# Efficient influence functions for Sobol' indices under two designs of experiments

Thierry Klein, Agnès Lagnoux, Paul Rochet, Thi Mong Ngoc Nguyen

# Efficient influence functions for Sobol' indices under two designs of experiments

Thierry Klein[1,2], Agnès Lagnoux[2,3], Thi Mong Ngoc Nguyen[4], and Paul Rochet[1]

[1]Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France.
[2]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS.
[3]UT2J, F-31058 Toulouse, France.
[4]Faculty of Mathematics & Computer Science, University of Science, VNU-HCMC, Ho Chi Minh City, Viet Nam.

July 20, 2024

**Abstract**

In this note, we are interested in the asymptotic efficiency of Sobol' indices estimators. After recalling the basis of asymptotic efficiency, we compute the efficient influence functions for Sobol' indices in two different contexts: the Pick-Freeze and the given-data settings.

## 1 Introduction

The use of complex computer models for the simulation and analysis of natural systems from physics, engineering, and other fields is by now routine. These models usually depend on many input variables, and it is thus crucial to understand which input parameter or which set of input parameters have an influence on the output. This is the aim of sensitivity analysis which has become an essential tool for system modeling and policy support (see, e.g., [19]). Global sensitivity analysis methods consider the input vector as random and propose a measure of influence of each subset of its components on the output of interest. We refer to the seminal book [20] for an overview on global sensitivity

analysis or to [6] for a synthesis of recent trends in the field. Among the different measures of global sensitivity analysis, variance-based measures are probably the most commonly used. The definition of the so-called Sobol' indices, introduced in [17] and later revisited in the framework of sensitivity analysis in [21, 22], is based on the Hoeffding decomposition of the variance [12]. More precisely, for the output $Y$ of a computer code $Y = G(V_1, \dots, V_p)$ where the inputs $V_i$ are assumed to be mutually independent, the Sobol' index of $Y$ with respect to a subset of inputs $X$ of dimension $d$ is defined by

$$S^X = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} = \frac{\mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2}{\mathrm{Var}(Y)}.$$

Since in practice computing explicitly the theoretical value of $S^X$ is out of reach, one of the main tasks in sensitivity analysis is to provide estimators of $S^X$, with guaranteed asymptotic properties such as consistency, rate of convergence, central limit theorem. In the recent years, a myriad of different estimators has been proposed, see [6, Chapter 4] for a complete review. To compare these different estimators, it is then relevant to define a notion of "optimality" using a concept similar to the Cramér-Rao bound in parametric statistics.

Optimality is assessed via the notion of *asymptotic efficiency* introduced in the seminal works [11, 13] in a parametric setting and further extended to semi and non-parametric models in [2, 16, 1] (see also [4, 23] for an extensive description of the theory of asymptotic efficiency). For Sobol' index inference specifically, the whole difficulty in showing asymptotic efficiency revolves around determining the so-called efficient influence function of the parameter $\psi = \mathbb{E}[\mathbb{E}[Y|X]^2]$. Once this is done, assessing the asymptotic efficiency of a particular estimator boils down to checking if its first order asymptotic Taylor expansion matches the empirical mean of the efficient influence function.

We tackle the issue of asymptotic efficiency in the two main frameworks related to Sobol' index inference, namely the Pick-Freeze and the given-data settings. When the context allows it, using the particular Pick-Freeze setting where two draws $Y, Y^X$ of the response are available for each realization of the input $X$, considerably simplifies the estimation process. The key to its success is to exploit a secondary expression of the parameter of interest given by

$$\psi = \mathbb{E}[\mathbb{E}[Y|X]^2] = \mathbb{E}[YY^X]$$

making a simple empirical estimator of $\psi$ available. By calculating the efficient influence function, we show this natural estimator to be asymptotically efficient in this context. The result remains valid under the sole assumption that the pair $(Y, Y^X)$ is exchangeable (i.e. $(Y, Y^X)$ and $(Y^X, Y)$ are identically distributed), thus providing a necessary and sufficient condition for asymptotic efficiency of the parameter $\psi = \mathbb{E}[YY^X]$ in the semi-parametric model of exchangeable bi-variate distributions.

In most situations however, the practitioner cannot afford the luxury of choosing the input's values when generating the data. The most common scenario is to deal with independent and identically distributed (i.i.d.) copies of $(X, Y)$, which constitutes a particular case of given-data. In this situation, we determine the efficient influence function in the non-parametric model on the distribution of the pair $(X, Y)$. Asymptotically efficient estimators of the Sobol' index have been proposed in the literature, although they usually require strong assumptions such as a low-dimensional input or an extensively smooth non-parametric regression function $x \mapsto \mathbb{E}[Y|X = x]$. In practice, building an asymptotically efficient estimator that lives up to its theoretical properties on numerical simulations (e.g. for high-dimensional inputs) remains somewhat of an open problem [7].

