Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces

Frédéric Cazals, Alix Lhéritier

Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces

Frédéric Cazals and Alix Lhéritier

Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces

Frédéric Cazals and Alix Lhéritier

Project-Team Algorithms-Biology-Structure

Abstract: Comparing two sets of multivariate samples is a central problem in data analysis. From a statistical standpoint, the simplest way to perform such a comparison is to resort to a non-parametric two-sample test (TST), which checks whether the two sets can be seen as i.i.d. samples of an identical unknown distribution (the null hypothesis). If the null is rejected, one wishes to identify regions accounting for this difference. This paper presents a two-stage method providing *feedback* on this difference, based upon a combination of statistical learning (regression) and computational topology methods.
Consider two populations, each given as a point cloud in $\mathbb{R}^d$. In the first step, we assign a label to each set and we compute, for each sample point, a discrepancy measure based on comparing an estimate of the conditional probability distribution of the label given a position versus the global unconditional label distribution. In the second step, we study the height function defined at each point by the aforementioned estimated discrepancy. Topological persistence is used to identify persistent local minima of this height function, their *basins* defining regions of points with high discrepancy and in spatial proximity.

Experiments are reported both on synthetic and real data (satellite images and handwritten digit images), ranging in dimension from $d = 2$ to $d = 784$, illustrating the ability of our method to localize discrepancies.

On a general perspective, the ability to provide feedback downstream TST may prove of ubiquitous interest in exploratory statistics and data science.

Key-words: Statistics, Information theory, Jensen-Shannon divergence, Data analysis, Data comparison, Point clouds, Nonparametric two-sample test, Effect size, Divergence estimation, Conditional probability estimation, Regression, Topological persistence.

Sur la localisation de la discrépance dans les espaces de grande dimension

Résumé : Comparer deux ensembles de données multivariées est un problème central en analyse de données. D'un point de vue statistique, les tests non-paramétriques d'homogénéité permettent de décider si les données peuvent être considérées comme émanant d'une même distribution (l'hypothèse nulle). Si celle-ci est rejetée, la question se posant est de localiser les régions rendant compte de la différence. Ce travail présente une méthode en deux étapes pour ce faire, combinant des outils d'analyse statistique (régression) et de topologique (persistance).

Considérons deux populations, chacune donnée comme une ensemble de points dans $\mathbb{R}^d$. Dans la première étape, un label est donné à chaque population, et on calcule pour chaque point une mesure de discrépance basée sur la comparaison d'une estimation de la probabilité conditionnelle du label étant donnée la position, et de la probabilité non conditionnelle du label. Dans la deuxième étape, on étudie la fonction hauteur définie en chaque point par la valeur de la discrépance. La persistance topologique est utilisée pour identifier les minima persistants de cette fonction, leurs bassins définissant des ensembles de points de forte discrépance voisins les uns des autres.

Des résultats expérimentaux sont présentés sur des données synthétiques et des images (satellitaires et de chiffres), allant de la dimension $d = 2$ à $d = 784$, illustrant la pertinence de l'approche pour localiser la discrépance.

Dans une perspective plus large, le complément d'information apporté aux tests à deux échantillons devrait s'avérer de grande importance en analyse exploratoire de données.

Mots-clés : Statistique, Théorie de l'information, Divergence de Jensen-Shannon, Analyse de données, Comparaison de données, Nuages de points, Test non-paramétrique d'homogénéité, Taille d'effet, Estimation de la divergence, Estimation de probabilités conditionnelles, Régression, Persistance topologique.

Contents

1   Introduction

1.1   Comparing Datasets in High Dimensional Spaces

Datasets represented as point clouds are ubiquitous in science and engineering, used for applications in 2D and 3D space (e.g. to represent laser scans in mock-up design) as well as in high dimensional spaces (e.g. to represent images and documents, physical or biological phenomena, etc). In manipulating such data, several classes of questions are faced, such as matching, topological inference, or comparison. This latter endeavor, which is the topic of this paper, calls for a discussion in three directions, namely statistics (two-sample tests), information theory and learning (divergence estimation), and geometry-topology.

From a statistical standpoint, a broad class of comparison methods, requiring tame assumptions on the data are nonparametric two-sample tests (TST), see, e.g., [13] and the references therein. In a nutshell, a TST is a statistical hypothesis test checking whether the two sets can be seen as i.i.d. samples of an identical unknown distribution (the null hypothesis). In accepting or rejecting the null hypothesis, under a level of statistical significance α, a TST summarizes the body of information encoded in the points' coordinates into a single boolean value [11]. However, this boolean information is in general of limited interest, for several reasons. First, it is unlikely that two real life datasets come from exactly the same distribution. Since consistent TST detect any kind of difference of any size if enough samples are given, the rejection of the null is expected. Second, the magnitude (and the nature) of the differences, known as *effect size*, usually conveys more information than the mere presence of a difference. Therefore, a reject decision should be just a signal to examine further the data in order to understand. Developing a notion of effect size for non-parametric two-sample tests in high dimensions has not been explored yet, and is the goal of this work.

From an information theoretical and probabilistic standpoint, the comparison can be phrased as the problem of estimating the global Kullback-Leibler divergence between unknown distributions using the samples in hand. For example, a difference between images has been proposed [10], by coupling a univariate Kullback-Leibler estimate (per pixel) and a decomposition of the discrepancy map thus defined using a watershed transform in image space. In a more general setting, there exist techniques to estimate this quantity that avoid density estimation in high dimensions (see, e.g., [22, 23, 19, 20]) and that could be amenable to decomposition, so as to determine a contribution of individual points or of groups of points. Nevertheless, this divergence lacks important properties (symmetry and boundedness), which makes it more difficult to further process it.

Finally, the comparison can also be tackled from the geometric and topological perspectives. In geometric terms, one may compute some distance or matching (one-to-one, one-to-many, many-to-many) between the data, see [7] and the references therein. While this procedure is informative for the two datasets in hand, the main difficulty consists in accommodating a probabilistic setting. In a more topological perspective, persistence theory [8], which aims at assessing the stability of topological features–generators of persistent homology groups, was recently used to compare *persistence landscapes* [3]. Such comparisons are clearly important since oblivious to geometric transformations, but our focus is clearly on geometry dependent features.

1.2   Contribution and Paper Overview

This paper proposes, to the best of our knowledge, the first attempt to model the differences (the effect size, see above), between two datasets for which one has rejected the null hypothesis stipulating that they share the same underlying distribution. In a nutshell, we aim at clustering

samples, based on two criteria, namely samples within a cluster should (i) contribute significantly to the difference between the two clouds, and (ii) form a connected region. Matching these goals yields a two-stage procedure. In the first stage, we model pointwise differences, which we call *discrepancies*, using the Jensen-Shannon divergence (JSD), which is symmetric and can be decomposed in terms of the conditional probability of belonging to one of the populations given a space location. This conditional probability can be naturally estimated using known techniques of non-parametric regression like the one based on k_n nearest neighbours, which possesses strong asymptotic guarantees of consistency. In the second stage, using a nearest neighbor graph defined over the samples, we study the height function defined by the estimated discrepancy, and design a clustering procedure based upon topological persistence. We note in passing that the second step is optional, as the JSD is of interest on its own to compare images, for example.

The paper is organized as follows. Sections 2 and 3 respectively present the two steps. Section 4 summarizes the various pieces of information provided by our analysis. Finally, sections 5 and 6 present experiments.

2 Estimating the Discrepancy between Datasets

We aim at modeling the discrepancy between two datasets $x^{(n_0)} \equiv \{x_1, \ldots, x_{n_0}\}$ and $y^{(n_1)} \equiv \{y_1, \ldots, y_{n_1}\}$, in some fixed dimension Euclidean space $\mathbb{R}^d$. We view these data as coming from two unknown densities f_X and f_Y with corresponding cumulative distributions functions F_X and F_Y.

2.1 Jensen-Shannon Divergence Decomposition using Conditional Distributions

Let $D_{\mathrm{KL}}(f \| g)$ be the Kullback-Leibler divergence (KLD) between two densities f and g defined as

$$D_{\mathrm{KL}}(f \| g) \equiv \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx \tag{1}$$

with the conventions $0 \log 0 = 0$ and $0 \log \frac{0}{0} = 0$.

