Convergence rates for estimators of geodesic distances and Fréchet expectations

Catherine Aaron, Olivier Bodart

Convergence rates for estimators of geodesic distances and Fréchet expectations

C. Aaron*, O. Bodart†

March 19, 2018

Abstract

Consider a sample $\mathcal{X}_n = \{X_1, \dots, X_n\}$ of i.i.d variables drawn with a probability distribution $\mathbb{P}_X$ supported on a set $M \subset \mathbb{R}^d$. This article mainly deals with the study of a natural estimator for the geodesic distance on M. Under rather general geometric assumptions on M, a general convergence result is proved. Assuming M to be a manifold of known dimension $d' \leq d$, and under regularity assumptions on $\mathbb{P}_X$, an explicit convergence rate is given. In the case when M has no boundary, the knowledge of the dimension d' is unnecessary to obtain this convergence rate. The second part of the work consists in building an estimator for the Fréchet expectations on M, and proving its convergence under regularity conditions, applying the previous results.

Keywords: Geometric inference, Geodesic distance, Statistics on manifolds, Fréchet expectations
AMS Classification : 62-07, 62G05, 62G20, 62H99

1 Introduction

Let $\mathbb{P}_X$ be a probability distribution supported on a set $M \subset \mathbb{R}^d$, $d \geq 2$, that is M is the smallest closed set in $\mathbb{R}^d$ of probability 1. Let $\mathcal{X}_n = \{X_1, \dots, X_n\}$ be a sample of i.i.d variables drawn on M with the distribution $\mathbb{P}_X$. The first

*Laboratoire de Mathématiques Blaise Pascal (UMR 6620 - CNRS) Université Clermont Auvergne (Clermont-Ferrand 2), 63177 Aubière cedex, France (Catherine.Aaron@uca.fr)

†Institut Camille Jordan Faculté des Sciences et Techniques, Université Jean Monnet, 42023 Saint-tienne Cedex 2 , France (Olivier.Bodart@univ-st-etienne.fr)

aim of this work is the study of a rather classical estimator of the geodesic distance on the unknown set M.

The way to build this estimator is quite intuitive (see e.g. [16]): given $r > 0$, build a graph interconnecting all the pairs (X_i, X_j) of the sample $\mathcal{X}_n$ such that $\|X_i - X_j\| \leq r$. The geodesic distance between any two points X_k and X_l of the sample is then estimated by the length of the shortest path connecting X_k and X_l in the graph (see the Definition 1 for details). This path (and its length) can be computed with optimal complexity by using Dijkstra's algorithm (see for example [3] for a presentation of this algorithm). As usual in such problems $r = r_n$ must be a conveniently chosen sequence. First it must converge to 0 as $n \to \infty$. Moreover this convergence has to be slow enough for the path realizing the estimator to be smooth enough.

To our knowledge, the asymptotic behavior of such an estimator has not been studied yet. We will show, under quite general assumptions on the support M, that choosing $r_n = d_h(\mathcal{X}_n, M)^{2/3}$ appears to be convenient (Theorem 1). Here and throughout the paper, $d_h(A, B)$ denotes the Hausdorff distance between the sets A and B:

$$d_h(A, B) = \max \left\{ \sup_{a \in A}(\inf_{b \in B} ||a - b||), \sup_{b \in B}(\inf_{a \in A} ||a - b||) \right\}.$$

Assuming that M is a d'-manifold, $d' \leq d$, and assuming some regularity for the distribution $\mathbb{P}_X$, it will be shown that $d_h(\mathcal{X}_n, M) = \mathcal{O}(\ln n/n)^{1/d'}$, allowing to give the convergence rate of our estimator when the dimension d' is known (Corollary 1). When d' is unknown, and M is supposed to have no boundary, the Corollary 2 presents an estimator of r_n which allows to obtain the same convergence rate.

Eventually we will apply these results to the estimation of the Fréchet expectations, as defined in [11], of the distribution $\mathbb{P}_X$ on M (Theorem 2).

Using the estimated geodesic distance in place of the euclidean distance has become frequent in different fields of application, in order to take the nonlinearity of the data into account. In [16], the authors propose to apply the multidimensional scaling (see e.g. [7]) to the array of geodesic distances between points. This idea opened the way to the use of the geodesic distance in dimension reduction (see [8], [5], [9], [14] and [10]). In [2] and [6] , the question of intrinsic dimension estimation using graph-based statistics is studied. In particular, in [6] the authors propose a generalization of the correlation dimension where the euclidean distance is replaced by the (estimated) geodesic distance. This approach has the advantage to be less sensitive to the (difficult) question of the choice of the parameter (see also [15]). In [11], the author rises the question of the generalization of classical statistical

quantities (such as the mean and median) to the case of data supported on Riemannian manifolds.

The paper is organized as follows: in Section 2, the general framweork, main definitions and the results are stated. The first subsection presents the results concerning the estimation of the geodesic distance on the support M (Theorem 1 and Corollaries 1 and 2), while the second states the Theorem for the Fréchet expectations estimator (Theorem 2). Section 3 is devoted to the proofs of the results.

2 General framework and Main results

2.1 Estimating geodesic distances

Let us first start with the definition of our estimator.

