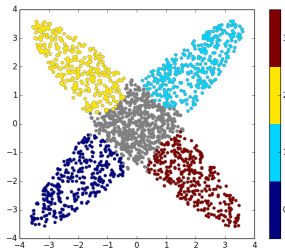
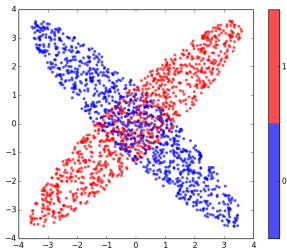


Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces

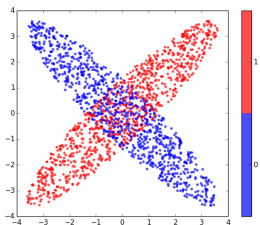
Frederic.Cazals@inria.fr, Alix.Lheritier@inria.fr
Inria Sophia Antipolis, Algorithms-Biology-Structure

- ▶ <http://team.inria.fr/abs>
- ▶ <http://sbl.inria.fr>

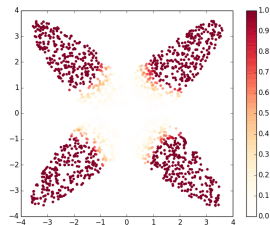


What do we provide?

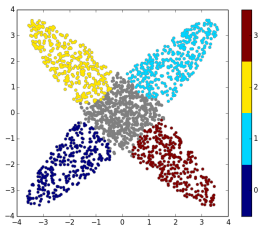
Given two point clouds,



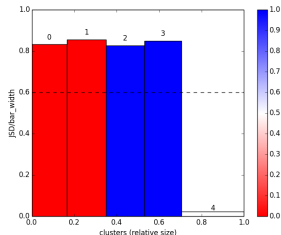
we localize the discrepancy,



to find spatially coherent regions
of high discrepancy,



and provide a cluster based
decomposed effect size.



Beyond two-sample tests

Goals

Step 1

Step 2

Step 3

Wrapping-up

More examples

Outlook

Data discrepancies: two-sample problem and effect size

▷ The two-sample test (TST) approach

- Two datasets $x^{(n_0)} \equiv \{x_1, \dots, x_{n_0}\}$ and $y^{(n_1)} \equiv \{y_1, \dots, y_{n_1}\}$ in \mathbb{R}^d as i.i.d. samples from two unknown densities f_X and f_Y
- Hypothesis testing: $H_0 : f_X = f_Y, H_1 : f_X$ and f_Y differ in some way.
→ accept/reject: summarizes difference in a single bit

▷ Effect size: “quantitative measure of the strength of a phenomenon”

- p -value gives magnitude of the statistical significance
but “Statistical significance = Effect size \times Sample size”
- The statistic of TST reflects the global discrepancy
and could be considered as a measure of the effect size

▷ Towards a nonparametric multivariate effect size:

- effect size must be standardized in some way in order to be comparable
- we seek to represent more general discrepancies,
in multidimensional spaces

Outline of our method: three steps

- ▶ **Step 1:** Estimate a measure of local discrepancy on each given point
- ▶ **Step 2:** Aggregate local discrepancy in a spatial coherent way, using topological persistence analysis to spot stable features, and produce clusters by removing low discrepancy points
- ▶ **Step 3:** Produce an effect size bar plot to summarize the discrepancy profile

Beyond two-sample tests

Goals

Step 1

Step 2

Step 3

Wrapping-up

More examples

Outlook

Pre-requisite: Jensen-Shannon divergence

▷ Kullback-Leibler divergence (KLD):

$$\begin{cases} D_{\text{KL}}(f\|g) \equiv \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx \\ D_{\text{KL}}(P\|Q) \equiv \sum_{I \in \mathcal{A}} P(I) \log \frac{P(I)}{Q(I)} \end{cases}$$

▷ The Jensen-Shannon divergence (JSD): symmetrizes and smoothes the KLD:

Consider $f \equiv (f_X + f_Y)/2$, then

$$JS(f_X\|f_Y) \equiv \frac{1}{2} (D_{\text{KL}}(f_X\|f) + D_{\text{KL}}(f_Y\|f))$$

▷ Main properties of JSD:

- JSD is symmetric
- JSD is bounded between 0 and 1
- Its square root yields a metric

▷Ref: Endres and Schindelin; IEEE Trans. Info. Theory, 2003

Step 1: Jensen-Shannon divergence and its decomposition

- ▷ **Notations:** two unknown densities f_X and f_Y , and the associated samples $x^{(n_0)}$ and $y^{(n_1)}$
- ▷ **Two random variables are implicitly defined:**
 - a position variable Z with density $f_Z \equiv f = (f_X + f_Y)/2$
 - a binary label $L \in \{0, 1\}$ with pmf $P(0) = 1/2$,
indicating from which density (f_X or f_Y) an instance of Z is obtained.