The Jensen-Shannon divergence (JSD) defined in [16], allows to symmetrize and smooth the KLD by taking the average KLD of f_X and f_Y to the average density $f \equiv {(f_X + f_Y)}/{2}$, that is:

$$JS(f_X \| f_Y) \equiv \frac{1}{2} \left(D_{\mathrm{KL}}(f_X \| f) + D_{\mathrm{KL}}(f_Y \| f) \right) \tag{2}$$

In addition to being symmetric, the JSD is bounded between 0 and 1 and its square root yields a metric. Note also that by taking the average, two random variables are implicity defined: a position variable Z with density $f_Z \equiv f$ and a binary label L that indicates from which original density (i.e. f_X or f_Y) an instance of Z is obtained. Formally, considering the alphabet $\mathcal{A} = \{0, 1\}$ and $X \sim F_X, Y \sim F_Y$, the following pair of random variables is defined:

$$(L, Z) = \begin{cases} (0, X) & \text{with prob. } \frac{1}{2} \\ (1, Y) & \text{with prob. } \frac{1}{2} \end{cases}. \tag{3}$$

In the sequel, we will consider the conditional and unconditional mass functions $P(l|z) = \mathbb{P}(L = l | Z = z)$ and $P(l) = \mathbb{P}(L = l) = \frac{1}{2}$ respectively, as well as the joint probability density $f_{L,Z}$. We will also use the notation f_l to denote f_X (resp. f_Y) when $l = 0$ (resp. $l = 1$).

Before establishing lemma 1, a key property for our comparison problem, we recall the definition of the Kullback-Leibler divergence between two discrete distributions P and Q over $\mathcal{A}$:

$$D_{\mathrm{KL}}\left(P \| Q\right) \equiv \sum_{l \in \mathcal{A}} P(l) \log \frac{P(l)}{Q(l)} \tag{4}$$

Lemma. 1. *One has:*

$$JS(f_X \| f_Y) = \int_{\mathbb{R}^d} f_Z(z) D_{\mathrm{KL}}\left(P(\cdot|z) \| P(\cdot)\right) dz \tag{5}$$

Proof of lemma 1. Recall that the JSD can be expressed as follows:

$$
\begin{aligned}
JS(f_X \| f_Y) &\equiv \frac{1}{2}\left(D_{\mathrm{KL}}\left(f_X \| f_Z\right) + D_{\mathrm{KL}}\left(f_Y \| f_Z\right)\right) \\
&= \frac{1}{2}\int_{\mathbb{R}^d} f_X(z) \log \frac{f_X(z)}{f_Z(z)} dz \\
&+ \frac{1}{2}\int_{\mathbb{R}^d} f_Y(z) \log \frac{f_Y(z)}{f_Z(z)} dz
\end{aligned}
$$

By linearity of integration and by Bayes' Theorem, and noting that the conditional densities $f\left(z|L=0\right) = f_X(z)$ and $f\left(z|L=1\right) = f_Y(z)$ we have then

$$
\begin{aligned}
JS(f_X \| f_Y) &= \int_{\mathbb{R}^d} \sum_l P(l) f\left(z|L=l\right) \log \frac{f\left(z|L=l\right)}{f_Z(z)} dz \\
&= \int_{\mathbb{R}^d} \sum_l P(l) \frac{f_{L,Z}(l,z)}{P(l)} \log \frac{f_{L,Z}(l,z)}{P(l) f_Z(z)} dz \\
&= \int_{\mathbb{R}^d} f_Z(z) \sum_l P(l|z) \log \frac{P(l|z)}{P(l)} dz \\
&\equiv \int_{\mathbb{R}^d} f_Z(z) D_{\mathrm{KL}}\left(P(\cdot|z) \| P(\cdot)\right) dz
\end{aligned}
$$

$\square$

The previous lemma shows that the JSD can be seen as the average, over $z \in \mathbb{R}^d$, of the KLD between the conditional and unconditional distribution of labels. More formally, we define:

Definition. 1. *The discrepancy at location z is defined as the KL divergence:*

$$\delta(z) \equiv D_{\mathrm{KL}}\left(P(\cdot|z) \| P(\cdot)\right). \tag{6}$$

Note that $\delta(z)$ ranges between 0 and 1 and is 0 iff $f_X(z) = f_Y(z)$. Note also that since $P(l)$ is known but $P(l|z)$ is not, the problem we consider now is the one of estimating $P(l|z)$ at each given location z.

2.2   Conditional Probability Estimation via Non-parametric Regression

In order to estimate the conditional distributions, we can use random design[1] non-parametric regression.

---

[1]In contrast to fixed design regression, where [14, Sec. 1.9]: "one observes values of some function at some fixed (given) points with additive random errors, and wants to recover the true value of the function at these points."

2.2.1 Generic Framework

First, we define the basic concepts (see, e.g., [14] for more details).

Definition. 2. *Given a random vector (Z, R), where $Z \in \mathbb{R}^d$ and the response variable $R \in \mathbb{R}$, the* regression function *is defined as*

$$m(x) = \mathbb{E}\left[R | Z = z\right]. \tag{7}$$

In the regression problem, the goal is to build an estimator $m_n(x)$ of $m(x)$ using a set of n i.i.d. realizations of (Z, R).

With respect to the guarantees that are provided for regressors, usually, the L_2 risk or mean squared error is considered, i.e.,

$$\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx)$$

where μ denotes the distribution of Z. Nevertheless, since our goal is to estimate the discrepancy $\delta(z)$, we seek pointwise guarantees for regressors. In particular, we will consider a strong form of consistency which is defined as follows.

Definition. 3. *Denoting μ the distribution of Z, a sequence of regression function estimates $\{m_n\}$ is* strongly pointwise consistent (s.p.c.) *if for μ-almost all $x \in \mathbb{R}^d$*

$$m_n(x) \xrightarrow{n \to \infty} m(x) \;\; a.s.. \tag{8}$$

In [14, Sec. 25.6], some s.p.c. regression estimates are presented. For example, regression estimates based on partitioning, kernel and nearest neighbors are s.p.c under certain conditions for their parameters, when the absolute value $|R| < M$, for some M.

Now we describe the s.p.c. k_n-nearest neighbor regression function estimate (see [14, Ch.6&25] for further details). Given the training data $\{Z_i, R_i\}_{i=1,\ldots,n}$, let us denote as $R_{(i,n)}(x)$ the response value corresponding to i-th nearest neighbor (with some tie-breaking rule) of x in Z^n. Then, the k_n-nearest neighbor (k_n-NN) regression function estimate is defined by

$$m_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} R_{(i,n)}(x). \tag{9}$$

Then we have the following theorem [14, Thm. 25.17]:

Theorem. 1 (Strong pointwise consistency of k-NN). *If $|R| < C$ for some $C < \infty$,*
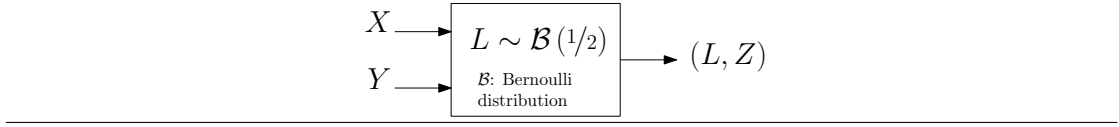
$$\frac{k_n}{\log n} \to \infty \;\; and \;\; \frac{k_n}{n} \to 0,$$

then the k_n-NN estimate using Euclidean distance is strongly pointwise consistent.

2.2.2 Application to Conditional Probability and Discrepancy Estimation

In order to apply this framework to our problem, note that the correspondence $R = L$ yields

$$m(z) = P(1|z). \tag{10}$$

Figure 1 Random multiplexer generating pairs (label, position).

Then, we can use the following estimator for $P(l|z)$

$$\hat{P}_n(l|z) \equiv |1 - l - m_n(z)|. \tag{11}$$

Note that it is required that $0 \leq m_n(z) \leq 1$, since we aim at estimating a conditional probability, and that it is satisfied by the k_n-nearest neighbor regressor.

Using Eq. (11), we finally obtain an estimator for $\delta(z)$:

$$\hat{\delta}_n(z) \equiv D_{\mathrm{KL}}\left(\hat{P}_n(\cdot|z) \| P(\cdot)\right) \tag{12}$$

Theorem. 2 (Consistency). *Let $\hat{P}_n$ be based on a s.p.c. sequence of regression estimates for (L, Z). Then,*

$$\hat{\delta}_n(z) \xrightarrow{n \to \infty} \delta(z) \ a.s.$$

for f-almost all $z \in \mathbb{R}^d$.