The article is organized as follows. In Section 2, we recall the definitions of efficient influence function and asymptotic efficiency as well as the useful Lemma 2.5 from [23] that links the two notions. Section 3 is devoted to the characterization of the efficient influence functions for Sobol' indices in the Pick-Freeze and in the given-data settings.

# 2 Asymptotic efficiency

Let us consider a set $\mathcal{P}$ of probability measures and a functional $\psi \colon \mathcal{P} \to \mathbb{R}$. Suppose that we observe an i.i.d. sample $Z_1, \ldots, Z_n$ from the distribution $P \in \mathcal{P}$ and we want to estimate the parameter $\psi(P)$. The aim is to define a notion of asymptotic optimality for an estimator of $\psi(P)$ based on the $n$-sample $(Z_1, \ldots, Z_n)$.

Inspired from [4, 23, 25], this section deals with asymptotic efficiency in semiparametric models that is the analog of the efficiency theory developed in the parametric setting in the sense of the Cramér-Rao bound (the minimal variance of an unbiased estimator of the natural parameter in an exponential family).

**Definition 2.1.** *A parametric submodel* $\{P_t,\ t \in [0, \varepsilon),\ \varepsilon > 0\} \subseteq \mathcal{P}$ *dominated by some measure* $\mu$ *is* differentiable in quadratic mean *at* $t = 0$ *with score function* $g \in L^2(P_0)$ *if*

$$\lim_{t \to 0^+} \int \left( \frac{\sqrt{f_t} - \sqrt{f_0}}{t} - \frac{1}{2} g \sqrt{f_0} \right)^2 d\mu = 0 \tag{1}$$

*where* $f_t = dP_t/d\mu$ *for* $t \in [0, \varepsilon)$.

Letting the maps $t \mapsto P_t$ range over all collections of differentiable submodels with a common root $P_0 = P \in \mathcal{P}$, we obtain a collection of score functions that defines the *tangent set* at $P$ denoted by $\dot{\mathcal{P}}_P \subset L^2(P)$.

A fundamental requirement to define asymptotic efficiency in semiparametric models is the pathwise differentiability of the target functional $\psi$. The motivation is to proceed to a distributional Taylor expansion of the parameter $\psi(P)$ around $P$ along differentiable

submodels. In the rest of the section, when considering a differentiable submodel $\{P_t,\ t \in [0,\varepsilon)\}$, it is always assumed implicitly that $g$ is the score function, $P_0 = P$ and the map $t \mapsto \psi(P_t)$ is differentiable at $t = 0$.

**Definition 2.2.** *A function $\psi_P \in L^2(P)$ is an* influence function *for estimating $\psi(P)$ if*

$$\lim_{t\to 0^+} \frac{\psi(P_t) - \psi(P)}{t} = \mathbb{E}_P[\psi_P(Z)g(Z)]$$

*for any differentiable submodel $\{P_t,\ t \in [0,\varepsilon)\}$. Moreover, the* efficient influence function *$\widetilde{\psi}_P$ is the unique influence function in the closure of the linear span of $\dot{\mathcal{P}}_P$ in $L^2(P)$.*

The notation $\mathbb{E}_P[h(Z)]$ means that the expectation is taken with respect to $P$ and thus $Z$ is assumed to be $P$-distributed. Remark that the efficient influence function can be obtained as the orthogonal projection of any influence function onto $\overline{\operatorname{lin}\dot{\mathcal{P}}_P}$ (the closure of the linear span of $\dot{\mathcal{P}}_P$ in $L^2(P)$). Furthermore, it entails that $\mathbb{E}_P[g(Z)] = 0$ and $\mathbb{E}_P[g^2(Z)] < \infty$ [24, Lemma 1.7].

For parametric models, the Cramer-Rao bound only applies to unbiased estimators which is far too restrictive in the semiparametric context. The notion of *regularity* defined below is used instead as a requirement for an estimator $\psi_n$ built from an i.i.d. sample $Z_1, ..., Z_n$ to be asymptotically efficient.

