{1}{n} \sup_{a \in \Lambda_{\mathcal{J}}} \sup_{(u, y) \in [0, 1] \times \mathbb{R}} \left| y \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} a_{jk} \psi_{jk}(u) - \mathbb{E} \left(Y_1 \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} a_{jk} \psi_{jk}(G_{\ell}(X_{\ell}^1)) \right) \right| \\ & \leq \frac{2M}{n} \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} |a_{jk}| 2^{j/2} \|\psi\|_{\infty} \leq \frac{2M \|\psi\|_{\infty}}{n} \sum_{j \in \mathcal{J}} \frac{\rho 2^{-j/2}}{z} 2^{j/2} = \frac{2M \|\psi\|_{\infty} \rho \text{Card}(\mathcal{J})}{n z}. \end{aligned} \quad (4.15)$$

For the expectation in the r.h.s. in the probability, we have:

$$\begin{aligned} \mathbb{E} \left(\sup_{a \in \Lambda_{\mathcal{J}}} |\bar{\gamma}_n(f)| \right) & \leq \mathbb{E}(\chi_n(\mathcal{J})) \leq \sqrt{\mathbb{E}(\chi_n^2(\mathcal{J}))} = \sqrt{\sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \mathbb{E}(\bar{\gamma}_n^2(\psi_{jk}))} \\ & = \sqrt{\sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \frac{1}{n} \text{Var}(Y_1 \psi_{jk}(G_{\ell}(X_{\ell}^1)))} \leq M \sqrt{\frac{2^{J_{\max}}}{n}} \end{aligned} \quad (4.16)$$

by using the Cauchy-Schwarz inequality and the fact that $\|\psi_{jk}\|_2^2 = 1$.

Because $\sup_{a \in \Lambda_{\mathcal{J}}} |\bar{\gamma}_n(f)| \geq \chi_n(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}(\rho) \cap \{\chi_n(\mathcal{J}) > z\}}$, Equations (4.13)-(4.16) become:

$$\mathbb{P} \left(\chi_n(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}(\rho) \cap \{\chi_n(\mathcal{J}) > z\}} \geq (1 + \eta) M \sqrt{\frac{2^{J_{\max}}}{n}} + \sqrt{\frac{2M^2 x}{n}} + \left(\frac{1}{3} + \frac{1}{\eta} \right) \frac{2M \|\psi\|_{\infty} \rho \text{Card}(\mathcal{J})}{n z} x \right) \leq e^{-x}.$$

Choosing $z = \sqrt{\frac{2x}{n}} \left(\frac{1}{3} + \frac{1}{\eta} \right) \|\psi\|_{\infty}$, we obtain:

$$\mathbb{P} \left(\chi_n(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}(\rho) \cap \{\chi_n(\mathcal{J}) > z\}} \geq (1 + \eta) M \sqrt{\frac{2^{J_{\max}}}{n}} + (1 + \rho) M \text{Card}(\mathcal{J}) \sqrt{\frac{2x}{n}} \right) \leq e^{-x}.$$

Choosing $\rho = \left(\frac{1}{3} + \frac{1}{\eta} \right) \|\psi\|_{\infty}$, we can get rid of the constraint $\{\chi_n(\mathcal{J}) > z\}$ to evaluate the above probability and choosing $x = x_{\mathcal{J}} + \xi$:

$$\begin{aligned} \mathbb{P} \left(\chi_n^2(\mathcal{J}) \mathbb{1}_{\Omega_{\mathcal{J}}(\rho)} - \frac{1}{n} \left[(1 + \eta)^2 M^2 2^{J_{\max}} + 2(1 + \rho)^2 \text{Card}^2(\mathcal{J}) x_{\mathcal{J}} \right. \right. \\ \left. \left. + 2(1 + \rho)(1 + \eta) M^2 2^{\frac{J_{\max} + 1}{2}} \text{Card}(\mathcal{J}) \sqrt{x_{\mathcal{J}}} \right] \geq h_{\mathcal{J}}(\xi) \right) \\ \leq e^{-x_{\mathcal{J}}} e^{-\xi}, \end{aligned}$$

where

$$h_{\mathcal{J}}(\xi) = \frac{2(1+\rho)M^2 \text{Card}(\mathcal{J})}{n} \left[(1+\rho)\text{Card}(\mathcal{J})\xi + (1+\eta)2^{\frac{J_{\max}+1}{2}} \sqrt{\xi} \right]. \quad (4.17)$$