Proof. Let us write $\hat{\delta}_n(z) = g(m_n(z))$, with $g(x) = x \log x/(1/2) + (1-x) \log(1-x)/(1/2)$. It is easy to see that $g(x)$ is a continuous function (composition, sum and product of continuous functions). Then, we apply the continuous mapping theorem [2], on every $z \in \mathbb{R}^d$ where $m_n(z) \xrightarrow{n \to \infty} m(z)$ to complete the proof. $\square$

2.3 Joint Distribution Compatible Sampling

In the regression framework, the samples must be i.i.d. from a joint distribution $f_{L,Z}$. In our original problem, we have two sets of samples drawn independently from f_X and f_Y. In order to ensure this condition, we will use the random multiplexer depicted in Fig. 1. On each input it receives i.i.d. samples from each of the populations. Then, it generates an instance l of L. According to the value of l (0 or 1), it consumes the corresponding input (X, Y resp.) and outputs it along with l.

The following lemma shows that the output has the desired joint density.

Lemma. 2. *An output pair from the random multiplexer has joint density $f_{L,Z}$.*

Proof. The joint density of an output pair (l, z) is

$$g(l, z) = g(z|L = l)P(l).$$

Since z is distributed as f_l, $g(z|L = l) = f_l(z)$. Therefore $g(l, z) = f_{L,Z}(l, z)$. $\square$

Remark 1. *In practice, there is a finite set of i.i.d. samples of X and Y available. Then, at some point the multiplexer can have no more data to consume on one of the inputs while there is still data available on the other one. Therefore, some samples of the original sets would not be used and, thus, some loss of information is to be expected. This can be alleviated by resampling B times as follows:*

1. *For $b \in 1..B$ do:*

   (a) *Generate a sequence $\{z'_i, l'_i\}_{i=1,...,n'}$ using the random multiplexer (until it fails to output, due to lack of data on one of the inputs)*

   (b) *Build $\hat{\delta}_{n'}^b(z)$ using some consistent estimator trained on $\{z'_i, l'_i\}_{i=1,...,n'}$*

2. *Define $\bar{\delta}_{n'}(z) \equiv median_{b \in 1..B}(\hat{\delta}_{n'}^b(z))$*

   Notice that $\bar{\delta}_{n'}(z)$ is also a consistent estimator.

3 Localizing the Discrepancy

3.1 Overview

Goals. We wish to identify groups of samples, called *clusters*, intuitively characterized by two properties: first, the discrepancy of such samples should be significant; second, samples within a cluster should be associated with regions where the discrepancy peaks. To meet the first goal, we assume the existence of a value δ_{max} stipulating that below δ_{max}, the discrepancy is not significant. As we shall see in Experiments, while this value is not unique in general, few *clusters* typically stand out from our persistence diagrams.

To meet the second goal, we resort to *mode clustering*, a general clustering strategy consisting of defining a cluster from the attraction basin of a local maximum of a density estimate [6, 5]. We therefore define a *landscape* consisting of the samples, their elevation being the value of the estimated discrepancy $\hat{\delta}_n(z)$, see Eq. (12). Practically though, we study the landscape whose elevation is $-\hat{\delta}_n(z)$ rather than $\hat{\delta}_n(z)$, which yields a more natural terminology since birth dates occur before death dates. In doing so, our clusters shall be defined from local minima and their attraction basin, i.e., stable manifolds (SM).

Varying the threshold δ_{max}: one versus many. Assigning samples of the landscape to persistent local minima is straightforward, and merely requires running a union-find algorithm [5]. The clusters obtained can then be filtered out so as to retain samples whose discrepancy is larger than δ_{max}. However, this procedure must be repeated upon changing the value δ_{max}. In the sequel, we present instead a procedure pre-processing the landscape so as to accommodate queries for multiple values of δ_{max}.

In a nutshell, our algorithm runs through three stages. First, critical points of the landscapes are identified using a k-nearest neighbor graph (k-NNG) [5]. The connections between these points define the Morse-Smale-Witten (MSW) complex restricted to local minima and index one saddles, from which we compute the persistence diagram (PD) of local minima [9]. (We note in passing that similarly to [5], our procedure shall be effective in high dimension since we only focus on index 0 and 1 critical points.) The PD is used to identify persistent minima whose critical value is at most $-\delta_{max}$, and we denote the corresponding sublevel set of the landscape $D_{\leq -\delta_{max}}$. Second, the sublevel set of the landscape is recursively simplified to retain persistent minima only [4]. We note in passing that this simplification requires more information than that defining the persistence pairs (Fig. 3). Third, the samples whose discrepancy is less than δ_{max} are removed from the stable manifolds of the remaining minima.

Output. The previous construction exploits a partition of the PD into five regions defined by three lines (Fig. 2), so that a local minimum m of the landscape (and its SM) gets qualified with respect to three criteria:

- Selected/rejected: m is selected provided that its birth date occurs before $-\delta_{max}$.

Figure 2 Partition of the persistence diagram exploited to define clusters. The partition of the domain $y \geq x$ is induced by three lines: (i) $y = x + \rho$ which specifies the persistence threshold (ii,iii) $x = -\delta_{max}, y = -\delta_{max}$, with δ_{max} the threshold on the significance of the discrepancy. See text for the specification of regions R_1 to R_5.
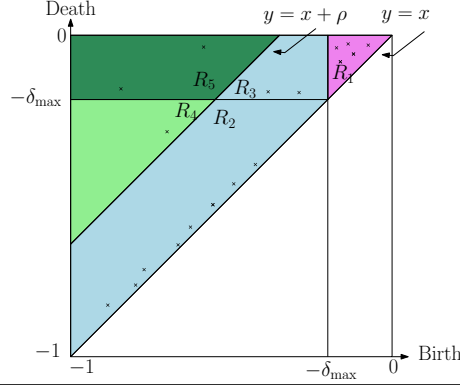
- Persistent/canceled: m is persistent if its persistence is $\geq \rho$, a user defined threshold.

- Filtered/un-filtered: the SM of m is filtered if the death date of m is larger than the threshold $-\delta_{max}$.

The possible combinations, illustrated on Fig. 2, are:

- $m \in R_1$: rejected. Such a local minimum is rejected, since its discrepancy is less than δ_{max}. No point of the SM of m is found in a cluster reported.

- $m \in R_2$: selected / canceled / un-filtered. Such a local minimum is selected, yet canceled by persistence. Because m dies before $-\delta_{max}$, all samples found in its SM shall be part of a cluster reported.

- $m \in R_3$: selected / canceled / filtered. A local minimum which is selected, yet canceled by persistence. However, because m dies after $-\delta_{max}$, only the portion of its SM belonging to the sublevel set $D_{\leq -\delta_{max}}$ shall be found in a cluster reported.

- $m \in R_4$: selected / persistent / un-filtered. Such a local minimum is selected, and is not canceled by persistence. Because m dies before $-\delta_{max}$, all the samples found in its SM are found in a cluster reported.

- $m \in R_5$: selected / persistent / filtered. This combination is similar to the previous case, except that samples whose discrepancy is less than δ_{max} are discarded. Note that the cluster associated with the global minimum belongs to this region even though it is not found on the PD since the global minimum never dies.

3.2   Algorithm

We now detail the three steps just outlined.

Step 1: Computing the MSW complex of the landscape. Morse theory is concerned with the study of a function defined on a manifold, and Morse homology with the homology of sublevel sets of this function. In particular, the Morse homology theorem stipulates that the homology

of a sublevel set can be computed from the Morse-Smale-Witten (MSW) complex, namely the incidence diagram between the critical points of the function [1].

A natural strategy to study a height function defined over a point cloud consists in using a k-NNG connecting the samples. One defines a negative pseudo-gradient from the star of each vertex, and a flow operator descending the pseudo-gradient until a fixed point is found [5]. Note that the set of all samples flowing to a local minimum makes up its stable manifold (SM). Given this pseudo-gradient, samples behaving like index 0 and index 1 critical points in the smooth setting are easily identified, and abusing terminology, we call such samples *critical points* in the sequel. An index 0 critical point is a sample having all its neighbors above it. An index 1 critical point is a sample p *flowing* to a local minimum, but having a neighbor q in the k-NNG flowing to a different local minimum. This latter situation is called a *bifurcation*, since intuitively, the line-segment $[p, q]$ intersects transversely the stable manifold of an index $d - 1$ critical point. Amidst all pairs p, q associated with the same two local minima, the sample with least elevation is termed a *saddle*. (Note that we do not make any claim on the relative position of that point and the real saddle, in case the landscape is associated with a differentiable height function.) If the landscape contains n_0 critical points and is connected, $n_0 - 1$ index one points suffice to compute the persistence of order 0. The process indeed boils down to running a Union-find algorithm, to infer merge events between the stable manifolds of local minima (Fig. 3(A,B,C)).

