# The Concentration of Fractional Distances

Damien François, Vincent Wertz, *Member*, *IEEE*, and Michel Verleysen, *Senior Member*, *IEEE*

**Abstract**—Nearest neighbor search and many other numerical data analysis tools most often rely on the use of the euclidean distance. When data are high dimensional, however, the euclidean distances seem to concentrate; all distances between pairs of data elements seem to be very similar. Therefore, the relevance of the euclidean distance has been questioned in the past, and fractional norms (Minkowski-like norms with an exponent less than one) were introduced to fight the concentration phenomenon. This paper justifies the use of alternative distances to fight concentration by showing that the concentration is indeed an intrinsic property of the distances and not an artifact from a finite sample. Furthermore, an estimation of the concentration as a function of the exponent of the distance and of the distribution of the data is given. It leads to the conclusion that, contrary to what is generally admitted, fractional norms are not always less concentrated than the euclidean norm; a counterexample is given to prove this claim. Theoretical arguments are presented, which show that the concentration phenomenon can appear for real data that do not match the hypotheses of the theorems, in particular, the assumption of independent and identically distributed variables. Finally, some insights about how to choose an optimal metric are given.

**Index Terms**—Nearest neighbor search, high-dimensional data, distance concentration, fractional distances.

## 1 Introduction

The search for nearest neighbors (NNs) is a crucial task in data management and data analysis. Content-based data retrieval systems, for example, use the distance between a query from the user and each element in a database to retrieve the most similar data [1]. However, the distances between elements can also be used to automatically classify data [2] or to identify clusters (natural groups of similar data) in the data set [3].

To define a measure of distance among data, the latter are often described in a euclidean space through a *feature vector*, and the euclidean distance is used. The euclidean distance is the euclidean norms of the difference between two vectors and is supposed to reflect the notion of similarity between them.

Nowadays, most data are getting more and more complex in the sense that a large number of features is needed to describe them; they are said to be high dimensional. For example, pictures taken by a standard camera consist of two to five million pixels, digital books contain thousands of words, DNA sequences are composed of tens of thousand bases, and so forth.

High-dimensional data must obviously be described in a high-dimensional space. In those spaces, however, the norm used to define the distance has the strange property to *concentrate* [4], [5]. As a consequence, all pairwise distances in a high-dimensional data set seem to be equal or at least very similar. This may lead to problems when searching for NNs [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. Therefore, the relevance of the norm, specifically of the euclidean norm, is questioned when measuring high-dimensional data similarity.

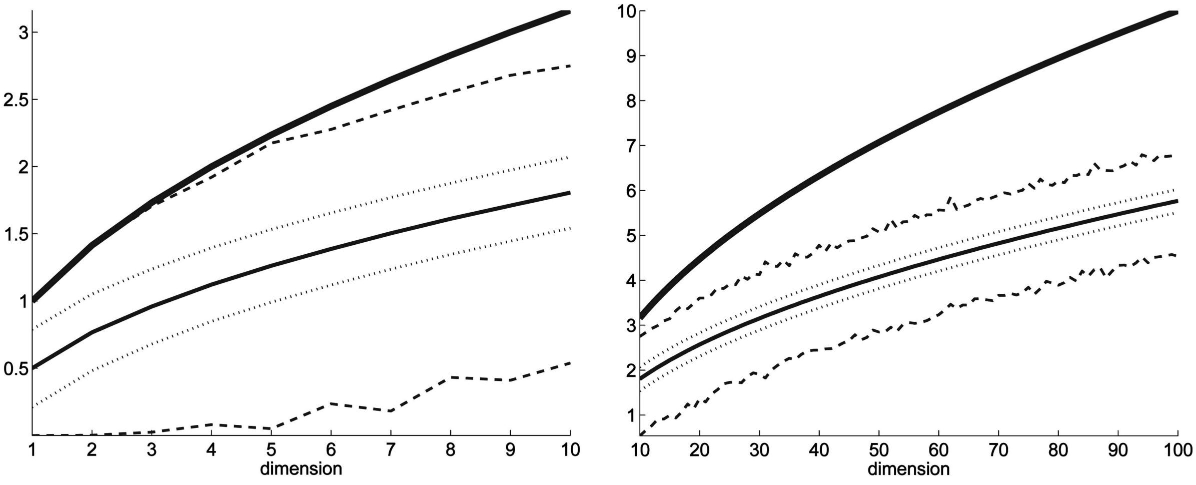
Some literature suggests using fractional norms, an extension of the euclidean norm, to counter concentration effects [23], [24]. Fractional norms have been studied and used [25], [26], [27], [28], [29], but many fundamental questions remain open about the concentration phenomenon. Does the concentration phenomenon occur when the data do not match the hypotheses of the theorems about concentration? Is the concentration phenomenon really intrinsic to the norm/distance, or can it be caused by some other counterintuitive effect of the high dimensionality? Are fractional norms always less concentrated than higher order norms? The aim of this paper is to provide answers to these questions. In particular, this paper will show that the norm of normalized data will concentrate even if the latter do not match the independent and identically distributed (i.i.d.) hypothesis. It will also show that the concentration phenomenon occurs even when an infinite number of data points are considered. Finally, this paper shows that, in contrast to what is generally acknowledged in the literature, fractional norms are not always less concentrated than higher order norms.

This paper is organized as follows: Section 2 introduces the concentration phenomenon and reviews the state of the art. Section 3 evokes the link between the concentration of the norm and the curse of dimensionality in database indexing. Section 4 discusses the limitations of current results and extends them. For the sake of clarity, this section will only present the new results; proofs and illustrations are gathered in Section 5. Finally, Section 6 proposes some ways to choose an optimal metric in particular situations.

- *D. François and V. Wertz are with the Department of Mathematical Engineering, Université catholique de Louvain, Georges Lemaître, 4, B-1348 Louvain-la-Neuve, Belgium.*
*E-mail: {francois, wertz}@inma.ucl.ac.be.*
- *M. Verleysen is with the Microelectronic Laboratory, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium. E-mail: verleysen@dice.ucl.ac.be.*

*Manuscript received 22 Aug. 2005; revised 15 Sept. 2006; accepted 11 Jan. 2007; published online 5 Mar. 2007.*
*For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0327-0805.*
*Digital Object Identifier no. 10.1109/TKDE.2007.1037.*

1041-4347/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society



(a)

(b)

Fig. 1. From bottom to top: minimum observed value, average minus standard deviation, average value, average plus standard deviation, maximum observed value, and maximum possible value of the Euclidean norm of a random vector. The expectation grows, but the variance remains constant. A small subinterval of the domain of the norm is reached in practice.

## 2 State of the Art

The concentration of distances in high-dimensional spaces is a rather counterintuitive phenomenon. It can be roughly stated as follows: In a high-dimensional space, all pairwise distances between points seem identical. This paper will study the concentration of the distances through the concentration of the norm of random vectors, as done in [4], [24], [30], and [31]. Given a data set supposedly drawn from a random variable Z, we consider Z$'$ that is distributed as Z and define X $=$ Z $-$ Z$'$. The probability density function of X is the convolution of the respective probability density functions of Z and $-$Z$'$. Studying the distribution of distances in the data set produced by Z, $\|\mathrm{Z}-\mathrm{Z}'\|$, is equivalent to studying the distribution of $\|\mathrm{X}\|$ the norm of X. Similarly, studying the pairwise distances in a data set with $n$ elements can be done by first building another data set with $n(n+1)/2$ elements corresponding to all pairs, whose attributes are the differences between two elements and then studying the norm of this new data.

This section illustrates the phenomenon and gives an overview of the main known results about the concentration of the norm/distance.

### 2.1 An Intuitive View of the Concentration Phenomenon

The following experiment will help in introducing the concept of concentration for the norm in high-dimensional spaces. The aim is to observe that the norms of high-dimensional vectors tend to be very similar to one another.

Let $\mathrm{X} = (X_1, \cdots, X_d)$ be a $d$-dimensional random vector taking values in the unit cube $[0,1]^d$. Each component $X_i$ will be referred to as *variable* $X_i$. We will denote by $\mathrm{X} \sim F$ a random vector X distributed according to the multivariate probability density function $F$. Let $\chi = \{x^{(j)}\}_{j=1}^{n} \subset \mathbb{R}^d$ be a finite sample drawn from X, that is, a set of independent realizations of X.

We consider the set $\{\|x^{(j)}\|\}_{j=1}^{n}$ of all the norms of the $x^{(j)}$. Obviously, the values of $\|x^{(j)}\|$ are bounded: $\|x^{(j)}\| \in [0, M]$, where $M = \|(1,\ldots,1)\|$.

In low-dimensional spaces ($d < 10$), if $n$ is not too small, $\min_j\{\|x^{(j)}\|\}_{j=1}^{n}$ will be close to zero, and $\max_j\{\|x^{(j)}\|\}_{j=1}^{n}$ will be close to $M$. However, in higher dimensional spaces ($d > 10$), this is not verified anymore. Fig. 1 shows the average value, empirical standard deviation, and maximum and minimum observed values of the norm of a uniformly randomly drawn sample of size $n = 10^5$ in spaces of growing dimension. The euclidean norm is considered, so $M = \sqrt{d}$.

We can observe that the average value of the norm increases with the dimension, whereas the standard deviation seems rather constant. When the dimension is low (Fig. 1a), we can see that the minimum and maximum observed values are close to the bounds of the domain of the norm, respectively, 0 and $\sqrt{d}$.

When the dimension is large, say, from dimension 10 onward, the maximum and minimum observed values tend to move away from the bounds. Indeed, even with a large number of points ($10^5$), all the observed norms seem to concentrate in a small portion of their domain. In addition, this portion gets smaller and smaller as the dimension grows, when compared to the size of the total domain.

This phenomenon is referred to as the *concentration of the norm*. Section 2.3 will review some results from the literature about the concentration of the norm in high-dimensional spaces; Minkowski norms will be introduced in Section 2.2.

