Global Sensitivity Analysis Via Optimal Transport

Emanuele Borgonovo* Alessio Figalli† Elmar Plischke‡
Giuseppe Savaré§

March 7, 2024

Abstract

We examine the construction of variable importance measures for multivariate responses using the theory of optimal transport. We start with the classical optimal transport formulation. We show that the resulting sensitivity indices are well-defined under input dependence, are equal to zero under statistical independence, and are maximal under fully functional dependence. Also, they satisfy a continuity property for information refinements. We show that the new indices encompass Wagner's variance-based sensitivity measures. Moreover, they provide deeper insights into the effect of an input's uncertainty, quantifying its impact on the output mean, variance, and higher-order moments. We then consider the entropic formulation of the optimal transport problem and show that the resulting global sensitivity measures satisfy the same properties, with the exception that, under statistical independence, they are minimal but not necessarily equal to zero. We prove the consistency of a given-data estimation strategy and test the feasibility of algorithmic implementations based on alternative optimal transport solvers. Application to the assemble-to-order simulator reveals a significant difference in the key drivers of uncertainty between the case in which the quantity of interest is profit (univariate) or inventory (multivariate). The new importance measures contribute to meeting the increasing demand for methods that make black-box models more transparent to analysts and decision-makers.

Keywords. Sensitivity Analysis, Computer Simulations, Variable Importance Measures

*Department of Decision Sciences, Bocconi University, 20136 Milan, Italy emanuele.borgonovo@unibocconi.it
†Department of Mathematics, ETH Zürich, 8092 Zürich, Switzerland alessio.figalli@math.ethz.ch
‡Institute of Disposal Research, Clausthal University of Technology, 38678 Clausthal-Zellerfeld, Germany elmar.plischke@tu-clausthal.de
§Department of Decision Sciences, Bocconi University, 20136 Milan, Italy giuseppe.savare@unibocconi.it

1. Introduction

Managerial decision-making is increasingly informed by forecasts produced by mathematical models. In many instances, these models calculate multiple quantities of interest: Future CO_2 emissions, temperature changes, and carbon prices are outputs of well-known integrated assessment models (Hu et al. 2012, Nordhaus 2017); the number of infected, hospitalized, or deceased individuals as well as policy-relevant economic quantities are simultaneously calculated by epidemiological models (Berger et al. 2022, Du et al. 2022).

The complexity of the problems and the large amount of data, however, often force analysts to implement sophisticated software architectures that make the resulting simulator black-boxes, with little hope of obtaining insights from intuition. Analysts should then transparently assess the stability of the simulator response and its sensitivity to the uncertain inputs (Kleijnen 2010, Barton 2016, Saltelli et al. 2020), before communicating results to stakeholders. Feature importance, that is, the understanding of the factors that drive a simulator behavior, becomes an essential insight for result explanation and communication. While there is a well-established set of methods for analyzing univariate responses, identifying key drivers for multivariate responses is an active area of research. For instance, well-known methods such as the variance-based approach of Wagner (1995) or the moment-independent approach of Baucells and Borgonovo (2013) are devised for the single output context.

We propose a novel approach to the global sensitivity analysis of multivariate responses grounded in the theory of optimal transport (OT) (Figalli and Glaudo 2021). We consider two classes of global sensitivity measures, based on the classical and the entropic formulation. We prove that indices based on the classical formulation possess key properties that ease their interpretation: They are equal to zero if and only if the (multivariate) output is independent of the input of interest and they are maximal if and only if the output is a deterministic function of the inputs. We also derive a monotonicity result according to which the value of the sensitivity index decreases if less refined information on the input is received. A key role in obtaining these properties is played not only by the convexity of the OT cost functional, but also, and far less obviously, by its strict-convexity on Dirac-δ measures.

We then study global sensitivity indices based on the entropic OT formulation (Cuturi 2013). The interest is twofold. On the one hand, the entropic formulation is often used as a substitute for the classical one because it allows for fast algorithmic implementations and, under mild conditions, the solution of the entropic problem approximates the classical solution well. On the other hand, due to its wide applicability, there is growing interest in studying the entropic formulation per se, independently of its use for approximating the classical problem (Genevay et al. 2018, Chen et al. 2021). However, the geometric properties of the entropic OT formulation are less known. We contribute by proving that the entropic cost functional is convex and, surprisingly, strictly convex on Dirac-δ masses. Then, entropic OT-based sensitivity indices are monotonically increasing for information refinements and maximal in the presence of a noiseless input-output dependence. They also attain their minimum value under independence, although the minimum may not necessarily be zero. They reach the same maximum value as sensitivity measures based on the classical OT formulation. However, as the value of the entropic regularization parameter increases, they tend to the maximum value for all inputs, thereby confounding the relative input importance.

We then focus on indices based on the 2-Wasserstein squared distance. We show that the corresponding indices allow for a transparent interpretation of the sensitivity measures. One can decompose them exactly into three terms. The first term and second terms account, respectively, for differences in the means and variance-covariance matrices of the model output. The third term is residual and is present when the effect of fixing an input impacts more than the first two moments of the output. In addition, we prove that the first term is the sum of the univariate variance-based sensitivity measures proposed by Wagner (1995) for the dependent as well as independent input cases and that, under input independence, it coincides with the multivariate indices proposed by Lamboni et al. (2011) and Gamboa et al. (2014).

To enable computation for realistic applications, we study an estimation design based on Pearson's given-data intuition (Pearson 1905), which makes the calculation cost linear in the sample size. We prove that the estimators are asymptotically unbiased and the estimates converge from below. The estimation design involves the solution of a series of data-driven OT problems. Here our work intersects with the fast-growing literature on efficient algorithmic solutions to OT problems, nowadays a topical research subject in machine learning (Altschuler et al. 2019, Janati et al. 2020). We implement and compare estimators with solvers that rely on alternative principles, namely, on the network simplex approach (Kuhn 1955), the partial orderings approach of Puccetti (2017), the Sinkhorn-based approach of Cuturi (2013), as well as algebraic estimators based on the Wasserstein-Bures approximation (Givens and Shortt 1984). Our goal is not to single out "a" (or the) best algorithm, but to assess whether numerical quantification is feasible in an amount of time that makes the method suitable for realistic applications. We evaluate the insights delivered by the new indices and the performance of the proposed estimators through several experiments. We start with the well-known Ishigami model and continue with a new univariate case in which dimensionality is increased up to about $10,000$ inputs. We then consider a multivariate normal test case, for which closed-form expressions of the OT-based sensitivity measures are available. Findings indicate that all the employed algorithms yield consistent estimates at reasonable sample sizes and with fast execution times. Also, their behavior is in line with the theoretical premises, with convergence from below. We then apply the new sensitivity measures to conduct a global sensitivity analysis of the well-known assemble-to-order (ATO) simulator of Hong and Nelson (2006). We consider both the system profit (univariate) and the final inventory (multivariate). The numerical investigation shows that the new indices yield additional insights, complementary to the ones produced by variance-based indices both in the univariate and the multivariate output cases.

2. Background

This section concisely reviews material on the theory of optimal transport (Section 2.1) and probabilistic sensitivity analysis (Section 2.2).

2.1. Optimal Transport and Wasserstein Distances

Optimal transport (OT, henceforth) is a classical research subject in operations research (Hitchcock 1940, Hillier and Lieberman 2012), and is actively studied across mathematics, statistics, and machine learning (Chen et al. 2021). We refer to the monographs of Villani

(2008), Peyré and Cuturi (2019), Figalli and Glaudo (2021) for a detailed treatment of theoretical and computational aspects.

Let Y be a random variable on measure space $(\Omega, \mathcal{B}, \mathbb{P})$, with support $\mathsf{Y} \subseteq \mathbb{R}^{n_Y}$. Let A be a measurable subset of Y. A marginal probability measure of Y denoted by ν, is a set function $\nu(A) = \mathbb{1}(Y \in A)$. For instance, if $A = \{y : y \in \mathsf{Y} \text{ and } y \leq y'\}$, then $\nu(A) = \mathbb{P}(Y \leq y') = F_Y(y')$, where $y \mapsto F_Y(y)$ is the cumulative distribution function of Y.

Consider two marginal distributions of Y, ν and ν': the optimal transport problem consists in transferring the first distribution into the second while minimizing a given cost function. Let $\pi(y, y')$ be a transfer plan and $\Pi(\nu, \nu')$ be the set of all transfer plans. Formally, an element of $\Pi(\nu, \nu')$ is a joint probability function whose marginal distributions are ν and ν', respectively. Posed a lower semi-continuous cost function $\mathsf{k} : \mathsf{Y} \times \mathsf{Y} \to [0, +\infty]$, let $\mathcal{K}(\pi) := \mathbb{E}_\pi[\mathsf{k}(Y, Y')] = \iint_{\mathsf{Y}\times\mathsf{Y}} \mathsf{k}(y, y') \mathrm{d}\pi(y, y')$ be the integral cost for transferring mass from ν to ν' under plan π. The Kantorovich formulation of the optimal transport problem consists of finding a transfer plan $\pi \in \Pi(\nu, \nu')$ that minimizes the integrated cost $\mathcal{K}$, i.e., in finding $K(\nu, \nu')$ such that

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

It can be shown that the Kantorovich problem in (1) has at least one solution if a transfer plan $\pi \in \Pi(\nu, \nu')$ with finite cost $\mathcal{K}(\pi) < \infty$ exists. A sufficient (and often encountered condition) is that $\mathsf{k}(y, y')$ is bounded by the sum of two nonnegative continuous and separate cost functions $\mathsf{a}_1(y)$, and $\mathsf{a}_2(y')$ such that $\mathbb{E}[\mathsf{a}_1(Y)] < +\infty$ and $\mathbb{E}[\mathsf{a}_2(Y)] < +\infty$ (bounded separate costs).

If Y and Y' are discrete random variables with probability mass functions given by, respectively, $Pr(Y = y_i) = s_i$ and $Pr(Y' = y'_j) = t_j$, with $s_i, t_j \geq 0$, $\sum_{i=1}^{I} s_i = \sum_{j=1}^{J} t_j = 1$, where I and J are natural numbers, then the Kantorovich problem amounts to solving the linear program

$$K(\nu, \nu') = \min \left\{ \sum_{ij} \mathsf{k}(y_i, z_j) p_{ij} : p_{ij} \geq 0,\ \sum_j p_{ij} = s_i,\ \sum_i p_{ij} = t_j \right\}. \tag{2}$$

When $\mathsf{k}(y, y') = \mathsf{d}^p(y, y')$ for a suitable continuous metric $\mathsf{d} : \mathsf{Y} \times \mathsf{Y} \to [0, +\infty)$, the Kantorovich problem

$$W_p^p(\nu, \nu') = \inf_{\pi \in \Pi(\nu, \nu')} \int \mathsf{d}^p(y, y') \, \mathrm{d}\pi(y, y') \tag{3}$$

defines the p-th power of the so-called Wasserstein distance of order p, W_p (henceforth Wasserstein distance).

Closed-form expressions for the Wasserstein distance are generally out of reach. However, in the multivariate case, $n_Y \geq 2$, let ν and ν' be two normal distributions with mean values m, m' and covariance matrices Σ, Σ' respectively. Then Givens and Shortt (1984) show that the squared 2-Wasserstein distance between ν and ν' is given by

$$\mathrm{WB}(\nu, \nu') = \|m - m'\|_2^2 + \mathrm{Tr}\left(\Sigma + \Sigma' - 2\left(\Sigma'^{1/2} \Sigma \Sigma'^{1/2}\right)^{1/2}\right), \tag{4}$$

where $\mathrm{Tr}(\cdot)$ denotes the matrix trace and $\Sigma^{1/2}$ is the symmetric square root of a symmetric and positive matrix. Equation (4) defines the Wasserstein-Bures semi-metric (Janati

et al. 2020) (henceforth denoted with $\mathrm{WB}(\nu, \nu')$). An interesting interpretation arises from the work of Gelbrich (1990), whose results show that Equation (4) can be interpreted as follows. The first term, $\|m - m'\|_2^2$ is the minimal cost for moving the distributions ν and ν' in such a way to match their first moments. The second term, $\mathrm{Tr}\left(\Sigma + \Sigma' - 2\left(\Sigma'^{1/2}\Sigma\Sigma'^{1/2}\right)^{1/2}\right)$ is the additional minimum cost for matching the second moments. In general, because we need to match more then the first two moments of ν and ν', we need to pay an extra cost and it is

$$W_2^2(\nu, \nu') \geq \mathrm{WB}(\nu, \nu'), \tag{5}$$

that is, the Wasserstein distance between ν and ν' is larger or equal to the Wasserstein-Bures distance. Here, a fresh look at the proofs of Gelbrich (1990) indicates that such inequality can be made sharper. In particular, under broad assumptions on ν and ν', it holds:

$$W_2^2(\nu, \nu') = \mathrm{WB}(\nu, \nu') + \Gamma(\nu, \nu') = \|m - m'\|_2^2 + \mathrm{Tr}\left(\Sigma + \Sigma' - 2\left(\Sigma'^{1/2}\Sigma\Sigma'^{1/2}\right)^{1/2}\right) + \Gamma(\nu, \nu'), \tag{6}$$

where $\Gamma(\nu, \nu') \geq 0$ is a non-negative residual term. Gelbrich (1990) proves that the residual term $\Gamma(\nu, \nu')$ is null when ν and ν' are two elliptical distributions with the same characteristic generator. Within the family of elliptical distributions with the same characteristic generator, $\mathrm{WB}(\nu, \nu') = 0$ implies that ν and ν' are the same distribution. Outside this family, $\mathrm{WB}(\nu, \nu') = 0$ implies only that they have identical means ($m = m'$) and variance-covariance matrices ($\Sigma = \Sigma'$).

In an influential work, Cuturi (2013) proposes to regularize the Kantorovich problem through a penalty term based on the Kullback-Leibler entropy of π w.r.t. a suitable reference probability measure ϑ

$$\mathrm{KL}(\pi|\vartheta) = \int \log\left(\frac{\mathsf{d}\pi}{\mathsf{d}\vartheta}\right) \mathsf{d}\pi, \tag{7}$$

with $\mathrm{KL}(\pi|\vartheta) = +\infty$ if π is not absolutely continuous w.r.t. ϑ.

A natural choice is to set ϑ as the product measure $\vartheta = \nu \times \nu'$, writing

$$K_\varepsilon(\nu, \nu') = \inf_{\pi \in \Pi(\nu, \nu')} \mathcal{K}(\pi) + \varepsilon \mathrm{KL}(\pi|\nu \times \nu'), \quad \varepsilon \geq 0, \tag{8}$$

where $\varepsilon \geq 0$ is called regularization parameter. Setting $\epsilon = 0$ recovers the unregularized problem. Problem (8) is referred to as the entropic OT problem. It admits a dual formulation, which can be expressed as:

$$\begin{aligned} K_\varepsilon(\nu, \nu') = \sup_{f_\epsilon, g_\epsilon \in \mathrm{C}_b(\mathsf{Y})} & \mathbb{E}[f_\epsilon(Y)] + \mathbb{E}[g_\epsilon(Z)] \\ & - \varepsilon\left(\iint \exp\left(\frac{f_\epsilon(y) + g_\epsilon(z) - \mathsf{k}(y, z)}{\varepsilon}\right) \mathsf{d}\nu(y)\, \mathsf{d}\nu'(z) - 1\right), \end{aligned} \tag{9}$$

where f_ϵ, g_ϵ belong to the class of continuous and bounded functions on Y. It is possible to prove that for $\varepsilon \to 0$ one regains the solution to the classical (Kantorovich) OT problem.

The results in Cuturi (2013) have paved the way to a flourishing research stream devoted to the algorithmic solution of problems (1) and (8). With some conceptual simplification,

one can consider three groups of algorithms, based respectively on linear programming, sorting and matrix scaling solvers. The first group contains 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). The second group relies on extending the univariate intuition that the one-dimensional Wasserstein distance can be obtained by suitable reordering of the data realizations. A multivariate algorithm that relies on a bubble-sort approach is presented in Puccetti (2017). The algorithm makes use of pairwise vector-comparisons and iterative swaps leading to an approximate solution of the classical OT problem in (1). These algorithms yield solutions of the Kantorovich problem in (1). The third class of algorithms solves the entropic problem in (8). 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 numerical solutions for the entropic problem in (8) or (9), which are approximating the solutions of the classical problem in (1).

Several other works in the management sciences have employed the Wasserstein distance as a metric. Mohajerin Esfahani and Kuhn (2018), Hanasusanto and Kuhn (2018) (see also (Chen et al. 2020)) use the Wasserstein distance to define ambiguity sets for robust optimization problems. Carlsson et al. (2018) and Wang et al. (2020) then apply it in their works regarding the definition and solution of the travelling salesman and the shortest path problems, respectively. Zhang et al. (2021) and Subramanyam et al. (2021) use the Wasserstein distance to obtain ambiguity sets for robust vehicle routing data-driven optimization. Recently, Bertsimas et al. (2022b) and Bertsimas et al. (2022a) employ the ∞-Wasserstein metric to define ambiguity sets for robust guarantees in two-stage and multi-stage stochastic optimization. Luo and Mehrotra (2019), Blanchet and Kang (2021) and Nguyen et al. (2022) use the Wasserstein distance in contexts at the intersection between optimization and statistical estimation. Luo and Mehrotra (2019) employ this metric to obtain an algorithm for robust optimization within a class of regression models; Blanchet and Kang (2021) employ the 2-Wasserstein distance to define a new robust inference approach called sample-out-of-sample inference; Nguyen et al. (2022) use the same metric to obtain robust data driven estimators of the inverse covariance matrix. To our knowledge, this manuscript is the first to explore the use of this metric in simulation experiments.

2.2. Global Sensitivity Analysis

In the management sciences, the term global sensitivity analysis appears for the first time in Wagner (1995). Wagner's approach starts with the generation of an input-output Monte Carlo sample. The sample is then post-processed via statistical methods and indications about the importance of the inputs are obtained by estimating variance-based sensitivity indices. Since then, the family of global sensitivity methods has expanded to include non-parametric regression approaches (Kleijnen and Helton 1999), moment-independent approaches (Baucells and Borgonovo 2013), value of information (Felli and Hazen 1998, Strong and Oakley 2013), Shapley values (Owen 2014) and other methods — See Razavi et al. (2021) for a perspective. In this section, we review the aspects of the literature that are most closely related to our work.

Let $Z = (X, Y)$ be a random variable on $(\Omega, \mathcal{B}, \mathbb{P})$, with support $\mathsf{Z} := \mathsf{X} \times \mathsf{Y}$, with Y and Y as previously defined, and X a random vector on $(\Omega, \mathcal{B}, \mathbb{P})$, with support $\mathsf{X} \subseteq \mathbb{R}^{n_X}$. Let us denote with $F_{XY}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ the joint cumulative probability distribution (cdf) function of (X, Y) and with $\mu(x)$, $F_X(x)$ the marginal probability measure and cumulative distribution function of X, respectively. A relevant role is played by the conditional probability distribution of Y given X. We denote with $\nu_x(y)$ and $F_{Y|X}(y)$, the corresponding probability measure and conditional cumulative distribution function.

Let $\mathscr{P}(\mathsf{Y})$ denote the set of all marginal probability distributions of Y. Consider a mapping $\zeta : \mathscr{P}(\mathsf{Y}) \times \mathscr{P}(\mathsf{Y}) \to [0, +\infty]$, whose value quantifies the discrepancy between two distributions in $\mathscr{P}(\mathsf{Y})$. We say that $\zeta(\cdot, \cdot)$ is a separation measurement, if it is null when the two marginal distributions are identical, i.e., $\zeta(\cdot, \cdot)$ satisfies $\zeta(\mathbb{Q}_Y, \mathbb{Q}_Y) = 0$ for all marginal distributions $\mathbb{Q}_Y \in \mathscr{P}(\mathsf{Y})$. Then, we define the global sensitivity index of X with respect to Y as

$$\xi^{\zeta}(Y, X) := \mathbb{E}_X \left[\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X}) \right]. \tag{10}$$

Several global sensitivity measures are written in the form of (10). If the output Y is a real number, and we select $\zeta^{\mathrm{V}}(\mathbb{P}_Y, \mathbb{P}_{Y|X}) = (\mathbb{E}[Y] - \mathbb{E}[Y|X])^2$ as separation measurement, we obtain the first order variance-based index of Wagner (1995), Sobol' (1993)

$$\xi^{W}(Y, X) = \mathbb{E} \left[(\mathbb{E}[Y] - \mathbb{E}[Y|X])^2 \right]. \tag{11}$$

Wagner's sensitivity measure $\xi^{W}(Y, X)$ represents the expected amount of reduction in output variance provided that we receive perfect information about X. Alternatively, if Y is absolutely continuous, one can use the L^1 norm between densities (Borgonovo et al. 2014), writing

$$\xi^{L_1}(Y, X) = \mathbb{E} \left[\frac{1}{2} \int_{\mathsf{Y}} \left| f_Y(y) - f_{Y|X}(y) \right| \mathrm{d}y \right]. \tag{12}$$

Equation (12) is also a representative of the family of global sensitivity measures based on Csiszar's divergences proposed in Rahman (2016).