- ▷ **Equivalently, one defines the following pair of random variables:**

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

- ▷ **Associated conditional and unconditional probability mass functions:**

$$\begin{cases} P(l|z) = \mathbb{P}(L = l | Z = z) \\ P(l) = \mathbb{P}(L = l) = \frac{1}{2} \end{cases}$$

- ▷ **Lemma:** the JSD can be expressed as:

$$JS(f_X \| f_Y) = \int_{\mathbb{R}^d} f_Z(z) D_{\text{KL}}(P(\cdot|z) \| P(\cdot)) dz$$

Step 1: the local discrepancy

▷ From

$$JS(f_X \| f_Y) = \int_{\mathbb{R}^d} f_Z(z) D_{\text{KL}}(P(\cdot|z) \| P(\cdot)) dz$$

▷ We define the *discrepancy at location z* as

$$\delta(z) \equiv D_{\text{KL}}(P(\cdot|z) \| P(\cdot)).$$

▷ Remarks:

- $\delta(z) \in [0, 1]$ and $\delta(z) = 0 \Leftrightarrow f_X(z) = f_Y(z)$.
- $P(I)$ is known but $P(I|z)$ is not:
 - we need to estimate $P(I|z)$ at each given location z .

Step 1: random design nonparametric regression

- ▶ **Consider random variables:** location $Z \in \mathbb{R}^d$, and response variable $R \in \mathbb{R}$
- ▶ **Associated regression function:**

$$m(z) \equiv \mathbb{E}[R|Z = z].$$

- ▶ **Consider data:** $\{(Z_i, R_i)\}_{i=1, \dots, n}$
- ▶ **k_n -nearest neighbors regressor:** upon sorting samples by increasing distance to z :

$$m_n(z) = \frac{1}{k_n} \sum_{i=1, \dots, k_n} R_{(i,n)}(z)$$

- ▶ **NB:** $m_n(z)$ is a random variables: some convergence assessment is in order.
- ▶ **Ref:** L. Györfi and A. Krzyżak; A distribution-free theory of nonparametric regression; 2002

Step 1: estimation via k -nearest neighbors

- ▶ Using the labels as response variable $R \equiv L$
- ▶ Estimate $P(\cdot|z)$ via random design nonparametric regression:
 - build an estimator $m_n(z)$ using n i.i.d. realizations of (L, Z) for:

$$m(z) = \mathbb{E}[L|Z = z] = P(1|z).$$

- Then, if $0 \leq m_n(z) \leq 1$, we can use the following estimator for $P(l|z)$:

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

- ▶ Thm: Using a k_n -nearest neighbors regressor, s.t. $\frac{k_n}{\log n} \rightarrow \infty$ and $\frac{k_n}{n} \rightarrow 0$:

$$\hat{\delta}_n(z) \equiv D_{\text{KL}}\left(\hat{P}_n(\cdot|z) \parallel P(\cdot)\right) \xrightarrow{n \rightarrow \infty} \delta(z) \text{ a.s.}$$

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

The random multiplexer to obtain i.i.d. realizations of (L, Z)

- ▶ A random sampler produces i.i.d. realizations of (Z, L) from $x^{(n_0)}$ and $y^{(n_1)}$:

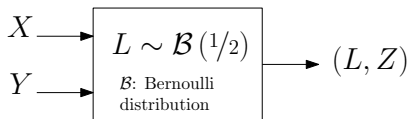


Figure: Random multiplexer generating pairs (label, position).

