Electronic Journal of Statistics
Vol. 17 (2023) 3854–3892
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2189

Affine invariant integrated rank-weighted statistical depth: properties and finite sample analysis

Stephan Clémençon

Telecom Paris, LTCI, Institut Polytechnique de Paris,
19 place Marguerite Perey, Palaiseau, 91120, France
e-mail: stephan.clemencon@telecom-paris.fr

Pavlo Mozharovskyi

Telecom Paris, LTCI, Institut Polytechnique de Paris,
19 place Marguerite Perey, Palaiseau, 91120, France
e-mail: pavlo.mozharovskyi@telecom-paris.fr

Guillaume Staerman*

LTCI, Télécom Paris, Institut Polytechnique de Paris,
19 place Marguerite Perey, Palaiseau, 91120, France
e-mail: guillaume.staerman@inria.fr

Abstract: Because it determines a center-outward ordering of observations in $\mathbb{R}^d$ with $d \geq 2$, the concept of statistical depth permits to define quantiles and ranks for multivariate data and use them for various statistical tasks (e.g. inference, hypothesis testing). Whereas many depth functions have been proposed *ad-hoc* in the literature since the seminal contribution of [50], not all of them possess the properties desirable to emulate the notion of quantile function for univariate probability distributions. In this paper, we propose an extension of the *integrated rank-weighted* statistical depth (IRW depth in abbreviated form) originally introduced in [40], modified in order to satisfy the property of *affine invariance*, fulfilling thus all the four key axioms listed in the nomenclature elaborated by [59]. The variant we propose, referred to as the affine invariant IRW depth (AI-IRW in short), involves the precision matrix of the (supposedly square integrable) d-dimensional random vector X under study, in order to take into account the directions along which X is most variable to assign a depth value to any point $x \in \mathbb{R}^d$. The accuracy of the sampling version of the AI-IRW depth is investigated from a non-asymptotic perspective. Namely, a concentration result for the statistical counterpart of the AI-IRW depth is proved. Beyond the theoretical analysis carried out, applications to anomaly detection are considered and numerical results are displayed, providing strong empirical evidence of the relevance of the depth function we propose here.

MSC2020 subject classifications: Primary 62G05; secondary 62H99, 62G30, 68Q32.
Keywords and phrases: Statistical depth, integrated rank-weighted depth, affine invariance, concentration inequalities, anomaly detection.

Received November 2022.

*Corresponding author.

Contents

1. Introduction

Since its introduction in [50], the concept of statistical depth has become increasingly popular in multivariate data analysis. For a distribution P on $\mathbb{R}^d$ with

$d > 1$, by transferring the natural order on the real line to $\mathbb{R}^d$, a depth function $D(., P) : \mathbb{R}^d \to \mathbb{R}_+$ provides a center-outward ordering of points in the support of P and can be straightforwardly used to extend the notions of (signed) rank or order statistics to multivariate data. It finds numerous applications in Statistics and Machine Learning such as robust inference [28], hypothesis testing [15] or novelty/anomaly detection [48], to name a few, see [34] and [35] for further examples. Numerous definitions have been proposed, as alternatives to the earliest proposal, the *halfspace* depth introduced in [50]: with simplicial [29], projection [30], majority [31], Oja [37], zonoid [23], spatial ([5] or [51]) and Monge-Kantorovich [8] depths being common examples, among many others. In order to compare systematically their merits and drawbacks, [59] have developed an axiomatic nomenclature of statistical depths, listing key properties that should be (ideally) satisfied by a 'proper' depth function. Roughly, as depth functions serve to define center-outward orderings, if a distribution P on $\mathbb{R}^d$ has a unique center $\theta \in \mathbb{R}^d$ (*i.e.*, a symmetry center in a defined sense, see [59] for details), the latter should be the deepest point. Further, for any deepest point, depth function should decrease along any fixed ray starting from it. One also expects that a depth function vanishes at infinity and does not depend on the coordinate system chosen. This latter property is usually formulated as *affine invariance*. (Section 2 below provides a thorough formulation of these four properties.) Beyond verifying these properties, the pros and cons of any data depth should be considered regarding the possible existence of algorithms for computation in the case of empirical distributions. In this respect, the extension of Tukey's halfspace depth recently introduced in [40] and referred to as the integrated rank-weighted (further IRW for shortness) depth offers many advantages. Rather than computing—for any point $x \in \mathbb{R}^d$—the minimum of the mass $P(H)$, $H \in \mathcal{H}_x$ taken over all closed halfspaces $\mathcal{H}_x = \{x' \in \mathbb{R}^d : \langle x' - x, u\rangle \leq 0, u \in \mathbb{S}^{d-1}\}$ with unit normal vector u and containing x, it is proposed to replace the infimum by the integral taken with respect to (w.r.t.) all possible directions u uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$ (following the footsteps of the general *integrated dual depth* approach developed in [10]). For an empirical or discrete distribution, IRW depth thus admits a weighted average representation. It can be easily approximated using Monte-Carlo methods in contrast to many other depth functions, whose values are defined as solutions to optimization problems, possibly complex ones in a high dimension. Beyond these computational aspects, it is shown in [40] that the IRW data depth satisfies several desirable properties (see Theorem 2 therein). Unfortunately, it does not fulfill the *affine invariance* property, crucial to the multivariate analysis of commensurable variables. Indeed, the values taken by the IRW depth may highly depend on the chosen coordinate system to represent available statistical information, ruining their interpretability, as will be shown on illustrative examples of Section 2. It is the main purpose of this paper to overcome—in a systemic way—the lack of affine invariance property in the definition of IRW depth by proposing a modified version of it, named AI-IRW. It consists in the IRW depth of the (square-integrable) random vector X with distribution P under study expressed in an orthogonal coordinate system such that its components are linearly uncor-

related. That is, such components are the principal components of X obtained by eigenvalue decomposition of its covariance matrix Σ. Under the assumption that Σ is positive-definite, the *affine invariant* version of IRW depth of X is the IRW depth of WX, denoting by W any *whitening matrix* (*i.e.* any square matrix W such that $W^\top W = \Sigma^{-1}$). In the case if Σ is not positive-definite, the proposed methodology should be naturally applied after a dimensionality reduction step, *i.e.* on an appropriate orthogonal projection of the original random vector X.

In this article, we show that the affine invariant version of IRW depth, further on referred to as the AI-IRW depth, is independent of the whitening matrix chosen, inherits all the properties and computational advantages of the IRW depth and aditionally satisfies the *affine invariance* property. Since its statistical counterpart based on a sample composed of independent copies of the random variable X is a complex function of the data which involves empirical version of an orthonormal transform of $\Sigma^{-1/2}$ (*i.e.*, the square root of the precision matrix), a finite-sample analysis is carried out here. Precisely, a concentration result for the sampling version of the AI-IRW depth is established. Beyond this theoretical outcome, the relevance of this depth notion is also supported by experimental results. When applied in various statistical tasks such as anomaly detection, it demonstrates superiority over the IRW and other existing depth measures, making it a strong contender in the field.

The article is structured as follows. In Section 2, the concept of data depth is briefly reviewed, and in particular the integrated rank-weighted depth [40], together with the axiomatic approach developed by [59] and illustrating examples; particular attention is paid to the affine invariance property. In Section 3, the AI-IRW depth is introduced, its properties are studied, and questions of approximation and estimation are discussed at length. The accuracy of the empirical version is investigated in Section 4 from a non-asymptotic perspective. Section 5 describes experimental results that empirically illustrate the advantages of the AI-IRW depth. Finally, concluding remarks are collected in Section 6. Proofs, additional technical details, and numerical results are deferred to the Appendix.

2. Background and motivations

The concept of depth function is motivated by necessity to extend the very useful notions of order and (signed) rank statistics in univariate statistical analysis to multivariate settings through depth-induced contours. Indeed, such statistics perform a wide variety of tasks, ranging from robust statistical inference to efficient statistical hypothesis testing. The earliest proposal is the halfspace depth developed in [50]. For any probability measure P_1 on $\mathbb{R}$, the univariate halfspace depth is defined by: $\forall\, t \in \mathbb{R}$,

$$D_{\mathrm{H},1}(t, P_1) = \min\left\{P_1\left(]-\infty, t]\right), P_1\left([t, +\infty[\right)\right\}.$$

Considering a multivariate r.v. X with probability distribution P on $\mathbb{R}^d$ with $d > 1$, its halfspace depth at $x \in \mathbb{R}^d$ is then defined as the infimum of the

probability mass taken over all possible closed halfspaces containing x:

$$D_{\mathrm{H}}(x,P)=\inf_{u\in\mathbb{S}^{d-1}}\mathbb{P}\left(\langle u,X\rangle\ \leq\ \langle u,x\rangle\right), \tag{1}$$

denoting by $\langle\cdot,\cdot\rangle$ and $||\cdot||$ the usual Euclidean inner product and norm on $\mathbb{R}^d$, and by $\mathbb{S}^{d-1}=\{z\in\mathbb{R}^d:\ ||z||=+1\}$ the unit sphere of $\mathbb{R}^d$ w.r.t. the Euclidean norm. Because of its appealing properties, halfspace depth (1) is undeniably the most documented notion of depth function in the statistical literature. It has been proved to fully characterize empirical and finitely discrete distributions in [49, 24]. Asymptotic properties such as consistency or asymptotic normality of its sampling version based on independent copies $X_1,\ldots,X_n$ of the generic r.v. X, obtained by replacing P in (1) with the empirical distribution $\widehat{P}=(1/n)\sum_{i=1}^n\delta_{X_i}$, where δ_x means the Dirac mass at any point x, are established in, e.g., [43, 12, 59]. Multivariate location estimators based on halfspace depth have been investigated in [13], and it has been shown to possess attractive robustness properties. For instance, the asymptotic breakdown point of the Tukey median, i.e., the barycenter of the deepest locations in the sense of (1), is equal to 1/3 for absolutely continuous centrosymmetric distributions, see [13]. Computational issues have also been extensively studied, see [33] or [32] for instance. However, as recalled in the introduction, many other notions of depth have been proposed during the last decades, far too numerous to be listed in an exhaustive manner here. We refer the reader to [34] or the Chapter 2 of [47] for excellent accounts of the statistical depth theory. A depth function has two arguments, it is a function $D:\mathbb{R}^d\times\mathcal{P}\to\mathbb{R}_+$, where $\mathcal{P}$ is some set of probability distributions on $\mathbb{R}^d$, which not necessarily contains all probability distributions on $\mathbb{R}^d$; see, e.g., [35]. In order to guarantee the "center-outward-ordering" interpretation of D, four key properties have been listed by [59], see also [14] and [34] for their different formulation. These are recalled below.

D$_1$ (Affine invariance) Denoting by P_X the distribution of a r.v. X taking its values in $\mathbb{R}^d$, we have

$$\forall x\in\mathbb{R}^d,\ \ D(Ax+b,P_{AX+b})=D(x,P_X),$$

for any d-dimensional r.v. X, any $d\times d$ nonsingular matrix A with real entries and any vector b in $\mathbb{R}^d$.

D$_2$ (Maximality at center) For any probability distribution P on $\mathbb{R}^d$ that possesses a symmetry center x_P (in a sense to be specified below), the depth function $D(\cdot,P)$ takes its maximum value at it:

$$D(x_P,P)=\sup_{x\in\mathbb{R}^d}D(x,P).$$

D$_3$ (Monotonicity relative to deepest point) Let $P\in\mathcal{P}$, and x_P be a deepest point of P. Then the depth function decreases on any ray that begins at x_P, i.e., for any $x\in\mathbb{R}^d$ and $\alpha\geq 0$

$$D(x_P,P)\geq D(x_P+\alpha(x-x_P),P).$$

D$_4$ (VANISHING AT INFINITY) For any probability distribution P on $\mathbb{R}^d$, the depth function D vanishes at infinity:

$$D(x, P) \to 0, \text{ as } ||x|| \to \infty.$$

It is worth mentioning that the most general notion of symmetry when analyzing data depth is the halfspace symmetry and shall be the one used throughout this article. Precisely, the probability distribution P is halfspace symmetric at x_P if $P(\mathcal{H}_{x_P}) \geq 1/2$ for any closed halfspace $\mathcal{H}_{x_P}$ passing through x_P. Various works have examined which of the properties, among those listed above, are satisfied by specific notions of depth introduced in the literature, see [59]. Some of them are constructed as an infimum—over projections on the unit sphere—of a univariate non-parametric statistics such as the projection depth proposed by [30] or those introduced in [56] or [58]. From a practical perspective, computing these projection-based depths involves using tools such as manifold optimization algorithms, facing various numerical difficulties as the dimension d increases, see [16]. In addition, the halfspace depth suffers from two major problems: (i) for each data point x, taking the direction achieving the minimum to assign a score to x possibly creates a significant sensitivity to noisy directions and (ii) the null score assigned to each new data point outside of the convex hull of the support of the distribution P makes the score of such points indistinguishable. A remedy based on Extreme Value Theory has been proposed in [17], which consists in smoothing the halfspace depth beyond the convex hull of the data. However, this variant relies on rather rigid parametric assumptions, is only approximately affine invariant and is confronted with the aforementioned limitation regarding the non-smoothed part of the data. Recently, alternative depth functions have been proposed, obtained by replacing the infimum over all possible directions with an integral, see [10]. In [40], a new data depth, referred to as the Integrated Rank-Weighted depth, is defined by substituting an integral over the sphere $\mathbb{S}^{d-1}$ for the infimum in (1). Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the spherical probability measure on $\mathbb{S}^{d-1}$ by ω_{d-1}, the $d \times d$ identity matrix by $\mathcal{I}_d$.

Definition 1 ([40])**.** *The Integrated Rank-Weighted (IRW) depth of $x \in \mathbb{R}^d$ w.r.t. a probability distribution P on $\mathbb{R}^d$ is defined as follows:*

$$\begin{aligned} D_{\text{IRW}}(x, P) &= \int_{\mathbb{S}^{d-1}} D_{\text{H},1}(\langle u, x\rangle, P_u)\ \omega_{d-1}(\mathrm{d}u) \\ &= \mathbb{E}\left[D_{\text{H},1}(\langle U, x\rangle, P_U)\right], \end{aligned} \tag{2}$$

where P_u is the pushforward distribution of P defined by the projection $x \in \mathbb{R}^d \mapsto \langle u, x\rangle$ and U is a r.v. uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.