### 2.2 The Euclidean Norm and the Minkowski Family

The Minkowski norms form a family of norms parametrized by their exponent $p = 1, 2, \ldots$:

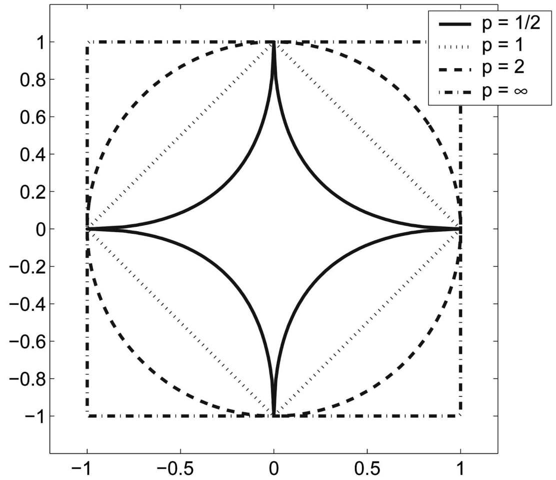


Fig. 2. Two-dimensional unit balls for several values of the parameter of the $p$-norm.

$$\|X\|_p = \left(\sum_i |X_i|^p\right)^{\frac{1}{p}}. \qquad (1)$$

When $p = 2$, the Minkowski norm corresponds to the euclidean one, which induces the euclidean distance. For $p = 1$, it is sometimes called "sum-norm" and induces the Manhattan or city-block metric. The limit for $p \to \infty$ is the "max-norm," which induces the Chebychev metric.

In the sequel, we will consider extensions of Minkowski norms to the case where $p$ is a positive real number. For $p \geq 1$, those extensions are indeed norms, but for $0 < p < 1$, the triangle inequality does not hold and, hence, they do not deserve the name; they are sometimes called prenorms. Actually, the inequality is reversed. A consequence is that the straight line is no longer the smallest path between two points, which may seem counterintuitive. In the remainder, we will denote $p$-norm a norm or prenorm of the form (1) with $p \in \mathbb{R}^+$. We will call a *fractional norm* a $p$-norm with $p < 1$.

Fig. 2 depicts the 2D unit balls (that is the set of $x^{(j)}$ for which $\|x^{(j)}\| = 1$) for values of $p$ equal to $\frac{1}{2}$, 1, 2, and infinity. We can see that, for $p \geq 1$, the balls are convex; for $0 < p < 1$, however, they are not. In Sections 2.3, 2.4, and 2.5, results from the literature about the concentration of Minkowski norms and fractional $p$-norms will be presented.

## 2.3 Concentration of the Euclidean Norm

In the experiment described at the beginning of this section, we observed that the expectation of the norm of a random vector increases with the dimension, whereas its standard deviation (and, hence, its variance) remains rather constant. Demartines [4] has theoretically confirmed this fact.

**Theorem 1: Demartines.** *Let* $\mathrm{X} \in \mathbb{R}^d$ *be a random vector with i.i.d. components:* $X_i \sim \mathcal{F}$. *Then,*

$$\mathrm{E}(\|X\|_2) = \sqrt{ad - b} + O(1/d) \text{ and} \qquad (2)$$

$$\mathrm{Var}(\|X\|_2) = b + O(1/\sqrt{d}), \qquad (3)$$

*where $a$ and $b$ are constants that do not depend on the dimension.*

The theorem is valid whatever the distribution $\mathcal{F}$ of the $X_i$ might be. Different distributions will lead to different values for $a$ and $b$, but the asymptotic results remain. The theorem proves that the expectation of the euclidean norm of random vectors increases as the square root of the dimension, whereas its variance is constant and independent of the dimension. Therefore, when the dimension is large, the variance of the norm is very small compared with its expected value.

Demartines concludes that when the dimension is large, vectors seem normalized: The relative error made while considering $\mathrm{E}(\|X\|_2)$ instead of the real value of $\|X\|_2$ becomes negligible. As a consequence, high-dimensional vectors appear to be distributed on a sphere of radius $\mathrm{E}(\|X\|_2)$. Demartines also notes that, since the euclidean distance is the norm of the difference between two random vectors, its expectation and variance follow laws (2) and (3); pairwise distances between points in high-dimensional spaces seem to be all identical.

Demartines mentions that if $X_i$ are not independent, the results are still valid provided that we replace $d$ with the actual number of "degrees of freedom."

The result from the work of Demartines is interesting in that it confirms the experimental results, but it is restricted to the euclidean norm and makes the rather strong hypothesis of independence and identical distributions.

## 2.4 Concentration of Arbitrary Norms

Independent of the results of Demartines' work, Beyer et al. explored the effect of dimensionality on the NN problem [5].

Whereas Demartines defines a data set $\chi$ as consisting of $n$ independent draws $x^{(j)}$ from a single random vector $X$, Beyer et al. consider $n$ random vectors $\mathrm{P}^{(j)}$; a data set is then made of one realization of each random vector.

The main result of Beyer et al.'s work is the following theorem. The original theorem is stated for an arbitrary distance measure; it is rewritten here with norms for illustration purposes.

**Theorem 2: Beyer et al., adapted.** *Let* $\mathrm{P}^{(j)} : 1 \leq j \leq n$ *be* $n$ *d-dimensional i.i.d. random vectors. If*

$$\lim_{d\to\infty} \mathrm{Var}\left(\frac{\|\mathrm{P}^{(j)}\|}{\mathrm{E}(\|\mathrm{P}^{(j)}\|)}\right) = 0 \qquad (4)$$

*then, for any* $\epsilon > 0$,

$$\lim_{d\to\infty} \mathbf{P}\left[\frac{\max_j \|\mathrm{P}^{(j)}\| - \min_j \|\mathrm{P}^{(j)}\|}{\min_j \|\mathrm{P}^{(j)}\|} \leq \epsilon\right] = 1.$$

The theorem is interpreted as follows: Suppose a set of $n$ data points, randomly distributed in the $d$-dimensional space. Some query point is supposed to be located at the origin, without loss of generality. Then, if hypothesis (4) is satisfied, independent of the distribution of the components of the $\mathrm{P}^{(j)}$, the difference between the largest and smallest distances to the query point becomes smaller and smaller when compared with the smallest distance when the dimension increases. The ratio

$$\frac{\max_j \|\mathrm{P}^{(j)}\| - \min_j \|\mathrm{P}^{(j)}\|}{\min_j \|\mathrm{P}^{(j)}\|}$$

is called the *relative contrast*.

Beyer et al. conclude that all points seem to be located at approximately the same distance from the query point; the concept of NN in a high-dimensional space is then less intuitive than in a lower dimensional one.

Beyer et al. explore some scenarios that satisfy (4) and some that do not. A proof that Minkowski norms and fractional norms satisfy the hypothesis will be given in Section 4.1.

### 2.5 Concentration of Minkowski Norms

Hinneburg et al. focused on the search for NN as well [23]. They produced the following theorem relative to Minkowski norms.

**Theorem 3: Hinneburg et al.** *Let* $\mathrm{P}^{(j)} : 1 \leq j \leq n$ *be* $n$ *d-dimensional i.i.d. random vectors and* $\|.\|_p$ *be the Minkowski norm with exponent p. If the* $\mathrm{P}^{(j)}$ *are distributed over* $[0,1]^d$, *then there exists a constant* $C_p$ *independent of the distribution of the* $\mathrm{P}^{(j)}$ *such that*

$$C_p \leq \lim_{d\to\infty} \mathrm{E}\left(\frac{\max_j \|\mathrm{P}^{(j)}\|_p - \min_j \|\mathrm{P}^{(j)}\|_p}{d^{\frac{1}{p}-\frac{1}{2}}}\right) \leq (n-1)\cdot C_p. \tag{5}$$

Hinneburg et al. note that a consequence of (5) is the surprising fact that, on average, the *contrast*

$$\max_j \|\mathrm{P}^{(j)}\|_p - \min_j \|\mathrm{P}^{(j)}\|_p \tag{6}$$

grows as $d^{1/p-1/2}$. They further conclude that, as a result, the contrast converges to a constant when the dimension increases and when the euclidean distance is used. For the $L_1$ norm, it increases as $\sqrt{d}$, for the euclidean norm ($p = 2$), it remains constant, and for norms with $p \geq 3$, it tends toward zero. Hinneburg et al. conclude that, for $L_p$ metrics with $p \geq 3$, the NN search in a high-dimensional space tends to be meaningless for Minkowski norms with exponent larger than or equal to 3, since the maximum observed distance tends toward the minimum one. The distance has lost its discriminative power between the notions of "close" and "far."

The conclusion we can draw from this theorem is the following: On average, the ratio between the contrast and $d^{1/p-1/2}$ is bounded. However, the bounds themselves depend on the value of $p$. Furthermore, if the number of points $n$ is large, the upper bound may be very large too. In practice, though, it may appear that the value of the ratio is much closer to the lower bound than to the upper one. A result independent of the number of points is presented in Section 4.2.

Yianilos [31] mentions that the standard deviation of the $p$-norm, $p \geq 1$, of a uniformly distributed random vector is $\Theta(d^{1/p-1/2})$ and sketches a proof. In contrast to the approaches by Beyer et al. and Hinneburg et al., Yianilos considers random vector distributions rather than realizations of random vectors. He also asserts that the *i.i.d.* hypothesis can be weakened. He then makes the same conclusions as Hinneburg et al. and investigates the consequences for excluded middle vantage point forests methods for similarity search in metric spaces. We will see a similar result in Section 4 without the restriction of the uniform distribution and on the values of $p$.

### 2.6 Concentration of Fractional Norms

Aggarwal et al. extend Hinneburg's result to fractional $p$-norms [24]. In a sense, if $p = 2$ is "better" than $p = 3$ and $p = 1$ is "better" than $p = 2$, why not to look at $p = \frac{1}{2}$ to see if it is "better" than $p = 1$? Aggarwal et al. produced the following theorem.

**Theorem 4: Aggarwal et al.** *Let* $\mathrm{P}^{(j)} : 1 \leq j \leq n$ *be* $n$ *d-dimensional independent random vectors, uniformly distributed over* $[0,1]^d$. *There exists a constant C independent of p and d such that*

$$C\sqrt{\frac{1}{2p+1}} \leq \lim_{d\to\infty} \mathrm{E}\left(\frac{\max_j \|\mathrm{P}^{(j)}\|_p - \min_j \|\mathrm{P}^{(j)}\|_p}{\min_j \|\mathrm{P}^{(j)}\|_p}\right)\cdot\sqrt{d} \leq (n-1)\cdot C\cdot\sqrt{\frac{1}{2p+1}}.$$

Aggarwal et al. note that the constant $\sqrt{1/(2p+1)}$ may play a valuable role in affecting the relative contrast and confirmed it experimentally with synthetic data sets. They then conclude that, on average, fractional norms provide better contrast than Minkowski norms in terms of relative distances.

The next section will provide a similar result independent of the number of points. That result will show that the conclusions of Theorem 4 cannot be extended to the case where the data are not uniformly distributed.

Independent of Aggarwal et al., Skala [30] has shown that the ratio

$$\rho_p(d) = \frac{\mathrm{E}(\|X\|_p)^2}{2\mathrm{Var}(\|X\|_p)}, \tag{7}$$

increases linearly with dimension $d$. Here, $X$ is a random vector whose components are *i.i.d.* distributed. The theorem does not require uniform densities; however, it is exact only in the $p = 1$ case and gives an approximate result for other values of $p$.

