Article

A Novel Nonparametric Distance Estimator for Densities with Error Bounds

Alexandre R.F. Carvalho [1,*], João Manuel R. S. Tavares [1] and Jose C. Principe [2]

[1] Instituto de Engenharia Mecânica e Gestão Industrial, Faculdade de Engenharia, Universidade do Porto; Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal; E-Mail: tavares@fe.up.pt

[2] Computational Neuro Engineering Laboratory, University of Florida; EB451 Engineering Building, University of Florida, Gainesville, FL 32611, USA; E-Mail: principe@cnel.ufl.edu

* Author to whom correspondence should be addressed; E-Mail: afcarval@gmail.com; Tel.: +351-225-081-719; Fax: +351- 225-081-440.

Abstract: The use of a metric to assess distance between probability densities is an important practical problem. In this work, a particular metric induced by an α-divergence is studied. The Hellinger metric can be interpreted as a particular case within the framework of generalized Tsallis divergences and entropies. The nonparametric Parzen's density estimator emerges as a natural candidate to estimate the underlying probability density function, since it may account for data from different groups, or experiments with distinct instrumental precisions, *i.e.*, non-independent and identically distributed (non-i.i.d.) data. However, the information theoretic derived metric of the nonparametric Parzen's density estimator displays infinite variance, limiting the direct use of resampling estimators. Based on measure theory, we present a change of measure to build a finite variance density allowing the use of resampling estimators. In order to counteract the poor scaling with dimension, we propose a new nonparametric two-stage robust resampling estimator of Hellinger's metric error bounds for heterocedastic data. The approach presents very promising results allowing the use of different covariances for different clusters with impact on the distance evaluation.

1. Introduction

Distances measures between two probability densities have been extensively studied in the last century [1]. These measures address two important main objectives: how difficult it is to distinguish between one pair of densities in the context of others and to assess the closeness of two densities, compared to others [2]. In learning scenarios essentially associated with the test of a single hypothesis, the use of a divergence to represent the notion of distance is efficient. However, in scenarios involving multiple hypotheses, such as clustering, image retrieval, or pattern recognition and signal detection, for instance, the non-symmetric and non-metric nature of divergences becomes problematic [3]. When deciding the closest or the farthest among three or more clusters, the use of a metric is important. In this work, a novel nonparametric metric estimator for densities with error bounds is presented. Shannon's entropy has a central role in information-theoretic studies. However, the concept of information is so rich that perhaps there is no single definition that will be able to quantify information properly [4]. The idea of using information theory functional, such as entropies or divergences, in statistical inference is not new. In fact, the so-called statistical information theory has been the subject of much research over the last half century [5]. How to measure the distance between two densities is an open problem with several proposals since the work of Hellinger in 1909 with Hellinger's distance [1], Kullback and Leibler (1951), with Kullback-Leibler's divergence [6], Bregman (1967) with the Bregman's divergence [7], Jeffreys (1974) with J-distance [8], RAO (1985) and Jianhua Lin (1991) with Jensen-Shannon's divergence [9,10], Menéndez *et al.* (1997) with (h, Φ)-entropy differential metric [11], Seth and Principe (2008) with correntropy [12], among others. This work looks into Hellinger's metric that is the preferred [13,14] or natural model metric [15]. In 2007 Puga found that Hellinger's metric is one particular α-divergence [16]. Here, we propose a new measure change to solve the nonparametric metric estimation and a two stage robust estimator with error bounds.

2. Theory Background

Following Hartley's (1928) and Shannon's (1948) works [17,18], Alfred Rényi introduced in 1960 the generalized α-entropy [19] of probability density function $f(x)$:

$$R_\alpha(f) = \frac{1}{1-\alpha} \ln \int f(x)^\alpha \, dx, \, \alpha > 0 \tag{1}$$

The corresponding generalized differential divergence between two densities $f_1(x)$ and $f_2(x)$ is:

$$D_\alpha^R(f_1, f_2) = \frac{1}{\alpha - 1} \ln \int \frac{f_1(x)^\alpha}{f_2(x)^{\alpha-1}} \, dx \tag{2}$$

Gell-Mann and Tsallis considered another family of α-entropies [20]:

$$T_\alpha(f) = \frac{1}{\alpha - 1} \left[1 - \int f(x)^\alpha \, dx \right] \tag{3}$$

being the corresponding α-divergences given as:

$$D_\alpha^T(f_1, f_2) = \frac{1}{1-\alpha} \left[1 - \int \frac{f_1(x)^\alpha}{f_2(x)^{\alpha-1}} \, dx \right] \tag{4}$$

Making $\alpha \to 1$, one easily can conclude that:

$$\lim_{\alpha \to 1} R_\alpha(f) = \lim_{\alpha \to 1} T_\alpha(f) = H_S(f) \tag{5}$$

and:

$$\lim_{\alpha \to 1} D_\alpha^R(f_1, f_2) = \lim_{\alpha \to 1} D_\alpha^T(f_1, f_2) = D_{KL}(f_1, f_2) \tag{6}$$

where:

$$H_S(f) = -\int f(x) \ln f(x) dx \tag{7}$$

is Shannon's differential entropy and:

$$D_{KL}(f_1, f_2) = \int f_1(x) \ln \frac{f_1(x)}{f_2(x)} dx \tag{8}$$

is Kulback-Leibler's divergence.

Another member of these families is the Rényi's quadratic entropy (α=2) that is defined as:

$$R_2(f) = -\ln \int f(x)^2 dx \tag{9}$$

while the respective divergence is:

$$D_2^R(f_1, f_2) = \ln \int \frac{f_1(x)^2}{f_2(x)} dx \tag{10}$$

Rényi's quadratic entropy, given by Equation (9), is particularly interesting because it accepts a close form nonparametric estimator, saving computational time compared to numerical integration or resampling [21,22].

α-Entropy families given by Equations (1) and (3) are monotonically coupled (Ramshaw [23]) through:

$$T_\alpha = \left(e^{(1-\alpha)R_\alpha} - 1\right) \big/ (1-\alpha) \tag{11}$$

Therefore, an optimization in one family has equivalence in the other.