As explained at length in [40], the name of the data depth (2) originates from the fact that it can be represented as a weighted average of a finite set of normalized center-outward ranks. It has many advantages over the original halfspace depth (1). First, by construction, it is *robust* to noisy directions, and *sensitive* to new data points outside of the convex hull of the training dataset

simultaneously, fixing the two problems mentioned above. Moreover, concerning numerical feasibility, the computation of the IRW depth does not require implementing any manifold optimization algorithm. It can be approximated using basic Monte Carlo techniques, providing confidence intervals as a by-product, see Remark 1 below. Its contours $\{D_{\mathrm{IRW}}(x,P)=\alpha\}$, $\alpha\in[0,1]$, also exhibit a higher degree of smoothness in general (the depth function (2) is continuous at any point $x\in\mathbb{R}^d$ that is not an atom for P, cf. Proposition 1 in [40]) and properties $\mathbf{D}_2$, $\mathbf{D}_3$ and $\mathbf{D}_4$ have been proved to be satisfied by (2) under mild assumptions, see Theorem 2 in [40].

Remark 1 (Monte Carlo approximation)**.** *Recall that a r.v. uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$ can be generated from a d-dimensional centered Gaussian random vector Z with the identity $\mathcal{I}_d$ as covariance matrix: if $Z\sim\mathcal{N}(0,\mathcal{I}_d)$, then $Z/||Z||\sim\omega_{d-1}$, see [25]. Hence, a basic Monte-Carlo method to approximate (2) would consist in generating $m\geq 1$ independent realizations $Z_1,\ldots,Z_m$ of $\mathcal{N}(0,\mathcal{I}_d)$ and compute*

$$\frac{1}{m}\sum_{j=1}^{m} D_{\mathrm{H},1}(\langle Z_j/||Z_j||,x\rangle, P_{Z_j/||Z_j||}), \tag{3}$$

refer to, e.g., [21] for an account of Monte Carlo integration methods.

However, it does not satisfy the crucial property $\mathbf{D}_1$ (affine invariance) in general, as illustrated in the two following examples (see also the next section and Section C.2 for additional numerical illustrations).

Example 1. *Here we provide an example of discrete distribution where IRW does not satisfy the affine-invariance property. Consider the discrete probability measure P assigning the weight $1/3$ to the bivariate points in $\mathcal{D}_3=\{(-1,2),(3,3),(2,1)\}$ and let us compute the IRW depth of $x=(0,1)$ and $y=(3,2)$ relative to P. It is easy to see that the mappings $u\in\mathbb{S}^1\mapsto D_{\mathrm{H},1}(\langle u,\ x\rangle,P_u)$ and $u\in\mathbb{S}^1\mapsto D_{\mathrm{H},1}(\langle u,\ y\rangle,P_u)$ take only two values, 0 or $1/3$. Identifying $\mathbb{S}^1$ as $[0,2\pi[$, the univariate halfspace depth of x relative to P is then null for any $u\in[\pi/4,\pi/2]\cup[5\pi/4,3\pi/2]$ and equal to $1/3$ if u belongs to the complementary set. In addition, $D_{\mathrm{H},1}(\langle u,\ y\rangle,P_u)$ is equal to 0 for any $u\in[3\pi/4,\pi]\cup[7\pi/4,2\pi]$ and equal to $1/3$ on the complementary set. One may easily check that $D_{IRW}(x,P)=D_{IRW}(y,P)=0.25$ and the same rank would be then assigned to each point by the IRW depth. Now, multiplying all ordinate values by 2, which is an affine transformation, the univariate halfspace depth of $\widetilde{x}=(0,2)$ is null for all u in $[\pi/8,\pi/2]\cup[9\pi/8,3\pi/2]$. At the same time, it remains equal to $1/3$ on the complementary set of this region. The depth of $\widetilde{x}$ is thus lower than 0.25. On the other hand, the univariate depth of $\widetilde{y}=(3,4)$ is now null on $[7\pi/8,\pi]\cup[15\pi/8,2\pi]$ while it remains equal to $1/3$ on the complementary set of this interval. It follows that $D_{IRW}(\widetilde{x})=5/24<0.25<7/24=D_{IRW}(\widetilde{y})$.*

Example 2. *This second example is illustrated numerically with a Gaussian distribution. To that end, we draw a sample $\mathbf{Z}$ with 1000 instances from a two-*

dimensional standard Gaussian distribution. We then apply a linear transformation of $\mathbf{Z}$ *by the matrix*

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 8 \end{bmatrix}$$

and define a second sample $A\mathbf{Z}$*. These two samples* $\mathbf{Z}$ *and* $A\mathbf{Z}$ *are depicted in Fig. 1 (top). We compute the approximation of the IRW depth, given in (2), on* $\mathbf{Z}$ *and* $A\mathbf{Z}$ *(for* $m = 10^6$ *to reduce the approximation error). The rank induced by the IRW depth is displayed through a color bar from yellow to dark blue (the darker it is, the higher the depth is). In addition, we depict the sorted depth values for the IRW depth computed on* $\mathbf{Z}$ *(red) and the corresponding IRW depth value for the same indices but computed on* $A\mathbf{Z}$ *(magenta). As we can see, the value of the IRW depth varies significantly between the sample and its linear transformation.*

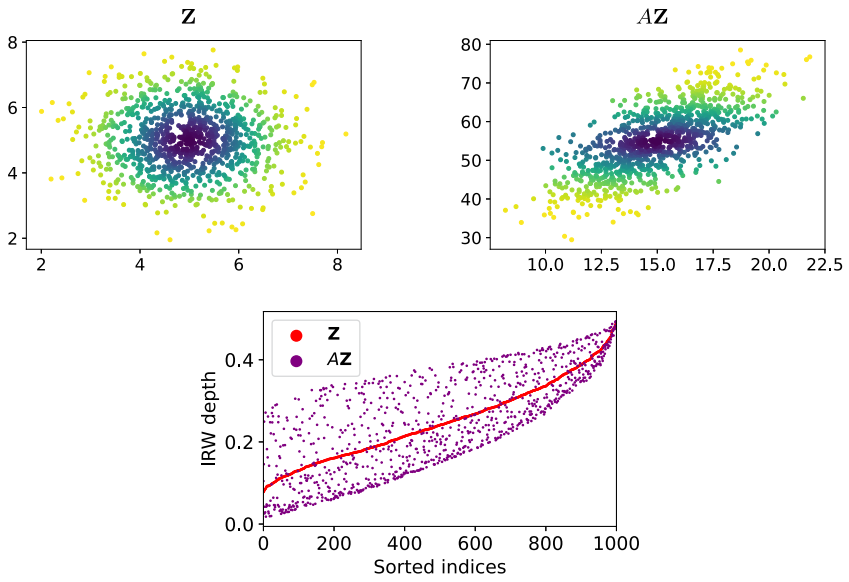

FIG 1. *Illustration of the non-affine-invariance of the IRW depth. The depth-induced ranks are highlighted by a color scale (the darker it is, the deeper it is) on the sample* $\mathbf{Z}$ *and its linear transformation* $A\mathbf{Z}$ *(top). The according to* $\mathbf{Z}$ *IRW-depth-sorted values are depicted for both samples (bottom).*

The property of affine invariance, i.e. insensitivity (in a proper sense) to linear transforms applied to both P and x refers to fundamental ideas of multivariate analysis and is important once linear combinations of different variables are considered. Although, generally speaking, expediency of commensurability can depend on a practical situation at hand, e.g., only orthogonal transforms can be applicable to certain types of data, such as real-world objects' coordinates, handling weighted sums of observations' variables is commonly used in statistics

since at least a century with linear discriminant analysis [19], principal component analysis [38], perceptron classifier [41] as well as logistic regression and a single neuron of an artificial neural network or support vector machine [9] being only very few examples. The fact that IRW depth is affected by non-uniform scaling is very problematic in practice regarding its interpretability in particular or its use for anomaly detection tasks, for instance, see Section 5 and is the main flaw of this approach as pointed out in [10, 40].

3. Affine invariant IRW depth – definition and properties

Here we propose to modify the depth function (2) in order to ensure that property $\mathbf{D}_1$ is always satisfied when the random vector X with distribution P under study is assumed to be square integrable with positive definite covariance matrix Σ. Precisely, rather than taking the expectation w.r.t. a random direction U uniformly distributed on $\mathbb{S}^{d-1}$ (i.e. integrating over all possible directions $u \in \mathbb{S}^{d-1}$), one considers the random projections defined by the eigenfunctions of the matrix Σ, i.e. the principal components of the r.v. X. In other words, the expectation is taken w.r.t. the distribution of the random vector $V = W^\top U/||W^\top U||$ valued in $\mathbb{S}^{d-1}$, where W is any *whitening* matrix, as formulated in the definition below.

Proposition 2 (Affine invariant IRW depth)**.** *Let X be a square integrable random vector with probability distribution P on $\mathbb{R}^d$ and positive definite covariance matrix Σ. Consider the function*

$$x \in \mathbb{R}^d \mapsto \mathbb{E}\left[D_{\mathrm{H},1}(\langle V, x\rangle, P_V)\right], \tag{4}$$

where $V = W^\top U/||W^\top U||$, U being uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$ and W a whitening matrix (i.e. a matrix W of the form $Q\Sigma^{-1/2}$ with $Q^\top Q = \mathcal{I}_d$). Then, the function (4) is independent from the whitening matrix W chosen. It is denoted by $D_{\text{AI-IRW}}(\cdot, P)$ and referred to as the affine invariant Integrated Rank-Weighted (AI-IRW) depth w.r.t. X.

The fact that (4) is independent from the whitening matrix W chosen (or, equivalently, from the orthonormal matrix $Q = W\Sigma^{1/2}$) results from a straightforward change of variable, details are left to the reader. Hence, any whitening matrix may be used to define the AI-IRW depth. For instance, a whitening matrix W can be obtained either by singular value decomposition or by Cholesky decomposition. Of course, in the case where the covariance matrix Σ of the supposedly square integrable r.v. X is not invertible, the AI-IRW depth notion should be applied to an orthogonal projection, after an appropriate dimensionality reduction step. From a computational perspective, the AI-IRW depth can be approximated by Monte Carlo methods in the same way as (2), see Remark 1. As revealed by the proposition stated below, the depth function (4) inherits all the properties of (2) under similar assumptions and is remarkably invariant under any affine transformation in addition.

Proposition 3 (Properties of the AI-IRW depth)**.** *The assertions below hold true for any probability distribution P of a square integrable r.v. X valued in $\mathbb{R}^d$ with positive definite covariance matrix.*

(*i*) *The AI-IRW depth satisfies the properties* $\mathbf{D}_1$ *and* $\mathbf{D}_4$*. In addition, the properties* $\mathbf{D}_2$ *and* $\mathbf{D}_3$ *are fulfilled for all halfspace symmetric distributions.*
(*ii*) *The AI-IRW depth function is continuous at each point x that is not an atom for P.*

The proof is detailed in Section B.1 of the Appendix. It is known that for elliptical distributions, affine invariant data depth level sets are concentric ellipsoids with the same center and orientation as the density level sets [31]. Therefore, the ordering returned by affine invariant data depths should be equal to that of the density function. Thus, in order to highlight the discrepancy between AI-IRW and IRW w.r.t. affine invariance, we propose to compare the ordering returned by AI-IRW and IRW to that of the density function on the Gaussian distribution (which belongs to the family of elliptical distributions). As illustrated by the Rank-Rank plots in Fig. 2, the ordering defined by the (empirical) AI-IRW depth is generally much closer to that induced by the underlying density than the order defined by the original (IRW depth) version.

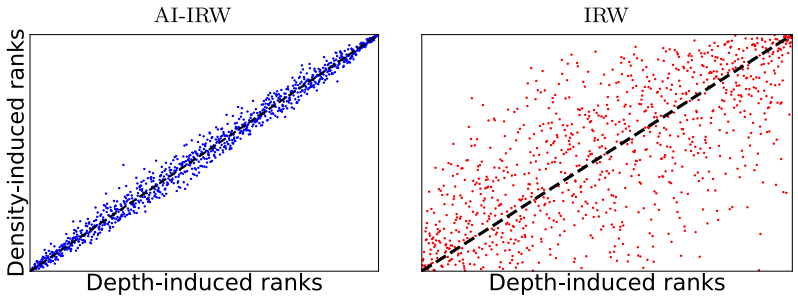

FIG 2. *Rank-Rank plots comparing the ranks of* 1000 *points sampled from a* 10*-d (anisotropic) Gaussian distribution with covariance matrix drawn at random from a Wishart distribution (with parameters $(d, \mathcal{I}_d)$) induced by the empirical depth (AI-IRW on the left, IRW on the right) and those induced by the Gaussian density.*

Sampling versions In practice, the distribution P is generally unknown as well as the covariance matrix Σ and only a sample $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ composed of $n \geq 1$ independent realizations of the distribution P is available. A statistical counterpart of the AI-IRW depth can be obtained by replacing P with the empirical measure $\widehat{P} = (1/n)\sum_{i=1}^n \delta_{X_i}$ and the whitening matrix W with a non-singular estimator $\widehat{W}$ based on $\mathcal{D}_n$ and plugging them next into formula (4), yielding: $\forall x \in \mathbb{R}^d$,

$$\widehat{D}_{\text{AI-IRW}}(x) = \mathbb{E}\left[D_{\text{H},1}(\langle \widehat{V}, x\rangle, \widehat{P}_{\widehat{V}}) \mid \mathcal{D}_n\right], \tag{5}$$

where $\widehat{V} = \widehat{W}^\top U/||\widehat{W}^\top U||$ and U is a r.v. uniformly distributed on $\mathbb{S}^{d-1}$ independent from the X_i's. From a practical perspective, the (conditional) expectation (5) can also be approximated by means of a basic Monte Carlo scheme, generating $m \geq 1$ i.i.d. random directions $U_1, \ldots, U_m$, copies of the generic r.v. U and independent from the original data $\mathcal{D}_n$: $\forall x \in \mathbb{R}^d$,

$$\widetilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(x) = \frac{1}{m} \sum_{j=1}^{m} \min \left\{ \widehat{F}_{\widehat{V}_j}\left(\langle \widehat{V}_j, x\rangle\right), 1 - \widehat{F}_{\widehat{V}_j}\left(\langle \widehat{V}_j, x\rangle\right) \right\}, \tag{6}$$

where, for all $j \in \{1, \ldots, m\}$ and $t \in \mathbb{R}$, we set

$$\widehat{V}_j = \widehat{W}^\top U_j/||\widehat{W}^\top U_j||, \quad \text{and} \quad \widehat{F}_{\widehat{V}_j}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\left\{\langle \widehat{V}_j, X_i\rangle \leq t\right\}}.$$

Putting aside temporarily the issue of estimating a whitening matrix W (discussed below), attention should be paid to the fact that the approximate sample version (6) is very easy to compute (see Algorithm 1 for the computation of $\widetilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(X_i)$ for all $i \leq n$) and involves no optimization procedure, in contrast to many other notions of depth function.

