Probabilistic Sensitivity with Optimal Transport

Emanuele Borgonovo[1], Alessio Figalli[2], Elmar Plischke[3], Giuseppe Savarè[1]
[1]Department of Decision Sciences, Bocconi University
[2]Department of Mathematics, ETH Zürich
[3]Faculty of Energy and Economic Sciences, Clausthal University of Technology

Abstract

The theory of optimal transport and the use of Wasserstein-type metrics are attracting increasing attention in statistics and machine learning. At the same time, the definition of measures of statistical association for multivariate responses is an open research topic, especially for feature selection in computer experiments. This work examines the construction of probabilistic sensitivity measures using the theory of optimal transport. We obtain a new family of indicators that are global, well posed in the presence of correlations and possess the zero-independence property. Closed form expressions are derived for the family of elliptical distributions. We study the connection between measures based on the Wasserstein-Bures approximation and previously introduced generalized variance-based indicators. For estimation, we employ a one-sample strategy that keeps computational burden under control. We prove the asymptotic consistency of the estimators. We compare estimators based on alternative algorithmic approaches developed in the machine learning literature for the solution of optimal transport problems. Findings show that consistent estimates are obtained at reasonable sample sizes and fast execution times.

1 Introduction

The pervasive use of scientific simulators requires approaches to transparently account for uncertainty in assumptions and inputs [Cockayne et al., 2019, Owhadi et al., 2013]. As part of uncertainty quantification, probabilistic sensitivity methods assist scientists in determining the key-drivers of output variability [Saltelli and Tarantola, 2002, Oakley and O'Hagan, 2004]. Although a rich set of tools and a consolidated theory are available for univariate responses, the analysis becomes challenging when the output is a multivariate random vector (see Marrel et al. [2017]). This problem is related to the definition of measures of statistical dependence in a multivariate or functional response context. Recent works such as Pan et al. [2020] and Chatterjee [2020] show a renewed interest in this topic. We investigate the use of optimal transport (OT) theory [Figalli and Glaudo, 2021] for defining probabilistic sensitivity measures for multivariate outputs. We discuss theoretical aspects first, with particular focus on positivity and the zero-independence property. We then address the analytical representation of the sensitivity measures via the

Wasserstein-Bures metric, showing that it is exact if the input and output distributions are elliptical. The resulting dependence measures can be decomposed into an advective and a diffusive contribution, with the advective part proportional to the generalized variance-based sensitivity indices of Gamboa et al. [2014]. We then address computation, defining given-data estimators that rely on Pearson [1905] partition-based design. From an algorithmic viewpoint, the design implies the solution of one OT-problem for each partition. This aspect, together with the related problem sizes make the estimation process a challenge for OT-solving algorithms, whose development is a hot topic in machine learning [Altschuler et al., 2019, Janati et al., 2020]. We test OT-solvers representatives of different families, namely, Puccetti [2017]'s proposal based on a partial orderings, two versions of Cuturi [2013]'s Sinkhorn-based entropic solver, and a direct implementation through the Wasserstein-Bures approximation. We prove asymptotic consistency of the estimators. We report results for several experiments on analytical test cases as well as on a well-known realistic simulator. Findings indicate that all the employed algorithms yield consistent estimates at reasonable sample sizes. Implementations are time-wise fast, with one exception for the Sinkhorn algorithm with numerical stabilization.
The remainder of the manuscript is organized as follows. Section 2 reviews related works on optimal transport theory and probabilistic sensitivity measures. Section 3 introduces the new family of OT-based dependence measures. Section 4 presents given-data estimators. Section 5 presents numerical experiments for univariate and multivariate analytical test cases. Section 6 presents results for a realistic application, the Bliznyuk et al. [2008] environmental model. Section 7 concludes the work. Appendix A reports all proofs.

2 Related Literature and Theoretical Foundations

The literature on optimal transport and probabilistic sensitivity analysis is broad and a review of both fields is outside our reach. In this section, we offer a concise overview due to space limitations, starting with optimal transport in Section 2.1.

2.1 Optimal Transport and Wasserstein Distances

Optimal transport (OT, henceforth) is a topical research subject intensively investigated in mathematics, statistics and machine learning. We refer to the monographs of Villani [2009], Panaretos and Zemel [2020], Peyré and Cuturi [2019], Figalli and Glaudo [2021] and the survey of Chen et al. [2021] for a detailed treatment of the theory and computational aspects, while we present a concise overview of the main principles.
Let $\mathcal{Y}$ be a Polish space (i.e., a separable topological space whose topology is induced by a complete metric), and denote by $\mathscr{P}(\mathcal{Y})$ the space of Borel probability measures on $\mathcal{Y}$. If $(\Omega, \mathcal{F}, \mathbb{P})$ is a standard Borel space and $Y : \Omega \to \mathcal{Y}$ is a (Borel) random variable, we will denote by $Y_\sharp\mathbb{P} = \mathbb{P}_Y \in \mathscr{P}(\mathcal{Y})$ the law or push-forward of Y in $\mathcal{Y}$, defined by $\mathbb{P}_Y(A) = Y_\sharp\mathbb{P}(A) = \mathbb{P}[Y \in A]$ for every Borel set $A \in \mathscr{B}(\mathcal{Y})$. If $\mathcal{Z}$ is another Polish space, $\nu \in \mathscr{P}(\mathcal{Y})$ and $\nu' \in \mathscr{P}(\mathcal{Z})$, we denote by $\Pi(\nu, \nu')$ the set of plans or couplings $\pi \in \mathscr{P}(\mathcal{Y}\times\mathcal{Z})$ whose marginals are ν and ν' respectively, i.e. $\nu = \mathrm{p}^{\mathcal{Y}}_\sharp\pi$, $\nu' = \mathrm{p}^{\mathcal{Z}}_\sharp\pi$, where $\mathrm{p}^{\mathcal{Y}}(y, z) = y$ and $\mathrm{p}^{\mathcal{Z}}(y, z) = z$. Given a (lower semicontinuous) cost function $c : \mathcal{Y} \times \mathcal{Z} \to [0, +\infty]$, the

Kantorovich formulation of the optimal transport problem

$$K(\nu,\nu') = \inf_{\pi\in\Pi(\nu,\nu')} \mathcal{C}(\pi) \tag{1}$$

consists of finding a transfer plan $\pi \in \Pi(\nu,\nu')$ minimizing the integrated cost $\mathcal{C}(\pi) = \int c(y,z)\,\mathrm{d}\pi(y,z)$. Whenever a transfer plan $\pi \in \Pi(\nu,\nu')$ with finite cost $\mathcal{C}(\pi) < \infty$ exists, it can be shown that the Kantorovich problem has (at least) a solution attaining the minimum of $\mathcal{C}$ in $\Pi(\nu,\nu')$. If $\nu = \sum_{i=1}^{I} m_i\delta_{y_i}$, $\nu' = \sum_{j=1}^{J} n_j\delta_{z_j}$ are discrete measures (where $m_i, n_j \geq 0$, $\sum_{i=1}^{I} m_i = \sum_{j=1}^{J} n_j = 1$) then

$$\Pi(\nu,\nu') = \left\{\pi = \sum_{i,j} p_{ij}\delta_{(y_i,z_j)} : p_{ij} \geq 0,\ \sum_j p_{ij} = m_i,\ \sum_i p_{ij} = n_j\right\} \tag{2}$$

is isomorphic to a bounded, closed, and convex polytope in $\mathbb{R}^{I\times J}$ and the Kantorovich problem amounts to solve the linear program

$$K(\nu,\nu') = \min\left\{\sum_{ij} c(y_i,z_j)p_{ij} : p_{ij} \geq 0,\ \sum_j p_{ij} = m_i,\ \sum_i p_{ij} = n_j\right\}. \tag{3}$$

A particular case occurs when $I = J = N$ and $m_i = n_j = \frac{1}{N}$: an application of Birkhoff's theorem yields

$$K(\nu,\nu') = \min\left\{\frac{1}{N}\sum_{i=1}^{N} c(y_i,z_{\sigma(i)}) : \sigma \in \mathrm{Sym}(N)\right\}, \tag{4}$$

where $\mathrm{Sym}(N)$ is the symmetric group of all the permutations of the first N integers. From (4), the solution of the linear program (3) is therefore a vertex of a high-dimensional polytope, i.e. it is always possible to find a permutation that solves the minimization problem.
When $\mathcal{Y} = \mathcal{Z}$ and $c(y,z) = \mathsf{d}^p(y,z)$ for a suitable continuous distance $\mathsf{d} : \mathcal{Y}\times\mathcal{Y} \to [0,+\infty)$, the p-th root of the optimal Kantorovich cost

$$W_p(\nu,\nu') = \sqrt[p]{\inf_{\pi\in\Pi(\nu,\nu')} \int \mathsf{d}^p(y,z)\,\mathrm{d}\pi(y,z)} \tag{5}$$

defines the so-called p-Wasserstein distance, which satisfies the axioms of a metric in the subset $\mathscr{P}_p(\mathcal{Y})$ whose measures have finite p-th moment $\int_{\mathcal{Y}} \mathsf{d}^p(y,y_0)\,\mathrm{d}\nu(y) < \infty$ for some (and thus any) $y_0 \in \mathcal{Y}$. Particularly important is the case $\mathcal{Y} = \mathbb{R}^{n_Y}$, $n_Y \in \mathbb{N}$, endowed with the Euclidean distance (or, more generally, the distance induced by a norm $\|\cdot\|$ in $\mathbb{R}^{n_Y}$).
Whenever $\nu = \mathbb{P}_Y$ and $\nu' = \mathbb{P}_Z$ then $\pi = \mathbb{P}_{(Y,Z)} = (Y,Z)_\sharp\mathbb{P}$ is a coupling between ν and ν' so that

$$W_p^p(\nu,\nu') \leq \mathbb{E}\Big[\mathsf{d}^p(Y,Z)\Big], \tag{6}$$

and using the dual formulation we also get

$$W_p^p(\nu,\nu') = \sup_{f,g}\Big\{\mathbb{E}\left[f(Y)\right] + \mathbb{E}\left[g(Z)\right] : f,g \in \mathrm{C}_b(\mathcal{Y}), \tag{7}$$

$$f(y) + g(z) \le \mathsf{d}^p(y,z) \text{ for all } y,z \in \mathcal{Y}\Big\}, \tag{8}$$

where the expressions $\mathbb{E}[f(Y)]$ e $\mathbb{E}[g(Z)]$ do not depend on Y and Z but only on ν,ν'. On the other hand, if $\mathbb{P}$ has no atoms then it can be shown that

$$W_p^p(\nu,\nu') = \inf_{X,Y}\Big\{\mathbb{E}\left[\mathsf{d}^p(Y,Z)\right] : Y_\sharp\mathbb{P} = \nu,\ Z_\sharp\mathbb{P} = \nu'\Big\}. \tag{9}$$

As a metric, the p-Wasserstein distance is actively studied and applied. We recall Wang et al. [2020] for an application in genomics, Puccetti et al. [2020] for recent results of Wasserstein baricenters, Deb and Sen [2021] for the use of the Wasserstein metric for statistical testing in high-dimensional settings. Panaretos and Zemel [2020] presents extensive details on the properties and applications of the Wasserstein metric in mathematics, statistics and artificial intelligence. We also recall the work of Berthet et al. [2020], that recently derive central limit theorems for univariate p-Wasserstein distances. For $n_Y = 1$, we have explicit formulations of W_p in terms of the cumulative distribution functions $F_\nu, F_{\nu'}$ induced by $\nu,\nu' \in \mathscr{P}(\mathbb{R})$ and their (pseudo-) inverses, the quantile functions. In fact, the cumulative distribution function provides the optimal transport map (also called the Monge map) to the standard uniform distribution, i.e., the Lebesgue measure λ on $[0,1]$, and conversely, the quantile function is the optimal transport map from the standard uniform distribution. It is a remarkable fact that they also provide the solution to the Kantorovich problems for a pair of measures.

Theorem 1 (Bobkov and Ledoux [2016])**.** *Let $\nu,\nu' \in \mathscr{P}_1(\mathbb{R})$, let λ be the uniform Lebesgue measure on $[0,1]$, and let us set $F_\nu(y) = \nu\big((-\infty,y]\big)$, $y \in \mathbb{R}$, and $Q_\nu(u) = \inf\big\{y \in \mathbb{R} : F_\nu(y) \ge u\big\}$, $u \in [0,1]$ (using the same notation with ν'). Then the coupling $(Q_\nu, Q_{\nu'})_\sharp\lambda$ belongs to $\Pi(\nu,\nu')$ and it is optimal for all the Wasserstein distances W_p (provided ν, ν' have finite p-moments). Hence, for all $p \ge 1$*

$$W_p^p(\nu,\nu') = \int_0^1 |Q_\nu(u) - Q_{\nu'}(u)|^p \, \mathrm{d}u. \tag{10}$$

In turn, this result yields a computational shortcut that makes the numerical calculation of p-Wasserstein distances straightforward: Given a sample of realizations of univariate random variables $Y \sim \nu$, $Y' \sim \nu'$, an estimate of their p-Wasserstein distance is found from ordering their realizations, taking the p-norm of their differences and averaging.
The multivariate case ($n_Y \ge 2$) is, instead, more difficult to handle from a computational as well as a theoretical viewpoint. Regarding theoretical aspects, we refer to Panaretos and Zemel [2019] for greater details and recall the distributional results for the 2-Wasserstein distance between discrete random vectors obtained in Sommerfeld and Munk [2018], as well as convergence results in Weed and Bach [2017]. Also, it is usually impossible to

obtain a closed form expression for the solution of Problem (1). A notable exception occurs when the involved distributions are elliptical. Let $\mathbb{P}$ and $\mathbb{Q}$ be two elliptical distributions and let $\mu_{\mathbb{P}}$, $\mu_{\mathbb{Q}}$ and $\Sigma_{\mathbb{P}}$, $\Sigma_{\mathbb{Q}}$ denote, respectively, their mean values and variance-covariance matrix, respectively. Then, Givens and Shortt [1984] and Gelbrich [1990] prove that

$$W_2(\mathbb{P},\mathbb{Q}) = \sqrt{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_2^2 + \mathrm{Tr}\left(\Sigma_{\mathbb{P}} + \Sigma_{\mathbb{Q}} - 2\left((\Sigma_{\mathbb{P}})^{1/2}\Sigma_{\mathbb{Q}}(\Sigma_{\mathbb{P}})^{1/2}\right)^{1/2}\right)}, \tag{11}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix and $\Sigma^{1/2}$ is the symmetric matrix square root operator. Note that, because variance-covariance matrices are positive definite for elliptical random variables, the matrix root square operation is well posed and results a unique real-valued positive definite matrix.