2.1. Square-Root Entropy

Let us consider $\alpha = 1/2$ in Equations (1)–(4). Then, the square-root entropy in the form of Tsallis is:

$$T_{1/2}(f) = 2\int \sqrt{f(x)} dx - 2 \tag{12}$$

with the corresponding divergence given as:

$$D_{1/2}^T(f_1, f_2) = 2 - 2\int \sqrt{f_1(x) f_2(x)} dx \tag{13}$$

In Rényi's form one finds, respectively, the entropy:

$$R_{1/2}(f) = 2\ln \int \sqrt{f(x)} dx, \tag{14}$$

and the divergence:

$$D_{1/2}^R(f_1, f_2) = -2\ln \int \sqrt{f_1(x) f_2(x)} dx \tag{15}$$

It should be noted that, from Equation (13), one obtains:

$$\sqrt{D_{1/2}^{T}(f_1, f_2)} = \sqrt{\int \left(\sqrt{f_1(x)} - \sqrt{f_2(x)}\right)^2 dx} = M(f_1, f_2) \tag{16}$$

where $M(f_1, f_2)$ is a information theoretic derived metric that, among other properties, verifies the triangular inequality. This particular α-divergence, by means of a monotonous transformation, induces the Hellinger's distance, which is a metric [13,14,24]:

$$M(f_1, f_2) = \sqrt{2 - 2I(f_1, f_2)} \tag{17}$$

On the other hand, information theoretic derived metrics given by Equations (15) and (16) are also related with Hellinger's affinity or Bhattacharya's coefficient ($0 \le I(f_1, f_2) \le 1$):

$$I(f_1, f_2) = \int \sqrt{f_1(x) f_2(x)} dx. \tag{18}$$

Considering the expected cross-value of two probability density functions $C(f_1, f_2)$:

$$C(f_1, f_2) = E_{f_1}(f_2) = E_{f_2}(f_1) = \int f_1(x) f_2(x) dx \tag{19}$$

the Hellinger's affinity given by Equation (18) can be then written as:

$$I(f_1, f_2) = \sqrt{C(f_1, f_2)} \int \sqrt{f_\Omega(x)} dx = \sqrt{C(f_1, f_2)} H(f_\Omega) \tag{20}$$

where $f_\Omega(x)$ is the normalized product density:

$$f_\Omega(x) = \frac{f_1(x) f_2(x)}{C(f_1, f_2)} \tag{21}$$

and $H(f_\Omega)$ the corresponding entropy of the information theoretic derived metric.

This metric has bounds that can be directly computed from the samples as shown by Puga [16]. These bounds often present overlapping hypothesis intervals, and resampling estimation is a necessary tool to remove ambiguities and access distances between densities.

2.2. Nonparametric Hellinger's Affinity Estimation

Let us focus on the application of the previous measures on two Parzen's nonparametric densities [25] from two data clusters $Cl^{(1)} = \{x_1^{(1)}, x_2^{(1)}, \dots, x_{N_1}^{(1)}\}$ and $Cl^{(2)} = \{x_1^{(2)}, x_2^{(2)}, \dots, x_{N_2}^{(2)}\}$:

$$f_1(x) = \frac{1}{N_1} \sum_{j=1}^{N_1} G\left(x, \sigma_1, x_j^{(1)}\right) \tag{22}$$

and:

$$f_2(x) = \frac{1}{N_2} \sum_{j=1}^{N_2} G\left(x, \sigma_2, x_j^{(2)}\right) \tag{23}$$

where $G(x, \sigma, \mu)$ is the Parzen's Gaussian kernel, also known as kernel bandwidth, with the approximation of covariance $\sigma^2 I$, and mean μ given as:

$$G(x, \sigma, \mu) = \prod_{i=1}^{\aleph} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x(i) - \mu(i))^2}{2\sigma^2}} \tag{24}$$

where $\aleph$ is the dimension. Notice that the two clusters in Equations (22) and (23) may have two different Gaussian kernel covariances. The kernel covariance may be obtained directly from the *a priori* knowledge of the instruments used to produce the data; for instance, two different instruments with different precisions may produce the same data, but the densities should reflect the measurements error through the bandwidth, (covariances). To estimate the bandwidth without instrumental apriori knowledge it is possible to estimate the kernel bandwidth with a suitable method, such as k-Nearest Neighbor (k-NN), Silverman [25] or Scott [26].

Now, let us adopt the summing convention $\sum\limits_{i,j} \equiv \sum\limits_{i=1}^{N_1} \sum\limits_{j=1}^{N_2}$ and define the following auxiliary variables:

$$\sigma^2 = \left(\sigma_1^2 + \sigma_2^2\right)/2 \tag{25}$$

$$\sigma_*^2 = \sigma_1^2 \sigma_2^2 / 2\sigma^2 \tag{26}$$

$$s_{i,j} = \left(\sigma_2^2 x_i^{(1)} + \sigma_1^2 x_j^{(2)}\right)/2\sigma^2 \tag{27}$$

$$d_{i,j} = \left(x_i^{(1)} - x_j^{(2)}\right)/2 \tag{28}$$

and:

$$F_D\left(d_{i,j}\right) = e^{-\frac{\|d_{i,j}\|^2}{\sigma^2}} \Big/ \sum_{k,l} e^{-\frac{\|d_{k,l}\|^2}{\sigma^2}} \tag{29}$$

The nonparametric estimator $\hat{f}_\Omega(\omega)$ results in:

$$\hat{f}_\Omega(\omega) = \sum_{i,j} F_D\left(d_{i,j}\right) G\left(\omega, \sigma_*^2, s_{i,j}\right) \tag{30}$$

2.3. The Resampling Estimator

The bootstrap resampling is reached through the distribution of probability given by Equation (30) combined with the random generation of samples (ω_k) from nonparametric Parzen's density with diagonal covariance, which is a well-established as well as a computationally efficient procedure [27]. Then, the synthesized samples are directly usable in the estimator:

$$H\left(f_\Omega\right) = \int \sqrt{f_\Omega(\omega)}\,d\omega = \int \frac{f_\Omega(\omega)}{\sqrt{f_\Omega(\omega)}}\,d\omega$$

$$= E_{f_\Omega}\left[\frac{1}{\sqrt{f_\Omega(\omega)}}\right] = \lim_{K \to \infty} \tilde{H}_K\left(f_\Omega\right) = \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\sqrt{\hat{f}_\Omega(\omega_k)}} \tag{31}$$

with $\omega_k \underset{i.i.d.}{\sim} f_\Omega$.