Definition 1. *Let $\mathcal{X}_n = \{X_1, \dots X_n\}$ be a set of n i.i.d. random variables with distribution $\mathbb{P}_X$ supported on a compact set $M \subset \mathbb{R}^d$, $d \geq 2$. Let, $r_n > 0$ being a given number, $\mathcal{G}_{r_n}(\mathcal{X}_n)$ be the graph which edges are the segments $[X_i, X_j]$ such that $\|X_i - X_j\| \leq r_n$.*

For $(i,j) \in \{1, \dots, n\}^2$,

let, if it exists, $\hat{\gamma}_{r_n}(X_i, X_j)$ be the shortest path (in euclidean norm) connecting X_i and X_j in $\mathcal{G}_{r_n}(\mathcal{X}_n)$, and $|\hat{\gamma}_{r_n}(X_i, X_j)|$ its length.

We aim at proving, for a class of convenient compact sets in $\mathbb{R}^d$, that $|\hat{\gamma}_{r_n}(X_i, X_j)|$ is an estimator of the geodesic distance $\gamma(X_i, X_j)$ on M, with good convergence properties.

Definition 2. *Let $M \subset \mathbb{R}^d$ be a compact set, M is said to be K_M-geodesically smooth (later denoted as GS) for some positive number K_M if:*

(i) for all $(x,y) \in M^2$ there exits a geodesic path $\gamma_{x \to y}$ of class $\mathcal{C}^1$ that links x to y;

(ii) there exits a real function β satisfying $\lim_{t \to 0} \beta(t) = 0$ such that $\forall (x,y) \in M^2$, $\|\gamma_{x \to y}\| \leq \beta(\|x - y\|)$;

(iii) let $\Gamma_{x \to y} : [0, |\gamma_{x \to y}|] \to \mathbb{R}^d$ be the parametrization of $\gamma_{x \to y}$ such that $\Gamma_{x \to y}(s)$ is the point of $\gamma_{x \to y}$ that is at a (curvilinear) distance s from x (along the geodesic curve). For all $(x,y) \in M^2$, the gradient of $\Gamma_{x \to y}$, denoted $\dot{\Gamma}_{x \to y}$, is K_M-Lipschitz continuous.

A compact manifold of class $\mathcal{C}^2$ with no boundary satisfies the assumptions of the Definition 2, but one can build more general examples of such sets (that is compact sets with $\mathcal{C}^1$ geodesic curves which have K_M-Lipschitz tangent maps). As an example, the Figure 1 depicts two examples of GS-Sets (sets 1 and 2), and one which is not. Notice that the second set, however satisfying the GS property, is not a manifold.

Figure 1: The sets are the colored areas. The first one is GS (with some geodesic curves depicted). The second is also GS (but is not a manifold). The third is not GS : some geodesic curves are not smooth enough.

Theorem 1. *Let $\hat{\gamma}_{r_n}$ be the estimator introduced in the Definition 1. Assume that there exists a sequence $\rho_n \xrightarrow{a.s.} 0$ such that $\rho_n \geq d_h(\mathcal{X}_n, M)$ (e.a.s), and let (r_n) be a sequence such that $r_n > 2\rho_n$ and $\rho_n / r_n \xrightarrow{a.s.} 0$. Then,*

$$\max_{i,j} \left| |\hat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}| \right| = \mathcal{O}\left(\max\left(r_n, \frac{\rho_n^2}{r_n^2} \right) \right) \quad e.a.s. \tag{1}$$

Assuming that $r_n > 2\rho_n$ ensures the existence of $|\hat{\gamma}_{r_n}(X_i, X_j)|$ for all i and j. The first part of the proof clearly illustrates this fact.

One would then assume the sequence $r_n = d_h(\mathcal{X}_n, M)^{2/3}$ to be an optimal choice. However, even though it is known that $d_h(\mathcal{X}_n, M) \to 0$ a.s. (see [4]), the rate of this convergence is unknown in general. Thus, in order to obtain a convergence rate for our estimator, we are going to make extra assumptions on the set M and the probability distribution $\mathbb{P}_X$.

Definition 3. *Let $\delta > 0$. A probability measure $\mathbb{P}_X$ supported on $M \subset \mathbb{R}^d$ is said to be $\delta-$standard with respect to a measure μ if there exists $\lambda > 0$ such that $\mathbb{P}_X(\mathcal{B}(x, \varepsilon)) \geq \delta \mu(\mathcal{B}(x, \varepsilon))$ for all $x \in M$ and $\varepsilon \in]0, \lambda]$.*

We then have the following result:

Corollary 1. *Let $M \subset \mathbb{R}^d$, $d \geq 2$ be a $d'-$dimensional manifold of class $\mathcal{C}^1$ satisfying the GS property for some number $K_M > 0$. Let $\mathbb{P}_X$ be a probability distribution on M. Assume, for some number $\delta > 0$, that $\mathbb{P}_X$ is $\delta-$standard with respect to the measure induced on M by the Lebesgue measure in $\mathbb{R}^d$.*

If the sequence (r_n) in the Definition 1 is such that

$$\left(A_0\frac{\ln n}{n}\right)^{2/3d'} \leq r_n \leq \left(A_1\frac{\ln n}{n}\right)^{2/3d'},$$

with $A_0 > 0$ and $A_1 > 0$, then

$$\max_{i,j}\left||\hat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}|\right| = \mathcal{O}\left(\left(\frac{\ln n}{n}\right)^{2/3d'}\right) \ e.a.s.$$

As usual when dealing with estimation problems, the sequence of radii (r_n) in the previous theorem remains abstract. In particular the dimension d' of the support is generally unknown. However, making extra assumptions on the support M and the density of the distribution, we can accurately estimate the sequence of radii, with no need for estimating d'. This last fact is indeed worth being emphasized, since the knowledge of the geodesic distance is known to be useful for a good estimation of the dimension of a manifold (see e.g. [6]).