With $\mathsf{Y} = \mathbb{R}$, setting

$$\xi^{\mathrm{Ku}}(Y, X) = \mathbb{E} \left[\sup_{y \in \mathbb{R}} \{F_Y(y) - F_{Y|X}(y)\} - \inf_{y \in \mathbb{R}} \{F_Y(y) - F_{Y|X}(y)\} \right], \tag{13}$$

one obtains the global sensitivity measure introduced in (Baucells and Borgonovo 2013), in which the separation measurement is the Kuiper distance, a generalization of the Kolmogorov-Smirnov metric. Gamboa et al. (2018) introduce a family of probabilistic sensitivity measures based on the Cramér-von Mises distance, defining

$$\xi^{\mathrm{CvM}}(Y, X) = \mathbb{E} \left[\int_{\mathsf{Y}} \left| F_Y(y) - F_{Y|X}(y) \right|^2 \mathrm{d}F_Y(y) \right]. \tag{14}$$

The probabilistic sensitivity framework of (10) does not require a functional relationship between Y and X. However, in simulation and machine learning, $Y = (Y_1, Y_2, \ldots, Y_{n_Y})$ is a quantity of interest (usually called output or target) calculated through a mathematical model, whose input is a random vector $X = (X_1, X_2, \ldots, X_{n_X})$, with X_i called input or feature. Then, we write the simulator input-output mapping that links Y to X as $Y = g(X, \mathcal{E}(X, \omega))$, where $g : \mathsf{X} \to \mathsf{Y}$ and where $\mathcal{E} : \mathsf{X} \times \Omega \to \mathsf{Y}$ is such that, for every value

of x, $\mathcal{E}(x)$ is a random vector on $(\Omega, \mathcal{B}, \mathbb{P})$. If $\mathcal{E}(X, \omega) \neq 0$ for some ω and X, the model is stochastic, deterministic otherwise. Then, let $\alpha \subseteq (1, 2, \ldots, n_X)$, $\alpha = (i_1, i_2, \ldots, i_k)$ be a subset of indices and let $X_\alpha = (X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ denote the corresponding group of features. If we are informed that $X_\alpha = x_\alpha$, then the model response becomes $Y = g(x_\alpha, X_{-\alpha}, \mathcal{E}(x_\alpha; X_{-\alpha}, \omega))$, where $-\alpha = \{1, 2, \ldots, n_y\} \setminus \alpha$ is the complementary set of α. Clearly, Y and $Y|X_\alpha = x_\alpha$ have probability measures $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_\alpha = x_\alpha}$.

The extension of variance-based indices $\xi^W(Y, X)$ to the multivariate case has been addressed in Lamboni et al. (2011) and Gamboa et al. (2014) with the introduction of generalized variance-based indices. In these works, independence among the inputs is assumed so that $\mu(x) = \prod_{i=1}^{n_X} \mu_i(x_i)$. Given $X = (X_1, X_2, \ldots, X_{n_X})$ and $Y = (Y_1, Y_2, \ldots, Y_{n_Y})$, we can define the variance-based importance measure of X_i with respect to any output Y_j, $i = 1, 2, \ldots, n_X$, $j = 1, 2, \ldots, n_Y$ via (11): $\xi^W(Y_j, X_i) = \mathbb{E}_{X_i}\left[(\mathbb{E}[Y_j] - \mathbb{E}[Y|X_i])^2\right]$. Then, let Σ_Y denote the variance-covariance-matrix of the output and $\mathbb{V}[Y]$ denote its trace, that is, the sum of the diagonal elements of Σ_Y. Assuming that $Y_j = g_j(X_1, X_2, \ldots, X_{n_X})$, i.e., the input-output mapping is deterministic, let the variance-based importance index of Lamboni et al. (2011) and Gamboa et al. (2014) be defined as

$$\xi^{LG}(Y, X_i) := \mathbb{V}[Y]^{-1} \sum_{t=1}^{n_Y} \xi^W(Y_t, X_i). \tag{15}$$

Thus, $\xi^{LG}(Y, X_i)$ is the fraction of the trace of the variance-covariance matrix of Y associated with X_i, when inputs are independent.

The works of Fraiman et al. (2020) and Gamboa et al. (2021) further extend ξ^{CvM} to the case in which the output belongs to a Riemannian manifold and to a metric space, respectively. Fort et al. (2021) address the sensitivity of models with stochastic output using these indices with the Wasserstein distance as a metric. An approach to create sensitivity measures for multivariate responses using distances between kernels is proposed in da Veiga (2021) and Barr and Rabitz (2022). This concise review shows that the definition of indices for vectorial outputs is an active research field, motivated also by industrial and machine learning applications (Marrel et al. 2017).

Recent studies by Chatterjee (2021), Wiesel (2022) and Deb et al. (2020) have refocused attention on the mathematical guarantees underlying the use of measures of statistical association. These guarantees include desirable properties such as zero-independence, max-functionality, and monotonicity. The first two properties originate from Postulates D and E in Rényi (1959). Postulate D (see also Axiom 1 in Móri and Székely (2019)) stipulates that a measure of statistical association is null if and only if Y is statistically independent of X. This property helps us to avoid the error of dismissing an input as unimportant when, in fact, it plays a role in the model. Postulate E (max-functionality) stipulates that the value of a global sensitivity measure is maximal if and only if Y is a deterministic function of X, i.e., if Y can be expressed as $g(X)$ for a mapping $g : \mathsf{X} \to \mathsf{Y}$. The third property, monotonicity, is associated with the following interpretation in our context: if we receive less refined information about an input, we require that such information is associated with a lower value of the global sensitivity measure than if we received more refined (or perfect) information on the same input. In the next section, we define global sensitivity measures based on OT functionals and discuss the conditions under which they possess these properties.

3. Global Sensitivity Measures based on Optimal Transport

This section introduces global sensitivity measures based on optimal transport, with the classical as well as entropic formulation of the cost function. It is structured as follows: definitions and properties are presented in Sections 3.1 and 3.2. Section D discusses in depth the interpretation of the new indices in light of their mathematical properties.

3.1. A family of OT-based indicators

Given the setup of Section 2.2, assume that the OT-functional in Equation (1) is associated with a continuous cost function $\mathsf{k} : \mathsf{Y} \times \mathsf{Y} \to [0, +\infty]$ which is null if and only if its two arguments are equal, that is $\mathsf{k}(y, y') = 0 \;\Leftrightarrow\; y = y'$.

Definition 1. Let X, Y be random variables with marginal distributions μ, ν respectively and let $(\nu_x^{\mathcal{F}})_{x\in\mathsf{X}}$ be the conditional distribution of Y generated by $(X, \mathcal{F})$. We call

$$\xi^K(Y, X|\mathcal{F}) := \mathbb{E}[K(\mathbb{P}_Y, \mathbb{P}^{\mathcal{F}}_{Y|X})] = \int_{\mathsf{X}} K(\nu, \nu_x^{\mathcal{F}})\, \mathrm{d}\mu(x) \tag{16}$$

the OT-based global sensitivity measure of $(X, \mathcal{F})$ with respect to Y.

A notable class of OT-based sensitivity measures is obtained using the p^{th}-power of Kantorovich-Rubinstein-Wasserstein distance W_p in (3) (as usual, we omit $\mathcal{F}$ when it coincides with $\mathscr{B}(\mathsf{X})$):

$$\xi^{\mathrm{W}_p^p}(Y, X) := \mathbb{E}\left[\inf_{\pi\in\Pi(\mathbb{P}_Y, \mathbb{P}_{Y|X})} \int \mathsf{d}^p(y, z)\, \mathrm{d}\pi(y, z)\right]. \tag{17}$$

In applications, it is often of interest to measure the relevance of a set of features/inputs, $(X_1, X_2, \ldots, X_{n_X})$. Ranking them by the magnitude of $\xi^K(Y, X_i)$ means to sort them based on the expected amount of work needed to optimally pass from the marginal (and current) probability measure of Y to the conditional (and updated) probability measure of Y given that we have received perfect information about X_i.

Remark 2. Equation (16) defines a family of global sensitivity measures. To illustrate, the well-known $\xi^{L_1}(Y, X)$ in Equation (12) can be reinterpreted as an OT-based sensitivity measure. In fact, if Y is equipped with the discrete metric, i.e., a metric such that for all $y,\ y' \in \mathsf{Y}$, $\mathsf{k}(y, y') = 0$ if $y = y'$ and $\mathsf{k}(y, y') = 1$ if $y \neq y'$, then, $\xi^K(Y, X) = \xi^{L_1}(Y, X)$ (see Appendix A for the calculations) — However in this case k is not continuous and the corresponding K is not strictly convex on Dirac-δ masses.

Proposition 3. *With the setup in Definition 1, $\xi^K(Y, X) \geq 0$ and $\xi^K(Y, X) = 0$ if and only if Y and X are statistically independent.*

(Please see Appendix A for all proofs).

Thus, the family of OT-based sensitivity measures in (16) possesses the zero-independence property. Proposition 3 then provides a lower-bound on $\xi^K(Y, X)$. In order to obtain an upper bound for $\xi^K(Y, X)$, we introduce the quantity

$$\mathbb{M}^K[Y] := \mathbb{E}[\mathsf{k}(Y, Y')] = \int_{\mathsf{Y}^2} \mathsf{k}(y, y')\, \mathrm{d}\nu(y)\, \mathrm{d}\nu(y'), \tag{18}$$

where Y' is an independent replica of Y. Notice that if k is the sum of two separately bounded cost functions then $\mathbb{M}^K[Y]$ is bounded.

For the next result, we recall that a Dirac measure δ_y centered at $y \in \mathsf{Y}$ is defined by $\delta_y(A) = 1$ if $y \in A$ and $\delta_y(A) = 0$ if $y \notin A$ for every $A \subset \mathsf{Y}$. The next Lemma states a useful property of the optimal transport cost with respect to Dirac measures.

Lemma 4. *Let K be the OT-functional in* (1) *with k continuous and let $\nu \in \mathscr{P}(\mathsf{Y})$ satisfy $\mathcal{K}(\nu \times \nu) < +\infty$. Then the function $K(\nu, \cdot)$ satisfies the following strict convexity inequality between Dirac measures in the support of ν:*

$$K\left(\nu, (1-t)\delta_{y_1} + t\delta_{y_2}\right) < (1-t)K\left(\nu, \delta_{y_1}\right) + tK\left(\nu, \delta_{y_2}\right) \tag{19}$$

for every $y_1, y_2 \in \operatorname{supp}(\nu)$, $y_1 \neq y_2$, $t \in (0,1)$.

Theorem 5. *Under the same assumptions of Lemma 4, for all random variables X, Y and every σ-algebra $\mathcal{F}$*

$$\xi^K(Y, X|\mathcal{F}) \leq \mathbb{M}^K[Y], \tag{20}$$

so that $\xi^K(Y, X|\mathcal{F})$ is finite if $\mathbb{M}^K[Y] < \infty$. Moreover $\xi^K(Y, X|\mathcal{F}) = \mathbb{M}^K[Y]$ if and only if Y is functionally dependent on X, i.e. $Y = f(X)$ $\mathbb{P}$-a.e. for some $\mathcal{F}$-measurable map $f : \mathsf{X} \to \mathsf{Y}$.

Theorem 5 states that the class of OT-based sensitivity satisfies the max-functionality property (Rényi's Postulate E) for a vast class of cost functions.

Remark 6. Equation 17 coincides with the numerator of the Wasserstein correlation coefficient defined by Wiesel (2022). Therein, an important result is Theorem 2.2 which shows that the Wasserstein correlation coefficient is maximal (equal to unity) in the case X and Y are related by a functional dependence and zero if they are statistically independent. Theorem 5 covers a slightly more general situation extending the result to general cost functionals k and offers a different perspective, based on the strict convexity of $K(\nu, \cdot)$ on Dirac measures stated in Lemma 4. This new approach allows us to extend this property to global sensitivity measures based on the entropic formulation of the OT problem.

Remark 7. Regarding the maximum value of (17), when $\mathsf{Y} = \mathbb{R}$ and p is an even integer, we have

$$\begin{aligned}\mathbb{M}^K[Y] = \int_{\mathbb{R}^2} (y-z)^p \,\mathrm{d}\nu(y)\,\mathrm{d}\nu(z) &= \sum_{k=0}^{p} (-1)^k \binom{p}{k} \int_{\mathbb{R}^2} y^k z^{p-k} \,\mathrm{d}\nu(y)\,\mathrm{d}\nu(z) \\ &= \sum_{k=0}^{p} (-1)^k \binom{p}{k} \mathbb{E}[Y^k]\,\mathbb{E}[Y^{p-k}]. \end{aligned} \tag{21}$$

When $\mathsf{Y} = \mathbb{R}^{n_Y}$ with $n_Y \geq 2$ and $\mathsf{k}(y, y') := \|y - y'\|_2^2$, we have

$$\mathbb{M}^{\mathrm{W}_2^2} = \int_{\mathsf{Y}^2} \|y - y'\|_2^2 \,\mathrm{d}\nu(y)\,\mathrm{d}\nu(y') = 2\int_{\mathsf{Y}} \|y - \mathbb{E}[Y]\|_2^2 \,\mathrm{d}\nu(y) = 2\mathbb{V}[Y], \tag{22}$$

where $\mathbb{V}[Y]$ denotes the trace of the variance-covariance matrix of Y.

The next result shows the monotonicity of ξ^K with respect to the information provided by $\mathcal{F}$ and can be applied to the case when we receive perfect information on a random variable U which is a transformation of X, $U = g(X)$.

Theorem 8. *For every σ-algebra $\mathcal{F} \subset \mathscr{B}(\mathsf{X})$ we have*

$$\xi^K(Y, X) \geq \xi^K(Y, X|\mathcal{F}). \tag{23}$$

In particular, if $g : \mathsf{X} \to \mathsf{U}$ is a Borel map with values in a Polish space U, $U := g \circ X$, and $\mathcal{F} = \sigma(g)$ is the σ-algebra generated by g, we have

$$\xi^K(Y, X) \geq \xi^K(Y, U) = \xi^K(Y,\, X|\mathcal{F}), \text{ where } \mathcal{F} = \sigma(g). \tag{24}$$

Moreover, if g is injective almost everywhere, then $\xi^K(Y, X) = \xi^K(Y, U)$.

Theorem 8 implies that receiving information in the form $U = g(X)$ has the same value as receiving direct information on X if the transformation g is injective. The fact that $\xi^K(Y, X)$ is greater than $\xi^K(Y, U)$ otherwise is consistent with the intuition that receiving direct information on X is more valuable than receiving "indirect" information via a transformation of X. Starting from Theorem 8, we can obtain an important continuity property with respect to an increasing family of σ-algebras in $\mathscr{B}(\mathsf{X})$.

Theorem 9. *Let $(\mathcal{F}^n)_{n\in\mathbb{N}}$ be an increasing family of sub-σ-algebras in $\mathcal{F}$ and let us denote by $\mathcal{F} = \bigvee_{n=1}^{\infty} \mathcal{F}^n$ the smallest σ-algebra containing each $\mathcal{F}_n$. We have*

$$\lim_{n\to\infty} \xi^K(Y,\, X|\mathcal{F}^n) = \xi^K(Y,\, X|\mathcal{F}). \tag{25}$$

Theorem 9 says that if we collect information on X in such a way to progressively refine the associated algebra towards $\mathcal{F}$, then $\xi^K(Y,\, X|\mathcal{F}^n)$ converges to $\xi^K(Y,\, X|\mathcal{F})$. Also, by Theorem 8, $\xi^K(Y, X|\mathcal{F}^n)$ is smaller than $\xi(Y, X|\mathcal{F})$ for any value of n, and thus $\xi^K(Y,\, X|\mathcal{F}^n)$ converges to $\xi^K(Y,\, X|\mathcal{F})$ from below. This result is also relevant for estimation, as discussed in Section 4.

We now consider separation measurements induced by the entropic OT functional (8). We then investigate whether these have the same properties as indicators based on the classical OT formulation. The next lemma shows that K_ε satisfies a property similar to Lemma 4.

Lemma 10. *Let K_ε be the entropic OT-functional in (8) with k continuous and let $\nu \in \mathscr{P}(\mathsf{Y})$ satisfy $\mathcal{K}(\nu \times \nu) < +\infty$. Then the function $K_\varepsilon(\nu, \cdot)$ satisfies the following strict convexity inequality between Dirac measures in the support of ν:*

$$K_\varepsilon\left(\nu, (1-t)\delta_{y_1} + t\delta_{y_2}\right) < (1-t)K_\varepsilon\left(\nu, \delta_{y_1}\right) + tK_\varepsilon\left(\nu, \delta_{y_2}\right) \tag{26}$$

for every $y_1, y_2 \in \operatorname{supp}(\nu)$, $y_1 \neq y_2$, $t \in (0, 1)$.

The next theorem summarizes results for $\xi^{K_\varepsilon}(Y, X)$.

Theorem 11. *Let $\varepsilon > 0$ and K_ε be the entropic OT-functional in (8). For all random variables X, Y the corresponding sensitivity index $\xi^{K_\varepsilon}(Y, X)$ satisfies*

$$\xi^K(Y, X|\mathcal{F}) \leq \xi^{K_\varepsilon}(Y, X|\mathcal{F}) \leq \mathbb{M}^K[Y], \tag{27}$$

so that $\xi^{K_\varepsilon}(Y, X|\mathcal{F})$ is finite if $\mathbb{M}^K[Y] < \infty$. Moreover $\xi^{K_\varepsilon}(Y, X|\mathcal{F}) = \mathbb{M}^K[Y]$ if and only if Y is functionally dependent on X, i.e. $Y = f(X)$ $\mathbb{P}$-a.e. for some $\mathcal{F}$-measurable map $f : \mathsf{X} \to \mathsf{Y}$. Eventually, $\xi^{K_\varepsilon}(Y, X)$ satisfies the same properties stated in Theorems 8, and 9.

Notice that (27) yields

$$\frac{\xi^K(X,Y)}{\mathbb{M}^K[Y]} \leq \frac{\xi^{K_\varepsilon}(Y,X)}{\mathbb{M}^K[Y]}. \tag{28}$$

This last inequality then allows a direct comparison of sensitivity measures based on classical and entropic formulations, as they are set on the same scale. Theorem 11 shows that global sensitivity measures based on the entropic OT, $\xi^{K_\epsilon}(Y,X)$, enjoy properties similar to those of indices based on the classical OT and stated in Theorems 5, 8, and 9. However, Proposition 3 does not hold for $\xi^{K_\varepsilon}(Y,X)$, since when $\text{supp}(\nu)$ is not reduced to a singleton (i.e. ν is not a Dirac measure) $K_\varepsilon(\nu,\nu) > 0$ and K_ε is not a strict separation measurement. When Y and X are independent, $\nu = \nu_x$ μ-a.e., and we have $\xi^{K_\varepsilon}(Y,X) = K_\varepsilon(\nu,\nu)$, which shows that ξ^{K_ε} does not possess the zero-independence property. The next result shows that $K_\varepsilon(\nu,\nu)$ is the minimum value of $\xi^{K_\varepsilon}(Y,X)$.

Proposition 12. *For every pair of random variables X, Y and every σ-algebra $\mathcal{F}$ we have*

$$\xi^{K_\varepsilon}(Y,X|\mathcal{F}) \geq K_\varepsilon(\nu,\nu). \tag{29}$$

Equality in (29) *is attained when Y and X are independent.*

Notice that $K_\varepsilon(\nu,\nu)$ is a minimum for $\xi^{K_\varepsilon}(Y,X)$, but there might exist ν' for which $K_\varepsilon(\nu,\nu') < K_\varepsilon(\nu,\nu)$. Supplementary Appendix B shows that equality in (29) does not imply that Y and X are independent.