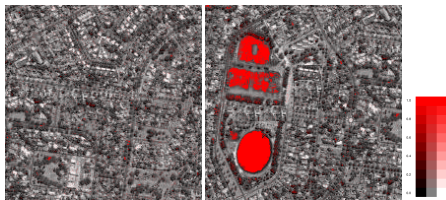
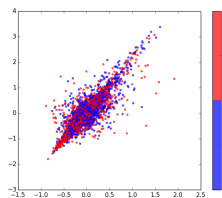
- ▶ The case of populations of uneven sizes:
 - the multiplexer will consume faster the *small* population, and halt
 - unused samples of the large population: detrimental since information loss
 - resample B times and take the median of estimates, on a per sample basis

Step 1: Illustration: statistical image comparison

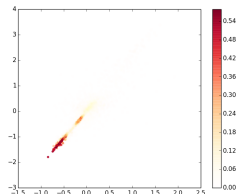
- ▷ **Images:** taking 2×2 blocks in each color channel (R,G,B) yields points in \mathbb{R}^{12} .
- ▷ **Interpolate** gray scale pixel color with red scale representing discrepancy at each pixel (upper left corner of the corresponding block) estimated with $k_n = n^{1/3}$

- ▷ **Multidimensional Scaling of parameter space:**

The two populations. . .



. . . colored with $\hat{\delta}$:



Beyond two-sample tests

Goals

Step 1

Step 2

Step 3

Wrapping-up

More examples

Outlook

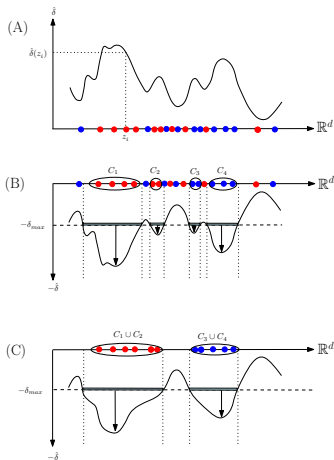
Step 2: Building the clusters from sublevel sets of $-\hat{\delta}(z)$

▷ Ingredients:

- ▶ Height function / landscape: estimated discrepancy $\hat{\delta}(z)$
- ▶ Parameter: significance threshold δ_{max}

▷ Construction:

- ▶ Idea: one cluster \sim one connected component of the sublevel set of $-\hat{\delta}(z)$ defined by δ_{max}
- ▶ Extra ingredient: smoothing the landscape to get rid of small clusters : smoothing using topological persistence at threshold ρ

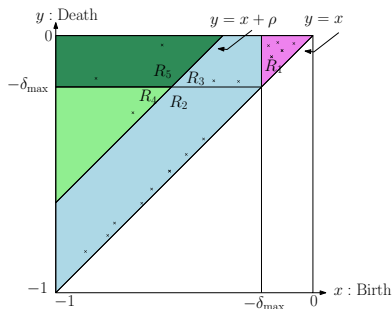


▷ **NB:** spurious samples removed from clusters due to filtering wrt δ_{max} .

Step 2: Building the clusters: persistence diagram

▷ Partition of the PD induced by:

- ▶ Significance threshold δ_{max}
- ▶ Persistence threshold ρ



▷ Local minimum m of $-\hat{\delta}(z)$:

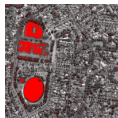
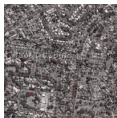
- ▶ Selected/rejected: m was born before $-\delta_{max}$.
- ▶ Persistent/canceled: $\text{persistence}(m) \geq \rho$
- ▶ Filtered (un-filtered): the catchment basin of m dies after (before) $-\delta_{max}$.

▷ Observation:

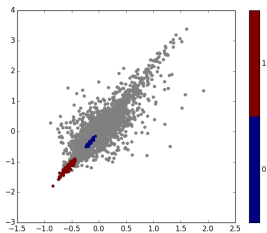
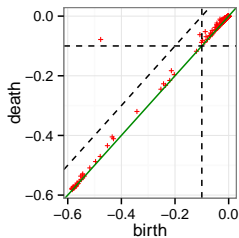
- ▶ # clusters : $1 + \#$ points in region R_5 of the PD.
- ▶ # persistent local minima : $1 + \text{num points in the region } R_4 \cup R_5$ of the PD.