However, the use of Equation (31) is associated with serious practical difficulties because the second moment:

$$\int d\omega - H\left(f_\Omega\right)^2 = \infty \tag{32}$$

has infinite variance, which is a condition where the central limit theorem is not valid. In this work, we use measure theory and propose the following change of measure:

$$z = f_\Omega(\omega) \tag{33}$$

with the associated density $f_Z(z)$:

$$E_{f_\Omega}\left[\frac{1}{\sqrt{f_\Omega(\omega)}}\right] = E_{f_z}\left[\frac{1}{\sqrt{z}}\right] = \lim_{K \to \infty} \sum_0^{z_k^{max}} \frac{1}{\sqrt{z_k}} \hat{f}_z(z_k) dz \text{ with } z_k \underset{i.i.d.}{\sim} f_Z \tag{34}$$

This new density presents a finite second moment:

$$\int_0^{z^{max}} \frac{1}{z} f_z(z) dz - H(f_\Omega)^2 \tag{35}$$

having f_z a limited support between 0 (zero) and z^{max}. This is a density with an abrupt jump in z^{max} end of the density. However, the approximation properties of a histogram are not affected by a simple jump at the end of the density [26], hence the histogram estimator was used to estimate $\hat{f}_Z(z_k)$ with $z_k = \hat{f}_\Omega(\omega_k)$.

The product probability density function (f_z) must be estimated from the random variable $z = f_\Omega(\omega)$, but it ensures finite variance, which is a requisite of the central limit theorem and the *t*-student confidence interval may be used Equation (39).

> **Figure 1.** A logarithmic scale for the σ_1/σ_2 coefficient variation and metric measure change between the two respective densities (f_1, f_2).
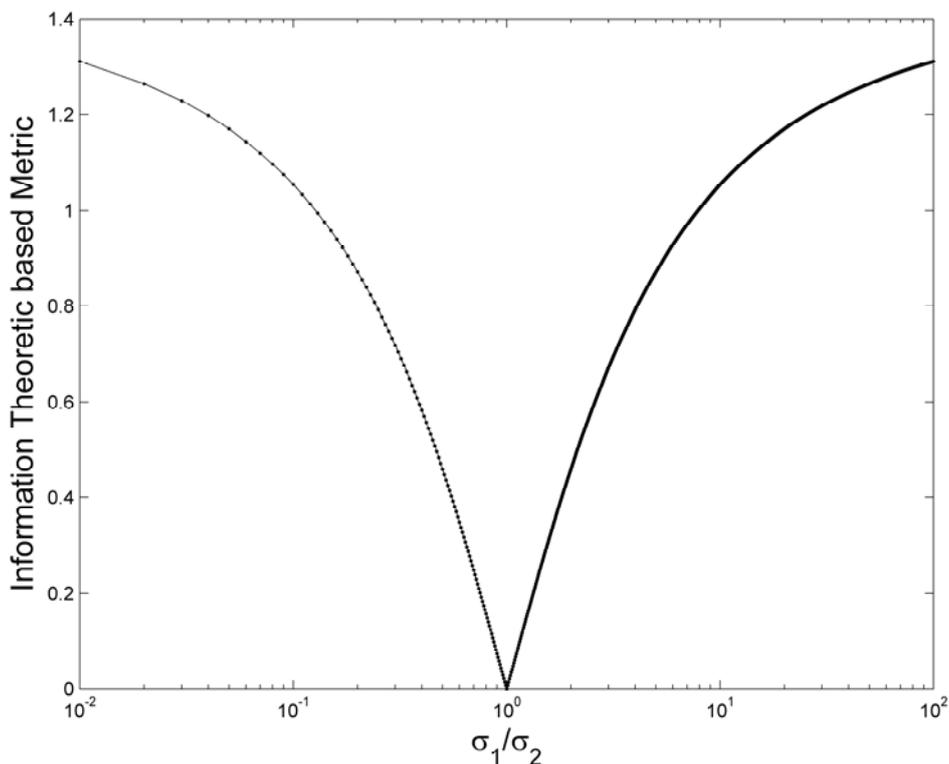
To test the algorithm, we consider the simplest case of Hellinger's metric (17) associated with the nonparametric densities of Equations (22) and (23). In this particular case, we have access to the analytical value of Hellinger's metric:

$$\tilde{M}(f_1, f_2) = \sqrt{2 - 2\left(\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}\right)^{\frac{N}{2}} e^{\frac{\|d_{1,2}\|^2}{\sigma^2}}} \tag{36}$$

Using Equations (34) and (36), we can quantify the computational behavior of the resampling estimator. Let us first consider the behavior of the Parzen's density estimator with two distinct kernel sizes: σ_1, σ_2. In the simplest case with only two kernels, located at the same coordinates, despite the same location, different Parzen's windows in Equation (36) provide different distances, as can be observed in Figure 1. It is possible to verify the symmetric behavior of the distance estimator and realize that the bandwidth of the Parzen's kernel is important to access the distance between clusters. This is a relevant characteristic, especially when the experimental data have different instrumental origins with different measurement precisions; the use of different bandwidths in the Parzen's kernels may reflect this importance feature of the density, and this implies that the data is heterocedastic.