**Definition 2.3.** *An estimator $\hat{\psi}_n$ is* regular *if there exists a probability distribution $L$ such that*

$$\sqrt{n}(\hat{\psi}_n - \psi(P_{1/\sqrt{n}})) \xrightarrow[n\to\infty]{\mathcal{L}} L \tag{2}$$

*for all differentiable submodels $\{P_t,\ t \in [0,\varepsilon)\}$, where for all $n \in \mathbb{N}$, the sample $Z_1, ..., Z_n$ is drawn from $P_{1/\sqrt{n}}$.*

By [23, Theorem 25.20], for a regular estimator $\hat{\psi}_n$ with centered Gaussian limit distribution $L = \mathcal{N}(0, \sigma^2)$, the variance $\sigma^2$ must satisfy

$$\sigma^2 \geqslant \frac{\mathbb{E}_P[\psi_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]},$$

for any submodel $\{P_t,\ t \in [0,\varepsilon)\} \subseteq \mathcal{P}$ and influence function $\psi_P$. This is the analog of the Cramer-Rao inequality in parametric models. If $\dot{\mathcal{P}}_P$ is a linear subspace of $L^2(P)$ (otherwise one must take its linear span $\operatorname{lin}\dot{\mathcal{P}}_P$ instead for the following equation to hold), taking the supremum over all submodels yields the optimized lower bound

$$\sigma^2 \geqslant \sup_{g\in\dot{\mathcal{P}}_P} \frac{\mathbb{E}_P[\psi_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]} = \mathbb{E}_P[\tilde{\psi}_P(Z)^2].$$

This leads to the very definition of asymptotic efficiency for a regular estimator.

**Definition 2.4.** *A regular estimator $\hat{\psi}_n$ is* asymptotically efficient at $P$ *if the limit distribution $L$ in* (2) *is the centered Gaussian with minimal variance :*

$$\sqrt{n}(\hat{\psi}_n - \psi(P_{1/\sqrt{n}})) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathbb{E}_P[\tilde{\psi}_P(Z)^2]). \tag{3}$$

Although the definition of asymptotic efficiency pertains to the limit distribution of an estimator $\hat{\psi}_n$, the important Lemma 25.23 from [23] reveals that it is in fact a condition in a, much stronger, probabilistic sense.

**Lemma 2.5.** *A regular estimator $\hat{\psi}_n$ is asymptotically efficient at $P$ if and only if the following expansion holds*

$$\hat{\psi}_n = \psi(P) + \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_P(Z_i) + o_{\mathbb{P}}(1/\sqrt{n}). \tag{4}$$

When considering several parameters $\psi^1(P), ..., \psi^k(P)$ simultaneously, the previous result has strong implications which directly stem from Definition 2.2. Firstly, marginal asymptotic efficiency of estimators $\hat{\psi}_n^1, ..., \psi_n^k$ implies joint asymptotic efficiency, since the efficient influence function of the vector parameter $\Psi = (\psi^1, ..., \psi^k)$ at $P$ is the vector $\widetilde{\Psi}_P = (\widetilde{\psi}_P^1, ..., \widetilde{\psi}_P^k)$ of the marginal efficient influence functions. Secondly, asymptotic efficiency is stable through smooth transformations : if $\phi$ is a differentiable function from $\mathbb{R}^k$ to $\mathbb{R}$, the efficient influence function of the parameter $\phi \circ \Psi(P)$ can be identified from Definition 2.2 in view of

$$\lim_{t\to 0^+} \frac{\phi \circ \Psi(P_t) - \phi \circ \Psi(P)}{t} = \mathbb{E}_P\Big[\Big(\nabla\phi(\Psi(P))^\top \widetilde{\Psi}_P(Z)\Big) g(Z)\Big],$$

where $\nabla$ denotes the gradient operator. As a direct consequence of the Delta method, the estimator $\phi(\hat{\psi}_n^1, ..., \hat{\psi}_n^k)$ thus satisfies the condition of Lemma 2.5 ,

$$\phi(\hat{\psi}_n^1, ..., \hat{\psi}_n^k) = \phi \circ \Psi(P) + \frac{1}{n}\sum_{i=1}^{n} \nabla\phi(\Psi(P))^\top \widetilde{\Psi}_P(Z_i) + o_{\mathbb{P}}(1/\sqrt{n}),$$

proving it is asymptotically efficient.

**How to show that an estimator of $\psi(P)$ is asymptotically efficient?**

- Determine the tangent set $\dot{\mathcal{P}}_P$.
- Compute the efficient influence function $\widetilde{\psi}_P \in \overline{\operatorname{lin} \dot{\mathcal{P}}_P}$ for estimating $\psi(P)$. This can be done by showing that

  $$\forall g \in \dot{\mathcal{P}}_P, \lim_{t\to 0^+} \frac{\psi(P_t) - \psi(P)}{t} = \mathbb{E}_P[\widetilde{\psi}_P(Z) g(Z)]$$

  for some differentiable submodel $P_t$ with score function $g$.