The square bracket in the l.h.s. inside the probability can be upper bounded by $n\text{pen}_1(\mathcal{J}) = K2^{J_{\max}}$, for a certain constant K that depends on $x_{\mathcal{J}}$, since $\text{Card}(\mathcal{J}) \leq J_{\max}$ and since $x^2 \leq 2^x$ for all integers $x \geq 1$. Then:

$$\mathbb{P}\left(\chi_n^2(\mathcal{J})\mathbb{1}_{\Omega_{\mathcal{J}}(\rho)} - \text{pen}_1(\mathcal{J}) \geq h_{\mathcal{J}}(\xi)\right) \leq e^{-x_{\mathcal{J}}} e^{-\xi}. \quad (4.18)$$

From this and (4.11),

$$A_1(\mathcal{J}) \leq \int_0^{+\infty} t e^{-x} e^{-h_{\mathcal{J}}^{-1}(t)} dt.$$

To upper bound the r.h.s., we have to lower bound $h_{\mathcal{J}}^{-1}(t)$ and hence upper bound $h_{\mathcal{J}}(t)$. The square bracket in (4.17) can be upper bounded by

$$\begin{cases} 2\sqrt{2}(1+\eta)2^{\frac{J_{\max}}{2}}\sqrt{\xi} & \text{if } \xi \leq 2\left(\frac{1+\eta}{1+\rho}\right)^2 \frac{2^{J_{\max}}}{\text{Card}^2 \mathcal{J}} \\ 2(1+\rho)\text{Card} \mathcal{J} \xi & \text{if } \xi > 2\left(\frac{1+\eta}{1+\rho}\right)^2 \frac{2^{J_{\max}}}{\text{Card}^2 \mathcal{J}}. \end{cases}$$

Then, for $t \geq 0$:

$$h_{\mathcal{J}}^{-1}(t) \geq \begin{cases} \frac{n^2 t^2}{32(1+\rho)^2 M^2 \text{Card}^2 \mathcal{J} (1+\eta)^2 2^{J_{\max}}} & \text{if } t \leq \frac{8M^2(1+\eta)^2 2^{J_{\max}}}{n} \\ \frac{nt}{4(1+\rho)^2 M^2 \text{Card}^2 \mathcal{J}} & \text{if } t > \frac{8M^2(1+\eta)^2 2^{J_{\max}}}{n}. \end{cases}$$

As a consequence,

$$\begin{aligned} A_1(\mathcal{J}) &\leq \int_0^{\frac{8M^2(1+\eta)^2 2^{J_{\max}}}{n}} t e^{-x_{\mathcal{J}}} \exp\left(-\frac{n^2 t^2}{32(1+\rho)^2 M^2 \text{Card}^2 \mathcal{J} (1+\eta)^2 2^{J_{\max}}}\right) dt \\ &\quad + \int_{\frac{8M^2(1+\eta)^2 2^{J_{\max}}}{n}}^{+\infty} t e^{-x_{\mathcal{J}}} \exp\left(-\frac{nt}{4(1+\rho)^2 M^2 \text{Card}^2 \mathcal{J}}\right) dt \\ &\leq e^{-x_{\mathcal{J}}} \frac{32(1+\rho)^2 M^2 \text{Card}^2 \mathcal{J} (1+\eta)^2 2^{J_{\max}}}{2n^2} \left[1 - \exp\left(-\frac{2M^2(1+\eta)^2 2^{J_{\max}}}{(1+\rho)^2 \text{Card}^2 \mathcal{J}}\right)\right] \\ &\quad + e^{-x_{\mathcal{J}}} \frac{16(1+\rho)^2 M^4 \text{Card}^2 \mathcal{J} \left(2(1+\eta)^2 2^{J_{\max}} + (1+\rho)^2 \text{Card}^2 \mathcal{J}\right)}{n^2} \exp\left(-\frac{2(1+\eta)^2 2^{J_{\max}}}{(1+\rho)^2 \text{Card}^2 \mathcal{J}}\right) \\ &\leq \frac{C 2^{2J_{\max}}}{n^2} e^{-x_{\mathcal{J}}}. \end{aligned} \quad (4.19)$$