Overall, the above results show that OT-based sensitivity measures (entropic and classical) possess intuitive properties that ease their interpretation: if Y is independent of X, information about X is irrelevant and reaches its lowest value (zero in the case of the classical OT formulation). Conversely, if Y is functionally dependent on X then the OT-based importance of X is maximal. In all other cases, the value of the OT-based sensitivity measure is in between these two extremes.

3.2. A Family of Sensitivity Indices

With the assumptions and analysis in the previous section, we can define the following sensitivity indices.

Definition 13. If $\mathbb{M}^K[Y] > 0$, we call

$$\iota^K(Y,X) = \frac{\xi^K(Y,X)}{\mathbb{M}^K[Y]} \tag{30}$$

and

$$\iota^{K_\epsilon}(Y,X) = \frac{\xi^{K_\epsilon}(Y,X)}{\mathbb{M}^K[Y]} \tag{31}$$

classical and entropic OT-based sensitivity index of X with respect to Y, respectively.

Then, by the zero-independence and max-functionality properties, we have $0 \leq \iota^K(Y,X) \leq 1$. The extreme values $\iota^K(Y,X) = 0$ and $\iota^K(Y,X) = 1$ indicate statistical independence and fully functional dependence, respectively. Differently, $\iota^{K_\epsilon}(Y,X)$ varies between its minimum and unity, with unity indicating functional dependence and the minimum being

reached when Y and X are independent. In the remainder, we shall focus on the case in which the cost is associated with the squared 2-Wasserstein distance, letting:

$$\iota(Y, X) = (2\mathbb{V}[Y])^{-1}\xi^{\mathrm{W}_2^2}(Y, X). \tag{32}$$

In general, closed form expressions for $\iota(Y, X)$ are out of reach. Nonetheless, a notable exception appears if the involved distributions are elliptical (Cambanis et al. 1981, Landsman and Valdez 2003). We say that Z follows an elliptically contoured distribution if its characteristic function can be represented in the form $\phi(z; \mu_Z, \Sigma_Z^*) = e^{iz^T\mu_Z^*}G(z^T\Sigma_Z^* z)$, where $G : \mathbb{R}^+ \to \mathbb{R}^+$ is called the characteristic generator (see (Cambanis et al. 1981, Theorem 2) for technical conditions), and μ_Z^* and Σ_Z^* are called location and dispersion parameters, respectively. One correspondingly writes $Z \sim \mathcal{EC}(\mu_Z^*, \Sigma_Z^*, G)$ where $\mathcal{EC}$ stands for elliptically contoured, as in Cambanis et al. (1981) — elliptical, for short. 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 $\Sigma_Z = -2G'(0+)\Sigma_Z^*$ (Cambanis et al. 1981, Theorem 4), where $G'(0+)$ is the right derivative of the characteristic generator at the origin. If a density function for an elliptical family exists, it has the form (Landsman and Valdez 2003, p. 58)

$$f_Z(z) = \frac{C}{\sqrt{|\Sigma_Z^*|}} G_d\left[\frac{1}{2}(z - \mu_Z^*)^T(\Sigma_Z^*)^{-1}(z - \mu_Z^*)\right], \tag{33}$$

where C is a normalizing constant, $|\cdot|$ stands for determinant and $G_d(\cdot)$ is the density generator.

Example 14. A representative of the family of elliptical distributions is the Gaussian family obtained with $G_d(\cdot) = e^{-\frac{1}{2}(\cdot)}$. In this case, we also have $\mu_Z^* = \mu_Z$ and $\Sigma_Z^* = \Sigma_Z$ and the density assumes the well-known expression

$$f_Z(z) = \frac{1}{(\sqrt{2\pi})^{n_Z}\sqrt{|\Sigma_Z|}} \exp\left[-\frac{1}{2}(z - \mu_Z)^T(\Sigma_Z)^{-1}(z - \mu_Z)\right], \tag{34}$$

where n_Z is the cardinality of Z. Other representatives are the Student-t, the logistic, and the exponential power distributions, which are obtained selecting alternative generators — please refer to (Landsman and Valdez 2003, p. 58-60) for the detailed expressions of these densities.

For OT-based sensitivity measures, a notable identity holds when both the marginal distribution and the conditional distributions of Y given X are elliptical.

Let us consider the global sensitivity index based on the Wasserstein-Bures semi-metric:

$$i^{\mathrm{WB}}(Y, X) := \frac{\mathbb{E}[\mathrm{WB}(\mathbb{P}_Y, \mathbb{P}_{Y|X})]}{2\mathbb{V}[Y]}, \tag{35}$$

where $\mathrm{WB}(\cdot, \cdot)$ is given in Equation (4).

Proposition 15. *Assume that the second moment of Y is finite. In general, it holds that*

$$\iota(Y, X) = i^{WB}(Y, X) + \frac{\mathbb{E}[\Gamma(\mathbb{P}_Y, \mathbb{P}_{Y|X})]}{2\mathbb{V}[Y]}, \tag{36}$$

so that $\iota(Y,X) \geq i^{WB}(Y,X)$. *If* $\mathbb{P}_Y$ *and* $\mathbb{P}_{Y|X}$ *are elliptical with the same characteristic generator* G *for values of* X *almost everywhere in* $\mathcal{X}$ *then*

$$\iota(Y,X) = i^{WB}(Y,X) = \text{Adv}(Y,X) + \text{Diff}(Y,X), \tag{37}$$

where

$$\text{Adv}(Y,X) = \frac{\mathbb{E}\left[\|\mathbb{E}[Y] - \mathbb{E}[Y|X]\|^2\right]}{2\mathbb{V}[Y]}, \tag{38}$$

and

$$\text{Diff}(Y,X) = \frac{\mathbb{E}\left[\text{Tr}\left(\Sigma_Y + \Sigma_{Y|X_i} - 2\left(\Sigma_{Y|X_i}^{1/2}\Sigma_Y\Sigma_{Y|X_i}^{1/2}\right)^{1/2}\right)\right]}{2\mathbb{V}[Y]}. \tag{39}$$

Equation (36) indicates that the OT-based sensitivity indices $\iota(Y,X)$ can be decomposed in two terms, the Wasserstein-Bures index in Equation (35) and a residual term given by $(2\mathbb{V}[Y])^{-1}\mathbb{E}[\Gamma(\mathbb{P}_Y, \mathbb{P}_{Y|X})]$. By Equations (38) and (39), the Wasserstein-Bures index is, in turn, the addition of two summands: $\text{Adv}(Y,X)$, that accounts for the difference in the first moments of $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X}$ and $\text{Diff}(Y,X)$ that involves the differences in their second moments (Σ_Y vs $\Sigma_{Y|X}$). These two terms are the expected optimal cost required for matching the first and second moments of $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X}$. If matching the first and second moment exhausts the transport of $\mathbb{P}_Y$ into $\mathbb{P}_{Y|X}$, then the residual term $\mathbb{E}[\Gamma(\mathbb{P}_Y, \mathbb{P}_{Y|X})]$ is null. Moreover, $\text{Adv}(Y,X)$ can be interpreted as an "advective part" that can be identified as a movement of the center of gravity, and $\text{Diff}(Y,X)$ as a "diffusive part" which leads to a dispersion (rotation) of the data points. — Supplementary Appendix D offers additional discussion on the interpretation of the advective and diffusive parts.

We also have a direct connection between the advective part of an optimal transport sensitivity measure and the generalized variance-based sensitivity measures of Lamboni et al. (2011) and Gamboa et al. (2014).

Proposition 16. *For the advective part of the Wasserstein-Bures global sensitivity measure, i.e.* $\text{Adv}(Y,X)$ *in* (38)*, it holds:*

$$\text{Adv}(Y,X) = (2\mathbb{V}[Y])^{-1}\sum_{t=1}^{n_Y}\xi^W(Y_t,X), \tag{40}$$

where $\xi^W(Y_t,X)$ *is Wagner's univariate sensitivity measure of* X *with respect to* Y_t *in Equation* (11) *and* $\xi^{LG}(Y,X)$*. Moreover, if we assume that the inputs are independent then*

$$\text{Adv}(Y,X) = \frac{1}{2}\sum_{t=1}^{n_Y} S(Y^t,X) = \frac{1}{2}\xi^{LG}(Y,X), \tag{41}$$

where $S(Y^t,X) = \dfrac{\xi^W(Y,X)}{\mathbb{V}[Y]}$ *is the Sobol' first order sensitivity index of* X *with respect to the* t^{th} *component of the output,* Y^t*.*

The first equality in Proposition 16 does not assume input independence and suggests that the numerator of the advective part of an OT-based sensitivity measure is the sum of the Wagner's univariate sensitivity measures of the output components $Y_1,\ldots,Y_{n_Y}$. If, in addition, we assume input independence, Equation (41) holds and $\text{Adv}(Y,X)$ differs only by a factor 1/2 from the sensitivity measures of Lamboni et al. (2011) and Gamboa et al. (2014).

Table 1: $\iota(Y, X_i)$, and associated decompositions into advective and diffusive parts for the model in Equation (44).

	$\iota(Y, X_i) = \iota^{WB}(Y, X_i)$	Adv_i	Diff_i	Perc. Adv.	$\iota_\epsilon(Y, X_i)$
X_1	0.492	0.294	0.198	60%	0.554
X_2	0.507	0.318	0.189	63%	0.575
X_3	0.117	0.107	0.01	91%	0.199

Corollary 17. *Let $X = (X_1, X_2, \ldots, X_{n_X})$, $X \sim \mathcal{EC}(m_X, \Sigma_X^*, G)$, $m_X = (m_1, m_2, \ldots, m_{n_X})$, with finite second moment. If $Y = AX+b$, where A is an $n_Y \times n_X$ matrix and $b \in \mathbb{R}^{n_Y}$, then the OT-based sensitivity measure between Y and X_i is given by (37) with $\Sigma_Y = A\Sigma_X A^T$, $\Sigma_{Y|X_i} = A\Sigma_i^c A^T$,*

$$\Sigma_i^c = (\sigma_{t,j}^i)_{t,j=1,2,\ldots,n_X}, \qquad \sigma_{t,j}^i = \sigma_{t,j} - \frac{\sigma_{t,i} \cdot \sigma_{i,j}}{\sqrt{\sigma_{i,i}}}, \tag{42}$$

and

$$m_{Y_k|X_i} = \sum_{j=1}^{n_X} a_{k,j} \left(m_j + (X_i - \mu_i) \frac{\sigma_{i,j}^i}{\sigma_{i,i}^i} \right), \tag{43}$$

for $k = 1, 2, \ldots, n_Y$.

Corollary 17 states that if the model output is a linear transformation of an elliptical input variable X, then we obtain a closed-form expression for $\iota(Y, X)$. This is due to the fact that all the involved distributions of Y, marginal and conditionals, are elliptical with the same characteristic generator. For instance, if X is a multivariate Gaussian or Student-t or logistic random variable, then all distributions of Y will be, respectively, Gaussian or Student-t or logistic, with parameters in Equations (42) and (43).

Example 18. Consider the input-output mapping $Y = g(X_1, X_2, X_3)$ given by:

$$\begin{cases} Y_1 = 4X_1 - 2X_2 + X_3 \\ Y_2 = 2X_1 + 5X_2 - X_3, \end{cases} \tag{44}$$

with X normally distributed, with mean $m_X = (1, 1, 1)$, and variance-covariance matrix $\Sigma_X = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$. Y is then normal with mean $m_Y = (3, 6)$ and variance-covariance matrix $\Sigma_Y = \begin{pmatrix} 15 & 7.5 \\ 7.5 & 33 \end{pmatrix}$. In this case, it is $\iota(Y, X_i) = \iota^{WB}(Y, X_i)$. The corresponding advective and diffusive parts are reported in Table 1. The third and fourth columns in Table 1 show that the advective part amounts at about 60% of the OT-based importance of X_1 and X_2, and at about 90% of the importance of $\iota(Y, X_3)$. The additional portion is due to the diffusive part, because all the involved distributions are elliptical. We use this example to illustrate the relationship between $\text{Adv}(Y, X)$ and Wagner's univariate sensitivity measures in Proposition 16. We calculate Wagner's variance-based importance measures for Y_1 and Y_2 separately.

The last column of Table 2 shows that, if we sum the Wagner's importance measures of the inputs with respect to each output, the sum exceeds the value of the corresponding

Table 2: Wagner's importance measures for the example.

	$\xi^W(\cdot, X_1)$	$\xi^W(\cdot, X_2)$	$\xi^W(\cdot, X_3)$	$\sum_{i=1}^{3} \xi^W(\cdot, X_i)$
Y_1	12.25	0.25	4	16.5
Y_2	16	30.25	6.25	42.5
$\xi^W(Y_1, X_i) + \xi^W(Y_2, X_i)$	28.5	30.50	10.25	

variance for both Y_1 and Y_2: this occurs because the inputs are not orthogonal. If we sum across the outputs, we get the values in the third row of Table 2. Dividing these sums by twice the diagonal of the variance-covariance matrix, we obtain the values of the advective part of the Wasserstein-Bures importance measure, in accordance with Proposition 16.

According to equation (37), an OT-based importance measure includes extra terms compared to a generalized variance-based index. Thus, the global impact of an input on the output distribution is more than just the sum of its individual variance-based sensitivities. By the properties of the Wasserstein distance, we know that the Wasserstein-Bures distance between ν and ν' is always lower than or equal to their squared Wasserstein-2 distance: $WB(\nu', \nu') \leq W_2^2(\nu', \nu')$. This inequality yields a corresponding inequality on the corresponding global sensitivity indices: $\iota^{WB}(Y, X) \leq \iota(Y, X)$. As a result, we can expect that if learning X only affects the first order moment m_Y of Y, then the importance of X is equal to $\mathrm{Adv}(Y, X)$, or the sum of univariate sensitivities. However, if there is also an impact on the second order moment Σ_Y, then a diffusive component is present. These two components sum to $\iota^{WB}(Y, X)$ and account for the input entire importance when all the involved distributions are elliptical with the same characteristic generator. The presence of an additional gap between $\iota^{WB}(Y, X)$ and $\iota(Y, X)$ suggests that information about X impacts the distribution of Y beyond its first two moments.

A recent result in Janati et al. (2020) allows us to obtain closed form expressions for the entropic OT-based sensitivity measures when the marginal and conditional distributions are normal. With the notation of Corollary 17, if X, Y and $Y|X$ are normally distributed, given $\varepsilon \geq 0$, then the entropic sensitivity index in (31) can be written as

$$\iota_\varepsilon^{\mathrm{WB}}(Y, X) = \mathrm{Adv}(Y, X) + \frac{\mathbb{E}[\mathrm{Tr}\left(\Sigma_Y + \Sigma_{Y|X} - D_\varepsilon\right) + L(D_\varepsilon, \varepsilon)]}{2\mathbb{V}[Y]}, \tag{45}$$

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

$$L(D_\varepsilon, \varepsilon) = \tfrac{\varepsilon}{2}\left(n_Y \cdot (1 - \log(\varepsilon)) + \log\det\left(D_\varepsilon + \tfrac{\varepsilon}{2} I\right)\right). \tag{46}$$

The terms D_ε and $L(D_\varepsilon, \varepsilon)$ appear in $\iota_\varepsilon(Y, X)$ rather than in $\iota^{\mathrm{WB}}(Y, X)$ in (37), as a consequence of the entropic penalty.

We close the investigation of the properties of OT-based sensitivity measures studying the behavior of entropic indices for large values of the regularization parameter.

Theorem 19. *Given $\iota_\varepsilon(Y, X_i)$ in Equation (31), we have:*

$$\lim_{\varepsilon \to \infty} \iota_\varepsilon(Y, X) = 1 \tag{47}$$

for any X.

Theorem 19 implies that the entropic importance of any random variable tends to the maximum value if the regularization parameter grows. Then, $\iota_\varepsilon(Y, X)$ becomes uninformative for large values of ε, as all X_i's are assigned the same value of $\iota_\varepsilon(Y, X)$.

Example 20 (Example 18 continued.). The last column of Table 1 reports the values of the entropic OT-based sensitivity indices $\iota_\varepsilon(Y, X_i)$ for the same input-output mapping and input distributions in Example 18. For illustrative purposes, we have set $\varepsilon = 1$. The values in Table 1 indicate that the entropic sensitivity measures are larger than the classical sensitivity measures for all inputs, in accordance with Theorem 11. The increases are systematic and the ranking is unchanged. However, there is no reassurance that this is maintained for any value of the regularization parameter. Increasing its value to $\varepsilon = 10$, we record $\iota_{\varepsilon=10}(Y, X_1) = 0.95$, $\iota_{\varepsilon=10}(Y, X_2) = 0.96$, $\iota_{\varepsilon=10}(Y, X_3) = 0.89$. While the ranking is maintained, the value of the entropic importance of X_3 increases notably. In agreement with Theorem 19, for higher values of ε we obtain $\iota_\varepsilon(Y, X_i) \approx 1$ for all the three inputs and we become unable to rank them.

4. Estimation

The estimation of global sensitivity measures in the common rationale of Equation (10) is widely recognized as a challenging task. A *brute force* implementation requires a double-loop of Monte Carlo simulations: an outer loop in which values of X are fixed and an inner loop in which the model is evaluated to obtain the conditional distribution of Y given X. The computational cost associated with this strategy is $C^{\text{Brute Force}} = n_X N_{\text{out}} N_{\text{inn}}$ model evaluations, where N_{out} and N_{inn} are the sample sizes allocated to the outer and inner loops, respectively. This cost is of the order of the square of the sample size and depends on the model input dimensionality n_X. *Nested estimation* is widely encountered in the management sciences. To illustrate, in the pricing of financial instruments the outer loop is needed to generate a set of scenarios in which a number the risk factors are fixed, while the inner loop calculates the future cash flows conditional on the scenario (Broadie et al. 2011, 2015). Several studies have addressed the reduction of numerical cost in problems involving nested estimation (see Gordy and Juneja (2010) for a review). Hong et al. (2017) propose smoothing approaches to reduce the computational burden of the inner conditional expectation. The same problem has been studied in parallel in the statistical literature for the estimation of global sensitivity measures, with the pick-and-freeze design as the first successful proposal to decrease the computational cost down to $\approx n_X(N + 1)$ model runs (Saltelli 2002, Gamboa et al. 2016). On the other hand, given-data (or once-through) designs bring the number of model evaluations down to N model evaluations, where N is the size of a single-loop Monte Carlo sample. The corresponding computational cost is then independent of the problem dimensionality n_X. Moreover, the design allows the calculation of global sensitivity measures also when the input-output sample come from data collection. We follow Pearson's intuition underlying the correlation ratio (Pearson 1905). One considers partitioning the support of X, X, into H non overlapping subsets, X_i^h, $h = 1, 2, \ldots, H$. Then, one writes an estimate of a global sensitivity measure in the common rationale of (10) as

$$\widehat{\xi}(Y, X_i; N, H) = \frac{1}{H} \sum_{h=1}^{H} \zeta(\mathbb{P}_Y^N, \mathbb{P}_{Y|X_i \in \mathsf{X}_h^i}^N), \tag{48}$$

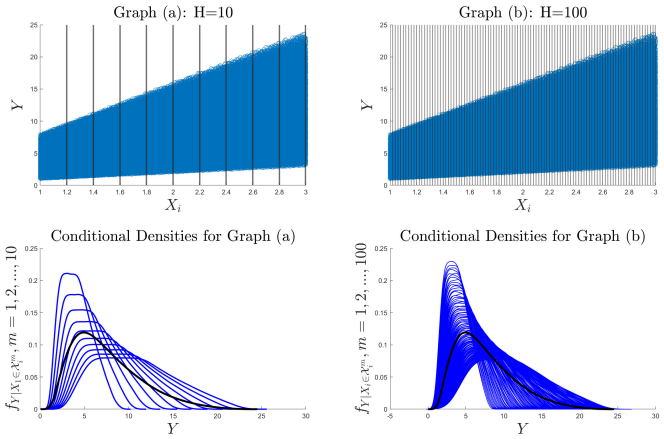


Figure 1: Scatterplot partitioning with $H = 10$ and $H = 100$ for hypothetical Y and X_i. The upper graphs display the scatterplot and the partitions, the lower graphs the corresponding empirical distributions (empirical densities are available in this case).

where $\zeta(\mathbb{P}_Y^N, \mathbb{P}^N_{Y|X_i \in \mathsf{X}_i^h})$ is an empirical estimate of the separation between the marginal and the conditional property of interest required by $\zeta(\cdot, \cdot)$, here denoted by $\mathbb{P}_Y^N$ and $\mathbb{P}^N_{Y|X_i \in \mathsf{X}_i^h}$, respectively. In the latter, the point condition $X_i = x_i$ is replaced by the bin condition $X_i \in \mathsf{X}_i^h$.