Let $L_n = \max_i(\min_{j\neq i}\|X_i - X_j\|)$ and let θ_n be the longest edge of the minimal spanning tree of the sample. Up to a rescaling of the data, we can suppose that $\max_i(\max_j||X_i - X_j||) \leq 1$. Then we have $L_n = \max_i(\min_{j\neq i}\|X_i - X_j\|) \leq 1$ and $\theta_n \leq 1$, hence $L_n^{2/3} \geq L_n$ and $\theta_n^{2/3} \geq \theta_n$.

Choosing a sequence of radii satisfying $r_n \geq \theta_n$ ensures the existence of the estimator $|\hat{\gamma}_{r_n}(X_i, X_j)|$. Conjecturing that the results of [12] can be generalized to the case of data drawn on a smooth manifold with a density close to the uniform one leads to the the choice of $r_n = c.\theta_n^{2/3}$ with $c \geq 1$. If the conjecture is correct, this would guarantee the existence of the estimator and provide optimal convergence rates. More practically, in order to prove a theoretical result we are led to chose r_n in relation to L_n. This only ensures the existence of our estimator asymptotically.

Corollary 2. *Let $M \in \mathbb{R}^d$, $d \geq 2$ be a $d'-$dimensional manifold, $d' < d$ of class $\mathcal{C}^2$ with no boundary and $\mathbb{P}_X$ be a probability distribution on M with continuous probability density $f_X \geq f_0 > 0$. Then, for any $c > 0$, setting $r_n = c.(\max_i(\min_{j\neq i}\|X_i - X_j\|))^{2/3}$ in the Definition 1, we have*

$$\max_{i,j}\left||\hat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}|\right| = \mathcal{O}\left(\left(\frac{\ln n}{n}\right)^{2/3d'}\right) \ e.a.s.$$

The assumptions of this Corollary imply those of the Theorem 1; they allow to explicitly build a convenient sequence of radii (r_n) only from the

sample. To prove this Theorem we use a result by Penrose (see [13]) which applies only in the case when M has no boundary. However, numerical simulations on $\mathcal{C}^2$ sets with boundary, satisfying the GS assumption, lead us to think that the result is also true for such sets.

The question of the choice of the sequence (r_n) remains a difficult subject. In our framework, we propose the following decision rule: first, in the absence of a priori knowledge on the data, and when the support M can have several arcwise connected components (ie data classes), we will choose r_n of order $L_n^{2/3}$. This sequence of radii will converge to 0 and allow to identify the different classes in the data with optimal convergence rate (although the existence of the estimator is only ensured asymptotically). If one knows a priori that the support is arcwise connected, choosing $r_n = c.\theta_n^{2/3}$ with $c \geq 1$ may be a convenient choice, even if the asymptotic properties of the estimator are conjecture-based.

2.2 Estimating Fréchet Expectations

In this section we assume the set M to be a a compact d'-manifold of class $\mathcal{C}^2$. Following the ideas of X. Pennec (see [11]), we consider the Fréchet expectations of the random variable X (which distribution is supported on M):

$$\mathbb{E}_k^{\mathrm{Fr}}(X) = \mathrm{argmin}_{x \in M} \mathbb{E}(|\gamma_{x \to X}|^k), \quad k \in \mathbb{N}^*, \tag{2}$$

which are generalizations of the expected value for $k = 2$ and of the median (or depth) for $k = 1$. As it is pointed out in [11], these expectations are not necessarily unique. For example, if M is a sphere and $\mathbb{P}_X$ the uniform distribution, then obviously all the points of M realize the minimum in (2) (for any $k \geq 1$).

To avoid dealing with such situations, we are going to make the following assumption, considering that k is fixed:

$$\begin{cases} \Phi(x) = \mathbb{E}(|\gamma_{x \to X}|^k) \text{ admits a unique minimum } x^* \in M, \\ \Phi \text{ is of class } \mathcal{C}^2 \text{ in a neighbourhood of } x^*, \\ H_\Phi(x^*) \text{ is positive definite,} \end{cases} \tag{3}$$

where H_Φ denotes the hessian matrix of Φ (i.e. $(H_\Phi)_{i,j} = \frac{\partial^2 \Phi}{\partial x_i, \partial x_j}$).

Remark: It must be noted that Φ is a continuous function on M. Indeed the triangle and Minkowski inequalities give $|\Phi(x)^{1/k} - \Phi(y)^{1/k}| \leq |\gamma_{x \to y}|$, for any $(x, y) \in M^2$. The extra (local) regularity in the conditions (3) is required for the sake of simplicity, allowing to apply basic differential calculus results at the optimal point x^*.