The latter upper bound determines the choice of J_{\max} . From the choice of $x_{\mathcal{J}}$ (2.8), we deduce that

$$\sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} A_1(\mathcal{J}) \leq \frac{C 2^{2J_n}}{n^2} = \frac{C}{n \log^2(n)}. \quad (4.20)$$

Upper bound of $A_2(\mathcal{J})$

For the term $A_2(\mathcal{J})$ of (4.9), noting that:

$$|\bar{\gamma}_n(\psi_{jk})| \leq M 2^{j/2} \|\psi\|_{\infty} + M 2^{-j/2} \int_{\mathbb{R}} \psi(u) du,$$

we have for a constant C that depends only on the choice of ψ_{-10} and ψ_{00} :

$$A_2(\mathcal{J}) \leq \left[C \sum_{j \in \mathcal{J}} \left(M 2^{j/2} \|\psi\|_\infty + M 2^{-j/2} \int_{\mathbb{R}} \psi(u) du \right)^2 \right]^2 \times \mathbb{P}\left(\Omega_{\mathcal{J}}^c(\eta)\right). \quad (4.21)$$

Since:

$$\begin{aligned} \sum_{i=1}^n \mathbb{E} \left[\left(\frac{Y_i \psi_{jk}(G_\ell(X_\ell^i)) - \mathbb{E}(Y_1 \psi_{jk}(G_\ell(X_\ell^1)))}{n} \right)^2 \right] &= \frac{\text{Var}(Y_1 \psi_{jk}(G_\ell(X_\ell^1)))}{n} \leq \frac{M^2}{n}, \\ \left| \frac{Y_i \psi_{jk}(G_\ell(X_\ell^i)) - \mathbb{E}(Y_1 \psi_{jk}(G_\ell(X_\ell^1)))}{n} \right| &\leq \frac{2M 2^{j/2} \|\psi\|_\infty}{n} \text{ a.s.} \end{aligned}$$

then we have by Bernstein's inequality (e.g. [18]):

$$\mathbb{P}\left(|\bar{\gamma}_n(\psi_{jk})| \geq \rho 2^{-j/2}\right) \leq 2 \exp\left(-\frac{n \rho^2 2^{-j}}{2(M^2 + 2M \|\psi\|_\infty \rho)}\right).$$

As a consequence,

$$\begin{aligned} \sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} A_2(\mathcal{J}) &\leq \sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} 2^{2J_{\max}} \mathbb{P}\left(\exists(j, k) \in \mathcal{J} \times \mathbb{Z}, |\bar{\gamma}_n(\psi_{jk})| \geq \rho 2^{-j/2}\right) \\ &\leq C \sum_{\mathcal{J} \subset \{-1, \dots, J_n\}} 2^{3J_{\max}} \exp\left(-\frac{n \rho^2 2^{-J_{\max}}}{2(M^2 + 2M \|\psi\|_\infty \rho)}\right). \end{aligned} \quad (4.22)$$

which is smaller than C/n^2 for sufficiently large n , as $J_{\max} \leq J_n = \log_2(\sqrt{n})$.

