

Total Effects with Constrained Features

Emanuele Borgonovo*

Elmar Plischke†

Clémentine Prieur‡

February 2, 2024

Abstract

Recent studies have emphasized the connection between machine learning feature importance measures and total order sensitivity indices (total effects, henceforth). Feature correlations and the need to avoid unrestricted permutations make the estimation of these indices challenging. Additionally, there is no established theory or approach for non-Cartesian domains. We propose four alternative strategies for computing total effects that account for both dependent and constrained features. Our first approach involves a generalized winding stairs design combined with the Knothe-Rosenblatt transformation. This approach, while applicable to a wide family of input dependencies, becomes impractical when inputs are physically constrained. Our second approach is a U-statistic that combines the Jansen estimator with a weighting factor. The U-statistic framework allows the derivation of a central limit theorem for this estimator. However, this design is computationally intensive. Then, our third approach uses derangements to significantly reduce computational burden. We prove consistency and central limit theorems for these estimators as well. Our fourth approach is based on a nearest-neighbour intuition and it further reduces computational burden. We test these estimators through a series of increasingly complex computational experiments with features constrained on compact and connected domains (circle, simplex), non-compact and non-connected domains (Sierpinski gaskets), we provide comparisons with machine learning approaches and conclude with an application to a realistic simulator.

Keywords Feature Importance; Constrained Features; Winding Stairs; U-Statistics

1 Introduction

Determining feature importance is a crucial task in machine learning and statistical investigations. In machine learning, it is an integral part of post-hoc explainability [Murdoch et al., 2019, Fisher et al., 2019], where it helps us understand the degree to which a model relies on the available features. This understanding has two final objectives [Genuer et al., 2010, Bénard et al., 2022]. The first objective is dimensionality reduction, which involves screening out features that do not contribute to the model’s predictions. The second objective is identifying the features that are most important for further modeling efforts or data collection by domain experts.

Over the years, several feature importance measures have been developed to perform this task. On the one hand, in machine learning, a particularly important family is represented by Breiman’s permutation importance measures [Breiman, 2001]. Breiman originally defines them based on the notion of mean decrease accuracy (MDA) [Bénard et al., 2022]. The intuition is as follows: A given machine learning model (e.g., a random forest) is fitted to a feature-target dataset, yielding a given predictive accuracy. The values of a specific feature are then permuted to break its relation to the target. The predictive accuracy is then reassessed for this perturbed dataset. The difference between the new (possibly degraded) and the original accuracies provides us with an indication about the importance of the feature. However, [Bénard et al., 2022, Proposition 2] show that there is no consensus on the exact mathematical formulation of the mean decrease accuracy and they prove that alternative software implementations yield different values. On the other hand, in statistics, a central role is played by measures of statistical association. Several indicators have been developed over the years: From the original Pearson linear correlation coefficient [Pearson, 1895], to the new correlation coefficient of Chatterjee [2021]. In this family, a significant role is played by the so-called total order sensitivity effects [Homma and Saltelli, 1996]— total effects for short.

Total effects are defined as the difference between the variance of the target and the portion that remains after all features have been fixed with the exception of the feature of interest. Homma and Saltelli [1996]

*Università Commerciale Luigi Bocconi and Bocconi Institute for Data Science and Analytics, Milano 20138, Italy. emanuele.borgonovo@ubocconi.it

†Technische Universität Clausthal, 38678 Clausthal-Zellerfeld, Germany. elmar.plischke@tu-clausthal.de

‡Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France. clementine.prieur@univ-grenoble-alpes.fr

show that, when features are independent, the total effect of a given feature equals the sum of all terms in the ANOVA expansion associated with that feature. Also, we can calculate total effects as the expectation of the squared difference of the values of the model output in two points that differ only in the value of the feature of interest. The new point can be obtained by a simple permutation of the values of the features in the dataset. This formulation is known as the Jansen’s estimation method [Jansen et al., 1994, Jansen, 1999] and has inspired the so-called pick-and-freeze designs. In the class of pick-and-freeze estimators the one proposed by Janon et al. [2014] is proven to be asymptotically efficient.

Hart and Gremaud [2018] show that even in the case of dependent features, total effects retain an interpretation from a relative error perspective. Under a squared loss function, a total index is the expected loss increase for approximating the input-output mapping with a function that does not contain the terms associated with the feature of interest. Also, B  nard et al. [2022] show that total effects are closely related to Breiman’s mean decrease accuracy. In particular, [B  nard et al., 2022, Proposition 2] show that the different software implementations do not converge to the total effects but to a quantity whose bias increases with dependence, and is potentially amplified by interactions. They propose corrections so that the calculation of Breiman’s mean decrease accuracy indeed converges to a total effect in the case the machine learning model is a random forest.

The presence of statistical dependence complicates the calculation of total effects (we refer to [Da Veiga et al., 2021, Ch. 5] for a thorough account). First, the interpretation in terms of the correspondence with the sum of terms in the ANOVA decomposition is lost. Second, also the possibility to use a Jansen-type estimator is not straightforward. In fact, while under independence the new points can be obtained with unrestricted permutations, the presence of dependencies challenges such procedure. The problem is similar to the one signalled by Hooker et al. [2021] in the machine learning literature: unrestricted permutations may make the new points fall in regions that are far from where the data lie, forcing the machine learning model to extrapolate. These difficulties are compounded when features are not only dependent but also constrained on non-Cartesian supports. Constraints arise in applications when physical or business reasons require features to be located in certain regions. Here, an unrestricted permutation could lead to a feature that falls outside a given constraint, making the evaluation of the model not only at risk of extrapolation, but also meaningless. Furthermore, if constraints give rise to disconnected feature domains, they make the functional ANOVA expansion ill-defined Owen and Prieur [2017]. Kucherenko et al. [2017] propose a numerical approach based on the combination of rejection sampling and quadrature for the calculation of variance-based indices, with focus on numerical aspects. However, a statistical analysis of possible estimators with constrained inputs is missing.

Our goal is to address the estimation of total effects with both dependent and constrained features, considering numerical as well as theoretical aspects. Here we assume, as typically encountered in sensitivity analysis, that the input distribution is known. In Section 6 we discuss alternatives for standard machine learning settings, where only a data sample is available to estimate feature importance.

We proceed as follows. First, we extend the estimator of Jansen [1999] to the case of dependent inputs. We show that it is still possible to estimate total effects under input dependence using a Jansen-like approach if the new value of the feature is obtained under conditional independence. We then propose a generalized winding stairs design based on the Knothe-Rosenblatt transform that can be used in association with a vast family of input dependencies. However, while this design conceptually pushes the boundaries of available methods for dependent inputs, it becomes impractical when inputs are constrained.

We then introduce a new estimator of total indices by applying a weighting factor (called density quotient) to the extended Jansen’s estimator. We show that the density quotient can be reinterpreted as a block-copula density, that vanishes when inputs are outside the constraints and that becomes unity when inputs are independent. We then formulate a U -statistic version of the estimator and obtain a central limit theorem. However, the new U -statistic estimator turns out to be computationally expensive as it requires the evaluation of the model at n^2 points. We then propose an alternative estimator based on a single permutation that reduces computational burden. We consider first the simplest estimator with the permutation given by a one-shift in the coordinate of interest and prove a central limit theorem of this estimator. We then extend the result for general derangements in the coordinate of interest. To further abate computational burden, we also introduce a nearest-neighbour estimator that makes the estimation cost independent of the number of features.

We derive analytical expressions for the estimators in the case of linear models and Gaussian inputs. We then challenge the estimators on test cases of increasing complexity, starting with Cartesian domains with dependent features, and moving to connected non-Cartesian domains (e.g., circle and simplex) to disconnected non-Cartesian domains (such as non-overlapping triangles and Sierpinski gaskets), we also provide comparison tests with machine learning approaches and finally conclude with an application to a realistic simulator, the flood example of de Rocquigny [2006].

2 Total Effects: Bridging Old and New

In this section, we review total effects from a fresh perspective. We discuss the covariance representation of total effects and establish a link with an early result by Fréchet. We underline the role of conditional independence in the estimation of total effects with independent as well as dependent inputs. The analysis allows us to propose a new estimator for total effects with dependent inputs based on winding stairs and the Knothe-Rosenblatt transformation. In Section 2.2, we highlight the roles of conditional independence for such a representation. In Section 2.3, we propose a new estimator based on winding stairs and on the Knothe-Rosenblatt transformation for the case in which input dependence can be expressed via a Gaussian copula.

2.1 A Fréchet Perspective

Let us consider a reference probability space $(\Omega, \mathcal{B}(\Omega), \Pr)$, where $\mathcal{B}(\Omega)$ is a Borel σ -algebra. Let also X, Y be random variables on $(\Omega, \mathcal{B}(\Omega), \Pr)$, with supports \mathcal{X}, \mathcal{Y} . We let $X = (X_1, X_2, \dots, X_d)$ be a d -dimensional random vector in \mathbb{R}^d , so that $\mathcal{X} \subseteq \mathbb{R}^d$ and consider a univariate Y , with $\mathcal{Y} \subseteq \mathbb{R}$. For the moment, we make the further assumption that the support of X is Cartesian, that is $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_d$, where \mathcal{X}_i is the support of X_i , $i = 1, 2, \dots, d$. We denote the cumulative distribution function and probability density functions of X and Y by $F_X(x)$, $f_X(x)$ and $F_Y(y)$, $f_Y(y)$, respectively. For notation simplicity, we regard X and Y as continuous in the remainder. We suppose that Y has finite second moment, $\mathbb{V}[Y] < \infty$. Let $u \subseteq [d] = \{1, \dots, d\}$, e.g., $u = \{i_1, i_2, \dots, i_k\}$ with $k \leq d$. Let x_u correspond to the $|u|$ -dimensional vector whose components are indexed by u , and x_{-u} the $(d - |u|)$ -dimensional vector whose components are indexed by the complement of u , $-u = [d] \setminus u$. For a single factor i , we have $u = i$ and we write the all-but-one set $[d] \setminus \{i\}$ as $-i$. The total effect of X_u is defined as [Homma and Saltelli, 1996]

$$\tau_u = \mathbb{E}[\mathbb{V}[Y|X_{-u}]] = \mathbb{V}[Y] - \mathbb{V}[\mathbb{E}[Y|X_{-u}]]. \quad (1)$$

The literature has also introduced the normalized total effect as $T_u = \tau_u / \mathbb{V}[Y]$. Using an argument of Fréchet [1934], we find the following useful equalities.

Lemma 1. *Let Y' be a replicate of Y conditionally independent on X_{-u} , i.e., Y and Y' have same distribution and satisfy $\Pr(Y \cdot Y'|X_{-u}) = \Pr(Y|X_{-u}) \Pr(Y'|X_{-u})$. Then,*

$$\tau_u = \mathbb{V}[Y] - \text{cov}(Y, Y') = \frac{1}{2} \mathbb{E}[(Y - Y')^2]. \quad (2)$$

Proof. The proof is postponed to Appendix A.1. \square

The first equality in (2) substitutes the possibly high-dimensional nonlinear regression $\mathbb{E}[Y|X_{-u}]$ in (1) with a covariance operation. When replacing the regression surface by Y' , the error term $Y' - \mathbb{E}[Y|X_{-u}]$ is uncorrelated to Y , because $\text{cov}(Y, Y' - \mathbb{E}[Y|X_{-u}]) = 0$. The second equality can be interpreted as a generalization of Jansen's equality for total effects [Jansen et al., 1994, Jansen, 1999] that does not require feature independence.

In simulation and machine learning Y is often a function of X , $Y = g(X)$, $g : \mathcal{X} \rightarrow \mathbb{R}$. Suppose that g is square integrable, and that it can be decomposed as

$$g(x) = g_0 + \sum_{u \in 2^{[d]}, u \neq \emptyset} g_u(X_u), \quad (3)$$

with $2^{[d]}$ the power set of $[d]$, $g_0 = \mathbb{E}[Y]$, and $g_u(x_u) = \mathbb{E}[Y|X_u = x_u] - \sum_{v \subsetneq u} g_v(x_v)$. Under input independence, we can expand $\mathbb{V}[Y]$ via the well-known functional ANOVA decomposition [Efron and Stein, 1981, Sobol', 1993, Oakley and O'Hagan, 2004, Sun et al., 2021]

$$\mathbb{V}[Y] = \sum_{u \in 2^{[d]}, u \neq \emptyset} \mathbb{V}[g_u(X_u)] \quad (4)$$

with $\mathbb{V}[g_u(X_u)]$ the variance of $g_u(X_u)$ in (3). In Homma and Saltelli [1996], Saltelli and Tarantola [2002], the total effect of input X_j is defined as the sum of all terms in the right hand side of (4) that contain index j and it is shown that

$$\tau_u = \mathbb{E}[\mathbb{V}[Y|X_{-u}]] = \mathbb{V}[Y] - \mathbb{V}[\mathbb{E}[Y|X_{-u}]] = \sum_{v \in 2^{[d]}, v \cap u \neq \emptyset} \mathbb{V}[g_v(X_v)]. \quad (5)$$

However, this identity does not hold if features are statistically dependent. Under dependence, τ_u remains defined as in (1) and enjoys an interpretation in terms of the L^2 approximation error, as established in Hart and Gremaud [2018]. In an argument similar to Rabitz and Alış [1999], Hart and Gremaud [2018] consider that the space L^2 can be decomposed into a direct sum $L^2(\mathcal{X}) = M_{-u} \oplus M_{-u}^\perp$ where M_{-u} contains all L^2

functions which solely depend on x_{-u} and M_{-u}^\perp is its orthogonal complement. In general, for dependent features, $M_{-u}^\perp \neq M_u$. Then we can write $g(x) = g_0 + g_{-u}(x_{-u}) + g_{-u}^\perp(x)$ with $g_{-u} \in M_{-u}$ and $g_{-u}^\perp \in M_{-u}^\perp$. If we ask the question of how accurately $g(x) - g_0$ can be approximated without the features in x_u , then the answer is

$$\|g - g_0 - g_{-u}\|_{L^2}^2 = \|g_{-u}^\perp\|_{L^2}^2 = \|g - g_0\|_{L^2}^2 - \|g_{-u}\|_{L^2}^2. \quad (6)$$

If we consider the L^2 norm weighted with the density of X , then we regain (1) from (6), because $g_{-u}(x_{-u}) = \mathbb{E}[g(X) - g_0 | X_{-u} = x_{-u}]$.

Conditional independence plays a central role in deriving (2): it is this property that enables one to replace $\mathbb{E}[Y | X_u]$ by Y' in (2). We show that it also plays a central role in estimating τ_u under input dependence via winding stairs and pick-and-freeze designs.

2.2 Conditional Independence and Total Effect Estimation

The gold standard for obtaining estimates for total effects *under input independence* is the Sobol' method, i.e., a pick-and-freeze design paired with Jansen's estimator [Jansen, 1999]. Let X, X' be d -dimensional input vectors. For $u \subseteq [d]$, we use the notation $X'_u : X_{-u}$ to denote the d -dimensional vector whose components indexed by u are taken from X' and whose components indexed by $-u$ are taken from X . Now, let X'_u be a replicate of X_u , independent of X_u conditionally on X_{-u} . Then, $(X'_u : X_{-u})$ and X are identically distributed and $Y = g(X)$ and $Y' = g(X'_u : X_{-u})$ are identically distributed and conditionally independent given X_{-u} . The second equality in Equation (2) can then be rewritten as

$$\tau_u = \frac{1}{2} \mathbb{E} \left[(g(X) - g(X'_u : X_{-u}))^2 \right]. \quad (7)$$

By Lemma 1, Equation (7) is true even under feature dependence. However, independence makes the design of an estimator for τ_u straightforward. One generates two independent samples of size n from the input distribution. Let us denote them by $X^A = (X^{A,i})_{i=1,\dots,n}$ and $X^B = (X^{B,i})_{i=1,\dots,n}$. The columns of these sample matrix blocks are recombined, copying input realizations for factor(s) $j \in u$ from the second sample (B) into the first sample (A) to form pick-and-freeze input sample blocks $X_u^B : X_{-u}^A$. The model is then evaluated to obtain the output samples $Y^A = g(X^A)$ and $Y_u^{BA} = g(X_u^B : X_{-u}^A)$. Combining them via Jansen's equality, we obtain the estimator

$$\hat{\tau}_u^{\text{PF}} = \frac{1}{2n} \sum_{i=1}^n \left(g(X^{A,i}) - g(X_u^{B,i} : X_{-u}^{A,i}) \right)^2 = \frac{1}{2n} \left(Y^{A,i} - Y_u^{BA,i} \right)^2. \quad (8)$$

After the introduction of this design in Sobol' [1993], Homma and Saltelli [1996], works such as Saltelli et al. [2000], Sobol' et al. [2007], Gatelli et al. [2009], Gamboa et al. [2016], Prieur and Tarantola [2017] have developed it further refining several aspects. Most of these works rely on the independence assumption, while we remove it in the remainder of this section.