The metric in the right-hand side of (11) is the Wasserstein-Bures semimetric [Janati et al., 2020], which we denote by $\mathrm{WB}(\cdot,\cdot)$. More in detail, let $\nu_1, \nu_2, \nu_3 \in \mathscr{P}(\mathcal{Y})$. Then, $\mathrm{WB} : \mathscr{P}(\mathcal{Y}) \times \mathscr{P}(\mathcal{Y}) \to \mathbb{R}$ is such that $\mathrm{WB}(\nu_1,\nu_2) \geq 0$, $\mathrm{WB}(\nu_1,\nu_2) = 0$ if $\nu_1 = \nu_2$, and $\mathrm{WB}(\nu_1,\nu_3) \leq \mathrm{WB}(\nu_1,\nu_2) + \mathrm{WB}(\nu_2,\nu_3)$. Now, let Y and Z be two random variables on $\mathcal{Y}$. Then, $\mathrm{WB}(\mathbb{P}_Y,\mathbb{P}_Z) = 0$ does not imply $Y = Z$, because for $\mathrm{WB}(\mathbb{P}_Y,\mathbb{P}_Z) = 0$ it is sufficient that $\mu_Y = \mu_Z$ and $\Sigma_Y = \Sigma_Z$, although $Y \neq Z$. However, on Gaussian distributions, WB is a metric and coincides with the 2-Wasserstein distance.

Regarding computation, in an influential work Cuturi [2013] proposes to regularize the Kantorovich problems through a penalty term based upon the Kullback-Leibler entropy of π w.r.t. a suitable reference probability measure ϑ

$$\mathrm{KL}(\pi|\vartheta) = \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\vartheta}\right)\mathrm{d}\pi, \quad \text{if } \pi \ll \vartheta, \tag{12}$$

($\mathrm{KL}(\pi|\vartheta) = +\infty$ if π is not absolutely continuous w.r.t. ϑ) added to the minimization problem. A natural choice is $\vartheta = \nu \otimes \nu'$, which yields

$$S_\varepsilon(\nu,\nu') = \inf_{\pi\in\Pi(\nu,\nu')} \mathcal{C}(\pi) + \varepsilon \mathrm{KL}(\pi|\nu\otimes\nu'), \quad \varepsilon \geq 0. \tag{13}$$

(13) is also called the entropic optimal transport problem. As for the Kantorovich problem, the entropic problem admits a dual formulation, which can also be expressed in terms of an arbitrary pair of random variables Y, Z with laws ν, ν':

$$\begin{aligned} S_\varepsilon(\nu,\nu') = \sup_{\substack{f\in C_b(\mathcal{Y}),\\ g\in C_b(\mathcal{Z})}} & \mathbb{E}[f(Y)] + \mathbb{E}[g(Z)] \\ & - \varepsilon\left(\iint \exp\left(\frac{f(y)+g(z)-c(y,z)}{\varepsilon}\right)\mathrm{d}\nu(y)\,\mathrm{d}\nu'(z) - 1\right). \end{aligned} \tag{14}$$

The entropic problem can be solved by a fixpoint iteration using Sinkhorn's algorithm. This functional fixpoint iteration is given by

$$g_{n+1}: \quad z \mapsto -\varepsilon \log \mathbb{E}\left[\exp\left(\frac{f_n(Y) - c(Y,z)}{\varepsilon}\right)\right], \tag{15}$$

$$f_{n+1}: \quad y \mapsto -\varepsilon \log \mathbb{E}\left[\exp\left(\frac{g_{n+1}(Z) - c(y,Z)}{\varepsilon}\right)\right]. \tag{16}$$

In the limit, $f_n \to f_*$, $g_n \to g_*$. The pair (f_*, g_*) then minimizes (14). For $\varepsilon \to 0$ one regains the solution to the Kantorovich OT problem (the pth power of the p-Wasserstein metric when $c = \mathsf{d}^p$). However, this limiting process introduces numerical instabilities. One option might be to use S_ε directly as objective function, thus retaining the solution of the entropic transport for a finite value of ε.
The development of efficient algorithms for the solution of classical and entropic OT problems is a very active research area [Peyré and Cuturi, 2019]. With some conceptual simplification, one can consider three groups of algorithms, based respectively on sorting, linear programming and matrix scaling. The first is inspired by the one-dimensional numerical sorting shortcut. A multivariate algorithm that approximates the 2-Wasserstein distance using iterative swaps is presented in Puccetti [2017]. The algorithm makes use of pairwise comparisons and leads to an approximate solution of the classical OT problem in (1). A second class comprises algorithms that solve the OT-linear program through specializations of the simplex method, which comprise variants of the Hungarian method [Kuhn, 1956], the network flow and the transportation simplex algorithms [Luenberger and Ye, 2016]. These algorithms yield the exact solution of the Kantorovich problem in (1). The third class of algorithms solves the entropic problem in (13). Cuturi [2013] revived interest in the Sinkhorn-Knopp method Knight [2008], yielding a computationally efficient fixpoint algorithm (see Peyré and Cuturi [2019] for a thorough treatment). Variants are discussed in articles such as Altschuler et al. [2017]. These algorithms provide a solution which is exact for the entropic problem (Equation (13)), and that can be regarded as an approximate solution of the classical problem (Equation (1)). For the Gaussian case, formulas for the solution of the exact and entropic OT are available, see Equations (11) and (36). Neglecting the Gaussian assumption, we obtain approximate solutions for the general case, leading to a fourth approach. A comparison of the performance of alternative algorithms (exact and approximated) can be found in the recent work of Dong et al. [2020], to which we also refer for further details on algorithmic aspects.

2.2 Probabilistic Sensitivity Analysis

The recent works of Pan et al. [2019, 2020], Chatterjee [2020] demonstrate a new attention in the statistical literature to the definition of measures of association between random variables. Within such family, probabilistic sensitivity measures play an important role in quantifying the relevance of random covariates on the response of computer simulators Oakley and O'Hagan [2004], Oakley [2009]. The literature is vast, and we refer to the monographs of [Saltelli et al., 2008, Sullivan, 2015], and to the handbook of [Ghanem et al., 2017] for broad overviews.
Let X and Y be random vectors/variables on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathcal{X} \subseteq \mathbb{R}^{n_X}$ and $\mathcal{Y} \subseteq \mathbb{R}^{n_Y}$ respectively. If $\alpha \subset \{1, \ldots, n_X\}$ is an ordered multiindex of length $a = |\alpha|$, we denote by $X_\alpha = \mathrm{p}_\alpha \circ X$ the random vector obtained by X by selecting the components indexed by α through the projection map $\mathrm{p}_\alpha(x_1, \cdots, x_n) = (x_{\alpha(1)}, \cdots x_{\alpha(a)})$.
We say that the mapping $\zeta : \mathscr{P}(\mathcal{Y}) \times \mathscr{P}(\mathcal{Y}) \to [0, +\infty)$ is a separation measurement if $\zeta(\nu, \nu') \geq 0$ and $\zeta(\nu, \nu) = 0$ for all $\nu, \nu' \in \mathscr{P}(\mathcal{Y})$. Now, if $\mathbb{P}_Y$ is the law of Y and $\mathbb{P}_{Y|X_\alpha}$ is the conditional law of Y given X_α, we define the probabilistic sensitivity index of X_α

with respect to Y as

$$\xi_\alpha^\zeta = \mathbb{E}\left[\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})\right]. \tag{17}$$

Equivalently, let $\pi = \mathbb{P}_{(X,Y)} \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$ be the coupling between $\mu = \mathbb{P}_X$ and $\nu = \mathbb{P}_Y$ induced by the two random vectors; denoting by $\mathcal{X}_\alpha = \mathrm{p}_\alpha(\mathcal{X}) \subset \mathbb{R}^a$, we can consider the measure $\mu_\alpha = (\mathrm{p}_\alpha)_\sharp \mu$, the coupling $\pi_\alpha = (\mathrm{p}_\alpha, \mathrm{id}_\mathcal{Y})_\sharp \pi = \mathbb{P}_{(X_\alpha,Y)}$ between μ_α and ν, and the disintegration of $(\nu_{x'})_{x'\in\mathcal{X}_\alpha} \in \mathscr{P}(\mathcal{Y})$ of π_α with respect to μ_α. We eventually get

$$\xi_\alpha^\zeta = \int_{\mathcal{X}_\alpha} \zeta(\nu, \nu_{x'})\,\mathrm{d}\mu_\alpha(x'). \tag{18}$$

One calls $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})$ the inner statistic of ξ_α^ζ. Several global sensitivity measures currently in use can be written in the form of (17) [Borgonovo et al., 2014]. To illustrate, for $n_Y = 1$, setting the inner statistic equal to $\zeta^{\mathrm{V}}(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha}) = (\mathbb{E}[Y] - \mathbb{E}[Y|X_\alpha])^2$, one obtains the well known first order variance-based sensitivity measure

$$\xi_\alpha^V \;=\; \mathbb{E}\left[(\mathbb{E}[Y] - \mathbb{E}[Y|X_\alpha])^2\right] \;=\; \mathbb{V}[\mathbb{E}[Y|X_\alpha]] \;=\; \int \left(\int_\mathcal{Y} y d\nu - \int_\mathcal{Y} y d\nu_x'\right)^2 \mathrm{d}\mu_\alpha(x'). \tag{19}$$

Normalized by $\mathbb{V}[Y]$, ξ_α^V in (19) is Pearson [1905] correlation ratio, and, under input independence, coincides with the first order Sobol' variance-based sensitivity index of X_i [Saltelli and Tarantola, 2002, Liu and Owen, 2006]. Alternatively, if Y is absolutely continuous, one can use the L^1 norm between densities [Borgonovo et al., 2014], writing

$$\xi_\alpha^{L_1} = \mathbb{E}\left[\int_\mathcal{Y} \left|f_Y(y) - f_{Y|X_\alpha}(y)\right| dy\right] = \int \|\nu - \nu_{x'}\|_{\mathrm{TV}} \mathrm{d}\mu_\alpha(x'), \tag{20}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm of a signed measure. Equation (20) is a representative of the family of global sensitivity measures based on Csiszar's divergences proposed in Rahman [2016]. Gamboa et al. [2018] introduce a family of probabilistic sensitivity measures based on the Cramér-von Mises distance, defining

$$\xi_\alpha^{\text{CvM-a}} = \mathbb{E}\left[\int_\mathcal{Y} \left(F_Y(y) - F_{Y|X_\alpha}(y)\right)^2 \mathrm{d}F_Y(y)\right] = \iint \left|F_\nu(y) - F_{\nu_{x'}}(y)\right|^2 \mathrm{d}\nu(y)\,\mathrm{d}\mu_\alpha(x'). \tag{21}$$

The sensitivity measures in Equation (20) and the family of Rahman [2016] naturally extend to the multivariate output case $n_Y \geq 2$. The extension of variance-based indices ξ_α^V has been made systematic in Gamboa et al. [2014] with the introduction of generalized indices based on the trace of the variance-covariance matrix of Y. Recent works that address generalized variance-based Sobol' indices also employing derivative-based methods are Lamboni [2019, 2020]. A variance-based approach is also employed in Alexanderian et al. [2020], where variance-based indices are modified to take into account the temporal variation of the output process variance. The works of Fraiman et al. [2020] and Gamboa et al. [2021] further extend $\xi_\alpha^{\text{CvM-a}}$ to the case in which the output belongs to a Riemannian manifold and to a metric space, respectively. Recently, Fort et al. [2021] address the sensitivity of models with stochastic output using these indices with the Wasserstein distance as a metric. We cannot enter into further details, but this concise review shows that the definition of indices for vectorial outputs is an active research field, motivated by industrial and machine learning applications [Marrel et al., 2011, Lamboni et al., 2011, Marrel et al., 2017, Betancourt et al., 2020]. In the next section, we propose a new family of indicators that exploits the theoretical foundations of optimal transport.

3 Defining Probabilistic Sensitivity Measures with Optimal Transport Theory

This section is divided into three parts. In the first part, we define a family of sensitivity indicators based on the OT framework. In the second, we discuss their properties in the univariate output case. In the third, we address the multivariate output case.

3.1 A family of OT-based indicators

In this subsection, we define a family of probabilistic sensitivity measures based on optimal transport. We follow the notation in Section 2, and consider an OT-problem with cost function $c : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty]$, lower semicontinuous, with $c(y, y') \geq 0$, $c(y, y) = 0$ for any $y, y' \in \mathcal{Y}$, and such that $c(y, y') = 0$ implies $y = y'$.

Lemma 2. *Let $\nu, \nu' \in \mathcal{P}$. The function $K(\cdot, \cdot) : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by* (1) *is a separation measurement and $K(\nu, \nu') = 0 \iff \nu = \nu'$.*

Definition 3. *We call*

$$\xi_\alpha^K = \mathbb{E}[K(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})] = \mathbb{E}\left[\inf_{\pi \in \Pi\left(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha}\right)} \mathcal{C}(\pi)\right] \tag{22}$$

the OT-based global sensitivity measure of X_α with respect to Y.

With the interpretation as cost function, ξ_α^K becomes the expected amount of work needed to optimally connect $\mathbb{P}_Y$ to $\mathbb{P}_{Y|X\alpha}$. Thus, the most important input (group) X_α is the one associated with the highest expected amount of work when we pass from the marginal (and current) probability measure of Y to the new probability measure of Y conditional on receiving information about X_α.

Proposition 4. *Suppose that ξ_α^K. Then, $\xi_\alpha^K \geq 0$ and $\xi_\alpha^K = 0$ if and only if Y and X_α are statistically independent.*

Thus, the family of OT-based sensitivity measures ξ_α^K comply with Renyi's postulate D for measures of statistical dependence Renyi [1959], also called zero-independence property in the recent works of Pan et al. [2019, 2020], Chatterjee [2020]. In particular, choosing as cost function the p^{th}-power of the Euclidean distance and then taking the $1/p$-power of the total cost, $\mathbb{R}^{n_y}$, ξ_α^K becomes:

$$\xi_\alpha^{W_p} = \mathbb{E}[W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})] = \mathbb{E}\left[\inf_{\pi \in \Pi\left(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha}\right)} \left\{\int ||y - y'||^p \, \mathrm{d}\pi(y, y')\right\}^{\frac{1}{p}}\right]. \tag{23}$$

We call $\xi_\alpha^{W_p}$ a p-Wasserstein based sensitivity measure. Additional properties are summarized next.

Proposition 5. *1. If $p \le q$ then $\xi_\alpha^{W_p} \le \xi_\alpha^{W_q}$;*

2. Let $\mathcal{Y}$ be equipped with the discrete metric, i.e., a metric such that for all y, $y' \in \mathcal{Y}$, $\mathrm{d}(y, y') = 0$ if $y = y'$ and $\mathrm{d}(y, y') = 1$ if $y \neq y'$. Then $\xi_\alpha^K = \xi_\alpha^{L_1}$ in (20).