Step 3: Upper bound of the second term in the r.h.s. of (4.5)

For the terms 2 to 4 of (4.4),

$$\begin{aligned} &2\langle \hat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}, h_{\mathcal{J}, \ell} \rangle - \|h_\ell - h_{\mathcal{J}, \ell}\|_2^2 - \zeta_n - \text{pen}_2(\mathcal{J}) \\ &= 2 \sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \bar{\gamma}_n(\psi_{jk}) \beta_{jk}^\ell - \|h_\ell - h_{\mathcal{J}, \ell}\|_2^2 - 2\bar{\gamma}_n(h_\ell) - \text{pen}_2(\mathcal{J}) \\ &= 2\bar{\gamma}_n\left(\sum_{j \in \mathcal{J}} \sum_{k \in \mathbb{Z}} \beta_{jk}^\ell \psi_{jk}\right) - \|h_\ell - h_{\mathcal{J}, \ell}\|_2^2 - 2\bar{\gamma}_n(h_\ell) - \text{pen}_2(\mathcal{J}) \\ &= 2\bar{\gamma}_n(h_{\mathcal{J}, \ell} - h_\ell) - \|h_\ell - h_{\mathcal{J}, \ell}\|_2^2 - \text{pen}_2(\mathcal{J}) \end{aligned} \quad (4.23)$$

$$\leq \left(\frac{\bar{\gamma}_n(h_{\mathcal{J}, \ell} - h_\ell)}{\|h_\ell - h_{\mathcal{J}, \ell}\|_2}\right)^2 - \text{pen}_2(\mathcal{J}) = \bar{\gamma}_n^2\left(\frac{h_{\mathcal{J}, \ell} - h_\ell}{\|h_\ell - h_{\mathcal{J}, \ell}\|_2}\right) - \text{pen}_2(\mathcal{J}), \quad (4.24)$$

by using the identity $2ab - b^2 \leq a^2$. Setting $\varphi_{\mathcal{J}} = \frac{h_{\mathcal{J}, \ell} - h_\ell}{\|h_\ell - h_{\mathcal{J}, \ell}\|_2}$ and using Bernstein's formula (see [18, p.25]), we have for all $x > 0$:

$$\mathbb{P}\left(\bar{\gamma}_n(\varphi_{\mathcal{J}}) \geq \sqrt{\frac{2M^2}{n}}x + \frac{2M \|\varphi_{\mathcal{J}}\|_\infty}{n}x\right) \leq e^{-x}. \quad (4.25)$$

Setting $x_{\mathcal{J}} + \xi$ as x in the above inequality and using that $(a+b)^2 \leq 2a^2 + 2b^2$, this implies that

$$\mathbb{P}\left(\bar{\gamma}_n^2(\varphi_{\mathcal{J}}) - \text{pen}_2(\mathcal{J}) \geq r_n(\xi)\right) \leq e^{-x_{\mathcal{J}}} e^{-\xi}, \quad (4.26)$$

where $\text{pen}_2(\mathcal{J})$ has been defined in (4.6) and

$$r_n(x, \xi) = \frac{4M^2 \|\varphi_{\mathcal{J}}\|_{\infty}^2 \xi^2}{n^2} + \frac{4M^2 \xi}{n}.$$

Then,

$$\begin{aligned} & \mathbb{E} \left(\left(2 \langle \widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}, h_{\mathcal{J}, \ell} \rangle - \|h_{\ell} - h_{\mathcal{J}, \ell}\|_2^2 - \zeta_n - \text{pen}_2(\mathcal{J}) \right)_+^2 \right) \\ & \leq \mathbb{E} \left(\left[\bar{\gamma}_n^2(\varphi_{\mathcal{J}}) - \text{pen}_2(\mathcal{J}) \right]^2 \right) \\ & \leq C \int_0^{+\infty} t \mathbb{P} \left(|\bar{\gamma}_n^2(\varphi_{\mathcal{J}}) - \text{pen}_2(\mathcal{J})| > t \right) dt \\ & \leq C e^{-x_{\mathcal{J}}} \int_0^{+\infty} t \exp \left(- \frac{n}{2 \|\varphi_{\mathcal{J}}\|_{\infty}^2} \left(\sqrt{1 + \frac{t \|\varphi_{\mathcal{J}}\|_{\infty}^2}{M}} - 1 \right) \right) dt \leq \frac{C e^{-x_{\mathcal{J}}}}{n^2}. \end{aligned} \quad (4.27)$$

The last inequality stems from the behaviour of the integrand when t is close to 0.