Under conditional independence not only the conditional probability measure can be written in product form, but the joint density also factors into the product of two terms, see Appendix A.1 for further details. As a direct consequence for the Jansen's estimator, when considering the joint distribution of X_u, X'_u and X_{-u} , we obtain the following result.

Proposition 2. *Let X' be a replicate of X , conditionally independent given X_{-u} . Letting $Y = g(X_u : X_{-u})$ and $Y' = g(X'_u : X_{-u})$, then Y is a replicate of Y' conditionally independent given X_{-u} . Written in density terms, we find two interchangeable representations of the total effect,*

$$\begin{aligned} \tau_u &= \frac{1}{2} \int_{\mathbb{R}^{d+|u|}} (g(x'_u : x_{-u}) - g(x_u : x_{-u}))^2 \cdot f_{u|-u}(x_u | x_{-u}) f(x'_u : x_{-u}) dx'_u dx_u dx_{-u} \\ &= \frac{1}{2} \int_{\mathbb{R}^{d+|u|}} (g(x_u : x_{-u}) - g(x'_u : x_{-u}))^2 \cdot f_{u|-u}(x'_u | x_{-u}) f(x_u : x_{-u}) dx'_u dx_u dx_{-u}. \end{aligned} \quad (9)$$

The second term in (9) is the numerator of Equation (2.11) in [Kucherenko et al., 2012, p. 939]. Equation (9) is an essential ingredient for the estimation of total Sobol' indices under feature dependence. In the next section, we exploit Equation (9) to create a generalized design. In Section 3, we use it for the definition of total effect estimators in the presence of constrained (i.e., non-Cartesian) input domains.

2.3 Winding stairs for Dependent Inputs with Gaussian Copula

We propose a new estimation strategy that combines Proposition 2 with the Knothe-Rosenblatt transformation [Knothe, 1957, Rosenblatt, 1952]. This transformation is proposed in the works of Mara and Tarantola [2012], Mara et al. [2015], Li and Rabitz [2017] in association with the challenging task of computing

variance-based sensitivity indices in the presence of dependent features. Formally, the Knothe-Rosenblatt transformation implies the following equations:

$$\begin{aligned} U_1 &= F_1(X_1), \\ U_2 &= F_{2|1}(X_2|X_1), \\ &\dots \\ U_d &= F_{d|1,\dots,d-1}(X_d|X_1, \dots, X_{d-1}), \end{aligned} \tag{10}$$

where $F_1(X_1)$ is the marginal cumulative distribution function of X_1 , $F_{2|1}(X_2|X_1)$ the conditional distribution function of X_2 given X_1 , etc. The transformation maps the dependent set of features \mathbf{X} into the independent set of variables U uniformly distributed in the unit hypercube. As a result, the transformed features U are independent. Considering the mapping from U to Y one can apply the theory and algorithms of the functional ANOVA expansion under independence. One can calculate global sensitivity indices on the independent coordinates in U . However, the transformation has two main drawbacks. The transformation is not unique, as there are as many transformations as the possible orderings of the features. The values of the global sensitivity measures and the corresponding ranking are then dependent on the chosen order. Moreover, results hold for the transformed variables U and reinterpreting results back on the X features is not straightforward.

We propose an intuition to use the Knothe-Rosenblatt transformation for calculating total indices that avoids the rank dependence on the feature ordering and allows us to remain within the original feature space. The key is to combine these two facts. The first is that, by Proposition 2, the total effects τ_j are associated with the conditional density $f_{j|-j}$. The second is that this coincides with the density of the last term in (10). Then, if this last term is available from the transformation, we can draw realizations of an independent standard uniform random variable and apply the inverse transformation $t_j : u \mapsto x_j = F_{j|-j}^{-1}(u|X_{-j} = x_{-j})$. This transformation is, indeed, the inverse of the last term in (10) (up to a re-ordering of input variables): we have an X_j which is conditional on all the remaining features.

We exploit this fact to introduce a generalized winding stairs total effect estimator for the case of dependent features. The term winding stairs originates with Jansen et al. [1994]. Please see also Chan et al. [2000] and Owen and Hoyt [2021] for further reviews. Assume that $X^{(0)}$ is a random copy of X , and U is a random vector of d independent standard uniformly distributed random variables, independent of $X^{(0)}$. In the classical winding stairs design, under independence, the j^{th} column in the feature sample matrix is replaced by an independent copy of X_j . Under dependence, using Lemma 1, we can cyclically replace the j^{th} entry in the input vector with a conditionally independent one. A way to obtain this conditionally independent sample is by a Knothe-Rosenblatt transformation of the following form: In the j^{th} step, the j^{th} component of $X^{(0)}$ is altered via

$$X_\ell^{(j)} = \begin{cases} F_{j|-j}^{-1}(U_j | X_{-j}^{(j-1)}), & \text{for } \ell = j, \\ X_\ell^{(j-1)}, & \text{otherwise.} \end{cases} \tag{11}$$

for $\ell, j = 1, 2, \dots, d$. When sampling, we obtain blocks of the type $X^{(j)} = (X_j^{(j)} : X_{-j}^{(j-1)})$ for $j = 1, \dots, d$. By construction, $Y^{(j)} = g(X^{(j)})$ is a replicate of $Y^{(j-1)}$ conditionally independent given X_{-j} . By Lemma 1, we obtain the following winding stairs total effect expression:

$$\tau_j^{\text{WS}} = \frac{1}{2} \mathbb{E} \left[\left(Y^{(j)} - Y^{(j-1)} \right)^2 \right], \quad j \in [d]. \tag{12}$$

The associated estimator is a variant of (8). As observed in Goda [2021], the winding stairs estimator is a sample average, so that the empirical variance of $\frac{1}{2} \left(Y^{(j)} - Y^{(j-1)} \right)^2$ is approximating the variance of the estimator $\hat{\tau}_j^{\text{WS}}$ in (12).

The computational cost associated with the calculation of a global sensitivity measure is expressed in terms of the required number of model evaluations. For a winding stairs design, the cost is $n(d+1)$, where n is the sample size and d is the input dimensionality (see the first row in Table 1, which reports the computational cost of all the estimators discussed in this work). The cost corresponds to the fact that for each of the n sampled locations we vary the inputs one-at-a-time.

In general, closed-form expressions for the Rosenblatt transformation are unavailable. However, analytical formulas exist when the input dependence is modeled via Gaussian copulas. Specifically, let Ψ_j , $j = 1, \dots, d$ be transformations from the marginal distribution of each input into the standard normal distribution and let Z_j be a standard normal random variable independent of X . Then there exist linear combinations such that the random vectors $[\Psi_1(X_1) \dots \Psi_d(X_d)]$ and

$$\left[\Psi_1(X_1) \dots \gamma_j^{(j)} Z_j + \sum_{\ell \neq j} \gamma_\ell^{(j)} \Psi_\ell(X_\ell) \dots \Psi_d(X_d) \right] \tag{13}$$

Table 1: Computational costs for the estimation of all d total effects with constrained inputs, block sample size n .

Design	Reference	Cost
Generalized Winding Stairs	Equation (12)	$n(d+1)$
Mix-and-reweight, U -statistics	Proposition 5	$d \cdot n^2 - d \cdot n + n$
Shift-and-reweight	Theorem 8	$n(d+1)$
Derange-and-reweight	Theorem 10	$n(d+1)$
Reweight with nearest-neighbour	Equation (26)	n

are identically $\mathcal{N}(0, \Sigma)$ distributed and conditionally independent given X_{-j} . These linear combinations can be extracted from a Cholesky decomposition of a reordered covariance matrix where the j^{th} row/column is moved to the last position (see the proof of Theorem 14 in Appendix A.5 for the computation of the coefficients $\gamma_\ell^{(j)}$ in the linear combination in (13)). Hence (11) becomes

$$X_j^{(j)} = \Psi_j^{-1} \left(\gamma_j^{(j)} Z_j + \sum_{\ell \neq j} \gamma_\ell^{(j)} \Psi_\ell \left(X_\ell^{(j-1)} \right) \right) = t_j(\Phi^{-1}(Z_j) | X_{-j}^{(j-1)}), \quad (14)$$

where Φ is the standard normal cumulative distribution function. From a numerical viewpoint, the computational cost associated with the winding stairs approach amounts at $n(d+1)$ model evaluations. This cost is explained as follows: we sample n random values of X and then consider one-at-a-time variations in each input for each of the n values.

This design generalizes a winding stairs approach to dependence structures that include the broad family of Gaussian copulas. However, for cases in which the Knothe-Rosenblatt transformation is not available, the generalized winding stairs design becomes impractical. This happens as soon as features do not leave on a support which is Cartesian. We then introduce alternative approaches that allow to generalize Lemma 1 to more complex dependence structures in the next sections.

3 Total Effects under Feature Dependence via Reweighting

Our purpose in this section is to introduce an estimation strategy that allows us to relax the traditional condition of a Cartesian domain, that is, we allow for $\mathcal{X} \neq \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_d$. We assume that the features are distributed with a joint density f such that $f(x) > 0$ if $x \in \mathcal{X}$ and $f(x) = 0$ if $x \notin \mathcal{X}$. In order to use an estimation strategy with a classical pick-and-freeze design, we start with the following definition.

Definition 3. Let X' be an independent copy of X . We call the function

$$\iota_u(X', X) = \frac{f(X'_u : X_{-u})}{f_u(X'_u) f_{-u}(X_{-u})} \quad (15)$$

the density quotient of X on \mathcal{X} for the feature list u .

By Proposition 2, the density quotient in (15) satisfies

$$\iota_u(X', X) = \frac{f_{u|-u}(X'_u | X_{-u})}{f_u(X'_u)} = \frac{f_{-u|u}(X_{-u} | X'_u)}{f_{-u}(X_{-u})} = \iota_{-u}(X, X').$$

To illustrate, for a Cartesian domain and independent inputs, we have $\iota_u(X', X) = 1$. Also, we have a compact expression for the case in which the dependence among two features can be expressed via a Gaussian copula.

Example 4. Under a bivariate Gaussian copula, the density quotient for pairwise dependence can be obtained in a compact form as follows. Let X_i and X_j be two random variables with a rank-correlation of ϱ . Setting $u_i = F_i(x_i)$ and $u_j = F_j(x_j)$, [Joe, 2014, Section 4.3.1] derives the density of the bivariate Gaussian copula as

$$\iota_i(u_i, u_j; \varrho) = \frac{\phi(\Phi^{-1}(u_i), \Phi^{-1}(u_j); \varrho)}{\phi(\Phi^{-1}(u_i))\phi(\Phi^{-1}(u_j))} = \frac{1}{\sqrt{1-\varrho^2}} e^{\frac{-\varrho}{2(1-\varrho^2)}(\varrho(z_i^2 + z_j^2) - 2z_i z_j)}, \quad (16)$$

where one uses the transformation $z_i = \Phi^{-1}(u_i) = \Phi^{-1}(F_i(x_i))$. The right hand side in (16) is the density quotient for a bivariate Gaussian copula.

Proposition 5. Let X and X' be i.i.d. random vectors. The following equality holds:

$$\tau_u = \mathbb{E} \left[\iota_u(X', X) \left(g(X_u : X_{-u}) - g(X'_u : X_{-u}) \right)^2 \right]. \quad (17)$$

Given n independent copies X^i of X , $i = 1, 2, \dots, n$, then an unbiased estimator of τ_u is

$$\hat{\tau}_{u,n}^U = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \frac{f(X_u^j : X_{-u}^i)}{f_u(X_u^j) f_{-u}(X_{-u}^i)} \cdot \left(g(X_u^i : X_{-u}^i) - g(X_u^j : X_{-u}^i) \right)^2. \quad (18)$$

Proof. See Appendix A.2 for the proofs of all results stated in this section. \square

Equation (18) combines a brute-force double-loop sample design together with Jansen's estimator modified by an importance sampling weight. The Jansen's approach would yield an estimator based on a pick-and-freeze design. However, a pure pick-and-freeze sample would imply generating the data from the product of their marginal distributions ignoring the probabilistic dependence generated by the presence of constraints. The density quotient in Proposition 5 introduces a correction factor that allows one to generate the data from the (correct) joint distribution. Moreover, the density quotient allows us to consider non-Cartesian input domains, as it vanishes for points outside the region where the inputs are defined. If features are independent we regain Jansen's classical estimator, because the density quotient is identically equal to one.

Lemma 6. *The mix-and-reweight estimator (18) of Proposition 5 is a U -statistic of order 2.*

We call the design associated with Proposition 5 mix-and-reweight. This design is associated with a computational cost of $dn^2 - nd + n$ evaluations (Table 1, second row). The cost depends quadratically on n , while the cost of the winding stairs design depends linearly on it. We can derive a central limit theorem for the mix-and-reweight estimator from the general theory of U -statistics [Hoeffding, 1948].

Lemma 7. *Let $\delta_1 = \mathbb{V}[\mathbb{E}[\Phi^s(W_1, W_2)|W_1]]$ and $\delta_2 = \mathbb{V}[\Phi^s(W_1, W_2)]$ with Φ^s defined by (51) and (52). Assume that $\mathbb{V}[\delta_2] < +\infty$. Then*

$$\mathbb{V}[\hat{\tau}_{u,n}^U] = \frac{2}{n(n-1)} (2(n-2)\delta_1 + \delta_2) = \frac{4}{n}\delta_1 + O(n^{-2}). \quad (19)$$

If $\delta_1 \neq 0$ then the U statistic is non-degenerate and $\sqrt{n}(\hat{\tau}_{u,n}^U - \tau_u) \rightarrow \mathcal{N}(0, 4\delta_1)$.

Equation (19) in Lemma 7 provides us with the asymptotic variance of the mix-and-reweight estimator. Empirically, one way to calculate this variance is to make use of the plug-in Jackknife estimator by Helmers [1991] (see also [Bose and Chatterjee, 2018, p. 106]), defined as

$$\hat{\mathbb{V}}[\hat{\tau}_{u,n}^U] = \frac{4}{n} \frac{n-1}{(n-2)^2} \cdot \sum_{i=1}^n \left(\frac{1}{n-1} \sum_{j \neq i} \Phi^s(X_u^i, X_{-u}^j) - \hat{\tau}_{u,n}^U \right)^2. \quad (20)$$

Alternatively, we can derive a bootstrap distribution of the estimator $\hat{\tau}_{u,n}^U$ from the sample of $\Phi^s(X_u^i, X_{-u}^j)$.

We find earlier accounts on the use of reweighting techniques in sensitivity analysis. Let us mention the estimation of first-order and total Sobol' indices in Sparkman et al. [2016] in which the already available sample of simulations is reweighted with importance sampling. Moreover, in [Badea and Bolado, 2008, Section 5.4], the authors discuss reweighting and rejection techniques to measure the potential impact of small changes in the input probability distribution on the output mean. In Kucherenko et al. [2017], a rejection technique to handle non-Cartesian input domains is implemented.

4 Derangement and Shift Estimators

The mix-and-reweight estimator of Lemma 7 possesses the clear theoretical advantages associated with the notion of U -statistics. However, the associated estimation cost may turn into a notable disadvantage in practical applications. To reduce the cost for estimating τ_u under input constraints, we propose two new estimators based on derangements and shifts. The intuition here is to take the difference between a given realization and another one (for instance randomly picked) instead of taking the differences against all other realizations.

We proceed as follows. Given a sample of size n , we introduce the cyclic shift-by-one of $\{1, \dots, n\}$ defined by $s_n(i) = i + 1$ for $i < n$, and $s_n(n) = 1$. We also define the acyclic shift-by-one by $s_{n-1}^a(i) = i + 1$ for $i < n - 1$, and $s_{n-1}^a(n - 1) = n$. Here, $s_{n-1}^a(\cdot)$ is a fixpoint-free map from $\{1, \dots, n - 1\}$ to $\{2, \dots, n\}$. Based on this idea, we introduce the shift-and-reweight total effect estimator for input j given by

$$\hat{\tau}_{j,n}^S = \frac{1}{2n} \sum_{i=1}^n \iota_j(X^{s_n(i)}, X^i) \left(g(X_j^i : X_{-j}^i) - g(X_j^{s_n(i)} : X_{-j}^i) \right)^2. \quad (21)$$

In the next result, we prove a central limit theorem for this estimator.