We define our MSW complex in a similar spirit. However, instead of collecting only incidences involving one of the aforementioned $n_0 - 1$ index one saddles, we collect incidences involving all index one saddles. The process yields a bipartite graph between index 0 and index 1 critical points (Fig. 3(D,E)), together with the stable manifolds of the local minima. Using this graph, we also compute the persistence diagram (PD) of sublevel sets. Note that in our case all pairs lie in the upper triangle of $[-1, 0] \times [-1, 0]$, since the function value, namely the negative discrepancy, lies in the range $[-1, 0]$.

Step 2: Simplifying the landscape. The general procedure to recursively simplify a landscape using the MSW complex has already been presented in the context of non manifold shape reconstruction [4]. In a nutshell, the cancellation of a pair of critical points (a, b) whose indices differ by one consists of rerouting the connections of a and b in the MSW complex, and of redistributing the stable manifold of a [4]. Note in particular that each remaining local minimum is endowed with two types of samples: those from its own SM and those from SM inherited from canceled local minima.

Step 3: Sub-level set extraction. The previous simplification yields a partition of the landscape into the SM of the persistent minima. We remove from these SM the samples whose discrepancy is less than δ_{max}, a task carried out in two steps. First, the samples from its own SM are filtered out. Second, the samples of basins inherited from the simplification are also filtered out, provided that such a basin was born before $-\delta_{max}$. In particular, inherited basins born after $-\delta_{max}$ are ruled out in constant time. The persistent local minima and their remaining samples, if any, form the clusters.

Remark 2. *The filtering step cannot be carried out before the construction of the nearest neighbor graph. Indeed, in doing so, one could deplete the neighborhood of samples whose height is close to $-\delta_{max}$, possibly forcing connections to samples located further away, thus jeopardizing the identification of critical points.*

Figure 3 Morse-Smale-Witten (MSW) complex versus disconnectivity graph (DG) in recursive landscape simplification. (A,B) Two landscapes, with critical points of indices 0 (disks) and 1 (squares). **(C)** The DG of both landscapes, which depicts the evolution of connected components of sublevel sets. Despite the differences between their MSW complexes, both landscapes share the same DG: upon passing the critical point e, the stable manifold of c merges with that born at a. **(D,E)** The MSW complexes of **(A,B)**, respectively. In cancelling the pair of critical points (c, e), one does not know from the DG with which (a or b) the basin of c should be merged. But the required information is found in the MSW complex: on the landscape (A), c is merged with a; on the landscape (B), c is merged with b.
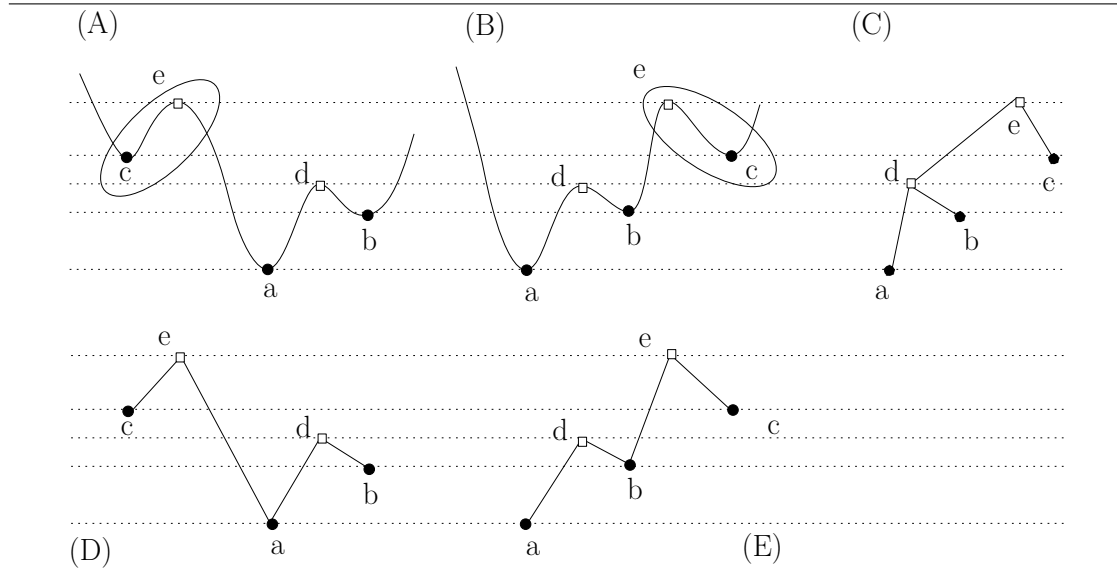
4 Combining Discrepancy Estimation and Localization

4.1 Qualifying the Clusters

We decompose the JSD by clusters of points that are defined by the method described in Section 3. Then, the contribution of a cluster C reads as:

$$JS_C\left(f_X \| f_Y\right) \equiv \frac{1}{n} \sum_{z \in z^n \cap C} \hat{\delta}_n(z). \tag{13}$$

Combining the analysis of sections 2 and 3 yields the workflow of Fig. 4.

4.2 Plots

The previous analysis are best exploited using the following plots:

- *Raw data embedding:* For samples embedded in 2D or 3D space, a plot of the points with a color to indicate the label (blue: 0; red: 1). For samples embedded in a higher dimensional space, a 2D embedding of these samples obtained using multi-dimensional scaling (MDS). In any case, the goal of this plot is to intuitively visualize the distributions of the two populations.

- *Discrepancy shaded data embedding — aka discrepancy plot:* A plot similar to the raw data plot, except that each sample is color coded using a heat palette from fully transparent white to red across yellow, as a function of the value of the estimated discrepancy $\hat{\delta}_n(z)$. That is, a point with $\hat{\delta}_n(z)$ equal to zero (resp. one) is colored fully transparent white (resp. non transparent red).

- *Persistence diagram:* A plot showing for each minima a red cross with coordinates (x, y) corresponding to its birth and death dates respectively, while analyzing the landscape whose elevation is the negative estimated discrepancy.

- *Clusters:* A plot similar to the raw data plot, with one color per cluster. The points not belonging to any cluster are colored in gray.

- *JSD decomposition plot:* A 1D plot presenting a synthetic view without relying on the MDS embedding. The x coordinate represents the sample space and always ranges from 0 to 1. The total estimated JSD is represented by the area under the dashed line. And the maximum possible JSD which is always 1 is represented by the area under the continuous line. One bar is depicted for each cluster plus another one (the last one) for the points not belonging to any cluster. The area of each bar represents the contribution of the corresponding cluster to the total JSD and its color corresponds to the proportion of samples 0 in the cluster.

4.3 Implementation

The random multiplexer was implemented in R. The discrepancy estimator was implemented in R using knn3 from the caret package [12]. The persistence based analysis of the height function defined by the estimated discrepancy was implemented in C++.

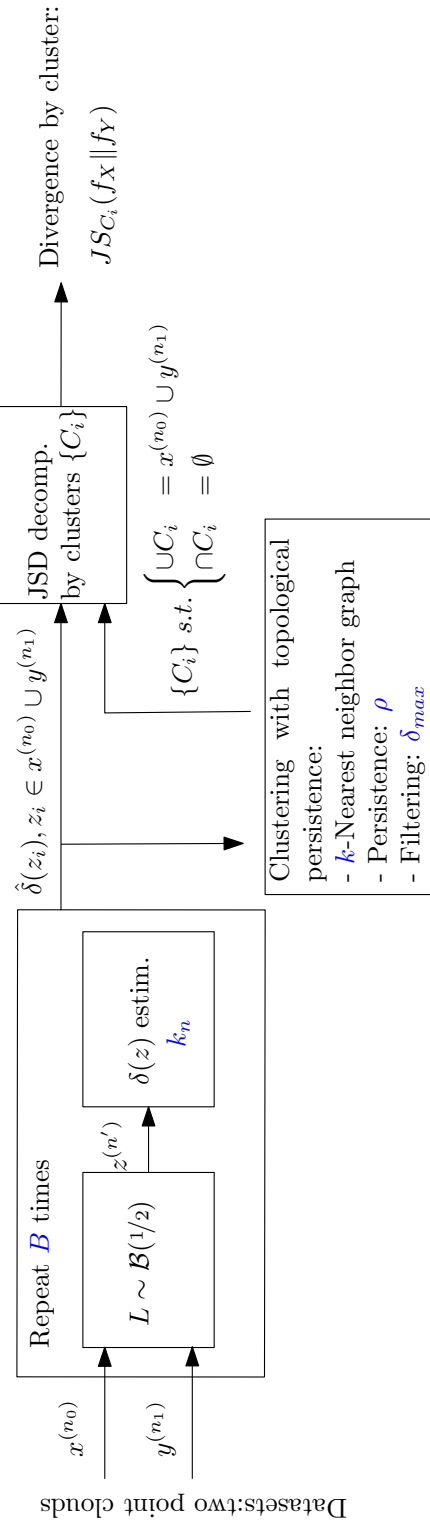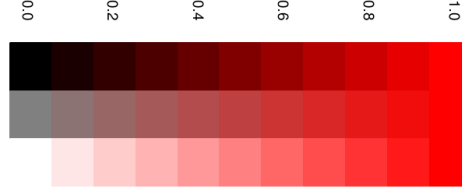**Figure 4** Workflow of the whole method. In blue: the parameters.