We can implement Equation (48) through the following steps. First, we build the scatterplot with X_i and Y on the horizontal and vertical axis, respectively. Next, we partition the horizontal axis into H bins X_i^h, $h = 1, 2, \ldots, H$, such that the union of all bins equals X_i and the intersection of any two bins is empty. Graph (a) in Figure 1 offers a visualization of this partitioning into ten intervals of the horizontal axis of a hypothetical scatterplot. The third step is to consider the separation between the empirical marginal distribution $\mathbb{P}_Y^N$ and the conditional marginal distribution $\mathbb{P}^N_{Y|X_i \in \mathsf{X}_i^h}$, that is, to compute $\zeta(\mathbb{P}_Y^N, \mathbb{P}^N_{Y|X_i \in \mathsf{X}_i^h})$. The estimate $\widehat{\xi}(Y, X_i; N, H)$ is then the average of these values.

Pearson's intuition suggests that, if the partition is sufficiently refined, the bin condition $\widehat{\mathbb{P}}_{Y|X_i \in \mathsf{X}_i^h}$ tends to the point condition $\widehat{\mathbb{P}}_{Y|X_i = x_i}$. (Graph (b) in Figure 1 shows a scatterplot partition with the cardinality increased to $H = 100$.) Then, if $\zeta(\mathbb{P}_Y^N, \mathbb{P}^N_{Y|X_i \in \mathsf{X}_i^h})$ is an accurate approximation of $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i \in \mathsf{X}_i^h})$, the value of $\widehat{\xi}(Y, X_i)$ should be close to $\xi(Y, X_i)$. More precisely, we expect that as the sample size N and the cardinality of the partition tend to infinity, then $\widehat{\xi}(Y, X_i)$ tends to $\xi(Y, X_i)$. The convergence depends on the properties of $\zeta(\widehat{\mathbb{P}}_Y^N, \widehat{\mathbb{P}}^N_{Y|X_i \in \mathsf{X}_i^h})$ and a general proof is nowadays missing.

However, we show in Supplementary Appendix A that if $\zeta(\widehat{\mathbb{P}}_Y^N, \widehat{\mathbb{P}}^N_{Y|X_i \in \mathcal{X}_i^m})$ is based on optimal transport, then

$$\lim_{H \to \infty, N \to \infty} \widehat{\xi}(Y, X_i; N, H) = \xi(Y, X_i). \tag{49}$$

A fundamental role in this result is played by the convexity and monotonicity of the OT functional in Equation (16). These two properties also imply that for N sufficiently large,

the estimates $\widehat{\xi}^K(Y,X;N,H)$ approximate the true value $\xi^K(Y,X)$ from below as the partition size increases. By Theorem 11, the same holds for the case in which the quantity to be estimated is an entropic-OT based sensitivity measure.

To complete the estimation procedure, we need an algorithm for solving the data-driven optimal transport problem between the two empirical distributions $\widehat{\mathbb{P}}_Y^N$ and $\widehat{\mathbb{P}}_{Y|X_i\in \mathsf{X}_i^h}^N$ for $h=1,2,\dots,H$. If the cost function is the squared Wasserstein metric, the problem is:

$$\begin{gathered}\inf_{\mathbf{s}} \sum_{k=1}^{N}\sum_{j:x_{j,i}\in \mathsf{X}_i^h} s_{k,j}\sum_{t=1}^{n_Y}(y_{k,t}-y_{j,t})^2 \\ \text{subject to} \\ \sum_{k=1}^{N} s_{k,j}=\tfrac{1}{N},\ \sum_{j:x_{j,i}\in \mathsf{X}_i^h} s_{k,j}=\tfrac{1}{N_h},\quad N_h=\#\{j:x_{j,i}\in \mathsf{X}_i^h\},\end{gathered} \tag{50}$$

for $h=1,2,\dots,H(M)$, where $\mathbf{s}$ is the set of (empirical) couplings, $\#\{\cdot\}$ denotes cardinality, N_m counts the realizations of X which are included in X_i^h; the realizations $y_{k,t}$ follow $\mathbb{P}_Y$, while the realizations $y_{j,t}$ follow $\mathbb{P}_{Y|X\in \mathsf{X}_i^h}$.

The algorithm that solves the OT problem in (50) is crucial because the estimation requires solving a conditional OT-problem in each partition set. However, if Y is univariate ($n_Y=1$), the solution is streamlined by results in works such as Vallender (1974), Cambanis et al. (1976). Given $u\in[0,1]$ let $Q_Y(u)$ be the u^{th} quantile of Y and $Q_{Y|X_i}(u)$ the u^{th} quantile of Y given X_i. By Cambanis et al. (1976) we can write:

$$W_2^2(\mathbb{P}_Y,\mathbb{P}_{Y|X})=\int_0^1\Big(Q_Y(u)-Q_{Y|X_i}(u)\Big)^2\,du. \tag{51}$$

Thus, the squared 2-Wasserstein distance can be found by integrating the squared difference of the quantile functions. Numerically, it is then enough to reorder the marginal and conditional quantiles of Y in each partition, calculate their squared differences, and take the average over the partitions.

If $n_Y\geq 2$, our work intersects with the growing body of literature on solvers for the optimal transport problem. The literature displays two main strategies. We can opt for an exact solver obtaining the exact value of $K(\widehat{\mathbb{P}}_Y^N,\widehat{\mathbb{P}}_{Y|X\in \mathsf{X}_i^h}^N)$. We use an implementation of the network simplex in our experiments. Alternatively, we can opt for an approximate solver. The proposal of Cuturi (2013) is to employ the entropic problem in (52), for which faster solvers are available. The given-data problem is:

$$\begin{gathered}\inf_{\mathbf{s}^\varepsilon} \sum_{k=1}^{N}\sum_{j:X\in \mathsf{X}_i^h}\left(s_{k,j}^{\varepsilon}\textstyle\sum_{t=1}^{n_Y}(y_{k,t}-y_{j,t})^2+\varepsilon\exp\left(-\frac{\sum_{t=1}^{n_Y}(y_{k,t}-y_{j,t})^2}{\varepsilon}\right)\right) \\ \text{such that} \\ \sum_{i=1}^{N} s_{i,j}^{\varepsilon}=\tfrac{1}{N},\ \sum_{j:X\in \mathsf{X}_i^h} s_{i,j}^{\varepsilon}=\tfrac{1}{N_h},\quad N_h=\#\{j:x_{j,i}\in \mathsf{X}_i^h\},\end{gathered} \tag{52}$$

Cuturi (2013)'s algorithm based on Sinkhorn iterations yields the solution of Problem (52) in computationally fast times. For small values of the regularization parameter, the obtained solution can then be used as a proxy for the solution of the classical given-data problem in (50). The trade-off is then between precision and speed. We also implement two further alternatives: the sorting approach of Puccetti (2017) which provides an approximate solution to Problem (50), and the Wasserstein-Bures approximation. A given-data

estimate of the Wasserstein-Bures index is given by

$$\hat{\iota}^{WB}(Y,X) = \frac{1}{2\widehat{\mathbb{V}}[Y]} \sum_{h=1}^{H(N)} \frac{N_h}{N} \left(\sum_{t=1}^{n_Y} (\hat{m}_{Y,t} - \hat{m}_{Y,t|X\in \mathsf{X}_i^h})^2 + \right.$$
$$\left. \mathrm{Tr}\left(\widehat{\Sigma}_Y + \widehat{\Sigma}_{Y|X\in\mathsf{X}_i^h} - 2\left(\sqrt{\widehat{\Sigma}_Y}\widehat{\Sigma}_{Y|X\in\mathsf{X}_i^h}\sqrt{\widehat{\Sigma}_Y} \right)^{1/2} \right) \right)^{\frac{1}{2}}, \quad (53)$$

where $\hat{m}_{Y,t}$ and $\hat{m}_{Y,t|X\in\mathsf{X}_i^h}$ are empirical means, $\widehat{\Sigma}_Y$ and $\widehat{\Sigma}_{Y|X\in\mathsf{X}_i^h}$ empirical covariance matrices. It is an immediate corollary of Theorem 21 that $\hat{\iota}^{WB}(Y,X)$ is asymptotically consistent provided that the variance-covariance matrix estimators are. The calculation of (53) is computationally fast, because it involves only linear algebra operations. However, when distributions are not elliptical $\hat{\iota}^{WB}(Y,X)$ in (53) cannot be regarded as an estimate of $\iota(Y,X)$.

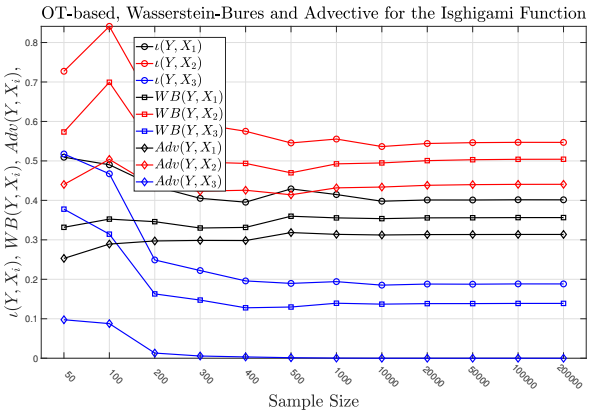
5. Experiments for Univariate and Multivariate Output Test Cases

This section is divided into two parts. In the first part, we discuss experiments for two univariate test cases. In the second part, we discuss a multivariate output test case in which it is possible to obtain $\iota(Y,X_i)$ analytically. All experiments are performed on a personal PC, with an Intel(R) Core(TM) i7-7700HQ CPU 2.80GHz processor and 64GRAM, subroutines implemented in MatLab.

5.1. Univariate Output Test Cases

Our first experiments are based on the well-known Ishigami function (Ishigami and Homma 1990). The input output mapping is given by $Y = \sin(X_1)(1 + 0.1X_3^4) + 7\sin(X_2)^2$ with X_1, X_2 and X_3 independent and uniformly distributed on $[-\pi,\pi]$. The values of variance-based sensitivity measures are analytically known, with $\xi^W(Y,X_3) = 0.31$, $\xi^W(Y,X_3) = 0.44$ and $\xi^W(Y,X_3) = 0$, a false negative. Analytical expressions of $\iota(Y,X_i)$ are out of reach, however calculations can be performed numerically. In fact, the Ishigami model is extremely fast to run, and we can study the estimates for large sample sizes. We apply the given data strategy with the estimator in Equation (48). Also, because the output is univariate, we can use the reordering strategy to find the Wasserstein-2 distance between the marginal and conditional distribution of Y in each partition. Figure 2 reports results for a numerical experiment in which the sample size is increased from $N = 50$ to $N = 200000$.

Table 2b reports the values of the estimates $\widehat{\iota}(Y,X_i)$, $\widehat{\iota}^{WB}(Y,X_i)$ and $\widehat{\mathrm{Adv}}(Y,X_i)$ at $N = 200000$. The fourth row shows that twice the values of the estimates of $\mathrm{Adv}(Y,X_i)$ coincide with the analytical values of the first-order variance-based indices, in agreement with Equation (41). The values of $\widehat{\iota}^{WB}(Y,X_i)$ are greater than the values of $\widehat{\mathrm{Adv}}(Y,X_i)$ for all three inputs, signaling the presence of a diffusive component. Relying on $\iota^{WB}(Y,X_i)$ already avoids the false negative for X_3, as $\widehat{\iota}^{WB}(Y,X_3) > 0$. The values of $\widehat{\iota}(Y,X_i)$ are, in turn, greater than the values of $\widehat{\iota}^{WB}(Y,X_i)$ for all three inputs. Because $\iota(Y,X_i)$ accounts for the complete transport between the marginal and conditional distributions,



(a) Asymptotic behavior of the estimates of $\iota(Y, X_i)$ (circle $\circ$), $\iota^{\mathrm{WB}}(Y, X)$ (Square $\square$), $\mathrm{Adv}(Y, X_i)$ (diamonds $\diamond$) for the Ishigami function. Red, Importance Measures of X_1; Blue, Importance Measures of X_2; Black, Importance Measures of X_3.

	X_1	X_2	X_3
$\widehat{\iota}(Y, X_i)$	0.40	0.55	0.19
$\widehat{\iota}^{WB}(Y, X_i)$	0.36	0.50	0.14
$2 \cdot \widehat{\mathrm{Adv}}(Y, X_i)$	0.31	0.44	0.00

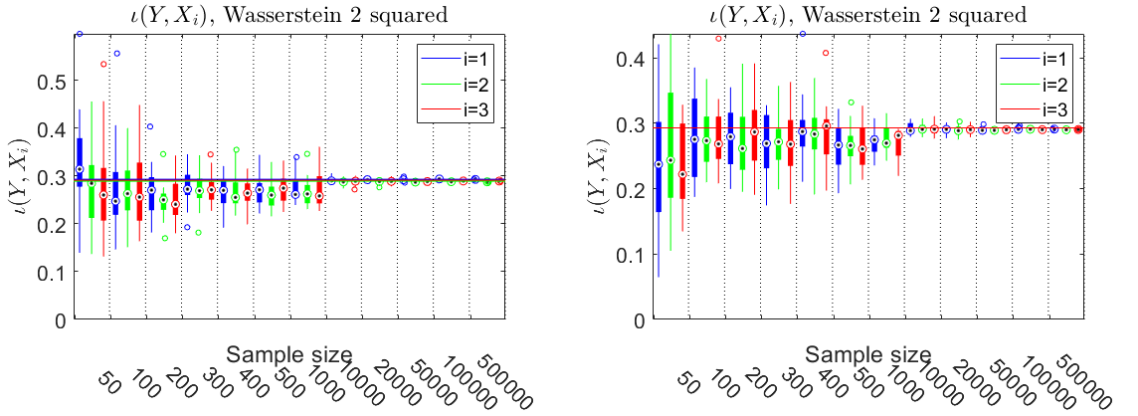
(b) Estimates of $\iota(Y, X_i)$, $\iota^{WB}(Y, X_i)$ and $\mathrm{Adv}(Y, X_i)$ at $N = 200,000$ for the Ishigami function.

Figure 2: Right graph: Estimates of $\iota(Y, X_i)$, $\iota^{WB}(Y, X_i)$ and $\mathrm{Adv}(Y, X_i)$ at increasing sample sizes and varying partitions. Left Table: Values at $N = 200000$ and $H = 60$.

these values indicate that the advective and diffusive parts do not fully explain the change in distributions for the case of the Ishigami function, in accordance with the fact that the involved distributions are not normal. Also, the value $\widehat{\iota}(Y, X_3) > 0$ confirms that Y is statistically dependent on X_3.

In our second test case, we perform experiments to analyze a higher dimensional setting. We consider a linear input-output mapping, $Y = aX^T$, with the number of inputs equal to $n_X = 999$ and $n_X = 9999$. We let X be a multivariate normal random vector, with pairwise correlations $\rho_{i,j} = 0.5$, $i, j = 1, 2, \ldots, n_X$ $(i \neq j)$. We then assign the weights as $a = [a_1, a_2, \ldots, a_{999}]$ 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 $i = 667, 668, \ldots, 999$. (A similar 3-groups split is performed for the 9999 case). Given this assignment, Y is correspondingly normal, with mean equal to 0 and variance $\mathbb{V}[Y] = 5.03\text{E}5$ and to $\mathbb{V}[Y] = 1.10\text{E}6$, for $n_X = 999$ and $N_X = 9999$, respectively. Because all conditional distributions are normal, for this test case it is possible to obtain the values of $\iota(Y, X_i)$ analytically. We calculate the expressions using the software MATHCAD. The values are $\iota(Y, X_i) = 0.293$, for the first input group, $\iota(Y, X_i) = 0.289$ for the second and $\iota(Y, X_i) = 0.291$ for the third, respectively, for $n_X = 999$. Thus, the inputs are ranked according to their weight, which is intuitive for linear models. However, the global sensitivity measures of the three input groups are close. This effect is due to the presence of correlations. For the case $n_Y = 9999$, the analytical calculations yields almost identical values for all three input groups with $\iota(Y, X_i) = 0.293$, $i = 1, 2, 3$.

Figure 3 reports estimates for samples generated using crude Monte Carlo with sizes from $N = 50$ to $N = 500000$, with 20 replicates at each sample size. As the sample size increases, we vary the partition cardinality from $H = 8$ to $H = 60$. Overall, the analysis takes 450 seconds in the $n_X = 999$ case and about 8 hours in the $n_X = 9999$ case. Figures



(a) Correlated normal test case with $n_X = 999$. (b) Correlated normal test case with $n_X = 9999$.

Figure 3: Numerical estimation of probabilistic sensitivity measures for the $n_X = 999$ and $n_X = 9999$ correlated normal random variables test case. Boxplots represent variability over 20 replicates. The sample size increases from $N = 50$ to $N = 500000$.

3a and 3b show that the estimates tend to the corresponding analytical values as the sample size increases.

5.2. Multivariate Normal Output Test Case

We report results of experiments aimed at illustrating Theorems 9 and 11 and the convergence from below of the estimates in Equation (48), for classical as well as entropic-OT-based global sensitivity indices. Results regarding the computational times needed to solve the given-data OT problems in Equations (50) and (52) follow. As a benchmark, we consider the input-output mapping and distributions in Example 18.

We fix the sample size at $N = 50000$ and implement the estimator in Equation (48) for partition cardinalities increasing from $H = 5$ to $H = 200$. Benchmarks for the numerical experiments are the analytical values of the sensitivity measures reported in Tables 1. Figure 4 displays the results.

The graphs in Figure 4 show that the estimates $\widehat{\iota}(Y, X_i)$ and $\widehat{\iota}_\epsilon(Y, X_i)$ tend to the corresponding analytical values from below. In fact, refining the partition can be interpreted as obtaining increasingly precise information on X_i and therefore as obtaining an algebra which is getting closer and closer to the algebra generated by X_i. We also observe that the estimates are almost insensitive to choices of the partition size H between 80 and 200. This *plateau* effect is in line with previous experiments on given-data estimators in Strong and Oakley (2013): For sufficiently large N one finds a range of values of H for which estimates show very little variability. Then, Strong and Oakley (2013) suggest to pick one of these values for reporting.

Next, we display results for a set of experiments aimed at investigating asymptotic behavior and run times when alternative algorithmic approaches (Table 3) are used to solve the given-data OT problem in Equation (50). For the Sinkhorn algorithm, we set the regularization parameter at $\varepsilon = \ell \cdot \|Q\|_\infty$, with $\ell = 0.001$ and Q is the cost matrix

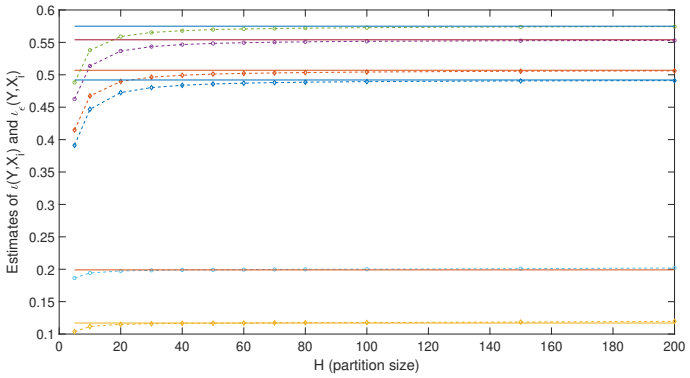


Figure 4: Vertical axis: Estimates of $\iota(Y, X_i)$ ($\diamond$) and $\iota_\epsilon(Y, X_i)$ ($\circ$) for the multivariate-output analytical test case. Horizontal axis: partition cardinality varies from $H = 5$ to $H = 200$. Dotted lines represent estimates, continuous lines analytical values (in Table 1).

