Statistical analysis: complementary topics

January 19, 2021

Frederic.Cazals@inria.fr

Statistical analysis: complementary topics

Beyond Two-Sample-Tests

- Problem
- Jensen-Shannon divergence and discrepancy
- Density based clustering

Comparing clusterings

- Motivation
- Problem statement
- Previous work
- D-family matching: problem
- Hardness
- Algorithms
- On the choice of D
- Experiments

Maximum Information Coefficient

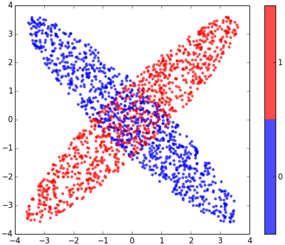
Statistical analysis: complementary topics

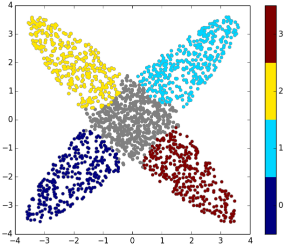
- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of *D*
  - Experiments
- Maximum Information Coefficient

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





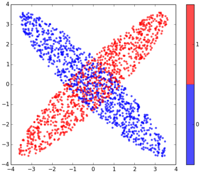
Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of D
  - Experiments
- Maximum Information Coefficient

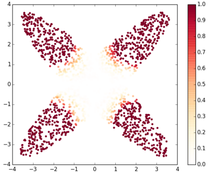
Localizing data discrepancies

- ▷ Problem: two populations differ in parameter/feature space: where are the differences?
- ▷ Contribution: density difference clustering based method

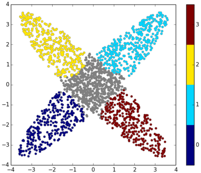
Given two point clouds,



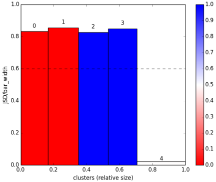
we localize the discrepancy,



to find spatially coherent regions of high discrepancy,

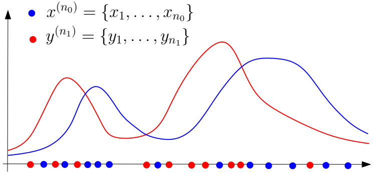


and provide a cluster based decomposed effect size.



Data discrepancies: two-sample problem and effect size

- The two-sample test (TST) approach
  - Two datasets i.i.d. samples from two unknown densities f_X and f_Y: $x^{(n_0)} \equiv \{x_1, \ldots, x_{n_0}\}$ and $y^{(n_1)} \equiv \{y_1, \ldots, y_{n_1}\}$ in $\mathbb{R}^d$



$$\begin{cases} \mathrm{H}_0 : f_X = f_Y, \\ \mathrm{H}_1 : f_X \text{ and } f_Y \text{ differ in some way} \end{cases} \tag{1}$$

- Classical TST
  - p-value gives magnitude of the statistical significance, but
  - (i) accept/reject: summarizes difference in a single bit
  - (ii) the statistic of TST reflects the global discrepancy / effect size
- Goal: towards a nonparametric multivariate effect size
  - (i') localize discrepancies accounting for the differences
  - (ii') provide standardized (normalized) effect size

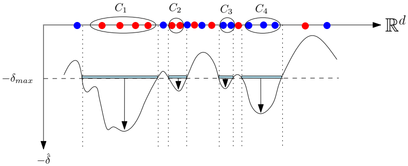
The three steps of the method

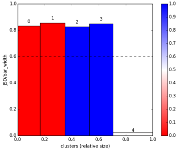
▷ **Step 1:** Estimate a measure of local discrepancy on each given point
Using $f \equiv {}^{(f_X+f_Y)}/_2$, define the Jensen-Shannon divergence:

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

▷ **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





Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of *D*
  - Experiments
- Maximum Information Coefficient

Pre-requisite: Jensen-Shannon divergence

▷ Kullback-Leibler divergence (KLD):

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

▷ The Jensen-Shannon divergence (JSD): symmetrizes and smoothes the KLD:
Consider $f \equiv {}^{(f_X + f_Y)}/_{2}$, then

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

▷ 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:

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

▷ Lemma: the JSD can be expressed as:

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

Step 1: the local discrepancy

- From

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

- We define the *discrepancy* at location z as

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

- 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:

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

▷ Consider data: $\{(Z_i, R_i)\}_{i=1,\ldots,n}$

▷ k_n-nearest neighbors regressor: upon sorting samples by increasing distance to z:

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

▷ NB: $m_n(z)$ is a random variables: some convergence assessment is in order.