Gathering the results of Steps 1 to 3, we have by (4.9) and (4.5) that the first term in the r.h.s. of (4.3) is smaller than $C/(n \log^2(n))$. This proves (4.1).

Step 4:

Let us now consider the term $\mathbb{E} \left[\left(-\widehat{\theta}_{\mathcal{J}, \ell} + V_{\ell} + \zeta_n \right)_+^2 \right]$ in (4.1). From (4.4) and (4.23):

$$\begin{aligned} & \mathbb{E} \left[\left(-\widehat{\theta}_{\mathcal{J}, \ell} + V_{\ell} + \zeta_n \right)_+^2 \right] \\ & = \mathbb{E} \left(\left(\|h_{\ell} - h_{\mathcal{J}, \ell}\|_2^2 - \|\widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}\|_2^2 + 2\bar{\gamma}_n(h_{\ell} - h_{\mathcal{J}, \ell}) + \text{pen}(\mathcal{J}) \right)_+^2 \right) \\ & \leq 4 \left(\|h_{\ell} - h_{\mathcal{J}, \ell}\|_2^4 + 4\mathbb{E} \left(\bar{\gamma}_n^2(h_{\ell} - h_{\mathcal{J}, \ell}) \right) + \mathbb{E} \left(\left[\|\widehat{h}_{\mathcal{J}, \ell} - h_{\mathcal{J}, \ell}\|_2^2 - \text{pen}_1(\mathcal{J}) \right]_+^2 \right) + \text{pen}_2^2(\mathcal{J}) \right), \end{aligned} \quad (4.28)$$

where $D_{\mathcal{J}}$ has been defined in (4.2).

For the second term in the r.h.s. of (4.28), we have:

$$\begin{aligned} \mathbb{E} \left(\bar{\gamma}_n^2(h_{\ell} - h_{\mathcal{J}, \ell}) \right) & = \text{Var} \left(\bar{\gamma}(h_{\ell} - h_{\mathcal{J}, \ell}) \right) \\ & \leq \frac{1}{n} \mathbb{E} \left(Y_1^2(h_{\ell}(G_{\ell}(X_{\ell}^1)) - h_{\mathcal{J}, \ell}(G_{\ell}(X_{\ell}^1)))^2 \right) \leq \frac{M^2 \|h_{\ell} - h_{\mathcal{J}, \ell}\|_2^2}{n} \\ & \leq C \left(\frac{1}{n^2} + \|h_{\ell} - h_{\mathcal{J}, \ell}\|_2^4 \right) \end{aligned} \quad (4.29)$$

by using that $2ab \leq a^2 + b^2$ for the last inequality.

The third term in the r.h.s. of (4.28) has been treated in (4.9) precedingly. We established an upper bound in $2^{\mathcal{J}_{\max}}/n^2$. The fourth term, $\text{pen}_2^2(\mathcal{J})$ is in $x_{\mathcal{J}}^2/n^2 \leq C \log^2(n)/n^2$ from (4.6). Gathering these results, we obtain (4.2) and then (2.26).

4.2 Proof of Corollary 2.8

Plugging (4.2) in (4.1), and using that

$$\mathbb{E}(\zeta_n^2) = \frac{2}{n} \text{Var} \left(Y_1 h_{\ell}(G_{\ell}(X_{\ell}^1)) \right) \leq \frac{2M^2 \|h_{\ell}\|_2^2}{n}, \quad (4.30)$$

we obtain:

$$\mathbb{E} \left[(\widehat{\theta}_\ell - V_\ell)^2 \right] \leq C \left[\inf_{\mathcal{J} \subset \{-1, \dots, J_n\}} \left(\|h_\ell - h_{\mathcal{J}, \ell}\|_2^4 + \frac{2^{J_{\max}}}{n^2} \right) + \frac{1 + \|h_\ell\|_2^2}{n} \right]. \quad (4.31)$$