The first property means that if we increase the power p in (23) then we obtain a higher expected cost for moving from $\mathbb{P}_Y$ to $\mathbb{P}_{Y|X_\alpha}$. The second property suggests that $\xi_\alpha^{L_1}$, a well-known moment-independent sensitivity measure, can be reinterpreted as an OT-based sensitivity measure if the output space is equipped with the discrete metric.
The probabilistic sensitivity framework of Equation (17) does not require a functional relationship between Y and X. However, in computer experiments and machine learning, we often seek to determine, or have available, an input output mapping of the form $y = g(x) + \mathcal{E}(x, \omega)$, with $g : \mathcal{X} \to \mathcal{Y}$ and where $\mathcal{E} : \mathcal{X} \times \Omega \to \mathcal{Y}$ is such that, for every value of x, $\mathcal{E}(x)$ is a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, let Y_α denote the simulator output conditional on fixing X_α. We have $Y_\alpha = g(x_\alpha; X_{-\alpha}) + \mathcal{E}(x_\alpha; X_{-\alpha}, \omega)$, where $-\alpha = \{1, 2, \ldots, n_y\} \setminus \alpha$ is the complementary set of α. Clearly, Y has probability measure $\mathbb{P}_Y$ and Y_α has probability measure $\mathbb{P}_{Y|X_\alpha = x_\alpha}$.

Remark 6. *The OT problem* (5) *is solved without taking the p^{th}-root of the cost function in several works. In this case, we write the corresponding sensitivity measure as*

$$\xi_\alpha^{W_p^p} = \mathbb{E}[W_p^p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})] = \mathbb{E}\left[\inf_{\pi \in \Pi\left(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha}\right)} \left\{\int ||y - y'||^p \, \mathrm{d}\pi(y, y')\right\}\right]. \tag{24}$$

Regarding the choice of p, the analyst has available a variety of possibilities. We note that if $y - y' > 1$, the higher the values of p, the more the cost function penalizes points that are further apart. The converse happens if $\mathcal{Y} \subseteq [0, 1]$. In the univariate output case ($n_Y \geq 1$), Vallender [1974] shows that when $p = 1$:

$$W_1(\nu, \nu') = \int_0^1 |Q_\nu(u) - Q_{\nu'}(u)| \, \mathrm{d}u = \int_{-\infty}^{+\infty} |F_\nu(y) - F_{\nu'}(y)| \, \mathrm{d}y, \tag{25}$$

Then, $p = 1$ is a convenient choice and it has been used in industrial applications [Liu and Homma, 2010]. For $n_Y \geq 2$, [Panaretos and Zemel, 2019, Section 2] list several properties that make $p = 2$ a convenient selection. On a historical note, the $2-$Wasserstein metric is also known as the Mallow distance, as it has been independently introduced and studied in Mallows [1972].

3.2 The Univariate Case

In the specific case of univariate responses, $n_Y = 1$, several results are available for the Wasserstein metric (some of them mentioned in the literature review section). In this section, we review some facts useful as a premise to the multivariate case. Recalling the notation of Theorem 1, we write $F_Y = F_{\mathbb{P}_Y}$, $Q_Y = Q_{\mathbb{P}_Y}$.

Proposition 7. *Let $n_Y = 1$ and let Y have finite moments up to order p; then the following hold:*

1. *For any* $p \geq 1$,

$$\xi_\alpha^{W_p} = \mathbb{E}\left[\left(\int_0^1 |Q_Y(u) - Q_{Y|X_\alpha}(u)|^p \,\mathrm{d}u\right)^{1/p}\right] = \mathbb{E}\left[\|Q_Y - Q_{Y|X_\alpha}\|_{L^p(0,1)}\right]. \quad (26)$$

2. *For* $p = 1$,

$$\xi_\alpha^{W_1} = \mathbb{E}\left[\int_y |F_Y(y) - F_{Y|X_\alpha}(y)|\,\mathrm{d}y\right] = \mathbb{E}\left[\int_0^1 |Q_Y(q) - Q_{Y|X_\alpha}(q)|\,\mathrm{d}q\right]. \quad (27)$$

As soon as the conditional and unconditional distributions are known, equations (26) and (27) yield analytical test cases for numerical experiments in the univariate case. These integral expressions can be readily implemented in a computer algebra system (e.g., MATHCAD, MATLAB or MATHEMATICA).
Consider now the univariate Gaussian case. We recall that for univariate normal random variables Y, Z in $\mathbb{R}$ with means μ_Y, μ_Z and standard deviations σ_Y, σ_Z, the 2-Wasserstein distance between $\mathbb{P}_Y$ and $\mathbb{P}_Z$ is given by the Wasserstein-Bures distance, that is

$$W_2(\mathbb{P}_Y, \mathbb{P}_Z) = \sqrt{(\mu_Y - \mu_Z)^2 + (\sigma_Y - \sigma_Z)^2}. \quad (28)$$

Conversely, for Y and Z with expectations μ_Y, μ_Z and standard deviations σ_Y, σ_Z, but not necessarily normally distributed, we have $W_2^2(\mathbb{P}_Y, \mathbb{P}_Z) \geq (\mu_Y - \mu_Z)^2 + (\sigma_Y - \sigma_Z)^2 \geq (\mu_Y - \mu_Z)^2$ Givens and Shortt [1984]. Then, we immediately have that $\xi^{W_2^2} \geq \xi^{\mathrm{V}}$. Moreover, given generic Y_1 and Y_2 with means μ_1, μ_2 and standard deviations σ_1, σ_2, their 2-Wasserstein distance is minimal when they are both normally distributed.

Proposition 8. *Let* $\alpha = \{i\}$ *for simplicity. Letting* $a = (a_1, a_2, ..., a_{n_X})$*, if the input-output mapping is of the form* $Y = aX^T$*, with* $X \sim N(\mu_X, \Sigma_X)$ *with mean* $\mu_X = (\mu_1, \mu_2, \ldots, \mu_{n_X})$*, and variance-covariance matrix* $\Sigma_X = (\sigma_{t,s})$*,* $t, s = 1, 2, \ldots, n_X$*, with* $\sigma_{t,s} = \sigma_{s,t}$*,* $\sigma_{t,t} = \sigma_t^2$ *and* $\sigma_{t,s} = \rho_{t,s}\sigma_t\sigma_s$*, where* $\rho_{t,s}$ *is the correlation coefficient between* X_t *and* X_s*, we have*

$$\xi_i^{W_2} = \mathbb{E}\left[\sqrt{\left(\mu_Y - \mu_{Y|X_i}\right)^2 + \left(\sigma_Y - \sigma_{Y|X_i}\right)^2}\right], \quad (29)$$

with $\mu_Y = a\mu_X^T$,

$$\mu_{Y|X_i = x_i} = \sum_{j=1}^{n_X} a_j \left(\mu_j + (x_i - \mu_i)\frac{\sigma_{i,j}^i}{\sigma_{i,i}^i}\right), \quad (30)$$

$$\Sigma_i^c = [\sigma_{t,j}^i = \sigma_{t,j} - \frac{\sigma_{t,i} \cdot \sigma_{i,j}}{\sqrt{\sigma_{i,i}}}, \quad t, j = 1, 2, \ldots, n_X], \quad (31)$$

and $\sigma_Y^2 = a\Sigma_X a^T$, $\sigma_{Y|X_i}^2 = a\Sigma_i^c a^T$.

Let us explore the link between $\xi_i^{W_2}$ in (29) and variance-based sensitivity measures further. In several works (e.g., Janati et al. [2020]), the square of the 2-Wasserstein distance is used. In the normal distribution case, we can write

$$\xi_i^{W_2^2} = \mathbb{E}\left[\left(\mu_Y - \mu_{Y|X_i}\right)^2 + \left(\sigma_Y - \sigma_{Y|X_i}\right)^2\right] = \xi_i^V + \mathbb{E}[(\sigma_Y - \sigma_{Y|X_i})^2] = \mathrm{Adv}_i^{W_2^2} + \mathrm{Diff}_i^{W_2^2}. \quad (32)$$

The second equality in (32) evidences that the squared 2-Wasserstein importance of X_i in the normal case is equal to the variance-based sensitivity measure of X_i plus a contribution coming from the difference in standard deviations. Hence the OT-based sensitivity measure decomposes into an "advective part" which acts constantly on all point, ($\text{Adv}_i^{W^2}$ in Equation (32)), and may therefore be identified as a movement of the center of gravity, i.e., the mean, and a "diffusive part"($\text{Diff}_i^{W^2}$) leading to a non-directional dispersion of the data. The equality in (32) suggests that, under the assumptions of Proposition 8, the advective and diffusive parts contribute additively to the input importance, if the squared 2-Wasserstein semimetric is considered as a separation measure. This is not true if the 2-Wasserstein metric is considered; decomposing (28) into these two parts, by Jensen's inequality and recalling that $\frac{1}{2}(|a|+|b|) \leq \sqrt{a^2+b^2} \leq |a|+|b|$, leads to

$$\frac{1}{2}\left(\mathbb{E}\left[\left|\mu_Y - \mu_{Y|X_i}\right|\right] + \mathbb{E}\left[\left|\sigma_Y - \sigma_{Y|X_i}\right|\right]\right) \leq \xi_i^{W_2} \leq \mathbb{E}\left[\left|\mu_Y - \mu_{Y|X_i}\right|\right] + \mathbb{E}\left[\left|\sigma_Y - \sigma_{Y|X_i}\right|\right]. \tag{33}$$

Example 9. *With the notation in Proposition 8, consider a linear model* $Y = \sum_{i=1}^{3} a_i X_i$, *with* $a_1 = 4$, $a_2 = -2$ *and* $a_3 = 1$. *Assuming* $\mu_X = (1,1,1)$, $\sigma_i^2 = 1$, *and* $\rho_{i,j} = 0.5$, *the values for* $\xi_i^{W_2}$, $\xi_i^{W_2^2}$, *and the corresponding advective and diffusive parts are reported in Table 1.*

Table 1: $\xi_i^{W_2}$, $\xi_i^{W_2^2}$ and associated decompositions into advective and diffusive parts for Example 9.

	$\xi_i^{W_2}$	$\mathbb{E}\left\|\mu_Y - \mu_{Y\|X_i}\right\|$	$\mathbb{E}\left\|\sigma_Y - \sigma_{Y\|X_i}\right\|$	$\xi_i^{W_2^2}$	$\text{Adv}_i^{W_2^2}$	$\text{Diff}_i^{W_2^2}$
X_1	3.79	2.79	2.21	17.15	12.25	4.90
X_2	0.40	0.40	0.03	0.25	0.25	0.00
X_3	1.75	1.60	0.56	4.31	4.00	0.31

Regarding ranking, Table 1 shows that X_1 *is the most important variable, followed by* X_3 *and* X_2, *under both the Wasserstein and the Wasserstein-squared metrics. The last three columns report the neat decomposition into advective and diffusive contributions showing that for all variables, the advective part is predominant: it accounts for* 71% *of* $\xi_1^{W_2^2}$, *and for more than* 90% *of* $\xi_2^{W_2^2}$, *and* $\xi_3^{W_2^2}$. *The first three columns also show that the advective part plays a major role in the Wasserstein-based importance,* $\xi_1^{W_2}$. *From the values in columns 2-4, note that* $\xi_1^{W_2} - \mathbb{E}\left|\mu_Y - \mu_{Y|X_i}\right| \approx 1$, *while* $\mathbb{E}\left|\sigma_Y - \sigma_{Y|X_i}\right| \approx 2.21$, *due to the non-additivity of the two contributions to* $\xi_1^{W_2}$.

3.3 The Multivariate Case

In the case $n_Y \geq 2$, closed form expressions similar to Equations (26) and (27) are generally not available, and the Wasserstein distance will have to be found solving a corresponding data-driven optimization problem that we discuss in Section 4.2. Nonetheless, in the particular case of elliptical distributions it is still possible to find the distance in a form that generalizes (29). Elliptical distributions have been widely studied in the statistical and actuarial sciences Cambanis et al. [1981], Landsman and Valdez [2003]. A random vector Z is elliptically distributed if its characteristic function can be represented in the form $\phi(z;\mu_Z,\Sigma_Z^*) = e^{iz^T\mu_Z^*}h(z^T\Sigma_Z^* z)$, where $h:\mathbb{R}^+\to\mathbb{R}^+$ is called the characteristic generator (see [Cambanis et al., 1981, Theorem 2] for technical conditions), and μ_Z^* and Σ_Z^* are called the location and dispersion parameters, respectively. One correspondingly writes $Z\sim\mathcal{EC}(\mu_Z^*,\Sigma_Z^*,h)$ where $\mathcal{EC}$ stands for elliptically contoured, as in Cambanis et al. [1981] — however, we will use elliptical, for short, henceforth. Note that, if the first moment, μ_Z, of Z exists then $\mu_Z=\mu_Z^*$; if the second moment exists then the variance-covariance matrix Σ_Z is related to the dispersion parameter Σ_Z^* as $\Sigma_Z=-2h'(0+)$ [Cambanis et al., 1981, Theorem 4], where $h'(0+)$ is the right derivative of the characteristic generator at the origin. The family of Gaussian distributions is retrieved for $h(\cdot)=e^{-\frac{1}{2}(\cdot)}$. The following result combines findings in Gelbrich [1990], Landsman and Valdez [2003].

Proposition 10. *Assume that the second moment of Y is finite. If $\mathbb{P}_Y$ is elliptical with generating function h, and $\mathbb{P}_{Y|X_\alpha}$ is elliptical with the identical generating function h for all values X_α, then $\xi_\alpha^{W_2}=\xi_\alpha^{\mathrm{WB}}$, where*

$$\xi_\alpha^{\mathrm{WB}}=\mathbb{E}\left[\sqrt{\|\mathbb{E}[Y]-\mathbb{E}[Y|X_\alpha]\|_2^2+\mathrm{Tr}\left(\Sigma_Y+\Sigma_{Y|X_\alpha}-2\left(\Sigma_Y^{1/2}\Sigma_{Y|X_\alpha}\Sigma_Y^{1/2}\right)^{1/2}\right)}\right]. \tag{34}$$

Proposition 10 suggests that if, in a probabilistic sensitivity analysis, all the involved distributions are of the same elliptical family then we have a closed form expression of $\xi_\alpha^{W_2}$, that becomes a probabilistic sensitivity measure based on the Wasserstein-Bures metric. If this condition is met, it is not necessary to actually solve the OT-problem in (5) to obtain the Wasserstein distance, but one can directly benefit from the closed form expression in (50).

Equation (50) can be used to expose a relationship between OT-based sensitivity measures and the generalized variance-based sensitivity indices of Gamboa et al. [2014]. Let us restrict attention to the argument under the square root in the right-hand side of (50): it is the sum of an advective part given by L^2 distance between the means and a diffusive part involving the covariance matrices.