Step 2: Illustration: statistical image comparison

▷ Images again:



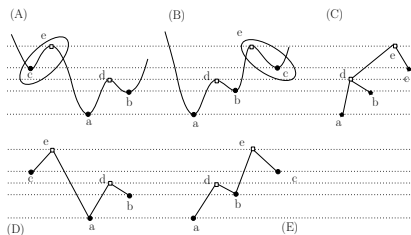
▷ Parameters: $k = 10$ (NNG), $\rho = 0.1$, $\delta_{max} = 0.1$



Landscape Simplification:

Union-find versus recursive simplification of the Morse-Smale-Witten complex

- ▶ Clustering: one versus many:
 - Work with critical points (instead of all samples)
 - Pre-process the c.p. to redo analysis at various δ_{max} threshold



- ▶ Ref: Chazal et al; Tomato; ACM SoCG 2011
- ▶ Ref: Banyaga, Hurtubise; Lectures on Morse Homology; 2004
- ▶ Ref: Cazals, Cohen-Steiner; CGTA, 2011

Beyond two-sample tests

Goals

Step 1

Step 2

Step 3

Wrapping-up

More examples

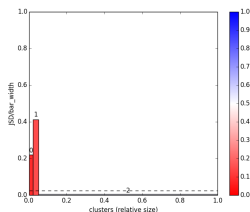
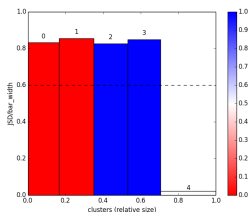
Outlook

Step 3: Effect size: discrepancy profile

- ▷ **Global estimated JSD**: area under dashed line
- ▷ **Maximum JSD**: area under continuous line (=1)
- ▷ **Contribution of each cluster C to JSD**: area of bar

$$JSC(f_X \| f_Y) \equiv \frac{1}{n_0 + n_1} \sum_{z \in (x^{(n_0)} \cup y^{(n_1)}) \cap C} \hat{\delta}(z).$$

- ▷ **Mass of each cluster**: bar width
 - ▷ **Population balance in each cluster**: bar color
-
- ▷ **Ellipses**:
 - Large global JSD (dashed line)
 - Contributed by **2+2** balanced clusters
 - ▷ **Images**:
 - Smaller global JSD (dashed line)
 - Contributed by **2** clusters



Beyond two-sample tests

Goals

Step 1

Step 2

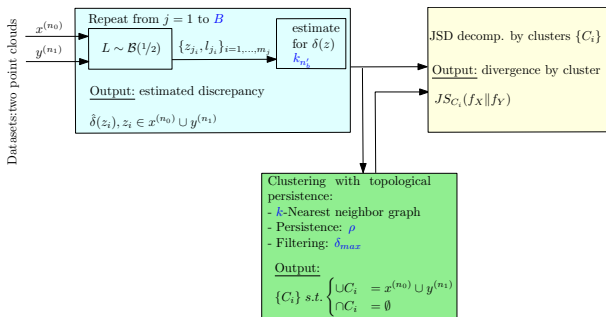
Step 3

Wrapping-up

More examples

Outlook

Wrapping-up: workflow



▷ **Compulsory parameters:**

k_n : regression parameter

δ_{max} : discrepancy significance threshold

ρ : persistence threshold

k : num. of nearest neighbors for the persistence based clustering

▷ **Optional parameter:**

B : num. repetition in case of unbalanced populations

Try me: <http://sbl.inria.fr>



Structural Bioinformatics Library

Template C++ / Python API for developing structural bioinformatics applications.

Home

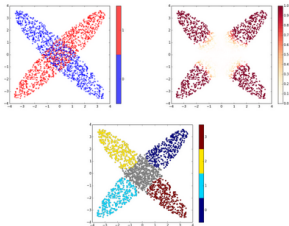
Packages

Classes

Search

Structural Bioinformatics

User Manual



Density_difference_based_clustering

Authors: A. Lheritier and F. Cazals

1. Goals

Comparing two sets of multivariate samples is a central problem in data analysis. From a statistical standpoint, one 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, denoted H_0).