Theorem 8. Let X be a sample of size n subject to the joint density f . Then, $\hat{\tau}_{j,n}^S$ is an unbiased estimator of τ_j . In addition, assume that $\sigma_j^2 = \mathbb{E}[V_j^1 V_j^1] + 2 \mathbb{E}[V_j^1 V_j^2] < +\infty$ with

$$V_j^i = \frac{1}{2} \iota_j(X^{s_n(i)}, X^i) \left(g(X_j^i : X_{-j}^i) - g(X_j^{i+1} : X_{-j}^i) \right)^2.$$

Then $\hat{\tau}_{j,n}^S$ defined by (21) satisfies the following central limit theorem:

$$\sqrt{n} \left(\hat{\tau}_{j,n}^S - \tau_j \right) \xrightarrow[n \rightarrow +\infty]{} \mathcal{N}(0, \sigma_j^2). \quad (22)$$

Proof. See Appendix A.3 for the proofs of results stated in this section. \square

Theorem 8 holds for the unnormalized version of the total index estimator. For the normalized version, let us define the shift-and-reweight normalized total effect estimator as

$$\hat{T}_j^S = \frac{\hat{\tau}_{j,n}^S}{\hat{S}_{Y,n}}, \quad (23)$$

where $\hat{S}_{Y,n}$ is the empirical n -sample variance of Y . A central limit theorem for the normalized estimator in (23) can be deduced using the so-called δ -method.

Corollary 9. Let X be a sample of size n subject to the joint density f . Assume $\mathbb{V}[Y] < +\infty$ and $\sigma_j^2 < +\infty$, with σ_j^2 defined in Theorem 8. Then we have:

$$\sqrt{n} \left(\hat{T}_{j,n}^S - T_j \right) \xrightarrow[n \rightarrow +\infty]{} \mathcal{N}(0, \sigma_{\text{norm},j}^2)$$

where $\sigma_{\text{norm},j}^2 = \rho_j^T \Sigma_j \rho_j$ with $\rho_j = \frac{1}{\mathbb{V}[Y]} [1, 2T_j \mathbb{E}[Y], -T_j]^T$ and

$$\Sigma_j = \begin{pmatrix} \mathbb{V}[V_j^1] & \text{cov}(V_j^1, Y^1) & \text{cov}(V_j^1, (Y^1)^2) \\ * & \mathbb{V}(Y^1) & \text{cov}(Y^1, (Y^1)^2) \\ * & * & \mathbb{V}[(Y^1)^2] \end{pmatrix}.$$

Theorem 8 deals with a cyclic shift-by-one strategy. However, the analyst may consider more general shifts or permutations. To define the corresponding estimator, we proceed as follows. Let $(\pi_n)_{n \geq 1}$ be a sequence of derangements (fixpoint-free permutations) of $\{1, \dots, n\}$. Then the derange-and-reweight total effect estimator for input j is defined as

$$\hat{\tau}_{j,n}^D = \frac{1}{2n} \sum_{i=1}^n \iota_j(X^{\pi_n(i)}, X^i) \left(g(X_j^i : X_{-j}^i) - g(X_j^{\pi_n(i)} : X_{-j}^i) \right)^2. \quad (24)$$

Theorem 10. Let X be a sample of size n subject to the joint density f . Then, the derange-and-reweight estimator in (24) is unbiased. In addition, suppose that there exists $\delta > 0$ such that $\mathbb{E}[|V_j^1|^{2+\delta}] < +\infty$ with V_j^1 defined as in Theorem 8 and $\lim_{n \rightarrow +\infty} m_n^{1+\delta} n^{-\delta/2} \rightarrow 0$, with m_n the number of cycles of π_n . Then we have the following central limit theorem:

$$\sqrt{n} \left(\hat{\tau}_{j,n}^D - \tau_j \right) \xrightarrow[n \rightarrow +\infty]{} \mathcal{N}(0, \sigma_j^2) \quad (25)$$

where $\sigma_j^2 = \mathbb{V}[V_j^1] + 2 \text{cov}(V_j^1, V_j^2)$ with

$$V_j^i = \frac{1}{2} \frac{f_{j,-j}(X_j^{i+1} : X_{-j}^i)}{f_j(X_j^{i+1})f_{-j}(X_{-j}^i)} \left(g(X_j^i : X_{-j}^i) - g(X_j^{i+1} : X_{-j}^i) \right)^2.$$

Theorem 10 proves a central limit result for the derange-and-reweight estimator (24).

Remark 11. The assumption that in the limit $\lim_{n \rightarrow +\infty} m_n^{1+\delta} n^{-\delta/2} \rightarrow 0$ holds does not seem too technical. Indeed, a permutation π_n of $\{1, \dots, n\}$ decomposes into cycles and a classical result in combinatorics lets us expect $1 + \frac{1}{2} + \dots + \frac{1}{n}$ cycles per permutation, and this harmonic series is approximately $\log(n)$.

Remark 12. It is possible to compute confidence intervals by block-bootstrapping. It is important to implement bootstrapping with blocks, in order to preserve the 1-dependence structure (see, e.g., Lahiri [2003] and references therein).

The shift-and-reweight and the derange-and-reweight estimators are associated with a computational cost equal to $n(d+1)$, the same as the winding stairs approach (Table 1, rows three and four).

5 Nearest-Neighbour Estimation

The designs we have introduced thus far (listed in rows 1 to 4 of Table 1) are based on Jansen’s intuition. We observe that, in a simulation setting, the costs in Table 1 are an upper bound: whenever the density quotient vanishes, there is no need to evaluate the model at those coordinates and we can save computational time. However, the estimators built on these designs are not given data. Indeed, given a design X and the evaluations of the model at X , the computation of all the estimators listed in rows 1 to 4 of Table 1 requires the simulator to be run at new design points (e.g., for the shift estimator, at points defined as $X_j^{s_n(i)} : X_{-j}^i$, $j = 1, \dots, n$). This is the reason why, while the costs of the most efficient estimators are linear in the sample size, they depend on the number of features (d) and thus can be exposed to the curse of dimensionality. We propose below a nearest-neighbour approach to get rid of the dependence of the cost in the input space dimension d . The construction is as follows. Consider two observations x and x' , and consider the recombined point $(x'_j : x_{-j})$. The first step is the evaluation of the density quotient $\iota_j(x', x)$ at this point. If the density quotient is non-negligible, then select the point which is closest to $(x'_j : x_{-j})$ with respect to some predefined Euclidean metric from the available data. More precisely, the point is the solution of

$$k^* = \operatorname{argmin}_{k \in [n]} \left\{ \|x_{k,j} - x'_j\|_2^2 + \|x_{k,-j} - x_{i,-j}\|_2^2 \right\}. \quad (26)$$

Hence, we use $g(x'_j : x_{-j}) \approx g(x_{k^*})$ in (18). Otherwise, if the quotient is small, we set the contribution of $\iota_j(x', x) \times (g(x'_j : x_j) - g(x))^2$ equal to zero. This step can be implemented by setting a threshold on the value of the density quotient and considering as negligible all values of the density quotient below the threshold.

All the required information for applying this design is contained in a sample generated by a once-through pass of a Monte Carlo simulation.

The nearest-neighbour approach serves here as a metamodel, predicting model outputs for the mixed input sample. This is a different use of the nearest-neighbour intuition than in Broto et al. [2020], Plischke et al. [2022], where nearest-neighbour is used to select a conditional stratum (see Devroye et al. [2013, 2018] for theoretical results). At this stage, also due the presence of the density quotient threshold, we do not furnish any theoretical results for our estimator based on (26). We will then evaluate this strategy based on empirical comparisons in a series of experiments in which we compare the performance of this design with the other estimators in Table 1.

6 The Link Between Total Indices and Breiman’s Permutation Importance with Feature Constraints

There is a close link between MDA_j and τ_j visible in [Hooker et al., 2021, Theorem 2] and discussed in great detail in B  nard et al. [2022]. In this section, we extend the relationship to the case in which inputs are constrained. Consider the problem of training an input-output mapping of the type $h(X, \theta)$, $h : \mathcal{X} \times \mathbb{R}^q \rightarrow \mathbb{R}$, where θ is a q -dimensional vector of auxiliary parameters. Let $\mathcal{L}(Y, g(X; \theta))$, $\mathcal{L} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ a loss function. The training problem can be defined as finding

$$\theta^* = \operatorname{argmin}_{\theta} \mathbb{E}[\mathcal{L}(Y, g(X; \theta))]. \quad (27)$$

We let $\mathbb{E}[\mathcal{L}(Y, g(X; \theta^*))]$ denote the nominal (and minimal) expected loss function for the machine learning problem. Then, Breiman’s importance of feature X_j is defined as

$$\text{MDA}_j = \mathbb{E}[\mathcal{L}(Y, g(X'_j : X_{-j}; \theta^*))] - \mathbb{E}[\mathcal{L}(Y, g(X; \theta^*))], \quad (28)$$

where $\mathbb{E}[\mathcal{L}(Y, g(X'_j : X_{-j}; \theta^*))]$ is the expected loss that we incur if feature X_j is permuted (using the model trained on the original data). The intuition is that if the machine learning model relies heavily on X_j for its predictions, then the loss in predictive accuracy should be high and consequently the difference between the nominal loss and the loss after permutation should be significant. After feature permutation we expect a decrease in model prediction accuracy, so that the expected loss after feature permutation is larger than the nominal expected loss, yielding $\text{MDA}_j \geq 0$.

Williamson et al. [2021, 2023] propose the following quantity as a model-agnostic variable importance measure:

$$\text{MDA}_j^W = \mathbb{E}[\mathcal{L}(Y, g_j(X_{-j}; \theta^{**}))] - \mathbb{E}[\mathcal{L}(Y, g(X; \theta^*))], \quad (29)$$

where $g_j(X_{-j}; \theta^{**})$ is the retrained predictor after feature X_j has been eliminated from the dataset and $g(X; \theta^*)$ is the model trained with all features in the dataset. The variable importance measure in (29) is called dropped variable importance in Hooker et al. [2021] and leave-one-out-covariate (LOCO) in Lei et al. [2018]. Notice that, if the model coefficient of determination R^2 is taken as a performance measure in (29), then MDA_j^W is the total index of X_j . The main difference between the approach to variable importance in

(29) and the feature importance in (28) is that the latter permutes the feature and uses the same model, while in the former one deletes the feature from the dataset and then uses a newly retrained model. The advantage of the remove-and-retrain approach is that it handles dependence and constraints. However, in case retraining is expensive the approach might become computationally heavy. The computational issues related to retraining are also noted in Lundberg and Lee [2017], who propose alternatives to avoid retraining in their introduction of the SHAP variable importance measure. Breiman’s approach to permute features and use the same model may then be computationally convenient. However, Breiman’s strategy is exposed to unrestricted permutations [Hooker et al., 2021], which could lead to violating the constraints.

Of relevance to us are also Equations (3.1) and (6.2) of Fisher et al. [2019]. Fisher et al. [2019] formulate Breiman’s importance in terms of model reliance, as a ratio between the expected loss after permutation over the expected loss before permutation. Using our notation, the definition of model reliance in their Equation (6.2) would read

$$\text{MR}_j = \frac{\mathbb{E}[\ell_j(X', X)\mathcal{L}(Y, g(X'_j : X_{-j}; \theta^*))]}{\mathbb{E}[\mathcal{L}(Y, g(X; \theta^*))]}.$$
 (30)

Rewritten in difference terms, (30) yields a dependence-aware version of (28)

$$\text{MDA}_j^{\text{FR}} = \mathbb{E}[\ell_j(X', X)\mathcal{L}(Y, g(X'_j : X_{-j}; \theta^*))] - \mathbb{E}[\mathcal{L}(Y, g(X; \theta^*))].$$
 (31)

Proposition 13. *Consider MDA_j^{FR} in (31). If $\mathcal{L}(Y, g(X))$ is a squared loss function and $g(X; \theta^*)$ is a perfect predictor, then $\text{MDA}_j^{\text{FR}} = \tau_j$.*

Proof. The proof is postponed to Appendix A.4. \square

Proposition 13 then suggests that, also under feature constraints, total indices can be reinterpreted in terms of mean decrease of a model’s accuracy.

Another popular alternative from the machine learning literature to measure variable importance is the model-X knockoffs from Candès et al. [2018]. Model-X knockoffs provide valid inference from finite samples in settings in which the conditional distribution of the response is unknown, but the input probability distribution (pdf) is known, or at least approximated. The approach relies on the construction of knockoff variables as far as on the proposition of feature statistics allowing false discovery rate control. In the framework of complex input dependence structure and complex input-output realisation these tasks are non trivial.

7 Analytical Total Effects for Linear Models with Gaussian Features

In this section, we derive analytical expressions for the values of total effects to be used as benchmarks in numerical experiments. In doing so, we also derive a formula for the density quotient in the case of Gaussian copulas.

Let us consider a linear mapping between Y and X , $Y = \beta^0 + \beta^T X$, $X \in \mathbb{R}^d$, $\beta^0 \in \mathbb{R}$, $\beta \in \mathbb{R}^d$ and let the features be normally distributed with mean μ and variance-covariance matrix Σ . Under this assumption all conditional distributions are Gaussian and all conditional expectations are linear. The pair (X, Y) is then also Gaussian with augmented covariance matrix, $\Sigma' = \begin{pmatrix} \Sigma & \Sigma\beta \\ \beta^T\Sigma & \Sigma\beta\beta^T + \Sigma\beta\beta^T \end{pmatrix}$.

Let m be a positive integer. Let u, v, w be pairwise disjoint index sets from $[m]$. For any m -dimensional multivariate normal distribution $Z \sim \mathcal{N}(\mu, \Gamma)$, the conditional distribution of Z_{u+v} given $Z_w = z_w$ is

$$\mathcal{N}(\mu_{u+v} + \Gamma_{u+v,w}\Gamma_{w,w}^{-1}(z_w - \mu_w), \Gamma_{u+v,u+v} - \Gamma_{u+v,w}\Gamma_{w,w}^{-1}\Gamma_{w,u+v}).$$
 (32)

Here, the correlation matrix in (32) is given in form of a Schur complement. We assume that the submatrix selected by w is invertible.

For a multivariate Gaussian distribution, the conditional independence $u \perp\!\!\!\perp v|w$ holds if and only if $\Gamma_{u,v} = \Gamma_{u,w}\Gamma_{w,w}^{-1}\Gamma_{w,v}$, as in this case the correlation matrix in (32) is block-diagonal, i.e.

$$\begin{bmatrix} \Gamma_{u,u} & \Gamma_{u,v} \\ \Gamma_{v,u} & \Gamma_{v,v} \end{bmatrix} - \begin{bmatrix} \Gamma_{u,w} \\ \Gamma_{v,w} \end{bmatrix} \Gamma_{w,w}^{-1} \begin{bmatrix} \Gamma_{w,u} & \Gamma_{w,v} \end{bmatrix} = \begin{bmatrix} \Gamma_{u,u} - \Gamma_{u,w}\Gamma_{w,w}^{-1}\Gamma_{w,u} & 0 \\ 0 & \Gamma_{v,v} - \Gamma_{v,w}\Gamma_{w,w}^{-1}\Gamma_{w,v} \end{bmatrix}.$$

Theorem 14. *Given $Y = \beta^0 + \beta^T X$, $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^d$, the unnormalized main and total effects are given by*

$$S_j = \beta^T \left(\frac{\Sigma_{[d],j}\Sigma_{[d],j}^T}{\Sigma_{j,j}} \right) \beta,$$
 (33)

$$T_j = \beta_j^2 \frac{\det(\Sigma)}{\det(\Sigma_{-j,-j})}.$$
 (34)

The output variance is $\mathbb{V}[Y] = \beta^T \Sigma \beta$.

Proof. The proof is postponed to Appendix A.5. \square

Alternative computations are offered in Mara and Tarantola [2012] for the case $d = 3$. These results can also be retrieved from the proof of [Owen and Prieur, 2017, Theorem 4.1].

The proof of Theorem 14 (see Appendix A.5) provides analytical formulas for main and total effects for linear models with Gaussian features. For the mix-and-reweight and the derange-and-reweight approaches, we obtain the density quotient

$$\iota_u(x', x) = \sqrt{\frac{\det(\Sigma_u) \det(\Sigma_{-u})}{\det(\Sigma)}} e^{-\frac{1}{2}((x'_u : x_{-u}) - \mu)^T (\Sigma^{-1} - (\Sigma_u^{-1} \Sigma_{-u}^{-1})) ((x'_u : x_{-u}) - \mu)}, \quad (35)$$

where $(\cdot : \cdot)$ is the out-of-order composition of vectors and block diagonal matrices. The expression for the density quotient readily generalizes to the case in which the factors are distributed with generic marginals correlated via a Gaussian copula. In the bivariate case, setting $\mu = 0$ and $\Sigma = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}$ in (35) yields (16).