- Find an asymptotically efficient estimator $\hat{\psi}_n$ of $\psi(P)$. This is equivalent to showing that it satisfies the following expansion by Lemma 2.5 :

$$\hat{\psi}_n = \psi(P) + \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_P(Z_i) + o_{\mathbb{P}}(1/\sqrt{n})$$

More generally, when the quantity of interest writes as $\phi(\psi^1(P), \ldots, \psi^k(P))$ where $\phi$ is a differentiable function from $\mathbb{R}^k$ to $\mathbb{R}$ and asymptotically efficient estimators $\hat{\psi}_n^1, ..., \hat{\psi}_n^k$ are available for each parameter, then $\phi(\hat{\psi}_n^1, ..., \hat{\psi}_n^k)$ is asymptotically efficient to estimate $\phi(\psi^1(P), \ldots, \psi^k(P))$, as a consequence of the two following theorems from [23] :

- Theorem 25.50 (efficiency in product spaces) ensures that $(\hat{\psi}_n^1, ..., \hat{\psi}_n^k)$ is (jointly) asymptotically efficient for estimating $(\psi^1(P), \ldots, \psi^k(P))$ ;
- Theorem 25.47 (efficiency and the Delta method) states that $\phi(\hat{\psi}_n^1, ..., \hat{\psi}_n^k)$ verifies the condition of Lemma 2.5 for asymptotic efficiency to estimate $\phi(\psi^1(P), \ldots, \psi^k(P))$.

# 3 Application to the estimation of Sobol' indices

We consider the following model

$$Y = G(X, W)$$

for some measurable function $G$, where $Y$ is a real-valued square integrable output and $X \in \mathbb{R}^d$ for some $d \in \mathbb{N}^*$ is a vector-valued input and $W$ a random term independent of $X$. In the context of sensibility analysis, an indicator commonly used to quantify the impact of one or several inputs on the output $Y$ is the so-called Sobol' index. Then the Sobol' index with respect to $X$ is defined by

$$S^X = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} = \frac{\mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2}{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}. \tag{5}$$

The difficult term to estimate is $\mathbb{E}[\mathbb{E}[Y|X]^2]$ since it involves a conditional expectation. Two procedures then arise. The first one uses the Pick-Freeze trick consisting in rewriting the variance of the conditional expectation in terms of a covariance. This procedure was the first method introduced and theoretically studied. The second procedure makes use of an non-parametric estimation of the regression function $m(x) = \mathbb{E}[Y|X = x]$ that is challenging to obtain. Because these two settings rely on different definitions of the underlying model and the parameter, the corresponding efficient influence functions are different and computed separately. In particular, the tangent sets $\dot{\mathcal{P}}_P$ are specific to the considered setting as proved in the following.

## 3.1 Efficient influence function for the Pick-Freeze setting

As mentioned previously, to avoid a direct estimation of the conditional expectation, the Pick-Freeze approach [9, 10, 14] relies on the rewriting:

$$\mathbb{E}[\mathbb{E}[Y|X]^2] = \mathbb{E}[YY^X],$$

where $Y^X = G(X, W')$ is obtained using a copy $W'$ of $W$ independent from $X, W$. Furthermore, since $Y$ and $Y^X$ are identically distributed, we have

$$\mathbb{E}[Y] = \frac{1}{2}(\mathbb{E}[Y] + \mathbb{E}[Y^X]) \quad \text{and} \quad \mathbb{E}[Y^2] = \frac{1}{2}(\mathbb{E}[Y^2] + \mathbb{E}[(Y^X)^2]).$$

Hence, given a particular sampling design where both $Y_i$ and $Y_i^X$ are observed for each input value $X_i$ for $i = 1, \dots, n$, the Pick-Freeze estimator of $S^X$ defined in Equation (5) can be built naturally by replacing the expectations $\mathbb{E}[YY^X]$, $\mathbb{E}[Y]$ and $\mathbb{E}[Y]^2$ by their empirical version using all the information: the $n$-sample of $Y$ together with the $n$-sample of $Y^X$. As shown in the sequel, this natural estimator is asymptotically efficient to estimate $S^X$.

More precisely, let $\mathcal{P}$ be the set of all distributions of exchangeable random vectors $(Y_1, Y_2)$ in $L^2(\mathbb{R}^2) : (Y_1, Y_2) \overset{\mathcal{L}}{=} (Y_2, Y_1)$. It is clear a random vector of $L^2(\mathbb{R}^2)$ is in $\mathcal{P}$ if and only if its cumulative distribution function $F$ is symmetric:

$$F(y_1, y_2) = F(y_2, y_1) \quad \forall (y_1, y_2) \in \mathbb{R}^2.$$

Let $P$ be the distribution of $(Y, Y^X)$. We check that $P \in \mathcal{P}$ thanks to [14, Lemma 2.4]. The tangent set of $\mathcal{P}$ at $P$ is given by:

$$\dot{\mathcal{P}}_P = \{g \in L^2(P) : \mathbb{E}_P[g(Y, Y^X)] = 0 \text{ and } g(y_1, y_2) = g(y_2, y_1), \forall (y_1, y_2) \in \mathbb{R}^2\}.$$