To quantify the error bounds estimation performance, we propose the generation of N_1 distance samples $\tilde{M}_m$ from resampling the density of Equation (34), and to estimate $\tilde{f}_Z$ we use a discrete histogram with N_1 bins, obtaining the ordered z_k and $\tilde{f}_{Z_k}$. As such, the metric $\tilde{M}$ estimator becomes:

$$\tilde{M}_m = \left\{ \sqrt{2 - 2\sqrt{C(f_1, f_2)} \frac{1}{\sqrt{z_k}} \tilde{f}_Z(z_k)\Delta z} \right\}_{m=1\ldots N_1} \tag{37}$$

which can be written as:

$$\tilde{M} = \sqrt{2 - 2\sqrt{C(f_1, f_2)} \sum_0^{z_k^{\max}} \frac{1}{\sqrt{z_k}} \tilde{f}_Z(z_k)\Delta z} \tag{38}$$

To assess the error bounds estimation, we use the t-student 95% confidence interval (39), which is a maximum entropy distribution [28,29] and provides a parametric approach to robust statistics [30], and allows the following calculation of the confidence limits:

$$[L, U] = \tilde{M} \pm t_{N_1 - 1, 0.5 + 0.95/2} \sqrt{\frac{1}{N_1(N_1 - 1)} \sum_{k=1}^{N_1} \left(\tilde{M}_m - \tilde{M}\right)^2} \tag{39}$$

We calculate the 95% confidence limits, the upper (U) and the lower (L) for the respective density resampling. The variance of this new estimator is well controlled in one dimension. The unexpected drawback of this estimator is its poor scaling performance with increased dimension, as depicted in Figure 2.

The new variable $z = f_\Omega(\omega)$ may be seen as a projection of the multidimensional Parzen's kernels into a 1-Dimensional function. This insight allowed the design of a two-stage estimator for $f_\Omega(\omega)$ that circumvents both problems: infinite variance and poor scalability with dimensionality.

Figure 2. Illustration of the metric estimator behavior for dimensions 1 (one) to 5.
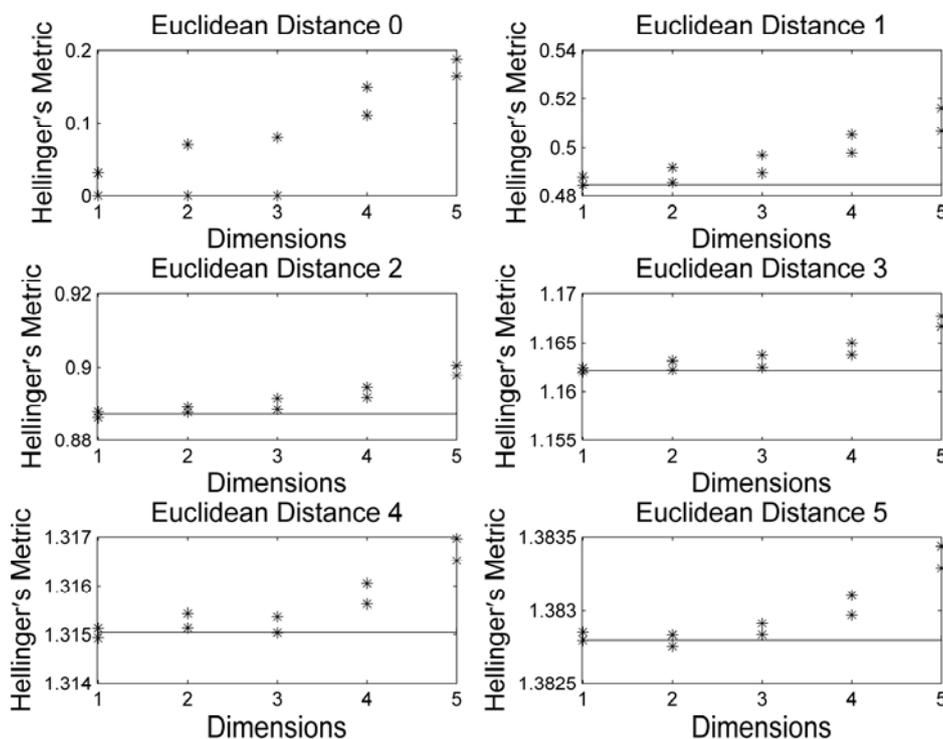
2.4. The Two Stage Resampling Estimator

We propose the generation of N_1 distance samples $\tilde{M}_k^{(n)}$ from resampling the density $f_\Omega(\omega_k)$, which constitutes one trial (n):

$$\tilde{M}_k^{(n)} = \left\{ \sqrt{2 - 2\frac{\sqrt{C(f_1, f_2)}}{\sqrt{f_\Omega(\omega_k^{(n)})}}} \right\}_{k=1\ldots N_1} \tag{40}$$

It is possible to estimate $\tilde{M}^{(n)}$ as:

$$\tilde{M}^{(n)} = \sqrt{2 - \frac{2}{N_1}\sum_{k=1}^{N_1}\sqrt{\frac{C(f_1, f_2)}{f_\Omega(\omega_k^{(n)})}}} \tag{41}$$

For each trial (n), the 95% confidence limits, the upper $U^{(n)}$ and the lower $L^{(n)}$ for the respective density resampling, can be calculated:

$$\left[L^{(n)}, U^{(n)}\right] = \tilde{M}^{(n)} \pm t_{N_1-1,0.5+0.95/2}\sqrt{\frac{1}{N_1(N_1-1)}\sum_{k=1}^{N_1}\left(\tilde{M}_k^{(n)} - \tilde{M}^{(n)}\right)^2} \tag{42}$$

It may seem that this step is enough to estimate the metric $\tilde{M}(f_1, f_2) = \sqrt{2 - 2\tilde{I}(f_1, f_2)}$, but the theoretically predicted undesired behavior associated to Equation (32), with large confidence intervals is present in this estimator. To demonstrate this drawback, we have simulated 100 trials of the simplest case of nonparametric Hellinger's metric, Equation (36), with Euclidean distance $d = \|x_1 - x_2\| = 1$. As can be observed in Figure 3, the large confidence intervals are present, hence the motivation for the two-stage error bound estimator.