If $h_\ell \in \mathcal{B}(\alpha, 2, \infty)$, then from Proposition 2.7, we have for $\mathcal{J} = \{-1, \dots, J_{\max}\}$ that $\|h_\ell - h_{\mathcal{J}, \ell}\|_2^4 \leq 2^{-4\alpha} J_{\max}$. Thus, for subsets \mathcal{J} of the form considered, the infimum is attained when choosing $J_{\max} = \frac{2}{4\alpha+1} \log_2(n)$, which yield an upper bound in $n^{8\alpha/(4\alpha+1)}$.

For h_ℓ in a ball of radius R , $\|h_\ell\|_2^2 \leq R^2$, and we can find an upper bound that does not depend on h . Because the last term in (4.31) is in $1/n$, the elbow effect is obtained by comparing the order of the first term in the r.h.s. ($n^{8\alpha/(4\alpha+1)}$) with $1/n$ when α varies. \square

A Properties of $\widehat{\beta}_{jk}$

Lemma A.1.

$$\lim_{n \rightarrow +\infty} \sqrt{n} (\widehat{\beta}_{jk}^\ell - \beta_{jk}^\ell) = \mathcal{N} \left(0, \text{Var}(Y \psi_{jk}(G_\ell(X_\ell))) \right) \quad (\text{A.1})$$

where

$$\text{Var}(Y \psi_{jk}(G_\ell(X_\ell))) = \text{Var}(g_\ell(X_\ell) \psi_{jk}(G_\ell(X_\ell))) + \mathbb{E}(\eta_\ell^2 \psi_{jk}^2(G_\ell(X_\ell))).$$

Proof. Recall that $Y^i = g_\ell(X_\ell^i) + \eta_\ell^i$. Then:

$$\begin{aligned} \widehat{\beta}_{jk}^\ell &= \frac{1}{n} \sum_{i=1}^n g_\ell(X_\ell^i) \psi_{jk}(G_\ell(X_\ell^i)) + \frac{1}{n} \sum_{i=1}^n \eta_\ell^i \psi_{jk}(G_\ell(X_\ell^i)) \\ &= \beta_{jk}^\ell + \frac{1}{n} \sum_{i=1}^n \left(g_\ell(X_\ell^i) \psi_{jk}(G_\ell(X_\ell^i)) - \mathbb{E}(g_\ell(X_\ell) \psi_{jk}(G_\ell(X_\ell))) \right) + \frac{1}{n} \sum_{i=1}^n \eta_\ell^i \psi_{jk}(G_\ell(X_\ell^i)). \end{aligned}$$

The second term in the r.h.s. is a bias term due to the approximation of β_{jk}^ℓ , defined as an integral, by a mean. The third term is due to the noise between Y^i and $g_\ell(X_\ell^i)$. The third term is centered as $\mathbb{E}(\eta_\ell | X_\ell) = 0$. Since the observations are i.i.d., we have by the central limit theorem that

$$\lim_{n \rightarrow +\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} g_\ell(X_\ell^i) \psi_{jk}(G_\ell(X_\ell^i)) - \mathbb{E}(g_\ell(X_\ell) \psi_{jk}(G_\ell(X_\ell))) \\ \frac{1}{n} \sum_{i=1}^n \eta_\ell^i \psi_{jk}(G_\ell(X_\ell^i)) \end{pmatrix} = \mathcal{N}(0, \Sigma) \quad (\text{A.2})$$

with

$$\Sigma = \begin{pmatrix} \text{Var}(g_\ell(X_\ell) \psi_{jk}(G_\ell(X_\ell))) & 0 \\ 0 & \text{Var}(\eta_\ell \psi_{jk}(G_\ell(X_\ell))) \end{pmatrix}.$$