Proposition 11. *Let $\alpha=\{i\}$, $i\in\{1,\dots,n_X\}$. For the advective part in the squared Wasserstein-Bures sensitivity measure, we have:*

$$Adv_i^{WB^2}=\mathbb{E}\left[\|\mathbb{E}[Y]-\mathbb{E}[Y|X_i]\|_2^2\right]=\mathbb{E}\left[\sum_{t=1}^{n_Y}(\mathbb{E}[Y_t]-\mathbb{E}[Y_t|X_i])^2\right]=\sum_{t=1}^{n_Y}\xi_i^{V,t}, \tag{35}$$

where $\xi_i^{V,t}$ is the univariate variance-based sensitivity measure (19) *of X_i with respect to Y_t. Moreover, if we assume that the inputs are independent then we have $\mathrm{Adv}_i^{\mathrm{WB}^2}=\sum_{t=1}^{n_Y}\mathbb{V}[Y^t]S_i^t$, where S_i^t is the Sobol' first order sensitivity measure of X_i with respect to Y^t.*

The equality $\mathrm{Adv}_i^{\mathrm{WB}^2} = \sum_{t=1}^{n_Y} \xi_i^{V,t}$ holds for generic input distributions. If we consider that inputs are independent, then $\mathrm{Adv}_i^{\mathrm{WB}^2}$ becomes the numerator of the variance-based global sensitivity measure introduced in Gamboa et al. [2014]. Let us denote this sensitivity measure by S_i^{Gamb}. Then, under independence, we can reinterpret S_i^{Gamb} as the advective part of $\xi_i^{\mathrm{WB}^2}$. Our results show that an OT-based importance measure contains one ore more additional terms compared to a generalized variance-based sensitivity measure. In the case of elliptical distributions this additional term is known, and is the diffusive part. For general distributions the additional terms are not generally known analytically, as the determination of OT-based sensitivity measures involves the full solution of the OT problem.

Corollary 12. *Let $X \sim \mathcal{EC}(\mu_X, \Sigma_X^*, h)$, with finite second moment. For notation simplicity, we consider the case $\alpha = \{i\}$. If $Y = AX + b$, where A is an $n_Y \times n_X$ matrix and $b \in \mathbb{R}^{n_Y}$, then $\xi_\alpha^{W_2} = \xi_\alpha^{\mathrm{WB}}$, with ξ_α^{WB} in Equation (50), where $\mu_Y = A\mu_X + b$, $\Sigma_Y = A\Sigma_X A^T$, $\Sigma_{Y|X_i} = A\Sigma_i^c A^T$, Σ_i^c as defined in Equation (31), and $\mu_{Y_k|X_i} = \sum_{j=1}^{n_X} a_{k,j} \left(\mu_j + (X_i - \mu_i) \frac{\sigma_{i,j}^i}{\sigma_{i,i}^i} \right)$, for $k = 1, 2, \ldots, n_Y$, with $\sigma_{i,j}^i$ as in (31).*
Moreover, if $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and given $\varepsilon \geq 0$ then the corresponding entropic probabilistic sensitivity measure (see Equation (13)) can be written as

$$\xi_\alpha^{S_\varepsilon} = \mathbb{E}\left[\sqrt{\sum_{t=1}^{m} (\mu_{Y,t} - \mu_{Y|X_\alpha,t})^2 + \mathrm{Tr}\left(\Sigma_Y + \Sigma_{Y|X_\alpha} - D_\varepsilon \right) + L(D_\varepsilon, \varepsilon)} \right], \tag{36}$$

where $D_\varepsilon = \left(4\Sigma_Y^{\frac{1}{2}} \Sigma_{Y|X_\alpha} \Sigma_Y^{\frac{1}{2}} + \frac{1}{4}\varepsilon^2 I \right)^{\frac{1}{2}}$, I is the identity matrix, and

$$L(D_\varepsilon, \varepsilon) = \tfrac{\varepsilon}{2} \left(n_Y \cdot (1 - \log(\varepsilon)) + \log\det\left(D_\varepsilon + \tfrac{\varepsilon}{2} I \right) \right). \tag{37}$$

The first part of Corollary 12 suggests that if the model output is a linear transformation of an elliptical input then we obtain closed form solutions for the corresponding OT-problems, with the conditional moments and variance-covariance matrices determined analytically. The second part benefits of a recent result in proven by [Janati et al., 2020] and suggests that, for linear models with normal random variables, we have a closed form solution for the entropic OT-based probabilistic sensitivity measure $\xi_\alpha^{S_\varepsilon}$ as well. Note that, setting $\varepsilon = 0$, one regains (50).

Example 13. *Consider the linear model $\begin{cases} Y_1 = X_1 + 2X_2 + 3X_3 \\ Y_2 = 2X_1 + 5X_2 - X_3. \end{cases}$ With the same distribution for X as in Example 9, Y is now multivariate normal with mean $\mu_Y = (3, 6)$ and variance-covariance matrix $\Sigma_Y = \left(\begin{smallmatrix} 15 & 7.5 \\ 7.5 & 33 \end{smallmatrix} \right)$. In Example 9, we have seen that X_1 is the most important input when the output is Y_1 and X_2 is the most important when the output is Y_2. When we consider the output vector (Y_1, Y_2) obtain the importance of the inputs is reported in Table 2. In these calculations, we used $\varepsilon = 2$ for the entropic OT.*

Table 2: Analytical values of $\xi_i^{W_2}$ and $\xi_i^{S_2}$ for the multivariate test case in Example 13.

Variable	$\xi_i^{W_2}$	$\xi_i^{S_2}$	$\sqrt{\text{Adv}_i^{WB^2}}$	$\xi_i^{W_2}$-$\sqrt{\text{Adv}_i^{WB^2}}$	Entropic Penalty
X_1	6.47	6.95	4.24	2.23	0.48
X_2	6.52	7.05	4.41	2.11	0.53
X_3	2.86	4.17	2.55	0.3	1.31

The second and third columns report ξ_α^{WB}, and $\xi_\alpha^{S_2}$, respectively. The fourth column in 2 reports the root square of the advective contribution of the inputs to ξ_α^{WB}, the fifth column reports the difference $\xi_i^{W_2}$-$\sqrt{Adv_i^{WB^2}}$, which can be interpreted as the residual diffusive part in the input importance, the sixth column the entropic penalty, as difference between $\xi_\alpha^{S_2}$ and ξ_α^{WB}, with $\varepsilon = 2$ in the entropic OT problem. We register $\xi_1^{\text{WB}} \approx \xi_2^{\text{WB}}$, and $\xi_1^{\text{WB}} \approx 2.2 \cdot \xi_3^{\text{WB}}$. In ξ_1^{WB} and ξ_2^{WB}, the advective contributions are about twice the diffusive contributions. For X_3, the diffusive component is 11% of ξ_3^{WB}. In $\xi_i^{S_2}$, the penalty contribution accounts for about 7% and 8% of $\xi_1^{S_2}$, $\xi_2^{S_2}$, and it increases at about 31% for $\xi_3^{S_2}$, respectively.

4 Estimation

The direct implementation of (24) for the estimation of OT-based probabilistic sensitivity measures requires two steps: an appropriate sampling strategy for generating realizations of X and Y that follow the appropriate distributions, and a procedure for obtaining the inner statistic via optimal transport. Regarding sampling, in simulation we have available a double-loop Monte Carlo estimation strategy, that results in a brute force numerical translation of (24). A second class consists of pick-and-freeze designs Gamboa et al. [2016] that grant a more efficient estimation of conditional quantities, or a given-data technique that uses input partitioning, and generalizes the intuition behind correlation ratio estimators of Pearson [1905]. This technique has been used in works such as Strong et al. [2012] for variance-based sensitivity measures and Strong and Oakley [2013] for value-of-information.
A second aspect to consider is the computational algorithm for solving the OT problem and thus quantifying $W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})$. We need to distinguish the case $n_Y = 1$ (univariate output) and $n_Y > 1$ (multivariate output).

4.1 Asymptotic Consistency

We focus on given-data estimation because this approach allows one to quantify probabilistic sensitivity measures from a sample of size N obtained within a plain Monte Carlo uncertainty quantification (thus, we need to run the simulator N times). In this respect, the approach is nominally advantageous when compared to a double loop approach, whose cost is of the order of $n_X \cdot N^2$ simulator evaluations. Also, a given data approach allows one to estimate global sensitivity measures from input-output samples coming from data collection, without the need of a simulator in the loop. The exploration of pick-and-freeze designs is left as an avenue of further research.
We recall that in a given-data context, $\mathbb{P}_Y$ is approximated by the empirical measure on

the Monte Carlo sample of the output, and $\mathbb{P}_{Y|X_\alpha}$ is approximated by subsampling from the available observations, as follows. Let $\mathcal{X}_\alpha$ denote the support of X_α and let $P(\mathcal{X}_\alpha)$ denote a partition of $\mathcal{X}_\alpha$ whose cardinality is n_P, i.e., $P(\mathcal{X}_\alpha) = \{\mathcal{X}_\alpha^1, \mathcal{X}_\alpha^2, \ldots, \mathcal{X}_\alpha^{n_P}\}$ such that $\mathcal{X}_\alpha = \bigcup_{m=1}^{n_P} \mathcal{X}_\alpha^m$, and $\mathcal{X}_\alpha^m \cap \mathcal{X}_\alpha^j = \emptyset$ for all $j \neq m$, and let $\delta_P = \max_{m=1}^{n_P} \operatorname{diam} \mathcal{X}_\alpha^m$ denote the diameter of the partition. Let also $P^N(\mathcal{X}_\alpha) = (\mathcal{X}_\alpha^1(N), \mathcal{X}_\alpha^2(N), \ldots, \mathcal{X}_\alpha^{n_P(N)}(N))$ denote a sequence of partitions indexed by N, and whose cardinality is determined by a function $n_P(N)$ that links the partition size n_P to the sample size N.

Definition 14. *We say that $P^N(\mathcal{X}_\alpha)$ is induced by a proper partition refining strategy if: (i) $P^{N'}(\mathcal{X}_\alpha)$ is finer than $P^N(\mathcal{X}_\alpha)$, whenever $N' > N$; (ii) $\delta_P(N) \to 0$ as $N \to \infty$; and (iii) $n_P(N)$ is increasing in N with $\lim_{N\to\infty} n_P(N) = \infty$ and $\lim_{N\to\infty} \frac{N}{n_P(N)} = \infty$.*

The above definition makes formal the intuition of a partition refining strategy (see Borgonovo et al. [2014], among others). Consider now a sample of size of N of realizations of X and Y, (x_N, y_N), with $x_N = \{x_{1,t}, x_{2,t}, \ldots, x_{N,t}\}$ for $t = 1, 2, \ldots, n_X$ and $y_N = \{y_{1,r}, y_{2,r}, \ldots, y_{N,r}\}$, for $r = 1, 2, \ldots, n_Y$. Consider also $y_N^m = \{y_{1,t}^m, y_{2,t}^m, \ldots, y_{N^m,t}\}$ the conditional realizations of Y given that $X_\alpha \in \mathcal{X}_\alpha^m(N)$, $m = 1, 2, \ldots, n_p(N)$. We can write the given-data estimator of $\xi_\alpha^{W_p}$ as

$$\hat{\xi}_\alpha^{W_p}(N) = \sum_{m=1}^{n_P(N)} \mathbb{P}[X_\alpha \in \mathcal{X}_\alpha^m(N)] W_p(y_N, y_N^m), \tag{38}$$

where $W_p(y_N, y_N^m)$ is the Wasserstein distance between the empirical versions of the measures $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_\alpha \in \mathcal{X}_\alpha^m(N)}$. We report the explicit expression of $W_p(y_N, y_N^m)$ in the next section. In the remainder, the hat $\hat{\cdot}$ indicates an estimator.

Proposition 15. *Consider $\hat{\xi}_\alpha^{W_p}(N)$ in Equation (38). Case 1): X_α is a discrete random variable: If $W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})$ is bounded for all values of X_α then $\lim_{N\to\infty} \hat{\xi}_\alpha^{W_p}(N) = \xi_\alpha^{W_p}$. Case 2): X_α is absolutely continuous: If $W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha})$ is bounded for almost all values of X_α and if $P_M(\mathcal{X}_\alpha; N)$ is a sequence of partitions induced by a proper partition refining strategy then*

$$\lim_{N\to\infty} \hat{\xi}_\alpha^{W_p}(N) = \xi_\alpha^{W_p}. \tag{39}$$

Proposition 15 reassures the analyst that the given-data estimator $\hat{\xi}_\alpha^{W_p}(N)$ is a consistent estimator of $\xi_\alpha^{W_p}$.
If we assume that the influence of an input on the output is restricted to the variation of the first moments and the diffusion contribution, we can resort to the Wasserstein-Bures approximation. A given data estimator of the Wasserstein-Bures importance measure is written as

$$\hat{\xi}_\alpha^{WB} = \sum_{m=1}^{n_P(N)} \frac{N_m}{N} \left(\sum_{t=1}^{n_Y} (\hat{\mu}_{Y,t} - \hat{\mu}_{Y|X_\alpha \in \mathcal{X}_\alpha^m, t})^2 + \operatorname{Tr}\left(\widehat{\Sigma}_Y + \widehat{\Sigma}_{Y|X_\alpha \in \mathcal{X}_\alpha^m} - 2 \left(\sqrt{\widehat{\Sigma}_Y} \widehat{\Sigma}_{Y|X_\alpha \in \mathcal{X}_\alpha^m} \sqrt{\widehat{\Sigma}_Y} \right)^{1/2} \right) \right)^{\frac{1}{2}}, \tag{40}$$

where $\hat{\mu}_{Y,t}$ and $\hat{\mu}_{Y|X_\alpha \in \mathcal{X}_\alpha^m,t}$ are empirical means, $\widehat{\Sigma}_Y$ and $\widehat{\Sigma}_{Y|X_\alpha \in \mathcal{X}_\alpha^m}$ empirical covariance matrices. A similar expression holds for a given-data estimator of $\xi_\alpha^{S_\varepsilon}$ in Equation (36):

$$\widehat{\xi}_\alpha^{S_\varepsilon} = \mathbb{E}\left[\sqrt{\sum_{t=1}^{m}(\widehat{\mu}_{Y,t} - \widehat{\mu}_{Y|X_\alpha,t})^2 + \mathrm{Tr}\left(\widehat{\Sigma}_Y + \widehat{\Sigma}_{Y|X_\alpha} - \widehat{D}_\varepsilon\right) + L(\widehat{D}_\varepsilon, \varepsilon)}\right], \qquad (41)$$

with obvious meaning for the symbol $\widehat{D}_\varepsilon$.

Proposition 16. *Assume that the estimators of the means and variance-covariance matrices in $\hat{\xi}_i^{\mathrm{WB}}$ and $\widehat{\xi}_\alpha^{S_\varepsilon}$ are consistent. If X_α is a discrete random variable then $\hat{\xi}_\alpha^{\mathrm{WB}}$ and $\widehat{\xi}_\alpha^{S_\varepsilon}$ are consistent estimators of ξ_α^{WB} and $\xi_\alpha^{S_\varepsilon}$. If X_α is an absolutely continuous random variable then $\hat{\xi}_i^{\mathrm{WB}}$ and $\widehat{\xi}_\alpha^{S_\varepsilon}$ are consistent provided that $P_M(\mathcal{X}_\alpha; N)$ is a proper partition refining strategy.*

Proposition 16 details the conditions under which the given-data estimators $\hat{\xi}_i^{WB}$ and $\widehat{\xi}_\alpha^{\mathrm{Sh}(\sigma)}$ are consistent. In the next section we provide additional details on the estimators in Equation (38).