Table 3: Average computational times (in seconds) for the calculations in Figure 5.

Sample Size	N=50	N=100	N=250	N=500	N=1000
Network Simplex	0.0319	0.0441	0.1691	0.6959	2.6304
Sinkhorn	0.0140	0.0014	0.0043	0.0065	0.0123
Swap	0.0089	0.0007	0.0009	0.0034	0.0086
Bures	0.0115	0.0005	0.0001	0.0001	0.0003

whose elements equal $Q_{k,j} = \sum_{t=1}^{n_Y} (y_{k,t} - y_{j,t})^2, k, j = 1, 2, \ldots, n_X$ and use the solution as a proxy for the classical OT-problem. Figure 5 shows results at increasing sample sizes (horizontal axis), with $N = (\,50,\, 100,\, 250,\, 500,\, 1000\,)$ and partitions set at $H(N) = (\,2,\, 5,\, 7,\, 8,\, 10\,)$. The first, second and third graphs report estimates (dotted lines) obtained when the OT problem in each partition is solved, respectively, with the network simplex, the Sinkhorn and the Swap algorithms. The fourth graph reports estimates using (53). All graphs show that at samples of about 250 realizations the estimates are close to the analytical values (continuous lines).

Table 3 reports the running times of the algorithms. Notice that at $N = 250$, with 7 partitions, we register a total of 21 optimization problems of size 250×35 to be solved, at $N = 1,000$ we have 30 problems of size $1000 \cdot 100$. The numbers in Table 3 show that the estimator that solves Problem (50) with the simplex algorithm is several times slower than the remaining algorithms. For instance, at $N = 1000$ the simplex algorithm takes on average ≈ 2.6 seconds to solve one instance, the Sinkhorn and the Swap algorithms about ≈ 0.012 seconds, the estimator in (53) about ≈ 0.0003 seconds.

We conclude by presenting the results of experiments at increasing values of the regularization parameter ε in the Sinkhorn approximation. We use a sample size of $N = 1000$ and, in addition to $\ell = 0.001$ adopted in the previous experiments, we consider $\ell = 0.01$, $\ell = 0.1$, $\ell = 1$ and $\ell = 10$. At $\ell = 0.01$ estimates increase of about 10% for the three inputs, at $\ell = 0.01$ they increase of about 65% for the first two inputs, and of about 400% for the third input. At $\ell = 1$ the estimates are close to 0.97 for all three inputs and at

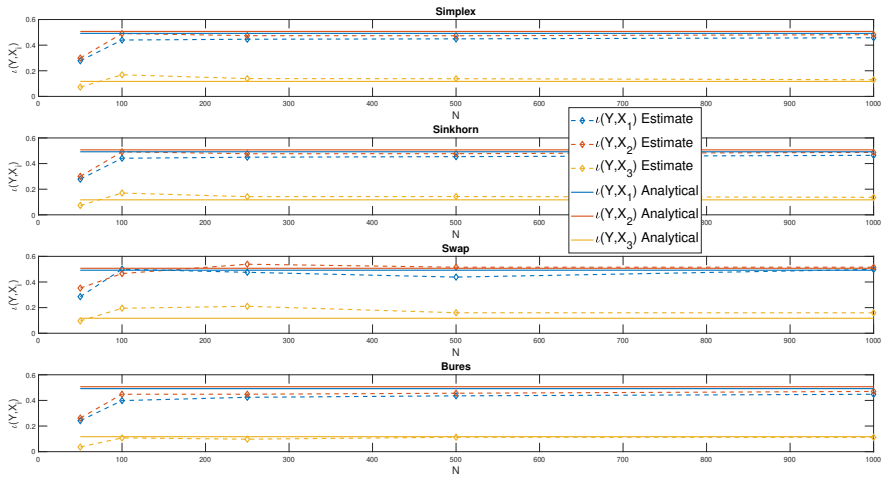


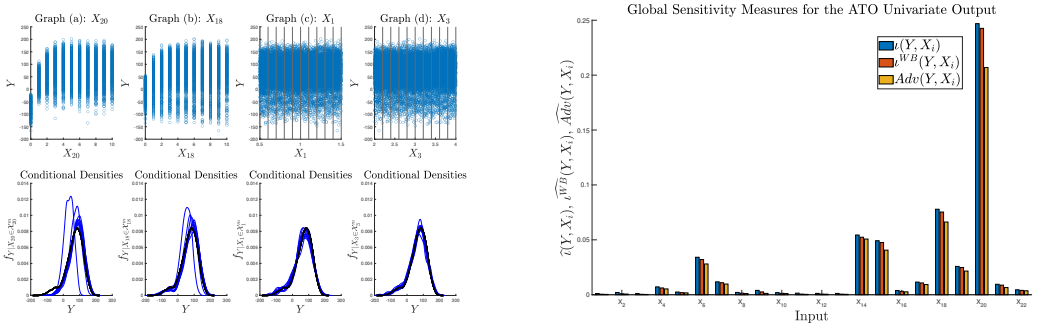
Figure 5: Estimates of $\iota(Y, X_i)$ (vertical axis) with four OT-solvers for the model in (44). Sample sizes (horizontal axis) vary from $N = 50$ to $N = 1,000$.

$\ell = 10$, they are almost close to unity. Also these results are in accordance with Theorem 19.

6. Application: The ATO Simulator

Premise: Univariate or Multivariate? The approach we have outlined applies to both a univariate and a multivariate output setting. While the distinction between the two settings may not always be clear-cut, it is important to ensure consistency between the quantity of interest and the decision-making problem at hand. For instance, if the model has been constructed to forecast the n_Y attributes of multicriteria utility function $U(Y) = u(Y_1, Y_2, \ldots, Y_{n_Y})$ that captures the decision-maker's preferences, then $U(Y)$ becomes the univariate output of interest. In this scenario, treating the outputs as a vector would be inconsistent with the problem setup. Nevertheless, there are situations where the output is inherently multivariate, as in the case of a vector in which each element is a different quantity, or a spatially or temporally distributed output, an image, or, in general, a list of outputs that cannot reasonably be incorporated into a multi-criteria utility function. In these cases, employing a multivariate sensitivity approach does not conflict with the overall decision-making problem setup. Additionally, the two approaches are not mutually exclusive, as in the case in which the analyst is interested in examining also the sensitivity of a specific output Y_i, for instance to verify the model's response vis-á-vis an underlying theory or business intuition.

The ATO Simulator. The theory of assemble-to-order (ATO) systems originates with Glasserman and Wang (1998), who use stochastic simulations to analyze the trade-off between stock reserve costs and service levels. In Hong and Nelson (2006), items (parts) are ordered and stocked, and products are then assembled based on the available items. Some of the items are key parts, without which the product cannot be assembled. Orders arrive



(a) First Row: scatterplot partitioning. Second Row: corresponding conditional densities.

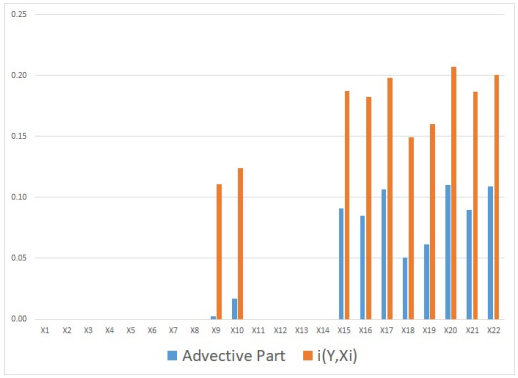
(b) Estimates of $\iota(Y, X_i)$, $\iota^{WB}(Y, X_i)$, $\mathrm{Adv}(Y, X_i)$ for the ATO final profit.

Figure 6: Global sensitivity analysis of the final profit of the ATO simulator.

stochastically, and if a key part is missing, a replenishment is ordered, and the product is completed after a random time interval. Xie et al. (2012) provide a software implementation of the model in Hong and Nelson (2006). The code, available at the SimOpt website, has been extensively used in simulation studies — see Binois et al. (2018) and Binois and Gramacy (2021) for reviews. We rely on the publicly available MatLab implementation, with the configuration of Hong and Nelson (2006): The inventory consists of eight items, from which five products are assembled. Simulator inputs are the prices for the eight items $(X_1, X_2, \ldots, X_8)$, the order arrival rates for the five products $(X_9, X_{10}, \ldots, X_{13})$, the items holding cost (X_{14}), the target levels for the eight inventory items $(X_{15}, X_{16}, \ldots, X_{22})$. The input ranges, as well as the assigned distributions, are the same as in Hong and Nelson (2006). We set the run length at 100 hours and consider as model outputs the final profit (univariate) and the final inventory (multivariate). While these outputs could be combined to form a nine-variate vector, we prefer to keep them separate because they have different managerial interpretations: Final profit quantifies economic performance, while final inventory informs us about both the quantities of items to be stocked and the overall storage capacity that allow the manager to achieve such economic performance.

We generate an input sample of size $N = 2^{13}$ with 5 replicates, for a total of 40960 model evaluations. The computing time for the model evaluations is about eight hours. We use given-data estimators for all the sensitivity measures applied, fixing the number of partitions at $H = 10$. We start with final profit as a quantity of interest. Figure 6a reports results for the scatterplot partitioning and the conditional model output distributions (empirical densities are available in this case) for selected inputs. In the second row of Figure 6a, the bold line represents the marginal distribution. The estimated mean profit value is $\widehat{\mathbb{E}}[Y] \approx 68$, the 5^{th} percentile at $y_{05} \approx -42$, the 95^{th} percentile at $y_{95} \approx -42$ and an estimated standard deviation of $\widehat{\sigma}_Y \approx 54$. The non-bold (blue) densities represent the conditional densities given one of the four inputs. To illustrate, in Graph (a) we have eleven conditional densities because the target level of inventory item six (input X_{20}) is a discrete random variable with support $\{0, 1, \ldots, 10\}$. The graph directly below evidences a left shift in the conditional densities after fixing this input: fixing the target level of inventory item six to zero yields a negative profit in all simulations (see the corresponding barplot in Graph (a). The visual impression from the second row in Figure 6a is that information about the target levels of inventory items six and four (input X_{18}), respectively, has a greater impact

on the final profit distribution than information about the prices of items one and three (inputs X_1 and X_3, respectively). This qualitative intuition is confirmed quantitatively by the estimates of the OT-based sensitivity measures (Figure 6b). The barplot in Figure 6b reports the values of the triplet of indices $\widehat{\iota}(Y, X_i)$, $\widehat{\iota}^{WB}(Y, X_i)$ and $\widehat{\text{Adv}}(Y, X_i)$. We obtain $\widehat{\iota}(Y, X_i) > \widehat{\iota}^{WB}(Y, X_i) > \widehat{\text{Adv}}(Y, X_i)$ for all inputs, in accordance with the theory, with the advective (variance-based) contribution amounting at a substantial portion of the overall input importance. The most important input is the target inventory level of item six, X_{20}, followed by the target inventory level for item four, X_{18}, and the items holding cost, X_{14}. The prices of the first three items (inputs X_1, X_2 and X_3) play a minor role. We observe an overall agreement between the ranking induced by $\widehat{\text{Adv}}(Y, X_i)$ and by $\iota(Y, X_i)$, with minor differences occurring for the least relevant inputs.



(a) Estimates of $\iota^{\text{WB}}(Y, X_i)$ and $\text{Adv}(Y, X_i)$ for the ATO inventory level.

(b) Inputs ranks on profit and final inventory of the ATO simulator.

Figure 7: Results for the multivariate output of the ATO simulator

Consider now the case in which inventory is the output. We report estimates of $\iota(Y, X_i)$ and $\text{Adv}(Y, X_i)$. By Proposition 16, ranking inputs with $\text{Adv}(Y, X_i)$ is equivalent to sorting them with the generalized variance-based importance measures. To calculate $\iota(Y, X_i)$, we solve the OT-problem in each partition using the Sinkhorn algorithm. The barplot in Figure 7a reports the estimates of $\widehat{\iota}(Y, X_i)$ as light colored (orange) bars and of $\text{Adv}(Y, X_i)$ as dark colored bars (blue). The values of $\widehat{\iota}(Y, X_i)$ are close to zero for the prices of the eight inventory items, (inputs $X_1, X_2, \ldots, X_8$), the order arrival rates of products three, four and five (inputs X_{11}, X_{12} and X_{13}, respectively), as well as for the holding cost (input X_{14}). By zero-independence these values indicate that the final multivariate inventory distribution is not sensitive to these inputs: gathering information about them brings little or no value to the decision-maker. Conversely, the non-zero values of the sensitivity indices of the prices of the first two products (inputs X_9 and X_{10}), and of the target inventory levels for all items (inputs $X_{15}, X_{16}, \ldots, X_{22}$) indicate that the multivariate output distribution is sensitive to these variables. However, by max-functionality we know that, because none of their OT-based indices is close to unity, the final inventory composition is not functionally dependent on any of these inputs separately, and gaining perfect information about any of them individually is not sufficient to remove uncertainty.

Comparing the values of $\text{Adv}(Y, X_i)$ and $\iota(Y, X_i)$ shows that the advective part accounts from a minimum of 2% up to a maximum of 54% of $\widehat{\iota}(Y, X_i)$, with the highest percentage associated with the most important input, the price of item six, X_{20}. Overall, in the

multivariate case the variance-contributions amount to a much lower fraction of the inputs' importance than in the univariate (final profit) case. We also observe greater discrepancies in the rankings generated by $\mathrm{Adv}(Y, X_i)$ and $\iota(Y, X_i)$ compared to those in the univariate case.

Figure 7b compares the input ranks for final profit (dashed blue line) against their ranks for final inventory (solid red line). Although the most important input, the target level of inventory item six (X_{20}), is the same for both the univariate and the multivariate output, the remaining ranks differ notably. The Spearman correlation coefficient between the two ranking is 0.54. For the final profit, the target levels of inventory items eight and three (X_{22} and X_{17}, respectively) rank tenth and seventh, while they rank second and third when the output is final inventory. Interestingly, the holding cost X_{14} ranks only 21^{st} for final inventory but third for the final profit. This suggests that uncertainty in inventory cost impacts profitability but not the final item allocation. The fact that the target level of inventory of item six, X_{20}, is ranked as the most important input for both final inventory and profit can be explained by its association with the only item that is key for all five products.

7. Conclusions

This work proposes an approach to global sensitivity analysis based on the theory of optimal transport that yields an elegant solution to the problem of determining key drivers of variability for multivariate responses. We have seen that the resulting sensitivity indices possess relevant properties such as zero-independence, max-functionality and monotonicity.

We have focused on the squared Wasserstein-2 distance and studied the case when marginal and conditional distributions are elliptical with the same characteristic generator. The closed-form expressions yield a global sensitivity measure based on the Wasserstein-Bures distance that extends Wagner's variance-based sensitivity measures and, under input independence, coincides with the generalized sensitivity indices of Lamboni et al. (2011) and Gamboa et al. (2014).

Recent literature is paying attention to the entropic formulation of OT problems, because it grants a computationally advantageous algorithmic implementation and its solution can be used as an approximation of the classical OT formulation. We have studied entropic-OT-based indices and we have seen that they possess the same properties as the ones based on the classical-OT formulation, with the difference that entropic OT-based indices are minimal but not necessarily zero in the case of statistical independence. However, for large values of the regularization parameter entropic OT-based indices tend to one and blur the inputs' importance, assigning the same value, independently of the underlying input-output mapping.

We have discussed a given-data estimation design that is linear in the sample size and does not depend on the problem dimensionality. We have proven that the given-data estimators are asymptotically unbiased and converge from below, both for the classical and entropic cases. We have conducted a series of numerical experiments aimed at testing the theoretical findings and at determining computational times using OT solvers that rely on alternative rationales, such as the network simplex, Cuturi's Sinkhorn, Puccetti's reordering, and the Wasserstein-Bures approximation. Results show that most estimators

yield accurate estimates within a short amount of time.

We have challenged the estimation strategy through several experiments, including the well-known ATO simulator, and an analytical test case with large number of inputs. Our results showed that the new indices provide valuable insights for both univariate and multivariate output settings, producing robust input rankings in both cases. Additionally, by calculating both OT-based and Wasserstein-Bures-indices simultaneously, analysts can avoid false negatives and determine the extent to which an input's importance is due to its impact on the first moments (advective part), on the second moments (diffusive part) or on other statistical properties of the output.

ore broadly, this work contributes to connecting sensitivity analysis with optimal transport, a branch of optimization receiving increasing attention in machine learning and statistics. Research in machine learning is rapidly progressing toward the development of efficient optimal transport solution algorithms (see Chen et al. (2022)). A natural continuation of this work is to invest in estimators based on these recent strategies. A further research avenue consists of applying theoretical findings on algorithmic convergence (such as those in Chizat et al. (2020) on the Sinkhorn algorithm) to obtain confidence bounds at finite sample sizes. Additionally, kernel-based indicators are currently being researched for use in multivariate problems. Comparing OT-based and kernel-based indices from both theoretical and numerical perspectives is a further research avenue following the present work.

Acknowledgements

The authors thank the editor the Associate Editor, and two anonymous reviewers for their constructive feedback and precious comments. We also wish to thank Barry L. Nelson and the participants of the I-Sim 2021 Symposium for their perceptive questions and comments on an earlier version of this work. We also thank Jason Altschuler and Promit Ghosal for their inputs and observations. A. Figalli acknowledges the support of the ERC Grant No.721675 "Regularity and Stability in Partial Differential Equations (RSPDE)" and of the Lagrange Mathematics and Computation Research Center. All files related to the experiments performed in this work can be retrieved at `https://github.com/emanueleborgonovo/OTsensitivity`.

References

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

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

Barr J, Rabitz H (2022) A generalized kernel method for global sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification* 10(1):27–54.

Barton RR (2016) Tutorial: Simulation metamodeling. *Proceedings - Winter Simulation Conference*, volume 2016-Febru, 1765–1779.

Baucells M, Borgonovo E (2013) Invariant Probabilistic Sensitivity Analysis. *Management Science* 59(11):2536–2549.

Berger D, Herkenhoff K, Huang C, Mongey S (2022) Testing and reopening in an SEIR model. *Review of Economic Dynamics* 43:1–21.

Bertsimas D, Shtern S, Sturt B (2022a) A Data-Driven Approach to Multistage Stochastic Linear Optimization. *Management Science* 69(1):51–74.

Bertsimas D, Shtern S, Sturt B (2022b) Technical Note-Two-Stage Sample Robust Optimization. *Operations Research* 70(1):624–640.

Binois M, Gramacy RB (2021) HetGP: Heteroskedastic Gaussian Process Modeling and Sequential Design in R. *Journal of Statistical Software* 98(13):1–44.

Binois M, Gramacy RB, Ludkovski M (2018) Practical Heteroscedastic Gaussian Process Modeling for Large Simulation Experiments. *Journal of Computational and Graphical Statistics* Forthc.(0):1–14, URL `http://dx.doi.org/10.1080/10618600.2018.1458625`.

Blanchet J, Kang Y (2021) Sample out-of-sample inference based on Wasserstein distance. *Operations Research* 69(3):985–1013.

Borgonovo E, Tarantola S, Plischke E, Morris MD (2014) Transformations and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society, Series B* 76:925–947.

Broadie M, Du Y, Moallemi CC (2011) Efficient risk estimation via nested sequential simulation. *Management Science* 57(6):1172–1194.

Broadie M, Du Y, Moallemi CC (2015) Risk estimation via regression. *Operations Research* 63(5):1077–1097.

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

Cambanis S, Simons G, Stout W (1976) Inequalities for Ek(X, Y) when the marginals are fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 36(4):285–294.

Carlsson JG, Behroozi M, Mihic K (2018) Wasserstein distance and the distributionally robust TSP. *Operations Research* 66(6):1603–1624.

Chatterjee S (2021) A New Coefficient of Correlation. *Journal of the American Statistical Association* 116(536):2009–2022.

Chen L, Kyng R, Liu YP, Peng R, Gutenberg MP, Sachdeva S (2022) Maximum Flow and Minimum-Cost Flow in Almost-Linear Time. *arXiv:2203.00671v2* April(1):1–112.

Chen Y, Georgiou TT, Pavon M (2021) Stochastic Control Liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger Bridge. *SIAM Review* 63(2):249–313.