Indeed, for all $P$-square-integrable and symmetrical function $g$, the submodel $\{P_t, t \geqslant 0\}$ whose Radon-Nikodym densities with respect to $P$ are given by

$$\forall t > 0\,, \; \frac{dP_t}{dP}(y_1, y_2) = \frac{c(t)}{1 + e^{-2tg(y_1, y_2)}} \text{ with } c(t) = \Big( \int \frac{1}{1 + e^{-2tg(y_1, y_2)}} dP(y_1, y_2) \Big)^{-1}$$

has score $g$ at $t = 0$ and clearly lies in $\mathcal{P}$. Moreover, the convergence in $L^2(\mu)$ of $(\sqrt{f_{t_n}} - \sqrt{f_0})/t_n$ towards $g\sqrt{f_0}/2$ with $t_n \to 0^+$ as $n \to \infty$ (i.e. the differentiability in the sense of Definition 2.1) implies the almost-sure convergence of a subsequence. Hence, the exchangeability condition in this model implies that all score functions $g$ are symmetrical $P$-almost surely.

To determine the efficient influence function, it is sufficient to consider differentiable submodels with a uniformly bounded score function $g$, since such functions are dense in

$\dot{\mathcal{P}}_P$ with respect to the $L^2(P)$-metric. Hence, we assume without loss of generality that $g$ is bounded, with $\|.\|_\infty$ designating its supremum norm restricted to the support of $P$ :

$$\|g\|_\infty := \sup_{(y_1,y_2)\in \mathrm{supp}(P)} |g(y_1,y_2)|.$$

Moreover, since all differentiable submodels $\{P_t, t \in [0,\varepsilon)\}$ with score $g$ lead to the same Riesz-representation

$$\lim_{t\to 0^+} \frac{\psi(P_t)-\psi(P)}{t} = \mathbb{E}_P[\widetilde{\psi}_P(Y,Y^X)g(Y,Y^X)]$$

where $\widetilde{\psi}_P \in \dot{\mathcal{P}}_P$ is the efficient influence function, considering the simple submodels $P_t = (1+tg)P$ is sufficient to determine $\widetilde{\psi}$. The boundedness of $g$ guarantees that the submodel $P_t = (1+tg)P$ is well defined for $\varepsilon$ small enough.

**Proposition 3.1** (Efficient influence function - Pick-Freeze [14])**.** *If $\mathbb{E}_P[(YY^X)^2] < \infty$, the efficient influence functions of $\mathbb{E}_P[YY^X]$, $\mathbb{E}_P[Y]$, and $\mathbb{E}_P[Y^2]$ at $P$ are respectively given by*

$$\begin{aligned}
(y_1,y_2) &\mapsto y_1y_2 - \mathbb{E}_P[YY^X],\\
(y_1,y_2) &\mapsto \frac{1}{2}(y_1+y_2) - \mathbb{E}_P[Y],\\
(y_1,y_2) &\mapsto \frac{1}{2}(y_1^2+y_2^2) - \mathbb{E}_P[Y^2].
\end{aligned}$$

*Proof of Proposition 3.1.* For any bounded function $g \in \dot{\mathcal{P}}_P$ and $P_t = (1+tg)P$, we have

$$\frac{\mathbb{E}_{P_t}[YY^X] - \mathbb{E}_P[YY^X]}{t} \;=\; \mathbb{E}_P[YY^X g(Y,Y^X)] = \mathbb{E}_P\left[(YY^X - \mathbb{E}_P[YY^X])g(Y,Y^X)\right].$$

Since the map $(y_1,y_2) \mapsto y_1y_2 - \mathbb{E}_P[YY^X]$ lies in $\dot{\mathcal{P}}_P$ (i.e. it is symmetrical and $P$-square-integrable by assumption), it is the efficient influence function for the parameter $\mathbb{E}_P[YY^X]$. Proceeding in the same way, we show that $(y_1,y_2) \mapsto y_1 - \mathbb{E}_P[Y]$ is an influence function for the parameter $\mathbb{E}_P[Y]$, as is $(y_1,y_2) \mapsto y_2 - \mathbb{E}_P[Y]$ by exchangeability. We can thus identify the efficient influence function

$$(y_1,y_2) \mapsto \frac{y_1+y_2}{2} - \mathbb{E}_P[Y]$$

as the only symmetrical influence function. The reasoning is the same for the third parameter $\mathbb{E}_P[Y^2]$. □