The first part of this assumption is very strong, but the second part is not. For example, when $d' = 1$ and M is homeomorphic to a segment, explicit computations show that (3) holds for $k = 1$ iff $f_X(x^*) \neq 0$. For $k = 2$, when M is a bounded closed convex set of dimension d, the geodesic distance on M coincides with the euclidean distance, the expectation $\mathbb{E}(X)$ lies in M, it minimizes the function $\Phi(x)$ and the condition (3) is satisfied (with $H_\Phi \equiv 2I_d$). This leads to think that, for $k = 2$, this condition is general enough and may hold for a wide class of regular submanifolds of $\mathbb{R}^d$.

In this section we aim at studying the behavior of the natural estimator of $\mathbb{E}_k^{\mathrm{Fr}}(X)$:

$$\hat{\mathbb{E}}_{k,r_n}^{\mathrm{Fr}}(\mathcal{X}_n) = \mathrm{argmin}_{X_i \in M} \frac{1}{n} \sum_j |\hat{\gamma}_{r_n}(X_i, X_j)|^k. \tag{4}$$

Theorem 2. *Assume that $M \subset \mathbb{R}^d$, $d \geq 2$ is a $d'-$dimensional manifold, $d' < d$ of class $\mathcal{C}^2$ with no boundary and that $\mathbb{P}_X$ is a probability distribution on M with continuous and bounded from below probability density f_X. Moreover, suppose that assumption (3) holds. Then, choosing $r_n = c(\max_i(\min_j ||X_i - X_j||))^{2/3}$ in the definition of $\hat{\gamma}_{r_n}$, we have*

$$|\mathbb{E}_k^{Fr}(X) - \hat{\mathbb{E}}_{k,r_n}^{Fr}(\mathcal{X}_n)| = \mathcal{O}\left(\left(\frac{\ln n}{n}\right)^{\min(1/4,1/3d')}\right) \ e.a.s.$$

3 Proofs of the results

Let us start with a result which is a direct consequence of the regularity of the set considered here.

Proposition 1. *If $M \subset \mathbb{R}^d$ is K_M-geodesically smooth then, there exist $r_M > 0$ and $A_M > 0$, depending only on M, such that*

$$\forall (x, y) \in M^2;\ ||x - y|| \leq r_M \Longrightarrow |\gamma_{x\to y}| \leq ||x - y|| + A_M||x - y||^2.$$

Proof. Let $(x, y) \in M^2$. Consider the parametrization $\Gamma_{x\to y}$ of the geodesic curve $\gamma_{x\to y}$ as in Definition 2. The map $\dot{\Gamma}$ being K_M-Lipschitz continuous, for all $t_0 \in [0, |\gamma_{x\to y}|]$, there exists $\varepsilon_{t_0}[0, |\gamma_{x\to y}|] \to \mathbb{R}^d$ such that:

$$\forall t \in [0, |\gamma_{x\to y}|],\ \dot{\Gamma}(t) = \dot{\Gamma}(t_0) + K_M|t - t_0|\varepsilon_{t_0}(t), \quad \|\varepsilon_{t_0}(t)\| \leq 1.$$

Thus,

$$\int_0^{|\gamma_{x\to y}|} \dot{\Gamma}(t)dt = \int_0^{|\gamma_{x\to y}|} \left(\dot{\Gamma}(t_0) + K_M|t - t_0|\varepsilon_{t_0}(t)\right) dt,$$

that is

$$y - x = \dot{\Gamma}(t_0)|\gamma_{x\to y}| + K_M \int_0^{|\gamma_{x\to y}|} |t - t_0|\varepsilon_{t_0}(t)\, dt.$$

Choosing $t_0 = |\gamma_{x\to y}|/2$, and noticing that with the chosen parametrization we have $\|\dot{\Gamma}(t_0)\| = 1$, we obtain

$$\forall (x, y) \in M^2,\ \|x - y\| \geq |\gamma_{x\to y}| - \frac{K_M}{4}|\gamma_{x\to y}|^2.$$

Now assuming that $\|x - y\| \leq K_M^{-1}$, the following alternative holds:

(i) either $|\gamma_{x\to y}| \geq \frac{2+2\sqrt{1-K_M\|x-y\|}}{K_M}$,

(ii) or $|\gamma_{x\to y}| \leq \frac{2-2\sqrt{1-K_M\|x-y\|}}{K_M}$.

For $\|x-y\|$ small enough, the first case is impossible because of the condition (ii) in the Definition 2. Therefore, there exists $r_M \leq K_M^{-1}$ such that, for $\|x - y\| \leq r_M$, the second case of the alternative holds. Making a Taylor expansion of $\|x - y\|$ ends the proof. □

3.1 Proof of Theorem 1

Let $(i, j) \in \{1, \dots n\}^2$, $i \neq j$, and let γ_{ij} be the geodesic curve between X_i and X_j. Consider a partition $\{x_0, \dots, x_K\}$ of γ_{ij} such that

$$x_0 = X_i,\ x_K = X_j, \tag{5}$$

$$K = \left\lceil \frac{|\gamma_{X_i\to X_j}|}{r_n - 2\rho_n} \right\rceil, \tag{6}$$

$$|\gamma_{x_k\to x_{k+1}}| = \frac{|\gamma_{X_i\to X_j}|}{K}, \tag{7}$$

so that

$$\begin{array}{l} |\gamma_{x_k\to x_{k+1}}| = r_n - 2\rho_n, \quad k = 0, \dots K-2, \\ |\gamma_{x_{K-1}\to x_K}| < r_n - 2\rho_n. \end{array} \tag{8}$$