Figure 3. t-Student 95% confidence intervals for Hellinger's metric defined by dots; the exact value is represented by a continuous line; the predicted large intervals are marked with triangles; and the miss-estimated intervals are marked with circles.
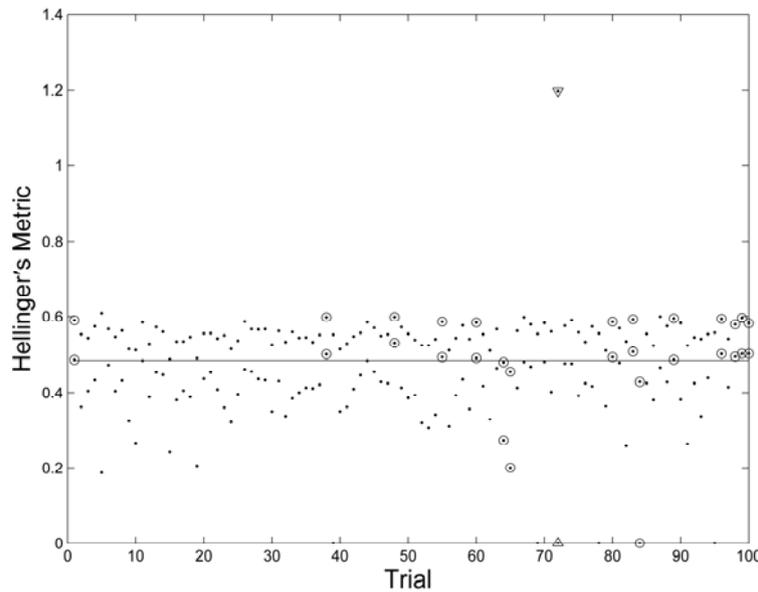
To achieve a robust error bound estimator $\left[\widetilde{L}_R, \widetilde{U}_R\right]$ for the expected value of $M(f_1, f_2)$ with similar results of $f_Z(z)$ in one dimension, we propose a new two-stage method. Comparing the results of the two densities resampling, we found that 31 selected trials out of 33 from $\hat{f}_\Omega(\omega_k)$ was in good agreement with $\widetilde{f}_Z(z)$. With 33 trails (n) generated with N_1 random samples each as:

$$\left\{\left\{\left\{\sqrt{2 - 2\frac{\sqrt{C(f_1, f_2)}}{\sqrt{\hat{f}_\Omega\left(\omega_k^{(n)}\right)}}}\right\}_{k=1\ldots N_1}\right\}_{n=1\ldots 33}\right\} \tag{43}$$

sorting the amplitude $\left|U^{(n)} - L^{(n)}\right|$ and keeping the 31 smallest intervals with the correspondent estimated affinities ($\widetilde{M}_s^{(n)}$), we obtain the estimator $\widetilde{M}_s(f_1, f_2)$ for the second stage with:

$$\widetilde{M}_s(f_1, f_2) = \frac{1}{31}\sum_{n=1}^{31}\widetilde{M}_s^{(n)} \tag{44}$$

Then, we calculated the respective t-student 95% confidence interval $\left[\widetilde{L}_s, \widetilde{U}_s\right]$ with the selected trials $\hat{M}_s^{(n)}$. To overcome the miss-estimated intervals, we have defined a second estimator for the lower limit of the interval ($\widetilde{L}_2$) and a second estimator for the upper limit of the interval ($\widetilde{U}_2$):

$$\left[\tilde{L}_2, \tilde{U}_2\right] = \left[\frac{1}{31}\sum_{n=1}^{31}L_s^{(n)}, \frac{1}{31}\sum_{n=1}^{31}U_s^{(n)}\right] \tag{45}$$

which is a potentially asymmetric interval, guided by the selected first-stage interval limits.

The robust estimator for the lower limit of the interval ($\widetilde{L}_R$) and the robust estimator for the upper limit of the interval ($\widetilde{U}_R$) were defined as:

$$\left[\tilde{L}_R, \tilde{U}_R\right] = \left[\min\left(\tilde{L}_s, \tilde{L}_2\right), \max\left(\tilde{U}_s, \tilde{U}_2\right)\right] \tag{46}$$

In Figure 4, we can find the intervals defined by Equation (46), and confirmed the robust interval estimator for the Hellinger's affinity.

Figure 4. Using the new robust two-stage resampling interval estimator the exact Hellinger's distance is more likely to be found within the interval; it should be noted that from dimension 1 to dimension 20, the exact value of the metric is always in the interval.
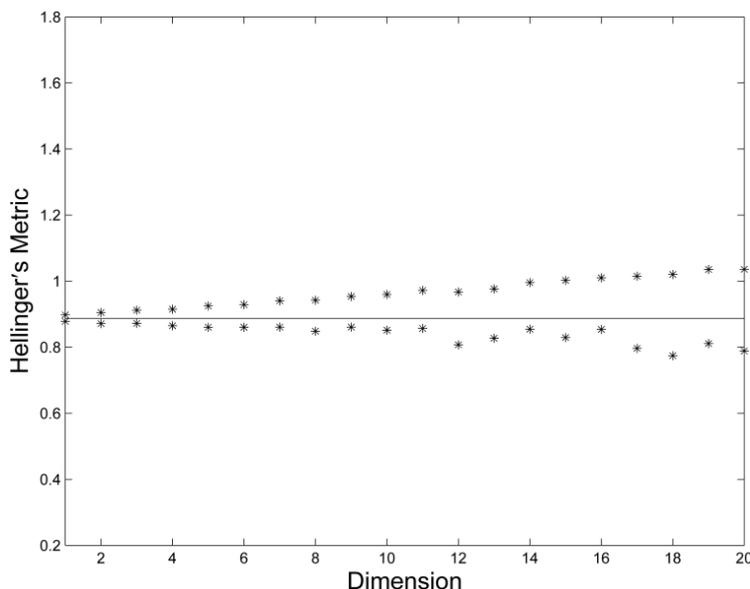
The detailed process of the two-stage estimator is presented in Algorithm 1. Notice that we studied up to dimension 20 with promising results. k-NN is a good alternative [31–36], but may present several difficulties, like the k determination [37], the distance measure choice [38] and the curse of dimensionality [39].