8 Experiments

In this section, we study the numerical implementation of the designs discussed in this work, using examples with constraints of growing complexity. The experiments are organized as follows. In Subsection 8.1, we start with linear models and correlated inputs to test the consistency of all estimators by comparison with the analytical benchmarks developed in Section 7. In Subsection 8.2, we provide a detailed study of reweighting estimators on the Ishigami function, a popular test case for sensitivity analysis. In Subsection 8.3, we study the case of inputs constrained to a circle and total Sobol' indices are estimated with nearest-neighbour and double-loop approaches, because a winding stairs approach becomes infeasible. In subsection 8.5 inputs are constrained by a non-connected domain (two separate triangles). Subsection 8.4 reports results for the case in which inputs are constrained on a simplex, based on the case study in Gilquin et al. [2015]. Subsection 8.6 reports results for the constraint represented by Sierpinski gaskets. Finally, Subsection 8.8 reports results for a realistic simulator. All experiments are run on personal computer with an Intel(R) i7-3770 CPU at 3.40GHz and 8GB RAM, using MATLAB R2022a. We rely on the MATLAB k-d-tree implementation for the nearest-neighbour search. In all the experiments we implemented the nearest-neighbour approach (see Section 5) with the low threshold equal to zero.

Because the methods are associated with different computational costs (third column in Table 1), to make experiments comparable, we fix the overall budget of the experiment and then calculate the corresponding sample size n . To illustrate, given a budget of $B = 10000$ model runs, for a $d = 3$ variable model we have sample sizes respectively of $n^{\text{winding stairs}} = 2500$ for the generalized winding stairs and the shift(derange)-and-reweight designs, and $n^{\text{U-statistic}} = 58$ for the U-statistic design, while the entire budget is available for the nearest-neighbour design $n^{\text{nearest-neighbour}} = B = 10000$.

8.1 Linear Model with Normal and Correlated Inputs

In this section, we consider a parameterization of a linear model with correlated inputs, for which analytical expressions are discussed in Section 7 (Theorem 14). Our goal is to test the performance of all estimators in Table 1. We use as a test case the model discussed in Kucherenko et al. [2012]. One sets $Y = g(X_1, X_2, X_3) = X_1 + X_2 + X_3$ with $X \sim \mathcal{N}(0, \Sigma)$, and correlation matrix $\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \varrho\sigma \\ 0 & \varrho\sigma & \sigma^2 \end{pmatrix}$. In the experiments, we set $\sigma = 2$ and assess the effect of increasing correlations, letting ϱ vary between $[-1, 1]$. We hypothesize a computational budget restriction, with a fixed budget $B = 1680$ and derive back the corresponding basic sample sizes n from Table 1. With $d = 3$ input factors, we find to $n^{\text{U-statistic}} = 24$, $n^{\text{Winding Stairs}} = n^{\text{Reweighting}} = 420$, and $nn^{\text{Nearest-Neighbour}} = 1680$. We perform the calculations with this given budget B for $\varrho \in \{-0.9, -0.75, -0.6, \dots, 0.6, 0.75, 0.9\}$. We also test the effect of changing the sample generator, comparing crude Monte Carlo (MC) and randomized Quasi-Monte Carlo (RQMC). We randomize the sequence generation process employing the MATLAB Sobol' scrambler discussed in Owen [1997], Matoušek [1998]. Figure 1 shows the results. Each of the panels in Figure 1 report the values of the total indices for each input τ_1, τ_2 , and τ_3 , respectively. The horizontal axis reports the values of the correlation coefficients. The analytical values are displayed as continuous lines, and the estimates as dashed lines. Confidence intervals are displayed as shaded areas around the point estimates. To compute them, for the U-statistics approach, we use the asymptotic normality result of Lemma 7 together with plug-in estimates for the estimator variance. For the winding stairs approach we follow Goda [2021]. For the shift-and-reweight approach (panels C and D) and for the derange-and-reweight approach (panels G and H) we use the upper and lower 2.5% quantiles from a block-bootstrap. We also used a normal approximation from the asymptotic results of Theorem 8 or Theorem 10 to compute confidence intervals. The results were similar and are therefore not reported here.

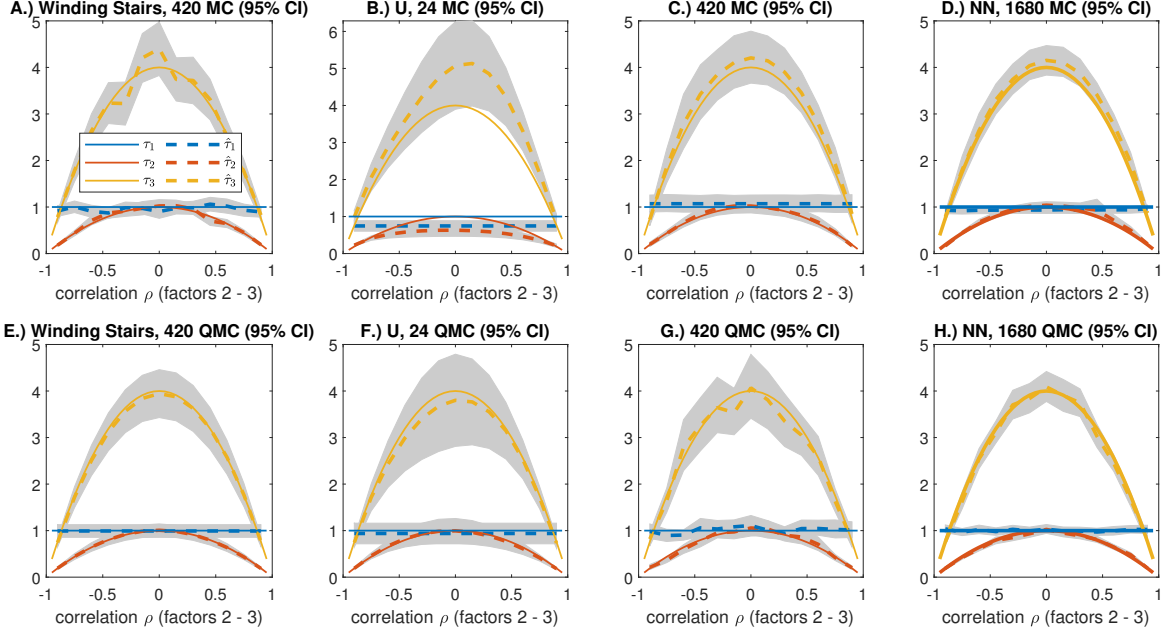


Figure 1: Total effects (unnormalized) for the linear model, including confidence bounds (gray areas): Winding stairs (panels A and E), U -statistics estimator (panels B and F), shift-and-reweight (panels C and D), derange-and-reweight (panels G and H). The analytical total effects are represented by solid lines.

Figure 1 shows that the point estimates from QMC designs (using scrambled sequences) generally perform better than those from a plain Monte-Carlo design, while the confidence bounds are comparable. Preliminary tests showed that the shift-and-reweight approach is not working well with QMC design. Instead, we replaced the shift with a permutation, and a derange-and-reweight approach has been used for panels G and H. Considering the mean squared error for different simulations one can conclude that for this example, a QMC design is an advantage for U -statistics and winding stairs methods, while the impact is not so clear on the derange-and-reweight methods (both with reevaluations and with nearest-neighbour approximations of the mixed sample).

In this example, investing the computational budget into one large sample and using a nearest-neighbour metamodeling approach seems to offer a good performance, only to be beaten by winding stairs QMC design that, as already remarked, is not necessarily available for general dependence in form of constraints of the input features (these visual findings are corroborated by corresponding mean squared errors, see Table 2).

Table 2: Averaged mean squared errors (with respect to ϱ) of the different methods and sampling designs for the Linear Model using 20 repetitions.

Factor	1	2	3
Winding Stairs MC	5.38e-03	2.52e-03	4.09e-02
Winding Stairs QMC	1.40e-04	7.61e-05	1.48e-03
U Statistics MC	6.19e-02	1.99e-02	5.91e-01
U Statistics QMC	1.43e-02	9.88e-03	9.14e-02
Derange MC	1.04e-02	5.02e-03	5.73e-02
Derange QMC	2.32e-03	2.17e-03	3.41e-02
NN Derange MC	3.27e-03	1.28e-03	2.53e-02
NN Derange QMC	7.47e-04	7.41e-04	1.11e-02

8.2 The Ishigami Function with Correlations under a Gaussian Copula

We further test the proposed designs on a well-known test case in simulation experiments [Kucherenko et al., 2012]. The domain is still Cartesian, the input-output mapping, however, presents interactions. Our goal is to obtain additional insights on the performance of the estimators used in the previous case study, and especially to test their behavior with respect to the random sample generator. The input-output mapping

is $Y = g(x_1, x_2, x_3) = \sin(x_1) + 7 \sin^2(x_2) + 0.1x_3^4 \sin(x_1)$ with X_i uniformly distributed on $[-\pi, \pi]$. As in [Kucherenko et al., 2012, Section 7.3], we introduce a statistical dependence between X_1 and X_3 by a pairwise Gaussian copula. We let the rank correlation coefficient $\rho(X_1, X_3)$ range from $\rho(X_1, X_3) = -0.9$ to $\rho(X_1, X_3) = 0.9$. There is no closed-form solution for the total indices at a generic value of $\rho(X_1, X_3)$. However, in the uncorrelated case, $\rho(X_1, X_3) = 0$ the total indices are analytically known, with values $T_1 = 0.56$, $T_2 = 0.44$ and $T_3 = 0.24$, respectively.

For the winding stairs estimator, we use the inverse Knothe–Rosenblatt transformation detailed in Example 4. For the reweight estimators, we implement the rank correlation using (14). We fix a budget of about $B = 8200$ simulations. This yields a basic sample size of $n = 2048$ for the generalized winding stairs and weight and derange approach, of $n = 53$ for the U -statistic and $n = 8192$ points for the nearest-neighbour estimator. We generate the sample first with crude MC and then with QMC.

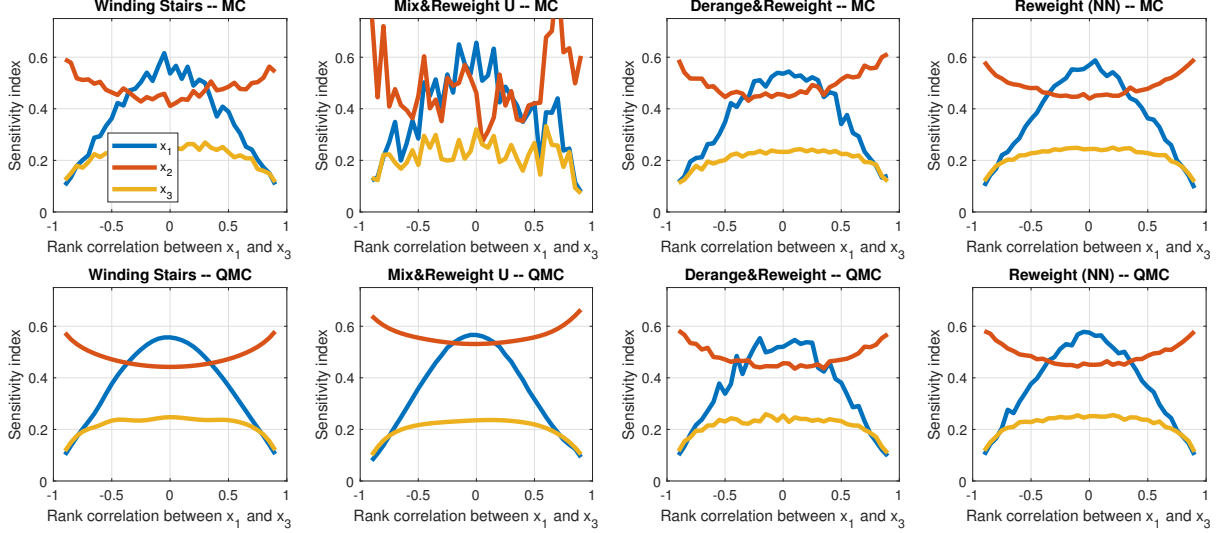


Figure 2: Ishigami function with rank correlation: normalized total (T_i) indices depending on the rank correlation $\rho^*(X_1, X_3)$. Total budget: 8200 model evaluations. Upper row: crude Monte Carlo sampling, lower row: Quasi Monte Carlo sampling.

The panels in Figure 2 show the estimates of the normalized total effects T_1 , T_2 , T_3 as a function of the correlation between X_1 and X_3 for the winding stairs, mix-and-reweight, derange-and-reweight and nearest-neighbour designs, respectively. In the upper row, we use crude MC to generate the input sample, while we use QMC in the lower row. The panels in the first row show that, when using crude MC, the winding stairs (first panel) and derange-and-reweight (third panel) designs yield comparable estimates. However, the mix-and-reweight estimator (second panel) yields highly unstable estimates, while the nearest-neighbour design yields more stable estimates. In the absence of correlations, at $\rho(X_1, X_3) = 0$, the estimates for all designs are close to the analytical values.

The panels in the second row show that, when using QMC, the winding stairs and U -statistic estimates exhibit increased regularity, while methods with a random derangement do not. This can be due to the random derangements breaking the properties of low discrepancy sequences. Despite the greater regularity of the U -statistic estimates as a function of $\rho(X_1, X_3)$, its estimates are upward biased for T_1 and most notably for T_2 , compared to the other approaches. A reason may be that at $n = 53$, the QMC Sobol' sequence does not populate a Latin Hypercube and the projections on the marginals are not uniform. To fill in a Latin Hypercube, we would need to increase the basic sample size to the next power of 2, $n = 64$. However, these additional eleven points in the sample block would propagate into a new budget of $B = 12160$, a nearly 50% increase in computational cost. Lastly, the rightmost panels show that the nearest-neighbour estimator performs similarly with both sample generation methods.

8.3 Features Constrained on a Circle

In the previous two test cases, the input domain was the Cartesian product of the individual input supports and the dependence structure was determined by probabilistic correlations. In this section, we continue the investigation of the performance of the estimators for an example in which the dependence structure cannot be modeled by a Gaussian copula. Specifically, we consider the inputs constrained on a circle. We consider

the two-dimensional input-output mapping

$$Y = g(X_1, X_2) = (X_1 - 1) \cdot (X_2 - 1), \quad (36)$$

with X_1 and X_2 uniformly distributed within a circle of radius π centered at the origin. This geometry rules

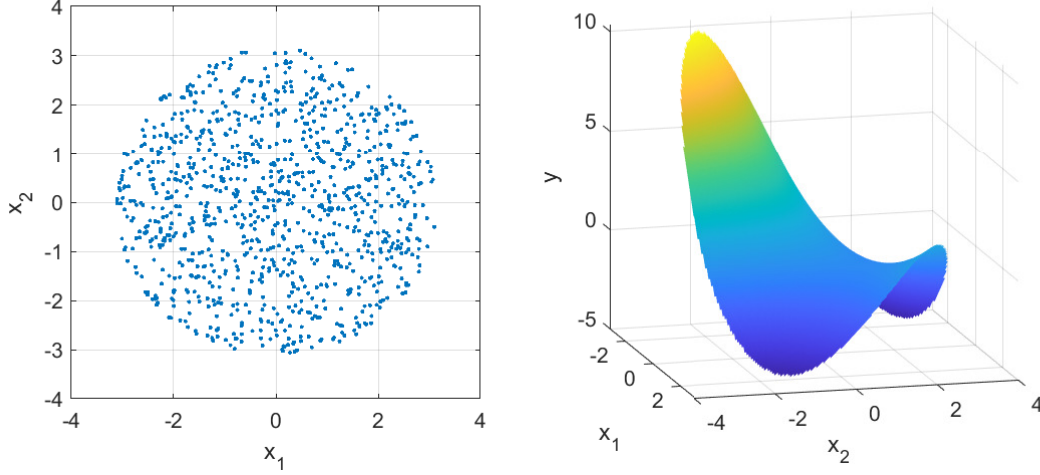


Figure 3: Uniform inputs constrained within a circle (left), model response (right).

out the application of a design based on the Knothe-Rosenblatt transform for calculating total effects. In fact, we would need to find the quantile function corresponding to the marginal cdf of X_1 ,

$$x \mapsto \frac{2}{\pi R^2} \left(\frac{x}{2} \sqrt{\max\{0, R^2 - x^2\}} + \frac{R^2}{2} \arcsin \left(\min \left\{ 1, \max \left\{ -1, \frac{x}{R} \right\} \right\} \right) + \frac{\pi R^2}{4} \right),$$

and plug this into the conditional quantile function of x_2 , $(u, x) \mapsto 2(u - \frac{1}{2})\sqrt{R^2 - x^2}$. Even for this simple geometry, this seems to be a tantalizing task. Conversely, in this case, we can find the density quotient analytically. Assuming a uniform unconstrained density $f_X(x)$ on the square of side $2R$ enclosing the circle of radius R , the joint density is $x \mapsto \frac{\mathbb{1}_C(x)f_X(x)}{\int_C f_X(x)dx}$ and the marginal distributions can be obtained accordingly. For the present test case, where the circle of radius R is centered at the origin, the density quotient is

$$\iota_1(x_1, x_2) = \iota_2(x_2, x_1) = \frac{\pi R^2}{4} \cdot \frac{\mathbb{1}\{x_1^2 + x_2^2 \leq R^2\}}{\sqrt{\max\{0, R^2 - x_1^2\}} \cdot \sqrt{\max\{0, R^2 - x_2^2\}}}. \quad (37)$$

Because the output density can be computed analytically, calculation with symbolic software (Mathcad Prime 8 in our case) yields unnormalized total effects $\tau_1 = \tau_2 = 6.53$. The equality follows by the problem symmetry. We use the experiments to investigate the rate of convergence of the nearest-neighbour and derange-and-reweight estimators. We also test the effect of the random number generator and use both Monte Carlo and quasi-Monte Carlo alternatives in our experiments. Because we have seen a poor performance of the shift-and-reweight estimator together with quasi-Monte Carlo sampling, we do not consider this estimator in these experiments. Figure 3 shows an input sample of size $n = 1024$ (left panel) and the model output response (right panel). For randomized QMC, two approaches are used: the scrambled sequence of Matoušek [1998], Owen [1997] and a “mingled” Halton sequence with linear scrambling [Bayouf and Mascagni, 2019].