Figure 5 Comparing images: interpolated color scale (Eq. (14)) used to represent the local discrepancy $\hat{\delta}_n(z) \in [0,1]$ of Eq. (12). The interpolation is illustrated on pixels of color black, gray or white.

5 Experiments: Using the Discrepancy for Statistical Image Comparison

When processing real images, two-sample testes are of mild interested since the null hypothesis is likely to be rejected. However, the JSD decomposition is still of interest to quantify the differences on a statistical basis.

Consider a digital image whose pixels use C color channels. For example, $C = 1$ in the monochrome case and $C = 3$ in the RGB color case (where the components of each vector correspond to the red/blue/green intensities). A digital image is a $r \times c$ matrix of pixels, such that each pixel takes values in $[0,1]^C$. (For a pixel, a value of 0 (resp. 1) represents the minimum (resp. the maximum) intensity for the corresponding color channel.) We follow the construction of [21] to build our samples, that is, by taking $b \times b$ pixel blocks yielding $(r-b)(c-b)$ samples, each being a vector of dimension Cb^2. Then, a discrepancy estimate $\hat{\delta}_n(z)$ is computed on each sample z and assigned to the pixel located in the upper left corner of the corresponding block. (NB: two bands of width $b-1$ on the right and bottom side of the image are not assessed.)

Using the satellite color images of [21] shown in Figure 6 and using the same block size $b = 2$, we compute the estimated discrepancy $\hat{\delta}_n(z)$ for each sample. Then, in order to visualize the result, we first convert the original image to grayscale by assigning to each pixel i with color vector $c_i = [r_i, g_i, b_i]$ the new color vector $c_i' = [a_i, a_i, a_i]$ where $a_i = (r_i+g_i+b_i)/3$. Then, we obtain the final color vector $c_i'' = [r_i'', g_i'', b_i'']$ by superimposing the corresponding discrepancy δ_i via red interpolation as follows (Fig. 5)

$$\begin{cases} r_i'' & = (1-\delta_i)a_i + \delta_i \\ g_i'' & = (1-\delta_i)a_i \\ b_i'' & = (1-\delta_i)a_i \end{cases}. \tag{14}$$

The results, using $k_n = n^{1/3}$, are shown in Figure 7. The methods evaluated in [21] aim at finding novelty in the second image with respect to the first one. Although in a different setting, our results are visually quite good in comparison to theirs: only scattered red points are shown in the first figure (which is consistent to their idea of background) and, in the second one, both the oval and the rectangular fields are clearly marked as divergent while none of the methods evaluated in [21] mark them both entirely.

Figure 6 Original satellite color images from [21].

Figure 7 Comparison between the images of Fig. 6. Results shown on grayscale converted image using the interpolation scheme of Eq. (14), illustrated on Fig. 5.
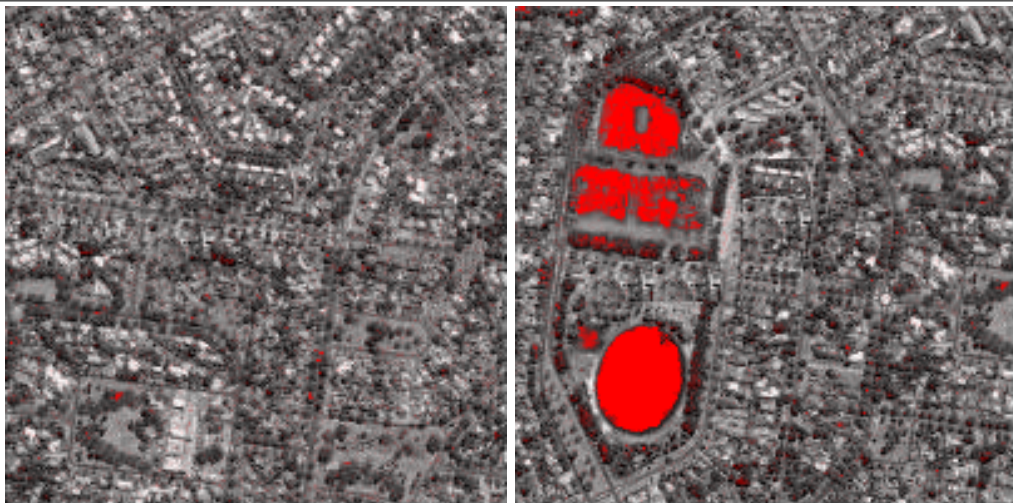
6 Experiments: Localizing Data Discrepancies

We present results on three datasets featuring various difficulties: low intrinsic dimension (crenels), varying intensity of discrepancy (mixture of Gaussians) , and real data embedded in high dimension (handwritten digits). For the sake of convenience, we refer to the two datasets to be compared as the blue and the red datasets. The number of neighbors was set to $k_n = n^{2/3}$. Note that the Maximum Mean Discrepancy two-sample test of [13] rejects the null hypothesis in all the cases for a significance level α lower than 1%.

6.1 Model: Crenels

Specification. The goal is to assess the ability of the method to spot local differences, and to cope with data of low intrinsic dimension (one) in a high dimensional space. We create two one dimensional datasets, which differ by two crenels.

More precisely, in this dataset, each point corresponds to a vector in $\mathbb{R}^{m \times m}$ encoding the pixels of a square grayscale image (0 = black, ≥ 1 = white) of size $m \times m$. The samples are the result of rotating the grayscale image i (supp. Fig. 12). Therefore, taking $m = 11$ yields a dataset of intrinsic dimension one embedded in dimension $d = 121$.

Each sample of the blue population is an instance of a RV X obtained by rotating i with a random uniformly distributed angle A_X, that is:

$$X = rotate(i, A_X), A_X \sim \mathcal{U}(s,t). \tag{15}$$

where the function *rotate* applies a bilinear filter to smooth the result (details in [18]).

For the red population, consider two Bernoulli random variables $B_1 \sim \mathcal{B}(p_1)$ and $B_2 \sim \mathcal{B}(p_2)$, and two uniform variables $U_1 \sim \mathcal{U}(a,b)$ and $U_1 \sim \mathcal{U}(c,d)$. Each sample of the red population is an instance of a RV Y defined as:

$$Y = rotate(i, A_Y), A_Y = B_1(B_2 U_1 + (1 - B_2)U_2) + (1 - B_1)A_X. \tag{16}$$

Note that the rotation used to obtain Y comes from the uniform distribution $\mathcal{U}(s,t)$ with probability $1 - p_1$, and that the discrepancy between both distributions are high in the angle ranges $[a,b]$ and $[c,d]$ (i.e. the *crenels*) and low everywhere else on the support, since, loosely speaking, points that are added to the crenels are missing from the rest of the support.

Practically, we used the following values: $m = 11$, $n_0 = 2000$, $n_1 = 2000$, $s = -15, t = 15$, $a = -4, b = -2, c = 9, d = 10$, $p_1 = 0.3, p_2 = 0.5$.