Algorithm 1 Approximation of the AI-IRW depth

Initialization: the number of projections m.

1: Construct $\mathbf{U} \in \mathbb{R}^{d\times m}$ by sampling uniformly m vectors $U_1, \ldots, U_m$ in $\mathbb{S}^{d-1}$
2: Compute a non-singular estimator $\widehat{\Sigma}$ of the covariance
3: Apply a whitening procedure (*e.g.* Cholesky decomposition or SVD) to $\widehat{\Sigma}$, yielding $\widehat{W}$
4: Compute $\mathbf{V} = \widehat{W}^\top \mathbf{U}/||\widehat{W}^\top \mathbf{U}||$
5: Compute $\mathbf{M} = \mathbf{XV}$
6: Compute the rank value $\sigma(i,j)$, the rank of index i in $\mathbf{M}_{:,j}$ for every $i \leq n$ and $j \leq m$
7: Set $\widetilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(X_i) = \frac{1}{m}\sum_{j=1}^{m} \sigma(i,j)$ for every $i \leq n$

Output: $\widetilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(X_i)$, $i \leq n$

On estimating a whitening matrix Consider the $d \times n$ matrix $\mathbf{X}_n = (X_1, \ldots, X_n)$ with the X_i's as columns. The simplest way of building an estimate $\widehat{W}$ consists in computing the empirical version $\widehat{\Sigma} = (1/n)\mathbf{X}_n\mathbf{X}_n^\top$ of the covariance matrix, which is a natural and nearly unbiased estimator, and applying next any whitening method (*e.g.* ZCA, PCA or Cholesky whitening) to it, when the latter is positive definite, producing a matrix of the form $\widehat{W} = Q\widehat{\Sigma}^{-1/2}$, where Q is a $d \times d$ orthonormal matrix. When the empirical covariance matrix $(1/n)\mathbf{X}_n\mathbf{X}_n^\top$ is not invertible, whitening is applied to a regularized, non-singular, version $\widehat{\Sigma}$ of it using *e.g.* Tikhonov regularization method. For simplicity, the estimator $\widehat{\Sigma}$ of Σ considered in the finite-sample study presented in the next section is the possibly regularized empirical covariance. However, alternative estimation techniques can be used, yielding possibly more efficient estimators under specific assumptions, in high-dimension especially. Shrinkage procedures

for covariance estimation under sparsity conditions have been investigated in e.g. [26, 7, 45], while a lasso method for direct estimation of the precision matrix, avoiding matrix inversion, is proposed in [20]. Robust covariance estimation techniques, tailored to situations where the data are possibly contaminated or heavy-tailed, have also been documented in the literature, see e.g. [42] and [44]. For the the sake of simplicity as well, empirical Cholesky whitening is considered to define and analyze the sampling version of the AI-IRW depth, but it is straightforward to see that the results hold true for any other whitening transformation of $\widehat{\Sigma}$.

Due to the presence of $\widehat{W}$ in (5) (respectively, in (6)), it is far from straightforward to assess the accuracy of the estimators of the AI-IRW depth proposed above. It is the purpose of the next section to study the uniform deviations between (4) and its empirical versions from a non-asymptotic perspective.

4. Finite-sample analysis – concentration bounds

We now investigate the accuracy of the sample version, as well as that of its Monte Carlo approximation, of the AI-IRW depth function introduced in the previous section in a non-asymptotic fashion. Keeping both the sample size n and the number of m of MC directions fixed, we establish a concentration bound for the maximal deviations between the true and estimated AI-IRW depth values, at some point x that holds with a probability $1-\delta$. To that end, we assume here that the estimator $\widehat{\Sigma}$ of the covariance Σ is the empirical covariance, if the latter is definite positive, and of any definite positive regularized version (e.g. Tikhonov) of the latter otherwise. The empirical whitening matrix W is the transpose of a Cholesky decomposition L of $\widehat{\Sigma}^{-1}$: $W = L^\top$, where L is a real lower triangular matrix with positive diagonal entries. The subsequent analysis requires additional hypotheses, listed below. The first assumption, classical when estimating a whitening matrix (see e.g. [4] or [18]), stipulates that the eigenvalues $\sigma_1, \ldots, \sigma_d$ of the covariance matrix Σ of the square integrable random vector X considered are bounded away from zero.

Assumption 1. *Assume that the smallest eigenvalue, $\varepsilon = \min_k \sigma_k$, is positive.*

The second assumption is technical, see [11]. It stipulates that Σ's eigenvalues are all of multiplicity 1 and that Σ's minimum eigengap is bounded away from zero.

Assumption 2. *Assume that all Σ's eigenvalues are different and their smallest difference, γ, is positive.*

We point out that, just like when Σ is not invertible, one always may bring back the analysis to a situation where Assumption 2 is fulfilled by means of a preliminary dimensionality reduction step. Notice incidentally that, when $\Sigma = \sigma\mathcal{I}_d$, with $\sigma > 0$, the AI-IRW reduces to IRW. The other assumptions correspond to smoothness conditions of Lipschitz type for the function $\phi : (u, x) \in \mathbb{S}^{d-1} \times \mathbb{R}^d \mapsto \mathbb{P}\{\langle u, X\rangle \leq \langle u, x\rangle\}$.

Assumption 3 (Uniform Lipschitz condition in projection)**.** *For all* $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$*, there exists* $L_p < +\infty$ *such that*

$$\sup_{u \in \mathbb{S}^{d-1}} |\phi(u, x) - \phi(u, y)| \leq L_p ||x - y||.$$

Assumption 4 (Uniform radial Lipschitz condition)**.** *For all* $(u, v) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$*, there exists* $L_R < +\infty$ *such that*

$$\sup_{x \in \mathbb{R}^d} |\phi(u, x) - \phi(v, x)| \leq L_R ||u - v||.$$

Notice that the same assumptions are involved in the non-asymptotic rate bound analysis carried out for the halfspace depth estimator in [2] and are used to establish limit results related to its approximation in [36]. The Lipschitz conditions are satisfied by a large class of probability distributions, for which Lipschitz constants L_R and L_p can be both explicitly derived. For instance, assume that the distribution P of X has compact support included in the ball $\mathcal{B}(0, r) = \{x \in \mathbb{R}^d : ||x|| \leq r\}$ relative to the Euclidean norm $||\cdot||$ with $r > 0$ and is absolutely continuous w.r.t. the Lebesgue measure with a density bounded by $M > 0$. Thus, the uniform Lipschitz conditions are then fulfilled with $L_R = MV_{d,r}$ and $L_p = MV_{d-1,r}$, where $V_{d,r} = \pi^{d/2} r^d / \Gamma(d/2 + 1)$ is the volume of the ball $\mathcal{B}(0, r)$ and $z \geq 0 \mapsto \Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ means the Gamma function. We refer the reader to the Appendix section for further details, see Lemmas 10 and 11 therein, and to [2] for additional examples. In contrast, a necessary condition for Assumption 3 to be satisfied is the absolute continuity of the measure P w.r.t. the Lebesgue measure, see Section 4 in [36]. The bounds stated in the theorem below reveal the accuracy of the statistical estimates (5) and (6) and highlight their behavior through explicit constants.

Theorem 4. *Suppose that the distribution* P *of the r.v.* X *is* τ *sub-Gaussian and satisfies Assumptions 1, 2, 3 and 4. The following assertions hold true.*

(i) For any $\delta \in \left(\max\{\Theta, 12.9^d\}\ e^{-\frac{n}{2}\min\{\alpha, \alpha^2, \alpha\Delta/8\}}, 1\right)$*, we have with probability at least* $1 - \delta$*:*

$$\sup_{x \in \mathbb{R}^d} \left| \widehat{D}_{\text{AI-IRW}}(x) - D_{\text{AI-IRW}}(x, P) \right| \leq \Delta \max_{s=1,2} \left(\frac{d + \log(2/\delta)}{n} \right)^{1/s} + \sqrt{\frac{8 \log(\Theta/\delta)}{n}},$$

where $\Delta = 512 L_R \tau^2 \max\{1/(\xi\varepsilon),\ 2\sqrt{2d}/(\gamma\varepsilon)\}$ *with* $\xi \in (0, \varepsilon)$*,* $\alpha = (\varepsilon - \xi)/(32\tau^2)$ *and* $\Theta = 12(2n)^{d+1}/(d+1)!$*.*

(ii) Let $r > 0$*. For any* $\delta \in \left(\max\{\Theta,\ 12.9^d\}\ e^{-n\min\{\alpha, \alpha^2, \alpha\Delta/8\}}, 1\right)$*, we have with probability at least* $1 - \delta$*:*

$$\sup_{x \in \mathcal{B}_r} \left| \widetilde{D}^{MC}_{\text{AI-IRW}}(x) - D_{\text{AI-IRW}}(x, P) \right| \leq \sqrt{\frac{128 \log(3\Theta/2\delta)}{9n}}$$

$$+2\sqrt{\frac{d\log(3rm)+\log(6/\delta)}{18m}}+\frac{4L_p}{3m}+\frac{8\Delta}{3}\max_{s=1,2}\left(\frac{d+\log(2/\delta)}{n}\right)^{1/s},$$

where the constants Θ, Δ, α *and the parameter* $\xi\in(0,\varepsilon)$ *are the same as those involved in* (i).

The upper confidence bound in assertion (i) is decomposed into two terms. The first term, of order $O(n^{-1/2})$, owes its presence to the replacement of $W^\top$ by its estimator. The second term, of order $O(\sqrt{\log(n)/n})$ and exhibiting a sublinear dependence in the dimension d, corresponds to the bound that would be obtained if $W^\top$ were known (it is then derived by means of the arguments used to study the concentration properties of the empirical halfspace depth, see chapter 26 in [46]). The upper confidence bound in assertion (ii) differs from that in assertion (i) in two respects. First, the additional terms clearly show the effect of the Monte Carlo approximation, which is negligible when $n\gg m$. Second, the maximal deviation is taken over a compact subset of $\mathbb{R}^d$. Furthermore, our theoretical analysis can be easily extended to the deviations of the sample version of IRW by simply omitting the term involving the square root of the precision matrix corresponding to the first term of (i) and the last term of (ii) leading to faster rates (see Section B.4 in the Appendix section).

Range of the confidence level The proof of the assertion (i) relies on controlling the deviations between the eigenvectors (resp., the inverses of the square-root eigenvalues) of $\widehat{\Sigma}$ and those of the true covariance matrix. The lower bound of the δ-range results from this control and is not limiting in practice since it decreases exponentially fast when the sample size increases.

About the constants Both upper bounds are provided with explicit constants. The explicit linear dependence on the dimension d is due to the operator norm that appears in the proof when controlling the eigenvectors of $\widehat{\Sigma}-\Sigma$. It implies an additional square root of d in the constant Δ following the classical inequality $||A||_{\text{op}}\leq\sqrt{d}||A||_1$ for any matrix $A\in\mathbb{R}^{d\times d}$ of full rank d. However, Lipschitz constants L_p and L_R, that are mandatory in order to derive bounds uniformly on $\mathbb{R}^d$ (or $\mathcal{B}_r$), appear to exhibit an implicit dependence on the dimension d. Indeed, these constants can be derived for r.v. valued in a compact support with bounded density exhibiting an exponential dependence on d. Unfortunately, this concern cannot be avoided unless removing the supremum involved in (i) and (ii). While the depth value at a single point $x\in\mathbb{R}^d$ is usually of limited importance, t is often more relevant in practice that an ensemble of depth values, i.e. the set $\{D(x,P),\ x\in\mathbb{R}^d\}$, are simultaneously well approximated by their empirical versions for comparison purposes. This implies estimation guarantees for the ranks induced by the depth function when computed on the whole sample $X_1,\ldots,X_n$, on which several applications such as anomaly detection fully rely on. The eigengap γ appears in the denominator due to the use of a variant of the Davis-Kahan theorem [11], so as to control the deviations between the eigenvectors of $\widehat{\Sigma}$ and those of Σ, and can not be

avoided. Observe that both upper-bounds explode as γ or ε vanish. These constants, related to the covariance matrix estimation, are often small in practice (see Section 5 where they are computed on real-world benchmarked datasets). However, they are often negligible w.r.t. the Lipschitz constant in the numerator that is $O(e^d)$ as mentioned above and is thus not limiting.

On optimality In absence of lower bound (and to the best of our knowledge, no such result is documented in the statistical depth literature yet), the optimality of the bounds above cannot be claimed of course. However, the proof partly consists in bounding the risk of the estimator of the covariance matrix Σ and involves the estimation rates given in Lemma 9 in the Appendix, which are known to be optimal for sub-Gaussian distributions [52]. It has been shown that faster rates for the estimation of the inverse of the covariance matrix can be established under additional sparsity assumptions (see e.g. Theorem 5 in [3]).

Choosing m The difficulty of approximating an integral over $\mathbb{R}^d$ by means of Monte-Carlo techniques grows with d. Our theoretical results, such as the upper bound in (ii), shed light on the behavior of m w.r.t. the dimension d. Indeed, focusing on the term $4L_p/(3m)$, L_p can be made explicit for density bounded distributions involving the volume of the unit sphere $\mathbb{S}^{d-1}$ that depends exponentially on d (see the paragraph above Theorem 4). Thus, m should be higher than $O(e^d)$ to yield a good statistical approximation. However, in practice, since computation times depend on m, a trade-off between statistical accuracy (the higher m, the better) and computational burden (the higher m, the heavier) must be found in practice, see Section 5. Regarding the terms in the upper-bound of (ii), if the dimension is high in comparison to the number of MC projections m, then $4L_p/(3m)$ will be the predominant term and the statistical error will be negligible compared to the approximation one, which may appear in practice. In contrast, if m is high enough compared to d, then the statistical and approximation errors will be close. Indeed, the constant in the first and second-term numerators behave similarly since $d\log(m) \approx \log(\Theta)$. The last term tends to be negligible compared to the previous two.