The mean squared errors which are averaged over all factors follow a $O(n^{-1})$ convergence rate for all sampling strategies, as seen in Figure 4, where the diagonal of the dashed triangle evidences the n^{-1} convergence rate. The estimators therefore are of rate $O(n^{-1/2})$, as for standard Monte Carlo estimation.

We observe a similar convergence rate across the alternative generators. A reason may be that the shifting strategy interferes with the regularity of the QMC structure, thus reducing the advantage of using a QMC generator in this context. Furthermore, a constraint may introduce discontinuities which are not compatible with functions of bounded variation in the sense of Hardy and Krause and in this case the Koksma-Hlawka Theorem does not provide an improved convergence rate.

8.4 Features Constrained on a Simplex

We have seen in the previous sections that the shift-and-reweight estimator does not combine well with a quasi-Monte Carlo sampling. We propose here a preprocessing consisting in permuting the order of the quasi-Monte Carlo realizations. With such a preprocessing, the mean squared error of the shift-and-reweight

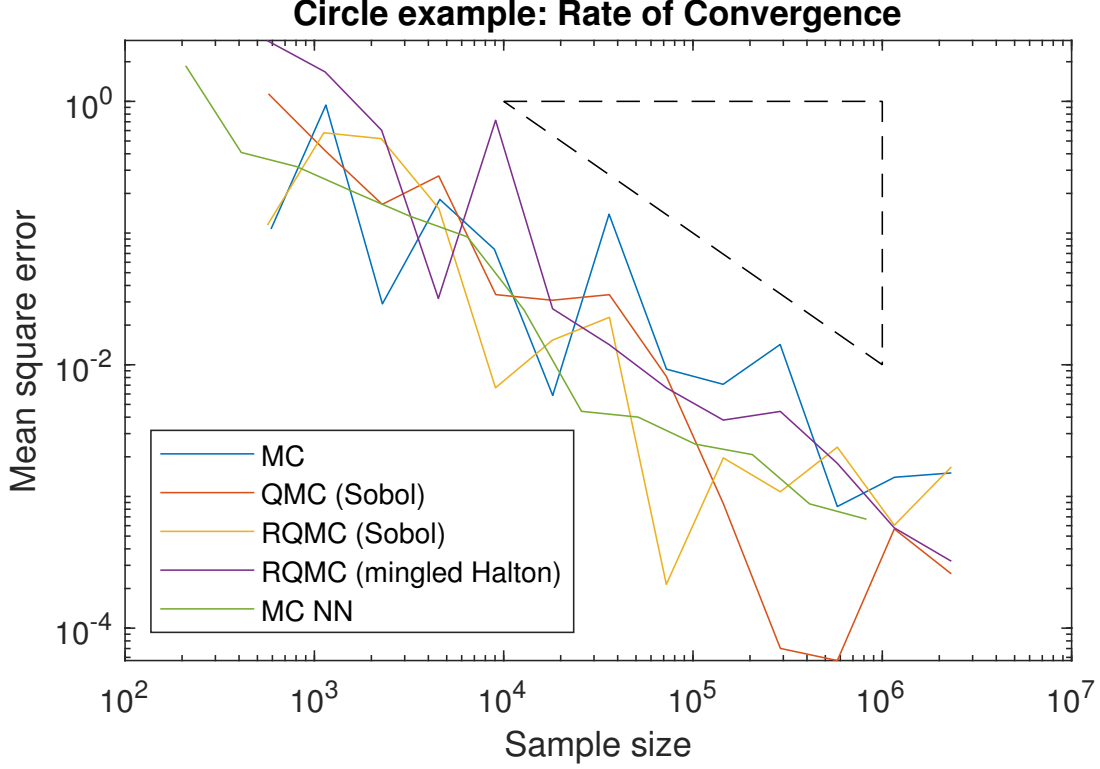


Figure 4: Mean squared errors for the unnormalized total effects over both factors with alternative sample generators, depending on the sample size (number of model evaluations) using basic sample sizes from 2^8 to 2^{20} .

estimator combined with QMC remains consistent with the Monte Carlo approximation error, and regain similar performance to the derange-and-reweight estimator.

We perform a series of tests for a setting in which it is still possible to obtain the total indices analytically via a symbolic calculation software (we used Maple software). More precisely, we consider features constrained on a simplex:

$$\{(x_1, \dots, x_d) \in [0, 1]^d : x_{d-1} \leq x_d\}. \quad (38)$$

Figure 5 provides a visualization of the constraint for the case $d = 3$. The density quotient between all pairs of inputs is constant and equal to one, except for the last two features: $\iota(x_{d-1}, x_d) = \frac{\mathbf{1}\{x_{d-1} \leq x_d\}}{2(1-x_{d-1})x_d}$.

As numerical test cases, we consider two models (with $d = 3$ and $d = 4$) introduced in Gilquin et al. [2015], where the authors tested a new space-filling sampling strategy for estimating grouped Sobol' indices in the framework of constrained inputs (see Jacques et al. [2006]), as well as the well-known Sobol' g -function. These models are described hereafter, together with the corresponding analytical values for total Sobol' indices under the simplex constrained given by (38):

- $g(x) = -x_1 + x_1x_2 - x_1x_2x_3 + x_1x_2x_3x_4$ with $T = (0.6300, 0.4861, 0.0064, 0.0064)$;
- $g(x_1, x_2, x_3) = x_1x_2 - x_3$ with $T = (0.125, 0.125, 0.375)$;
- Sobol' g -function with parameter vector $a = (0, 1, 3, 6)$, such that $T = (0.766, 0.236, 0.052, 0.017)$.

The numerical experiments aim to compare estimates obtained with the shift-and-reweight estimator with data generated either with a Monte Carlo sampling or with a quasi-Monte Carlo sampling (here a scrambled Sobol' sequence) with preprocessing. The preprocessing consists of a random permutation of the sample realizations. We report results for the the mean squared errors (average over 20 replicates) in Figure 6. For comparison, the figure also reports the estimates of the nearest-neighbour estimator with crude Monte Carlo sampling. The results demonstrate that preprocessing restores the performance of the shift-and-reweight estimator also with a quasi-Monte Carlo data-generation. However, there is no advantage in combining a shift-and-reweight estimator with quasi-Monte Carlo sampling rather than with crude Monte Carlo. Moreover, in two out of three examples the nearest-neighbour approach presents the best performance, while for the Sobol' g -functions the three estimators perform similarly, with a slightly better performance for the shift-and-reweight estimator.

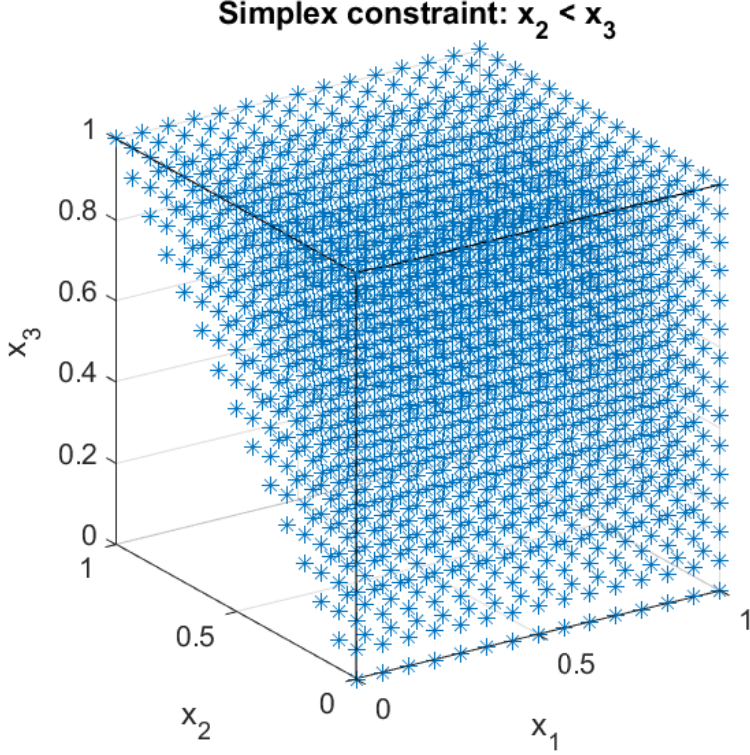


Figure 5: An input constraint in form of a simplex condition.

Table 3: A product model on the support formed by two separate triangles, normalized total effects, shift-and-reweight method of Theorem 8.

Size	Runs	Plain Monte Carlo					Nearest-Neighbour				
		T_1	T_2	$\sigma(T_1)$	$\sigma(T_2)$		T_1	T_2	$\sigma(T_1)$	$\sigma(T_2)$	
		.037*	.27**	10^{-2} .			.037*	.27**	10^{-2} .		
474	960	.0328	.2447	.5575	3.7109	474	.0330	.2445	.5545	3.7247	
701	1445	.0325	.2401	.5041	3.0781	701	.0320	.2396	.5028	3.0652	
990	2063	.0376	.2768	.4063	2.4153	990	.0375	.2768	.4060	2.4113	
1969	3953	.0471	.2592	.3825	1.7471	1969	.0473	.2595	.3858	1.7512	
3522	7138	.0401	.2475	.2603	1.3221	3522	.0400	.2476	.2580	1.3237	

8.5 Features Constrained on Two Disconnected Triangles

In this section, we further challenge the estimators with experiments for a test case in which the inputs are on a disconnected domain. We hypothesize the input-output mapping as $g(x_1, x_2) = (x_1 - 1) \cdot (x_2 - 1)$ and let the features lie in the 2-dimensional region $\mathcal{X} = \{(x_1, x_2) \in [0, 1]^2 : x_2 \leq \frac{1}{2} - x_1 \vee x_2 \leq 1 - \frac{1}{4}x_1\}$ (Figure 7). We assign a uniform density $f_{12} = 4$ within the triangles, which vanishes outside the triangles. It is possible to compute analytically the marginal and conditional densities of the inputs, as well as the density quotients. From this knowledge, the values of the total indices can be analytically obtained and are equal to $T_1 = 0.037$ and $T_2 = 0.27$ (the total variance is 0.117). In the next series of numerical tests, we employ random sampling with rejection, starting from a uniform distribution on the unit square. Figure 7 provides a visualization of the input space, when the data are generated using crude Monte Carlo after sampling 2,000 points, and about $\frac{1}{4}$ of the points are left in the domain.

We perform a series of experiments at increasing sample sizes, from $n = 474$ to $n = 3522$ as reported in Table 3 (due to rejection sampling the sample sizes do not form a regular progression). We then apply the shift-and-reweight, the derange-and-reweight and the nearest-neighbour estimators. Because the first two estimators produce similar results, we only display the values of the shift-and-reweight estimates. The estimator variances in Table 3 are computed from a plug-in estimator as per Lemma 7.

The values in Table 3 show that the estimators exhibit similar accuracy, with a decreasing variance of the estimates. However, the nearest-neighbour estimator is associated with a much lower computational cost. We believe this good performance is allowed by the low dimension of the input space and we are to challenge it in later experiments (Section 8.7).

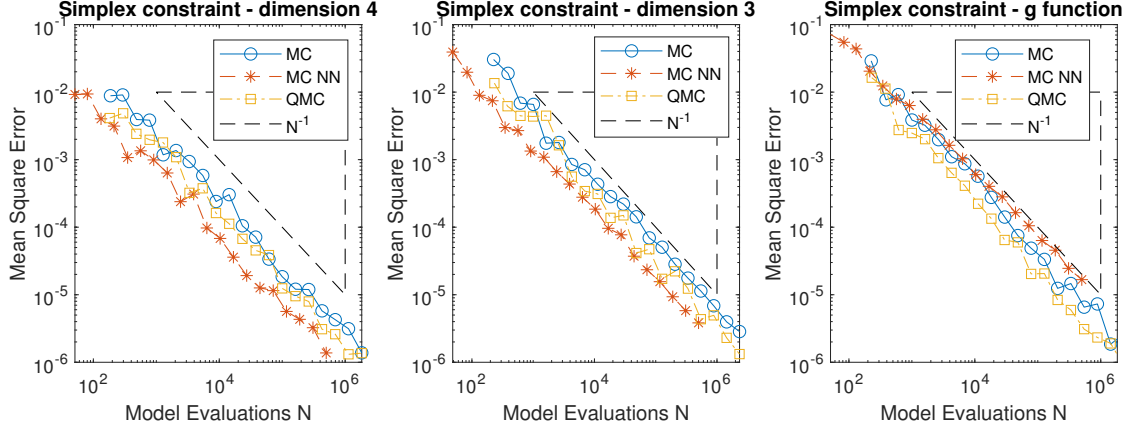


Figure 6: MSE at increasing sample sizes for total effect estimation for models with a simplex constraint on the inputs.

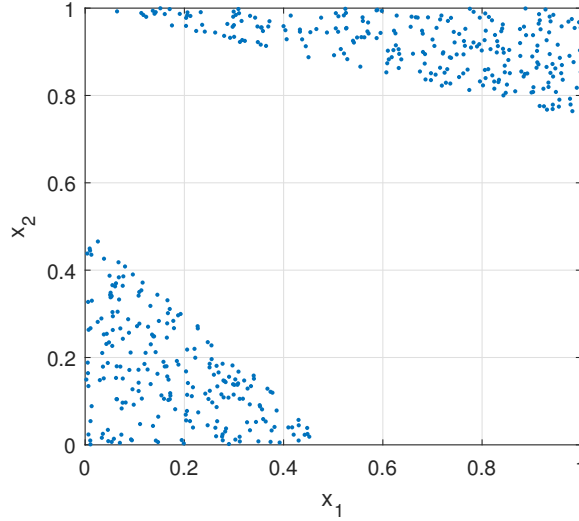


Figure 7: An input domain in the form of two separate triangles. About 500 points remain in the domain after rejection.

8.6 Features Constrained on the Sierpinski Gasket

In this section, we consider experiments in which the structure of the feature support is progressively complicated. We start with a Cartesian support and remove parts of the support to move towards a disconnected structure until we reach a Sierpinski gasket, that contains holes and is not star-shaped connected. Figure 8 provides a visualization of the regions. The second and third domains are created by cutting corners of the unit square. The former (second panel) consists of all points on the $[0, 1]$ with the exclusion of the pairs (x_1, x_2) such that $x_1 + x_2 > 3/2$, the latter (third panel) also excludes points of the type $x_1 + x_2 < 1/2$. The fourth domain approximates a fractal structure, the Sierpinski gasket, by excluding all realizations (x_1, x_2) that satisfy the following three conditions for $k = 1, 2, \dots, 5$:

- $\text{mod}(2^{k-1}(x_1 + x_2), 2) > 1$,
- $\text{mod}(2^{k-1}x_1, 2) \leq 1$, and
- $\text{mod}(2^{k-1}x_2, 2) \leq 1$,

where $x \mapsto \text{mod}(x, 2) = x - 2 \lfloor \frac{x}{2} \rfloor$ denotes the rest after an integer division by 2.

We consider a mapping of the form $Y = X_1 + \beta X_2$, with β varying in $\{-2, -1, 0, 1, 2\}$ for the experiments. With this test case we aim to unveil some of the subtleties that appear when the input domain becomes increasingly more disconnected.