4.2 The given-data OT-Problem and Corresponding Algorithms

It is possible to write the OT problem explicitly for each partition in the given-data estimator of Equation (38). With the definitions in the previous section, consider a sample of size N and a partition set $\mathcal{X}_m(N)$ as defined above, we have:

$$\begin{gathered}\inf_{\mathbf{s}} \sqrt[p]{\sum_{k=1}^{N} \sum_{j: x_{j,i} \in \mathcal{X}_\alpha^m(N)} s_{k,j} \sum_{t=1}^{n_Y} (y_{k,t} - y_{j,t})^p} \\ \text{subject to} \\ \sum_{k=1}^{N} s_{k,j} = \tfrac{1}{N}, \quad \sum_{j: x_{j,i} \in \mathcal{X}_\alpha^m(N)} s_{k,j} = \tfrac{1}{N_m}, \quad N_m = \#\{j : x_{j,i} \in \mathcal{X}_\alpha^m(N)\},\end{gathered} \qquad (42)$$

where $\#\{\cdot\}$ denotes cardinality of a set, so that N_m counts the realizations of X_α which are included in $\mathcal{X}_\alpha^m(N)$; note that the realizations $y_{k,t}$ follow $\mathbb{P}_Y$, while the realizations $y_{j,t}$ follow $\mathbb{P}_{Y|X_\alpha \in \mathcal{X}_\alpha^m(N)}$.
The corresponding entropic optimization problem is, then,

$$\begin{gathered}\inf_{\mathbf{s}} \sum_{k=1}^{N} \sum_{j: X_\alpha \in \mathcal{X}_\alpha^m(N)} \left(s_{k,j} \sum_{t=1}^{n_Y} (y_{k,t} - y_{j,t})^2 + \varepsilon \exp\left(-\frac{\sum_{t=1}^{n_Y}(y_{k,t} - y_{j,t})^2}{\varepsilon}\right)\right) \\ \text{such that} \\ \sum_{i=1}^{N} s_{i,j} = \tfrac{1}{N}, \quad \sum_{j: X_\alpha \in \mathcal{X}_\alpha^m(N)} s_{i,j} = \tfrac{1}{N_m}, \quad N_m = \#\{j : x_{j,i} \in \mathcal{X}_\alpha^m(N)\}.\end{gathered} \qquad (43)$$

Following Cuturi [2013], the problem in (43) can be solved by applying the Sinkhorn algorithm.
Note that the number of realizations of Y available for the estimation of the unconditional distribution differs from the number available for the conditional estimation. To illustrate,

suppose that we have $N = 10,000$ and $M(N) = 20$, with equally populated partitions. Then, we have $10,000$ realizations for estimating the marginal distribution while we have $N_m = 500$ realizations for the conditional distributions in each partition. This aspect is particularly relevant for sensitivity measures based on distances between density functions. When kernel density estimation is used, a few realizations of Y in a partition may lead to a too rough approximation of the conditional density. However, if we cast the estimation of the distance between $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X}$ in an OT framework, the issue is avoided, because one works directly with realizations of Y.

The literature is actively studying algorithms to solve problems (42) and (43). The analyst needs to make an up-front choice, depending on whether she aims at an exact or at an approximated solution of the OT-problem. In the first case, she needs to solve Problem (42) in each partition with an exact solver (e.g., the Hungarian method). In the second case, she may consider solving Problem (43) instead of Problem (42), for which faster algorithms are available. These algorithms yield an exact solution of the entropic OT-problem which, for small values of the penalty, can be used as approximation of the original OT-problem. The trade-off is then the one between precision and time.

We examine alternative algorithmic implementations. The first is the sorting algorithm of Puccetti [2017]. The second is a direct encoding of the Sinkhorn iteration algorithm, following the works of Cuturi [2013]. This implementation is computationally fast, but prone to numerical instability. We then propose an alternative implementation, that eliminates this instability by considering a logarithmic transformation of the entropy function. As a further alternative, we consider the Wasserstein-Bures approximation, with the calculation of the given-data estimators in (40) and (41). These estimators are computationally convenient, as they involve only linear algebra operations. All these algorithms have been implemented in corresponding MATLAB subroutines, to allow for a uniform comparison. The codes can be retrieved at `https://github.com/emanueleborgonovo/OTsensitivity`.

5 Numerical Experiments with Analytical Test Cases

This section is divided in two parts. In the first part, we discuss experiments in a univariate setting, in the second part we address a multivariate setting with new analytical test cases as benchmarks.

5.1 Univariate Output Test Case

In the case $n_Y = 1$, the solution of the OT-problem in each partition is simplified by the possibility of using the convenient reordering strategy we have discussed. For illustrative purposes, we report results for the case of an input-output mapping with several inputs, $n_X = 999$, response given by $Y = aX^T$, with $a = [a_1, a_2, \ldots, a_{999}]$ and $X = [X_1, X_2, \ldots, X_{999}]$, with $a_i = 4$ for $i = 1, 2, \ldots, 333$, $a_i = -2$ for $i = 334, 335, \ldots, 666$, $a_i = 1$ for $i = 1, 668, \ldots, 999$. The inputs are correlated normal random variables, with means and standard deviations equal to unity and correlations given by $\rho_{i,j} = 0.5$, $i, j = 1, 2, \ldots, 999$, $i \neq j$. Correspondingly Y is normal, with $\mathbb{E}[Y] = 999$ and $\mathbb{V}[Y] = 5.025 \cdot 10^5$. Table 3 reports analytical results for the OT-based importance measures, both in the form of the Wasserstein-Bures distance in (29) and in the quantile-integral form of (26), which lead to identical results in this case (the expressions have been encoded in MATHCAD).

Table 3: Probabilistic Sensitivity Measures for the $n_X = 999$, $n_Y = 1$ normal 50% correlated test case.

Parameter	ξ_i^{W1}	ξ_i^{W2}	ξ_i^{W8}	$\xi_i^{L^1}$	ξ_i^{V}
$X_1, X_2, \ldots, X_{333}$	433	474	634	0.306	.500
$X_{334}, X_{335}, \ldots, X_{666}$	430	471	628	0.303	.495
$X_{667}, X_{668}, \ldots, X_{999}$	431	472	631	0.305	.498

The second column reports the values of $\xi_i^{W_1}$ for the three inputs. We register similar values of $\xi_i^{W_1}$ across the three groups, a consequence of the input correlations. The same is registered for the other sensitivity measures in Table 3. Note that $\xi_\alpha^{W_1} \leq \xi_\alpha^{W2} \leq \xi_i^{W_8} \leq \xi_i^{W_\infty}$ for all i, in line with item 1 of Proposition 5. For comparison, Table 3 also reports the values for two already introduced global sensitivity measures based on the L^1 norm and on the contribution to variance, respectively. Note that the all sensitivity measures agree on the input ranking. Figure 1 reports results for the numerical estimates at increasing sample sizes.

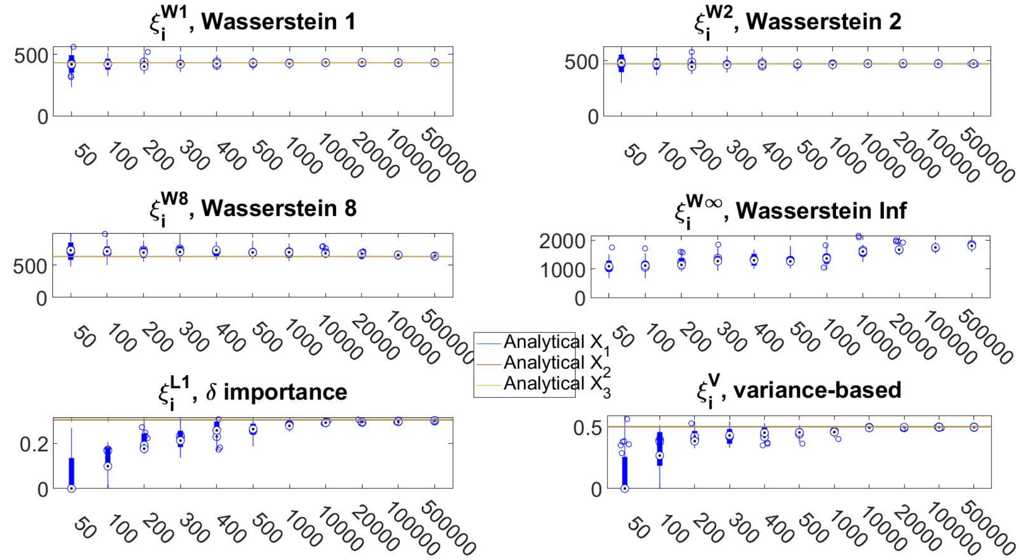


Figure 1: Numerical estimation of probabilistic sensitivity measures for the $d = 999$ correlated normal random variables test case. Horizontal axis: size of the Monte Carlo Sample. Vertical axes: values of the probabilistic sensitivity measure.

The input samples of sizes ranging from $N = 50$ to $N = 500000$ are generated with crude Monte Carlo. We use the partition refining strategy

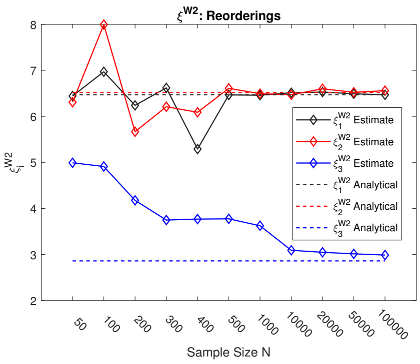
$$M(N) = \min\left\{\left\lceil \frac{N^2}{7 + \tanh(\frac{1500-N}{500})} \right\rceil, 48\right\}. \quad (44)$$

At each N perform 20 replicates of the experiment. The whole analysis takes 525 seconds on a personal computer, with Intel(R) Core(TM) i7-7700HQ CPU, 2.80GHz processor and 64GB RAM.

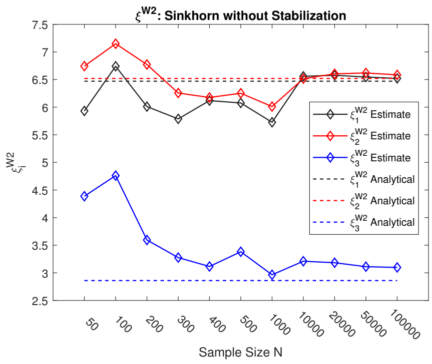
Figure 1 shows that all estimates tend to the corresponding analytical values as the sample size increases. In the fourth panel, we report values for $\xi_{\alpha}^{W_\infty}$, for which, however, analytical values cannot be obtained. Nonetheless, we register $\xi_{\alpha}^{W_\infty} \approx 1850$, higher than the ones of the other $\xi_{\alpha}^{W_p}$ with lower values of p, consistently with item 1 in Proposition 5. Overall, these experiments (and additional ones we performed not reported for brevity) show that the 1-dimensional computation of OT-based sensitivity measures is straightforward.

5.2 Multivariate Normal Output: Numerical Experiments

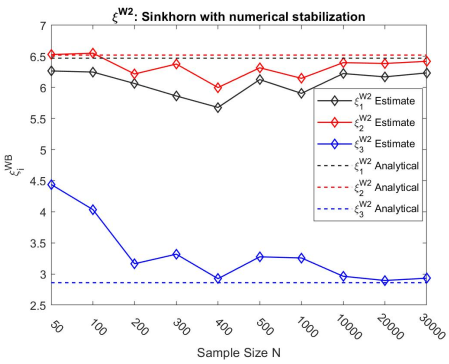
In this section, we consider the same inputs and outputs of Example 13. We compare results for estimation of ξ^{W_2} implementing the sorting algorithm of Puccetti [2017], the Sinkhorn algorithm with and without provisions for numerical stability, and the Wasserstein-Bures estimator in (40) [see the end of Section 4.2]. Our main goal is to investigate asymptotic behavior and timing.



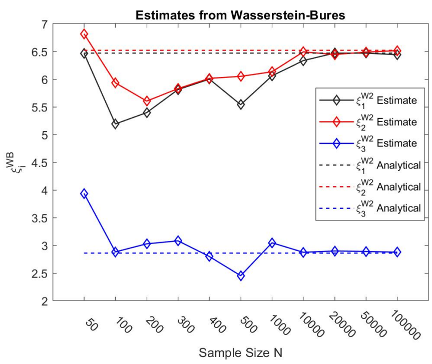
(a) $\hat{\xi}_i^{W_2}$ with the sorting algorithm of Puccetti [2017].



(b) $\hat{\xi}_i^{W_2}$ after solving entropic transport via Sinkhorn without stabilization.



(c) $\hat{\xi}_i^{W_2}$ after solving entropic transport via Sinkhorn with stabilization.



(d) $\hat{\xi}_i^{WB}$ via Wasserstein-Bures.

Figure 2: Multivariate-output analytical test case: results for alternative estimators.

Figure 2 reports results at increasing sample sizes. The starting sample size is $N = 50$, while the end sample size depends on the algorithm; we use $M(N)$ in (44). Figure 2(a) shows the estimates obtained implementing the sorting algorithm of Puccetti [2017]. Overall (i.e., from sample N=50 to N=100,000), the analysis takes about 4 minutes on the above mentioned personal computer. The estimates converge towards the analytical values as the sample size increases. At $N = 500$ the ranking of the inputs is correctly identified, and the estimation error for the two most important inputs is negligible. For all $N \geq 500$, the values of $\widehat{\xi}_1^{W_2}(N)$ remain close but larger than those of $\widehat{\xi}_2^{W_2}(N)$ with no ranking reversals, $\widehat{\xi}_3^{W_2}(N)$ remains consistently lower than $\widehat{\xi}_2^{W_2}(N)$ and $\widehat{\xi}_1^{W_2}(N)$.

Figure 2(b) displays results when the fast implementation of the Sinkhorn algorithm is used. In this case, the code solves one entropic OT problem in (43) per partition. To illustrate at $N = 10000$ we have $M = 22$ partitions, and 3 inputs, which leads to a total of 66 optimal transport problems of size 454×10000 to be solved. The average time to solve one of these problems with the fast implementation is 0.51 seconds, so that the overall time is about 36 seconds. At the largest sample size $N = 100,000$, we have $M = 48$ partitions, and we need to solve 144 problems of size 2128×100000. The average solution time is 20.36 seconds and overall 47 minutes are needed. Similarly to the results in Figure 2(a), the ranking of the inputs is identified already at the smallest sample size (in which the the analysis takes less than a second), and does not changes as the sample size increases. At the largest sample size, the estimation errors are of the order of 0.8%, 1% and 8%, for $\xi_1^{W_2}$, $\xi_2^{W_2}$ and $\xi_3^{W_2}$, respectively.

Figure 2(c) displays results when the Sinkhorn algorithm with numerical stabilization is used. The sample sequences ends at $N = 30000$, because the calculations at larger sample sizes fail due to an out-of-memory problem. At $N = 30000$, we have 96 optimal transport problems of size 937×30000. The average time for the solution of each problem is 511 seconds per problem, for a total of about 14 hours. At this sample size, the errors are at about 3.7%, 1.5% and 2.5%, respectively for $\xi_1^{W_2}$, $\xi_2^{W_2}$ and $\xi_3^{W_2}$. However, note that the inputs are correctly ranked at all sample sizes, even at the smallest ones.