Remark 2 (Related work)**.** *We point out that non-asymptotic results about the accuracy of sample versions of statistical depths, such as those stated above, are seldom in the literature. To the best of our knowledge, rate bounds have only been derived in the halfspace depth case before. The first result (see [46] chapter 26), where uniform rates of the sample version are provided, uses the fact that the set of halfspaces in* $\mathbb{R}^d$ *is of finite* VC *dimension. Recently, this result has been refined under the Assumptions 3 and 4 in [2]. The convergence rate of the Tukey depth corresponds to that of the AI-IRW regarding the sample size. Asymptotic rates of convergence for the Monte Carlo approximation of the halfspace depth, i.e., when the minimum over the unit hypersphere is approximated from a finite number of directions, have been recently established in [36]. In contrast to the finite-sample framework, uniform asymptotic rates have been proved in several settings. Unfortunately, approximating a minimum over the*

unit sphere $\mathbb{S}^{d-1}$ using a Monte Carlo scheme is not optimal. Indeed when the distribution is assumed to belong to a bounded subset of $\mathbb{R}^d$ with bounded density, the authors obtain slow rates of order $O((\log(m)/m)^{1/(d-1)})$ suffering from the curse of dimensionality. Futhermore, they show that obtaining uniform rates of the halfspace depth approximation is not possible in absence of the bounded density assumption, see Section 4.2 in [36].

5. Numerical experiments

The advantages of the novel notion of depth introduced in Section 3 are supported by various experimental results in this part. First, we explore empirically the behavior of the returned ranks as the number m of sampled projections increases. A robust estimator of the AI-IRW is also introduced using the well-known Minimum Covariance Determinant (MCD) estimator [42] of the covariance matrix (used in the third line of the Approximation Algorithm 1). Second, the application of the AI-IRW depth to anomaly detection is considered, illustrating clearly the improvement on the performance attained.

5.1. On approximating the AI-IRW depth

The accuracy of Monte Carlo approximation is assessed for the empirical versions of the AI-IRW depth. In addition, robustness and computation time (see Section C.1) of the proposed approximations are investigated.

5.1.1. Accuracy of the AI-IRW approximation

The accuracy of Monte Carlo approximation, depending on the number m of random directions uniformly sampled, is evaluated for the empirical versions of the AI-IRW depth. The experiment is based on samples of size $n = 1000$ drawn from a centered Gaussian distribution with an identity covariance matrix and one sampled from a Wishart distribution (with parameters $(d, \mathcal{I}_d)$), where the dimension d varies in the range $\{2, 5, 10, 15, 20, 30, 40, 50\}$. We compute $\widetilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}$ and $\widetilde{D}^{\mathrm{MC}}_{\text{IRW}}$ on these samples by varying the number of projections m between 100 and 7000. Note that for an affine-invariant depth function, the depth contours of an elliptical distribution equal its density contours. As none of AI-IRW and IRW can be expressed by means of a closed analytical form, we propose to evaluate the quality of the returned ranks considering those of the density of sampled distribution as the "true" depth.

The coherence between ranks is assessed using the popular Kendall tau correlation coefficient, see [22]. The procedure is repeated ten times, and the averaged results are reported in Fig. 3. As expected, the approximation quality increases with m and decreases with d. Sharp approximations are obtained with far less than $O(e^d)$ projections for the three cases in the example involving a standard Gaussian distribution. We can notice that estimating the covariance matrix and

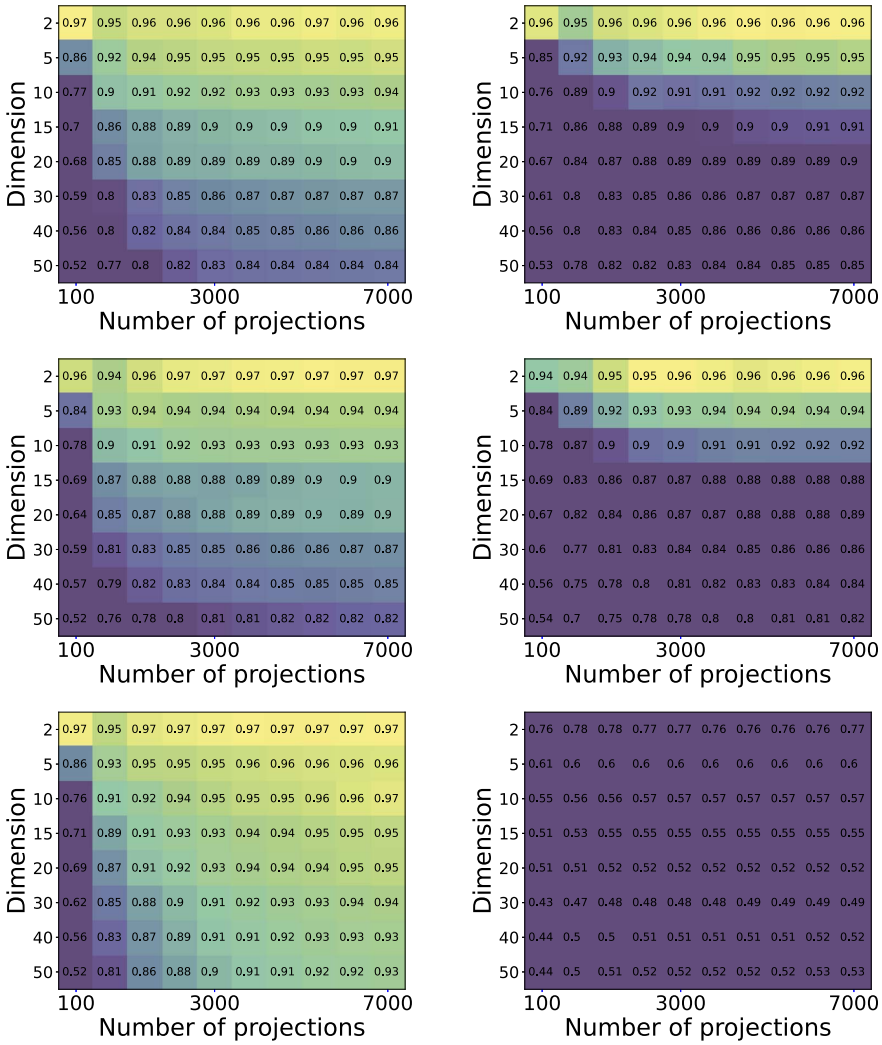

FIG 3. *Kendall correlation between population density ranks and the approximated ranks of AI-IRW using SC (top), MCD estimates (middle) and those of IRW (bottom), depending on the number of approximating projections m and the dimension d, for standard (left) and correlated (right) Gaussian distribution.*

inverting it numerically reduces slightly the convergence speed of the ranks as the dimension increases. Furthermore, sharp approximations are obtained with far less than $O(e^d)$ projections, also in the case of correlated Gaussian data. Indeed, in the worst case, i.e. when $d = 50$, a correlation of 0.93 is attained for AI-IRW, using both sample covariance (SC) and MCD (with support fraction set to $(n+d+1)/2$) estimators, with only 5500 directions which is roughly $100 \times d$ while $e^{50} \approx 10^{21}$. In low dimension, few projections are needed to obtain a corre-

lation higher than 0.98. Kendall correlations of IRW are close to those of AI-IRW with a slight advantage to the IRW depth as expected due to the presence of an additional covariance estimate term. In view of these results and because of the computation time of the approximations (documented in Section C.1 of the Appendix), choosing $m = 100 \times d$ appears as a good compromise between statistical accuracy and computation time, as done in the next experiments. All the computations are performed using a computer with 3.2 GHz Intel processor and 32 GB of RAM.

5.1.2. Robustness as the proportion of outliers increases

In this part, we examine the robustness of the ordering produced. It is based on the construction of two contaminated datasets from samples of size $n = 1000$ drawn from the multivariate standard Gaussian distribution (standard, so that affine non-invariant depths are not disadvantaged) in dimension $d = 2$. To build corrupted dataset, the two following contaminated models are used. The first is based on adding "isolated outliers" where each of them is defined as $(0, a)$ where a is sampled uniformly between $[4, 8]$. The second is based on adding "aggregated outliers" by randomly and uniformly drawing a location b in $[4, 8]$ and then drawing anomalies following the Gaussian distribution $\mathcal{N}(\mathbf{b}, \mathcal{I}_2)$ where $\mathbf{b}$ is the vector (b, b). Therefore, each dataset is constructed as follows: a proportion of outliers $\alpha \in [0, 0.15]$ is added to the normal data, represented by the standard Gaussian distribution, following one of the two aforementioned contamination models and thus yields two settings. The AI-IRW depth using SC and MCD estimators as well as the IRW depth are computed on these contaminated datasets, all with the number of Monte Carlo projections set to $m = 1000$. The Kendall tau distance is used to measure the deviation between the "true" ranks that are computed on samples without corruption and those computed on samples with corruption w.r.t. a proportion of anomalies α. The averaged Kendall tau's (over 100 runs) are displayed in Fig. 4.

As expected, results show that the MCD estimator provides robustness to the AI-IRW depth while the sample covariance estimator breaks down after only 1% of anomalies. Interestingly, the MCD estimator does not bring more robustness than the underlying robustness of the IRW depth. It highlights somehow a "worst case" robustness between the estimator of the covariance matrix and the underlying IRW depth which is reached by the latter. Therefore, we emphasize that AI-IRW, despite introducing affine invariance to IRW depth, does not enhance its robustness.

5.2. Application to anomaly detection

In this section, we study the performance of the proposed depth when dedicated to the anomaly detection task. First, we show that the affine invariance increases the ability to of AI-IRW to detect anomalies over IRW through a synthetic experiment. Second, we benchmark AI-IRW with on various real-world datasets.

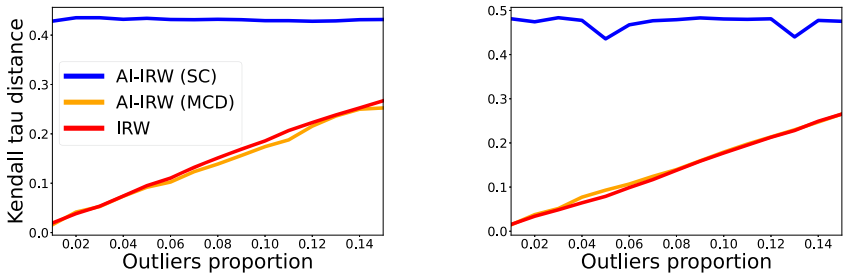

FIG 4. *Coherence of the returned rank measured by Kendall tau depending on outliers proportion for Gaussian distribution with isolated outliers (left) and aggregated outliers (right) for AI-IRW (using SC and MCD estimates) and IRW.*

5.2.1. Anomaly detection: a comparison on a toy dataset

This part compares the AI-IRW depth using the sample covariance estimator and the IRW depth regarding their performance for anomaly detection on a simulated dataset. To conduct this experiment, we build a toy-contaminated dataset (see Fig. 5, top) where five aggregated outliers (crossed points) and five isolated outliers (triangles) are added to 1000 points stemming from a 2-dimensional standard *Gaussian* distribution transformed adding the vector $u = (15, 55)$ and multiplying by the linear transformation described by matrix $A = \begin{bmatrix} 2 & 3 \\ 1 & 8 \end{bmatrix}$. Further, we compute AI-IRW (SC) and IRW depths on this dataset with $m = 10^5$ to reduce the depth approximation by Monte Carlo. The scores for the two benchmarked data depths, w.r.t. the index associated with the data, are depicted in Fig. 5 (bottom). Outliers are indexed from 1000 to 1010. Two dotted lines represent the lowest scores assigned by depth functions to normal data. The figure shows that while IRW fails to give the lowest score to these anomalies (most of the triangles and crosses are below the lowest normal score), AI-IRW succeeds in assigning the ten lowest scores to the ten anomalies (triangles and crosses are all below the lowest normal score). AI-IRW can assign the lowest depth to these anomalies, while IRW fails to identify them. Thus, the affine invariance strengthens the robustness and ability to detect anomalies even when using a non-robust covariance matrix estimator. It is worth noting that contrary to the experiment in Section 5.1.2, outliers are not that far from the normal distribution and thus do not deteriorate the estimation of $\widehat{\Sigma}$.