We have

$$\begin{aligned} \text{Var}(g_\ell(X_\ell) \psi_{jk}(G_\ell(X_\ell))) &= \mathbb{E}(g_\ell^2(X_\ell) \psi_{jk}^2(G_\ell(X_\ell))) - \beta_{jk}^2 \\ &= \int_0^1 \left(\sum_{jk} \beta_{jk} \psi_{jk}(u) \right)^2 \psi_{jk}^2(u) du - \beta_{jk}^2, \end{aligned}$$

and as $\mathbb{E}(\eta_\ell | X_\ell) = 0$, we have

$$\text{Var}(\eta_\ell \psi_{jk}(G_\ell(X_\ell))) = \mathbb{E}(\eta_\ell^2 \psi_{jk}^2(G_\ell(X_\ell))) = \int_{\mathbb{R}} \mathbb{E}(\eta_\ell^2 | X_\ell = G_\ell^{-1}(2^{-j}(v+k))) \psi^2(v) dv.$$

Using the Slutsky lemma concludes the proof. \blacksquare

B Sobol indices

The Sobol indices are based on the following decomposition for f (see Sobol [21]). We recall the formulas here, with the notation X_{p+1} for the random variable ε :

$$Y = f(X_1, \dots, X_p, \varepsilon) = f_0 + \sum_{\ell=1}^{p+1} f_\ell(X_\ell) + \sum_{1 \leq \ell_1 < \ell_2 \leq p+1} f_{\ell_1 \ell_2}(X_{\ell_1}, X_{\ell_2}) + \dots + f_{1, \dots, p+1}(X_1, \dots, X_p, \varepsilon) \quad (\text{B.1})$$

where $f_0 = E[Y]$, $f_\ell(X_\ell) = E[Y|X_\ell] - E[Y]$,
 $f_{\ell_1 \ell_2}(X_{\ell_1}, X_{\ell_2}) = E[Y|X_{\ell_1}, X_{\ell_2}] - E[Y|X_{\ell_1}] - E[Y|X_{\ell_2}] - E[Y]$, ...

Then, the variance of Y can be written as:

$$\text{Var}(Y) = \sum_{\ell=1}^{p+1} V_\ell + \sum_{1 \leq \ell_1 < \ell_2 \leq p+1} V_{\ell_1 \ell_2} + \dots + V_{1 \dots p+1} \quad (\text{B.2})$$

where

$$V_\ell = \text{Var}(E[Y|X_\ell]), \quad V_{\ell_1 \ell_2} = \text{Var}(E[Y|X_{\ell_1}, X_{\ell_2}]) - V_{\ell_1} - V_{\ell_2}, \dots$$

$$V_{1 \dots p+1} = \text{Var}(Y) - \sum_{\ell=1}^{p+1} V_\ell - \sum_{1 \leq \ell_1 < \ell_2 \leq p+1} V_{\ell_1 \ell_2} - \dots - \sum_{1 \leq \ell_1 < \dots < \ell_p \leq p+1} V_{\ell_1 \dots \ell_p} \quad (\text{B.3})$$

The first order indices are then defined as:

$$S_\ell = V_\ell / \text{Var}(Y) = \text{Var}(E[Y|X_\ell]) / \text{Var}(Y) \quad (\text{B.4})$$

S_ℓ corresponds to the part of the variance that can be explained by the variance of Y due to the variable X_ℓ alone. In the same manner, we define the second order indices, third order indices, etc. by dividing the variance terms by $\text{Var}(Y)$.