Figure 2(d) displays results when the Wasserstein Bures estimator in Equation (40) is used. One notes a convergence to the analytical values, with percentage error in the estimate approaching zero for $N \geq 10000$. The time required to carry out the entire analysis from $N = 50$ to $N = 100000$ is about 0.6 minutes.

6 An Environmental Test Case

This sections applies the method to the environmental simulator in Bliznyuk et al. [2008], to which we refer for a full description of the problem that originates this model. We make use of the MATLAB code available at `https://www.sfu.ca/~ssurjano/environ.html`. The model simulates the dispersion of chemical pollutants in the soil at S spatial locations in T points in time as a function of four inputs: the mass of pollutant spilled at each location (Ma), the diffusion rate in the channel (D), the location of the second spill (Lo), and the time of the second spill (τ). These inputs are assigned uniform distributions over the ranges $\text{Ma} \in [7, 13]$, $\text{Di} \in [.02, 0.12]$, $\text{Lo} \in [0.01, 3]$, and $\tau \in [0.01, 30.295]$. Regarding the output grid, the default option foresees five locations and 200 points in time, so that

$n_Y = 1000$. We run the simulator with a $N = 30000$ quasi-Monte Carlo sample based on Halton sequences Owen [2006] to perform an initial uncertainty quantification. From this sample, we estimate the variance-based sensitivity measure of the four inputs at each location, through a given-data strategy, with $M = 32$ partitions (Figure 3).

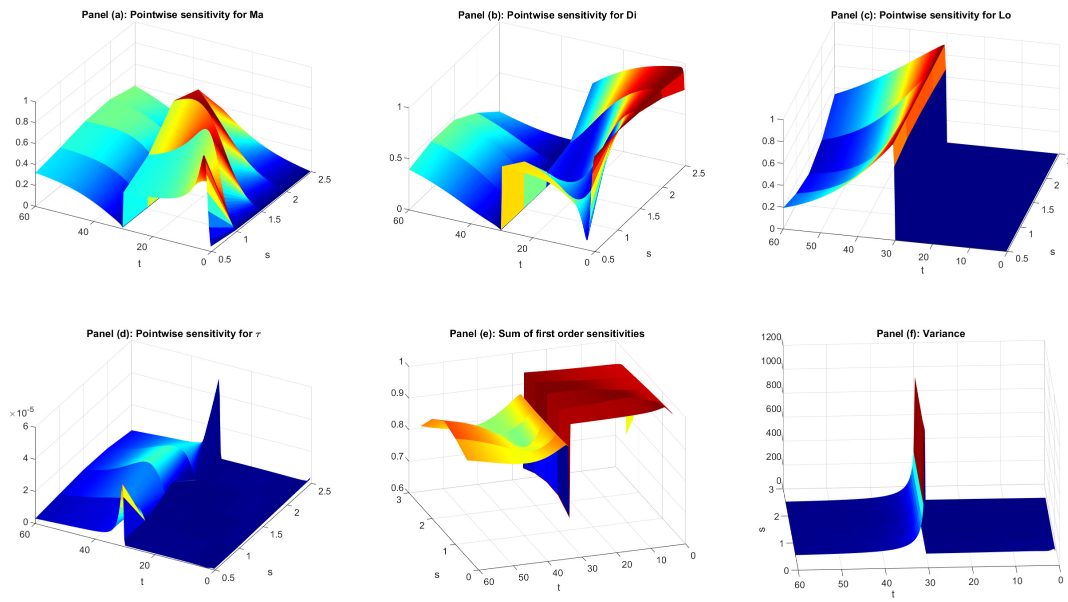


Figure 3: Sensitivity Maps for $\xi_i^{V,q}/\mathbb{V}[Y^q] = S_i^q$, $q = 1, 2, \ldots, 1000$, $i = 1, 2, 3, 4$.

Panels (a)-(d) in Figure 3 display the sensitivity measures $\xi_i^{V,q}/\mathbb{V}[Y^q]$, $q = 1, 2, \ldots, 1000$ for each time and space location for the four inputs. One notes the sharp discontinuity in the values of $\xi_i^{V,q}/\mathbb{V}[Y^q]$ for all inputs at $t = 30$. This discontinuity is generated by the emergence of the second pollutant spill. Panel (a) displays the sensitivity map for input Ma; one observes an increase in the importance in this input up to time $t = 30$, then a sudden decrease and a further increase for $t > 30$. Panel (b) Di; one observes a decrease with time of the importance of this input, till $t = 30$; a further sudden decrease at $t = 30$ and an increase for $t > 30$. Panel (c) shows that the location of the second spill (Lo), is inactive up to $t = 30$, then has a very high value of the variance-based sensitivity measure, being almost the only key-driver of model output variability around $t = 30$. Its importance then decreases as t increases. Panel (d) shows that τ has an overall lower importance than the other three inputs. Panels (e) and (f) display, respectively, the sum of the variance-based sensitivity measures and the spatio-temporal behavior of the model output variances. Panel (e) shows that the sum of the sensitivity measures is close to unity at almost all times preceding $t = 30$, has a discontinuity at $t = 30$ and is lower then unity for $t > 30$. This result means that, with the exclusion of the discontinuity, the output response is close to be additive on a global scale before $t = 30$ and then interactions emerge after such time. Panel (f) evidences the jump in model output variance at $t = 30$. This point-wise analysis provides information on the spatio-temporal behavior of the model response, but does not yield a univocal indication of the overall importance of the inputs. To obtain this indication, we compute the OT-based sensitivity measures. We rely on the same algorithmic approaches of Section 5.2. The reordering algorithm, the fast Sinkhorn implementation and the Wasserstein-Bures estimators allow us to consider the entire sample of size $N = 30000$. With 4 inputs and 32 partitions, the algorithms take 14, 15 and 3 minutes, respectively. The Sinkhorn implementation with numerical stabilization requires 54 hours, with a sample of size $N = 15000$ and 17 partitions, for a total of 68 OT-problems solved.

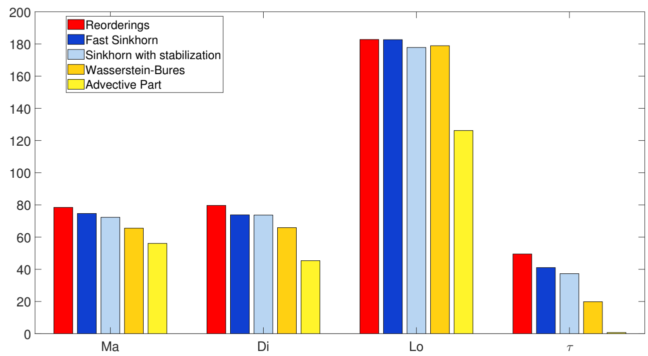


Figure 4: Estimates of ξ_i^{W2} with alternative methods.

The final estimates of the algorithms are, indeed, similar, leading to $\hat{\xi}_{\text{Ma}}^{W_2} \approxeq 72$, $\hat{\xi}_{\text{Di}}^{W_2} \approxeq 73$, $\hat{\xi}_{\text{Lo}}^{W_2} \approxeq 179$, $\hat{\xi}_{\tau}^{W_2} \approxeq 35$. Thus, the location parameter Lo is consistently identified as the most important input, followed by Ma, Di, and with τ playing a minor role. This last result confirms the visual impression of Panel (d) in Figure 3, while the much higher relevance of the location parameter with respect to the mass and diffusion parameters was not granted by a visual inspection of Panels (a), (b) and (c) in Figure 3. This high relevance is attributable to the abrupt change in system behavior caused by the insurgence of the second spill. The fifth bar in each group of Figure 4 reports the values of $\sqrt{\text{Adv}_i^{WB^2}}$, whose square, as we have shown, differs for a normalization factor from S_i^{Gamb}. We observe that the advective part (i.e., the variance-based contribution) accounts for about 0.86%, 0.69%, 0.70% of $\hat{\xi}_{\text{Ma}}^{WB}$, $\hat{\xi}_{\text{Di}}^{WB}$, $\hat{\xi}_{\text{Lo}}^{WB}$, respectively, while it accounts for only 3% of $\hat{\xi}_{\tau}^{WB}$.

7 Conclusions

The theory of optimal transport nested into a probabilistic sensitivity analysis frame leads to an elegant approach for the sensitivity analysis of multivariate responses. The resulting approach applies to samples generated by computer simulators or retrieved by data collection, and does not require the assumption that inputs are independent. The associated dependence measures comply with Renyi's postulate D for measures of statistical dependence and analytical expressions become available when the input and output distributions are elliptical. In this case, we have shown that the separation measurement becomes the Wasserstein-Bures metric and that its advective part measure coincides with the variance-based sensitivity indices of Gamboa et al. [2014]. Exploiting the continuity of the Wasserstein metric, we have proven that given-data estimators are asymptotically consistent. The sequential nature of the estimation challenges OT-solvers, evidencing a trade-off between accuracy and time. However, the advances in the algorithmic solution of OT-problems recently achieved in machine learning lead to fast implementations, with very promising results. In particular, we have seen that given-data estimators based on the Sinkhorn iteration, on reorderings and on the Wasserstein-Bures approximation yield accurate estimates at reasonable sample sizes and with fast execution times. These results pave the way to the application of OT-based sensitivity measures within uncertainty quantification and feature selection for multivariate output contexts. Possible further research avenues comprise, on the one hand, the comparison and integration with existing approaches for multivariate output sensitivity and the application to large dimensional data-driven problems in machine learning and industrial applications.

A Proofs

Proof of Lemma 2. By (1), we have that $K(\nu,\nu') = \inf_{\pi\in\Pi(\nu,\nu')}\mathcal{C}(\pi)$ is greater than or equal to zero, and, in particular, $K(\nu,\nu')$ is null if $\nu=\nu'$. Thus, $K(\nu,\nu')$ complies with the definition of separation measurement. In addition, if $K(\nu,\nu')=0$ for some $\pi^*\in\Pi(\nu,\nu')$, then it must be $\mathcal{C}(\pi^*) = \int c(y,y')\,\mathrm{d}\pi^*(y,y') = 0$. Because the integrand is non-negative, by the monotonicity of integrals it must hold that $c(y,y')=0$ on a set of non-null measure with respect to π^*. Then, because $c(y,y')=0$ implies $y=y'$, it holds that $\nu=\nu'$. ∎

Proof of Proposition 4. The first property, $\mathbb{E}[K(\mathbb{P}_Y,\mathbb{P}_{Y|X})]\geq 0$, follows immediately from the fact that $K(\mathbb{P}_Y,\mathbb{P}_{Y|X})\geq 0$. For the second property, independence of Y and X_α is equivalent to $\mathbb{P}_{Y|X_\alpha}=\mathbb{P}_Y$. Hence, applying Lemma 2 shows that $\xi^K_\alpha = \mathbb{E}[K_\alpha(\mathbb{P}_{Y|X_\alpha},\mathbb{P}_Y)] = 0 \iff Y$ and X_α are independent. ∎

Proof of Proposition 5. **Item 1**. By [Panaretos and Zemel, 2020, Equation (2.1)], Wasserstein metrics satisfy $W_p(\mathbb{P}_Y,\mathbb{P}_{Y|X})\leq W_q(\mathbb{P}_Y,\mathbb{P}_{Y|X})$ if $p\leq q$. Then, taking the expectations of both sides leads to the desired result. **Item 2**. Let Y be absolutely continuous. Let $(\mathcal{Y},d)$ be a metric space with the discrete metric. Also, let $\mathbb{P}$, $\mathbb{Q}$ be two probability measures on $(\Omega,\mathcal{F})$, with densities $f_{\mathbb{P}}(y)$, $f_{\mathbb{Q}}(y)$. Dobrushin [1970] proves that in the case Y is univariate and the metric for the optimal transport problem is the discrete metric then the Wasserstein metric is given by $W(\mathbb{P},\mathbb{Q}) = \sup_{B\in\mathcal{B}(\Omega)}|\mathbb{P}(B)-\mathbb{Q}(B)|$. Then, by Scheffè's theorem Scheffé [1947], we have

$$W(\mathbb{P},\mathbb{Q}) = \sup_{B\in\mathcal{B}(\Omega)}|\mathbb{P}(B)-\mathbb{Q}(B)| = \int_{\mathbb{R}}|f_P(y)-f_Q(y)|dy. \tag{45}$$

Hence, $\xi^{L^1}_X$ is an OT-based sensitivity measure between density functions. ∎

Proof of Proposition 7. The proof follows straightforwardly by combining (25) and (10) with (17). ∎

Proof of Proposition 8. If $X \sim N(\mu_X, \Sigma_X)$ and $Y = aX^T + b$, then $Y \sim N(\mu_Y, \sigma_Y)$ with $\mu_Y = a\mu_X^T$ and $\sigma_Y^2 = a\Sigma_X a^T$. Also, $Y|X_i \sim N(\mu_{Y|X_i}, \Sigma_{Y|X_i})$, with $\mu_{Y|X_i}$ given in (30) and $\Sigma_{Y|X_i} = \Sigma_i^c$, in Equation (31). Therefore, Y and $Y|X_i$ are normal random variables for all values of X_i. Consequently, by [Givens and Shortt, 1984, Proposition 7], we have that $W_2(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})$ is the Wasserstein-Bures distance which is given by

$$W_2(\mathbb{P}_Y, \mathbb{P}_{Y|X_i}) = \mathrm{WB}(\mathbb{P}_Y, \mathbb{P}_{Y|X_i}) = \sqrt{\left(\mu_Y - \mu_{Y|X_i}\right)^2 + \left(\sigma_Y - \sigma_{Y|X_i}\right)^2}. \tag{46}$$

Then, combining (46) and (17), we obtain Equation (29). ∎

Proof of Proposition 10. By Theorem 2.1 of Gelbrich [1990], if $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_\alpha}$ are elliptical with the same generating function h then the Wasserstein distance between $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_\alpha}$ is given by

$$W_2(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha}) = \sqrt{||\mu_Y - \mu_{Y|X_\alpha}||_2^2 + \mathrm{Tr}\left(\Sigma_Y + \Sigma_{Y|X_\alpha} - 2\left((\Sigma_Y)^{1/2}\Sigma_{Y|X_\alpha}(\Sigma_Y)^{1/2}\right)^{1/2}\right)} = \mathrm{WB}(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha}). \tag{47}$$

By assumption, this holds for all values of X_α, and inserting this expression in the common rationale of (17) completes the proof. ∎

Proof of Proposition 11. To prove Equation (35), note that

$$\mathrm{ADV}_i^{\mathrm{WB}^2} = \mathbb{E}\left[\sum_{t=1}^{n_Y}(\mu_{Y,t} - \mu_{Y|X_i,t})^2\right] = \sum_{t=1}^{n_Y}\mathbb{E}[(\mu_{Y,t} - \mu_{Y|X_i,t})^2] = \sum_{t=1}^{n_Y}\xi_i^{V,t}. \tag{48}$$