We have

$$|\gamma_{X_i\to X_j}| = \sum_{k=0}^{K-1} |\gamma_{x_k\to x_{k+1}}| \geq \sum_{k=0}^{K-1} \|x_k - x_{k+1}\|. \tag{9}$$

From the definition of ρ_n, for any $k \in \{0, \dots, K\}$, there exists $i_k \in \{1, \dots n\}$ such that $\|X_{i_k} - x_k\| \leq \rho_n$. Let us denote, for the sake of simplicity,

$$Y_k = X_{i_k}, \quad \varepsilon_k = Y_k - x_k, \quad U_k = \frac{x_k - x_{k+1}}{\|x_k - x_{k+1}\|}.$$

Recall that

$$\|\varepsilon_k\| \leq \rho_n,\ k = 0 \dots K-1. \tag{10}$$

For $k \in \{0, \dots, K-1\}$,

$$\begin{aligned}
\|Y_k - Y_{k+1}\|^2 &= \|\varepsilon_k + (x_k - x_{k+1}) - \varepsilon_{k+1}\|^2 \\
&= \|x_k - x_{k+1}\|^2 + 2\left\langle x_k - x_{k+1} \,|\, \varepsilon_k - \varepsilon_{k+1}\right\rangle + \|\varepsilon_k - \varepsilon_{k+1}\|^2 \\
&= \|x_k - x_{k+1}\|^2 \times \left(1 + 2\,\frac{\langle U_k \,|\, \varepsilon_k - \varepsilon_{k+1}\rangle}{\|x_k - x_{k+1}\|} + \frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|^2}\right),
\end{aligned}$$

that is, taking the square root of this equality, and noticing that $\sqrt{1+t} \leq 1 + t/2$, $t \geq -1$,

$$\begin{aligned}
\|Y_k - Y_{k+1}\| &\leq \|x_k - x_{k+1}\| \times \left(1 + \frac{\langle U_k \,|\, \varepsilon_k - \varepsilon_{k+1}\rangle}{\|x_k - x_{k+1}\|} + \frac{1}{2}\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|^2}\right) \\
&\leq \|x_k - x_{k+1}\| + \langle U_k \,|\, \varepsilon_k - \varepsilon_{k+1}\rangle + \frac{1}{2}\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|}.
\end{aligned}$$

In view of (5), (6),(7) and (8) $\|x_{K-1} - x_K\|$ is not bounded from bellow hence we shall treat the cases $k < K-1$ and $k = K-1$ separately. From (9) we have

$$|\gamma_{X_i \to X_j}| \geq ||x_{K-1} - x_K|| + \sum_{k=0}^{K-2} \|Y_k - Y_{k+1}\| - \frac{1}{2}S_1 - S_2, \tag{11}$$

with

$$S_1 = \sum_{k=0}^{K-2} \frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|}, \quad S_2 = \sum_{k=0}^{K-2} \langle U_k \,|\, \varepsilon_k - \varepsilon_{k+1}\rangle\,. \tag{12}$$

Let us first study S_1. From (8) and the Proposition 1, we have, for $k \in \{0, \dots, K-2\}$,

$$r_n - 2\rho_n - A_M\,(r_n - 2\rho_n)^2 \leq \|x_k - x_{k+1}\| \leq r_n - 2\rho_n \leq r_n, \tag{13}$$

with $A_M > 0$ only depending on M.

Then, for n large enough to have $u_n = \frac{2\rho_n}{r_n} + A_M \frac{(r_n - 2\rho_n)^2}{r_n} < 1$ and applying that $\frac{1}{1-u} \leq 1 + u$ when $u \in [0, 1[$ we have, for all $k \in \{0, \dots, K-2\}$

$$\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|} \leq \frac{4\rho_n^2}{r_n - 2\rho_n - A_M\,(r_n - 2\rho_n)^2} \leq \frac{4\rho_n^2}{r_n}\left(1 + \frac{2\rho_n}{r_n} + A_M\frac{(r_n - 2\rho_n)^2}{r_n}\right).$$

Thus

$$\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|} \leq \frac{4\rho_n^2}{r_n}\,(1 + o(1))$$

uniformly in K, X_i, X_j.