5.2.2. Benchmarking AI-IRW using real-world datasets

To illustrate the performance improvement due to introduction of affine invariance to the IRW, we conduct a comprehensive comparative study of anomaly detection on 10 widely used datasets in the literature[1]: *Mulcross*, *Shuttle*, *Thy-*

[1] http://odds.cs.stonybrook.edu/

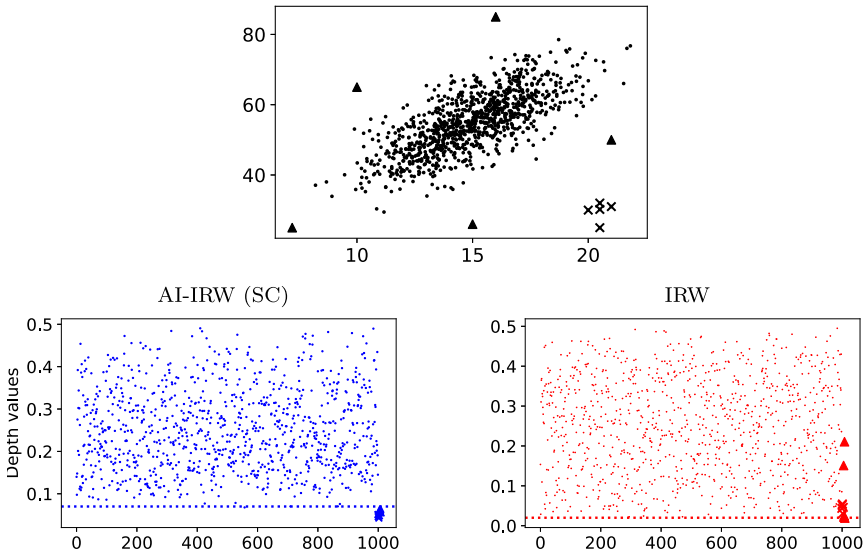

FIG 5. *Toy dataset (top) with aggregated outliers (crosses) and isolated outliers (triangles) and the associated AI-IRW (SC) and IRW depth values (bottom). Values are plotted according to their index in the data. The dotted line represents the lowest assigned score to normal data.*

TABLE 1

Information on datasets considered for the performance comparison: the number n of instances, the number d of attributes, $\hat{\gamma}$ and $\hat{\varepsilon}$ the eigengap and the smallest eigenvalue of the SC estimator respectively.

	n	d	% of anomaly	$\hat{\gamma}$ ($\times 0.01$)	$\hat{\varepsilon}$ ($\times 0.01$)
Ecoli	195	5	26	0.3	0.2
Shuttle	49097	9	7	9	5.7
Mulcross	262144	4	10	100	10^{-10}
Thyroid	3772	6	2.5	0.01	0.1
Wine	129	13	7.7	0.9	0.9
Http	567479	3	0.4	19	2.9
Smtp	95156	3	0.03	3.9	36
Breastw	683	9	35	80	20
Musk	3062	166	3.2	9.4	6
Satimage	5803	36	1.2	283	2.6

roid, *Wine*, *Http*, *Smtp*, *Ecoli*, *Breastw*, *Musk* and *Satimage* varying in size and dimension. Information of the benchmarked datasets such as the size, their percentage of anomalies and their estimated eigengap and smallest eigenvalues are given in Table 1. We place ourselves in the unsupervised setting. We train all methods on unlabeled data, and we use labels only to assess the performance of the methods by Area Under the Receiver Operation Characteristic curve (AUROC). We contrast the proposed approach with the affine non-invariant

version, the original halfspace depth (T), halfspace mass depth (HM; [6]), the AutoEncoder (AE; [1]) where the reconstruction error is used as anomaly score and one of the most used multivariate anomaly detection algorithms: Isolation Forest (IF; [27]). The related hyperparameters are set by default for simplicity. Based on the previous experiment, AI-IRW, IRW, and halfspace depths are calibrated with $m = 100 \times d$. From Table 2 one observes that AI-IRW uniformly and significantly in many cases improves on standard IRW. This is rather comparable with Isolation Forest and the halfspace mass depth. AI-IRW, IRW, HM and Tukey are implemented from scratch in python using `numpy` python library. Isolation Forest is implementated using `scikit-learn` python library [39] while the AutoEncoder implementation is based on`pyod` python library [57]. All the computations are made by means of a computer with 3.2 GHz Intel processor and 32 GB of RAM. The computation time used to perform the anomaly detection benchmark is displayed in Table 3.

TABLE 2
AUROCs of benchmarked anomaly detection methods.

	AI-IRW	IRW	HM	Tukey	IF	AE
Ecoli	0.85	0.83	**0.88**	0.68	0.77	0.64
Shuttle	0.99	0.99	0.99	0.86	0.99	0.99
Mulcross	**1**	0.98	**1**	0.87	0.96	**1**
Thyroid	**0.98**	0.80	0.84	0.92	0.97	0.97
Wine	0.96	0.96	**0.99**	0.71	0.8	0.72
Http	**1**	0.95	0.97	0.99	**1**	**1**
Smtp	**0.96**	0.77	0.74	0.85	0.90	0.82
Breastw	0.97	0.97	**0.99**	0.84	**0.99**	0.91
Musk	**1**	0.84	0.97	0.77	**1**	**1**
Satimage	**0.99**	0.96	0.98	0.95	**0.99**	0.98

TABLE 3
Computation time of benchmarked anomaly detection methods in seconds.

	AI-IRW	IRW	HM	T	IF	AE
Ecoli	0.04	0.005	0.02	0.005	0.13	9
Shuttle	20	6.8	1.5	6.8	1.4	469
Mulcross	75	27	6.2	27	5.9	2383
Thyroid	1	0.21	0.2	0.2	0.18	42
Wine	0.05	0.01	0.06	0.008	0.12	8.1
Http	97	45	11	45	11	5197
Smtp	22	4.5	1	4.5	1.88	903
Breastw	0.46	0.04	0.06	0.04	0.14	17.2
Musk	20.5	5.23	2.5	5.2	0.43	103
Satimage	6.3	2.63	0.9	2.6	0.31	76

6. Conclusion

In this paper, we have introduced a novel notion of statistical depth (AI-IRW), modifying the original Integrated Rank-Weighted (IRW) depth proposed in [40]. It has been shown that the AI-IRW depth does not only inherit all the compelling features of the IRW depth, its theoretical properties and its computational advantages (no optimization problem solving is required to compute it), but also fulfills in addition the affine invariance property, crucial regarding interpretability issues. The natural idea at work consists in averaging univariate Tukey halfspace depths computed from random projections of the data onto (nearly) uncorrelated lines, defined by the (empirical) covariance structure of the data, rather than projections onto lines fully generated at random. Though the AI-IRW sample version exhibits a complex probabilistic structure, an estimator of the precision matrix being involved in its definition, a non-asymptotic analysis has been carried out here, revealing its good concentration properties around the true AI-IRW depth. The merits of the AI-IRW depth have been illustrated by encouraging numerical experiments, for anomaly detection in particular, offering the perspective of a widespread use for various statistical learning tasks.

This Appendix is organized as follows.

- Useful preliminary results are stated and proved in Appendix A.
- The proofs of the results stated in the paper are given in Appendix B.
- Additional experiments are displayed in Appendix C.

Appendix A: Preliminary results

First, we recall some lemmas on linear algebra, halfspace depth and covariance matrix estimation, used in the subsequent proofs.

A.1. Basics of linear algebra

Here useful results of linear algebra are recalled for clarity.

Lemma 5 ([53], Theorem 4.1)**.** *Let A and B be two invertible matrices of size $d \times d$ and $||A||_{op}$ be the operator norm of matrix A. Then it holds:*

$$||A^{-1} - B^{-1}||_{\text{op}} \leq ||A^{-1}||_{\text{op}}||B^{-1}||_{\text{op}}||A - B||_{\text{op}}. \tag{7}$$

Lemma 6 ([52], Lemma 2.2)**.** *Let A be a matrix of size $d \times d$ and N_ρ be an ρ-net of $\mathbb{S}^{d-1}$. Then it holds:*

$$||A||_{\text{op}} \leq \frac{1}{1-2\rho} \max_{v \in N_\rho} |v^\top A v|.$$

Lemma 7. *Let A_1 and A_2 be two real symmetric and invertible matrices of dimension $d \times d$ with $O_1 D_1 O_1^\top$ and $O_2 D_2 O_2^\top$ their eigenvalues decomposition in orthornormal bases. Denotes $W_1 = D_1^{-1/2} O_1^\top$ and $W_2 = D_2^{-1/2} O_2^\top$. Then it holds:*

$$||W_1 - W_2||_{\text{op}} \leq ||D_2^{-1/2}||_{\text{op}} \left(||D_1^{1/2} - D_2^{1/2}||_{\text{op}} \, ||D_1^{-1/2}||_{\text{op}} + ||O_1 - O_2||_{\text{op}} \right).$$

Proof.

$$\begin{aligned} ||W_1 - W_2||_{\text{op}} &\leq ||O_1||_{\text{op}} \, ||D_1^{-1/2} - D_2^{-1/2}||_{\text{op}} + ||O_1 - O_2||_{\text{op}} \, ||D_2^{-1/2}||_{\text{op}} \\ &\leq ||D_1^{-1/2} - D_2^{-1/2}||_{\text{op}} + ||D_2^{-1/2}||_{\text{op}} ||O_1 - O_2||_{\text{op}} \\ &\overset{(i)}{\leq} ||D_2^{-1/2}||_{\text{op}} \left(||D_1^{1/2} - D_2^{1/2}||_{\text{op}} \, ||D_1^{-1/2}||_{\text{op}} + ||O_1 - O_2||_{\text{op}} \right), \end{aligned}$$

where (i) holds due to Lemma 5. □

A.2. Non-asymptotic rates on halfspace depth and sample covariance matrix

We now recall useful results on maximum deviations of the halfspace depth estimator as well as the sample covariance matrix.

Lemma 8 ([46], Chapter 26)**.** *Let $P \in \mathcal{P}(\mathbb{R}^d)$. Let $X_1, \ldots, X_n$ a sample from P with empirical measure $\widehat{P} = (1/n) \sum_{i=1}^n \delta_{X_i}$. Denote by F_u and $\widehat{F}_u$ the cdf of P_u and $\widehat{P}_u$ respectively. Then, for any $t > 0$, it holds:*

$$\mathbb{P}\left(\sup_{\substack{x \in \mathbb{R}^d \\ u \in \mathbb{S}^{d-1}}} \left| \widehat{F}_u(u^\top x) - F_u(u^\top x) \right| > t \right) \leq \frac{6(2n)^{d+1}}{(d+1)!} \exp(-nt^2/8).$$

Lemma 9 (Variant of [52], Proposition 2.1)**.** *Let Σ be the covariance matrix of a τ sub-Gaussian random variables X that takes its values in $\mathbb{R}^d$. Let $X_1 \ldots, X_n$ be a sample from X and denote by $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$ the SC estimator of Σ. Then it holds:*

$$\mathbb{P}\left(||\widehat{\Sigma} - \Sigma||_{\text{op}} > t \right) \leq 2 \times 9^d \exp\left\{ -\frac{n}{2} \min\left\{ \frac{t^2}{(32\tau^2)^2}, \frac{t}{32\tau^2} \right\} \right\}.$$

Let $\sigma_d \geq \cdots > \sigma_1$ and $\widehat{\sigma}_d > \cdots > \sigma_1$ be respectively the ordered eigenvalues of Σ and $\widehat{\Sigma}$. Using Weyl's Theoreom [54], it holds:

$$\mathbb{P}\left(\max_{1 \leq k \leq d} |\widehat{\sigma}_k - \sigma_k| > t \right) \leq 2 \times 9^d \exp\left\{ -\frac{n}{2} \min\left\{ \frac{t^2}{(32\tau^2)^2}, \frac{t}{32\tau^2} \right\} \right\}.$$

Proof. Let N_ρ be an ρ-net of the sphere $\mathbb{S}^{d-1}$. Applying Lemma 6 on $\widehat{\Sigma} - \Sigma$, for any $t, \rho > 0$, we have

$$\mathbb{P}\left(||\widehat{\Sigma} - \Sigma||_{\text{op}} > t \right) \leq \mathbb{P}\left(\frac{1}{1 - 2\rho} \max_{v \in N_\rho} |v^\top (\widehat{\Sigma} - \Sigma) v| > t \right)$$

$$\leq |N_\rho| \, \mathbb{P}\left(|v^\top (\widehat{\Sigma} - \Sigma) v| > (1 - 2\rho) \, t \right),$$

where $|N_\rho|$ stands for the cardinal of the set N_ρ. Noticing that $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$ is a sum of independent matrices we have

$$v^\top (\widehat{\Sigma} - \Sigma) v = \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E} Z_i,$$

where $Z_i = (v^\top X_i)^2$ for every $1 \leq i \leq n$ and $Z_i - \mathbb{E} Z_i$ are i.i.d random variables that are $((16\tau)^2, 16\tau^2)$ sub-exponential.

Choosing $\rho = 1/4$, noticing that $N_{1/4} \leq 9^d$ and applying the sub-exponential tail bound lead to the desired result. □

Appendix B: Technical proofs of the main results

We now prove the main results stated in the paper.

B.1. Proof of Proposition 3

B.1.1. Affine invariance

Let $A \in \mathbb{R}^{d \times d}$ be a non-singular matrix and $b \in \mathbb{R}^d$. Let Σ_X and Σ_{AX} the covariance matrix of X and AX respectively. Define the Cholesky decomposition as $\Sigma_X = \Lambda_X \Lambda_X^\top$ and $\Sigma_{AX} = A \Lambda_X \Lambda_X^\top A^\top = \Lambda_{AX} \Lambda_{AX}^\top$. It holds:

$$\begin{aligned}
&D_{\text{AI-IRW}}(Ax + b, AX + b) \\
&= \frac{1}{V_d} \int_{\mathbb{S}^{d-1}} D_{\text{H},1}(\langle \frac{\Lambda_{AX+b}^{-\top} u}{||\Lambda_{AX+b}^{-\top} u||}, Ax + b \rangle, \, \langle \frac{\Lambda_{AX+b}^{-\top} u}{||\Lambda_{AX+b}^{-\top} u||}, AX + b \rangle) \, \mathrm{d}u \\
&= \frac{1}{V_d} \int_{\mathbb{S}^{d-1}} D_{\text{H},1}(\langle \Lambda_{AX+b}^{-\top} u, Ax + b \rangle, \, \langle \Lambda_{AX+b}^{-\top} u, AX + b \rangle) \, du \\
&= \frac{1}{V_d} \int_{\mathbb{S}^{d-1}} D_{\text{H},1}(\langle \Lambda_{AX}^{-\top} u, Ax \rangle, \, \langle \Lambda_{AX}^{-\top} u, AX \rangle) \, \mathrm{d}u \\
&= \frac{1}{V_d} \int_{\mathbb{S}^{d-1}} D_{\text{H},1}(\langle u, \Lambda_X^{-1} x \rangle, \langle u, \Lambda_X^{-1} X \rangle) \, \mathrm{d}u \\
&= \frac{1}{V_d} \int_{\mathbb{S}^{d-1}} D_{\text{H},1}(\langle \frac{\Lambda_X^{-\top} u}{||\Lambda_X^{-\top} u||}, x \rangle, \langle \frac{\Lambda_X^{-\top} u}{||\Lambda_X^{-\top} u||}, X \rangle) \, \mathrm{d}u \\
&= D_{\text{AI-IRW}}(x, P).
\end{aligned}$$

The same reasoning applies for any whitening matrices.