Moreover, we report the results in [Gamboa et al., 2014, Section 3.1]. First, note that under independence one can write

$$g(X) = g_i(X_i) - g_i(X_{-i}) + g_{-i,j}(X_i, X_{-i}) - \mathbb{E}[Y], \tag{49}$$

where $g_i(X_i) = \mathbb{E}[Y|X_i] - \mathbb{E}[Y]$, $g_{-i}(X_{-i}) = \mathbb{E}[Y|X_{-i}] - \mathbb{E}[Y]$, and $g_{i,-i}(X_i, X_{-i}) = g(X) - g_i(X_i) - g_{-i}(X_{-i})$. Under independence, the variance of Y can be decomposed as $\Sigma_Y = \Sigma_i^Y + \Sigma_{-i}^Y + \Sigma_{i,-i}^Y$, and the generalized Sobol' sensitivity index of X_i is then $S_i = \dfrac{tr(\Sigma_i^Y)}{tr(\Sigma_Y)}$, where $tr(\Sigma_i^Y)$ equals the sum of the individual contributions of X_i to the variance of Y^t, that is $tr(\Sigma_i^Y) = \sum_{t=1}^{n_Y}\xi_i^{V,t}$. ∎

Proof of Corollary 12. Let $Y = AX + b$, with Y a random vector in $\mathbb{R}^m$ on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $A = (a_{i,j})$, $i = 1, 2, \ldots, n_X$, $j = 1, 2, \ldots, n_Y$, and $b = (b_1, b_2, \ldots, b_{n_Y})$. First, let $X \in \mathbb{R}^{n_X}$, $X \sim \mathcal{EC}(\mu_X, \Sigma_X^*, h)$ with finite second order moment. If A is an $n_Y \times n_X$ matrix and $b \in \mathbb{R}^{n_Y}$, then $Y \sim \mathcal{EC}(A\mu_X + b, A\Sigma_X^* A^T, h)$ and $Z = Y|X_i$ is elliptical [see Landsman and Valdez [2003] among others]. At the same time, as proven in Cambanis et al. [1981], if X is elliptical then the random variable $U = X|X_i$ is elliptical and $U \sim \mathcal{EC}(\mu_{X|X_i}, \Sigma_i^c, h)$, with Σ_i^c as given in (31). Therefore, $(Y|X_i) \sim \mathcal{EC}(A\mu_{Y|X_i} + b, A\Sigma_i^c A^T, h)$. Then, Y and $(Y|X_i)$ are both elliptical random variables with generating function h.

Then, the 2-Wasserstein metric between $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_i}$ $W_2(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})$ is equal to the Wasserstein-Bures metric $WB(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})$ in (11) for every X_i. Then, we have

$$WB(\mathbb{P}_Y, \mathbb{P}_{Y|X_i}) = \sqrt{||\mu_Y - \mu_{Y|X_i}||_2^2 + \text{Tr}\left(\Sigma_Y + \Sigma_{Y|X_i} - 2\left(\Sigma_Y^{1/2}\Sigma_{Y|X_i}\Sigma_Y^{1/2}\right)^{1/2}\right)} \tag{50}$$

for all values of X_i, with $\mu_Y = A\mu_X + b$, $\Sigma_Y = A\Sigma_X A^T$, $\Sigma_{Y|X_i} = A\Sigma_i^c A^T$, Σ_i^c as defined in Equation (31), and $\mu_{Y_k|X_i} = \sum_{j=1}^{n_X} a_{k,j}\left(\mu_j + (X_i - \mu_i)\frac{\sigma_{i,j}^i}{\sigma_{i,i}^i}\right)$, for $k = 1, 2, \ldots, n_Y$, with $\sigma_{i,j}^i$ as in (31).

For the second part of the Corollary, the following holds. If X is normally distributed, then the linear combination of normal random variables is still normal. Then, $Sk_\varepsilon(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})$ has the closed form expression recently proven in [Janati et al., 2020], namely

$$S_\varepsilon(\mathbb{P}_Y, \mathbb{P}_{Y|X_i}) = \sqrt{||\mu_Y - \mu_{Y|X}||_2^2 + \text{Tr}\left(\Sigma_Y + \Sigma_{Y|X_i} - D_\varepsilon\right) + L(D_\varepsilon, \varepsilon)}, \tag{51}$$

with D_ε and $L(D_\varepsilon, \varepsilon)$ as given after Equation (36) and in Equation (37), respectively, where μ_Y, $\mu_{Y|X}$ and Σ_Y, $\Sigma_{Y|X_i}$ are formally identical as in the first part of the proof. Then, substituting the right hand side of Equation (51) into Equation (22) completes the proof.

∎

Proof of Proposition 15. In the proof convergence is meant as convergence in probability.
Case 1: Discrete X_α. In this case, the support of X_α is a countable collection of realizations of X_α, $\mathcal{X}_\alpha = \{x_\alpha^1, x_\alpha^2, \ldots, x_\alpha^M\}$, with corresponding probability mass function whose elements are the values $\mathbb{P}[X_\alpha = x_\alpha^m] \geq 0$, for all $m = 1, 2, \ldots, M$. Correspondingly, $\xi_\alpha^{W_p}$ can be rewritten as

$$\xi_\alpha^{W_p} = \sum_{m=1}^{M} \mathbb{P}[X_\alpha = x_\alpha^m] W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha = x_\alpha^m}). \tag{52}$$

Then, consider a dataset of size N of input output realizations. Let $y_N = \{y_{1,t}, y_{2,t}, \ldots, y_{N,t}\}$, $t = 1, 2, \ldots, n_Y$, be the set of realizations of Y. Let also $y_N^m = \{y_{1,t}^m, y_{2,t}^m, \ldots, y_{N^m,t}\}$ denote the conditional realizations given that $X_\alpha = x_\alpha^m$. We write the given data estimator of $\xi_\alpha^{W_p}$ in Equation (52) as

$$\widehat{\xi}_\alpha^{W_p}(N) = \sum_{m=1}^{M} \widehat{\mathbb{P}}[X_\alpha = x_\alpha^m] \widehat{W}_p(y_N, y_N^m), \tag{53}$$

where by $\widehat{W}_p(y_N, y_N^m)$ we indicate that the Wasserstein distance is computed on the sample. Then, let $\widehat{\mathbb{P}}[X_\alpha = x_\alpha^m]$ be any consistent estimator of $\mathbb{P}[X_\alpha = x_\alpha^m]$. $\widehat{\mathbb{P}}[X_\alpha = x_\alpha^m]$ can be $\mathbb{P}[X_\alpha = x_\alpha^m]$ itself, if this is known. This is typically the case in simulation settings, where the joint input distribution is assigned by the analyst. Alternatively, the ratio $\frac{N_m}{N}$, where N_m denotes the number of realizations of X_α equal to x_α^m in the dataset, is a consistent estimator of $\mathbb{P}[X_\alpha = x_\alpha^m]$ by the law of large numbers. Then, note that the continuity of the Wasserstein distance (Sommerfeld and Munk [2018] and [Villani, 2009, p. 109, Corollary 6.11]) implies that $\widehat{W}_p(y_N, y_N^m) \to W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha=x_\alpha^m})$ as $N \to \infty$. Hence, for every m,

$$\widehat{\mathbb{P}}[X_\alpha = x_\alpha^m]\widehat{W}_p(y_N, y_N^m) \underset{N\to\infty}{\to} \mathbb{P}[X_\alpha = x_\alpha^m]W_p(y_N, y_N^m), \tag{54}$$

so that

$$\begin{gathered} \lim_{N\to\infty} \widehat{\xi}_\alpha^{W_p}(N) = \lim_{N\to\infty} \textstyle\sum_{m=1}^M \widehat{\mathbb{P}}[X_\alpha = x_\alpha^m]\widehat{W}_p(y_N, y_N^m) \\ = \textstyle\sum_{m=1}^M \lim_{N\to\infty} \widehat{\mathbb{P}}[X_\alpha = x_\alpha^m]\widehat{W}_p(y_N, y_N^m) = \sum_{m=1}^M \mathbb{P}[X_\alpha = x_\alpha^m]W_p(y_N, y_N^m) = \xi_\alpha^{W_p}. \end{gathered} \tag{55}$$

Case 2: Absolutely Continuous X_α. $\xi_\alpha^{W_p}$ is now written as

$$\xi_\alpha^{W_p} = \int_{\mathcal{X}_\alpha} W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha=x_\alpha})f_\alpha(x_\alpha)dx_\alpha \tag{56}$$

where the integral in the above expression is meant in a Riemann-Stieltjes multivariate sense and the density $f_\alpha(x_\alpha)$ exists for the absolute continuity of X_α. We introduce a sequence of refining partitions $P_M^{\delta_P}(\mathcal{X}_\alpha) = \{\mathcal{X}_\alpha^m(N),\, m = 1, 2, \ldots, M(\delta_P)\}$. One says that

$$\xi_\alpha^{W_p} = \lim_{\delta_P\to 0} \sum_{m=1}^{M(\delta_P)} W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha\in\mathcal{X}_\alpha^m(\delta_P)})\mathbb{P}[X_\alpha \in \mathcal{X}_\alpha^m(\delta_P)] \tag{57}$$

if the limit exists. Then, consider a fixed partition $P_M^N(\mathcal{X}_\alpha)$ of size $M(N)$ for a given sample of size N. For such fixed partition, the given-data estimator of $\xi_\alpha^{W_p}$ is written as

$$\widehat{\xi}_\alpha^{W_p}(N) = \sum_{m=1}^{M(N)} \widehat{\mathbb{P}}[X_\alpha \in \mathcal{X}_\alpha^m(N)]\widehat{W}_p(y_N, y_N^m), \tag{58}$$

where, with similar notation as in the first part of this proof, y is the set of realizations of Y in the sample and $y_N^m = \{y_{1,t}^m, y_{2,t}^m, \ldots, y_{N^m,t}\}$ denotes the conditional realizations given that $X_\alpha \in \mathcal{X}_\alpha^m(N)$. Fixing the partition size M, the first part of the proof implies that this estimator is consistent for every fixed partition, that is

$$\sum_{m=1}^{M} \widehat{\mathbb{P}}[X_\alpha \in \mathcal{X}_\alpha^m(N)]\widehat{W}_p(y_N, y_N^m) \underset{N\to\infty}{\to} \sum_{m=1}^{M} \mathbb{P}[X_\alpha \in \mathcal{X}_\alpha^m(N)]W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha\in\mathcal{X}_\alpha^m}). \tag{59}$$

Then, we consider altering the partition size in accordance with the sequence of refining partitions $P_M^N(\mathcal{X}_\alpha)$ as the sample size increases. For a proper refinement strategy, the maximal diameter in the partition decreases with increasing N, and (58) and (59) then convergence to the Riemann-Stieltjes integral

$$\widehat{\xi}_\alpha^{W_p}(N) \underset{N\to\infty}{\to} \int_{\mathcal{X}_\alpha} W_p\left(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha}\right) d\mathbb{P}[X_\alpha] \tag{60}$$

where we used $\widehat{W}_p(y_N, y_N^m) \to W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha\in\mathcal{X}_\alpha^m(\delta_P)})$ which holds by continuity of the Wasserstein distance and the strong law of large numbers (see Sommerfeld and Munk [2018], and [Villani, 2009, p. 109, Corollary 6.11]). The integral (60) then exists by the assumption that $W_p(\mathbb{P}_Y, \mathbb{P}_{Y|X_\alpha\in\mathcal{X}_\alpha^m(\delta_P)})$ is bounded. ∎

Proof of Proposition 16. We use the notation from the proof of Proposition 15. In order to save space, we write suggestively $\mathrm{Tr}\sqrt{A,B} = \mathrm{Tr}\left(A + B - 2\left(B^{1/2}AB^{1/2}\right)^{1/2}\right)$ for the diffusive part. In the present proof we start noting that if Y is normal, then ξ^{WB} coincides with ξ^{W_2} and we fall under Proposition 15. For all other cases, we consider the discussion below. *Case 1:* X_α discrete. Let X_α have a countable number of realizations $x_\alpha^1, x_\alpha^2, \ldots, x_\alpha^M$. Then,

$$\xi_\alpha^{\mathrm{WB}} = \sum_{m=1}^{M} \mathbb{P}[X_\alpha = x_\alpha^m] \cdot \sqrt{\sum_{t=1}^{n_Y} \left(\mu_{Y,t} - \mu_{Y|X_\alpha = x_\alpha^m, t}(N)\right)^2 + \mathrm{Tr}\sqrt{\Sigma_Y(N), \Sigma_{Y|X_\alpha \in \mathfrak{X}_\alpha^m}(N)}}. \quad (61)$$

Let Y_t, $t = 1, 2, \ldots, n_Y$, denote a generic component of the multivariate output vector Y. Let $x = (x_{j,i})$, where $i = 1, 2, \ldots, n_X$, $j = 1, 2, \ldots, N$ denotes a sample of realizations of X and let $y = (y_{r,t})$, $r = 1, 2, \ldots, N$, $t = 1, 2, \ldots, n_Y$ denote the corresponding dataset of output realizations. The estimator $\hat{\mu}_{Y,t}(N) = N^{-1} \sum_{r=1}^{N} y_{r,t}$ is consistent. Also,

$$\widehat{\Sigma}_Y(N) = (\widehat{\sigma}_{r,s}) = \left(\frac{1}{N-1} \sum_{t=1}^{n_Y} (y_{r,t} - \hat{\mu}_{Y,t})(y_{s,t} - \hat{\mu}_{Y,t}(N))\right)_{r,s=1,\ldots,N} \quad (62)$$

is a consistent estimator of Σ_Y. Let $y_{r|X_\alpha = x_\alpha^m, t}$ denote the realizations of Y_t conditional on the fact that $X_\alpha = x_\alpha^m$ and let N_m denote their cardinality. Then, for the mean, we have $\hat{\mu}_{Y|X_\alpha,t}(N) = N_m^{-1} \sum_{r=1}^{N_m} y_{r|X_\alpha = x_\alpha^m, t}$. For the covariance matrix, we have

$$\widehat{\Sigma}_{Y|X_\alpha}(N) = [\widehat{\sigma}_{r,s} = \frac{1}{N-1} \sum_{r=1,s=1}^{N_m} (y_{r,t} - \hat{\mu}_{Y|X_\alpha = x_\alpha^m, t})(y_{s,t} - \hat{\mu}_{Y|X_\alpha = x_\alpha^m, t}(N)). \quad (63)$$