The definition of ρ_n implies that $r_n - 2\rho_n \sim r_n$. Moreover, since the set M is compact and satisfies the GS assumption, $\gamma_{X_i \to X_j}$ is uniformly bounded for all $(i,j) \in \{1, \dots n\}^2$. Hence, there exists $L_M > 0$ such that

$$0 < K \leq \frac{L_M}{r_n}, \tag{14}$$

where K is defined by (6), and we have

$$S_1 \leq L_M \left(\frac{4\rho_n^2}{r_n^2} + o\left(\frac{\rho_n^2}{r_n^2} \right) \right). \tag{15}$$

Now, since the set M is smooth we can write, for $k \in \{0, \dots, K-2\}$,

$$U_k = \overset{\bullet}{\Gamma}_{x_0 \to x_k}(k(r_n - 2\rho_n)) + \mathcal{O}(\|x_k - x_{k+1}\|) = \overset{\bullet}{\Gamma}_{x_0 \to x_k}(k(r_n - 2\rho_n)) + \mathcal{O}(r_n),$$

and

$$U_k - U_{k+1} = \overset{\bullet}{\Gamma}_{x_0 \to x_k}(k(r_n - 2\rho_n)) - \overset{\bullet}{\Gamma}_{x_0 \to x_k}((k+1)(r_n - 2\rho_n)) + \mathcal{O}(r_n).$$

Then, $\overset{\bullet}{\Gamma}$ being Lipschitz continuous,

$$\|U_k - U_{k+1}\| = \mathcal{O}(r_n). \tag{16}$$

We can now rewrite S_2 as

$$S_2 = \sum_{k=1}^{K-2} \langle U_k - U_{k-1} \,|\, \varepsilon_k \rangle + \langle U_0 \,|\, \varepsilon_0 \rangle - \langle U_{K-2} \,|\, \varepsilon_{K-1} \rangle ,$$

hence, in view of (10), (13), (14), we have

$$S_2 = \mathcal{O}(\rho_n).$$

Combining this last inequality with (11),(12),(15), we obtain

$$|\gamma_{X_i \to X_j}| \geq \sum_{k=0}^{K-2} \|Y_k - Y_{k+1}\| - \mathcal{O}(\rho_n) - 2L_M \frac{\rho_n^2}{r_n^2} + o\left(\frac{\rho_n^2}{r_n^2} \right).$$

Thus

$$|\gamma_{X_i \to X_j}| \geq \sum_{k=0}^{K-1} \|Y_k - Y_{k+1}\| - \|Y_{K-1} - Y_K\| - \mathcal{O}(\rho_n) - 2L_M \frac{\rho_n^2}{r_n^2} + o\left(\frac{\rho_n^2}{r_n^2} \right).$$

Recall that, for all $k \in \{0, \dots, K-1\}$ we have $\|Y_k - Y_{k+1}\| = \|(x_k - x_{k+1}) - (\varepsilon_k - \varepsilon_{k+1})\|$, hence the triangle inequality, (8) and (10) yield

$$\|Y_k - Y_{k+1}\| \leq r_n,\ k \in \{0, \dots, K-1\}. \tag{17}$$

Applying (17) for $k = K-1$ we first obtain

$$|\gamma_{X_i \to X_j}| \geq \sum_{k=0}^{K-1} \|Y_k - Y_{k+1}\| - r_n - \mathcal{O}(\rho_n) - 2L_M \frac{\rho_n^2}{r_n^2} + o\left(\frac{\rho_n^2}{r_n^2}\right). \tag{18}$$

By (17) the path $Y_0, \dots Y_K$ belongs to the graph $\mathcal{G}_{r_n}(\mathcal{X}_n)$ so we clearly have:

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \leq \sum_{k=0}^{K-1} \|Y_k - Y_{k+1}\|,$$

therefore, since $\rho_n = o(r_n)$ and in view of (18), we have

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \leq |\gamma_{X_i \to X_j}| + r_n + o(r_n) + 2L_M \frac{\rho_n^2}{r_n^2} + o\left(\frac{\rho_n^2}{r_n^2}\right). \tag{19}$$

We are now going to prove the following inequality :

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \geq |\gamma_{X_i \to X_j}| - 2\, A_M L_M r_n. \tag{20}$$

For the sake of clarity, let us omit the superscripts in the definition 1 and denote $Z_0 = X_i, Z_1, \dots Z_{L_1}, Z_L = X_j$ the nodes of the graph $\mathcal{G}_n^{i,j}$ realizing the path $\widehat{\gamma}_{r_n}(X_i, X_j)$. Proposition 1 yields

$$|\gamma_{X_i \to X_j}| \leq \sum_{k=0}^{L-1} |\gamma_{Z_k \to Z_{k+1}}| \leq \sum_{k=0}^{L-1} \left(\|Z_k - Z_{k+1}\| + A_M \|Z_k - Z_{k+1}\|^2 \right).$$

Noticing, from Definition 1, that $\|Z_k - Z_{k+1}\| \leq r_n$, we obtain

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \geq |\gamma_{X_i \to X_j}| - A_M L r_n^2. \tag{21}$$

Let us now obtain a bound for the number of nodes L in the path $\widehat{\gamma}_{r_n}(X_i, X_j)$. Necessarily, by construction, we have

$$\|Z_k - Z_{k+1}\| + \|Z_{k+1} - Z_{k+2}\| > r_n, \quad k = 0, \dots, L-2. \tag{22}$$

Indeed, if it was not the case, we would have $\|Z_k - Z_{k+2}\| \leq r_n$, hence the path $\{Z_k, Z_{k+2}\}$ would be shorter in the graph $\mathcal{G}_n^{i,j}$ than the path $\{Z_k, Z_{k+1}, Z_{k+2}\}$

which is impossible. Therefore, summing up (22) for $k \in \{0, \dots L-2\}$, we obtain

$$Lr_n^2 \leq 2\,|\widehat{\gamma}_{r_n}(X_i, X_j)| + r_n^2,$$

hence, in view of (19), (21) and reminding that $|\gamma_{X_i \to X_j}|$ is uniformly bounded, we get (20).