It is possible to derive the expression of the density quotient semi-analytically. Let $x_1, x_2, x \in [0, 1]$. For

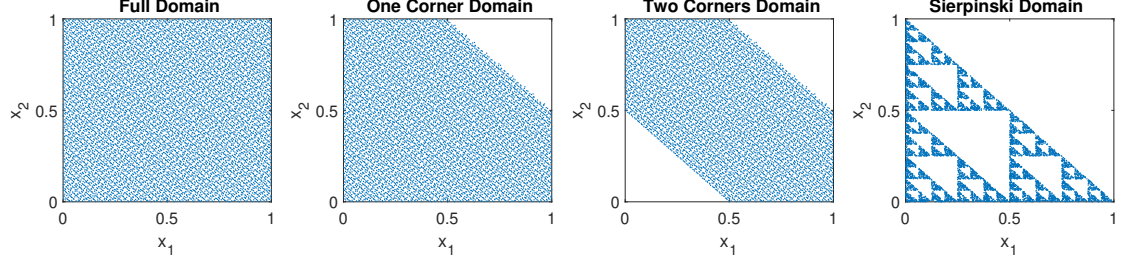


Figure 8: Feature constraints: left one-corner, center two-corner, right Sierpinski with depth $k = 5$.

the cut-one-corner constraint, joint density is defined by

$$f_{12}^{OC}(x_1, x_2) = \frac{7}{8} \cdot \mathbb{1} \left\{ x_1 + x_2 \leq \frac{3}{2} \right\} \quad (39)$$

and marginal densities by

$$f_1^{OC}(x) = f_2^{OC}(x) = \frac{7}{8} \cdot \left(1 - \left(x - \frac{1}{2} \right)^+ \right). \quad (40)$$

For the cut-two-corners domain, joint density is given by

$$f_{12}^{TC}(x_1, x_2) = \frac{3}{4} \cdot \mathbb{1} \left\{ \frac{1}{2} \leq x_1 + x_2 \leq \frac{3}{2} \right\} \quad (41)$$

and marginal densities by

$$f_1^{TC}(x) = f_2^{TC}(x) = \frac{3}{4} \cdot \left(1 - \left| x - \frac{1}{2} \right| \right). \quad (42)$$

The rejection rate is present in both the marginal distributions and the joint one, so that the inverse of the rejection rate is a multiplicative constant in the density quotient. For the Sierpinski gasket, the joint distribution is the product of indicator functions of the complements of the Sierpinski sets listed above. From these distributions, we can calculate the density quotient by marginal integration.

The shapes of the constraints rule out an approach based on the winding stairs design. We are then left with reweighting (shift or derange) and nearest-neighbour approaches. To proceed with numerical experiments, we use a rejection method. We generate data using crude Monte-Carlo sampling, and reject feature realizations that do not satisfy the constraints. For the cut-one-corner and cut-two-corners domains, we start with a sample of size $n = 10240$ on the $[0, 1]^2$ full Cartesian domain. After rejection of realizations falling outside the domain, we are left with samples of sizes $n = 8938$ and $n = 7690$, respectively for the the cut-one-corner and cut-two-corners domains. These numbers are in line with the theoretical acceptance rates of $7/8$ and $3/4$, respectively. For the Sierpinski gasket, the first step in the rejection process retains half of the observations, while each further step retains $3/4$ of the observations. Hence after five iterations, we reach a sample size of $n \approx \frac{1}{2} \cdot \left(\frac{3}{4} \right)^4 n_0$, where n_0 is the starting sample size. In this experiment, we start with an initial sample size of $n_0 = 50000$ and, after rejection, we are left with $n = 7932$ observations. Figure 9 shows

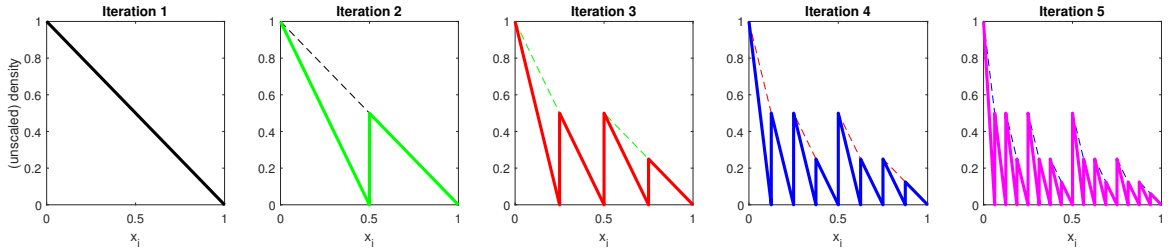


Figure 9: Sierpinski marginal density derivation: evolution of the marginal integral with rejections.

the empirical marginal densities of the features. Iteration $k = 1$ removes $1/2$, the further iterations $k > 1$ remove each $1/4$ of the mass. One needs to rescale these curves to obtain the marginal probability densities for both factors which are then piecewise linearly defined. The two-corner design and the Sierpinski gasket both satisfy $\text{cov}(X_1, X_2) = -0.5$ and have roughly the same input variances. In a linear model, only the input variances and covariances enter into the computation of total effects; thus, we expect similar values of the normalized total indices for the two corners and the Sierpinski gasket cases.

Table 4: Estimates of the normalized total effects for the linear model $Y = X_1 + \beta X_2$ (S&R: shift-and-reweight, NN: nearest-neighbour method).

	Full Design				One Corner				Two Corners				Sierpinski			
	S&R		NN		S&R		NN		S&R		NN		S&R		NN	
β	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
-2	0.20	0.80	0.20	0.80	0.15	0.63	0.16	0.63	0.11	0.45	0.11	0.44	0.11	0.42	0.11	0.43
-1	0.50	0.50	0.50	0.50	0.37	0.38	0.38	0.38	0.25	0.26	0.25	0.25	0.25	0.25	0.26	0.25
0	1.01	0	1.00	0	0.92	0	0.93	0	0.74	0	0.76	0	0.75	0	0.77	0
1	0.50	0.50	0.50	0.50	0.60	0.61	0.60	0.60	0.72	0.77	0.74	0.74	0.77	0.76	0.76	0.75
2	0.20	0.80	0.20	0.80	0.22	0.92	0.23	0.91	0.24	1.02	0.25	0.99	0.26	1.01	0.25	1.00

Table 4 reports the results obtained with the shift-and-reweight and the nearest-neighbour approaches. The resulting values indicate only small differences in the estimates. Similar values are also obtained with the derange-and-reweight estimator (not reported). Overall, the estimators exhibit a similar performance. The results in Table 4 can then be used to obtain some further indications about the behavior of total indices under constraints. For the fully connected domain, it is $T_1 + T_2 = 1$, as expected for a linearly additive model with independent inputs. However, we cannot expect this equality to hold when the inputs are constrained, as constraints make the inputs statistically dependent. For instance, for the two-corners domain and $\beta = -2$ we find $T_1 + T_2 = 0.56$, while for $\beta = 2$, we find $T_1 + T_2 = 1.26$. Similar values are obtained for the Sierpinski gasket. Results for the case $\beta = 0$ are also interesting. The total effect of X_2 is zero which correctly asserts that X_2 is inactive. However, the fact that the total effect of X_1 is less than 1 is due to input dependence, and must not be erroneously interpreted as the presence of an interaction in the model.

8.7 Comparison Tests with Machine Learning Approaches

The previous experiments have focused on a computer modeling setting. Indeed, by construction, the estimators we discussed are well suited for a setting in which the model and data distributions are known to the analyst. However, we have seen in Section 6 that there are links between our estimators and feature importance measures of the machine learning literature. In the present section, we present experiments aimed at shedding initial light on the behaviour of our estimators in comparison with the remove-and-retrain estimator discussed by Hooker et al. [2021] and Williamson et al. [2021, 2023]. We also aim to address the performance of our estimators for problems of larger dimensionality than in the previous sections.

As a first test case, we consider the following version of the Bratley et al. [1992] function defined as:

$$g(x) = \sum_{i=1}^d (-1)^i \prod_{j=i}^d x_{d+1-i}, \quad (43)$$

with inputs constrained on a simplex defined from (38) with $d = 10$. For this test case, it is possible to obtain the value of the total indices analytically, still using analytical calculations from Maple. We start with a sample of size 100,000 generated using crude Monte Carlo, with independent inputs and employ the rejection method to implement the simplex constraint. For the remove-and-retrain estimator, we proceed as follows. We use an artificial neural network (with 10 hidden layers) as machine learning model. The accuracy of the neural network is high, with a coefficient of model determination R^2 close to unity. As foreseen by the algorithm, we then remove each feature, retrain the neural network and measure the difference in R^2 . The resulting value is an estimate of the total indices of the features. The analytical values, as far as the results obtained with the derange-and-reweight method, the nearest-neighbour method and the remove-and-retrain method are reported Table 5.

Table 5: Performance results for the 10-dimensional model described in (43) (Ana.: analytical results, D&R: derange and reweight, NN: derange with nearest-neighbour, MM: metamodel fit using a feed-forward neural network with 10 hidden layers). The size of the MC sampling is 49808.

Type	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	[s]
Ana.	.0001	.0001	.0006	.0013	.0046	.013	.040	.118	.359	.821	
D&R	.0002	.0001	.0006	.0014	.0049	.013	.041	.119	.358	.814	28
NN	.0398	.0399	.0405	.0411	.0437	.051	.073	.133	.310	.676	139
MM	-.0002	-.0000	-.0001	.0009	.0044	.013	.041	.122	.356	.827	323

Let us now analyze the results. For the derange-and-reweight estimator (D&R), we observe estimates very close to the analytical values, and an overall estimation time of 28 seconds (last column). The estimates obtained with the nearest-neighbour design (NN) are distorted. Moreover, the estimation is notably time-consuming due to the long time required by the nearest neighbour search. The deteriorated performance

shows that the increased dimensionality negatively impacts this estimator. Finally, the estimates of the remove-and-retrain estimator (MM) are accurate, although one obtains negative values for features X_1 , X_2 and X_3 , whose analytical values are close to zero. Overall, the analysis takes 323 seconds, due to the retraining of the machine-learning model.

We then report results for a second test case, namely,

$$y = \sin \left(\pi \sum_{i=1}^d x_i \right), \quad (44)$$

with inputs still constrained on the simplex defined from (38) with $d = 10$. By construction the total indices of the first 8 features are equal and the last two. Analytically, we find $T_1 = T_2, \dots = T_8 = 0.595$, and $T_9 = T_{10} = 0.332$.

To study the performance of the estimators, we generate a first sample of size $n = 2,000$ and after rejecting the realizations violating the constraint we are left with about half of the realizations (987). Results are presented in Table 6.

Table 6: Performance for the 10-dimensional model described in (44) (Ana.: analytical results, D&R: derange and reweight, NN: derange with nearest-neighbour, MM: metamodel fit using a feed-forward neural network with 10 hidden layers). The size of the MC sampling is 987.

Type	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	[s]
Ana.	.595	.595	.595	.595	.595	.595	.595	.595	.332	.332	
D&R	.624	.610	.601	.596	.579	.640	.559	.615	.328	.299	0.51
NN	.309	.283	.240	.300	.300	.306	.248	.267	.205	.198	0.89
MM	-.216	-.292	-.340	-.311	.012	-.294	-.326	.006	-.671	-.502	12.1

The derange-and-reweight approach (D&R) provides estimates close to the analytical values, while the nearest-neighbour estimates (NN) are again distorted. They are equal to about half of the analytical values for all inputs. However, they still signal that X_9 and X_{10} are less important than the other eight features. The reject-and-retrain estimator (MM) shows a poor performance in this case, with several negative estimates, and fails in indicating the true feature importance in this case. We believe the reason is the poor fit of the machine-learning model. In fact, the neural network achieves an R^2 of about 7%, signaling that the emulator does not capture the input-output mapping in this test case.

In conclusion, the results in these two experiments seem not to recommend the nearest-neighbour design as the input space dimension increases, because it produces biased estimates. The main advantage of the reject-and-retrain estimator is that it can be applied both in a simulation and in a purely data-driven context because it does not require knowledge of the density quotients. However, it has to be considered with caution because its performance strongly depends on having an accurate machine-learning model, whose training time might make the estimation computationally expensive.

8.8 Application: the Flood Model of De Rocquigny (2006)

In this section, we apply the estimators to a realistic example, the flood model in de Rocquigny [2006], Chastaing et al. [2012]. The model calculates the maximum annual overflow, given eight input features (Table 7). We assume the same dependence structure in the input features as in Chastaing et al. [2012].

Table 7: Flood model feature list.

No.	Symb.	Description	Unit	Distribution and Truncation
1	Q	Maximal annual flow rate	$\frac{m^3}{s}$	Gumbel(1013,558) on (500,3000)
2	K_s	Strickler coefficient	—	Normal(30,8) on (15, $+\infty$)
3	Z_v	River downstream level	m	Triangular(49,50,51)
4	Z_m	River upstream level	m	Triangular(54,55,56)
5	H_d	Dyke height	m	Uniform(7,9)
6	C_b	Bank level	m	Triangular(55,55.5,56)
7	L	Length of river stretch	m	Triangular(4990,5000,5010)
8	B	River width	m	Triangular(295,300,305)

The correlation between the pair of features (1,2) is set to 0.5, and the correlation between (3,4) and (7,8) to 0.3 each, via Gaussian copula. The density quotients for each pair are given in (16). We calculate total indices with the derange-and-reweight approach of Theorem 8 and nearest-neighbour, fixing a budget of $B = 9,000$ model evaluations. The basic sample size is then $n = 1000$ Monte Carlo realisations. The

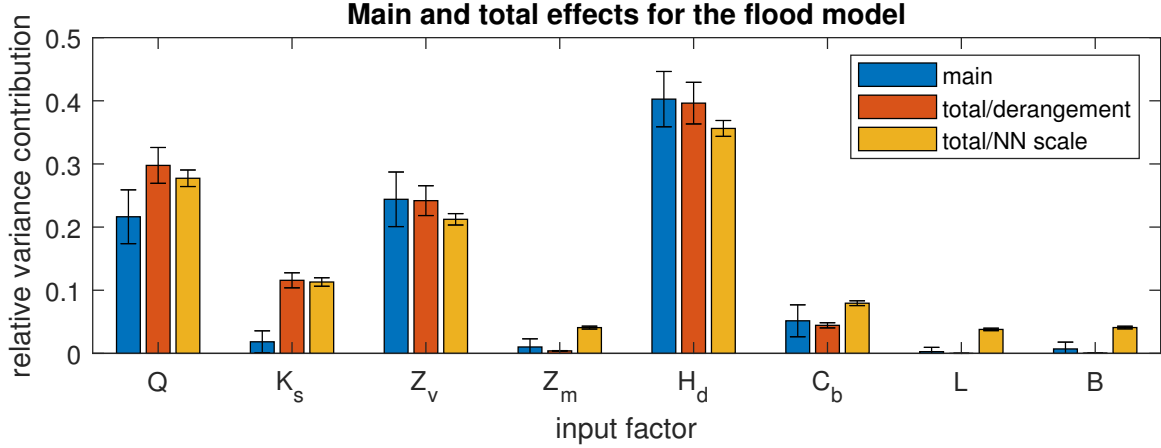


Figure 10: Sensitivity results for the flood model (error bars represent the 95% confidence band).

correlation structure in the basic sample is implemented via the Iman-Conover method [Iman and Conover, 1982, Mainik, 2015]. Main effects are estimated from this basic sample using a discrete cosine transformation Plischke [2012] with 8 harmonics. A jackknife is used to derive confidence bounds. For nearest-neighbour, a Monte Carlo sample of size $n = 9,000$ (same budget) is used. Because of the different scales in the inputs, we standardize the input sample using the empirical standard deviations as scaling factors. Figure 10 displays the results in the form of a barplot, reporting the main and total effects for each feature, as well as the error bands on top of the bars. The error bands for the main effects and for the shift/derangement approach are calculated as 1.96 times the square root of the plug-in variance estimates, using a normal approximation for a 5% confidence bound. Regarding the feature ranking, Feature 6 (H_d) is identified as the most important, followed by Features 1 (Q), 3 (Z_v), 2 (K_s) and 7 (C_b); the remaining features play a minor role. This result is in accordance with the findings in Chastaing et al. [2012]. The values of the main effects and total effects (estimated with the derange-and-reweight approach) are close. Because Feature 6 is stochastically independent of the remaining features, this equality signals that H_d is not involved in relevant interactions. In contrast, features 1 and 2 are noticeably different under main and total effects, with total effects larger than their main effects. Also, if ranked according to main effects, Feature 3 would rank second most important, switching place with Feature 1. Overall, the derange-and-reweight approach that evaluates the model at the mixed realizations shows comparable results to the nearest-neighbour approach. However, the nearest-neighbour approach seems to exhibit a bias for the least important inputs. By the theoretical results of Devroye et al. [2018] for dimensions larger than four, the nearest-neighbour estimator bias dominates the estimator variance. One insight gained from this application is to use both a derangement and a nearest-neighbour approach, when possible, to confirm the estimates of both methods. This is because the former is less affected by bias as dimensionality increases.

9 Final Remarks

Estimating total effects under feature dependence and constraints is a challenging task. We have proposed a set of estimators that accommodate increasingly complex constraints. For all estimators we have addressed both theoretical and numerical aspects. We have first analyzed the performance of a winding stairs approach that relies on a Knothe-Rosenblatt transformation for the case of dependent features. While the estimator accommodates a broad family of input dependences, it becomes inapplicable in the presence of non-Cartesian domains.