## 3 Concentration and Similarity Search in Databases

In order to better understand the impact of concentration on NN search and indexing, this section reviews some example studies available in the literature. It further describes the context of this research and justifies the importance of the study of concentration and alternative metrics.

The major consequence of concentration on NN search is that indexing methods, which have an expected logarithmic cost, actually sometimes perform no better than simple linear scanning. This has been described among others in [32], [33], [34], and [35] and was acknowledged by many others as the *curse of dimensionality*.

Consequently, new indexing structures, specially designed for high-dimensional data were suggested, like the X-tree [34], TV-tree [35], and so forth, and an approximate search was proposed as a solution [36]. A survey of these methods can be found in [37].

It seems that Brin [38] was one of the pioneers to actually relate the curse of dimensionality to the particular shape of distance distributions in high-dimensional spaces. He proposed the Geometric Near-neighbor Access Tree (GNAT) indexing structure that was built taking into account the minimum and maximum distances in the data set.

Whereas Berchtold et al. [39] suggested that the decrease of performances is due to the boundary effects, coming from the fact that all points seem located in the "corners" of the space, Weber et al. [40] proposes a review of counterintuitive properties of high-dimensional geometrical spaces and relates them to the losses of performances.

In view of their theorem (Theorem 2), Beyer et al. explain the decreases of performances of the indexing methods by the instability of the search for NNs. They propose to use less concentrated metrics and further question the intrinsic relevance of NN search, *independent of performance issues*, for concentrated metrics.

Whereas many of the responses to the curse of dimensionality are to introduce a new indexing method, Hinneburg et al. [23] and Aggarwal et al. [24] propose to use alternative metrics that are less concentrated. Their main purpose, however, is to use a more *"meaningful"* metric in high-dimensional spaces and not to accelerate the search.

In both cases, using less-concentrated norms either to perform better with indexing structures or to get a more meaningful NN search thus assumes that the concentration phenomenon is intrinsic to the notion of the norm.

The aim of this paper is to show that the concentration is indeed intrinsic to the norm. To this end, we will consider results that are independent of the number of points. Considering a high-dimensional bounded space with some points scattered over it, it seems reasonable to think that the less points you have, the more concentrated the distances are akin to be since the maximum distance will decrease while the minimum distance will increase. The "number of points" parameter either has to be taken into account or has to be ruled out of the equation. In the following, we propose to do this by considering data distributions instead of data sets. The conclusion of our study is that, indeed, the concentration phenomenon is intrinsic to the norm and, thus, justifies Aggarwal et al.'s work on fractional metrics to reduce concentration.

From Aggarwal et al.'s results, it is often extrapolated that using fractional distances with small values of the $p$ parameter will decrease the concentration phenomenon. The latter fact has subsequently been used as an argument to use fractional norms in all sorts of situations without first checking for applicability. However, the restriction is that the data must be uniformly distributed. As we are dealing here with high-dimensional spaces, it appears quite evident that real data can only sparsely populate high-dimensional spaces because of their finite number and because they often lie on a submanifold. Therefore, data are far from being uniformly spread over the space. This study confirms that the distribution of data has to be taken into account to estimate the concentration. Depending on the distribution, the evolution of the concentration with respect to the value of parameter $p$ may be increasing as $p$ gets smaller; in other cases, it also may have a local maximum for some value of $p$, as will be shown later.

## 4 Further Theoretical Results

Section 2 reviewed major results from the literature. This section will describe new results in order to better understand the phenomenon of norm concentration in high-dimensional spaces. All proofs and examples are postponed to Section 5 for ease of readability.

### 4.1 The Finite and Bounded Sample

A finite number of points will most probably be sparsely distributed in a high-dimensional space. Most points will be far away from each other, and the density will be very low over the whole space. This is referred to as the *Empty Space Phenomenon* [41]. Furthermore, if the points live in a closed (bounded) region of the space, for instance, the $[0,1]^d$ hypercube, then the maximum distance is bounded too. In such a situation, it may happen that the relative contrast is very low. The questions are then: Is the concentration phenomenon a side effect of the Empty Space Phenomenon, just because we consider a finite number of points in a bounded portion of a high-dimensional space? Would the conclusions still hold if an infinite number of points (in other words a distribution) spanning the whole space was considered?

Unfortunately, the results by Beyer et al., Hinneburg et al., and Aggarwal et al. cannot be extended to the case where the number of data points is arbitrarily large. Indeed, the bounds on the relative contrast depend on the number of points. Furthermore, if the values the data points $\mathrm{P}^{(j)}$ can take are unbounded, the notion of relative contrast may not be relevant anymore since it relies on maximum and minimum values.

In contrast to Beyer et al.'s, Hinneburg et al.'s, and Aggarwal et al.'s results, Demartines' and Yianilos' ones do not refer to a finite number of points but rather to a distribution. Unfortunately, they consider Minkowski norms only. An interesting result would be to extend their results to fractional $p$-norms.

For that purpose, it is proposed to use the ratio

$$RV_{\mathcal{F},p} = \frac{\sqrt{\mathrm{Var}(\|X\|_p)}}{\mathrm{E}(\|X\|_p)} \tag{8}$$

as a measure of the concentration; $RV_{\mathcal{F},p}$ will be called the *relative variance* of the norm. It is nonnegative; a small value indicates that the distribution of the norm is concentrated, whereas a larger one indicates a wide effective range of norm values. Intuitively, we can see that $RV_{\mathcal{F},p}$ measures the concentration by relating a measure of spread (standard deviation) to a measure of location (expectation). In that sense, it is similar to the relative contrast that also relates a measure of spread (range) to a measure of location (minimum).

The main result of this section is that, regarding the relative variance, all $p$-norms concentrate as the dimensionality of the space increases. This is stated more precisely in Theorem 5, deduced from two lemmas that, respectively, characterize the individual behaviors of the variance and of the expectation of the norm. Those lemmas are presented here while their respective proofs can be found in Section 5.1.

**Lemma 1.** *Let* $\mathrm{X} = (X_1, \cdots, X_d)$ *be a random vector with i.i.d. components:* $X_i \sim \mathcal{F}$. *Then,*

$$\lim_{d\to\infty} \frac{\mathrm{E}(\|X\|_p)}{d^{1/p}} = c \tag{9}$$

*with $c$ being a constant independent of the dimension but related to $p$ and to the distribution $\mathcal{F}$.*

Therefore, it appears that the expectation of the $p$-norm of a random vector grows as the $p$th root of the dimension. Note that this result is consistent with (2) in the euclidean case $(p = 2)$.

The second lemma concerns the variance of the norm.

**Lemma 2.** *Let $X = (X_1, \cdots, X_d)$ be a random vector with i.i.d. components: $X_i \sim \mathcal{F}$. Then,*

$$\lim_{d\to\infty} \frac{\mathrm{Var}(\|X\|_p)}{d^{\frac{2}{p}-1}} = c' \tag{10}$$

*with $c'$ being a constant independent of the dimension but related to $p$ and to the distribution $\mathcal{F}$.*

The dependency on $d^{\frac{2}{p}-1}$ indicates that the variance remains constant with the dimension for the euclidean distance. We can also see that the variance decreases when the dimension increases for $p$-norms with $p > 2$ and increases for $p$-norms with $p < 2$. This is consistent with Demartines' and Hinneburg et al.'s results.

From these Lemmas, it can be shown that all $p$-norms concentrate as the dimension increases.

**Theorem 5.** *Let $X = (X_1, \cdots, X_d)$ be a random vector with i.i.d. components: $X_i \sim \mathcal{F}$. Then,*

$$\lim_{d\to\infty} \frac{\sqrt{\mathrm{Var}(\|X\|_p)}}{\mathrm{E}(\|X\|_p)} = 0,$$

*that is, the relative variance decreases toward zero when the dimension grows.*

The proof can be found in Section 5.1. From Theorem 5, it can be concluded that the concentration of the norms in high-dimensional spaces is really an intrinsic concentration property of the norms and not a side effect of the finite sample size or from the Empty Space Phenomenon. This result extends Demartines' one to all $p$-norms and does not depend on the sample size.

Moreover, although all $p$-norms concentrate as the dimension increases, they do not concentrate in the same way. Characterizing these differences with respect to $p$ is the topic of Section 4.2.

### 4.2 Impact of the Value of $p$ on the Concentration

In the previous section, it has been shown that all $p$-norms concentrate, whatever is the value of $p (p > 0)$. In this section, the relationship between the relative variance and the value of $p$ for a given dimension will be made explicit.

**Theorem 6.** *Let $X = (X_1, \cdots, X_d)$ be a random vector with i.i.d. components: $X_i \sim \mathcal{F}$. Then,*

$$\lim_{d\to\infty} \sqrt{d} \cdot \frac{\sqrt{\mathrm{Var}(\|X\|_p)}}{\mathrm{E}(\|X\|_p)} = \frac{1}{p}\frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}},$$

*where $\mu_{\mathcal{F},p} = \mathrm{E}(|X_i|^p)$, and $\sigma^2_{\mathcal{F},p} = \mathrm{Var}(|X_i|^p)$.*

If $d$ is large, we can thus approximate the relative variance by

$$RV_{\mathcal{F},p} = \frac{\sqrt{\mathrm{Var}(\|X\|_p)}}{\mathrm{E}(\|X\|_p)} \simeq \frac{1}{\sqrt{d}} \cdot \left(\frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}}\right).$$

This means that, for a fixed (large) dimension, the relative variance evolves with $p$ as

$$\mathcal{K}_{\mathcal{F},p} = \frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}}. \tag{11}$$

The evolution of $\mathcal{K}_{\mathcal{F},p}$ with $p$ determines the value of the relative variance for a fixed dimension $d$. Aggarwal et al. has shown that, if the points are uniformly distributed, the relative contrast at fixed dimension increases as $p$ decreases. We will show below and prove in Section 5.3 that, under the uniform distribution hypothesis, the relative variance increases as $p$ decreases too.

However, in general, the function described in (11) is not always strictly decreasing with $p$ as stated in the following proposition:

**Proposition 1.** *The relative variance (11) is 1) a strictly decreasing function of $p$ when the variables are distributed according to a uniform distribution $\mathcal{F}$ over the interval [0, 1], but 2) there exists distributions $\mathcal{F}$ for which it is not the case. Fractional norms are not always less concentrated than higher order norms.*

The proof is given in Section 5.2. There are thus data for which the 1-norm is more concentrated than the 2-norm, for instance; in general, a higher order norm can be less concentrated than a fractional norm by having a higher relative contrast or relative variance. In conclusion, using fractional norms does not always bring less concentrated distances.