These are consistent estimators of $\mu_{Y|X_\alpha}$ and $\widehat{\Sigma}_{Y|X_\alpha}$, respectively. Noting that $N_m \to \infty$ as $N \to \infty$, and that by the law of large numbers $\frac{N_m}{N} \to \mathbb{P}[X_\alpha = x_\alpha^m]$ as $N \to \infty$, we have

$$\begin{aligned}
&\lim_{N\to\infty} \hat{\xi}_\alpha^{\mathrm{WB}}(N) = \\
&\quad \lim_{N\to\infty} \sum_{m=1}^{M} \frac{N_m}{N} \sqrt{\sum_{t=1}^{n_Y} (\hat{\mu}_{Y,t}(N) - \hat{\mu}_{Y|X_\alpha = x_\alpha^m, t}(N))^2 + \mathrm{Tr}\sqrt{\widehat{\Sigma}_Y(N)), \mathrm{Tr}(\widehat{\Sigma}_{Y|X_\alpha \in \mathfrak{X}_\alpha^m}(N))}} \\
&\quad = \sum_{m=1}^{M} \mathbb{P}[X_\alpha = x_\alpha^m] \sqrt{\sum_{t=1}^{n_Y} (\mu_{Y,t} - \mu_{Y|X_\alpha = x_\alpha^m, t})^2 + \mathrm{Tr}\sqrt{\Sigma_Y, \Sigma_{Y|X_\alpha \in \mathfrak{X}_\alpha^m}}} = \xi_\alpha^{\mathrm{WB}}. \quad (64)
\end{aligned}$$

Case 2: X_α absolutely continuous. In this case, ξ_α^{WB} is written as

$$\xi_\alpha^{\mathrm{WB}} = \int_{\mathfrak{X}_\alpha} \sqrt{\sum_{t=1}^{n_Y} (\mu_{Y,t} - \mu_{Y|X_\alpha,t}(N))^2 + \mathrm{Tr}\sqrt{\Sigma_Y(N), \Sigma_{Y|X_\alpha}(N)}} dF_{X_\alpha}. \quad (65)$$

By definition of Riemann-Stieltjes sum, we can write

$$\xi_\alpha^{\mathrm{WB}} = \lim_{\delta_P \to 0} \sum_{m=1}^{M(\delta_P)} \mathbb{P}[X_\alpha \in \mathcal{X}_\alpha^m] \sqrt{\sum_{t=1}^{n_Y} (\mu_{Y,t} - \mu_{Y|X_\alpha \in \mathcal{X}_\alpha^m, t}(N))^2 + \operatorname{Tr} \sqrt{\Sigma_Y(N), \Sigma_{Y|X_\alpha \in \mathcal{X}_\alpha^m}(N)}}. \tag{66}$$

Let then $P_M^N(\mathcal{X}_\alpha)$ be a sequence of partitions determined by a partition refining strategy. Then, $\lim_{N\to\infty}$ implies $\lim_{\delta(P)\to 0}$. Let us now write $\xi_\alpha^{\mathrm{WB}} = \lim_{N\to\infty} \sum_{m=1}^{M(N)} S_m(N)$, where the summand $S_m(N)$ is given by

$$S_m(N) = \mathbb{P}[X_\alpha \in \mathcal{X}_\alpha^m(N)] \cdot \sqrt{\sum_{t=1}^{n_Y} (\mu_{Y,t} - \mu_{Y|X_\alpha \in \mathcal{X}_\alpha^m(N), t}(N))^2 + \operatorname{Tr} \sqrt{\Sigma_Y(N), \Sigma_{Y|X_\alpha}(N)}}. \tag{67}$$

The given-data approximation of $S_m(N)$ is

$$\widehat{S}_m(N) = \widehat{\mathbb{P}}[X_\alpha \in \mathcal{X}_\alpha^m(N)] \sqrt{\sum_{t=1}^{n_Y} (\widehat{\mu}_{Y,t} - \widehat{\mu}_{Y|X_\alpha \in \mathcal{X}_\alpha^m(N), t}(N))^2 + \operatorname{Tr} \sqrt{\widehat{\Sigma}_Y(N), \widehat{\Sigma}_{Y|X_\alpha}(N)}}. \tag{68}$$

Similarly to the previous part of the proof, letting $y_{r|X_\alpha \in \mathcal{X}_\alpha^m(N), t}$ denote the realizations of Y_t conditional on $X_\alpha \in \mathcal{X}_\alpha^m(N)$, the estimators

$$\hat{\mu}_{Y|X_\alpha \in \mathcal{X}_\alpha^m, t}(N) = \frac{1}{N_m} \sum_{r=1}^{N_m} y_{r|X_\alpha \in \mathcal{X}_\alpha^m(N), t}, \tag{69}$$

and

$$\widehat{\Sigma}_{Y|X_\alpha \in \mathcal{X}_\alpha^m} = (\widehat{\sigma}_{r,s})_{r,s=1,\ldots,n} = \frac{1}{N_m - 1} \sum_{t}^{n_Y} (y_{r,t} - \hat{\mu}_{Y,t}(N))(y_{r,t} - \hat{\mu}_{Y,t}(N)) \tag{70}$$

are consistent estimators of $\mu_{Y|\alpha \in \mathcal{X}_\alpha^m, t}(N)$ and $\Sigma_{Y|X_\alpha \in \mathcal{X}_\alpha^m}$, respectively for any $P_M(\mathcal{X}_\alpha; N)$; also $\frac{N_m}{N}$ is a consistent estimator of $\widehat{\mathbb{P}}[X_\alpha \in \mathcal{X}_\alpha^m(N)]$. Thus, $\widehat{\mathbb{P}}[X_\alpha \in \mathcal{X}_\alpha^m(N)] \to \mathbb{P}[X_\alpha \in \mathcal{X}_\alpha^m(N)]$. By the consistency of $\hat{\mu}_{Y|X_\alpha \in \mathcal{X}_\alpha^m, t}(N)$ and of $\hat{\mu}_{Y|X_\alpha = x_\alpha, t}(N)$, as $N \to \infty$ both $|\hat{\mu}_{Y|X_\alpha \in \mathcal{X}_\alpha^m, t}(N) - \mu_{Y|X_\alpha \in \mathcal{X}_\alpha^m, t}(N)| \to 0$ and $|\hat{\mu}_{Y|X_\alpha = x_\alpha, t}(N) - \mu_{Y|X_\alpha = x_\alpha, t}(N)| \to 0$. A similar reasoning applies to the variance-covariance estimators. Thus, as $N \to \infty$, we have by consistency that it converges to the Stietljes integral

$$\widehat{\xi}_\alpha^{\mathrm{WB}}(N) \to \int_{X_\alpha} \sqrt{\sum_{t=1}^{n_Y} (\mu_{Y,t} - \mu_{Y|X_\alpha})^2 + \operatorname{Tr} \sqrt{\Sigma_Y, \Sigma_{Y|X_\alpha}}} d\mathbb{P}[X_\alpha] = \xi_\alpha^{\mathrm{WB}} \tag{71}$$

Hence, if the limits of $\widehat{S}_m(N)$ for $N \to \infty$ exist then (71) is well-defined. ∎

References

A. Alexanderian, P. A. Gremaud, and R. C. Smith. Variance-based sensitivity analysis for time-dependent processes. *Reliability Engineering & System Safety*, 196:106722, 2020.

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 1965–1975, 2017.

J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *Advances in Neural Information Processing Systems*, volume 32, pages 1–11, 2019.

P. Berthet, J.-C. Fort, and T. Klein. A central limit theorem for Wasserstein type distances between two distinct univariate distributions. *Annales de l'institut Henri Poincare (B) Probability and Statistics*, 56(2):954–982, 2020.

J. Betancourt, F. Bachoc, T. Klein, D. Idier, R. Pedreros, and J. Rohmer. Gaussian process metamodeling of functional-input code for coastal flood hazard assessment. *Reliability Engineering and System Safety*, 198, 2020.

N. Bliznyuk, D. Ruppert, C. Shoemaker, R. Regis, S. Wild, and P. Mugunthan. Bayesian Calibration and Uncertainty Analysis for Computationally Expensive Models Using Optimization and Radial Basis Function Approximation. *Journal of Computational and Graphical Statistics*, 17(2):270–294, 2008.

S. G. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics and Kantrovich transport distances.* American Mathematical Society, 2016.

E. Borgonovo, S. Tarantola, E. Plischke, and M. D. Morris. Transformations and Invariance in the Sensitivity Analysis of Computer Experiments. *Journal of the Royal Statistical Society, Series B*, 76(5):925–947, 2014.

S. Cambanis, S. Huang, and G. Simons. On the Theory of Elliptically Contoured Distributions. *Journal of Multivariate Analysis*, 11:368–385, 1981.

S. Chatterjee. A New Coefficient of Correlation. *Journal of the American Statistical Association*, 2020.

Y. Chen, T. T. Georgiou, and M. Pavon. Stochastic control liaisons: Richard Sinkhorn meets gaspard monge on a schrödinger bridge. *SIAM Review*, 63(2):249–313, 2021.

J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *NeurIPs*, 26:1–9, 2013.

N. Deb and B. Sen. Multivariate Rank-based Distribution-free Nonparametric Testing using Measure Transportation. *Journal of the American Statistical Association*, pages 1–52, 2021. doi: 10.1080/01621459.2021.1923508.

R. Dobrushin. Prescribing a System of Random Variables by Conditional Distributions. *Theory of Probability and Its Applications*, 15:458–486, 1970.

Y. Dong, Y. Gao, R. Peng, I. Razenshteyn, and S. Sawlani. A Study of Performance of Optimal Transport. *ArXiv*, 2005(01182v1):1–9, 2020.

A. Figalli and F. Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. European Mathematical Society, 2021. ISBN 978-3-98547-010-5. doi: 10.4171/ETB/22.

J.-C. Fort, T. Klein, and A. Lagnoux. Global Sensitivity Analysis and Wasserstein Spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):880–921, 2021.

R. Fraiman, F. Gamboa, and L. Moreno. Sensitivity indices for output on a riemannian manifold. *International Journal for Uncertainty Quantification*, 10(4):297–314, 2020.

F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8:573–603, 2014.

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.

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

F. Gamboa, T. Klein, A. Lagnoux, and L. Moreno. Sensitivity Analysis in General Metric Spaces. *Reliability Engineering & System Safety*, 212:107611, 2021.

M. Gelbrich. On a Formula for the L2 Wasserstein Metric Between Measures on Euclidean Hilbert Spaces. *Mathematical Nachrichten*, 147:185–203, 1990.

R. Ghanem, D. Higdon, and H. Owhadi, editors. *Handbook of Uncertainty Quantification*. Springer Verlag, 2017.

C. R. Givens and R. M. Shortt. A Class of Wasserstein Metrics for Probability Distributions. *Michigan Mathematical Journal*, 31:231–240, 1984.

H. Janati, B. Muzellec, G. Peyré, and M. Cuturi. Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form. *ArXiv*, 2006.02572:1–37, 2020.

P. A. Knight. The Sinkhorn-Knopp Algorithm: Convergence and Applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

H. W. Kuhn. Variants of the hungarian method for assignment problems. *Naval Research Logistics Quarterly*, 3(4):253–258, dec 1956. ISSN 00281441.

M. Lamboni. Multivariate sensitivity analysis: Minimum variance unbiased estimators of the first-order and total-effect covariance matrices. *Reliability Engineering and System Safety*, 187:67–92, 2019.

M. Lamboni. Derivative-based generalized sensitivity indices and Sobol' indices. *Mathematics and Computers in Simulation*, 170:236–256, 2020.

M. Lamboni, H. Monod, and D. Makowski. Multivariate Sensitivity Analysis to Measure Global Contribution of Input Factors in Dynamic Models. *Reliability Engineering & System Safety*, 96:450–459, 2011.

Z. M. Landsman and E. A. Valdez. Tail Conditional Expectations for Elliptical Distributions. *North American Actuarial Journal*, 7(4):55–71, 2003.

Q. Liu and T. Homma. A New Importance Measure for Sensitivity Analysis. *Journal of Nuclear Science and Technology*, 47(1):53–61, 2010.

R. Liu and A. B. Owen. Estimating Mean Dimensionality of Analysis of Variance Decompositions. *Journal of the American Statistical Association*, 101(474):712–721, 2006.

D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Springer, Cham, fourth edition, 2016. ISBN 978-3-319-18841-6.

C. L. Mallows. A Note on Asymptotic Joint Normality. *Annals of Mathematical Statistics*, 43:508–515, 1972.

A. Marrel, B. Iooss, M. Jullien, B. Laurent, and E. Volkova. Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics*, 22(3):383–397, 2011.

A. Marrel, N. Saint-Geours, and M. De Lozzo. Sensitivity analysis of spatial and/or temporal phenomena. In *Handbook of Uncertainty Quantification*, pages 1327–1357. 2017.

J. 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.

J. E. Oakley. Decision-theoretic Sensitivity Analysis for Complex Computer Models. *Technometrics*, 51(2):121–129, 2009.

A. B. Owen. Halton sequences avoid the origin. *SIAM Review*, 48(3):487–503, 2006.

H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz. Optimal Uncertainty Quantification. *Siam Review*, 55(2):271–345, 2013.

W. Pan, X. Wang, W. Xiao, and H. Zhu. A Generic Sure Independence Screening Procedure. *Journal of the American Statistical Association*, 114(526):928–937, 2019.

W. Pan, X. Wang, H. Zhang, H. Zhu, and J. Zhu. Ball Covariance: A Generic Measure of Dependence in Banach Space. *Journal of the American Statistical Association*, 115 (529):307–317, 2020.

V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019.

V. M. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space*. Cham, 2020.

K. Pearson. *On the General Theory of Skew Correlation and Non-linear Regression*, volume XIV of *Mathematical Contributions to the Theory of Evolution, Drapers' Company Research Memoirs*. Dulau & Co., London, 1905.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):206–355, 2019.

G. Puccetti. An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *Journal of Mathematical Analysis and Applications*, 451 (1):132–145, 2017.

G. Puccetti, L. Rüschendorf, and S. Vanduffel. On the computation of Wasserstein barycenters. *Journal of Multivariate Analysis*, 176, 2020.

S. Rahman. The f-Sensitivity Index. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):130–162, 2016.

A. Renyi. On Measures of Statistical Dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451, 1959.

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, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis – The Primer*. Chichester, 2008.

H. Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438, 1947.

M. Sommerfeld and A. Munk. Inference for Empirical Wasserstein Distances on Finite Spaces. *Journal of the Royal Statistical Society Series B*, 80(1):219–238, 2018.

M. Strong and J. E. Oakley. An efficient method for computing partial expected value of perfect information for correlated inputs. *Medical Decision-Making*, 33:755–766, 2013.

M. Strong, J. E. Oakley, and J. Chilcott. Managing Structural Uncertainty in Health Economic Decision Models: a Discrepancy Approach. *Journal of the Royal Statistical Society, Series C*, 61(1):25–45, 2012.

T. J. Sullivan. *Introduction to Uncertainty Quantification*. Springer Verlag, 2015. ISBN 978-3-319-23394-9.

S. S. Vallender. Calculation of the Wasserstein Distance between Probability Distributions on the Line. *Theory of Probability and its Applications, SIAM*, 18(4):784–786, 1974.

C. Villani. *Optimal Transport: Old and New*. Springer Verlag, Berlin, 2009.

S. Wang, T. T. Cai, and H. Li. Optimal Estimation of Wasserstein Distance on a Tree With an Application to Microbiome Studies. *Journal of the American Statistical Association*, 2020.

J. Weed and F. Bach. Sharp Asymptotic and Finite-Sample Rates of Convergence of Empirical Measures in Wasserstein Distance. *ArXiv*, 1707.(00087v1):1–35, 2017.