**Proposition 3.2** (Asymptotically efficient estimation of $S^X$ - Pick-Freeze [14])**.** *The estimator $S^X_{n,PF}$ defined by*

$$S^X_{n,PF} = \frac{\frac{1}{n}\sum_{i=1}^n Y_i Y_i^X - \left(\frac{1}{2n}\sum_{i=1}^n (Y_i + Y_i^X)\right)^2}{\frac{1}{2n}\sum_{i=1}^n (Y_i^2 + (Y_i^X)^2) - \left(\frac{1}{2n}\sum_{i=1}^n (Y_i + Y_i^X)\right)^2}$$

*is asymptotically efficient for estimating $S^X$ for $P \in \mathcal{P}$.*

*Proof of Proposition 3.2.* We proceed as explained in Section 2. Observe that

$$S^X = \phi(\mathbb{E}_P[YY^X], \mathbb{E}_P[Y], \mathbb{E}_P[Y^2]) \quad \text{where } \phi(x,y,z) = \frac{x-y^2}{z-y^2}.$$

By Lemma 2.5, the empirical estimators

$$\frac{1}{2n}\sum_{i=1}^n (Y_i + Y_i^X) \quad \frac{1}{2n}\sum_{i=1}^n (Y_i^2 + (Y_i^X)^2) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n Y_i Y_i^X$$

are asymptotically efficient to estimate $\mathbb{E}_P[Y], \mathbb{E}_P[Y^2]$ and $\mathbb{E}_P[YY^X]$ respectively. The asymptotic efficiency of $S^X_{n,\mathrm{PF}}$ follows from the differentiability of $\phi$ by Theorems 25.47 and 25.50 of [23], as explained in the end of Section 2. □

## 3.2 Efficient influence function for the given-data setting

We now consider the given-data setting where we observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn independently from some distribution $P$. The model $\mathcal{P}$ contains all the distributions on $\mathbb{R}^d \times \mathbb{R}$ that are square integrable with respect to their second argument : $\mathbb{E}_P[Y^2] < \infty$. In this model, the tangent set $\dot{\mathcal{P}}_P$ at $P \in \mathcal{P}$ is the maximal tangent set, containing all $P$-square-integrable functions $g$ with zero integral

$$\dot{\mathcal{P}}_P = \{g \in L^2(P) : \mathbb{E}_P[g(X,Y)] = 0\}.$$

Indeed, for all $P$-square-integrable function $g$, the submodel $\{P_t, t \geqslant 0\}$ whose Radon-Nikodym densities with respect to $P$ are given by

$$\forall t > 0\,,\ \frac{dP_t}{dP}(x,y) = \frac{c(t)}{1 + e^{-2tg(x,y)}} \ \text{ with } \ c(t) = \left(\int \frac{1}{1 + e^{-2tg(x,y)}} dP(x,y)\right)^{-1}$$

has score $g$ at $t = 0$ and clearly lies in $\mathcal{P}$ for all $t > 0$ in view of

$$\mathbb{E}_{P_t}[Y^2] \leqslant c(t)\mathbb{E}_P[Y^2] < \infty.$$

It remains to determine the efficient influence functions of the three parameters $P \mapsto \mathbb{E}_P[Y]$, $P \mapsto \mathbb{E}_P[Y^2]$ and $P \mapsto \mathbb{E}_P[\mathbb{E}_P[Y|X]^2]$. As discussed in Section 3.1, it is sufficient for this purpose to consider the submodels of the form $P_t = (1 + tg)P$ with $g$ uniformly bounded. Because the first two parameters are linear, they are particularly easy to deal with, e.g.

$$\frac{\mathbb{E}_{P_t}[Y] - \mathbb{E}_P[Y]}{t} = \mathbb{E}_P[Y g(X,Y)] = \mathbb{E}_P\Big[(Y - \mathbb{E}_P[Y])g(X,Y)\Big].$$

We verify easily that the efficient influence functions of $\mathbb{E}_P[Y]$ and $\mathbb{E}_P[Y^2]$ are respectively

$$(x,y) \mapsto y - \mathbb{E}_P[Y] \quad \text{and} \quad (x,y) \mapsto y^2 - \mathbb{E}_P[Y^2],$$

although the latter requires the additional condition $\mathbb{E}_P[Y^4] < \infty$ for it to lie in the tangent set $\dot{\mathcal{P}}_P$. The empirical means $\frac{1}{n}\sum_{i=1}^n Y_i$ and $\frac{1}{n}\sum_{i=1}^n Y_i^2$ are thus proved to be asymptotically efficient by Equation 4 in Lemma 2.5.