B.1.2. Proving maximality at the center

Assume that P is halfspace symmetric about a unique β, i.e., $\mathbb{P}(X \in \mathcal{H}_\beta) \geq \frac{1}{2}$ for every closed halfspace $\mathcal{H}_\beta$ such that $\beta \in \partial \mathcal{H}$ with $\partial \mathcal{H}$ the boundary of $\mathcal{H}$.

Thus, it is easy to see that $D_{\text{AI-IRW}}(\beta, P) \geq \frac{1}{2}$. The uniqueness of β and the fact that $D_{\text{AI-IRW}}$ is lower than $1/2$ for any element in $\mathbb{R}^d$ by definition imply that

$$\beta = \operatorname*{argsup}_{x \in \mathbb{R}^d} D_{\text{AI-IRW}}(x, P).$$

B.1.3. Vanishing at infinity

The proof is a particular case of the proof of theorem 1 in [10]. We detail it for the sake of clarity. Let U be a random variable following ω_{d-1}, the uniform measure on the unit sphere $\mathbb{S}^{d-1}$. Defines $V = W^\top U/||W^\top U||$ and ν_{d-1} its probability distribution. Let $\theta > 0$ and $x \in \mathbb{R}^d$, then $r(\theta) := \nu_{d-1}\{v : \ \frac{|\langle v,x\rangle|}{||x||} \leq \theta\}$ goes to zero when $\theta \to 0$. For any $x \in \mathbb{R}^d \backslash \{0\}$, we have

$$\begin{aligned}
D_{\text{AI-IRW}}(x, P) &= \int_{\mathbb{R}^d} \min\left\{F_v(v^\top x), 1 - F_v(v^\top x)\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&\leq \int_{\mathbb{R}^d} \mathbb{I}\left\{v : \ \frac{|\langle v, x\rangle|}{||x||} \leq \theta\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&+ \int_{\mathbb{R}^d} F_v(v^\top x) \, \mathbb{I}\left\{v : \ \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle \leq 0\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&+ \int_{\mathbb{R}^d} (1 - F_v(v^\top x)) \, \mathbb{I}\left\{v : \ \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle > 0\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&\leq r(\theta) + \int_{\mathbb{R}^d} F_v(-\theta||x||) \, \mathbb{I}\left\{v : \ \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle \leq 0\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&+ \int_{\mathbb{R}^d} (1 - F_v(\theta||x||)) \, \mathbb{I}\left\{v : \ \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle > 0\right\} \, \mathrm{d}\nu_{d-1}(v).
\end{aligned}$$

Now, when $||x|| \to \infty$, the dominated convergence theorem ensures that

$$\lim \sup_{||x|| \to \infty} D_{\text{AI-IRW}}(x, P) \leq r(\theta) \underset{\theta \to 0}{\to} 0.$$

B.1.4. Decreasing along rays

The proof is a slight modification of the proof of Assertion (iii) of Theorem 2 in [40]. Details are left to the reader.

B.1.5. Continuity

For any $P \in \mathcal{P}(\mathbb{R}^d)$, the continuity of the inner product and the cdf ensure continuity of $D_{\text{H},1}(v^\top x, v^\top X)$ for any $v \in \mathbb{S}^{d-1}$. Therefore, the continuity of $x \mapsto D_{\text{AI-IRW}}(x, P)$ follows from dominated convergence.

B.2. Proof of Theorem 4

We now prove the main results of the paper. Defines the following SVD decomposition of the covariance matrix $\Sigma = ODO^\top$. To derive our results, we set our whitening matrix as $W = D^{-1/2}O^\top$. Our results holds true for any whitening matrix.

B.2.1. Assertion (i)

Introducing terms, using the fact that $z \mapsto \min(z, 1-z)$ is 1-Lipschitz and using triangle inequality, it holds:

$$\sup_{x\in\mathbb{R}^d} \left|\widehat{D}_{\text{AI-IRW}}(x) - D_{\text{AI-IRW}}(x,P)\right| \leq \underbrace{\sup_{x\in\mathbb{R}^d} \mathbb{E}\left|\widehat{F}_{\widehat{V}}(\widehat{V}^\top x) - F_{\widehat{V}}(\widehat{V}^\top x)\right|}_{(1)} + \underbrace{\sup_{x\in\mathbb{R}^d} \mathbb{E}\left|F_{\widehat{V}}(\widehat{V}^\top x) - F_V(V^\top x)\right|}_{(2)}.$$

Now, the first term (1) can be controlled using the bound for the deviations of halfspace depth deferred in Lemma 8. Thus, for any $t > 0$ it holds:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathbb{R}^d} \mathbb{E}\left|\widehat{F}_{\widehat{V}}(\widehat{V}^\top x) - F_{\widehat{V}}(\widehat{V}^\top x)\right| > t/2\right) \\
&\leq \mathbb{P}\left(\sup_{\substack{y\in\mathbb{R}^d \\ u\in\mathbb{S}^{d-1}}} \left|\widehat{F}_u(u^\top y) - F_u(u^\top y)\right| > t/2\right) \\
&\leq \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32). \qquad (8)
\end{aligned}$$

The second term (2) relies on the influence of the deviations of the sample covariance matrix. First remark that:

$$\begin{aligned}
\sup_{x\in\mathbb{R}^d} \mathbb{E}\left|F_{\widehat{V}}(\widehat{V}^\top x) - F_V(V^\top x)\right| \leq \sup_{\substack{x\in\mathbb{R}^d \\ u\in\mathbb{S}^{d-1}}} \Bigg|&\mathbb{P}\left(\left\langle \frac{\widehat{W}^\top u}{||\widehat{W}^\top u||}, X - x\right\rangle \leq 0 \;\middle|\; \mathcal{S}_n\right) \\
&- \mathbb{P}\left(\left\langle \frac{W^\top u}{||W^\top u||}, X - x\right\rangle \leq 0\right)\Bigg|.
\end{aligned}$$

Now, since X is radially Lipschitz continuous, we have:

$$\begin{aligned}
&\left|\mathbb{P}\left(\left\langle \frac{\widehat{W}^\top u}{||\widehat{W}^\top u||}, X - x\right\rangle \leq 0 \;\middle|\; \mathcal{S}_n\right) - \mathbb{P}\left(\left\langle \frac{W^\top u}{||W^\top u||}, X - x\right\rangle \leq 0\right)\right| \\
&\quad \leq L_R \left|\left| \frac{\widehat{W}^\top u}{||\widehat{W}^\top u||} - \frac{W^\top u}{||W^\top u||}\right|\right|.
\end{aligned}$$

Introducing terms and using triangle inequality leads to:

$$\left|\left|\frac{\widehat{W}^\top u}{||\widehat{W}^\top u||} - \frac{W^\top u}{||W^\top u||}\right|\right| \leq \frac{||\widehat{W} - W||_{\text{op}}}{||Wu||} + ||\widehat{W}u|| \left(\frac{1}{||\widehat{W}u||} - \frac{1}{||Wu||}\right)$$
$$\leq \frac{2||\widehat{W} - W||_{\text{op}}}{||Wu||},$$

yielding:

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\left|F_{\widehat{V}}(\widehat{V}^\top x) - F_V(V^\top x)\right| \leq \frac{2L_R}{\sqrt{\varepsilon}}\, ||\widehat{W} - W||_{\text{op}}. \tag{9}$$

Assume that $ODO^\top$ and $\widehat{O}\widehat{D}\widehat{O}^\top$ are the eigenvalues decomposition of Σ and $\widehat{\Sigma}$ in orthonormal bases. Thus, thanks to Lemma 7, we have:

$$||\widehat{W} - W||_{\text{op}} \leq ||D^{-1/2}||_{\text{op}} \left(||\widehat{D}^{1/2} - D^{1/2}||_{\text{op}}\, ||\widehat{D}^{-1/2}||_{\text{op}} + ||\widehat{O} - O||_{\text{op}}\right).$$

Now, since $\min_{k\leq d} \sqrt{\widehat{\sigma}_k} \geq \sqrt{\varepsilon} - \max_{k\leq d} |\sqrt{\widehat{\sigma}_k} - \sqrt{\sigma}_k|$ and $\max_{k\leq d} |\sqrt{\widehat{\sigma}_k} - \sqrt{\sigma}_k| \leq \frac{1}{\sqrt{\varepsilon}} \max_{1\leq k\leq d} |\widehat{\sigma}_k - \sigma_k|$, using Weyl's inequality leads to:

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\left|F_{\widehat{V}}(\widehat{V}^\top x) - F_V(V^\top x)\right| \leq \frac{2L_R}{\varepsilon}\left(\frac{||\widehat{\Sigma} - \Sigma||_{\text{op}}}{\varepsilon - ||\widehat{\Sigma} - \Sigma||_{\text{op}}} + ||\widehat{O} - O||_{\text{op}}\right).$$

Let $\mathcal{A}_\xi = \left\{||\widehat{\Sigma} - \Sigma||_{\text{op}} < \varepsilon - \xi\right\}$ for any $\xi \in [0, \varepsilon)$. Using union bound and combining (9) with the previous equation, for any $t > 0$ and $\xi \in (0, \varepsilon)$ it holds:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^d} \mathbb{E}\left|F_{\widehat{V}}(\widehat{V}^\top x) - F_V(V^\top x)\right| > t/2\right) \leq \mathbb{P}\left(\frac{2L_R}{\xi\varepsilon}||\widehat{\Sigma} - \Sigma||_{\text{op}} > t/4\right)$$
$$+ \mathbb{P}\left(\frac{2L_R}{\varepsilon}\, ||\widehat{O} - O||_{\text{op}} > t/4\right)$$
$$+ \mathbb{P}\left(\mathcal{A}_\xi^c\right),$$

where $\mathcal{A}_\xi^c$ stands for the complementary event of $\mathcal{A}_\xi$. Applying Lemma 9 gives:

$$\mathbb{P}\left(||\widehat{\Sigma} - \Sigma||_{\text{op}} > \frac{\xi\varepsilon t}{8L_R}\right) \leq 2 \times 9^d \exp\left\{-\frac{n}{2}\min\left\{\frac{(\xi\varepsilon t)^2}{(256L_R\tau^2)^2}, \frac{\xi\varepsilon t}{256L_R\tau^2}\right\}\right\}, \tag{10}$$

and

$$\mathbb{P}\left(\mathcal{A}_\xi^c\right) \leq 2 \times 9^d \exp\left\{-\frac{n}{2}\min\left\{\frac{(\varepsilon - \xi)^2}{(32\tau^2)^2}, \frac{\varepsilon - \xi}{32\tau^2}\right\}\right\}. \tag{11}$$

Furthermore, it is easy to see that $||\widehat{O} - O||_{\text{op}} \leq \sqrt{d} \max_{k\leq d} ||\widehat{O}_k - O_k||$ where O_k is the k-th column of the matrix O. Let γ be the minimum eigengap, following a variant of the Davis-Kahan theorem [11] (see Corollary 1 in [55]), it holds:

$$||\widehat{O} - O||_{\text{op}} \leq \frac{2\sqrt{2d}||\widehat{\Sigma} - \Sigma||_{\text{op}}}{\gamma}.$$

Using Lemma 9 again leads to:

$$\mathbb{P}\left(\frac{4L_R\sqrt{2d}||\widehat{\Sigma}-\Sigma||_{\text{op}}}{\gamma\varepsilon} > t/4\right) \leq 2\times 9^d \exp\left\{-\frac{n}{2}\min\left\{\frac{(\gamma\varepsilon t)^2}{(512L_R\sqrt{2d}\tau^2)^2}, \frac{\gamma\varepsilon t}{512L_R\sqrt{2d}\tau^2}\right\}\right\}. \tag{12}$$

Combining (10), (11) and (12) it holds:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^d}\mathbb{E}\left|F_{\widehat{V}}(\widehat{V}^\top x)-F_V(V^\top x)\right| > t/2\right) \leq 6\times 9^d\exp\left(-\frac{n}{2}\min\left\{(\beta t)^2,\beta t\right\}\right),$$

for any $t \leq (\varepsilon-\xi)/(32\tau^2\beta)$ where $\beta = \frac{\varepsilon}{256L_R\tau^2}(\xi\wedge\frac{\gamma}{2\sqrt{2d}})$. Finally, for any $t\leq(\varepsilon-\xi)/(32\tau^2\beta)$ it holds:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\text{AI-IRW}}(x)-D_{\text{AI-IRW}}(x,P)\right| > t\right) \leq 6.9^d\exp\left(-\frac{n}{2}\min\left\{(\beta t)^2,\beta t\right\}\right) + \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32). \tag{13}$$

Bounding each term in the right side by $\delta/2$ and reverting the equation lead to the desired result.