### 4.3 The i.i.d. Hypothesis

All theorems presented in Section 4 rely on the fact that data are supposed to be i.i.d. What happens if it is not the case in practice (actually it will hardly ever be)? Are these assumptions really needed, or do they merely make the proofs easier or even feasible? This section addresses these issues.

First, the "identically distributed" assumption is considered. In the previous sections, we have supposed that all $X_i$ are distributed according to the same distribution: $X_i \sim \mathcal{F}$. If variables $X_i$ are not identically distributed, it means that each $X_i$ is distributed according to some distribution noted $\mathcal{F}_i$: $\forall i : X_i \sim \mathcal{F}_i$.

**Proposition 2.** *If the data are not identically distributed, then the conclusion of Theorem 5 still holds provided that the data are normalized.*

The proof is given in Section 5.3. Normalizing means here to subtract the mean from the variables and divide them by their standard deviation so that $\forall i : \mathrm{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 1$. Normalizing data is often considered as important when using norms and distances, because it ensures that all variables $X_i$ will have equal influence on the computation of the norm. If this it is not the case, the variables with the largest variances will have the largest influence on the distance value, whereas the variables with low variances will have little or no influence on the computation of the norm.

Actually, norms will concentrate for nonidentically distributed data if they are normalized. However, if the data are not normalized, some variables may have too little effect on the distance value.

The "independent" assumption may not be encountered in practice. If $X$ has components $X_i$ that are not independent, it means that the joint distribution $F$ of $\mathrm{X} = (X_1, \cdots, X_d)$ differs from the product of the marginal distributions $\mathcal{F}_i$. The existence of relations between the elements of $X$ means that $X$ lies on a (possibly nonlinear) submanifold of $\mathbb{R}^d$. If the submanifold is nonlinear, the euclidean norm, for instance, measures the distances between data points using "short cuts," that is, through a straight line in the space; this can be very different from a geodesic distance measured along the manifold [42], [43]. However, the manifold can be represented in a vector space $\mathbb{R}^{d_{int}}$ whose dimension is the number of degrees of freedom of the manifold. Dimension $d_{int}$ is called the *intrinsic dimension* and $d$ the *embedding* dimension. Each realization of $X$ in the embedding space may then be mapped to a unique realization in a $d_{int}$-dimensional projection space.

One may thus expect that the asymptotic properties of the norms behave similarly in both spaces. Actually, the measure of intrinsic dimensionality $\rho$ proposed by Chavez et al. [1] is precisely the inverse of the square of the relative variance. Korn et al. [22] relate the concentration to the fractal dimension of the data, which is another way to measure its intrinsic dimensionality.

The fact that the concentration depends on the intrinsic dimension of the data is often admitted [2], even if there is no consensus about the actual definition of the intrinsic dimension.

As a consequence, high-dimensional data that present a lot of correlation or dependencies between variables will be much less concentrated than if all variables are independent. The conclusion we can draw is that the concentration phenomenon depends on the intrinsic dimension of the data more than on the dimension of the embedding space.

## 5 Proofs and Illustrations

In this section, proofs are given for Lemmas 1 and 2, for Theorems 5 and 6, and for Propositions 1 and 2.

### 5.1 Proof of Theorem 5

The proof of Theorem 5 is based on Lemmas 1 and 2. Therefore, the proofs of the lemmas are given first. □

#### *5.1.1 Proof of Lemma 1*

The proof requires two steps. First, it will be shown that, under the assumptions of Lemma 1, we have

$$\mathbf{P}\left[\lim_{d\to\infty}\frac{\|X\|_p}{d^{1/p}} \quad = \quad c\right] = 1, \tag{12}$$

where $c$ is a constant independent of the dimension $d$ of $X$ (Step 1). Then, this result will be extended to the expectation of the norm (Step 2).

**Step 1**. Let $S_i = |X_i|^p$ for $i = 1, \ldots, d$. The $S_i$ are i.i.d. as well; let $\mu_{\mathcal{F},p}$ be their expectation.

The Strong Law of Large Numbers (SLLN) allows us to write

$$\mathbf{P}\left[\lim_{d\to\infty}\frac{1}{d}\cdot\sum_{i=1}^{d} S_i = \mu_{\mathcal{F},p}\right] = 1.$$

Then,

$$\mathbf{P}\left[\lim_{d\to\infty}\left(\frac{1}{d}\cdot\sum_{i=1}^{d} S_i\right)^{1/p} = {\mu_{\mathcal{F},p}}^{1/p}\right] = 1$$

or, by definition of $S_i$,

$$\mathbf{P}\left[\lim_{d\to\infty}\frac{1}{d^{1/p}}\cdot\left(\sum_{i=1}^{d}|X_i|^p\right)^{1/p} = {\mu_{\mathcal{F},p}}^{1/p}\right] = 1.$$

This is nothing else than

$$\mathbf{P}\left[\lim_{d\to\infty}\frac{\|X\|_p}{d^{1/p}} = {\mu_{\mathcal{F},p}}^{1/p}\right] = 1,$$

which concludes the proof with $c = {\mu_{\mathcal{F},p}}^{1/p}$.

**Step 2**. By (12), for any realization $\xi$ of $X$ except some $\xi \in \Omega$, a subset of $\mathbb{R}^d$ with measure 0, we have

$$\lim_{d\to\infty}\frac{\|\xi\|_p}{d^{1/p}} = {\mu_{\mathcal{F},p}}^{1/p}.$$

Thus,

$$\begin{aligned}\int_{\mathbb{R}^d\backslash\Omega} F(\xi)\cdot\lim_{d\to\infty}\frac{\|\xi\|_p}{d^{1/p}}\mathrm{d}\xi \quad &= \quad \int_{\mathbb{R}^d\backslash\Omega} F(\xi)\cdot{\mu_{\mathcal{F},p}}^{1/p}\mathrm{d}\xi & (13)\\ &= \quad {\mu_{\mathcal{F},p}}^{1/p} & (14)\end{aligned}$$

with the last equality coming from the fact that ${\mu_{\mathcal{F},p}}^{1/p}$ is bounded. In addition, we have

$$\begin{aligned}\int_{\mathbb{R}^d\backslash\Omega} F(\xi)\cdot\lim_{d\to\infty}\frac{\|\xi\|_p}{d^{1/p}}\mathrm{d}\xi \quad &= \quad \int_{\mathbb{R}^d} F(\xi)\cdot\lim_{d\to\infty}\frac{\|\xi\|_p}{d^{1/p}}\mathrm{d}\xi & (15)\\ &= \quad \lim_{d\to\infty}\int_{\mathbb{R}^d} F(\xi)\cdot\frac{\|\xi\|_p}{d^{1/p}}\mathrm{d}\xi & (16)\end{aligned}$$

since $\Omega$ is of measure 0. Combining (14) and (16) gives

$$\lim_{d\to\infty}\int_{\mathbb{R}^d}\frac{\|\xi\|_p}{d^{1/p}}\mathrm{d}\xi = \lim_{d\to\infty}\frac{\mathrm{E}(\|X\|_p)}{d^{1/p}} = {\mu_{\mathcal{F},p}}^{1/p}, \tag{17}$$

which concludes the proof since the right-hand side of (17) is independent of the dimension $d$. □

#### *5.1.2 Proof of Lemma 2*

Let $\mu_{\mathcal{F},p}$ be the expectation of variable $|X_i|^p$ and $\sigma_{\mathcal{F},p}$ its variance. For random vectors $X$ of dimension $d$ and a given $p$-norm, we have

$$\begin{aligned}\frac{\mathrm{Var}\|X\|_p}{d^{2/p-1}} &= \frac{\mathrm{E}\left[\left(\|X\|_p - E\|X\|_p\right)^2\right]}{d^{2/p-1}}\\ &= \mathrm{E}\left[\left(\frac{\|X\|_p - E\|X\|_p}{d^{1/p-1/2}}\right)^2\right].\end{aligned}$$

Since

$$\|X\|_p - E\|X\|_p = \frac{\|X\|_p^p - \left(E\|X\|_p\right)^p}{\sum_{r=0}^{p-1}\|X\|_p^{p-r-1}\cdot\left(E\|X\|_p\right)^r},$$

we have

$$\frac{\|X\|_p - E\|X\|_p}{d^{1/p-1/2}} = \frac{\left(\|X\|_p^p - \left(E\|X\|_p\right)^p\right)/\sqrt{d}}{\sum_{r=0}^{p-1}\left(\frac{\|X\|_p}{d^{1/p}}\right)^{p-r-1}\cdot\left(\frac{E\|X\|_p}{d^{1/p}}\right)^r}.$$

Therefore,

$$\frac{\mathrm{Var}\|X\|_p}{d^{2/p-1}} = \mathrm{E}\left[\left(\frac{(\|X\|_p^p - \left(E\|X\|_p\right)^p)/\sqrt{d}}{\sum_{r=0}^{p-1}\left(\frac{\|X\|_p}{d^{1/p}}\right)^{p-r-1}\cdot\left(\frac{E\|X\|_p}{d^{1/p}}\right)^r}\right)^2\right].$$

Using the theorem by Lebesgue and Fatou about the convergence of integrable functions, we can swap the limit and the expectation operators:

$$\begin{aligned}&\lim_{d\to\infty}\frac{\mathrm{Var}\|X\|_p}{d^{2/p-1}}\\ &= \mathrm{E}\left[\lim_{d\to\infty}\frac{\left((\|X\|_p^p - \left(E\|X\|_p\right)^p)/\sqrt{d}\right)^2}{\left(\sum_{r=0}^{p-1}\left(\frac{\|X\|_p}{d^{1/p}}\right)^{p-r-1}\cdot\left(\frac{E\|X\|_p}{d^{1/p}}\right)^r\right)^2}\right]\\ &= \mathrm{E}\left[\frac{\lim_{d\to\infty}\left((\|X\|_p^p - \left(E\|X\|_p\right)^p)/\sqrt{d}\right)^2}{\lim_{d\to\infty}\left(\sum_{r=0}^{p-1}\left(\frac{\|X\|_p}{d^{1/p}}\right)^{p-r-1}\cdot\left(\frac{E\|X\|_p}{d^{1/p}}\right)^r\right)^2}\right].\end{aligned}$$

The transition from the former to the latter is allowed since the denominator does not tend toward zero.