Regarding the estimation of $\psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y|X]^2]$, the efficient influence function has been given (without proof) in [8]. In [5], the authors recover the efficient influence function when $X$ is one-dimensional and the distribution of $(X,Y)$ is absolutely continuous with respect to the Lebesgue measure. We believe that their proof is still valid for a multidimensional input $X$. For completeness, we here calculate the efficient influence function in the general case.

**Proposition 3.3** (Efficient influence function - given-data setting)**.** *If $\mathbb{E}_P[Y^4] < \infty$, the efficient influence function of $\psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y|X]^2]$ at $P$ is given by*

$$(x,y) \mapsto (2y - m(x))m(x) - \psi(P) \tag{6}$$

*where $m(x) = \mathbb{E}_P[Y|X = x]$.*

*Proof of Proposition 3.3.* Let $P_t = (1 + tg)P$ with $g$ uniformly bounded. We recall that the conditional expectation function of $Y$ knowing $X$ under $P_t$

$$m_t : x \mapsto \mathbb{E}_{P_t}[Y|X = x]$$

satisfies

$$\mathbb{E}_{P_t}[Y h(X)] = \mathbb{E}_{P_t}[m_t(X)h(X)] \tag{7}$$

for all measurable function $h : \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}_{P_t}[|Y h(X)|] < +\infty$. From

$$\psi(P_t) := \mathbb{E}_{P_t}[Y m_t(X)] = \mathbb{E}_P[Y m_t(X)] + t\,\mathbb{E}_P[Y m_t(X) g(X,Y)],$$

we deduce

$$\frac{\psi(P_t) - \psi(P)}{t} = \mathbb{E}_P\Big[\frac{m_t(X) - m(X)}{t} Y\Big] + \mathbb{E}_P\Big[Y m_t(X) g(X,Y)\Big]$$

recalling that $\mathbb{E}_P[Ym(X)] = \mathbb{E}_P[m^2(X)]$. Taking $h(x) = m(x)$ in (7) yields in particular

$$\mathbb{E}_P\Big[Ym(X)\Big(1 + tg(X,Y)\Big)\Big] = \mathbb{E}_P\Big[m_t(X)m(X)\Big(1 + tg(X,Y)\Big)\Big]$$

leading to

$$\mathbb{E}_P\Big[\frac{m_t(X) - m(X)}{t}m(X)\Big] = \mathbb{E}_P\Big[(Y - m_t(X))m(X)g(X,Y)\Big]$$

whence

$$\frac{\psi(P_t) - \psi(P)}{t} = \mathbb{E}_P\Big[\Big((Y - m_t(X))m(X) + Ym_t(X)\Big)g(X,Y)\Big].$$

Using Lemma 3.4 below, we get

$$\begin{aligned}\Big|\frac{\psi(P_t) - \psi(P)}{t} &- \mathbb{E}_P\Big[\Big(2Y - m(X)\Big)m(X)g(X,Y)\Big]\Big| \\ &= \Big|\mathbb{E}_P\Big[(m_t(X) - m(X))(Y - m(X))g(X,Y)\Big]\Big| \\ &\leqslant \|g\|_\infty\sqrt{\mathbb{E}_P[(Y - m(X))^2]}\,\|m_t - m\|_P \xrightarrow[t\to 0^+]{} 0\end{aligned}$$

Hence,

$$\lim_{t\to 0^+}\frac{\psi(P_t) - \psi(P)}{t} = \mathbb{E}_P\Big[\Big((2Y - m(X))m(X) - \psi(P)\Big)g(X,Y)\Big]$$

revealing $(x,y) \mapsto (2y - m(x))m(x) - \psi(P) \in \dot{\mathcal{P}}_P$ as the efficient influence function. □

**Lemma 3.4.** *Let $P_t$ be a measure absolutely continuous with respect to $P$ with Radon-Nicodym density $f_t(x,y), (x,y) \in \mathrm{supp}(P)$ and such that $\int y^2 dP_t(x,y) < \infty$. If $f_t$ converges uniformly towards 1 as $t \to 0^+$, then $m_t$ converges towards $m$ in $L^2(P)$ :*

$$\|m_t - m\|_P^2 := \int (m_t(x) - m(x))^2 dP(x,y) \xrightarrow[t\to 0^+]{} 0.$$