B.2.2. Assertion (ii)

Let $\mathcal{B}_r$ a centered ball of $\mathbb{R}^d$ with radius $r>0$ and assume that X satisfies assumption 2 for any $x\in\mathcal{B}_r$. Introducing terms and using triangle inequality, it holds:

$$\sup_{x\in\mathcal{B}_r}\left|\widetilde{D}^{\text{MC}}_{\text{AI-IRW}}(x)-D_{\text{AI-IRW}}(x,P)\right| \leq \underbrace{\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\text{AI-IRW}}(x)-D_{\text{AI-IRW}}(x,P)\right|}_{(1)} + \underbrace{\sup_{x\in\mathcal{B}_r}\left|D^{\text{MC}}_{\text{AI-IRW}}(x,P)-D_{\text{AI-IRW}}(x,P)\right|}_{(2)}.$$

The first term (1) can be bounded using assertion (i) while controlling the approximation term (2) relies on classical chaining arguments. As the function $z\mapsto\min(z,1-z)$ is 1-Lipschitz for any $z\in(0,1)$ and by triangle inequality, for any y in $\mathcal{B}_r$ we have:

$$\left|D^{\text{MC}}_{\text{AI-IRW}}(y,P)-D_{\text{AI-IRW}}(y,P)\right| \leq \frac{1}{m}\sum_{j=1}^m\Big|\mathbb{P}\left\{\langle V_j,X-y\rangle\leq 0\mid V_j\right\} - \mathbb{E}\left[\mathbb{P}\left\{\langle V_j,X-y\rangle\leq 0\right\}\right]\Big|.$$

Since it is an average of bounded and i.i.d random variables, combining Hoeffding inequality and union bound, for any $t > 0$ and any y in $\mathcal{B}_r$ it holds:

$$\mathbb{P}\left(\left|D^{\mathrm{MC}}_{\text{AI-IRW}}(y,P) - D_{\text{AI-IRW}}(y,P)\right| \ > \ t/2\right) \leq 2\exp\left(-mt^2/2\right). \tag{14}$$

As X is uniformly continuous Lipschitz in projection for any $u \in \mathbb{S}^{d-1}$, observing that $\forall (x,y) \in \mathcal{B}_r^2$, it holds:

$$\begin{aligned}\left|D^{\mathrm{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| &\leq \left|D^{\mathrm{MC}}_{\text{AI-IRW}}(x,P) - D^{\mathrm{MC}}_{\text{AI-IRW}}(y,P)\right| \\ &\quad + \left|D^{\mathrm{MC}}_{\text{AI-IRW}}(y,P) - D_{\text{AI-IRW}}(y,P)\right| \\ &\quad + \left|D_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(y,P)\right| \\ &\leq 2L_p\ ||x-y|| + \left|D^{\mathrm{MC}}_{\text{AI-IRW}}(y,P) - D_{\text{AI-IRW}}(y,P)\right|.\end{aligned} \tag{15}$$

Now let $\zeta > 0$ and $y_1, \ldots, y_{\mathcal{N}(\zeta,\mathcal{B}_r,||.||_2)}$ be a ζ-coverage of $\mathcal{B}_r$ with respect to $||.||_2$. We have:

$$\log\left(\mathcal{N}(\zeta,\mathcal{B}_r,||.||_2)\right) \leq d\log\left(3r/\zeta\right). \tag{16}$$

Set $\mathcal{N} = \mathcal{N}\left(\zeta,\mathcal{B}_r,||.||_2)\right)$ for simplicity. There exists $\ell \leq \mathcal{N}$ such that $||x - y_\ell||_2 \leq \zeta$. Thus, (15) leads to

$$\left|D^{\mathrm{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| \leq 2L_p\ \zeta + \left|D^{\mathrm{MC}}_{\text{AI-IRW}}(y_\ell,P) - D_{\text{AI-IRW}}(y_\ell,P)\right|.$$

Applying (14) to every y_ℓ and the union bound, for any $t > 0$, we get:

$$\mathbb{P}\left(\sup_{\ell \leq \mathcal{N}} \left|D^{\mathrm{MC}}_{\text{AI-IRW}}(y_\ell,P) - D_{\text{AI-IRW}}(y_\ell,P)\right| > t/2\right) \leq 2\mathcal{N}\exp\left(-mt^2/2\right),$$

yielding:

$$\mathbb{P}\left(\sup_{x\in\mathcal{B}_r} \left|D^{\mathrm{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| > t/2\right) \leq 2\mathcal{N}\exp\left(-2m\left(t/2 - 2L_p\zeta\right)^2\right).$$

Using (13), the union bound and (16), we obtain:

$$\begin{aligned}&\mathbb{P}\left(\sup_{x\in\mathcal{B}_r} \left|\widetilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(x) - D_{\text{AI-IRW}}(x,P)\right| \ > \ t\right) \\ &\leq \mathbb{P}\left(\sup_{x\in\mathcal{B}_r} \left|\widehat{D}_{\text{AI-IRW}}(x) - D_{\text{AI-IRW}}(x,P)\right| \ > \ t/2\right) \\ &\quad + \mathbb{P}\left(\sup_{x\in\mathcal{B}_r} \left|D^{\mathrm{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| \ > \ t/2\right) \\ &\leq 6.9^d \exp\left(-\frac{n}{2}\min\left\{\left(\beta t/2\right)^2, \beta t/2\right\}\right) + \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/128)\end{aligned}$$

$$+ 2\left(\frac{3r}{\zeta}\right)^d \exp\left(-2m\left(t/2 - 2L_p\zeta\right)^2\right).$$

Choosing $\zeta \sim m^{-1}$, bounding each term on the right-hand side by $\delta/3$ and reverting the previous equation lead to the desired result.

B.3. Geometrical results on the Lipschitz constants involved in Assumptions 3 and 4

Lemma 10. *Let $r > 0$ and denote by $V_{d,r}$ the volume of the d-ball $\mathcal{B}(0,r)$. Assume that X takes its values in $\mathcal{B}(0,r)$ and has an M- bounded density w.r.t. the Lebesgue measure λ. The r.v. X is uniformly* Radially Lipschitz Continuous *with constant $L_R = MV_{d,r}$.*

Proof. Let $x \in \mathbb{R}^d$. By $||.||_g$ it means the geodesic norm on the unit sphere of $\mathbb{R}^d$. It holds:

$$\begin{aligned}
&|\phi(u,x) - \phi(v,x)| \\
&\leq \mathbb{P}\left\{X \in \mathcal{B}(0,r) : \langle u, X - x\rangle \text{ and } \langle v, X - x\rangle \text{ are of opposite sign}\right\} \\
&\leq M\,\lambda\left\{z \in \mathcal{B}(-x,r) : \langle u, z\rangle \text{ and } \langle v, z\rangle \text{ are of opposite sign}\right\} \\
&\overset{(i)}{\leq} M\,V_{d,r} \times \frac{2}{\pi}\arccos\left(\langle u, v\rangle\right) \\
&= M\,V_{d,r} \times \frac{2}{\pi}\,||u - v||_g \\
&\leq M\,V_{d,r}\,||u - v||,
\end{aligned}$$

where (i) arises from the fact that the volume of $\mathcal{E}_{u,x,y} = \{z \in \mathcal{B}(-x,r) : \langle u, z\rangle$ and $\langle v, z\rangle$ are of opposite sign$\}$ is the volume of two cones of angle $||u - v||_g$, as depicted in Fig. 6. □

Lemma 11. *Let $r > 0$ and assume that X takes its values in $\mathcal{B}(0,r)$ and has M-bounded density w.r.t. the Lebesgue measure λ. Thus X is uniformly* Lipschitz continuous in projection *with constant $L_p = MV_{d-1,r}$.*

Proof. Let $u \in \mathbb{S}^{d-1}$. By $||.||_g$ it means the geodesic norm on the unit sphere of $\mathbb{R}^d$. It holds:

$$\begin{aligned}
&|\phi(u,x) - \phi(u,y)| \\
&\leq \mathbb{P}\left\{X \in \mathcal{B}(0,r) : \langle u, X - x\rangle \text{ and } \langle u, X - y\rangle \text{ are of opposite sign}\right\} \\
&\leq M\,\lambda\left\{z \in \mathcal{B}(0,r) : \langle u, z - x\rangle \text{ and } \langle u, z - y\rangle \text{ are of opposite sign}\right\} \\
&\overset{(i)}{\leq} M\,V_{d-1,r} \times |\langle u, x\rangle - \langle u, y\rangle| \\
&\leq M\,V_{d-1,r}\,||x - y||,
\end{aligned}$$

where (i) arises from the fact that we encompass $\mathcal{F}_{u,x,y}$ by an hyper-cylinder of length $|\langle u, x\rangle - \langle u, y\rangle|$ where $\mathcal{F}_{u,x,y} = \{z \in \mathcal{B}(0,r) : \langle u, z - x\rangle$ and $\langle u, z - y\rangle$ are of opposite sign$\}$, as illustrated in Fig. 7. □

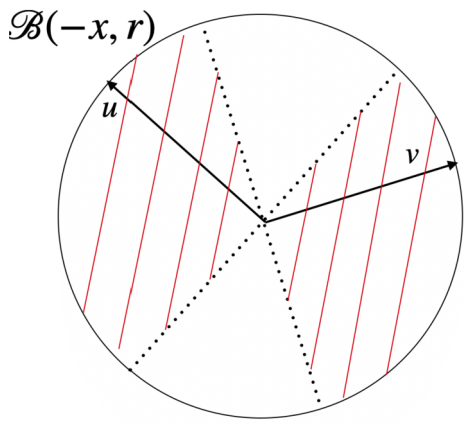

FIG 6. *Illustration of the set* $\mathcal{E}_{u,x,y}$ *in* $\mathbb{R}^2$. *It corresponds to the portion of* $\mathcal{B}(-x, r)$ *hatched in red.*

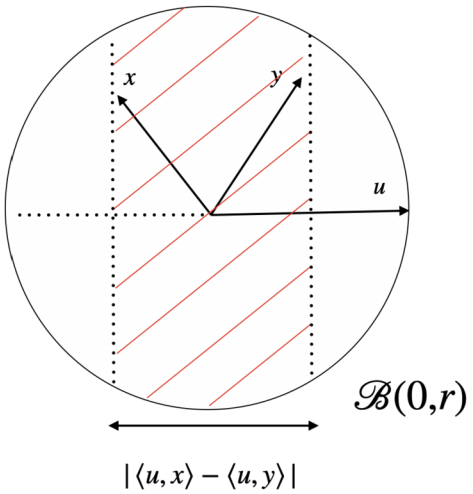

FIG 7. *Illustration of the set* $\mathcal{F}_{u,x,y}$ *in* $\mathbb{R}^2$. *It corresponds to the portion of* $\mathcal{B}(0, r)$ *hatched in red.*

B.4. Finite-sample analysis of the IRW depth

A finite sample analysis on IRW can be derived from our results on AI-IRW as it is described in the next corollary.

Corollary 12. *Suppose that the distribution* P *of the r.v.* X *satisfies Assumptions 3 and 4. Then, for any* $\delta \in (0, 1)$, *it holds:*

$$\sup_{x \in \mathcal{B}_r} \left| \widehat{D}_{IRW}^{MC}(x) - D_{IRW}(x, P) \right| \leq \sqrt{\frac{8 \log(\Theta/\delta)}{n}} + 2\sqrt{\frac{d \log(3rm) + \log(6/\delta)}{8m}} + \frac{2L_p}{m},$$

where $\Theta = 12(2n)^{d+1}/(d+1)!$.

Proof. First notice that:

$$\sup_{x\in\mathcal{B}_r}\left|\widehat{D}^{\mathrm{MC}}_{\mathrm{IRW}}(x) - D_{\mathrm{IRW}}(x,P)\right| \leq \underbrace{\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\mathrm{IRW}}(x) - D_{\mathrm{IRW}}(x,P)\right|}_{(1)} + \underbrace{\sup_{x\in\mathcal{B}_r}\left|D^{\mathrm{MC}}_{\mathrm{IRW}}(x,P) - D_{\mathrm{IRW}}(x,P)\right|}_{(2)}.$$

Now, the first term (1) can be controlled using the bound for the deviations of Halfspace Depth deferred in Lemma 8. Thus, for any $t>0$, it holds:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\mathrm{IRW}}(x) - D_{\mathrm{IRW}}(x,P)\right| > t/2\right) \leq \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32). \tag{17}$$

The second term can be bounded following the same reasoning than for the Monte-Carlo approximated term of AI-IRW described in Section B.2.2. Thus, with the same notations, for any $t>0$, we have:

$$\mathbb{P}\left(\sup_{x\in\mathcal{B}_r}\left|D^{\mathrm{MC}}_{\mathrm{IRW}}(x,P) - D_{\mathrm{IRW}}(x,P)\right| > t/2\right) \leq 2\mathcal{N}\exp\left(-2m\left(t/2 - 2L_p\zeta\right)^2\right). \tag{18}$$

Using (17) and (18), one gets:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathcal{B}_r}\left|\widehat{D}^{\mathrm{MC}}_{\mathrm{IRW}}(x) - D_{\mathrm{IRW}}(x,P)\right| > t\right)\\
&\leq \mathbb{P}\left(\sup_{x\in\mathcal{B}_r}\left|\widehat{D}_{\mathrm{IRW}}(x) - D_{\mathrm{IRW}}(x,P)\right| > t/2\right)\\
&\quad + \mathbb{P}\left(\sup_{x\in\mathcal{B}_r}\left|D^{\mathrm{MC}}_{\mathrm{IRW}}(x,P) - D_{\mathrm{IRW}}(x,P)\right| > t/2\right)\\
&\leq \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32) + 2\left(\frac{3r}{\zeta}\right)^d\exp\left(-2m\left(t/2 - 2L_p\zeta\right)^2\right).
\end{aligned}$$

Choosing $\zeta \sim m^{-1}$, bounding each term on the right-hand side by $\delta/2$ and reverting the previous equation lead to the desired result. □

Appendix C: Additional experiments

C.1. Computation time of the AI-IRW depth using both SC and MCD estimators

Computation times related to the first experiment of Section 5.1 are displayed in Fig. 8 for the AI-IRW depth using both SC and MCD estimators as well as the IRW depth.