Practically, we used:

- $m = 11$

- $n_0 = 2000$, $n_1 = 2000$

- $s = -15, t = 15$

- $a = -4, b = -2, c = 9, d = 10$

- $p_1 = 0.3, p_2 = 0.5$

Results. Figure 8 shows the result of our method when applied to this dataset. The linear shape of the 2D MDS embedding illustrates the one dimensional nature of the data. The discrepancy plot hints at the crenels created by the two uniform distributions in Eq. (16). On the persistence diagram (built with $k = 30$), we see a group of low discrepancy minima that are removed by

filtering out with δ_{max} (dashed vertical line). One also identifies one persistent local minimum corresponding to the longer and, thus, the less red-concentrated crenel $[a, b]$. The other crenel corresponds to the global minimum which does not appear in the plot since it death date is infinite. The clusters yielded by the simplification and filtering steps correspond to the crenels. In the divergence decomposition plot, we observe these two crenels with high discrepancy produced by a high proportion of red points and also a non negligible total discrepancy given by the rest of the points, which has a higher proportion of blue points.

6.2   Model: Gaussian Mixture

Specification. The goal is to assess the ability of the method to spot regions of different intensity of discrepancy.

Two Gaussian mixture models where randomly generated using MixSim R package [17]. The distributions for X and Y consist in two mixtures of four different two-dimensional Gaussians with equal weight and with some degree of overlap. In this example, $n_0 = n_1 = 2000$.

Results. The results are presented on Fig. 9 and 10 corresponding respectively to thresholds $\delta_{max_1} = 0.1$ and $\delta_{max_2} = 0.24$. The discrepancy plot clearly shows regions of different intensities. The persistence diagram (built with $k = 6$) highlights four persistent minima plus the global one, calling for simplification. On the other hand, the critical values associated with the local minima of the elevation are quite scattered. Visually, three groups emerge, so that we investigate the clusters using two thresholds $\delta_{max_1} = 0.1 < \delta_{max_2} = 0.24$, yielding different compositions for the five clusters. Consider, e.g., the cluster 1 associated with δ_{max_1}. This cluster contains a high proportion of red points, whence a high contribution to the JSD. In moving from δ_{max_1} with δ_{max_2}, more points get discarded, the remaining points revealing the *core* of the clusters.

In any case, note that the regions do not necessarily coincide with original components of the mixture, i.e., they are a result of the comparison only.

6.3   Model: Mixture of Handwritten Digits

Specification. The goal is assess the ability of the method to spot local differences, and to cope with real life high dimensional data ($d = 784$).

This dataset is based on the MNIST dataset [15] that contains examples of handwritten digits. For our experiment, we used the digits 3,6 and 8 and we built our populations by sampling with replacement from these three populations. The following table summarizes the number of samples taken from each digit set for each population:

digit	blue	red
3	100	1000
6	500	500
8	1000	100

Results. The results are presented on Fig. 11. The triangle shape of MDS embedding clearly shows the regions corresponding to each digits and the JSD decomposition plot highlights two of them as expected. Then, the persistence diagram (built with $k = 30$) hints at one persistent local minima plus the global one, yielding two clusters corresponding to digits 3 and 8 as expected.
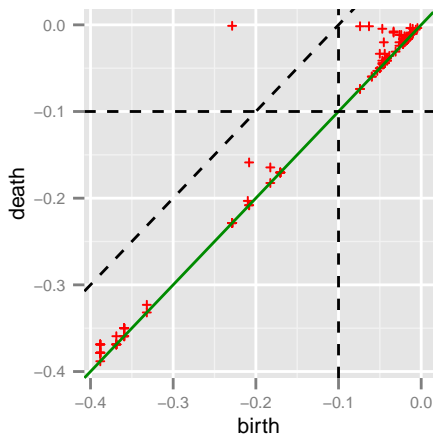
Figure 8 Model: Crenels. Figs, from top to bottom: Raw data embedding, Discrepancy plot, Persistence diagram, Clusters plot, Divergence decomposition
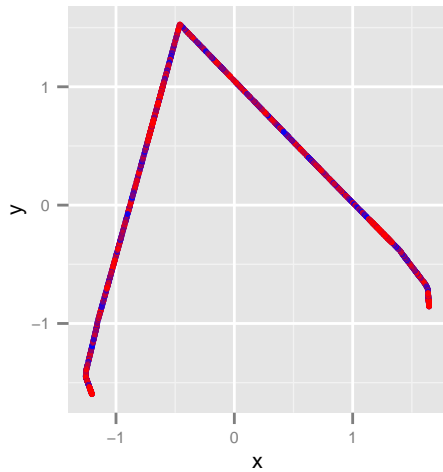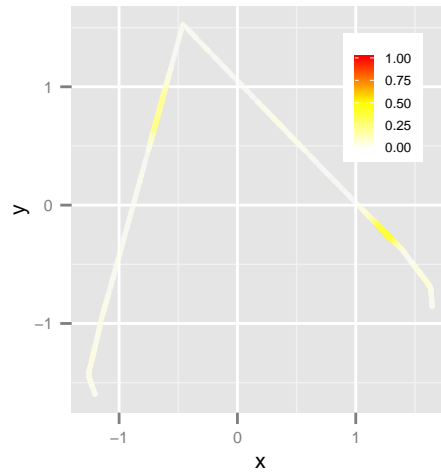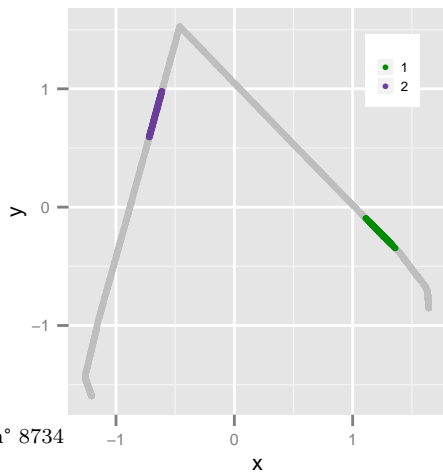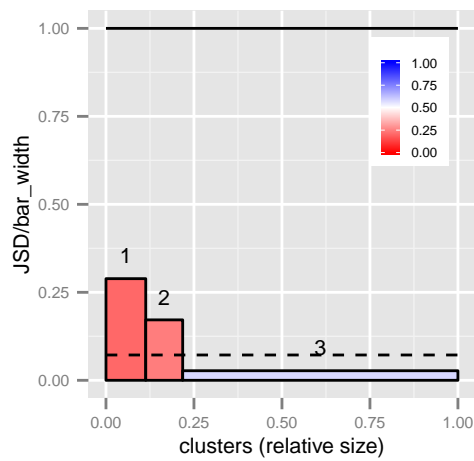