Because of Lemma 1, we know that $\frac{\mathrm{E}\|X\|_p}{d^{1/p}}$ tends toward a constant as $d$ increases. Furthermore, we have seen that *almost surely*, that is, with probability 1, $\frac{\|X\|_p}{d^{1/p}}$ also tends toward a constant.

Therefore, the denominator almost surely tends toward a constant as the dimension grows:

$$\mathbf{P}\left[\lim_{d\to\infty}\sum_{r=0}^{p-1}\left(\frac{\|X\|_p}{d^{1/p}}\right)^{p-r-1}\cdot\left(\frac{E\|X\|_p}{d^{1/p}}\right)^r = p\cdot\mu_{\mathcal{F},p}^{(p-1)/p}\right] \tag{18}$$

is 1, which means that, with probability 1,

$$\lim_{d\to\infty}\frac{\mathrm{Var}\|X\|_p}{d^{2/p-1}} = \frac{\mathrm{E}\left[\lim_{d\to\infty}\left((\|X\|_p^p - \left(E\|X\|_p\right)^p)/\sqrt{d}\right)^2\right]}{\left(p\cdot\mu_{\mathcal{F},p}^{(p-1)/p}\right)^2}.$$

If we focus on the numerator now, we notice that

$$\begin{aligned}\|X\|_p^p - \left(E\|X\|_p\right)^p &= \sum_{i=1}^d |X_i|^p - \left(E\|X\|_p\right)^p\\ &= \sum_{i=1}^d\left(|X_i|^p - \frac{\left(E\|X\|_p\right)^p}{d}\right).\end{aligned}$$

Using the result from Lemma 1, we write

$$\lim_{d\to\infty}\left(\|X\|_p^p - \left(E\|X\|_p\right)^p\right) = \lim_{d\to\infty}\sum_{i=1}^d\left(|X_i|^p - \mu_{\mathcal{F},p}\right).$$

The numerator can now be written as

$$\mathrm{E}\left[\left(\sum_{i=1}^d\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\right)^2/d\right].$$

Since

$$\begin{aligned}\mathrm{E}\left[\sum_{i=1}^d\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\right] &= \sum_{i=1}^d\mathrm{E}\left[\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\right]\\ &= \sum_{i=1}^d\left(\mathrm{E}[|X_i|^p] - \mu_{\mathcal{F},p}\right)\\ &= \sum_{i=1}^d\left(\mu_{\mathcal{F},p} - \mu_{\mathcal{F},p}\right)\\ &= 0,\end{aligned}$$

we have

$$\mathrm{E}\left[\left(\sum_{i=1}^d\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\right)^2\right] = \mathrm{Var}\left(\sum_{i=1}^d\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\right).$$

However,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^d\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\right) &= \sum_{i=1}^d\mathrm{Var}\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\\ &= \sum_{i=1}^d\mathrm{Var}(|X_i|^p)\\ &= d\cdot\sigma_{\mathcal{F},p}^2\end{aligned}$$

leading to the conclusion that

$$\mathrm{E}\left[\left(\sum_{i=1}^d\left(|X_i|^p - \mu_{\mathcal{F},p}\right)\right)^2/d\right] = \sigma_{\mathcal{F},p}^2. \tag{19}$$

Using expression (19) for the numerator and (18) for the denominator, the result is that

$$\mathbf{P}\left[\lim_{d\to\infty}\frac{\mathrm{Var}\|X\|_p}{d^{2/p-1}} = \frac{\sigma_{\mathcal{F},p}^2}{(p\cdot\mu_{\mathcal{F},p}^{(p-1)/p})^2}\right] = 1. \tag{20}$$

The same arguments as in **Step 2** from the proof of Lemma 1 can be used to get

$$\lim_{d\to\infty} \frac{\mathrm{Var}\|X\|_p}{d^{2/p-1}} = \frac{\sigma^2_{\mathcal{F},p}}{\left(p \cdot \mu_{\mathcal{F},p}^{(p-1)/p}\right)^2}, \tag{21}$$

which proves the lemma since the right-hand side of (21) is independent of the dimension $d$. □

### 5.1.3 Proof of Theorem 5

Lemmas 1 and 2 are used to prove Theorem 5.

Writing

$$\frac{\sqrt{\mathrm{Var}\|X\|_p}}{\mathrm{E}\|X\|_p} = \frac{\frac{\sqrt{\mathrm{Var}\|X\|_p}}{d^{1/p-1/2}}}{\frac{\mathrm{E}\|X\|_p}{d^{1/p}}} \cdot d^{-1/2}$$

and taking the limit

$$\lim_{d\to\infty} \frac{\sqrt{\mathrm{Var}\|X\|_p}}{\mathrm{E}\|X\|_p} = \lim_{d\to\infty} \frac{\frac{\sqrt{\mathrm{Var}\|X\|_p}}{d^{1/p-1/2}}}{\frac{\mathrm{E}\|X\|_p}{d^{1/p}}} \cdot d^{-1/2},$$

we have, by Lemma 1 and Lemma 2

$$\lim_{d\to\infty} \frac{\sqrt{\mathrm{Var}\|X\|_p}}{\mathrm{E}\|X\|_p} = \frac{\sqrt{c'}}{c} \cdot \lim_{d\to\infty} d^{-1/2} = 0.$$

□

## 5.2 Proof of Theorem 6 and Proposition 1

Similar to Theorem 5, the proof of Theorem 6 is based on Lemmas 1 and 2.

Theorem 6 is proven as follows:

From Lemmas 1 and 2,

$$\lim_{d\to\infty} \sqrt{d} \cdot \frac{\sqrt{\mathrm{Var}\|X\|_p}}{\mathrm{E}\|X\|_p} = \lim_{d\to\infty} \frac{\frac{\sqrt{\mathrm{Var}\|X\|_p}}{d^{1/p-1/2}}}{\frac{\mathrm{E}\|X\|_p}{d^{1/p}}} = \frac{\sqrt{c'}}{c}. \tag{22}$$

Using the values of $c$ and $c'$, respectively, from (17) and (21), we have

$$\frac{\sqrt{c'}}{c} = \frac{\sqrt{\frac{\sigma_{\mathcal{F},p}}{\left(p \cdot \mu_{\mathcal{F},p}^{\frac{p-1}{p}}\right)^2}}}{\mu_{\mathcal{F},p}^{\frac{1}{p}}} = \frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}}. \tag{23}$$

Proposition 1 contains two claims; the proofs are, respectively, given in Part a and Part b.

**Part a**. If $\mathcal{F}$ is the uniform distribution over the interval [0, 1], $\mu_{\mathcal{F},p}$ is given by

$$\mu_{\mathcal{F},p} = \frac{1}{p+1}.$$

Since $\sigma^2_{\mathcal{F},p} = \mu_{\mathcal{F},2p} - {\mu_{\mathcal{F},p}}^2$, we have

$$\sigma_{\mathcal{F},p} = \frac{p}{p+1} \cdot \sqrt{\frac{1}{2p+1}}.$$

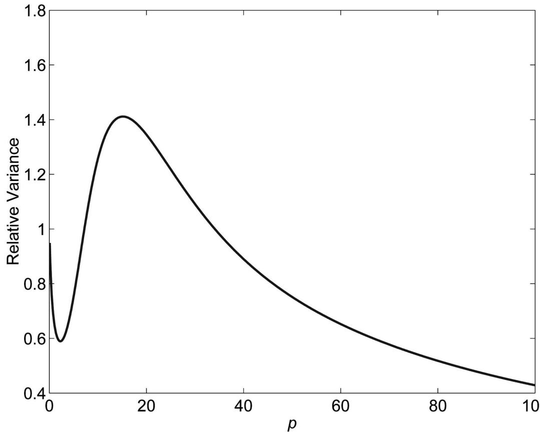


Fig. 3. Relative variance for data distributed as $\mathcal{F}^*$, as a function of $p$. We can see that a maximum is obtained for a rather large value of $p$, far from 1.

Therefore,

$$\frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}} = \sqrt{\frac{1}{2p+1}}. \tag{24}$$

We can conclude that, under the uniform distribution and large dimension $d$ hypotheses, the relative variance decreases with $p$. The concentration of the norm thus increases as $p$ grows.

**Part b**. A counterexample is provided to prove assertion b of the proposition. Let us consider a situation where data are dispatched into two Gaussian clusters with variance $\sigma^2 = 1$ and, respectively, centered on 1 and $-1$. The marginal distribution of each $X_i$ is then

$$\mathcal{F}^*(\xi) = \frac{1}{cst} \cdot \left(e^{-\left(\frac{\xi-1}{\sigma}\right)^2} + e^{-\left(\frac{\xi+1}{\sigma}\right)^2}\right). \tag{25}$$

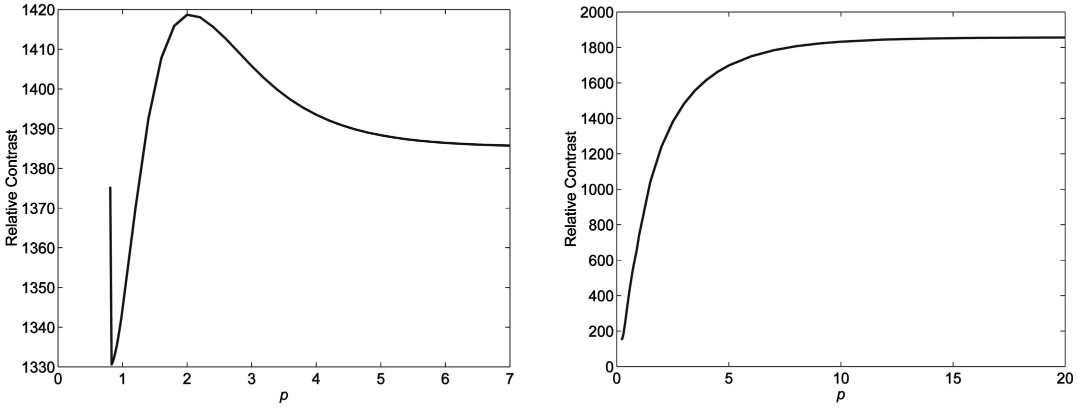
In this example, the relative variance is higher for higher order norms than for fractional norms, as illustrated in Fig. 3; consequently, fractional norms are more concentrated than higher order norms with values of $p \in [8, 30]$.