▷Ref: L. Györfi and A. Krzyzak; A distribution-free theory of nonparametric regression; 2002

Step 1: estimation via k-nearest neighbors

- ▷ Using the labels as reponse 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:

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

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

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

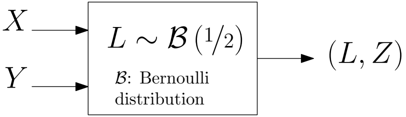
- ▷ Thm: Using a k_n-nearest neighbors regressor, s.t. $\frac{k_n}{\log n} \to \infty$ and $\frac{k_n}{n} \to 0$:

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

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 $y^{(n_1)}$:



X

Y

$L \sim \mathcal{B}\left(1/2\right)$

$\mathcal{B}$: Bernoulli distribution

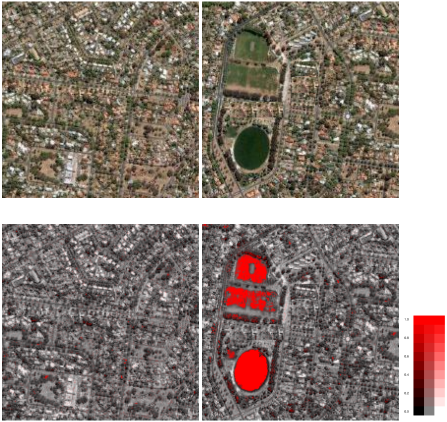
(L, Z)

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

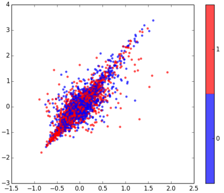
Step 1: Illustration: statistical image comparison

- ▷ 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 $k_n = n^{1/3}$

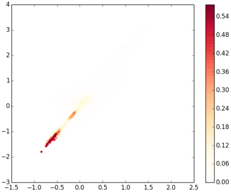


▷ Multidimensional Scaling of parameter space:

The two populations. . .



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

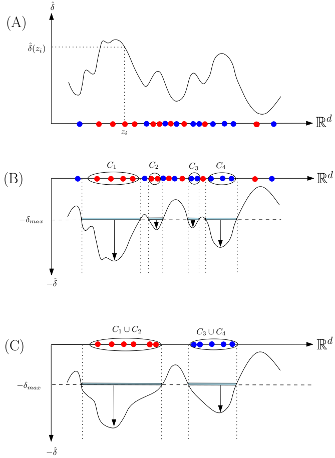


Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of D
  - Experiments
- Maximum Information Coefficient

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

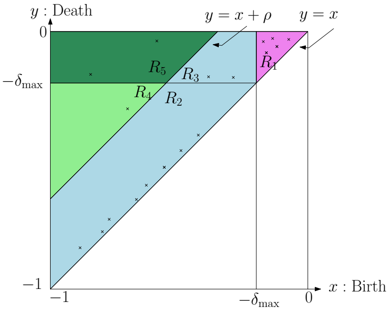
- ▷ Ingredients:
  - ▶ Height function / landscape: estimated discrepancy $\bar{\delta}_z(z)$
  - ▶ Parameter: significance threshold δ_{max}
- ▷ Construction:
  - ▶ Idea: one cluster $\sim$ one connected component of the sublevel set of $-\bar{\delta}_z(z)$ defined by δ_{max}
  - ▶ Extra ingredient: smoothing the landscape to get rid of small clusters : smoothing using topological persistence at threshold ρ



▷ NB: spurions 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

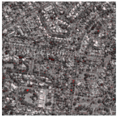


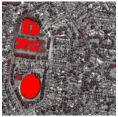
- Local minimum m of $-\bar{\delta}_z(z)$:
  - Selected/rejected: m was born before $-\delta_{max}$.
  - Persistent/canceled: persistence(m) $\geq \rho$
  - Filtered (un-filtered): the catchment basin of m dies after (before) $-\delta_{max}$.

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

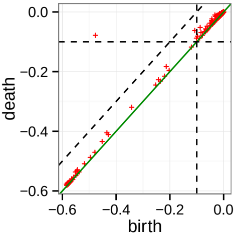
Step 2: Illustration: statistical image comparison

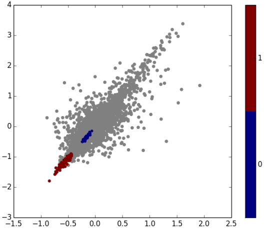
▷ Images again:





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





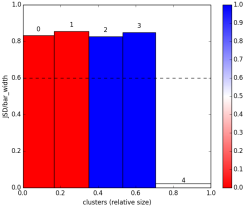
Step 3: Effect size: discrepancy profile

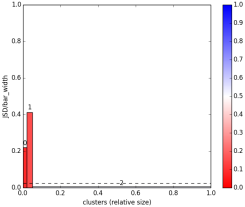
- Global estimated JSD: area under dashed line
- Maximum JSD: area under continuous line (=1)
- Contribution of each cluster C to JSD: area of bar