We have then studied estimators based on pairing Jansen’s design with a reweighting factor. We have formulated a U-statistic estimator, for which a central limit theorem is immediately derived. To abate the computational burden, we have proposed two alternatives based on shifts and derangements, for which we have proven central limit theorems. We have also considered a nearest-neighbour approach for which, however, theoretical results seem out of reach. We have tested the behavior of the estimators through numerous experiments with feature constraints of increasing complexity.

We have also studied the connection of these approaches with the calculation of feature importance measures in the machine-learning literature. On the theoretical side, we have derived the link between total indices under constraints and Breiman’s permutation feature importance measures. We have compared our approach with the removal-and-retrain method of Williamson et al. [2023], whose importance measures are, in fact, total indices. We have seen that our method and removal-and-retrain produce similar results when the machine learning model captures well the simulation response. However, when the performance of the

machine learning model is poor the removal-and-retrain method fails in computing total Sobol’ indices. This investigation is, however, preliminary and future research is needed to come to more definitive result. In this perspective, future research aims at further studying the performance of the method in a data-driven context. We expect some challenges to emerge. On the one hand, the nearest-neighbour approach is expected to suffer from the curse of dimensionality, as our experiments have shown. On the other hand, estimation accuracy is expected to depend on the accurate calculation of density ratios, which, instead, are known in the simulation context. There, a starting point is also the work of Fisher et al. [2019], where density ratios are approximated.

References

- A. Badea and R. Bolado. Milestone M.2.1.D.4: Review of sensitivity analysis methods and experience. Technical report, PAMINA Project, Sixth Framework Programme, European Commission, 2008. <http://www.ip-pamina.eu/downloads/pamina.m2.1.d.4.pdf>.
- M. Bayoucef and M. Mascagni. A computational investigation of the optimal Halton sequence in QMC applications. *Monte Carlo Methods and Applications*, 25(3):187–207, 2019.
- C. B  nard, S. D. Veiga, and E. Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika*, 109(4):881–900, 2022.
- A. Bose and S. Chatterjee. *U-Statistics, M_m -Estimators and Permutations*. Springer Verlag, Singapore, 2018.
- P. Bratley, B. L. Fox, and H. Niederreiter. Implementation and tests of low-discrepancy sequences. *ACM Transactions on Modeling and Computer Simulation*, 2(3):195–213, 1992.
- L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- B. Broto, F. Bachoc, and M. Depecker. Variance reduction for estimation of Shapley effects and adaptation to unknown input distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):693–716, 2020.
- E. Cand  s, Y. Fan, L. Janson, and J. Lv. Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society, Series B*, 3(80):551–577, 2018.
- K. Chan, A. Saltelli, and S. Tarantola. Winding stairs: A sampling tool to compute sensitivity indices. *Statistics and Computing*, 10(3):187–196, 2000.
- G. Chastaing, F. Gamboa, and C. Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012.
- S. Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536): 2009–2022, 2021.
- S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. *Basics and Trends in Sensitivity Analysis: Theory and Practice in R*. SIAM, Philadelphia PA, 2021.
- E. de Rocquigny. La ma  trise des incertitudes dans un contexte industriel. 1^{re} partie: une approche m  thodologique globale bas  e sur des exemples. *Journal de la Soci  t   Fran  aise de Statistique*, 147(3): 33–71, 2006.
- L. Devroye, P. G. Ferrario, L. Gy  rfi, and H. Walk. Strong universal consistent estimate of the minimum mean squared error. In *Empirical Inference*, pages 143–160. Springer Verlag, 2013.
- L. Devroye, L. Gy  rfi, G. Lugosi, and H. Walk. A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics*, 12:1752–1778, 2018.
- P. H. Diananda. The central limit theorem for m-dependent variables asymptotically stationary to second order. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 50, pages 287–292. Cambridge University Press, 1954.
- B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981.
- A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20:1–81, 2019.

- M. Fréchet. Sur le coefficient, dit de corrélation et sur la corrélation en général. *Revue de l'Institut International de Statistique*, 1(4):16–23, 1934.
- 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.
- D. Gatelli, S. Kucherenko, M. Ratto, and S. Tarantola. Calculating first-order sensitivity measures: A benchmark of some recent methodologies. *Reliability Engineering&System Safety*, 94:1212–1219, 2009.
- R. Genuer, V. Michel, E. Eger, and B. Thirion. Random forests based feature selection for decoding fmri data. In *Proceedings Compstat*, volume 267, pages 1–8, 2010.
- L. Gilquin, C. Prieur, and E. Arnaud. Replication procedure for grouped sobol’indices estimation in dependent uncertainty spaces. *Information and Inference: A Journal of the IMA*, 4(4):354–379, 2015.
- T. Goda. A simple algorithm for global sensitivity analysis with Shapley effects. *Reliability Engineering&System Safety*, 213:107702, 2021.
- J. Hart and P. A. Gremaud. An approximation theoretic perspective of Sobol’ indices with dependent variables. *Int. J. Uncertainty Quantification*, 8(6):483–493, 2018.
- R. Helmers. On the Edgeworth expansion and the bootstrap approximation for a Studentized U -statistic. *The Annals of Statistics*, 19:470–484, 1991.
- W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.
- T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering&System Safety*, 52(1):1–17, 1996.
- G. Hooker, L. Mentch, and S. Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):82:1–16, 2021.
- R. L. Iman and W. J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, 11(3):311–334, 1982.
- J. Jacques, C. Lavergne, and N. Devictor. Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering&System Safety*, 91(10–11):1126–1134, 2006.
- 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, 2014.
- M. J. W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 117(1–2):35–43, 1999.
- M. J. W. Jansen, W. A. H. Rossing, and R. A. Daamen. Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. *Predictability and nonlinear modelling in natural sciences and economics*, pages 334–343, 1994.
- H. Joe. *Dependence Modeling with Copulas*. CRC Press, Boca Raton, 2014.
- H. Knothe. Contributions to the theory of convex bodies. *Michigan Math. J.*, 4(1):39–52, 1957.
- S. Kucherenko, S. Tarantola, and P. Annoni. Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183(4):937–946, 2012.
- S. Kucherenko, O. V. Klymenko, and N. Shah. Sobol’ indices for problems defined in non-rectangular domains. *Reliability Engineering&System Safety*, 167:218–231, 2017.
- S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer Science+Business Media, New York, 2003.
- J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111, 2018.
- G. Li and H. Rabitz. Relationship between sensitivity indices defined by variance- and covariance-based methods. *Reliability Engineering&System Safety*, 167:136–157, 2017.
- S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. preprint, arXiv, 2017. <https://arxiv.org/abs/1705.07874>.

- G. Mainik. Risk aggregation with empirical margins: Latin hypercubes, empirical copulas, and convergence of sum distributions. *Journal of Multivariate Analysis*, 141:197–216, 2015.
- T. A. Mara and S. Tarantola. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering&System Safety*, 107:115–121, 2012.
- T. A. Mara, S. Tarantola, and P. Annoni. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling&Software*, 72:173–183, 2015.
- J. Matoušek. On the L^2 -discrepancy for anchored boxes. *Journal of Complexity*, 14(4):527–556, 1998.
- W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.
- M. H. Neumann. A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics. *ESAIM: Probability and Statistics*, 17:120–134, 2013.
- J. E. Oakley and A. O’Hagan. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 66(3):751–769, 2004.
- A. B. Owen. Scrambled net variance for integrals of smooth functions. *The Annals of Statistics*, 25(4):1541–1562, 1997.
- A. B. Owen and C. R. Hoyt. Efficient estimation of the ANOVA mean dimension, with an application to neural net classification. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2), 2021.
- A. B. Owen and C. Prieur. On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.
- K. Pearson. Notes on regression and inheritance in the case of two parents. *Proc. Royal Soc. London*, 58:240–242, 1895.
- E. Plischke. How to compute variance-based sensitivity indicators with your spreadsheet software. *Environmental Modelling&Software*, 35:188–191, 2012.
- E. Plischke, G. Rabitti, and E. Borgonovo. Has the spell been broken? Estimating global sensitivity measures via nearest neighbors. 2022. In Preparation.
- C. Prieur and S. Tarantola. Variance-based sensitivity analysis: Theory and estimation algorithms. In *Handbook of Uncertainty Quantification*, pages 1217–1239. Springer Verlag, Cham, 2017.
- H. Rabitz and Ö. F. Alış. General foundations of high-dimensional model representations. *J. Math. Chem.*, 25(2–3):197–233, 1999.
- M. Rosenblatt. Remarks on a multivariate transformation. *Ann. Math. Statist.*, 23(3):470–472, 1952.
- A. Saltelli and S. Tarantola. On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. *Journal of the American Statistical Association*, 97(459):702–709, 2002.
- A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. John Wiley&Sons, Chichester, 2000.
- I. M. Sobol’. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling & Computational Experiments*, 1:407–414, 1993.
- I. M. Sobol’, S. Tarantola, D. Gatelli, S. S. Kucherenko, and W. Mauntz. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering&System Safety*, 92:957–960, 2007.
- D. M. Sparkman, J. E. Garza, H. R. Millwater, Jr., and B. P. Smarslok. Importance sampling-based post-processing method for global sensitivity analysis. In *18th AIAA Non-Deterministic Approaches Conference. 4-8 January 2016, San Diego, California, USA*. AIAA SciTech, 2016. Paper #AIAA 2016-1444.
- X. Sun, W. Zhong, and P. Ma. An asymptotic and empirical smoothing parameters selection method for smoothing spline anova models in large samples. *Biometrika*, 108(1):149–166, 2021.
- A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.

B. D. Williamson, P. B. Gilbert, M. Carone, and N. Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77:9–22, 2021. With discussion.

B. D. Williamson, P. B. Gilbert, N. R. Simon, and M. Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

A Proofs

A.1 More on conditional independence

In this appendix we detail more the role of conditional independence for the computation of total Sobol’ indices. In particular, we provide the proof of Lemma 1, Proposition 2 stated in Section 2.

Proof of Lemma 1. From Fréchet [1934] we know that:

$$\mathbb{V}[\mathbb{E}[Y|X_{-u}]] = \mathbb{E}[\mathbb{E}[Y|X_{-u}]^2] - \mathbb{E}[Y]^2 = \mathbb{E}[\mathbb{E}[Y \cdot \mathbb{E}[Y|X_{-u}] - \mathbb{E}[Y]^2|X_{-u}]] = \text{cov}(Y, \mathbb{E}[Y|X_{-u}]).$$

Then, we deduce the following covariance representation of τ_u :

$$\tau_u = \mathbb{E}[\mathbb{V}[Y|X_{-u}]] = \mathbb{V}[Y] - \text{cov}(Y, \mathbb{E}[Y|X_{-u}]). \quad (45)$$

Now, let Y' be an independent replicate of Y conditionally on X_{-u} . As a consequence of conditional independence, $\mathbb{E}[Y \cdot Y'|X_{-u}] = \mathbb{E}[Y|X_{-u}] \mathbb{E}[Y'|X_{-u}]$. Hence

$$\text{cov}(Y, Y') = \mathbb{E}[\mathbb{E}[Y \cdot Y' - \mathbb{E}[Y]^2|X_{-u}]] = \mathbb{E}[\mathbb{E}[Y|X_{-u}] \cdot \mathbb{E}[Y'|X_{-u}] - \mathbb{E}[Y]^2] = \mathbb{V}[\mathbb{E}[Y|X_{-u}]],$$

so that

$$\tau_u = \mathbb{V}[Y] - \text{cov}(Y, Y') = \frac{1}{2} (\mathbb{V}[Y] + \mathbb{V}[Y']) - \text{cov}(Y, Y') = \frac{1}{2} \mathbb{E}[(Y - Y')^2]$$

where $\mathbb{V}[Y] = \mathbb{V}[Y']$ as both are identically distributed. \square

Let us shed some more light on conditional independence. For this, let Z be a general m -dimensional random vector that takes its values on a Cartesian product space. Again for notation simplicity, we assume that the joint probability distribution of Z has a density function $h(z)$ with respect to Lebesgue measure. For u, v two disjoint subsets of $[m] = \{1, 2, \dots, m\}$, we denote the disjoint union of u and v by $u + v$.

Definition 15. Let u, v, w be pairwise disjoint index sets in $[m]$. Then Z_u and Z_v are conditionally independent given Z_w ($u \perp\!\!\!\perp v|w$, for short) if for all $z_u \in \mathcal{Z}_u, z_v \in \mathcal{Z}_v, z_w \in \mathcal{Z}_w$, the density h satisfies

$$h_{u+v|w}(z_{u+v}|z_w) = h_{u|w}(z_u|z_w) \cdot h_{v|w}(z_v|z_w). \quad (46)$$

Multiplying (46) by $h_w(z_w)$ we find

$$h_{u+v+w}(z_{u+v+w}) = h_{u|w}(z_u|z_w) \cdot h_{v|w}(z_v|z_w) h_w(z_w) = \begin{cases} h_{v|w}(z_v|z_w) \cdot h_{u+w}(z_{u+w}), \\ h_{u|w}(z_u|z_w) \cdot h_{v+w}(z_{v+w}). \end{cases} \quad (47)$$

Proof of Proposition 2. The proof of Proposition 2 follows straightforwardly from these observations. \square

A.2 Proof of the results stated in Section 3

Proof of Proposition 5. Consider an independent copy X' of X . Projections onto index subsets u and $-u$ keep independence intact, i.e., $X'_u = P_u(X')$ and $X_{-u} = P_{-u}(X)$ are independent. The random vector obtained by glueing these two vectors together therefore has a density $f_u \cdot f_{-u}$, breaking the inter-block dependence. Hence generally for a measurable function $h : \mathbb{R}^d \rightarrow \mathbb{R}$, we have

$$\mathbb{E}[h(X'_u : X_{-u})] = \iint h(x'_u : x_{-u}) f_{-u}(x_{-u}) f_u(x'_u) dx_{-u} dx'_u.$$

Now, in order to compare the expectation of $h(X)$ with X possibly being dependent, we split the argument X into two arguments via projections onto subdimensions indexed by u and $-u$, so that we may write the joint density as product of marginal and conditional density,

$$\mathbb{E}[h(X_u : X_{-u})] = \iint h(x_u : x_{-u}) f_{u|-u}(x_u|x_{-u}) f_{-u}(x_{-u}) dx_u dx_{-u}.$$

Considering all three terms in a function $h_2 : \mathbb{R}^{|u|} \times \mathbb{R}^{d-|u|} \times \mathbb{R}^{|u|} \rightarrow \mathbb{R}$ and taking its expectation then leads to

$$\mathbb{E}[h_2(X_u, X_{-u}, X'_u)] = \iiint h_2(x_u, x_{-u}, x'_u) f_{u|-u}(x_u|x_{-u}) f_{-u}(x_{-u}) f_u(x'_u) dx_u dx_{-u} dx'_u.$$

Let us now consider the weighted Jansen's estimator,

$$h_2(X_u, X_{-u}, X'_u) = \frac{1}{2} \frac{f_{u,-u}(X'_u : X_{-u})}{f_u(X'_u) f_{-u}(X_{-u})} (g(X_u : X_{-u}) - g(X'_u : X_{-u}))^2.$$

Then by (9),

$$\begin{aligned} \mathbb{E}[h_2(X_u, X_{-u}, X'_u)] &= \frac{1}{2} \iiint \frac{f_{u,-u}(x'_u : x_{-u})}{f_u(x'_u) f_{-u}(x_{-u})} (g(x_u : x_{-u}) - g(x'_u : x_{-u}))^2 \\ &\quad f_{u|-u}(x_u|x_{-u}) f_{-u}(x_{-u}) f_u(x'_u) d(x_u, x_{-u}, x'_u) \\ &= \frac{1}{2} \iiint f_{u|-u}(x'_u|x_{-u}) (g(x_u : x_{-u}) - g(x'_u : x_{-u}))^2 \\ &\quad f_{u|-u}(x_u|x_{-u}) f_{-u}(x_{-u}) d(x_u, x_{-u}, x'_u) = \tau_u. \end{aligned} \quad (48)$$

The last equality follows from (9). By definition, a sample consists of realizing n independent copies of X . Hence two different copies of X , X^i and X^j , $i, j = 1, \dots, n$, $i \neq j$, are independent. Then, an estimator of τ_u is

$$\hat{\tau}_{u,n}^U = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} h_2(X_u^i, X_{-u}^i, X_u^j), \quad (49)$$

which yields (18). \square

Proof of Lemma 6. Let us define the random vector $W = (W_1, W_2) = (X_u, X_{-u})$ (the random vector X split according to the index set u). Let W^i , $i = 1, \dots, n$, be identical copies of W . We then write the estimator as follows:

$$\hat{\tau}_{u,n}^U = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \Phi(W^i, W^j) = \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j > i} \phi^s(W^i, W^j) \quad (50)$$

with

$$\Phi^s(W^i, W^j) = \frac{1}{2} (\Phi(W^i, W^j) + \Phi(W^j, W^i)), \quad (51)$$

$$\Phi(W^i, W^j) = \frac{1}{2} i_u(W_1^i, W_2^j) \left(g(W_1^i : W_2^j) - g(W_1^j : W_2^i) \right)^2. \quad (52)$$