*Proof of Lemma 3.4.* For $t > 0$ sufficiently small so that $f_t(x,y) > 1/2$ for all $(x,y) \in \mathrm{supp}(P)$, we have

$$\|m_t\|_P^2 = \int m_t^2(x)dP(x,y) \leqslant 2\int m_t^2(x)dP_t(x,y) \leqslant \int y^2 dP_t(x,y) < \infty,$$

guaranteeing that $\|m_t - m\|_P < \infty$ as $t \to 0^+$. Moreover, remark that

$$\begin{aligned}\|m_t - m\|_P^2 &= \int (m_t(x) - m(x))^2(dP_t - dP)(x,y) + \int (m_t(x) - m(x))y\,(dP_t - dP)(x,y) \\ &\leqslant \|f_t - 1\|_\infty\left(\|m_t - m\|_P^2 + \|m_t - m\|_P\sqrt{\mathbb{E}_P[Y^2]}\right)\end{aligned}$$

which concludes the proof. □

**Asymptotically efficient estimators in the literature** In [8], the authors considered a truncated version of $\psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y|X]^2]$. To estimate $\psi(P)$, they first estimate the regression function $m$ by a kernel estimator and then use a one-step procedure to improve the corresponding plug-in estimator. For a one-dimensional input $X$, an asymptotically efficient estimator of $\psi(P)$ that relies on a preliminary kernel estimator of the input's density was given in [5], while a simpler alternative approach based on ordered statistics can be found in [15]. More recently, combining the approaches of [8] and mirror transformations (see [3] and [18]), asymptotically efficient estimators of $\psi(P)$ are provided in [7] for an input $X$ of any dimension, under adequate regularity conditions.

# References


[1] J. M. Begun, W. J. Hall, W.-M. Huang, and J. A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. The Annals of Statistics, 11(2):432–452, 1983.

[2] R. Beran. Robust location estimates. The Annals of Statistics, pages 431–444, 1977.

[3] K. Bertin, N. Klutchnikoff, J. R. Léon, and C. Prieur. Adaptive density estimation on bounded domains under mixing conditions. Electronic Journal of Statistics, 14(1):2198 – 2237, 2020.

[4] P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. Efficient and adaptive estimation for semiparametric models, volume 4. Springer, 1993.

[5] S. Da Veiga. Global sensitivity analysis with dependence measures. J. Stat. Comput. Simul., 85(7):1283–1305, 2015.

[6] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. Basics and Trends in Sensitivity Analysis: Theory and Practice in R. SIAM, 2021.

[7] S. Da Veiga, F. Gamboa, A. Lagnoux, T. Klein, and C. Prieur. A mirror adaptation for one-step estimation of sobol' indices. arXiv preprint arXiv:2303.17832, 2023.

[8] K. Doksum and A. Samarov. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. The Annals of Statistics, pages 1443–1473, 1995.

[9] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol Pick-Freeze Monte Carlo method. Statistics, 50(4):881–902, 2016.

[10] F. Gamboa, T. Klein, and A. Lagnoux. Sensitivity analysis based on Cramér von Mises distance. SIAM/ASA Journal on Uncertainty Quantification, 6(2):522–548, Apr. 2018.

[11] J. Hájek. A characterization of limiting distributions of regular estimates. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 14(4):323–330, 1970.

[12] W. Hoeffding. A class of statistics with asymptotically normal distribution. Ann. Math. Statistics, 19:293–325, 1948.

[13] N. Inagaki. On the limiting distribution of a sequence of estimators with uniformity property. Annals of the Institute of Statistical Mathematics, 22(1):1–13, 1970.

[14] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol index estimators. ESAIM: Probability and Statistics, 18:342–364, 1 2014.

[15] T. Klein and P. Rochet. Efficiency of the averaged rank-based estimator for first order sobol index inference. Statistics & Probability Letters, 207:110015, 2024.

[16] B. Y. Levit. On the efficiency of a class of non-parametric estimates. Theory of Probability & Its Applications, 20(4):723–740, 1976.

[17] K. Pearson. On the partial correlation ratio. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 91(632):492–498, 1915.

[18] L. Pujol. Nonparametric estimation of a multivariate density under kullback-leibler loss with ISDE. arXiv preprint arXiv:2205.03199, 2022.

[19] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J. H. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, and H. Maier. The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. Environmental Modelling and Software, 137:104954, Mar. 2021.

[20] A. Saltelli, K. Chan, and E. Scott. Sensitivity analysis. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000.

[21] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. Math. Modeling Comput. Experiment, 1(4):407–414 (1995), 1993.

[22] I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation, 55(1-3):271–280, 2001.

[23] A. W. van der Vaart. Asymptotic statistics, volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[24] A. W. van der Vaart. Semiparametric statistics. In Lectures on probability theory and statistics (Saint-Flour, 1999), pages 331–457. Springer, 2002.

[25] A. Yiu, E. Fong, C. Holmes, and J. Rousseau. Semiparametric posterior corrections. arXiv preprint arXiv:2306.06059, 2023.