Algorithm 1—Two-stage resampling estimator

---

(1) COMMENT [To find the bandwidth of a cluster.
     Use the apriori knowledge from the instrumental data to estimate the bandwidth.
     If the precision of the instrumental data is not available, then use one of the preferred method to estimate bandwidth [26]; here it is used the Silverman rule and a cross validation search for maximum likelihood density [25].
(2) COMMENT [To estimate one trial.
     Determine the number of random samples to generate.
     Bootstrap method from Parzen's kernels with random generation of samples.
     Obtain the metric estimate given by Equation (41).
(3) COMMENT [To estimate the robust bonds, in the second step.
     Repeat the estimate.
     Calculate the 95% t-student interval from the estimates using Equation (42).
     Select the best intervals amplitudes given by Equation (44).
     Calculate the mean of the lower and the upper interval using Equation (45).
     Retain the maximum of the upper bound and the minimum of the lower bound given by Equation (46).


The implemented algorithm is available upon request.

3. Results and Discussion

To study the proposed resampling estimator behavior, we addressed several dimensions $(\aleph)$, different Parzen's coefficients $(\sigma = \sigma_1 = \sigma_2)$ as well as distinct Euclidean distances $(d = \|x_1 - x_2\|)$.

Firstly, we studied the estimator from dimension 1 (one) to dimension 20 and obtained the results shown in Figure 5, which let us verify that the exact value was always within the estimated interval.

Figure 5. Behavior of the new robust two-stage resampling interval estimator regarding dimensions from 1 to 20. (The exact value of the nonparametric Hellinger's metric is represented by a continuous line and is always contained in the estimated interval.)
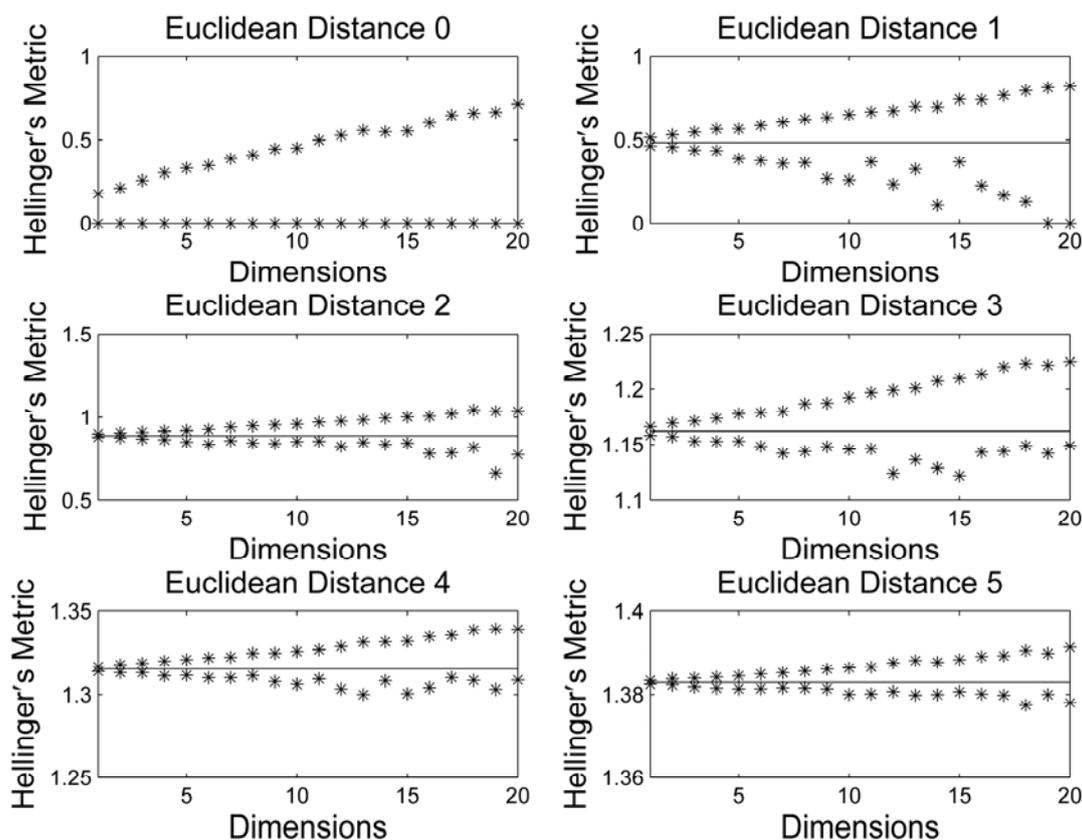
If the precision needed is not enough to generate disjoint intervals in competitive scenarios composed by multiple hypotheses, then the two-stage resampling can be repeated using a higher N_1, see Figure 6.

One can see in Figure 6 that the interval decreases with the increase of random samples, and that the exact value of nonparametric Hellinger's metric, which is represented by a continuous line, is always contained in the estimated interval.

To verify the behavior of the resampling Estimator with the Parzen's window σ^2 variation, we studied the results for 0.1 to 2 with 0.1 increases, Figure 7. In all the cases, the exact value is within the estimated error bound. Hence, the error bound estimator proposed here leads to robust intervals estimation.