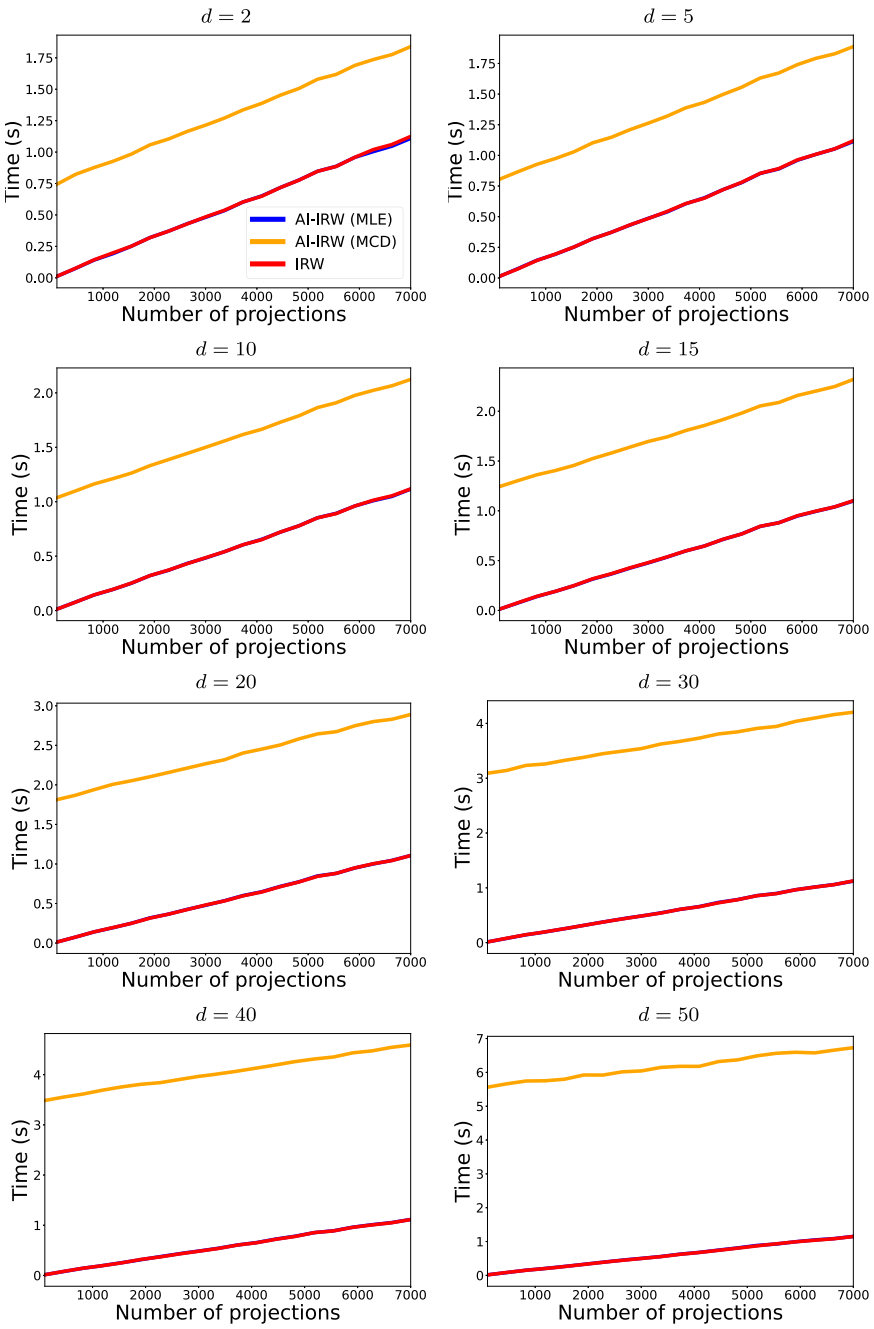

FIG 8. *Computation time of the AI-IRW depth using both SC and MCD estimators and the IRW depth depending on the number of projections for various dimensions. AI-IRW and IRW have the same computation time since the computation of the sample covariance matrix is negligible w.r.t. the computation of the IRW depth.*

C.2. Illustration of affine (non-)invariance

Fig. 9 illustrates the affine non-invariance of the IRW and the affine invariance of the AI-IRW. To that end, we simulate 10000 points stemming from a 2-dimensional centred Gaussian distribution with a covariance matrix drawn from a Wishart distribution. Further, we compute the AI-IRW depth with the sample covariance estimator and the IRW depth. In Fig. 9, we display the data with the score returned by the two depth functions such that the lighter it is, the farther from the centre it is. We can see that the score returned by the IRW and its contours are spherical, while those returned by the AI-IRW depth are ellipsoidal like those of the true underlying distribution.

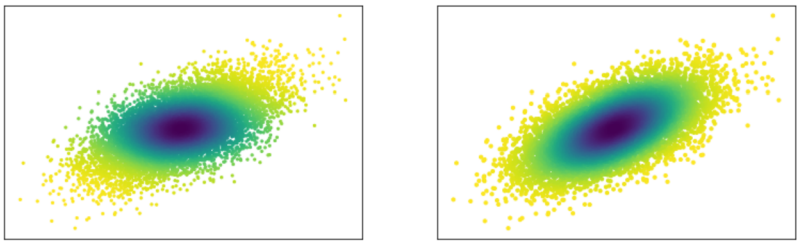

FIG 9. *The IRW depth (left) and the AI-IRW (right) depth on a Gaussian distribution. The darker the point, the higher the depth.*

C.3. Variance of the AI-IRW score

C.3.1. Variance with respect to sample realizations

We compare the stability of the approximation estimator AI-IRW measuring its variance. For 100 points stemming from a 10-dimensional Gaussian distribution with zero mean and covariance matrix drawn from the Wishart distribution (with parameters $(d, \mathcal{I}_d)$) on the space of definite matrices, the variance of the returned score is computed on two points, denoted by x_1 and x_2, drawn randomly from the 100 points previous points. The score is computed for AI-IRW, IRW, halfspace mass and halfspace depths each approximated using $m = 1000$ directions. Fig. 10 illustrates that (i) no additional variance is introduced by the affine invariant version. It further shows (ii) closeness of the three scores (due to absence of correlation) as well as (iii) their higher concentrations compared to halfspace mass and halfspace depth.

C.3.2. Variance w.r.t. noisy directions

The experiment in Section C.3.1 is repeated with different level of Gaussian noise that are added to sampled directions, i.e. $U = \frac{Z+\varepsilon\mathcal{N}(\mathbf{0},\mathcal{I}_d)}{||Z+\varepsilon\mathcal{N}(\mathbf{0},\mathcal{I}_d||}$. This experiment is

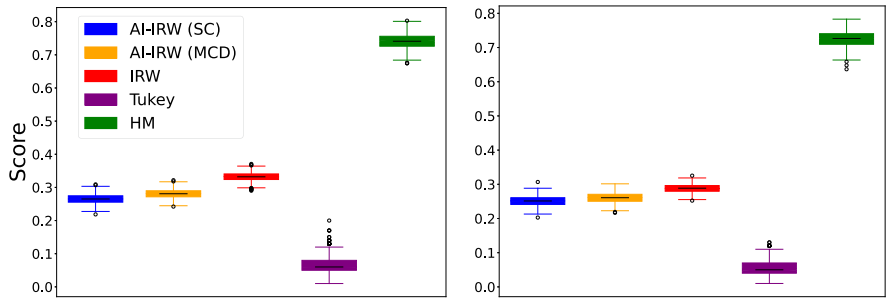

FIG 10. *Variance of the score of x_1, x_2 (from left to right) over 1000 repetitions for the AI-IRW, IRW, halfspace mass (HM) and halfspace (Tukey) depths.*

conducted with AI-IRW, IRW, HM and halfspace depth using $m = 1000$ sampled directions. The root mean square variance (over 100 repetitions) between the returned score and the original score (without noise) are computed for x_1, x_2 (same as those in Section C.3.1), see Fig. 11. Results show that AI-IRW (using the SC estimator) shares very few differences with IRW while the superiority of AI-IRW (and IRW) over the existing methods depth such as haflspace and halfspace mass is highlighted.

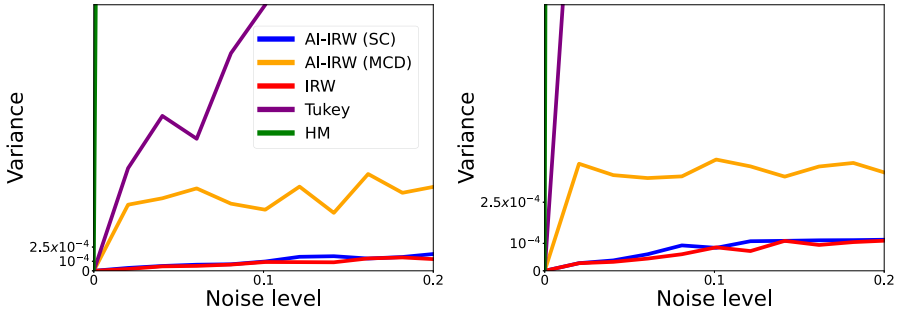

FIG 11. *Variance of the score of x_1, x_2 (from left to right) over the noise level induced in sampled directions with 1000 repetitions for the AI-IRW, IRW, Tukey depth.*

Funding

G.S. was supported by BPI France in the context of the PSPC Project Expresso (2017-2021) and the industrial chair DSAIDIS of Télécom Paris.

References

[1] AGGARWAL, C. C. (2015). Outlier analysis. *Data Mining.* MR3024573

[2] Burr, M. A. and Fabrizio, R. J. (2017). Uniform convergence rates for halfspace depth. *Statistics and Probability Letters* **124** 33–40. MR3608209
[3] Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38** 2118–2144. MR2676885
[4] Cai, T. T. and Zhou, H. H. (2013). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics* **40** 2389–2420. MR3097607
[5] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association* **91** 862–872. MR1395753
[6] Chen, B., Ting, K. M., Washio, T. and Haffari, G. (2015). Half-space mass: A maximally robust and efficient data depth method. *Machine Learning* **100** 677–699. MR3383987
[7] Chen, Y., Wiesel, A., Eldar, Y. C. and Hero, A. O. (2010). Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing* **58** 5016–5029. MR2722661
[8] Chernozhukov, V., Galichon, A., Hallin, M. and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics* **45** 223–256. MR3611491
[9] Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning* **20** 273–297.
[10] Cuevas, A. and Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data. *Journal of Multivariate Analysis* **100** 753–766. MR2478196
[11] Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. *SIAM Journal on Numerical Analysis* **7** 1–46. MR0264450
[12] Donoho, D. L. (1982). Breakdown Properties of Multivariate Location Estimators, PhD thesis, Harvard University.
[13] Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics* **20** 1803–1827. MR1193313
[14] Dyckerhoff, R. (2004). Data depths satisfying the projection property. *AStA – Advances in Statistical Analysis* **88** 163–190. MR2074729
[15] Dyckerhoff, R., Ley, C. and Paindaveine, D. (2015). Depth-based runs test for bivariate central symmetry. *Annals of the Institute of Statistical Mathematics* **67** 917–941. MR3390173
[16] Dyckerhoff, R., Mozharovskyi, P. and Nagy, S. (2021). Approximate computation of projection depths. *Computational Statistics & Data Analysis* **157**. MR4207999
[17] Einmahl, J. H. J., Li, J. and Liu, R. Y. (2015). Bridging centrality and extremity: Refining empirical data depth using extreme value statistics. *The Annals of Statistics* **43** 2738–2765. MR3405610
[18] Fan, J., Liao, Y. and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* **19** C1–C32. MR3501529

[19] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.
[20] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
[21] Kalos, M. H. and Whitlock, P. A. (2008). *Monte Carlo Methods*. Wiley-Blackwell. MR2503174
[22] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93. MR0138175
[23] Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics* **25** 1998–2017. MR1474078
[24] Koshevoy, G. A. (2002). The Tukey depth characterizes the atomic measure. *Journal of Multivariate Analysis* **83** 360–364. MR1945958
[25] Krantz, S. G. and Parks, H. R. (2008). *Geometric Integration Theory*. Birkhäuser. MR2427002
[26] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88** 365–411. MR2026339
[27] Liu, F. T., Ting, K. M. and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*.
[28] Liu, R. Y., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics* **27** 783–858. MR1724033
[29] Liu, R. Y. (1990). On a notion of data depth based upon random simplices. *The Annals of Statistics*. MR1041400
[30] Liu, R. Y. (1992). *Data depth and multivariate rank tests* In L_1-*Statistical Analysis and Related Methods* 279–294. North-Holland, Amsterdam. MR1214839
[31] Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association* **88** 252–260. MR1212489
[32] Liu, X., Mosler, K. and Mozharovskyi, P. (2018). Fast computation of Tukey trimmed regions and median in dimension $p > 2$. *Journal of Computational and Graphical Statistics*. In press. MR4007750
[33] Liu, X. and Zuo, Y. (2014). Computing halfspace depth and regression depth. *Communications in Statistics – Simulation and Computation*. MR3215761
[34] Mosler, K. (2013). Depth Statistics. In *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather* (C. Becker, R. Fried and S. Kuhnt, eds.) 17–34. Springer. MR3137661
[35] Mosler, K. and Mozharovskyi, P. (2022). Choosing among notions of multivariate depth statistics. *Statistical Science* **37** 348–368. MR4444371
[36] Nagy, S., Dyckerhoff, R. and Mozharovskyi, P. (2020). Uniform convergence rates for the approximated halfspace and projection depth. *Electronic Journal of Statistics* **14** 3939–3975. MR4165498
[37] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statis-*

tics & Probability Letters **1** 327–332. MR0721446
[38] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2** 559–572.
[39] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12** 2825–2830. MR2854348
[40] Ramsay, K., Durocher, S. and Leblanc, A. (2019). Integrated rank-weighted depth. *Journal of Multivariate Analysis* **173** 51–69. MR3920995
[41] Rosenblatt, F. (1957). The Perceptron—a perceiving and recognizing automaton. Report 85-460-1. Cornell Aeronautical Laboratory. MR0122606
[42] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79** 871–880. MR0770281
[43] Rousseeuw, P. J. and Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing* **8** 193–203. MR1702314
[44] Rousseeuw, P. J. and van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.
[45] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**. MR2183942
[46] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons. MR0838963
[47] Staerman, G. (2022). Functional anomaly detection and robust estimation, PhD thesis, Institut polytechnique de Paris.
[48] Staerman, G., Adjakossa, E., Mozharovskyi, P., Hofer, V., Gupta, J. S. and Clémençon, S. (2023). Functional anomaly detection: a benchmark study. *International Journal of Data Science and Analytics* **16** 101–117.
[49] Struyf, A. J. and Rousseeuw, P. J. (1999). Halfspace depth and regression depth characterize the empirical distribution. *Journal of Multivariate Analysis* **69** 135–153. MR1701410
[50] Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (R. D. James, ed.) **2** 523–531. MR0426989
[51] Vardi, Y. and Zhang, C.-H. (2000). The multivariate L1-median and associated data depth. *Proceedings of the National Academy of Sciences* **97** 1423–1426. MR1740461
[52] Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability* **25** 655–686. MR2956207
[53] Wedin, P.-A. (1973). Perturbation theory for pseudo-inverses. *IT Numerical Mathematics*. MR0336982

[54] Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen. *Mathematische Annalen* **71** 441–479. MR1511670

[55] Yu, Y., Wang, T. and Samworth, R. J. (2014). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006

[56] Zhang, J. (2002). Some extensions of Tukey's depth function. *Journal of Multivariate Analysis* **82** 134–165. MR1918618

[57] Zhao, Y., Nasrullah, Z. and Li, Z. (2019). PyOD: A Python toolbox for scalable outlier detection. *Journal of Machine Learning Research* **20** 1–7.

[58] Zuo, Y. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics* **31** 1460–1490. MR2012822

[59] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics* **28** 461–482. MR1790005