Raw data embedding
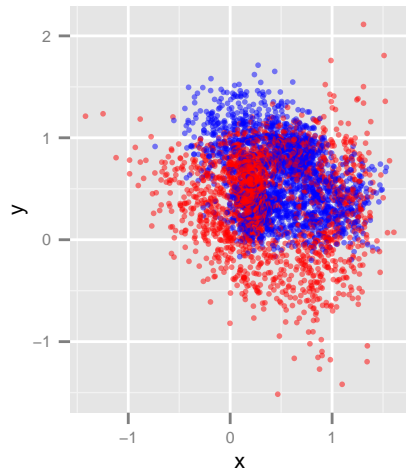

Discrepancy plot

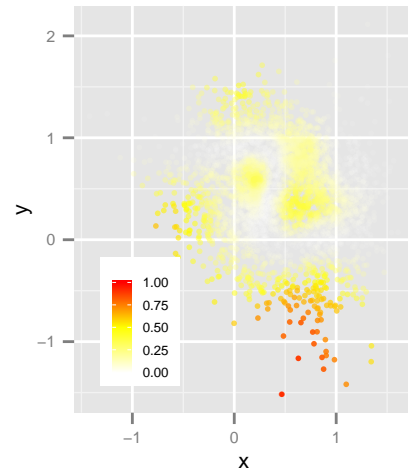Persistence diagram

Clusters

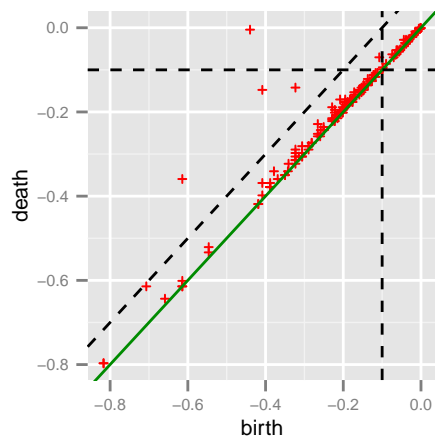Divergence decomposition plot

Figure 9 Model: Gaussian mixture. Figs, from top to bottom: Raw data embedding, Discrepancy plot, Persistence diagram, Clusters plot, Divergence decomposition
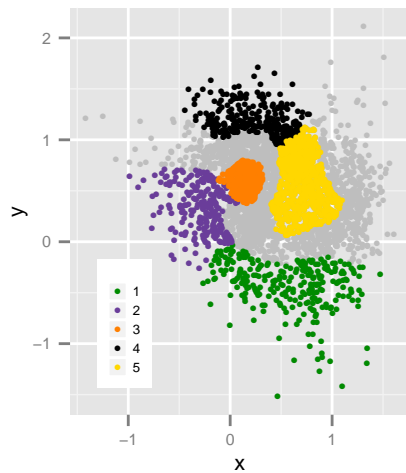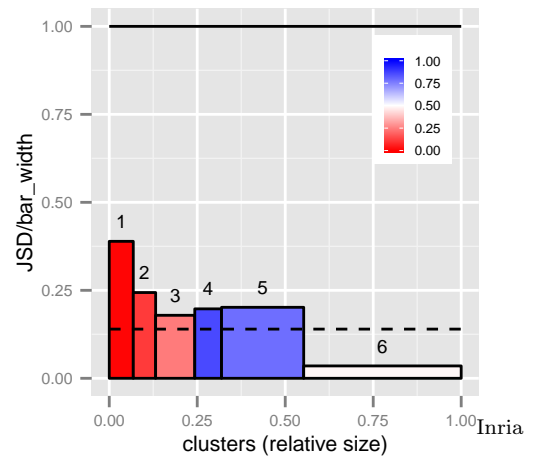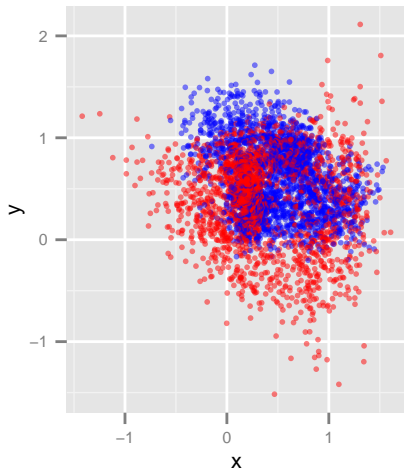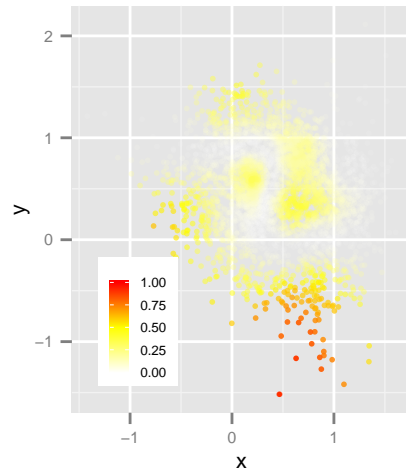

Raw data embedding

Discrepancy plot

Persistence diagram

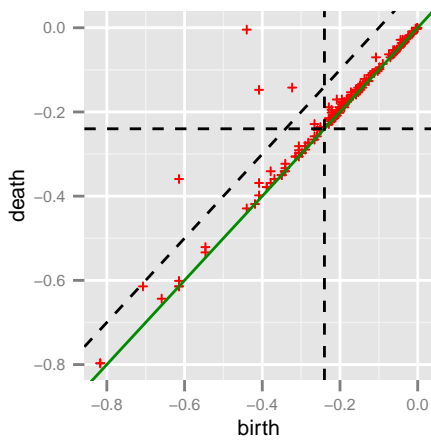Clusters

Divergence decomposition plot

Figure 10 Model: Gaussian mixture. Figs, from top to bottom: Raw data embedding, Discrepancy plot, Persistence diagram, Clusters plot, Divergence decomposition
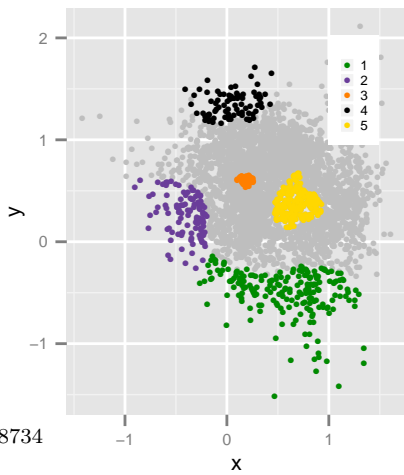

Raw data embedding

Discrepancy plot

Persistence diagram

Clusters

Divergence decomposition plot

Figure 11 Model: Handwritten digits. Figs, from top to bottom: Raw data embedding, Discrepancy plot, Persistence diagram, Clusters plot, Divergence decomposition

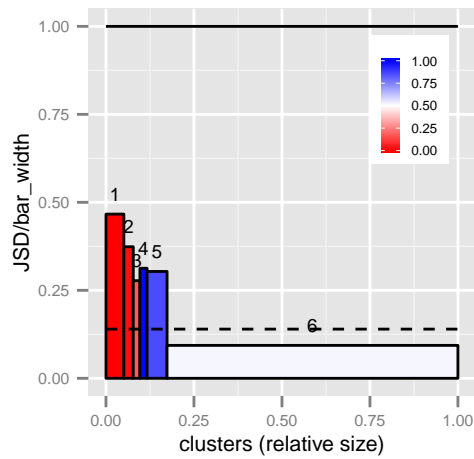Raw data embedding
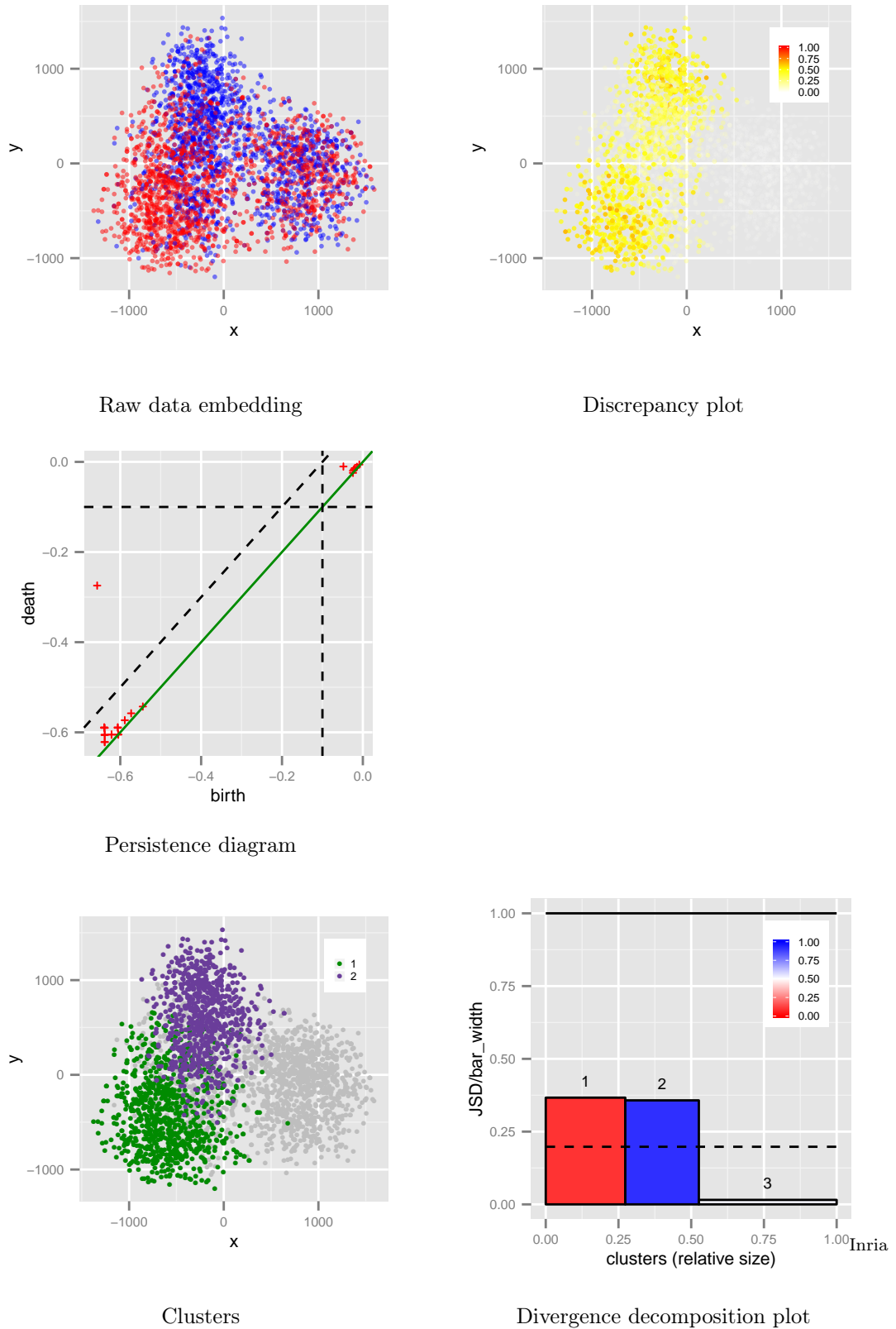

Discrepancy plot

Persistence diagram

Clusters

Divergence decomposition plot

Inria

7 Conclusion

This paper proposes the first method to model the difference between datasets given as point clouds, for which there is evidence showing that they do not have the same underlying distribution. The method relies on a pointwise estimation of an integrand related to the Jensen-Shannon divergence (JSD), a symmetric version of the Kullback-Leibler divergence. An estimate of the JSD is obtained for each sample using a non-parametric regression method relying on k_n nearest neighbors estimates. Topological persistence is then used to gather samples in groups associated with local maxima of the JSD. All in all, our method delivers groups of samples with *significant* contribution to the JSD, and associated with local maxima of the JSD.