Chen Z, Sim M, Xiong P (2020) Robust stochastic optimization made easy with RSOME. *Management Science* 66(8):3329–3339.

Chizat L, Roussillon P, Léger F, Vialard FX, Peyré G (2020) Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. *Advances in Neural Information Processing Systems*, volume 2020-Decem.

Cuturi M (2013) Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems* 26:2292–2300.

da Veiga S (2021) Kernel-based ANOVA decomposition and Shapley effects – Application to global sensitivity analysis. preprint, Hyper articles en ligne, hal-03108628.

Deb N, Ghosal P, Sen B (2020) Measuring Association on Topological Spaces Using Kernels and Geometric Graphs. *arXiv:* 2010(.01768v2):1–66.

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

Du Z, Wang L, Bai Y, Wang X, Pandey A, Fitzpatrick MC, Chinazzi M, Pastore y Piontti A, Hupert N, Lachmann M, Vespignani A, Galvani AP, Cowling BJ, Meyers LA (2022) Cost-effective proactive testing strategies during COVID-19 mass vaccination: A modelling study. *The Lancet Regional Health - Americas* 8:1–10.

Felli JC, Hazen GB (1998) Sensitivity analysis and the expected value of perfect information. *Medical Decision Making* 18:95–109.

Figalli A, Glaudo F (2021) *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows* (European Mathematical Society, Zurich), ISBN 978-3-98547-010-5.

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

Fraiman R, Gamboa F, Moreno L (2020) Sensitivity Indices for Output on a Riemannian Manifold. *International Journal for Uncertainty Quantification* 10(4):297–314.

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

Gamboa F, Janon A, Klein T, Lagnoux A, Prieur C (2016) Statistical inference for Sobol pick-freeze Monte Carlo method. *Statistics* 50(4):881–902.

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

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

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

Genevay A, Peyré G, Cuturi M (2018) Learning generative models with Sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, 1608–1617.

Givens CR, Shortt RM (1984) A Class of Wasserstein Metrics for Probability Distributions. *Michigan Mathematical Journal* 31:231–240.

Glasserman P, Wang Y (1998) Leadtime-Inventory Trade-Offs in Assemble-to-Order Systems. *Operations Research* 46(6):858–871.

Gordy MB, Juneja S (2010) Nested simulation in portfolio risk measurement. *Management Science* 56(10):1833–1848.

Hanasusanto GA, Kuhn D (2018) Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research* 66(3):849–869.

Hillier FS, Lieberman G (2012) *Introduction to Operations Research* (New York, NY: McGraw-Hill), 7th edition, ISBN 978-0072535105.

Hitchcock FL (1940) The Distribution of a Product from Several Sources to Numerous Localities. *MIT Journal of Mathematics and Physics* 20:224–230.

Hong LJ, Juneja S, Liu G (2017) Kernel smoothing for nested estimation with application to portfolio risk measurement. *Operations Research* 65(3):657–673.

Hong LJ, Nelson BL (2006) Discrete Optimization via Simulation Using COMPASS. *Operations Research* 54(1):115–129.

Hu Z, Cao J, Hong LJ (2012) Robust Simulation of Global Warming Policies Using the DICE Model. *Management Science* 58(12):2190–2206.

Ishigami T, Homma T (1990) An Importance Quantification Technique in Uncertainty Analysis for Computer Models. *Uncertainty modelling and analysis*.

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

Kleijnen JPC (2010) Sensitivity analysis of simulation models. *Encyclopedia of Operations Research and Management Science*, 1–10 (Wiley).

Kleijnen JPC, Helton JC (1999) Statistical analyses of scatterplots to identify important factors in large-scale simulations, 2: Robustness of techniques. *Reliability Engineering&System Safety* 65(2):187–197.

Knight PA (2008) The Sinkhorn-Knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications* 30(1):261–275.

Kuhn HW (1955) The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2(1-2):83–97.

Kuhn HW (1956) Variants of the Hungarian Method for Assignment Problems. *Naval Research Logistics Quarterly* 3(4):253–258.

Lamboni M, Monod H, Makowski D (2011) Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering&System Safety* 96(4):450–459.

Landsman ZM, Valdez EA (2003) Tail Conditional Expectations for Elliptical Distributions. *North American Actuarial Journal* 7(4):55–71.

Luenberger DG, Ye Y (2016) *Linear and Nonlinear Programming* (Cham: Springer), fourth edition, ISBN 978-3-319-18841-6.

Luo F, Mehrotra S (2019) Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models. *European Journal of Operational Research* 278(1):20–35.

Marrel A, Saint-Geours N, De Lozzo M (2017) Sensitivity analysis of spatial and/or temporal phenomena. *Handbook of Uncertainty Quantification*, 1327–1357 (Cham: Springer Verlag).

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1-2):1–52.

Móri TF, Székely GJ (2019) Four simple axioms of dependence measures. *Metrika* 82(1):1–16.

Nguyen VA, Kuhn D, Esfahani PM (2022) Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator. *Operations Research* 70(1):490–515.

Nordhaus W (2017) Integrated Assessment Models of Climate Change. *NBER Reporter* 3:1–4, URL `https://www.nber.org/reporter/2017number3/integrated-assessment-models-climate-change`.

Owen AB (2014) Sobol' Indices and Shapley Value. *SIAM/ASA Journal on Uncertainty Quantification* 2(1):245–251.

Pearson K (1905) *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* (London: Dulau & Co.).

Peyré G, Cuturi M (2019) Computational optimal transport. *Foundations and Trends in Machine Learning* 11(5–6):355–607.

Puccetti G (2017) 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.

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

Razavi S, Jakeman A, Saltelli A, Prieur C, Iooss B, Borgonovo E, Plischke E, Lo Piano S, Iwanaga T, Becker W, Tarantola S, Guillaume JHA, Jakeman J, Gupta H, Melillo N, Rabitti G, Chabridon V, Duan Q, Sun X, Smith S, Sheikholeslami R, Hosseini N, Asadzadeh M, Puy A, Kucherenko S, Maier HR (2021) The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling and Software* 137(104954):1–22.

Rényi A (1959) On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae* 10:441–451.

Saltelli A (2002) Making Best Use of Model Valuations to Compute Sensitivity Indices. *Computer Physics Communications* 145:280–297.

Saltelli A, Bammer G, Bruno I, Charters E, Di Fiore M, Didier E, Espeland WN, Kay J, Lo Piano S, May D, Pielke RJ, Portaluri T, Porter TM, Puy A, Rafols I, Ravetz JR, Reinert E, Sarewitz D, Start PB, Stirling A, van der Sluijs JP, Vineis P (2020) Five Ways to Ensure that Models Serve Society: a Manifesto. *Nature* 582:482–484.

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

Schwartz L (1973) Surmartingales régulières à valeurs mesures et désintégrations régulières d'une mesure. *J. Analyse Math.* 26:1–168.

Sobol' IM (1993) Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling & Computational Experiments* 1:407–414.

Strong M, Oakley JE (2013) An efficient method for computing single-parameter partial expected value of perfect information. *Medical Decision Making* 33(6):755–766.

Subramanyam A, Mufalli F, Laínez-Aguirre JM, Pinto JM, Gounaris CE (2021) Robust multiperiod vehicle routing under customer order uncertainty. *Operations Research* 69(1):30–60.

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

Villani C (2008) *Optimal Transport: Old and New* (Berlin: Springer Verlag).

Wagner HM (1995) Global Sensitivity Analysis. *Operations Research* 43,6:948–969.

Wang Z, You K, Song S, Zhang Y (2020) Wasserstein distributionally robust shortest path problem. *European Journal of Operational Research* 284(1):31–43.

Wiesel J (2022) Measuring Association with Wasserstein Distances. *Bernoulli* 28(4):2816–2832.

Xie J, Frazier P, Chick S (2012) Assemble to Order Simulator. *SimOptWebsite* .

Zhang Y, Zhang Z, Lim A, Sim M (2021) Robust data-driven vehicle routing with time windows. *Operations Research* 69(2):469–485.

A. Proofs

Proof. Calculations for Remark 2. Let Y be absolutely continuous. Let (Y, k) be a metric space with the discrete metric. Also, let $\mathbb{P}$, $\mathbb{Q}$ be two probability measures on $(\Omega, \mathcal{B})$, 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

$$W(\mathbb{P}, \mathbb{Q}) = \sup_{B \in \mathcal{B}(\Omega)} |\mathbb{P}(B) - \mathbb{Q}(B)| = \frac{1}{2} \int_{\mathbb{R}} |f_P(y) - f_Q(y)| dy. \tag{54}$$

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

Proof. Proof of Proposition 3. Positivity: $\mathbb{E}[K(\nu, \nu_x)] \geq 0$, follows immediately from the fact that $K(\nu, \nu') \geq 0$ for any $\nu, \nu' \in \mathscr{P}(\mathsf{Y})$. Zero: independence of Y and X is equivalent to $\nu = \nu_x$ for μ-a.e. $x \in \mathsf{X}$. Then, under independence it is $\mathsf{k}(y, y') = 0$ almost everywhere in X, which leads to $\xi^K(Y, X) = 0$. Conversely, if $K(\nu, \nu') = 0$ for some $\pi^* \in \Pi(\nu, \nu')$, then it must be $\mathcal{K}(\pi^*) = \int \mathsf{k}(y, y') \, \mathrm{d}\pi^*(y, y') = 0$. Because the integrand is non-negative, it must hold that $\mathsf{k}(y, y') = 0$ on a set of π^*-probability 1. Then, because $\mathsf{k}(y, y') = 0$ implies $y = y'$, we have that $\nu_x = \nu$ almost everywhere. □

Proof. Proof of Lemma 4. We argue by contradiction and we assume that there exists $\nu \in \mathscr{P}$, $y_1, y_2 \in \mathrm{supp}(\nu)$ with $y_1 \neq y_2$, and $t \in (0,1)$ such that

$$K\Big(\nu, (1-t)\delta_{y_1} + t\delta_{y_2}\Big) = (1-t)K(\nu, \delta_{y_1}) + tK(\nu, \delta_{y_2}) < \infty. \tag{55}$$

We set $\nu' := (1-t)\delta_{y_1} + t\delta_{y_2}$ and $\pi := \nu \times \nu'$. We then have:

$$\begin{aligned}
\mathcal{K}(\pi) &= \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')d\pi(y,y') = \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')d\nu(y)d\nu'(y') \\
&= \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')d\nu(y)\left((1-t)\delta_{y_1}(y') + t\delta_{y_2}(y')\right)dy' = (1-t)\int_{\mathsf{Y}} \mathsf{k}(y,y_1)d\nu(y) + t\int_{\mathsf{Y}} \mathsf{k}(y,y_2)d\nu(y) \\
&= (1-t)K(\nu,\delta_{y_1}) + tK(\nu,\delta_{y_2}) = K(\nu,\nu'),
\end{aligned}$$

so that π is an optimal coupling between ν and ν' for the cost k. Then, by the properties of optimal plans and the continuity of k, $\mathrm{supp}(\pi) = \mathrm{supp}(\nu) \times \{y_1, y_2\}$ is k-cyclically monotone (see (Villani 2008, Ch. 5)). Since $\{y_1, y_2\} \subset \mathrm{supp}(\nu)$ the pairs (y_1, y_2) and (y_2, y_1) belong to $\mathrm{supp}(\pi)$; however

$$\mathsf{k}(y_1, y_2) + \mathsf{k}(y_2, y_1) > 0 = \mathsf{k}(y_1, y_1) + \mathsf{k}(y_2, y_2), \tag{56}$$

which shows that there is a cycle that improves the integral cost, contradicting the optimality of π. □

Proof. Proof of Theorem 5. For the inequality, it is sufficient to observe that from the theory of optimal transport

$$K(\nu,\nu') \leq \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')\,\mathrm{d}\nu(y)\,\mathrm{d}\nu'(y') = \mathcal{K}(\nu \times \nu'). \tag{57}$$

In our context, then we have

$$K(\nu,\nu_x) \leq \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')\,\mathrm{d}\nu(y)\,\mathrm{d}\nu_x(y') = \mathcal{K}(\nu \times \nu_x). \tag{58}$$

Taking the integral with respect to μ we have

$$\xi^K(Y,X) = \int_{\mathsf{X}} K(\nu,\nu_x)\,\mathrm{d}\mu(x) \leq \int_{\mathsf{X}} \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')\,\mathrm{d}\nu(y)\,\mathrm{d}\nu_x(y')\,\mathrm{d}\mu(x) \tag{59}$$

$$\leq \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')\,\mathrm{d}\nu(y) \int_{\mathsf{X}} \mathrm{d}\nu_x(y')d\mu(x) = \iint_{\mathsf{Y}^2} \mathsf{k}(y,y')\,\mathrm{d}\nu(y)\,\mathrm{d}\nu(y') = \mathbb{M}^K[Y], \tag{60}$$

where $\mathbb{M}^K[Y]$ is in (18). Then, in the case of functional dependence, we have $Y = f(X)$ for some $\mathcal{F}$-measurable map $f : \mathsf{X} \to \mathsf{Y}$, $\nu = f_\sharp\mu$, and $\nu_x^{\mathcal{F}} = \delta_{f(x)}$; since

$$K(\nu, \delta_{f(x)}) = \int_{\mathsf{Y}} \mathsf{k}(y, f(x))\mathrm{d}\nu(y),$$

we obtain

$$\begin{aligned}
\xi^K(Y,X|\mathcal{F}) &= \int_{\mathsf{X}} K(\nu,\nu_x^{\mathcal{F}})\,\mathrm{d}\mu(x) = \int_{\mathsf{X}} K(\nu,\delta_{f(x)})\,\mathrm{d}\mu(x) = \int_{\mathsf{X}} \Big(\int_{\mathsf{Y}} \mathsf{k}(y,f(x))\mathrm{d}\nu(y)\Big)\,\mathrm{d}\mu(x) \\
&= \int_{\mathsf{Y}} \Big(\int_{\mathsf{Y}} \mathsf{k}(y,y')\,\mathrm{d}\nu(y)\Big)\,\mathrm{d}\nu(y') = \mathbb{M}^K[Y].
\end{aligned}$$

Thus, we have proven that the maximum is reached in the case of a functional dependence. For the converse implication, let us suppose that $\xi^K(Y, X|\mathcal{F}) = \mathbb{M}^K[Y] = \int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu(y)$, that is,

$$\int_{\mathsf{X}} K(\nu, \nu_x)\,\mathrm{d}\mu(x) = \mathbb{M}^K[Y] = \int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu(y), \tag{61}$$

and let us show that $\nu_x^{\mathcal{F}}$ is concentrated into a Dirac mass for μ-a.e. $x \in \mathsf{X}$. By disintegration we can write the last term as

$$\int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu(y) = \int_{\mathsf{X}} \Big(\int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu_x(y) \Big)\,\mathrm{d}\mu(x). \tag{62}$$

Since for every $x \in \mathsf{X}$ we have

$$K(\nu, \nu_x^{\mathcal{F}}) \le \int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu_x^{\mathcal{F}}(y),$$

comparing (62) and (61) we get

$$K(\nu, \nu_x^{\mathcal{F}}) = \int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu_x^{\mathcal{F}}(y) \quad \text{for } \mu\text{-a.e. } x \in \mathsf{X}. \tag{63}$$

We want to show that equality in (63) can hold only if $\nu_x^{\mathcal{F}}$ is a Dirac measure using Lemma 4 (with $t = 1/2$) and the fact that the support of $\nu_x^{\mathcal{F}}$ is contained in the support of ν. We argue by contradiction: if for some $x \in \mathsf{X}$ the measure $\nu_x^{\mathcal{F}}$ is not concentrated at a unique point, then we can write $\nu_x^{\mathcal{F}} = \frac{1}{2}\nu_x^1 + \frac{1}{2}\nu_x^2$ for some probability measures $\nu_x^1, \nu_x^2 \in \mathscr{P}(\mathsf{Y})$, $\nu_x^1 \neq \nu_x^2$, which clearly have support contained in $S := \mathrm{supp}(\nu)$. Defining $\sigma := \nu_x^1 \times \nu_x^2$ we can represent $\nu_x^{\mathcal{F}}$ as

$$\nu_x^{\mathcal{F}} = \int_{S\times S} (\frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2})\,\mathrm{d}\sigma(y_1, y_2). \tag{64}$$

Since $K(\nu, \cdot)$ is convex and l.s.c. and $\frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$ are supported in S for every $(y_1, y_2) \in S \times S$, Jensen inequality yields

$$K(\nu, \nu_x^{\mathcal{F}}) \le \int_{S\times S} K\big(\nu, \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}\big)\,\mathrm{d}\sigma(y_1, y_2). \tag{65}$$

Since σ is not concentrated on the diagonal of $\mathsf{Y} \times \mathsf{Y}$, the strict convexity of ζ on Dirac masses yields

$$\begin{aligned}
\int_{S\times S} K\big(\nu, \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}\big)\,\mathrm{d}\sigma(y_1, y_2) &< \int_{S\times S} \Big(\frac{1}{2}K\big(\nu, \delta_{y_1}\big) + \frac{1}{2}K\big(\nu, \delta_{y_2}\big)\Big)\,\mathrm{d}\sigma \\
&= \frac{1}{2}\int_{\mathsf{Y}} K(\nu, \delta_{y_1})\,\mathrm{d}\nu_x^1(y_1) + \frac{1}{2}\int_{\mathsf{Y}} K(\nu, \delta_{y_2})\,\mathrm{d}\nu_x^2(y_2) \\
&= \int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu_x^{\mathcal{F}}(y) < \infty.
\end{aligned}$$

Combining the last strict inequality with (65) we thus obtain

$$K(\nu, \nu_x^{\mathcal{F}}) < \int_{\mathsf{Y}} K(\nu, \delta_y)\,\mathrm{d}\nu_x^{\mathcal{F}}(y),$$

a contradiction with (63). We eventually obtain that $\nu_x^{\mathcal{F}} = \delta_{f(x)}$ for some Borel map $f : \mathsf{X} \to \mathsf{Y}$; since $\nu_x^{\mathcal{F}}$ is $\mathcal{F}$-measurable, we conclude that f is $\mathcal{F}$-measurable as well. □

Proof. Proof of Theorem 8.

Let us call $\mu^{\mathcal{F}}_{x'}$ be the disintegration of μ w.r.t. $\mathcal{F}$ and let $\nu^{\mathcal{F}}_{x'}$ the disintegration of π w.r.t. X and the σ-algebra $\mathcal{F}$. By (Schwartz 1973, Theorem 3,1) we have

$$\nu^{\mathcal{F}}_{x'} = \int_{\mathsf{X}} \nu_x \, \mathrm{d}\mu^{\mathcal{F}}_{x'}(x) \quad \text{for } \mu\text{-a.e. } x' \in \mathsf{X}. \tag{66}$$

One can also directly check the validity of (66): denoting by $(\vartheta_{x'})_{x' \in X}$ the measure given by the right-hand side of (66), we have for every Borel set $B \in \mathscr{B}(\mathsf{Y})$

$$\vartheta_{x'}(B) = \int_{\mathsf{X}} \nu_x(B) \, \mathrm{d}\mu^{\mathcal{F}}_{x'}(x) = \mathbb{E}_\mu[\nu_\cdot(B)|\mathcal{F}] \tag{67}$$

by definition of the conditional measure $\mu^{\mathcal{F}}_{x'}$; thus $x' \mapsto \vartheta_{x'}(B)$ is $\mathcal{F}$-measurable for every Borel set $B \in \mathscr{B}(\mathsf{Y})$. On the other hand, integrating (67) w.r.t. μ on a arbitrary set $A \in \mathcal{F}$ we get

$$\int_A \vartheta_{x'}(B) \, \mathrm{d}\mu(x') = \int_A \Big(\int_{\mathsf{X}} \nu_x(B) \, \mathrm{d}\mu^{\mathcal{F}}_{x'}(x) \Big) \, \mathrm{d}\mu(x') = \int_A \nu_x(B) \, \mathrm{d}\mu(x) = \pi(A \times B)$$

by definition of $\mu^{\mathcal{F}}_{x'}$, so that $\vartheta_{x'}(B)$ should coincide with $\nu^{\mathcal{F}}_{x'}$ for μ-a.e. $x' \in X$ by the (essential) uniqueness of the regular conditional laws.