Table of Contents

- ↓ Goals
- ↓ Using the programs
 - ↓ Pre-requisites
 - ↓ Step 1
 - ↓ Step 2
 - ↓ Step 3
 - ↓ Plots
- ↓ Input: Specifications and File Types
 - ↓ Step 1: using `sbl-ddbc-step-1-discrepancy.py` and `sbl-ddbc-disc-based-colored-embedding.py`
 - ↓ Step 2: using `sbl-ddbc-step-2-clustering.py`
 - ↓ Step 3: using `sbl-ddbc-step-3-cluster-plots.py`
- ↓ Output: Specifications and File Types
- ↓ Examples
 - ↓ Mixture of Gaussians
 - ↓ Higher dimensional case
- ↓ Algorithms and Methods
 - ↓ Discrepancy estimation
 - ↓ Persistence analysis
- ↓ Programmer's Workflow
- ↓ External dependencies

Beyond two-sample tests

Goals

Step 1

Step 2

Step 3

Wrapping-up

More examples

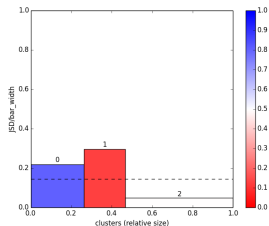
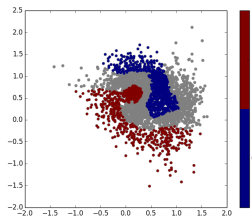
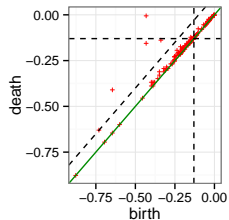
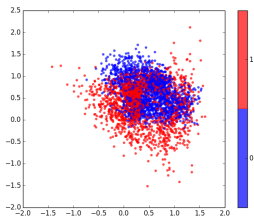
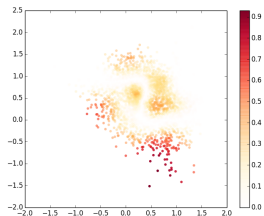
Outlook

Gaussian mixture: specification

- ▷ **Goal:** ability to spot regions of different intensity of discrepancy.
- ▷ **Data:**
 - distributions for X and Y : two mixtures of four 2D Gaussians
 - $n_0 = n_1 = 2000$.

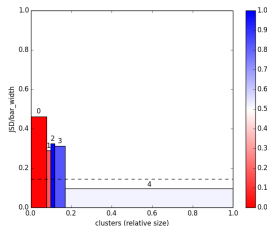
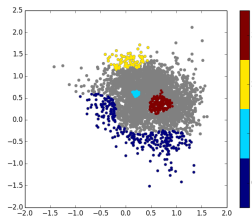
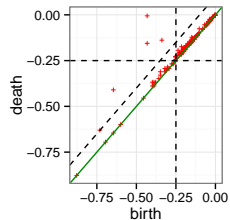
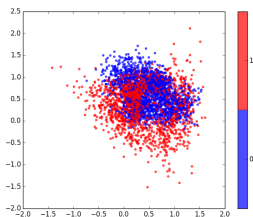
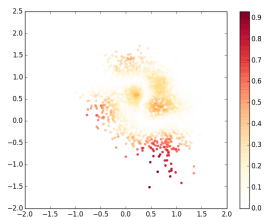
Gaussian mixture: results I

- $k = 6$ (NNG)
- $\delta_{max} = 0.13$ yields two large clusters



Gaussian mixture: results II

- $k = 6$ (NNG)
- $\delta_{max} = 0.25$ yields four small clusters



Crenels: specification

▷ **Goal:** coping with data of low intrinsic dimension (in fact: 1 in $d=121$)

▷ **Data:**

– Points:

- consider the pixels (0/1) of a $m \times m$ grayscale image T
- rotating the image I yields a point cloud (1 point per image)
- $m = 11$ yields $d = 121$; but intrinsic dimension is one

– Populations:

red points: from RV $X = rotate(I, A_X)$ with uniform RV $A_X \sim \mathcal{U}(s, t)$,

blue points: from RV $Y = rotate(I, A_Y)$, with:

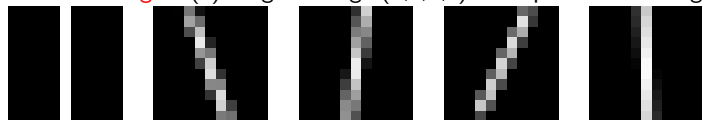
consider two Bernoulli RV $B_1 \sim \mathcal{B}(p_1)$ and $B_2 \sim \mathcal{B}(p_2)$,

two uniform RV $U_1 \sim \mathcal{U}(a, b)$ and $U_2 \sim \mathcal{U}(c, d)$.

Define: $A_Y = B_1(B_2 U_1 + (1 - B_2) U_2) + (1 - B_1) A_X$.