This inequality and (19) finally imply

$$\left||\widehat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}|\right| \leq C_M \, \max\left(\frac{\rho_n^2}{r_n^2}, r_n\right), \tag{23}$$

where the constant $C_M > 0$ only depends on the manifold M. This yields the estimate (1) and concludes the proof.

3.2 Proof of Corollary 1

Reasoning as in [1], since M is of class $\mathcal{C}^1$, one can cover M with $\nu_n \leq C\,n$ (with $C > 0$) deterministic balls of radius $\varepsilon_n = (1/n)^{1/d'}$ with centers $x_i \in M$, $i \in \{1, \dots, \nu_n\}$. Let $\omega_{d'}$ be the volume of the $d' - dimensional$ unit ball. We then classically have

$$\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'}\right) = \mathbb{P}_X\left(\exists x \in M; \mathcal{B}\left(x, \left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'}\right) \cap \mathcal{X}_n = \emptyset\right).$$

The triangular inequality thus implies,

$$\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'}\right) \leq \mathbb{P}_X\left(\exists i; \mathcal{B}\left(x_i, \left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'} - \varepsilon_n\right) \cap \mathcal{X}_n = \emptyset\right).$$

The probability distribution being standard with respect to the the d' dimensional measure, we have

$$\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'}\right) \leq \nu_n\left(1 - \delta\omega_{d'}\left(\left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'} - \varepsilon_n\right)^{d'}\right)^n.$$

The Taylor expansion gives

$$\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'}\right) \leq C\,n^{1-2\lambda+o(1)},$$

for any $\lambda > 0$.

Applying the Borel-Cantelli Lemma, we deduce that, for any $\lambda > 1$,

$$d_h(\mathcal{X}_n, M) \leq \left(\frac{2\lambda}{\delta\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'} \qquad e.a.s.$$

Applying the Theorem 1 ends the proof.

3.3 Proof of Corollary 2

Let

$$t_n = \max_i (\min_j \|X_i - X_j\|).$$

Applying Theorem 5.1 p.958 in [13], we have

$$\frac{n\,\omega_{d'}\,t_n^{d'}}{\ln n} \xrightarrow{a.s.} f_0^{-1}.$$

Therefore, one can easily deduce that

$$\left(\frac{1}{2f_0\,\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'} \leq t_n \leq \left(\frac{2}{f_0\,\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'} \quad e.a.s.$$

Since we have $r_n = (c\,t_n)^{2/3}$, the assumptions of the Corollary 1 are fulfilled, which allows to conclude the proof.

3.4 Proof of Theorem 2

In view of (3) and (4), let us introduce the following estimators :

$$\begin{aligned} \overline{\Phi}(x) &= \frac{1}{n}\sum_i (|\gamma_{x\to X_i}|^k), \\ \hat{\Phi}(x) &= \frac{1}{n}\sum_i (|\hat{\gamma}_{r_n}(x, X_i)|^k). \end{aligned} \tag{24}$$

At first, let us prove that there exits a deterministic constant $D > 0$ such that

$$\max_i |\hat{\Phi}(X_i) - \Phi(X_i)| = D\left(\frac{\ln n}{n}\right)^{\min\{2/3d', 1/2\}} \text{ e.a.s} \tag{25}$$

Indeed, the manifold M being compact, one can apply the Hoeffding inequality and obtain that

$$\forall x \in M, \quad \mathbb{P}_X(|\overline{\Phi}(x) - \Phi(x)| \geq \varepsilon_n) \leq 2\exp(-2n\varepsilon_n^2/L^{2k}),$$

$L > 0$ being the constant introduced in the proof of Theorem 1. Hence,

$$\mathbb{P}_X(\exists i \in \{1, \ldots, n\}\,;\, |\overline{\Phi}(X_i) - \Phi(X_i)| \geq \varepsilon_n) \leq 2n\exp(-2n\varepsilon_n^2/L^{2k}).$$

Setting $\varepsilon_n = \sqrt{2}L^k\sqrt{\ln n/n}$ in this last inequality yields

$$\mathbb{P}_X(\max_i |\overline{\Phi}(X_i) - \Phi(X_i)| \geq \varepsilon_n) \leq 2n^{-3}$$

so that the Borel-Cantelli Lemma allows to conclude that

$$\max_i |\overline{\Phi}(X_i) - \Phi(X_i)| = \mathcal{O}\left(\frac{\ln n}{n}\right)^{1/2} \text{ e.a.s.} \tag{26}$$

Now, noticing that the assumptions of Corollary 2 are fulfilled, we have

$$\max_i |\hat{\Phi}(X_i) - \overline{\Phi}(X_i)| = \mathcal{O}\left(\frac{\ln n}{n}\right)^{2/3d'} \text{ e.a.s.}$$

Combining this with (26), we obtain (25).