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

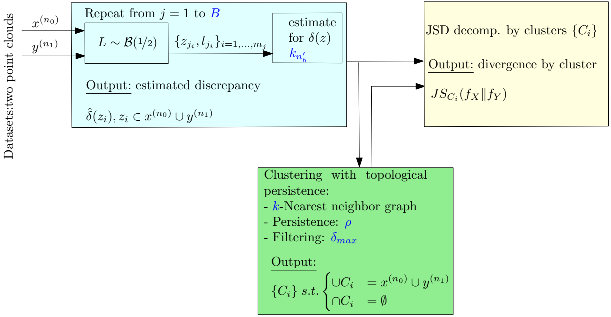
- Mass of each cluster: bar width
- Population balance in each cluster: bar color

- Ellipses:
  - Large global JSD (dashed line)
  - Contributed by **2**+**2** balanced clusters
- Images:
  - Smaller global JSD (dashed line)
  - Contributed by **2** clusters





Wrapping-up: workflow



Datasets:two point clouds

$x^{(n_0)}$

$y^{(n_1)}$

Repeat from $j = 1$ to B

$L \sim \mathcal{B}(1/2)$

$\{z_{j_i}, l_{j_i}\}_{i=1,\ldots,m_j}$

estimate for $\delta(z)$ $k_{n'_b}$

Output: estimated discrepancy

$\hat{\delta}(z_i), z_i \in x^{(n_0)} \cup y^{(n_1)}$

JSD decomp. by clusters $\{C_i\}$

Output: divergence by cluster

$JS_{C_i}(f_X \| f_Y)$

Clustering with topological persistence:
- k-Nearest neighbor graph
- Persistence: ρ
- Filtering: δ_{max}

Output:

$$\{C_i\}\ s.t. \begin{cases} \cup C_i & = x^{(n_0)} \cup y^{(n_1)} \\ \cap C_i & = \emptyset \end{cases}$$

- 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

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.

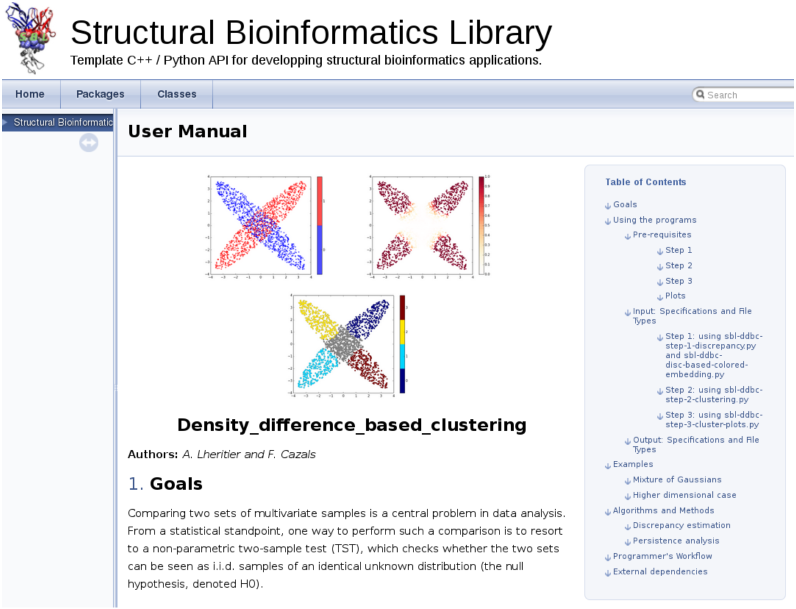
▷Ref: Kpotufe; k-NN regression adapts to local intrinsic dimension; NIPS 2011

▷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

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



Structural Bioinformatics Library

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

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 H0).

Table of Contents

Consistency of sequence regression estimates $\{m_n\}$ Based on L_2 norm

▷ Consider the following RV–induced by the data D_n:

$$\int \mid m_n(x) - m(x) \mid^2 \mu(dx). \tag{2}$$

▷ Def: The sequence $\{m_n\}$ is **weakly consistent** for a certain distribution of (X, Y) if

$$\lim_{n\to\infty} \mathrm{E}\Big[\int (m_n(x) - m(x))^2 \mu(dx)\Big] = 0. \tag{3}$$

▷ Def: The sequence $\{m_n\}$ is **strongly consistent** for a certain distribution of (X, Y) if

$$\lim_{n\to\infty} \int (m_n(x) - m(x))^2 \mu(dx) = 0 \text{ with proba. one.} \tag{4}$$

▷ Def: The sequence $\{m_n\}$ is **weakly universally consistent** if it is weakly consistent for all distributions of (X, Y) with $\mathbb{E}\left[Y^2\right] < \infty$.
▷ Def: The sequence $\{m_n\}$ is **strongly universally consistent** if it is strongly consistent for all distributions of (X, Y) with $\mathbb{E}\left[Y^2\right] < \infty$.