$n_0 = 2000, n_1 = 2000$

▷ **Rotated images:** (a) Original image (b,c,d,e) Example rotated images



(a)

(b)

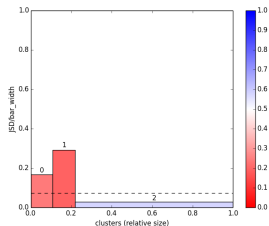
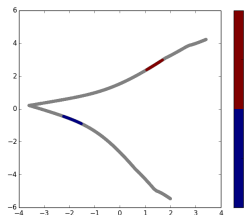
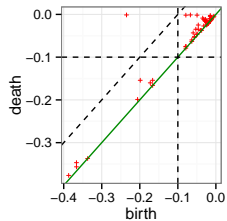
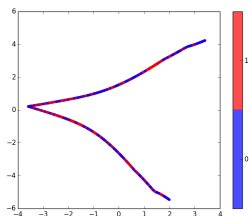
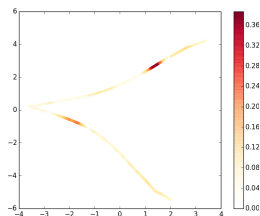
(c)

(d)

(e)

Crenels: results

- $k = 30$ (NNG),
- $\delta_{max} = 0.1$



Handwritten digits: specification

▷ Handwritten digits:

- One digit: 28×28 grayscale image: $d = 784$
- Populations: $n_0 = n_1 = 1600$

▷ More specifically: two mixtures of 3s, 6s and 8s

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

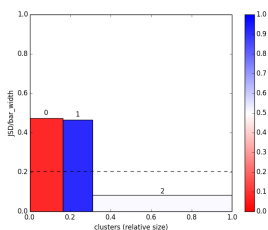
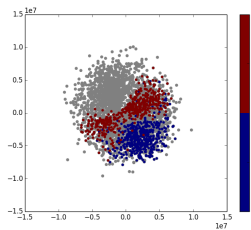
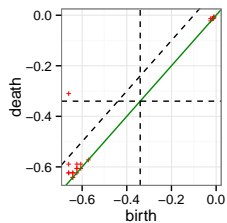
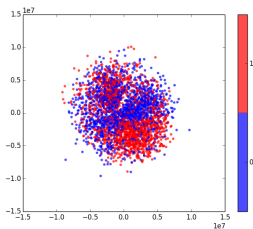
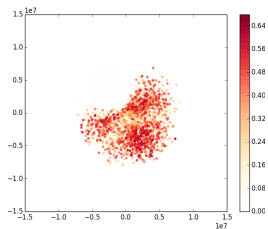
▷ Images from: <http://www.cs.nyu.edu/~roweis/data.html>:



▷Ref: LeCun and Cortes; The MNIST database of handwritten digits, 1998

Handwritten: results

- $k = 30$ (NNG),
- $\delta_{max} = 0.35$



Beyond two-sample tests

Goals

Step 1

Step 2

Step 3

Wrapping-up

More examples

Outlook

Outlook: about regression

- ▶ k-NN based regressors: adapt to local intrinsic dimension: convergence results proved (L_2 sense) for marginals μ which are doubling measures.
- ▶ random projection tree based regressors: convergence results proved (L_2 sense) when \mathcal{X} has Assouad dimension d . NB: more efficient than k-NN since cells of RPT have constant size.
- ▶ Open problem (AFAIK): strong pointwise consistency using RPTrees.

▷Ref: Kpotufe; k-NN regression adapts to local intrinsic dimension; NIPS 2011

▷Ref: Kpotufe and Dasgupta; A tree-based regressor that adapts to intrinsic dimension; J. of Computer and System Sciences, 2012

Outlook: general

- ▶ About p-values:
 - ▶ Use a classical test, possibly Maximum Mean Discrepancy (Gretton et al).
 - ▶ Also: the k-NN estimator used in a sequential way can be used to compute a p-value in a flexible way—the number of samples to process need not be known in advance.
- ▶ More applications:
 - ▶ Finding clusters with low discrepancy: study $\hat{\delta}$.
 - ▶ Goodness-of-fit analysis: sampling from a given model, then comparing data to spot discrepancies
- ▶ Feedback versus feature based selection: Compare to NIPS 2015 paper *Principal differences analysis: feature based identification in the context of TST*