Next, Φ being continuous on the compact set M, and in view of assumptions (3), the gradient of Φ vanishes at the (unique) minimum point x^*, hence there exist $r_0 > 0$, $c_0 > 0$, $c_1 > 0$ and ε_0 such that

$$\forall x \in M \cap B^c,\ \Phi(x) \geq \Phi(x^*) + \varepsilon_0, \tag{27}$$

$$\forall x \in M \cap \overline{B},\ c_0||x - x^*||^2 \leq \Phi(x) - \Phi(x^*) \leq c_1||x - x^*||^2, \tag{28}$$

where $B = \mathcal{B}(x^*, r_0)$ is the open ball in $\mathbb{R}^d$ of center x^* and radius r_0. The second inequality holds due to the positiveness of the Hessian matrix $H_\Phi(x^*)$.

Now, since the assumptions of Corollary 2 are satisfied, there exists $C > 0$ such that $d_h(\mathcal{X}_n, M) \leq C(\ln n/n)^{1/d'}$. Thus, *e.a.s.*

$$\exists i_0 \in \{1, \ldots, n\}\,;\, \|X_{i_0} - x^*\| \leq C\left(\frac{\ln n}{n}\right)^{1/d'}.$$

For n large enough, we have $r_0 > C\,(\ln n/n)^{1/d'}$, hence, in view of (28), X_{i_0} satisfies

$$\Phi(X_{i_0}) \leq \Phi(x^*) + c_1 C^2 \left(\frac{\ln n}{n}\right)^{2/d'},$$

and by (25):

$$\hat{\Phi}(X_{i_0}) \leq \Phi(x^*) + c_1 C^2 \left(\frac{\ln n}{n}\right)^{2/d'} + D\left(\frac{\ln n}{n}\right)^{2\alpha},$$

with

$$\alpha = \min\left\{\frac{1}{3d'}, \frac{1}{4}\right\},$$

that is, for n large enough, we have

$$\exists i_0 \in \{1, \ldots, n\}, \quad \hat{\Phi}(X_{i_0}) \leq \Phi(x^*) + 2D\left(\frac{\ln n}{n}\right)^{2\alpha}. \tag{29}$$

Assume now that n is large enough so that that $\varepsilon_0 > 4D(\ln n/n)^{2\alpha}$. For any $i \in \{1, \ldots, n\}$ such that

$$\|X_i - x^*\| \geq \frac{2D}{\sqrt{c_0}} \left(\frac{\ln n}{n}\right)^{\alpha},$$

in view of (27) and (28), this point satisfies

$$\Phi(X_i) \geq \Phi(x^*) + 4D \left(\frac{\ln n}{n}\right)^{2\alpha}.$$

Thus, in view of (25), we have

$$\|X_i - x^*\| \geq \frac{2D}{\sqrt{c_0}} \left(\frac{\ln n}{n}\right)^{\alpha} \Rightarrow \hat{\Phi}(X_i) \geq \Phi(x^*) + 3D \left(\frac{\ln n}{n}\right)^{2\alpha}. \tag{30}$$

Finally, let $i^* \in \{1, \ldots, n\}$ such that X_{i^*} realizes the minimum (4). From (24) and (29), it is clear that

$$\hat{\Phi}(X_{i^*}) \leq \Phi(x^*) + 2D \left(\frac{\ln n}{n}\right)^{2\alpha},$$

and (30) allows to conclude the proof.

References

[1] A. Baillo, A. Cuevas, and A. Justel. Set estimation and nonparametric detection. *The Canadian Journal of Statistics*, 28:765–782, 2000.

[2] M.R. Brito, A.J. Quiroz, and J.E. Yukich. Intrinsic dimension identification via graph-theoretic methods. *Journal of Multivariate Analysis*, 116:263–277, 2013.

[3] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, McGraw-Hill, 2001.

[4] A. Cuevas and A. Rodríguez-Casal. On boundary estimation. *Advanced in Applied Probability*, 36:340–354, 2004.

[5] P. Demartines and J. Herault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, 1997.

[6] D. Granata and V. Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Nature Scientific Reports*, 6(31377), 2016.

[7] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[8] J.A. Lee, A. Lendasse, and M. Verleyzen. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*, 57:49–76, 2004.

[9] M. Lennon, G. Mercier, M.C. Mouchot, and L. Hubert-Moy. Curvilinear component analysis for nonlinear dimensionality reduction of hyperspectral images. In S.B. Serpico, editor, *Image and Signal Processing for Remote Sensing VII*, volume 4541, 2001.

[10] J. Nilsson, T. Fioretos, M. Höglund, and M. Fontes. Approximate geodesic distances reveal biologically relevant structures in microarray data. *Bioinformatics*, 20(6):874–880, 2004.

[11] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

[12] M.D. Penrose. The longest edge of the random minimal spanning tree. *The Annals of Applied Probability*, 7(2):340–361, 1997.

[13] M.D. Penrose. A strong law for the largest nearest-neighbour link between random points. *Journal of the London Mathematical Society (Second Series)*, 60(3):951–960, 1999.

[14] L.K. Saul and S.T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2002.

[15] F. Takens. On the numerical determination of the dimension of an attractor. In *Dynamical systems and bifurcations (Groningen, 1984)*, volume 1125 of *Lectures Notes in Math.*, pages 99–106. Springer, Berlin, 1985.

[16] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality a global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.