Using the convexity of $K(\nu, \cdot)$ and Jensen inequality we have

$$\begin{aligned}
\xi^K(Y, X|\mathcal{F}) &= \int_{\mathsf{X}} K(\nu, \nu^{\mathcal{F}}_{x'}) \, \mathrm{d}\mu(x') = \int_{\mathsf{X}} K\Big(\nu, \int_{\mathsf{X}} \nu_x \, \mathrm{d}\mu^{\mathcal{F}}_{x'}(x)\Big) \, \mathrm{d}\mu(x') \\
&\le \int_{\mathsf{X}} \Big(\int_{\mathsf{X}} K(\nu, \nu_x) \, \mathrm{d}\mu^{\mathcal{F}}_{x'}(x)\Big) \, \mathrm{d}\mu(x') = \int_{\mathsf{X}} K(\nu, \nu_x) \, \mathrm{d}\mu(x) = \xi^K(Y, X),
\end{aligned}$$

which yields (23).

In the case $\mathcal{F} = \sigma(g)$, we can set $\sigma := g_\sharp \mu = U_\sharp \mathbb{P} \in \mathscr{P}(\mathsf{U})$, and we can decompose μ as $\mu = \int_{\mathsf{U}} \mu_u \, \mathrm{d}\sigma(u)$. If ν'_u is the disintegration of π given g, we have as in (67)

$$\nu'_u = \int_{\mathsf{X}} \nu_x \, \mathrm{d}\mu_u(x) \tag{68}$$

and the previous estimate reads as

$$\begin{aligned}
\xi^K(Y, U) &= \int_{\mathsf{U}} K(\nu, \nu'_u) \, \mathrm{d}\sigma(u) = \int_{\mathsf{U}} K\Big(\nu, \int_{\mathsf{X}} \nu_x \, \mathrm{d}\mu_u(x)\Big) \, \mathrm{d}\sigma(z) \\
&\le \int_{\mathsf{U}} \Big(\int_{\mathsf{X}} K(\nu, \nu_x) \, \mathrm{d}\mu_u(x)\Big) \, \mathrm{d}\sigma(u) = \int_{\mathsf{X}} K(\nu, \nu_x) \, \mathrm{d}\mu(x) = \xi^K(Y, X).
\end{aligned}$$

Equality holds if for σ-a.e. $u \in \mathsf{U}$

$$K\Big(\nu, \int_{\mathsf{X}} \nu_x \, \mathrm{d}\mu_u(x)\Big) = \int_{\mathsf{X}} K(\nu, \nu_x) \, \mathrm{d}\mu_u(x)$$

and this happens, e.g., if g is injective, so that μ_u is a Dirac mass concentrated in $g^{-1}(u)$ for σ-a.e. u and $\nu_x = \nu'_{g(x)}$.

□

Proof. Proof of Theorem 9. It is clear that $\limsup_{n\to\infty} \xi^K(Y,\, X|\mathcal{F}_n) \le \xi^K(Y, X|\mathcal{F})$; we thus have to show that $\liminf_{n\to\infty} \xi^K(Y,\, X|\mathcal{F}_n) \ge \xi^K(Y, X|\mathcal{F})$.

Let us consider the family of σ-algebras in $\mathsf{X}\times\mathsf{Y}$ $\mathcal{F}'_n := \{A\times\mathsf{Y} : A\in\mathcal{F}_n\}$ and $\mathcal{F}'_\infty := \{A\times\mathsf{Y} : A\in\mathcal{F}\}$, Let $(\nu_x^\infty)_{x\in\mathsf{X}}$ (resp. $(\nu_x^n)_{x\in\mathsf{X}}$) be the conditional measures in $\mathscr{P}(\mathsf{Y})$ of π with respect to $\mathcal{F}'_\infty$ (resp. $\mathcal{F}'_n$).

For every $f\in L^1(\mathsf{X}\times\mathsf{Y},\pi)$, the conditional expectations $f_n := \mathbb{E}_\pi[f,\mathcal{F}'_n]$ (resp. $f_\infty := \mathbb{E}_\pi[f,\mathcal{F}'_\infty]$) can be identified with the $\mathcal{F}_n$ (resp. $\mathcal{F}$) measurable functions given by

$$f_n(x) = \mathbb{E}^{\nu_x^n}[f] = \int_{\mathsf{Y}} f(x,y)\,\mathrm{d}\nu_x^n(y), \quad f_\infty(x) = \mathbb{E}^{\nu_x^\infty}[f] = \int_{\mathsf{Y}} f(x,y)\,\mathrm{d}\nu_x^\infty(y) \quad \text{for } \mu\text{-a.e. } x\in X,$$

and we have $f_n\to f_\infty$ π (and thus μ) almost everywhere.

Choosing $f(x,y)=\varphi(y)$ with $\varphi\in \mathrm{C}_b(\mathsf{Y})$ and observing that

$$\mathbb{E}^{\nu_x^n}[f] = \int_{\mathsf{Y}} \varphi\,\mathrm{d}\nu_x^n(y),$$

we get

$$\lim_{n\to\infty}\int_{\mathsf{Y}} \varphi\,\mathrm{d}\nu_x^n(y) = \int_{\mathsf{Y}} \varphi\,\mathrm{d}\nu_x^\infty(y),$$

for μ-a.e. $x\in X$. Choosing a countable collection of functions in $\mathrm{C}_b(\mathsf{Y})$ which determine weak convergence we conclude that $\nu_x^n\to\nu_x^\infty$ for μ-a.e. $x\in\mathsf{X}$.

Since K is lower semicontinuous w.r.t. weak convergence we have

$$\liminf_{n\to\infty} K(\nu,\nu_x^n) \ge K(\nu,\nu_x^\infty) \quad \text{for } \mu\text{-a.e. } x\in\mathsf{X}$$

and therefore Fatou's Lemma yields

$$\liminf_{n\to\infty} \xi^K(Y,X|\mathcal{F}_n)) = \liminf_{n\to\infty}\int_{\mathsf{X}} K(\nu,\nu_x^n)\,\mathrm{d}\mu(x) \ge \int_{\mathsf{X}} K(\nu,\nu_x^\infty)\,\mathrm{d}\mu(x) = \xi^K(Y,X|\mathcal{F}).$$

□

Proof. Proof of Lemma 10 We first observe that

$$K_\varepsilon(\nu,\delta_{y'}) = K(\nu,\delta_{y'}) = \int_{\mathsf{Y}} \mathsf{k}(y,y')\,\mathrm{d}\nu(y), \tag{69}$$

because in this case the unique coupling in $\Pi(\nu,\delta_{y'})$ is $\nu\times\delta_{y'}$ and $\mathrm{KL}(\pi,\nu\times\delta_{y'})=0$.

We use a similar argument by contradiction as in the case of K: let us suppose that there exists $\nu\in D(K_\varepsilon)$, $y_1,y_2\in\mathrm{supp}(\nu)$ with $y_1\neq y_2$, and $t\in(0,1)$ such that

$$K_\varepsilon\Big(\nu,(1-t)\delta_{y_1}+t\delta_{y_2}\Big) = (1-t)K_\varepsilon(\nu,\delta_{y_1}) + tK_\varepsilon(\nu,\delta_{y_2}) < \infty.$$

We set $\nu' := (1-t)\delta_{y_1}+t\delta_{y_2}$ and $\pi := \nu\times\nu'$. It is easy to check that

$$\begin{aligned}\mathcal{K}_\varepsilon(\pi) &= (1-t)\int_{\mathsf{Y}} \mathsf{k}(y,y_1)\,\mathrm{d}\nu(y) + t\int_{\mathsf{Y}} \mathsf{k}(y,y_2)\,\mathrm{d}\nu(y)\\ &= (1-t)K(\nu,\delta_{y_1}) + tK(\nu,\delta_{y_2}) = (1-t)K_\varepsilon(\nu,\delta_{y_1}) + tK_\varepsilon(\nu,\delta_{y_2}) = K_\varepsilon(\nu,\nu'),\end{aligned}$$

so that π is an optimal coupling between ν and ν' for the entropic optimal transport problem. It follows that the density $\mathrm{d}\pi/\mathrm{d}(\nu \times \nu') = 1$ should admit a k, ε cyclically invariant version: in particular

$$\mathsf{k}(z_1, y_1) + \mathsf{k}(z_2, y_2) = \mathsf{k}(z_1, y_2) + \mathsf{k}(z_2, y_1),$$

for z_1, z_2 in a dense subset of $\mathrm{supp}(\nu)$. Using the continuity of k and approximating y_i with a sequence of points z_i, we get $0 = \mathsf{k}(y_1, y_1) + \mathsf{k}(y_2, y_2) = \mathsf{k}(y_1, y_2) + \mathsf{k}(y_2, y_1)$, a contradiction. □

Proof. Proof of Theorem 11. We first observe that $K_\varepsilon(\nu, \nu')$ is lower semicontinuous, and convex in its second argument, analogously to $K(\nu, \nu')$. Equation (27) follows by integrating (69) w.r.t. ν. We can then apply, with minor modifications, the arguments used in the proofs of Theorems 5, 8, and 9 for K to complete the proof (these steps are omitted for brevity). □

Proof. Proof of Proposition 12. In general, by the convexity of $K_\varepsilon(\nu, \nu')$ in its second argument, we have

$$\xi^{K_\varepsilon}(Y, X) = \int_{\mathcal{X}} K_\varepsilon(\nu, \nu_x)\, \mathrm{d}\mu(x) \geq K_\varepsilon(\nu, \int_{\mathcal{X}} \nu_x\, \mathrm{d}\mu(x)) = K_\varepsilon(\nu, \nu),$$

where the equality $\nu = \int_{\mathcal{X}} \nu_x\, \mathrm{d}\mu(x)$ follows by disintegration. Thus, $K_\varepsilon(\nu, \nu)$ is a minimum for $\xi^{K_\varepsilon}(Y, X)$. Moreover, if Y and X are independent, then $\nu_x = \nu$ for μ-a.e. $x \in \mathsf{X}$, so that one gets the equality in the previous formula and in (29). □

Proof. Proof of Proposition 15. Inserting Equation (6) into Equation (30) we find Equation (37). Then, Theorem 2.1 in Gelbrich (1990) shows that when $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X}$ are elliptical with the same generator for all values of X, then the Wasserstein distance between $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X}$ is given by Equation (4) almost everywhere in $\mathcal{X}$, because the residual term $\Gamma(\mathbb{P}_Y, \mathbb{P}_{Y|X})$ is null almost everywhere. Then, taking the expectation, Equation (37) holds. Equation (38) is an immediate consequence of Equation (4), while Equation (39) requires the additional observation that $\mathbb{E}[\mathrm{Tr}\, \Sigma_{Y|X}] = \mathrm{Tr}\, \Sigma_Y$ and $\mathrm{Tr}\, \Sigma_Y = \mathbb{V}[Y]$ so that $\mathrm{Diff}(Y, X) = 1 - \mathbb{V}[Y]^{-1}\mathbb{E}[\mathrm{Tr}\left(\Sigma_Y^{1/2}\Sigma_{Y|X}\Sigma_Y^{1/2}\right)^{1/2}]$. □

Proof. Proof of Proposition 16. To prove (40), note that

$$\mathrm{Adv}(Y, X) = \mathbb{E}\left[\sum_{t=1}^{n_Y} (\mathbb{E}[Y_t] - \mathbb{E}[Y_t|X])^2\right] = \sum_{t=1}^{n_Y} \mathbb{E}[(\mu_{Y,t} - \mu_{Y|X_i,t})^2] = \sum_{t=1}^{n_Y} \xi_i^{V,t}. \tag{70}$$

To prove Equation (41), we report the results in (Gamboa et al. 2014, Section 3.1). First, note that one can write

$$g(X) = g_i(X_i) - g_i(X_{-i}) + g_{-i,j}(X_i, X_{-i}) - \mathbb{E}[Y], \tag{71}$$

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 $S(Y, X_i) = \dfrac{\mathrm{Tr}(\Sigma_i^Y)}{\mathrm{Tr}(\Sigma_Y)}$, where $\mathrm{Tr}(\Sigma_i^Y)$ equals the sum of the individual contributions of X_i to the variance of Y^t, that is $\mathrm{Tr}(\Sigma_i^Y) = \sum_{t=1}^{n_Y} \xi_i^{V,t}$, so that .

$= \sum_{t=1}^{n_Y} S(Y_t, X_i)$

□

Proof. Proof of Corollary 17. Let $Y = AX + b$, with Y a random vector in $\mathbb{R}^m$ on probability space $(\Omega, \mathcal{B}, \mathbb{P})$, $A = (a_{i,j})$, $i = 1, 2, \ldots, n_X$, $j = 1, 2, \ldots, n_Y$, and $b = (b_1, b_2, \ldots, b_{n_Y})$. First, let $X \in \mathbb{R}^{n_X}$, $X \sim \mathcal{EC}(m_X, \Sigma_X^*, h)$ with finite second order moment. If A is an $n_Y \times n_X$ matrix and $b \in \mathbb{R}^{n_Y}$, then $Y \sim \mathcal{EC}(Am_X + b, A\Sigma_X^* A^T, h)$ 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 $U \sim \mathcal{EC}(m_{X|X_i}, \Sigma_i^c, h)$, with Σ_i^c as given in (42). Therefore, $(Y|X_i) \sim \mathcal{EC}(Am_{Y|X_i} + b, A\Sigma_i^c A^T, h)$. Then, Y and $(Y|X_i)$ are both elliptical random variables with characteristic generator G. Then, the 2-Wasserstein metric between $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_i}$ is equal to the Wasserstein-Bures metric $WB(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})$ in (4) for every X_i. □

Proof. Proof of Theorem 19 We recall that in Equation (31) the numerator is equivalent to writing

$$\xi^{K_\varepsilon}(Y, X) = \mathbb{E}[K_\varepsilon(\mathbb{P}_Y, \mathbb{P}_{Y|X})], \tag{72}$$

with $K_\varepsilon(\nu, \nu')$ defined by Equation (8). The value $K_\varepsilon(\cdot, \cdot)$ is determined by the solution of the entropic-OT problem. The theory of the entropic-OT states that the optimal transport plan is the product plan $\nu \times \nu'$ when $\varepsilon \to \infty$. We have seen that, in correspondence of this solution, $K_\varepsilon(\nu, \nu')$ is maximal and equal to $\mathbb{M}^K[Y]$. That is $K_\varepsilon(\nu, \nu_x) \to \int\int k(y, y')\, \mathrm{d}\nu(y)\mathrm{d}\nu_x(y') = \mathcal{K}(\nu \times \nu_x)$, as $\varepsilon \to \infty$. Thus, for every value $X = x$, $\varepsilon \to \infty$, $K_\varepsilon(\mathbb{P}_Y, \mathbb{P}_{Y|X=x}) \to \mathbb{M}^K[Y]$. Then, $\mathbb{E}[K_\varepsilon(\mathbb{P}_Y, \mathbb{P}_{Y|X})] \to \mathbb{E}[\mathbb{M}^K[Y]] = \mathbb{M}^K[Y]$. Then, by Equation (31) we have $\iota_\infty(Y, X) = 1$. □

Estimation: Detailed Treatment and Proof Consider a sequence of random variables (X^N, Y^N), $N \in \mathbb{N}$, defined on $(\Omega^N, \mathfrak{B}^N, \mathbb{P}^N)$ with values in $\mathsf{X} \times \mathsf{Y}$ such that the joint laws $\pi^N = (X^N, Y^N)_\sharp \mathbb{P}^N$ is weakly converging to $\pi = (X, Y)_\sharp \mathbb{P}$. When k is not bounded, we will also assume that

$$\lim_{N \to \infty} \mathbb{E}^N[\mathsf{a}(Y^N)] = \mathbb{E}[\mathsf{a}(Y)] < \infty, \tag{73}$$

where $\mathsf{a} = \mathsf{a}_1 + \mathsf{a}_2$, with the separate cost functions bounded, so that $\mathbb{M}^K[Y] < \infty$. A typical example is given by $\Omega^N := \{1, 2, \cdots, N\}$ with the uniform measure and $X^N(n) := X_n(\omega), Y^N(n) := Y_n(\omega)$, $n = 1, \cdots, N$, are obtained by the evaluation of a sequence $(X_n, Y_n)_{n \in \mathbb{N}}$ of mutually independent random variables sharing the same joint law of (X, Y).

Here, we can distinguish the simpler case when X takes values in a finite set (or, equivalently, $\mathcal{F}$ is finite) from the general one. In the discrete case, we will assume that $\mathsf{X} = \{x^1, x^2, \cdots, x^H\}$ is a finite set, $\mathcal{F} = 2^{\mathsf{X}}$, and we consider the quantity

$$\xi_N^K := \xi^K(Y^N, X^N). \tag{74}$$

In the general case, we introduce a countable collection of measurable partitions of X, $\mathscr{X}^M = \{\mathsf{X}_h^M\}_{h=1,\cdots,H(M)}$, $M \in \mathbb{N}$, generating a corresponding family of σ-algebras $\mathcal{F}^M = \sigma(\mathscr{X}^M)$ satisfying

$$\mathcal{F}^M \subset \mathcal{F}^{M+1}, \quad \bigvee_{M \in \mathbb{N}} \mathcal{F}^M = \mathcal{F}, \quad \mu(\partial \mathsf{X}_h^M) = 0 \quad \text{for every } M \in \mathbb{N},\ 1 \le h \le H(M). \tag{75}$$

We then rewrite the estimator in (48) as

$$\widehat{\xi}^K(M,N) = \sum_{h=1}^{H(M)} \widehat{\mathbb{P}}^N[X \in \mathsf{X}_i^h] K(\widehat{\mathbb{P}}_Y^N, \widehat{\mathbb{P}}_{Y|X\in\mathsf{X}_i^h}^N), \tag{76}$$

where, $\widehat{\mathbb{P}}_Y$ and $\widehat{\mathbb{P}}_{Y|X}$ are the empirical marginal and conditional measures estimated from the available observations, and consider the quantities

$$\xi_{M,N}^K := \xi^K(Y^N, X^N|\mathcal{F}^M). \tag{77}$$

Theorem 21. *Under the above conditions:*

1. *If* X *is finite then*

$$\lim_{N\to\infty} \xi_N^K = \xi^K(X,Y). \tag{78}$$

2. *In the general case, if* (75) *holds true,*

$$\lim_{M\to\infty}\lim_{N\to\infty} \xi_{M,N}^K = \xi^K(X,Y|\mathcal{F}). \tag{79}$$

Proof. Proof of Theorem 21. *Case 1:* X *is finite* It is not restrictive to assume that $\mathbb{P}[X = x^h] > 0$, $h = 1, \cdots, H$. In this case, the OT-based sensitivity measures $\xi^K(Y,X)$, ξ_N^K can be rewritten as

$$\xi^K = \sum_{h=1}^H \mathbb{P}[X = x^h] K(\mathbb{P}_Y, \mathbb{P}_{Y|X=x^h}) = \sum_{h=1}^H \mu(\{x^h\}) K(\nu, \nu_{x^h}), \tag{80}$$

$$\xi_N^K = \sum_{h=1}^H \mathbb{P}^{\mathbb{N}}[X^N = x^h] K(\mathbb{P}_{Y^N}, \mathbb{P}_{Y^N|X^N=x^h}) = \sum_{h=1}^H \mu^N(\{x^h\}) K(\nu^N, \nu_{x^h}^N), \tag{81}$$

where $\nu_{x^h}(B) = (\mu(\{x^h\}))^{-1}\pi(\{x^h\}\times B)$ for every Borel set B of Y and a similar formula holds for $\nu_{x^h}^N$.

The weak convergence assumption on π^N and the fact that X is finite yields $\nu^N \to \nu$ and, for every h, $\mu^N(\{x^h\}) \to \mu(\{x^h\})$, $\nu_{x^h}^N \to \nu_{x^h}$ as $N \to \infty$. We can then easily pass to the limit in (81) as $N \to \infty$ obtaining (80).