The next examples are taken from real data sets used by Aggarwal et al. in [24]: the segmentation data set and the Wisconsin Diagnostic Breast Cancer (WDBC) data set from the University of California, Irvine (UCI) Machine Learning Repository [44]. Fig. 4 shows a plot of the relative contrast as a function of $p$. In Fig. 4a (the segmentation data set), between $p = 1$ and $p = 2$, the relative contrast is actually increasing. This is a real example where the euclidean norm is actually less concentrated than the 1-norm. Fig. 4b is even more interesting. Here, the relative contrast is consistently better for higher order norms than for fractional norms. □

## 5.3 Proofs for Proposition 2

In Section 5.1, the proof of Theorem 5 was built using the Law of Large Numbers. Although this law is often stated with the *i.i.d.* hypothesis, the "identically distributed" one is sufficient but not necessary. Actually, if the data are normalized, the assumption of identical distributions is not necessary.

It has been shown that a less restrictive sufficient condition for the Law of Large Numbers to hold is that



(a)

(b)

Fig. 4. (a) Relative contrast for the segmentation and (b) the WDBC data sets. The relative contrast is higher for higher order norms than that for the fractional norms.

the set of $X_i$ must be uniformly integrable [45]. This means that, for any $\epsilon > 0$, there exists some $K < \infty$ such that for every $X_i$

$$\int_{|X_i| \geq K} \mathcal{F}_i(\xi)|\xi| d\xi < \epsilon. \tag{26}$$

Note that (26) is actually the expectation of $|X_i|$ restricted to the values $|X_i| \geq K$. It is equivalent to [45]

$$\exists k > 1 : \sup_i \mathrm{E}|X_i|^k < \infty. \tag{27}$$

If the data are normalized, $\mathrm{E}(X_i) = 0$, and $\mathrm{Var}(X_i) = 1$ for all $i$. We then have

$$1 = \mathrm{Var}\ X_i = \mathrm{E}\ X_i^2 - \mathrm{E}\ X_i^2 = \mathrm{E}\ X_i^2 = \mathrm{E}|X_i|^2.$$

Taking $k = 2$ in (27) leads to the conclusion that if the data are centered and reduced, as often done in practice, then they are uniformly integrable. In this case, the Law of Large Numbers still holds without the assumption of equal distributions. Therefore, all subsequent results are valid provided that the data are normalized; this proves Proposition 2.

To show that the "independent" part of the *i.i.d.* hypothesis is not necessary either, we will compare the concentration of norms in real data sets with the same embedding dimensions but different intrinsic dimensions.

Suppose we have some data described in a $d$-dimensional space for which we know that the intrinsic dimension $d_{int}$ is lower. Such a situation occurs, for example, when the variables are significantly correlated. We will see that, for such data, the value of the relative variance is much more similar to the one of a $d_{int}$-dimensional data set with independent variables than of a $d$-dimensional data set with independent variables. For that purpose, we will ensure that marginal distributions, that is, the distribution of each variable taken separately, are identical in both data sets.

Suppose we have $\chi = \{x^{(j)}\}_{j=1}^n$, a sample drawn from $\mathrm{X} = (X_1, \cdots, X_d) \sim F(\xi_1, \xi_2, \ldots, \xi_d)$, where $F$ is the multivariate probability density function of X. The marginal distributions of $X_i$ are

$$\mathcal{F}_i(\xi_i) = \int\int F(\xi_1, \ldots \xi_i, \ldots, \xi_d)\ \mathrm{d}\xi_1 \ldots\ \mathrm{d}\xi_{i-1}\ \mathrm{d}\xi_{i+1} \ldots\ \mathrm{d}\xi_d.$$

To produce a data set $\chi'$ that is marginally identically distributed as $\chi$, we propose the following: If we consider a matrix where each row corresponds to a data element $x^{(j)}$ and each column to a variable $X_i$, the values in each column are randomly permuted. As a consequence, the marginal distributions of each variable will not change. By contrast, all relationships between variables are destroyed in the process. Therefore, we obtain a sample $\chi'$ that is marginally distributed as $\chi$, but where the components are now independent; the intrinsic dimension of $\chi'$ is thus equal to its embedding dimension.

Let us denote by $\widehat{RV_{\chi,2}}$ the estimation of $RV_{\mathcal{F},p}$ with the euclidean norm given the data set $\chi$ with high embedding dimension and low intrinsic dimension. We will compare this value to the value of the relative variance of $\chi'(\widehat{RV_{\chi',2}})$ that has a high intrinsic dimension. Moreover, we will compare those values to the value of the relative variance for a data set made of a small number (typically 20 times lower than the original number of variables) of low-correlated variables from $\chi(\widehat{RV_{\chi s,2}})$ (small embedding dimension). The relative variance of a uniformly distributed synthetic data set of dimensionality $d : (\widehat{RV_{rand,2}})$ (high intrinsic dimension) is also computed. It is expected that the relative variances for data sets with low intrinsic dimension will be similar while being much lower than the relative variances of the data sets with high intrinsic dimension.

The relative variance for a data set $\chi = \{x^{(j)}\}_{j=1}^n$ is estimated as follows:

$$\widehat{RV_{\mathcal{X},2}} = \frac{\sqrt{\sum_{j=1}^n \left( \|x^{(j)}\|_2 - \sum_{j=1}^n \|x^{(j)}\|_2 \right)^2}}{\sum_{j=1}^n \|x^{(j)}\|_2}.$$

The data sets mentioned in Table 1 are high-dimensional data coming from various domains. The first two data sets are the near-infrared spectra of apple and meat,

TABLE 1
Relative Variances for Several Real Data Sets Compared with the Relative Variances of Artificial *i.i.d.* Data Sets

| dataset | $d$ | $\widehat{RV_{\mathcal{X},2}}$ | $\widehat{RV_{\mathcal{X}',2}}$ | $\widehat{RV_{\mathcal{X}s,2}}$ | $\widehat{RV_{rand,2}}$ |
|---|---|---|---|---|---|
| apples spec. | 110 | 0.6061 | 0.0585 | 0.6331 (5) | 0.0708 |
| meat spec. | 100 | 0.7616 | 0.0779 | 0.7741 (5) | 0.0738 |
| yesno | 8192 | 0.3459 | 0.0305 | 0.3656 (400) | 0.0077 |
| boatgoat | 8192 | 0.1832 | 0.0171 | 0.2021(400) | 0.0083 |
| musk | 167 | 0.1740 | 0.0437 | 0.2757 (8) | 0.0561 |
| ionosphere | 34 | 0.2855 | 0.0816 | 0.2170 (5) | 0.1214 |

*The numbers between parentheses are the number of variables used to build $\mathcal{X}_s$.*

respectively,[1] the next two data sets are recorded sounds of "yes" and "no," and of "boat" and "goat" [46].[2] The ionosphere and the musk data sets come from the UCI machine learning repository.[3] The dimensionality goes from 34 to more than 8,000.

As expected, it is seen in Table 1 that the relative variance of the original data set (column 1) is very similar to the relative variance of a subset of its variables (column 3). In contrast, the relative variance for the crafted data set $\chi'$ (column 2) is very similar to the relative variance of a random uniformly distributed data set of the same dimension (column 4). □

# 6 About the Optimal Value of $p$

Throughout the preceding sections, we saw that all fractional distances concentrate in high-dimensional spaces and that concentration depends nontrivially on the value of $p$. This concentration has negative consequences on indexing methods and leads to the questioning of the meaningfulness of the NN search when the distances seem to be all identical. This motivates the use of alternative metrics to the widely used euclidean one.

## 6.1 Fractional Norms and Non-Gaussian Noise

One more argument can motivate the use of fractional norms. It can easily be shown that the concept of euclidean distance and the concept of Gaussian white noise are intimately tied. A Gaussian white noise is a noise that has a normal distribution with zero mean and equal variance for all variables. Looking for the most similar data element to a query point by minimizing the euclidean distance is equivalent to choosing the NN to be the point most likely to be the query point under the Gaussian white noise scheme.

In low-dimensional spaces, most often, a Gaussian white noise is an acceptable assumption. However, when the dimension increases, other noise schemes might be more appropriate. The white Gaussian noise is a model that describes small alterations gently distributed over the coordinates. Obviously, in low-dimensional spaces, this is the only kind of noise we can cope with. Imagine now a noise scheme that models large alterations of only some of the coordinates instead of small alterations of all coordinates. This kind of noise will generate the so-called "outliers" in low-dimensional spaces, but in higher dimensions, where more information is available for each data element, it can just be handled as a noise scheme. Examples of such noise schemes are the so-called Salt and Pepper noise on images and Burst noise on time series; encoding errors may also be viewed as noise that sometimes dramatically affects a small number of the coordinates.

In [47], a "colored" noise scheme is studied, which concentrates its effects on some of the coordinates while leaving the other ones nearly unchanged. The experiments on high-dimensional data show that, for such a noise, fractional norms are better at identifying the "real" NNs, that is, the original point when the noise is removed. For Gaussian white noise, however, the euclidean norm gives better results.

Similarly, in [24], the experiments show that fractional norms are better suited for classification when masking noise is applied. This noise scheme randomly changes some values of the coordinates; it is very different from Gaussian noise.

All these results clearly illustrate the fact that the "optimal" value of $p$ is highly application dependent.

## 6.2 Choosing the Norm for Regression/Classification

In a prediction or classification problem, the value of $p$ could be chosen so as to get the best model performances, according to the expected error in predicting the response value or the class label. It would thus be considered as an additional parameter to the model: The norm that is chosen is the norm that minimizes the differences between the true response values or class labels and the predicted ones. However, this would necessitate multiplying the computation times by as many different values of $p$ are tested.

An elegant alternative is to choose the value prior to the building of the model. In this case, the norm is chosen before any prediction model is constructed. It is chosen according to a statistical measure of relevance for each $p$-norm that is considered. This statistical measure gives each $p$-norm a score based on how well similar data elements according to the $p$-norm relate to similar response values or class labels. If the data elements that are close in the data space, that is, similar according to the norm, are also close in terms of associated response value, then the norm is considered relevant as a measure of similarity for those data. Nonparametric noise estimators such as the Gamma test [48] or the Differogram [49] are suitable for this, as well as the performances of a 1-NN model.

This latter strategy (1-NN model) is illustrated in the following experiments. We consider real data, namely, the Housing[4] data set, and the Wine, Tecator, and Apple[5] data sets. The objective in the Housing data set is to predict the values of houses (in thousand dollars) described by 13 attributes representing the demographic statistics of the area around each house. The data set contains 506 instances split into 338 learning examples and 169 for testing. The

[1] Available at http://www.ucl.ac.be/mlg.

[2] Available at http://nathalie.vialaneix.free.fr/maths/article.php3?id_article=20.

[3] Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

[4] Available at the UCI repository.

[5] All of them are available at http://www.ucl.ac.be/mlg.