Hence, $\hat{\tau}_{u,n}^U$ defines a U -statistic of order 2 for $\tau_u = \mathbb{E}[\Phi(W^i, W^j)] = \mathbb{E}[\hat{\tau}_{u,n}^U]$. \square

Proof of Lemma 7. The proof of (19) follows directly from (50) using the computations in [Bose and Chatterjee, 2018, Section 1.2.1]. The asymptotic normality is a consequence of [Bose and Chatterjee, 2018, Theorem 1.1]. \square

A.3 Proof of the results stated in Section 4

Proof of Theorem 8. First, the i th realization and the $s_n(i)$ th one in the input sample are independent. We thus deduce from (48) that for any $1 \leq i \leq n$,

$$\frac{1}{2} \mathbb{E} \left[\iota_j(X^{s_n(i)}, X^i) \left(g(X_j^i : X_{-j}^i) - g(X_j^{s_n(i)} : X_{-j}^i) \right)^2 \right] = \tau_j \quad (53)$$

thus the estimator $\hat{\tau}_{j,n}$ is unbiased. To prove the central limit theorem, we first decompose $\hat{\tau}_{j,n}^S$ as:

$$\hat{\tau}_{j,n}^S = \frac{n-1}{n} \tilde{\tau}_{j,n-1} + \frac{1}{2n} \frac{f_{j,-j}(X_j^1 : X_{-j}^n)}{f_j(X_j^1) f_{-j}(X_{-j}^n)} (g(X_j^n : X_{-j}^n) - g(X_j^1 : X_{-j}^n))^2$$

where $\tilde{\tau}_{j,n-1}$ stands for the estimator built from Formula (21) on (X_1, \dots, X_{n-1}) and by replacing the cyclic shift-by-one s_n by the acyclic one s_{n-1}^a . Then, $\tilde{\tau}_{j,n-1} = \frac{1}{n-1} \sum_{i=1}^{n-1} V_j^i$ and the sequence $(V_j^i)_{i \geq n}$ is stationary and 1-dependent. Thus, the limit $\sqrt{n}(\tilde{\tau}_{j,n-1} - \tau_j) \rightarrow \mathcal{N}(0, \sigma_j^2)$ follows from [Diananda, 1954, Theorem 5]. Finally, noting that

$$\mathbb{E} \left[\left| \sqrt{n} \frac{1}{2n} \frac{f_{j,-j}(X_j^1 : X_{-j}^n)}{f_j(X_j^1) f_{-j}(X_{-j}^n)} (g(X_j^n : X_{-j}^n) - g(X_j^1 : X_{-j}^n))^2 \right| \right] = \frac{\tau_j}{\sqrt{n}} \xrightarrow{n \rightarrow +\infty} 0 \quad (54)$$

and applying Slutsky's Theorem we get (22). \square

Proof of Corollary 9. The result follows from Theorem 8 and by applying the Delta method (see, e.g., van der Vaart [1998]). First, by mimicking the proof of Theorem 8, it is possible to prove that for any α, β, γ , a central limit theorem holds true for $\alpha \hat{\tau}_{j,n}^S + \beta \frac{1}{n} \sum_{i=1}^n g(X^i) + \gamma \frac{1}{n} \sum_{i=1}^n g^2(X^i)$. It yields a central limit theorem for $(U_{1,n}, U_{2,n}, U_{3,n}) = (\hat{\tau}_{j,n}^S, \frac{1}{n} \sum_{i=1}^n g(X^i), \frac{1}{n} \sum_{i=1}^n g^2(X^i))$, namely

$$\sqrt{n} \left((U_{1,n}, U_{2,n}, U_{3,n})^T - \theta_j \right) \xrightarrow{n \rightarrow +\infty} \mathcal{N}(0, \Sigma_j)$$

with $\theta_j = (\tau_j, \mathbb{E}[Y], \mathbb{E}[Y^2])$ and

$$\Sigma_j = \begin{pmatrix} \mathbb{V}[V_j^1] & \text{cov}(V_j^1, Y^1) & \text{cov}(V_j^1, (Y^1)^2) \\ * & \mathbb{V}(Y^1) & \text{cov}(Y^1, (Y^1)^2) \\ * & * & \mathbb{V}[(Y^1)^2] \end{pmatrix}.$$

Then we can prove the central limit theorem for T_j , using the Delta method on $\psi(x, y, z) = \frac{x}{z - y^2}$ and $\theta_j = (\tau_j, \mathbb{E}[Y], \mathbb{E}[Y^2])$. More precisely, let ρ_j denote the gradient of ψ at θ_j . We have $\rho_j = \nabla \psi(\theta_j) = \frac{1}{\mathbb{V}[Y]} [1, 2T_j \mathbb{E}[Y], -T_j]^T$. Thus

$$\sqrt{n} (\hat{T}_{j,n}^S - T_j) \xrightarrow{n \rightarrow +\infty} \mathcal{N}(0, \rho_j^T \Sigma_j \rho_j).$$

It concludes the proof of Corollary 9. \square

Proof of Theorem 10. To prove that $\hat{\tau}_{j,n}^D$ is unbiased, we use the same arguments as the ones used to prove that $\hat{\tau}_{j,n}^S$ defined by (21) in Theorem 8 is unbiased, additionally noting that $\pi_n(i) \neq i$ for all $i = 1, \dots, n$. To prove the central limit theorem, we first decompose, for each n , the permutation π_n in cycles $C_{1,n}, \dots, C_{m_n,n}$. Let us arbitrarily fix the first element in each cycle. We then form p_n blocks, with $p_n = \max_{1 \leq k \leq m_n} \ell_{k,n}$ and $\ell_{k,n}$ the length of cycle $C_{k,n}$. For each n , we re-order the X^i 's so that the first $b_{1,n} = m_n$ re-ordered variables are the first element of each cycle, the next $b_{2,n}$ re-ordered variables are the second element of each cycle with length at least two and so on until the $n - \sum_{v=1}^{p_n-1} b_{v,n}$ last $b_{p_n,n}$ re-ordered variables which are the last element in each cycle of length p_n . Here, $1 \leq b_{p_n,n} \leq \dots \leq b_{1,n} = m_n$. For each n , we denote the re-ordered sequence of X^i 's by $X^{i,n}$, $1 \leq i \leq n$, $n \geq 1$. We then define $S_{v,n} = \sum_{i=k_{v-1,n}+1}^{k_{v,n}} \tilde{V}_j^{i,n}$, with $k_0 = 0$, $k_{v,n} = \sum_{w=1}^v b_{w,n}$ and $\tilde{V}_j^{i,n}$ defined as $V_j^i - \tau_j/\sqrt{n}$ but with the $X^{i,n}$'s in place of the X^i 's. More precisely, we have to use the trick in the proof of Theorem 8 by replacing first the last variable X^i in cycle $C_{1,n}$ by X^{n+1} , \dots , the last variable X^i in cycle $C_{m_n,n}$ by X^{n+m_n} . As $\lim_{n \rightarrow +\infty} m_n/\sqrt{n} = 0$, we prove that the remaining term decreases fast enough not to perturb the result of the central limit theorem (see the proof of Theorem 8 for more details). Now, as $\lim_{n \rightarrow +\infty} m_n/\sqrt{n} = 0$, then $p_n \geq n/m_n \geq \sqrt{n}/m_n \rightarrow +\infty$. For each $n \geq 1$, the sequence $(S_{v,n})_{1 \leq v \leq p_n}$ is a sequence of 1-dependent variables. Then applying [Neumann, 2013, Theorem 2.1], we get the central limit theorem, as soon as there exists $\delta > 0$ such that $\mathbb{E}[|V_j^1|^{2+\delta}] < +\infty$ for some $\delta > 0$. Indeed, as variables inside each block are centered, independent and identically distributed,

$$\sum_{v=1}^{p_n} \mathbb{E}[S_{v,n}^2] = \sum_{v=1}^{p_n} \frac{b_{v,n}}{n} \mathbb{V}[V_j^1] = \mathbb{V}[V_j^1] < +\infty.$$

Then, Assumption (2.1) in [Neumann, 2013, Theorem 2.1] is true due to the stationarity of $V_j^i V_j^{i+1}$. Indeed, due to 1-dependence,

$$\begin{aligned} \mathbb{V} \left[\sum_{v=1}^{p_n} S_{v,n} \right] &= \sum_{v=1}^{p_n} \mathbb{V}[S_{v,n}] + 2 \sum_{1 \leq v < w \leq p_n} \text{cov}(S_{v,n}, S_{w,n}) = \sum_{v=1}^{p_n} b_{v,n} \frac{\mathbb{V}[V_j^1]}{n} + 2 \sum_{v=1}^{p_n-1} \text{cov}(S_{v,n}, S_{v+1,n}) \\ &= \frac{\mathbb{V}[V_j^1]}{n} \sum_{v=1}^{p_n} b_{v,n} + \frac{2}{n} \sum_{v=1}^{p_n-1} b_{v+1,n} \text{cov}(V_j^1, V_j^2) = \mathbb{V}[V_j^1] + 2 \frac{n - m_n}{n} \text{cov}(V_j^1, V_j^2) \xrightarrow{n \rightarrow +\infty} \sigma_j^2 < +\infty. \end{aligned}$$

Let us now prove that Assumption (2.2) of [Neumann, 2013, Theorem 2.1] holds. For any $\varepsilon > 0$ we have

$$\begin{aligned}
\sum_{v=1}^{p_n} \mathbb{E} \left[S_{v,n}^2 \mathbb{I}_{|S_{v,n}| > \varepsilon} \right] &\leq \sum_{v=1}^{p_n} \mathbb{E} \left[|S_{v,n}|^{2+\delta} \right]^{\frac{2}{2+\delta}} (\Pr(|S_{v,n}| > \varepsilon))^{\frac{\delta}{2+\delta}} \quad (\text{using Hölder Inequality}) \\
&\leq \varepsilon^{-\delta} \sum_{v=1}^{p_n} \mathbb{E} \left[|S_{v,n}|^{2+\delta} \right] \quad (\text{using Markov Inequality}) \\
&\leq \varepsilon^{-\delta} \sum_{v=1}^{p_n} \mathbb{E} \left[\left(\sum_{i=k_{v-1,n}+1}^{k_{v,n}} n^{-1/2} |V_j^i - \tau_j| \right)^{2+\delta} \right] \\
&= \varepsilon^{-\delta} \sum_{v=1}^{p_n} \frac{b_{v,n}^{2+\delta}}{n^{1+\frac{\delta}{2}}} \mathbb{E} \left[\left(\sum_{i=k_{v-1,n}+1}^{k_{v,n}} b_{v,n}^{-1} |V_j^i - \tau_j| \right)^{2+\delta} \right] \\
&\leq \varepsilon^{-\delta} \sum_{v=1}^{p_n} \frac{b_{v,n}^{2+\delta}}{n^{1+\frac{\delta}{2}}} \mathbb{E} \left[\sum_{i=k_{v-1,n}+1}^{k_{v,n}} b_{v,n}^{-1} |V_j^i - \tau_j|^{2+\delta} \right] \quad (\text{using Jensen Inequality}) \\
&\leq \frac{1}{\varepsilon^\delta} \frac{1}{n^{1+\frac{\delta}{2}}} \sum_{v=1}^{p_n} b_{v,n}^{1+\delta} \sum_{i=k_{v-1,n}+1}^{k_{v,n}} \mathbb{E} \left[|V_j^i - \tau_j|^{2+\delta} \right] \\
&\leq \frac{1}{\varepsilon^\delta} \frac{1}{n^{1+\frac{\delta}{2}}} m_n^{1+\delta} \sum_{v=1}^{p_n} b_{v,n} \mathbb{E} \left[|V_j^1 - \tau_j|^{2+\delta} \right] \quad (\text{by stationarity of } (|V_j^i - \tau_j|)_{i \geq 1}) \\
&\leq \frac{1}{\varepsilon^\delta} \frac{m_n^{1+\delta}}{n^{\frac{\delta}{2}}} \mathbb{E} \left[|V_j^1 - \tau_j|^{2+\delta} \right] \xrightarrow{n \rightarrow +\infty} 0
\end{aligned}$$

The limit holds as $\mathbb{E} [|V_j^1|^{2+\delta}] < +\infty$ and $\lim_{n \rightarrow +\infty} \frac{m_n^{1+\delta}}{\sqrt{n}^\delta} \rightarrow 0$. This concludes the proof of Theorem 10. \square

A.4 Link with Breiman's Permutation Importance

Proof of Proposition 13. Let us write the squared loss function as $\mathcal{L}(Y, g(X; \theta)) = (Y - g(X; \theta))^2$. Then, (31) becomes

$$\text{MDA}_j^{\text{FR}} = \mathbb{E}[\ell_j(X', X)(Y - g(X'_j : X_{-j}; \theta))^2] - \mathbb{E}[(Y - g(X; \theta))^2]. \quad (55)$$

Using the assumption that $g(X; \theta)$ is a perfect predictor, we have $Y = g(X; \theta)$ for all values of X , so that $\mathbb{E}[(Y - g(X; \theta))^2] = \mathbb{E}[0] = 0$. Then, (55) becomes

$$\text{MDA}_j = \mathbb{E}[\ell_j(X', X)(g(X, \theta) - g(X'_j : X_{-j}; \theta))^2] = \tau_j.$$

Hence total effects and MDA with a weighted squared loss are the same. \square

A.5 Proof of Theorem 14

Proof of Theorem 14. Main effect: We consider the Gaussian multivariate distribution of (X, Y) . Setting $u = \{d+1\}$, $v = \emptyset$, $w = \{j\}$ in (32), then the variance of Y conditionally on X_j is

$$\Sigma_{Y,Y} - \Sigma_{Y,j} \Sigma_{j,j}^{-1} \Sigma_{j,Y} = \beta^T \Sigma \beta - \beta^T (\Sigma_{[d],j} \Sigma_{j,j}^{-1} \Sigma_{j,[d]}) \beta, \quad (56)$$

using here the special structure of the augmented matrix. This variance is constant (as it does not depend on the value of X_j), so that $\mathbb{E}[\mathbb{V}[Y|X_j]] = \mathbb{V}[Y|X_j]$. However, for main effects we are interested in the variance of the conditional expectation, so we have to subtract the value in (56) from the total variance which yields (33).

Total effect: The Rosenblatt transform in the Gaussian case uses the Cholesky decomposition of the covariance matrix. The Cholesky matrix $C = \text{chol}(\Sigma)$ is an upper triangular matrix such that $C^T C = \Sigma$. If $Z \sim \mathcal{N}(0, I)$ then $X = \mu + C^T Z \sim \mathcal{N}(\mu, \Sigma)$. One can define the Cholesky decomposition recursively,

$$\text{chol}(\Sigma) = \begin{pmatrix} \Sigma_{1,1}^{1/2} & \Sigma_{1,1}^{-1/2} \cdot \Sigma_{1,-1} \\ 0 & \text{chol}(\Sigma_{-1,-1} - \Sigma_{1,-1}^T \Sigma_{1,1}^{-1} \Sigma_{1,-1}) \end{pmatrix}$$

where the index -1 denotes all coordinates but the first (if Σ is a scalar then $\text{chol}(\Sigma) = \Sigma^{1/2}$). By reordering the input factors, we may assume without loss of generality that $j = d$. Then define $Y = \beta^0 + \beta^T(\mu + C^T Z)$ and $Y' = \beta^0 + \beta^T(\mu + C^T Z')$ with $Z, Z' \sim \mathcal{N}(0, I)$, differing only in their last coordinate independent from each other. Then Y and Y' are identically (but not independently) distributed, and we know from Section 2.2 that $\text{cov}(Y, Y') = \frac{1}{2} \mathbb{E}[(Y - Y')^2] = \frac{1}{2} \beta_d^2 C_{d,d}^2 \mathbb{E}[(Z_d - Z'_d)^2]$. But Z_d and Z'_d are iid. standard normal, i.e., $\frac{1}{2} \mathbb{E}[(Z_d - Z'_d)^2] = 1$. We are left with the identification of the last diagonal entry of the Cholesky matrix.

Because of its hierarchical triangular structure, it keeps subdeterminants intact, and $C_{d,d} = \sqrt{\frac{\det \Sigma}{\det \Sigma_{-d,-d}}}$. \square

B Software Availability

The repository

<https://gitlab.gwdg.de/elmar.plischke/global-sensitivity-analysis-collection/>

contains implementations of the generalized winding stairs total estimator (`windsi.m`, `windsi.R`), of the mix-and-reweight U-statistics estimator (`totalsdep.m`), and of the derange-and-reweight and shift-and-reweight pick-and-freeze estimators (`totalsdep_pnf.m`, `totalsdep_pnfshift.m`) for use with MATLAB/OCTAVE or R.