Figure 6. Illustration of an asymptotic study of the novel robust two-stage resampling interval estimator between the upper and lower interval limits for different number of random samples (a detailed view regarding the Euclidean distances between 9 and 10 was added to the 1,000 samples graphic so the behavior of the estimator can be easily confirmed.)
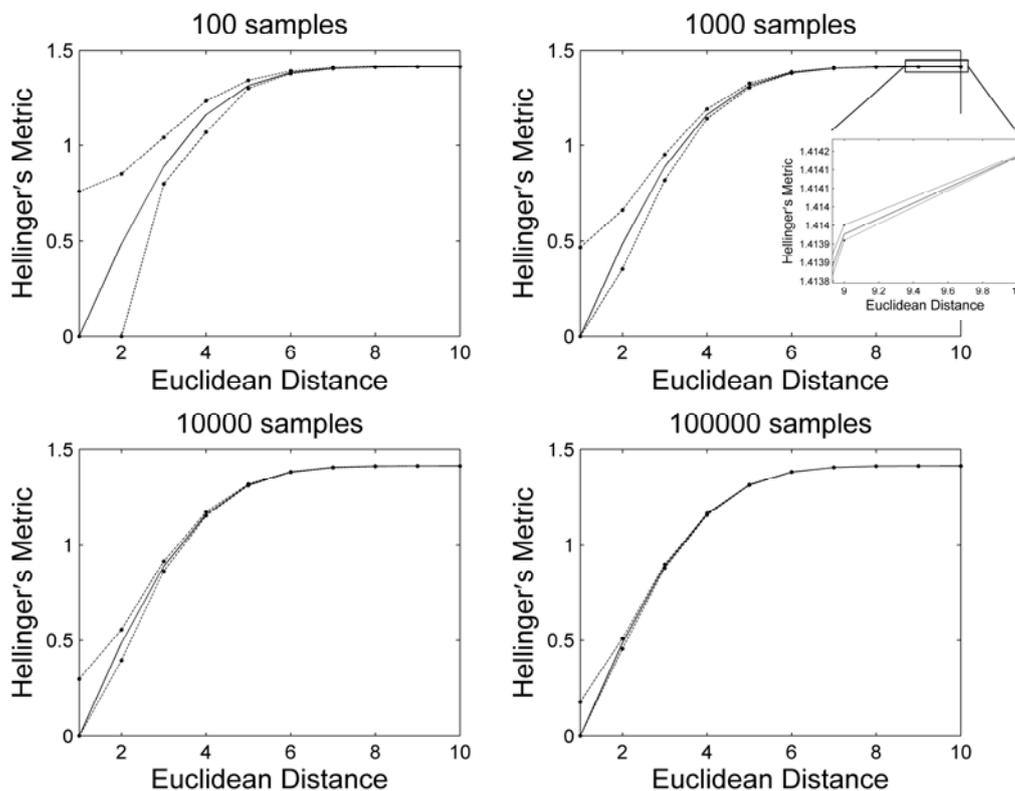

Figure 7. Illustration of the behavior of the new robust two-stage resampling interval estimator with the Parzen's window variation from 0.1 to 2; the graphics regard Euclidean distances from 0 (zero) to 5; the upper and lower interval limits always contain the exact value that is represented by a continuous line.
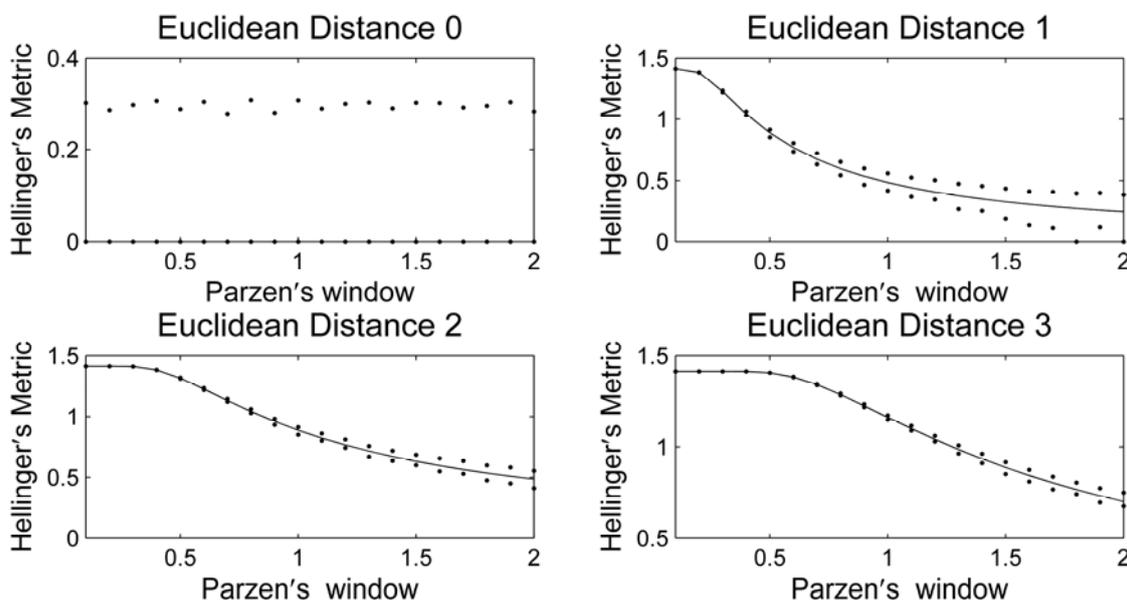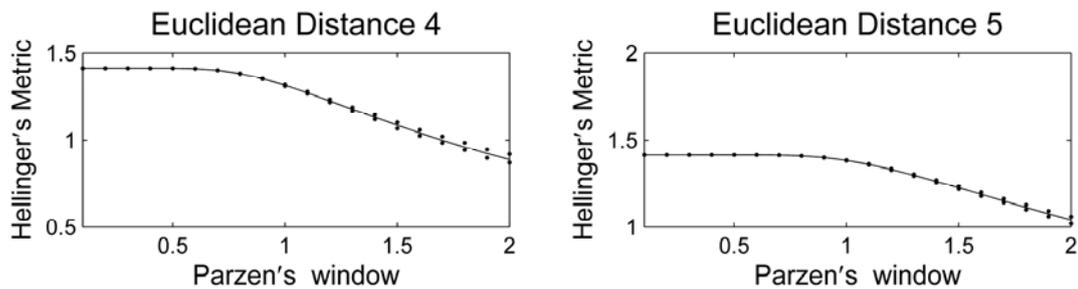
Figure 7. *Cont.*

4. Conclusions

Hellinger's metric was obtained from the generalized differential entropies and divergences. A nonparametric metric estimator based on Parzen's window was introduced. We proposed a change of measure to allow a resampling method. With the change of measure proposed, it was possible to design a new two-stage resampling error bound estimator. The resampling error bound estimator also has the advantage of resampling just one density (the sum of normalized product densities) given by a nonparametric Parzen's density with diagonal covariance, with asymptotic behavior. The new algorithm presented a robust behavior and very promising results. The asymptotic behavior allows to use this metric, in a competitive scenario with three or more densities, like clustering and image retrieval, to obtain disjoint intervals, simply by increasing the number of resampling samples. As to possible future work, two possible paths seem interesting: to evaluate Hellinger's metric behavior on medical image processing and analysis as in [40,41], and to assess k-NN entropy estimation capability with the metric and heterocedastic data addressed here.