TABLE 2
Performances of the RBFN on Several Data Sets with the Euclidean Norm and with Fractional Norms

| Dataset | d/Learning/test | Euclidean norm | Fractional norm |
|---|---|---|---|
| Housing | 13 / 400 / 106 | $p = 2 \rightarrow 9.7371$; | $p = 1/8 \rightarrow 7.0561$ |
| Wine | 256 / 94 / 34 | $p = 2 \rightarrow 0.3883$; | $p = 1 \rightarrow 0.1067$ |
| Tecator | 100 / 172 / 43 | $p = 2 \rightarrow 8.2202$; | $p = 1/8 \rightarrow 6.7685$ |
| Apple | 110 / 250 / 87 | $p = 2 \rightarrow 1.6389$; | $p = 1/2 \rightarrow 1.5191$ |

*The data are altered with burst noise. The norm in the RBFN is chosen according to the leave-one-out performances of a 1-NN.*

other three data sets are spectral data: for the Wine, the alcohol content must be predicted, for the Tecator, it is the fat content, and for the Apple data set, it is the sugar content. A *burst noise* was added to the data: Some coordinates were altered with a multiplicative noise of high amplitude. With such a noise scheme, strongly non-Gaussian fractional norms are better suited to measure the similarity. The leave-one-out error of a 1-NN is used to measure the relevance of each norm and to choose the most relevant one. Then, a Radial Basis Function Network (RBFN) [50] is built with the chosen norm. The parameters of the RBFN are chosen by fourfold cross validation. Table 2 reports the Root Mean Square Error (RMSE) of prediction on an independent test set for all the data sets. For each of these data sets, the fractional norm, chosen prior to the building of the model, gives better results than the euclidean norm.

### 6.3 Choosing the Norm in Content-Based Similarity Search

In many multimedia database systems, the user feedback can be used to estimate the relevance of the result of the search for similar elements. This can be translated into the relevance of the metric used to get similar elements. For instance, in image and text retrieval, the relevance measure is used to weigh the coordinates in the computation of the distance to better reflect the perceived (subjective) similarity between objects [51], [52].

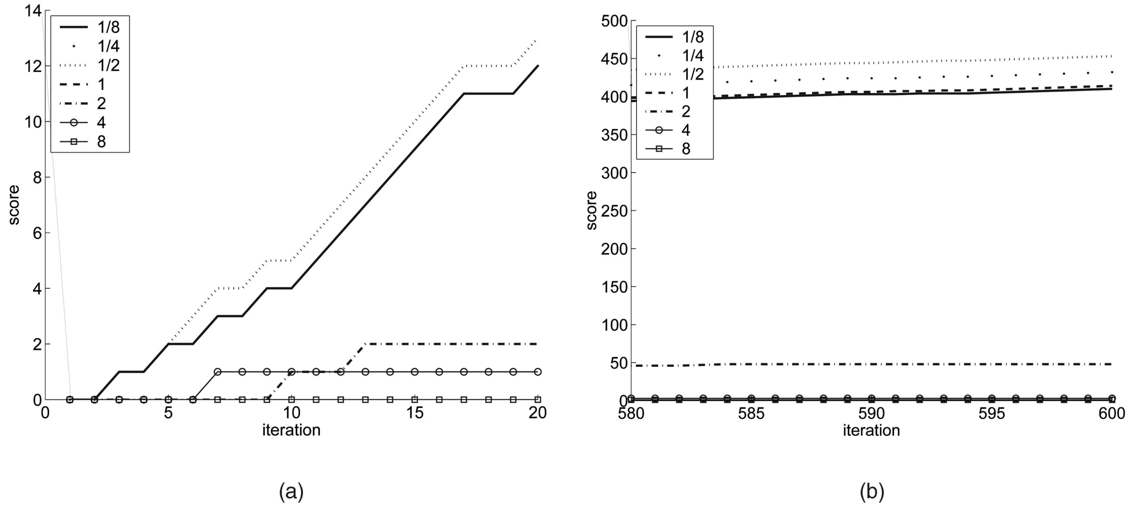
A relevance feedback algorithm is then given as follows: Suppose that all elements are described by a feature vector; given a query $Q$, the system retrieves the nearest or the 2-NNs according to several $p$-norms (with $p = 0.1, 0.5, 1, 2, 4,$ for instance). The user then identifies the most relevant results, and the score of the $p$-norms corresponding to those results is increased. After several iterations, the $p$-norm with the highest score is chosen.

To illustrate this procedure, the XM2VTS database is used. This database is comprised of 600 photographs, altered with Salt and Pepper noise (some pixels are set to black or white randomly), of 200 individuals (that is, three pictures per individual.) At each iteration, a picture from the data set is considered as a query point. Its NN according to several $p$-norms are retrieved, and the score of the norms for which the retrieved image corresponds to the same individual is increased by one. Fig. 5 shows the evolution of the score for each $p$-norm at iterations 1 to 20 and iteration 580 to 600. We can see that the $1/2$-norm is the one with the highest score after 20 iterations and still the one after all 600 iterations. The $1/4$ and $1/8$-norms also have high scores in contrast to higher order norms that perform poorly. The same experiment was repeated 100 times with only 10 iterations. In 86 percent of these experiments, the $1/2$-norm was identified as the most relevant. The $1/8$-norm was chosen as the most relevant three times, the $1/4$-norms only once, the 1-norm 10 times, and the other (higher order) norms were never chosen.

In conclusion, fractional norms should be used when they are a more relevant measure of similarity and, hence, increase the performances, rather than because of concentration considerations.

## 7 Conclusions

A comparison between data elements is often performed using the euclidean norm and distance. In high-dimensional spaces, however, norms concentrate. This means that all pairwise distances in a data set are very similar and can



(a)

(b)

Fig. 5. The evolution of the score of each norm as a function of the number of iterations for the face image database. The 1/2-norm appears as the most relevant from iteration 10 onward.

lead to questioning the relevance of the euclidean distance for measuring similarities in high-dimensional spaces.

This paper considers the problem of the concentration of the distances independently from the number of data and shows that the concentration is indeed an intrinsic property of the distances and not due to finite sample size. This paper furthermore shows that

- all p-norms concentrate, even when an infinite number of data (that is, a distribution) is considered, including when the distribution is unbounded;
- the value of $p$ and the shape of the distribution of high-dimensional data influence the value of the relative variance, which is derived for any distribution and used as a measure of concentration;
- consequently, the exponent $p$ of the norm can be adjusted to fight the concentration phenomenon;
- there exist distributions for which the relative contrast does not increase when the exponent $p$ of the norm decreases;
- the identically distributed hypothesis in the concentration of the norm theorems is not necessary as soon as the variables are normalized;
- the concentration phenomenon is more related to the intrinsic dimension of data than to their embedding dimension, which makes its consequences in practical situations less severe than mathematically expected; and
- the optimal metric is highly application-dependent, and some sort of supervision is needed to optimally choose the metric.

Fractional norms are not always less concentrated than other norms. They seem, however, to be more relevant as a measure of similarity when the noise affecting the data is strongly non-Gaussian.

## ACKNOWLEDGMENTS

The work of Damien Francois is funded by a grant from the Belgian FRIA. Part of his work and of the work of Vincent Wertz is supported by the Interuniversity Attraction Pole (IAP), initiated by the Belgian Federal State, Ministry of Sciences, Technologies and Culture. The scientific responsibility rests with the authors.

## REFERENCES

[1] E. Chavez, G. Navarro, R.A. Baeza-Yates, and J.L. Marroquin, "Searching in Metric Spaces," *ACM Computing Surveys*, vol. 33, no. 3, pp. 273-321, Sept. 2001.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. Academic Press Professional, 1990.

[3] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, Sept. 1999.

[4] P. Demartines, "Analyse de Données par Réseaux de Neurones Auto-Organisés," PhD dissertation, Institut Nat'l Polytechnique de Grenoble, Grenoble, France, 1994 (in French).

[5] K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is "Nearest Neighbor" Meaningful," *Proc. Seventh Int'l Conf. Database Theory*, pp. 217-235, Jan. 1999.

[6] Y. Tao, J. Sun, and D. Papadias, "Analysis of Predictive Spatio-Temporal Queries," *ACM Trans. Database Systems*, vol. 28, no. 4, pp. 295-336, Dec. 2003.

[7] N. Katayama and S. Satoh, "Distinctiveness-Sensitive Nearest-Neighbor Search for Efficient Similarity Retrieval of Multimedia Information," *Proc. 17th Int'l Conf. Data Eng. (ICDE '01)*, pp. 493-502, Apr. 2001.

[8] N. Katayama and S. Satoh, "Similarity Image Retrieval with Significance-Sensitive Nearest-Neighbor Search," *Proc. Sixth Int'l Workshop Multimedia Information Systems (MIS '00)*, pp. 177-186, Oct. 2000.

[9] J. Yang, A. Patro, S. Huang, N. Mehta, M.O. Ward, and E.A. Rundensteiner, "Value and Relation Display for Interactive Exploration of High Dimensional Datasets," *Proc. IEEE Symp. Information Visualization (InfoVis '04)*, W.A. Burks, ed., pp. 73-80, Oct. 2004.

[10] E. Tuncel, H. Ferhatosmanoglu, and K. Rose, "Vq-Index: An Index Structure for Similarity Searching in Multimedia Databases," *Proc. 10th ACM Int'l Conf. Multimedia*, pp. 543-552, Dec. 2002.

[11] N. Mamoulis, D.W. Cheung, and W. Lian, "Similarity Search in Sets and Categorical Data Using the Signature Tree," *Proc. 19th Int'l Conf. Data Eng. (ICDE '03)*, pp. 75-86, Mar. 2003.

[12] P. Ciaccia and M. Patella, "Pac Nearest Neighbor Queries: Approximate and Controlled Search in High-Dimensional and Metric Spaces," *Proc. 16th Int'l Conf. Data Eng. (ICDE '02)*, W.A. Burks, ed., pp. 244-255, Mar. 2000.

[13] M. Demirbas and H. Ferhatosmanoglu, "Peer-to-Peer Spatial Queries in Sensor Networks," *Proc. Third IEEE Int'l Conf. Peer-to-Peer Computing (P2P '03)*, pp. 32-39, Sept. 2003.

[14] D.S. Johnson, S. Krishnan, J. Chhugania, S. Kumar, and S. Venkatasubramanian, "Compressing Large Boolean Matrices Using Reordering Techniques," *Proc. 30th Int'l Conf. Very Large Data Bases (VLDB '04)*, pp. 13-23, Sept. 2004.

[15] K. Chakrabarti and S. Mehrotra, "The Hybrid Tree: An Index Structure for High Dimensional Feature Spaces," *Proc. 15th Int'l Conf. Data Eng. (ICDE '99)*, pp. 440-447, Feb. 1999.