On the theoretical side, several questions are of major interest. A first goal will be to characterize the clusters returned by our procedure, based upon assumptions on the distributions underlying the data. This problem is related to the robustness of a recent clustering method combining mode seeking and topological persistence [5], since under suitable conditions, persistent modes of the density and those defining clusters have been shown to match. Coming up with a similar line of argumentation is more challenging in our case since two densities are involved, and the magnitude of the JSD and those of these densities are independent quantities. A second goal will consist of generalizing the method to data associated with a (non Euclidean) metric space.

On the practical side, we believe that our method goes well beyond statistical analysis based on two-sample tests, which essentially summarizes the information contained in all coordinates into a single boolean value (accept or reject the null hypothesis). It should therefore prove of interest wherever two-sample tests are used.

References

[1] A. Banyaga and D. Hurtubise. *Lectures on Morse Homology*. Kluwer, 2004.

[2] P. Billingsley. *Convergence of probability measures (2nd Edition)*. John Wiley & Sons, 2013.

[3] P. Bubenik. Statistical topological data analysis using persistence landscapes. *J. of Machine Learning Research*, 16:77–102, 2015.

[4] F. Cazals and D. Cohen-Steiner. Reconstructing 3D compact sets. *Computational Geometry Theory and Applications*, 45(1-2):1–13, 2011.

[5] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. In *ACM SoCG*, pages 97–106. ACM, 2011.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

[7] H. Ding and J. Xu. FPTAS for minimizing earth mover's distance under rigid transformations. In *Algorithms–ESA 2013*, pages 397–408. Springer, 2013.

[8] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. AMS, 2010.

[9] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse complexes for piecewise linear 2-manifolds. In *ACM Symp. on Computational Geometry*, pages 70–79, 2001.

[10] R. Filipovych, S.M. Resnick, and C. Davatzikos. Semi-supervised cluster analysis of imaging data. *NeuroImage*, 54(3):2185–2197, 2011.

[11] J. Friedman. On multivariate goodness-of-fit and two-sample testing. *Proceedings of Phystat2003, http://www. slac. stanford. edu/econf C*, 30908, 2004.

[12] M. Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, and the R Core Team. *caret: Classification and Regression Training*, 2014. R package version 6.0-35.

[13] A. Gretton, K.M. Borgwardt, J.R. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[14] L. Györfi and A. Krzyzak. *A distribution-free theory of nonparametric regression*. Springer, 2002.

[15] Y. LeCun and C. Cortes. The mnist database of handwritten digits, 1998.

[16] J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[17] V. Melnykov, W-C. Chen, and R. Maitra. MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, 51(12):1–25, 2012.

[18] G. Pau, A. Oles, M. Smith, and O. Sklyar andW. Huber. *EBImage: Image processing toolbox for R*. R package version 4.4.0.

[19] F. Pérez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 1666–1670. Ieee, 2008.

[20] J. Silva and S. Narayanan. Universal consistency of data-driven partitions for divergence estimation. In *IEEE International Symposium on Information Theory*, pages 2021–2025. Citeseer, 2007.

[21] L. Song, C.H. Teo, and A.J. Smola. Relative novelty detection. In *International Conference on Artificial Intelligence and Statistics*, pages 536–543, 2009.

[22] Q. Wang, S.R. Kulkarni, and S. Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *Information Theory, IEEE Transactions on*, 51(9):3064–3074, 2005.

[23] Q. Wang, S.R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *Information Theory, IEEE Transactions on*, 55(5):2392–2405, 2009.
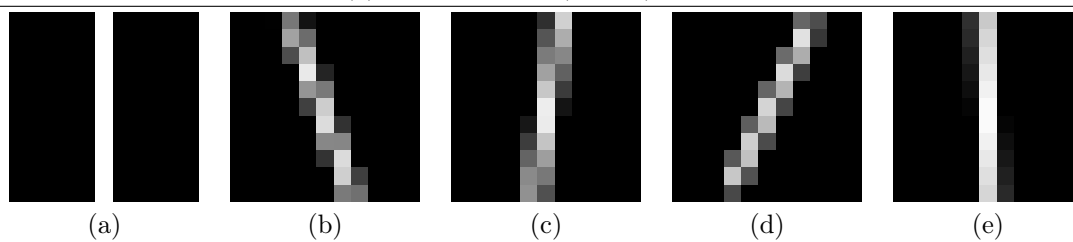
8 Supplemental: Data Sets

8.1 Crenels

Summary:

- $n_0 = n_1 = 2000$

- $d = 121$

- Rationale: A dataset to assess the ability of the method to spot local differences, and to cope with data of low intrinsic dimension in a high dimensional space.

Figure 12 Rotated images (a) Orignal image (b,c,d,e) Example rotated images

| (a) | (b) | (c) | (d) | (e) |

8.2 Gaussian mixture

Summary:

- $n_0 = n_1 = 2000$

- $d = 2$

- Rationale: A simple and easy to visualize dataset with regions of different intensities of divergence

The first model is a mixture of four spherical gaussians with equal probability, i.e.,

$$X = \sum_{i=1}^{4} 1_{\{i=U\}} N_0[i]$$

where U is a uniform discrete RV which takes values in $\{1..4\}$ and

$$N_0[i] \sim \mathcal{N}(\mu_0[i], \sigma_0[i])$$

The randomly generated parameters by MixSim R Package were:

$$\mu_0[1] = (0.9556715, 0.3617815)$$
$$\mu_0[2] = (0.6207539, 0.8498296)$$
$$\mu_0[3] = (0.3077166, 0.2886823)$$
$$\mu_0[4] = (0.1496965, 0.9773699)$$

$$\sigma_0[1] = 0.04771839 I_2, \ \sigma_0[2] = 0.02111844 I_2, \ \sigma_0[3] = 0.03342965 I_2, \ \sigma_0[4] = 0.07551961 I_2$$

where I_2 is the 2×2 identity matrix.

The second model is a mixture of four non-spherical gaussians with equal probability defined by the following randomly generated parameters (using analogous notation):

$$\mu_1[1] = (0.00677118, 0.07022882)$$
$$\mu_1[2] = (0.21864233, 0.46602229)$$
$$\mu_1[3] = (0.99020950, 0.20540745)$$
$$\mu_1[4] = (0.29765334, 0.80943535)$$

$$\sigma_1[1] = \begin{bmatrix} 0.1522406 & -0.1031709 \\ -0.1031709 & 0.1509098 \end{bmatrix} \sigma_1[2] = \begin{bmatrix} 0.006279164 & -0.001738228 \\ -0.001738228 & 0.032324880 \end{bmatrix}$$

$$\sigma_1[3] = \begin{bmatrix} 0.05161954 & 0.02275865 \\ 0.02275865 & 0.25600142 \end{bmatrix} \sigma_1[4] = \begin{bmatrix} 0.11359079 & 0.02248852 \\ 0.02248852 & 0.02819918 \end{bmatrix}$$

8.3　Mixture of Handwritten Digits

Summary:

- $n_0 = n_1 = 1600$

- $d = 784$

- Rationale: A dataset to assess the ability of the method to spot local differences, and to cope with real life high dimensional data.

Figure 13 Subsets of handwritten digits used. Images cropped from http://www.cs.nyu.edu/~roweis/data.html.