Acknowledgments

Conflict of Interest

The authors declare no conflict of interest.

References

1. Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Crelle* **1909**, 210–271.
2. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Series B*. **1966**, *28*, 131–142.
3. Ullah, A. Entropy, divergence and distance measures with econometric applications. *J. Stat. Plan Inferace* **1996**, *49*, 137–162.
4. Principe, J.C. *Information Theoretic Learning Renyi's Entropy and Kernel Perspectives*; Springer: New York, NY, USA, 2010.

5. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2005; p. 483.

6. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.

7. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217.

8. Jeffreys, H. Fisher and inverse probability. *Int. Stat. Rev.* **1974**, *42*, 1–3.

9. Rao, C.R.; Nayak, T.K. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Trans. Inf. Theory* **1985**, *31*, 589–593.

10. Lin, J.H. Divergence measures based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.

11. Menéndez, M.L.; Morales, D.; Pardo, L.; Salicrú, M. (h, Φ)-entropy differential metric. *Appl. Math.* **1997**, *42*, 81–98.

12. Seth, S.; Principe, J.C. Compressed signal reconstruction using the correntropy induced metric. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, March 31–April 4, 2008; pp. 3845–3848.

13. Topsoe, F. Some inequalities for information divergence and Related measures of discrimination. *IEEE Trans. Inf. Theory* **2000**, *46*, 1602–1609.

14. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412.

15. Cubedo, M.; Oller, J.M. Hypothesis testing: A model selection approach. *J. Stat. Plan. Inference* **2002**, *108*, 3–21.

16. Puga, A.T.; Non-parametric Hellinger's Metric. In Proceedings of CMNE/CILANCE 2007, Porto, Portugal, 13–15 June 2007.

17. Hartley, R.V.L. Transmission of information. *Bell Syst. Tech. J.* **1928**, *7*, 535–563.

18. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

19. Rényi, A. *On Measures of Entropy and Information*, *Fourth Berkeley Symposium on Math. Statist. and Prob*; University of California: Berkeley: CA, USA, 1961; Volume 1, pp. 547–561.

20. *Nonextensive Entropy: Interdisciplinary Applications*; Gell-Mann, M.; Tsallis, C., Eds.; Oxford University Press: New York, NY, USA, 2004.

21. Wolf, C. Two-state paramagnetism induced by Tsallis and Renyi statistics. *Int. J. .Theor. Phys.* **1998**, *37*, 2433–2438.

22. Gokcay, E.; Principe, J.C. Information theoretic clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 158–171.

23. Ramshaw, J.D. Thermodynamic stability conditions for the Tsallis and Renyi entropies. *Phys. Lett. A* **1995**, *198*, 119–121.

24. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **2002**, *70*, 419–435.

25. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: London, UK, 1986.

26. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; Wiley: New York, NY, USA, 1992.

27. Devroye, L. *Non-Uniform Random Variate Generation*; Springer-Verlag: New York, NY, USA, 1986.

28. Preda, V.C. The student distribution and the principle of maximum-entropy. *Ann. Inst. Stat. Math.* **1982**, *34*, 335–338.

29. Kapur, J.N. *Maximum-Entropy Models in Science and Engineering*; Wiley: New York, NY, USA, 1989.

30. *The Probable Error of a Mean*. Available online: http://www.jstor.org/discover/10.2307/2331554?uid=2&uid=4&sid=21102107492741/ (accessed on 28 April 2013).

31. Leonenko, N.; Pronzato, L.; Savani, V. A class of Renyi information estimators for multidimensional densities. *Ann. Stat.* **2008**, *36*, 2153–2182.

32. Li, S.; Mnatsakanov, R.M.; Andrew, M.E. k-nearest neighbor based consistent entropy estimation for hyperspherical distributions. *Entropy* **2011**, *13*, 650–667.

33. Penrose, M.D.; Yukich, J.E. Laws of large numbers and nearest neighbor distances. In *Advances in Directional and Linear Statistics*; Wells, M.T., SenGupta, A., Eds.; Physica-Verlag: Heidelberg, Germany, 2011; pp. 189–199.

34. Misra, N.; Singh, H.; Hnizdo, V. Nearest neighbor estimates of entropy for multivariate circular distributions. *Entropy* **2010**, *12*, 1125–1144.

35. Mnatsakanov, R.; Misra, N.; Li, S.; Harner, E. k-Nearest neighbor estimators of entropy. *Math. Method. Stat.* **2008**, *17*, 261–277.

36. Wang, Q.; Kulkarni, S.R.; Verdu, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Inf. Theory* **2009**, *55*, 2392–2405.

37. Hall, P.; Park, B.U.; Samworth, R.J. Choice of neighbor order in nearest-neighbor classification. *Ann. Stat.* **2008**, *36*, 2135–2152.

38. Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J.B.O. Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412–2422.

39. Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When Is "Nearest Neighbor" Meaningful? In Proceedings of 7the International Conference on Database Theory, Jerusalem, Israel, 10–12 January 1999; pp. 217–235.

40. Vemuri, B.C.; Liu, M.; Amari, S.I.; Nielsen, F. Total bregman divergence and its applications to DTI analysis. *IEEE Trans. Med. Imag.* **2011**, *30*, 475–483.

41. Liu, M.; Vemuri, B.; Amari, S. I.; Nielsen, F. Shape retrieval using hierarchical total bregman soft clustering. *IEEE T. Pattern Anal.* **2012**, *34*, 2407–2419.