[16] K. Chakrabarti and S. Mehrotra, "Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces," *Proc. 26th Int'l Conf. Very Large Data Bases (VLDB '00)*, pp. 89-100, Sept. 2000.

[17] T. Liu, A. Moore, A. Gray, and K. Yang, "An Investigation of Practical Approximate Nearest Neighbor Algorithms," *Proc. 18th Ann. Conf. Neural Information Processing Systems (NIPS '04)*, pp. 825-832, Dec. 2004.

[18] K.P. Bennett and U. Fayyad Dan Geiger, "Density-Based Indexing for Approximate Nearest-Neighbor Queries," *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 233-243, Aug. 1999.

[19] S. Berchtold, C. Böhm, H.V. Jagadish, H.-P. Kriegel, and J. Sander, "Independent Quantization: An Index Compression Technique for High-Dimensional Data Spaces," *Proc. 16th Int'l Conf. Data Eng. (ICDE '00)*, pp. 577-588, Mar. 2000.

[20] K.Y. Yip, D.W. Cheung, and M.K. Ng, "A Highly-Usable Projected Clustering Algorithm for Gene Expression Profiles," *Proc. Workshop Data Mining in Bioinformatics (BIOKDD '03)*, pp. 41-48, Aug. 2003.

[21] M. Vlachos, D. Gunopulos, and G. Kollios, "Robust Similarity Measures for Mobile Object Trajectories," *Proc. 13th Int'l Workshop Database and Expert Systems Applications (DEXA '02)*, pp. 721-728, Sept. 2002.

[22] F. Korn, B.-U. Pagel, and C. Faloutsos, "On the "Dimensionality Curse" and the "Self-Similarity Blessing"," *IEEE Trans. Knowledge and Data Eng.*, vol. 13, no. 1, pp. 96-111, Jan./Feb. 2001.

[23] A. Hinneburg, C.C. Aggarwal, and D.A. Keim, "What Is the Nearest Neighbor in High Dimensional Spaces," *Proc. 26th Int'l Conf. Very Large Data Bases (VLDB '00)*, A. El Abbadi, M.L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, and K.-Y. Whang, eds., pp. 506-515, Sept. 2000.

[24] C.C. Aggarwal, A. Hinneburg, and D.A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Spaces," *Proc. Eighth Int'l Conf. Database Theory*, J. Van den Bussche and V. Vianu, eds., pp. 420-434, Jan. 2001

[25] C.C. Aggarwal, "Re-Designing Distance Functions and Distance-Based Applications for High Dimensional Data," *Proc. ACM Int'l Conf. Management of Data (SIGMOD '01)*, vol. 30, no. 1, pp. 13-18, Mar. 2001.

[26] K. Doherty, R. Adams, and N. Davey, "Non-Euclidean Norms and Data Normalisation," *Proc. 12th European Symp. Artificial Neural Networks*, M. Verleysen, ed., Apr. 2004.

[27] P. Howarth and S.M. Rüger, "Fractional Distance Measures for Content-Based Image Retrieval," *Proc. 27th European Conf. Information Retrieval Research (ECIR '05),* J.M. Fernandez-Luna and D.E. Losada, eds., pp. 447-456, Mar. 2005

[28] H. Jin, B.C. Ooi, H. Shen, C. Yu, and A. Zhou, "An Adaptive and Efficient Dimensionality Reduction Algorithm for High-Dimensional Indexing," *Proc. 19th Int'l Conf. Data Eng. (ICDE '03),* pp. 87-98, Mar. 2003.

[29] C. Elkan, "Using the Triangle Inequality to Accelerate K-Means," *Proc. 20th Int'l Conf. Machine Learning (ICML '03),* pp. 147-153, Aug. 2003.

[30] M. Skala, "Measuring the Difficulty of Distance-Based Indexing," *Proc. 12th Int'l Conf. String Processing and Information Retrieval (SPIRE '05),* M.P. Consens and G. Navarro, eds., pp. 103-114, Nov. 2005.

[31] P. Yianilos, "Excluded Middle Vantage Point Forests for Nearest Neighbor Search," technical report, NEC Research Inst., Jan. 1999, presented at *Proc. Sixth Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) Implementation Challenge: Near Neighbor Searches Workshop.*

[32] S. Berchtold, C. Boehm, B. Braunmueller, D.A. Keim, and H.-P. Kriegel, "Fast Similarity Search in Multimedia Databases," *Proc. ACM Int'l Conf. Management of Data (SIGMOD '97),* J. Peckham, ed., May 1997.

[33] N. Katayama and S. Satoh, "The Sr-Tree: An Index Structure for High-Dimensional Nearest Neighbor Queries," *Proc. ACM Int'l Conf. Management of Data (SIGMOD '97),* J. Peckham, ed., pp. 369-380, May 1997.

[34] S. Berchtold, D.A. Keim, and H.-P. Kriegel, "The X-Tree: An Index Structure for High-Dimensional Data," *Proc. 22nd Int'l Conf. Very Large Data Bases (VLDB '96),* T.M. Vijayaraman, A.P. Buchmann, C. Mohan, and N.L. Sarda, eds., pp. 28-39, Sept. 1996.

[35] K.-I. Lin, H.V. Jagadish, and C. Faloutsos, "The TV-Tree: An Index Structure for High-Dimensional Data," *VLDB J.,* vol. 3, no. 4, pp. 517-542, 1994.

[36] P. Indyk and R. Motwani, "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality," *Proc. 30th Ann. ACM Symp. Theory of Computing,* pp. 604-613, May 1998.

[37] C. Böhm, S. Berchtold, and D.A. Keim, "Searching in High-Dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases," *ACM Computing Surveys,* vol. 33, no. 3, pp. 322-373, Sept. 2001.

[38] S. Brin, "Near Neighbor Search in Large Metric Spaces," *Proc. 21st Int'l Conf. Very Large Data Bases (VLDB '95),* U. Dayal, P.M.D. Gray, and S. Nishio, eds., pp. 574-584, Sept. 1995.

[39] S. Berchtold, C. Bohm, D.A. Keim, and H.-P. Kriegel, "A Cost Model for Nearest Neighbor Search in High-Dimensional Data Space," *Proc. 16th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems (PODS '97),* pp. 78-86, May 1997.

[40] R. Weber, H.-J. Schek, and S. Blott, "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces," *Proc. 24th Int'l Conf. Very Large Data Bases (VLDB '98),* A. Gupta, O. Shmueli, and J. Widom, eds., pp. 194-205, Aug. 1998.

[41] R. Bellmann, *Adaptive Control Processes: A Guided Tour.* Princeton Univ. Press, 1961.

[42] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science,* vol. 290, no. 5500, pp. 2319-2323, Dec. 2000.

[43] J.A. Lee, A. Lendasse, and M. Verleysen, "Nonlinear Projection with Curvilinear Distances: Isomap versus Curvilinear Distance Analysis," *Neurocomputing,* vol. 57, pp. 49-76, 2003.

[44] C.L Blake, D.J. Newman, S. Hettich, and C.J. Merz, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[45] D. Landers and L. Rogge, "Laws of Large Numbers for Pairwise Independent Uniformly Integrable Random Variables," *Math. Nachrichten,* vol. 130, pp. 189-192, 1987.

[46] G. Biau, F. Bunea, and M.H. Wegkamp, "Functional Classification in Hilbert Spaces," *IEEE Trans. Information Theory,* vol. 51, no. 6, pp. 2163-2172, 2005.

[47] D. Francois, V. Wertz, and M. Verleysen, "Non-Euclidean Metrics for Similarity Search in Noisy Datasets," *Proc. European Symp. Artificial Neural Networks (ESANN '05),* pp. 339-344, 2005.

[48] A.J. Jone, A. Stefansson, and N. Koncar, "A Note on the Gamma Test," *Neural Computing and Applications,* vol. 5, no. 3, pp. 131-133, Sept. 1997.

[49] K. Pelckmans, J. De Brabanter, J.A.K. Suykens, and B. De Moor, "The Differogram: Nonparametric Noise Variance Estimation and Its Use for Model Selection," *Neurocomputing,* vol. 69, no. 1, pp. 100-122, Dec. 2005.

[50] M. Powell, "Radial Basis Functions for Multivariable Interpolation: A Review," *Algorithms for Approximation,* M.G. Cox and J.C. Mason, eds., pp. 143-167, Clarendon Press, 1987

[51] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval," *IEEE Trans. Circuits and Systems for Video Technology,* special issue on segmentation, description, and retrieval of video content, vol. 8, no. 5, pp. 644-655, Sept. 1998.

[52] G. Salton and C. Buckley, "Mproving Retrieval Performance by Relevance Feedback," *J. Am. Soc. for Information Science,* vol. 41, pp. 288-297, June 1990.

**Damien François** received the MSc degree in computer science and the PhD degree in applied mathematics from the Université catholique de Louvain, Belgium, in 2002 and 2007, respectively. He is now a research assistant at the Center for System Engineering and Applied Mechanics (CESAME) and a member of the Machine Learning Group at the Université catholique de Louvain. His main interests include high-dimensional data analysis, distance-based prediction models, and feature/model selection.

**Vincent Wertz** received the engineering degree in applied mathematics in 1978 and the PhD degree in control engineering in 1982 from the Université catholique de Louvain, Louvain-la-Neuve, Belgium. He is now a professor at the Université catholique de Louvain after having held a permanent position with the Belgian National Fund for the Scientific Research (NFSR). His main research interests are in the fields of identification, predictive control, fuzzy control, and industrial applications. Lately, he has also been involved in a major pedagogical reform of the first and second year teaching program in the School of Engineering. He is a member of the IEEE.

**Michel Verleysen** received the MS and PhD degrees in electrical engineering from the Université catholique de Louvain, Belgium, in 1987 and 1992, respectively. He was an invited professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 1992, at the Université d'Evry Val d'Essonne, France, in 2001, and at the Université Paris I-Panthéon-Sorbonne in 2002, 2003, and 2004, respectively. He is now a research director at the Belgian Funds National de la Recherche Scientique (FNRS) and lecturer at the Université catholique de Louvain. He is the editor-in-chief of *Neural Processing Letters*, the chairman of the Annual European Symposium on Artificial Neural Networks (ESANN), an associate editor of the *IEEE Transactions on Neural Networks*, and a member of the editorial board and program committee of several journals and conferences on neural networks and learning. He is an author or coauthor of about 200 scientific papers in international journals and books or communications to conferences with reviewing committees. He is the coauthor of the scientific popularization book on artificial neural networks in the series *Que Sais-Je?* in French. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis. He is a senior member of the IEEE.