Case 2 We want to prove first that

$$\lim_{N\to\infty} \xi_{M,N}^K = \xi^K(Y,X|\mathcal{F}^M). \tag{82}$$

The situation here is quite similar to the previous case: we set $\mu_{M,h} := \mu(\mathsf{X}_h^M)$, $\mu_{M,h}^N := \mu^N(\mathsf{X}_h^M)$, $\mathrm{H}(M) := \{h \in \mathbb{N} : 1 \le h \le H(M),\ \mu_{M,h} > 0\}$, $\mathrm{H}^N(M) := \{h \in \mathbb{N} : 1 \le h \le H(M),\ \mu_{M,h}^N > 0\}$ and for every Borel set B of Y $h \in \mathrm{H}(M)$ $\nu_{M,h}(B) = (\mu_{M,h})^{-1}\pi(\mathsf{X}_h^M \times B)$; similarly, if $h \in \mathrm{H}^N(M)$ $\nu_{M,h}^N(B) = (\mu_{M,h}^N)^{-1}\pi(\mathsf{X}_h^M \times B)$.

$\xi^K(Y,X)$, ξ_N^K can be rewritten as

$$\xi^K(Y,X|\mathcal{F}^M) = \sum_{h\in\mathrm{H}(M)} \mu(\mathsf{X}_h^M) K(\nu, \nu_{M,h}), \tag{83}$$

$$\xi_N^K = \sum_{h\in\mathrm{H}^N(M)} \mu^N(\mathsf{X}_h^M) K(\nu^N, \nu_{M,h}^N). \tag{84}$$

The weak convergence assumption on π^N and the fact that the boundaries of X_h^M are μ-negligible, yield $\nu^N \to \nu$, $\mu^N(\mathsf{X}_h^M) \to \mu(\mathsf{X}_h^M)$ for every $h \in \{1, \cdots, H(M)\}$, and $\nu_{M,h}^N \to \nu_{M,h}$ as $N \to \infty$ whenever $h \in \mathrm{H}(M)$. Moreover, we have

$$\lim_{N\to\infty} \int_{\mathsf{Y}} \mathsf{a}(y)\, \mathrm{d}\nu^N(y) = \int_{\mathsf{Y}} \mathsf{a}(y)\, \mathrm{d}\nu(y)$$

so that a is uniformly integrable w.r.t. ν^N, i.e.

$$\forall\, \epsilon > 0\ \exists\, L > 0 : \quad \int_{A(L)} \mathsf{a}(y)\, \mathrm{d}\nu^N(y) \le \epsilon \quad \text{for every } N \in \mathbb{N}, \tag{85}$$

where $A(L) := \{y \in \mathsf{Y} : \mathsf{a}(y) \ge L\}$.

If $h \in \mathrm{H}(M)$ we know that $(\mu_{M,h}^N)^{-1} < c$ for sufficiently large N, and we deduce that a is uniformly integrable w.r.t. $\nu_{M,h}^N$ as well, since $\nu_{M,h}^N \le c\nu^N$; we thus get

$$\lim_{N\to\infty} \int_{\mathsf{Y}} \mathsf{a}(y)\, \mathrm{d}\nu_{M,h}^N(y) = \int_{\mathsf{Y}} \mathsf{a}(y)\, \mathrm{d}\nu_{M,h}(y) \quad \text{for every } h \in \mathrm{H}(M)$$

and therefore

$$\lim_{N\to\infty} K(\nu^N, \nu_{M,h}^N) = K(\nu, \nu_{M,h}) \quad \text{for every } h \in \mathrm{H}(M). \tag{86}$$

When $\mu_{M,h} = 0$, we can use (85) to get

$$\mu_{M,h}^N K(\nu^N, \nu_{M,h}^N) \le \mu_{M,h}^N \int_{\mathsf{Y}} \mathsf{a}\, \mathrm{d}\nu^N + \int_{\mathsf{Y}\times\mathsf{X}_h^M} \mathsf{a}(y)\, \mathrm{d}\pi^N \le \mu_{M,h}^N \int_{\mathsf{Y}} \mathsf{a}\, \mathrm{d}\nu^N + \mu_{M,h}^N L + \epsilon$$

so that since $\mu_{M,h}^N \to 0$ as $N \to \infty$

$$\limsup_{N\to\infty} \mu_{M,h}^N K(\nu^N, \nu_{M,h}^N) \le \epsilon;$$

since $\epsilon > 0$ is arbitrary, we conclude that

$$\lim_{N\to\infty} \mu_{M,h}^N K(\nu^N, \nu_{M,h}^N) = 0.$$

We can then easily pass to the limit in (83) as $N \to \infty$ obtaining (84).

The last step

$$\lim_{M\to\infty} \xi^K(Y, X|\mathcal{F}^M) = \xi^K(Y, X|\mathcal{F})$$

eventually follows by Theorem 9. □

B. Counterexample for the Minimization of the Entropic Transport

Let Y be a random variable taking values in the finite set $\mathsf{Y} = \{y_1, y_2, y_3, y_4\}$ with uniform distribution and let X be valued in $\{0, 1\}$, so that $X = 0$ if $Y \in \{y_1, y_2\}$ and $X = 1$ if $Y \in \{y_3, y_4\}$. We consider a symmetric cost k on Y, vanishing on the diagonal, and such that

$$\mathsf{k}(y_1, y_2) = \mathsf{k}(y_3, y_4) = h, \quad \mathsf{k}(y_1, y_3) = \mathsf{k}(y_2, y_3) = \mathsf{k}(y_2, y_4) = \mathsf{k}(y_1, y_4) = k,$$

for parameters $h, k > 0$ satisfying

$$0 < k < \ln 2, \quad h := \ln\Big(\frac{\mathrm{e}^k}{2 - \mathrm{e}^k}\Big). \tag{87}$$

Just to fix ideas, one can always choose the parameter k so that $\mathsf{Y} \subset \mathbb{R}^2$ is given by the vertexes of the unit square $(0,0), (\sqrt{k}, \sqrt{k}), (0, \sqrt{k}), (\sqrt{k}, 0)$ and the cost k coincides with the squared Euclidean distance so that h satisfies the additional constraint $h^2 = 2k^2$. We have $\nu_0 = \nu_{X=0} = \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$, $\nu_1 = \nu_{X=1} = \frac{1}{2}(\delta_{y_3} + \delta_{y_4})$, and $\nu = \frac{1}{2}(\nu_0 + \nu_1)$, the law of Y, is the uniform distribution on Y. In this case, we have

$$\xi^{K_1}(Y, X) = K_1(\nu, \nu) = k, \tag{88}$$

but X, Y are not independent. In fact, because k takes the constant value k among pairs of points in the support of ν_0 and ν_1 it is easy to check that $\pi^{1,2} := \nu_0 \times \nu_1$ is the optimal coupling for ν_0 and ν_1 with constant optimal potentials $f = g \equiv \phi$ so that $K_1(\nu_0, \nu_1) = k = 2\phi$. We look for the optimal coupling for $K_1(\nu_0, \nu_0)$ among measures $\pi^{1,1} := \alpha(\delta_{11} + \delta_{22}) + \beta(\delta_{12} + \delta_{21})$, $\beta = 1/2 - \alpha$, imposing that the density $\mathrm{d}\pi^{1,1}/\mathrm{e}^{-\mathsf{k}}(\nu_0 \times \nu_0)$ coincides with $\mathrm{e}^{2\phi} = \mathrm{e}^k$. We thus obtain the conditions $4\alpha = \mathrm{e}^k, \quad 4\beta\mathrm{e}^h = \mathrm{e}^k$ and we get

$$\alpha = \frac{\mathrm{e}^k}{4}, \quad \beta = \frac{\mathrm{e}^{k-h}}{4},$$

which, thanks to (87), satisfy $0 < \alpha < 1/2,\ 2\alpha + 2\beta = 1$, and

$$K_1(\nu_0, \nu_0) = 2\beta h + 2\alpha \ln(4\alpha) + 2\beta \ln(4\beta) = 2\beta h + 2\alpha k + 2\beta(k - h) = k.$$

Notice that the computation of the dual problem (9) yields

$$2\phi - \frac{1}{4}\Big(2\mathrm{e}^{2\phi} + 2\mathrm{e}^{2\phi - h}\Big) + 1 = k - \frac{1}{2}\Big(\mathrm{e}^k + \mathrm{e}^{k-h}\Big) + 1 = k - 2(\alpha + \beta) + 1 = k.$$

By symmetry we also get $K_1(\nu_1, \nu_1) = k$. We now claim that $K_1(\nu, \nu_0) = k$ is attained by $\pi^1 := \frac{1}{2}(\pi^{1,1} + \pi^{1,2})$ In fact, by the convexity of the functional defining K_1 we have $K_1(\nu, \nu_0) \le \frac{1}{2}(K_1(\nu_0, \nu_0) + K_1(\nu_1, \nu_0)) = k$. On the other hand, choosing constant potentials $f = g \equiv \phi$, we obtain

$$K_1(\nu, \nu_0) \ge 2\phi + \frac{1}{2}\Big(\int \mathrm{e}^{2\phi - \mathsf{k}}\,\mathrm{d}(\nu_0 \times \nu_0) + \int \mathrm{e}^{2\phi - \mathsf{k}}\,\mathrm{d}(\nu_1 \times \nu_0)\Big) + 1 = k.$$

A similar argument shows that $K_1(\nu, \nu_1) = K_1(\nu, \nu) = k$. It is now sufficient to observe that $\xi^{K_\varepsilon}(Y, X) = \frac{1}{2}(K_1(\nu, \nu_0) + K_1(\nu, \nu_1)) = k$ as well.

C. Counterexample: Variance-based sensitivity does not posses information monotonicity

The next example illustrates a case in which we have two random variables that have the same variance-based importance measure, although one is a non-monotonic transformation of the other, and thus is less informative.

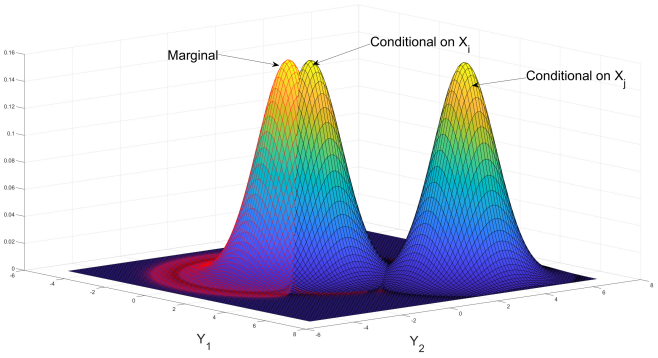


Figure 8: Example of change in distribution after receiving information on X_i or on X_j. The original (marginal) distribution $\mathbb{P}_Y$ is $\mathcal{N}(Y;[0,0],\Sigma)$, then $\mathbb{P}_{Y|X_i}$ is $\mathcal{N}(Y;[0.5,0.5],\Sigma)$ and $\mathbb{P}_{Y|X_j}$ is $\mathcal{N}(Y;[4,4],\Sigma)$, with Σ equal to the 2×2 identity matrix.

Example 22. Take $m \in [0,1]$ and let X_m be uniformly distributed in $[m-1, m+1]$. Then $\mathbb{E}[X_m] = m$ and $X_m = m + X_0$, where the law of X_0 is $F_{X_0}(x) = \frac{1}{2}\min\{\max\{x+1,2\},0\}$. Let Z be discretely uniformly distributed in $\{-1,1\}$ independently of X_m, with law $F_z(z) = \frac{1}{2}(\delta(z+1)+\delta(z-1))$ and expectation $\mathbb{E}[Z] = 0$. We write the positive and negative parts of a random variable X as $X^+ = \max\{0,X\}$ and $X^- = -\min\{0,X\}$, respectively.

We set $Y = X_m^- \cdot Z + X_m^+$. Then by independence

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[X_m^-]\cdot\mathbb{E}[Z] + \mathbb{E}[X_m^+] = \mathbb{E}[X_m^+] = \frac{1}{2}\int_{-m}^{1}(x+m)dx = \frac{1}{4}(1+m)^2,\\
\mathbb{E}[Y|X_m] &= \mathbb{E}[X_m^-|X_m]\cdot\mathbb{E}[Z] + \mathbb{E}[X_m^+|X_m] = X_m^+,\\
\mathbb{E}[Y|X_m^+] &= \mathbb{E}[X_m^-|X_m^+]\cdot\mathbb{E}[Z] + \mathbb{E}[X_m^+|X_m^+] = X_m^+.
\end{aligned}$$

Hence, we have $\mathbb{V}[\mathbb{E}[Y|X_m]] = \mathbb{V}[\mathbb{E}[Y|X_m^+]]$, so that $\xi^W(Y,X_m) = \xi^W(Y,X_m^+)$ and the variance- based sensitivity indices of both X_m and X_m^+ coincide. However, intuitively, receiving information on the positive part of X_m has a lower value than receiving information on X_m itself. In fact, some tedious calculations show that $\iota(Y,X_m) > \iota(Y,X_m^+)$.

D. More on Interpretation and Properties of the Wasserstein-Bures Semi-Metric

To provide further insights into the interpretation of Equation 16, it is helpful to consider the marginal distribution of Y as a representation of our current state of knowledge about Y (our degree-of-belief in a Bayesian framework). To illustrate, let us suppose that $\mathbb{P}_Y$ is bivariate normal with mean $m_Y = [\ 0 \quad 0\ ]$ and variance-covariance matrix $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Consider next that we receive information on two uncertain inputs X_i and X_j. Suppose that after receiving information on $X_i = x_i$, the conditional distribution $\mathbb{P}_{Y|X_i=x_i}$ is still normal but with different means $m_{Y|X_i=x_i} = [\ 0.5 \quad 0.5\ ]$ and identical Σ.

Receiving information about X_j leads (hypothetically) to a conditional normal distribution with $m_{Y|X_j=x_j} = [\ 4 \quad 4\ ]$ and identical Σ. Figure 8 visualizes these distributions. As shown in Figure 8 conditioning on $X_j = x_j$ causes our degree-of-belief to deviate more from the original marginal distribution than conditioning on $X_i = x_i$. Using (4), we find, in fact, $WB(\mathbb{P}_Y, \mathbb{P}_{Y|X_i=x_i}) = 0.5$ and $WB(\mathbb{P}_Y, \mathbb{P}_{Y|X_j=x_j}) = 32$. In a global sensitivity setting, we consider the expectation of the separation over all possible values of X_i and X_j. Then, $\iota(Y, X_i) > \iota(Y, X_j)$ means that, in expectation, learning X_j brings our degree of belief about Y further away from the current belief than learning X_j. How far we move is quantified by the expected amount of work associated with passing from $\mathbb{P}_Y$ to $\mathbb{P}_{Y|X_i}$.

The properties of zero-independence and max-functionality help us with the interpretation of the two extremes of the scale: when receiving information on X_i has either no value or maximal value.

Example 23. Consider the following input-output mapping, $g : \mathbb{R}^4 \to \mathbb{R}^2$,

$$\begin{cases} Y_1 = X_1 X_2 + 0 \cdot X_4 \\ Y_2 = X_1 X_3 + 0 \cdot X_4, \end{cases} \tag{89}$$

with X_1 uniformly distributed on $[-1, 1]$, X_2, X_3 and X_4 uniformly distributed on $[0, 1]$. Then, we obtain the values of the variance-based and optimal transport-based global sensitivity measures in Table 4.

Table 4: Global sensitivity measures for this example

	X_1	X_2	X_3	X_4
$\xi^{LG}(Y, X_i)$	0.75	0	0	0
$\iota(Y, X_i)$	0.63	0.10	0.10	0

The zero values of $\xi^{LG}(Y, X_2)$, $\xi^{LG}(Y, X_3)$ and $\xi^{LG}(Y, X_4)$ may give the false impression that these inputs are irrelevant. To avoid this false impression, we can rely on the values of $\iota(Y, X_i)$, which indicate that Y is dependent on X_1, X_2 and X_3, with a stronger dependence on X_1, but independent of X_4.

Zero-independence is crucial in avoiding false negatives. This is because even if Y is a function of X_i, the value of a global sensitivity measure may be zero in the absence of this property. The max-functionality property helps us interpreting the other extreme of the scale: The closer $\iota(Y, X_i)$ is to unity the closer information about X_i is to eliminate uncertainty in Y. In Example 23, the input whose value is closer to unity is X_1, and can therefore be judged as the most important variable. However, because $\iota(Y, X_1) < 1$, information about this input does not eliminate uncertainty in Y. We need to learn simultaneously the values of X_1, X_2 and X_3, because $\iota(Y, X_2)$ as well as $\iota(Y, X_3)$ are greater than zero.

Monotonicity helps with another intuition: if we receive less accurate information about X_i, the value of the importance measure decreases.

Example 24 (Example 23 continued). Suppose we receive information about X_1 in the distorted form $Z_1 = X_1^2$. The distorted information hides the negative part of X_1 and it is less accurate than direct information on X_1. Calculating its OT-based importance, we obtain $\iota(Y, Z_1) = 0.19$, which is about 70% lower then $\iota(Y, X_1)$.

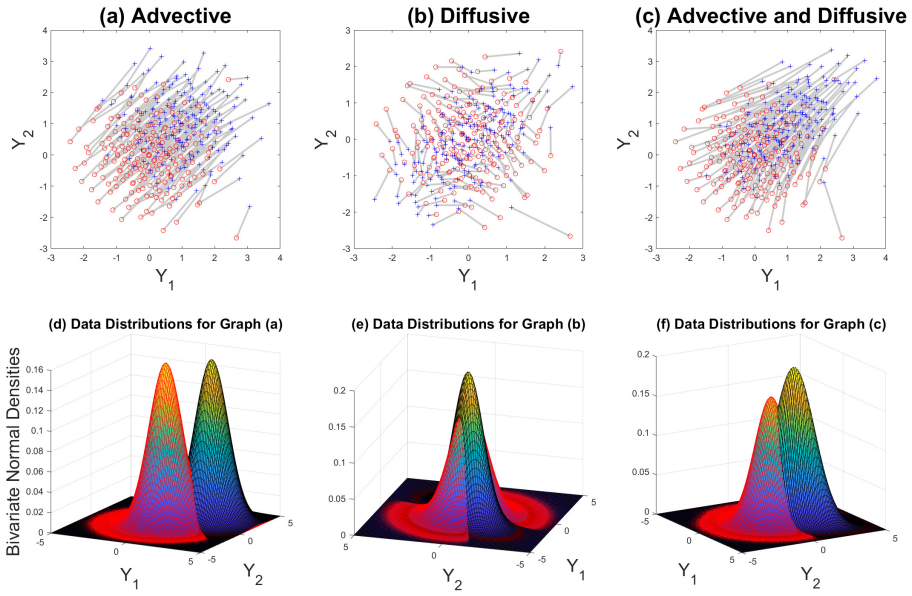


Figure 9: Graphs (a), (b), (c): circles (∘), original data.

Variance-based indices $\xi^W(Y, X_i)$ do not have this property. Example 22 in Appendix C illustrates a situation where two inputs, say Z and X, have the same variance-based importance. However, Z is a non-monotonic transformation of X, and is, therefore, less informative. This difference is instead signaled by their OT-based importance measures, according to which it is $\iota(Y, Z) < \iota(Y, X)$. Figure 9 helps us discussing further the interpretation of the advective and diffusive terms in (38) and (39). Graph (a) in Figure 9 shows data generated with two normal distributions whose parameterization differs only in their means namely $m_Y = [0\ 0]$ and $m_{Y'} = [1\ 1]$. The variance-covariance matrices are both equal to $\Sigma = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. (The corresponding densities are displayed in Graph (d)). The circles (∘) correspond to realizations sampled from the first distribution, the plus signs (+) to realizations from the second distribution. The lines joining pairs of these points (a ∘ and a +) show the optimal couplings. Graph (b) shows a transport between two distributions with identical means, $m_Y = m_{Y'} = [0\ 0]$, and different variance-covariance matrices, namely $\Sigma_Y = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $\Sigma_{Y'} = \left[\begin{smallmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{smallmatrix}\right]$ — see Graph (e) for the corresponding densities. In this case, the advective transport is null and the transport is purely diffusive. A visual comparison of Graphs (a) and Graph (b) shows that in Graph (a) translations play a prevailing role in the optimal coupling, while rotations play a major role in Graph (b). Finally, Graph (c) displays data generated from two multivariate normal distributions that differ in both their means and variance covariance matrices and the corresponding optimal couplings. (The distribution parameters are $m_Y = [0\ 0]$, $\Sigma_Y = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $m_{Y'} = [1\ 1]$, $\Sigma_{Y'} = \left[\begin{smallmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{smallmatrix}\right]$.) Because we have both an advective and a diffusive component in the optimal transport, translational (advective) as well as rotational (diffusive) effects appear.