References

- [1] H. Anderson and T. Britton. *Stochastic Epidemic models and Their Statistical Analysis*, volume 151 of *Lecture Notes in Statistics*. Springer, New York, 2000.
- [2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.
- [3] G. Castellán. *Sélection d’histogrammes ou de modèles exponentiels de polynômes par morceaux à l’aide d’un critère de type Akaike*. PhD thesis, Université d’Orsay, 2000.
- [4] G. Chagny. Penalization versus Goldenshluger-lepski strategies in warped bases regression. *ESAIM: P&S*, 17:328–358, 2013.
- [5] A. Cousien, V.C. Tran, S. Deuffic-Burban, M. Jauffret-Roustide, J.S. Dhersin, and Y. Yazdanpanah. Hepatitis C treatment as prevention of viral transmission and level-related morbidity in persons who inject drugs. *Hepatology*, 2015.
- [6] A. Cousien, V.C. Tran, S. Deuffic-Burban, M. Jauffret-Roustide, G. Mabileau, J.S. Dhersin, and Y. Yazdanpanah. Cost-effectiveness of risk reduction interventions and improvements in the cascade of care of chronic hepatitis C in people who inject drugs. 2015. Manuscript in preparation.
- [7] Anthony Cousien, Viet Chi Tran, Sylvie Deuffic-Burban, Marie Jauffret-Roustide, Jean-Stéphane Dhersin, and Yazdan Yazdanpanah. Hepatitis C treatment as prevention of viral transmission and liver-related morbidity in persons who inject drugs. *Hepatology*, 2015.
- [8] J.-C. Fort, T. Klein, A. Lagnoux, and B. Laurent. Estimation of the Sobol indices in a linear functional multidimensional model. *Journal of Statistical Planning and Inference*, 143(9):1590–1605, 2013.

- [9] Jason Grebely, Gail V Matthews, Andrew R Lloyd, and Gregory J Dore. Elimination of hepatitis c virus infection among people who inject drugs through treatment as prevention: feasibility and future requirements. *Clinical infectious diseases*, 57(7):1014–1020, 2013.
- [10] W. Härdle, G. Kerkycharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications*, volume 129 of *Lecture Notes in Statistics*. Springer, New York, 1987.
- [11] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In *Proceedings of the ISUMA'90, First International Symposium on Uncertainty Modelling and Analysis*, University of Maryland, USA, pages 398–403, December 3-5 1990.
- [12] A. Janon, M. Nodet, and C. Prieur. Uncertainties assessment in global sensitivity indices estimation from metamodels. *International Journal for Uncertainty Quantification*, 4(1):21–36, 2014.
- [13] M.J.W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 117:35–43, 1999.
- [14] G. Kerkycharian and D. Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004.
- [15] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.
- [16] J.-M. Loubes, C. Marteau, and M. Solís. Rates of convergence in conditional covariance matrix estimation. 2014. ArXiv:1310.8244.
- [17] A. Marrel, B. Iooss, S. Da Veiga, and M. Ribatet. Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing*, 22(3):833–847, 2012.
- [18] P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Jean Picard, Berlin Heidelberg, springer edition, 2007. Ecole d’Eté de Probabilités de Saint-Flour XXXIII-2003.
- [19] A. Saltelli, K. Chan, and E.M. Scott. *Sensitivity analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2000.
- [20] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global sensitivity analysis*. John Wiley & Sons, Chichester, 2008.
- [21] I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment*, 1(4):407–414, 1993.
- [22] M. Solís. *Conditional covariance estimation for dimension reduction and sensitivity analysis*. Phd thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, Toulouse, France, 2014.
- [23] Lorna E Thorpe, Lawrence J Ouellet, Ronald Hershov, Susan L Bailey, Ian T Williams, John Williamson, Edgar R Monterroso, and Richard S Garfein. Risk of hepatitis c virus infection among young adult injection drug users who share injection equipment. *American journal of epidemiology*, 155(7):645–653, 2002.
- [24] V.C. Tran. *Une ballade en forêts aléatoires. Théorèmes limites pour des populations structurées et leurs généalogies, étude probabiliste et statistique de modèles SIR en épidémiologie, contributions à la géométrie aléatoire*. Habilitation à diriger des recherches, Université de Lille 1, 11 2014. <http://tel.archives-ouvertes.fr/tel-01087229>.
- [25] A.B. Tsybakov. *Introduction à l’estimation non-paramétrique*, volume 41 of *Mathématiques & Applications*. Springer, 2004.
- [26] WHO. Hepatitis C Fact Sheet 164, 2014.