▷`Ref: book`

Consistency of sequence regression estimates $\{m_n\}$ based on pointwise convergence

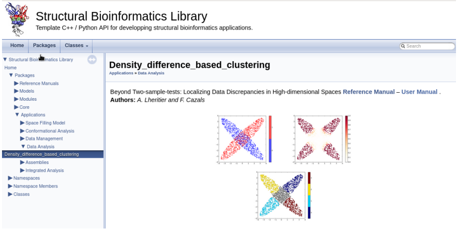
▷ Def: The sequence $\{m_n(x)\}$ is called **strongly pointwise consistent** is $m_n(x) \to m(x)$ a.s.

▷ Def: The sequence is called **strongly universal pointwise consistent** if it is strongly pointwise consistent for all distributions of (X, Y) with $\mathbb{E}\left[Y^2\right] < \infty$.

▷Ref: book

Structural Bioinformatics Library

Package Density Difference Based Clustering @ `http://sbl.inria.fr`



Structural Bioinformatics Library

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

Density_difference_based_clustering

Applications » Data Analysis

Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces Reference Manual – User Manual .
Authors: *A. Lheritier and F. Cazals*

- User manual `https://sbl.inria.fr/doc/Density_difference_based_clustering-user-manual.html`
- General entry: `http://sbl.inria.fr`

▷Ref: Cazals and A. Lhéritier, IEEE/ACM DSAA, 2015

▷Ref: Kim, Lee, Lei, Electronic Journal of Statistic, 2019

Statistical analysis: complementary topics

Beyond Two-Sample-Tests

- Problem
- Jensen-Shannon divergence and discrepancy
- Density based clustering

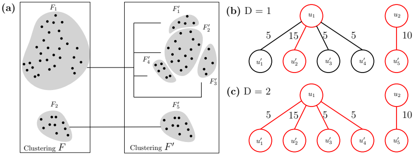
Comparing clusterings

- Motivation
- Problem statement
- Previous work
- D-family matching: problem
- Hardness
- Algorithms
- On the choice of D
- Experiments

Maximum Information Coefficient

Comparing two clusterings using matchings between clusters of clusters

F. Cazals, D. Mazauric, R. Tetley, and R. Watrigant
ACM Trans. Exp. Algorithms, 2019
`https://sbl.inria.fr/doc/D_family_matching-user-manual.html`



Statistical analysis: complementary topics

Beyond Two-Sample-Tests

- Problem
- Jensen-Shannon divergence and discrepancy
- Density based clustering

Comparing clusterings

- Motivation
- Problem statement
- Previous work
- D-family matching: problem
- Hardness
- Algorithms
- On the choice of D
- Experiments

Maximum Information Coefficient

Clustering algorithms

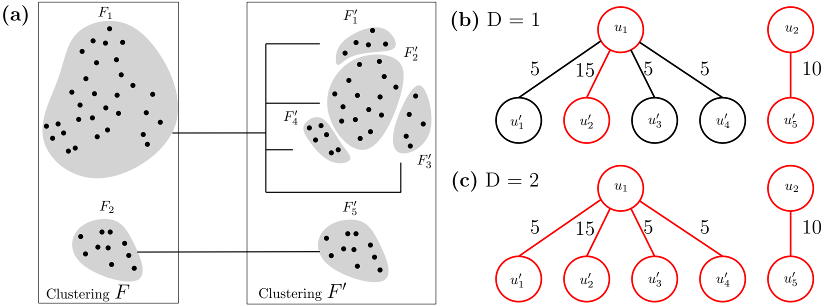
- Many algorithms: which one?
- Many parameters: which ones?
- Many clustering: are they consistent? A problem of scale...

Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of D
  - Experiments
- Maximum Information Coefficient

Grouping clusters into **metaclusters:** problem formalization in terms of intersection graph

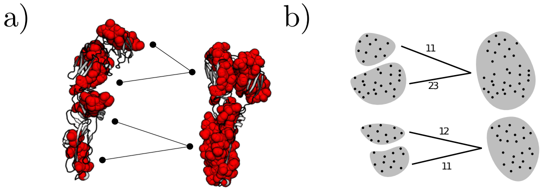
- Goal: recovering some coherence between groups of clusters
  - as a function of a scale parameter *D*



- Rationale: many-to-many
  - Aggregating many clusters, map to many clusters
  - Characterize the *scale* at which clusters merge

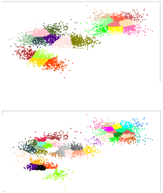
Structurally conserved motifs in protein structures
Many-to-many correspondence between clusters

- Handling small and conserved structural motifs in proteins

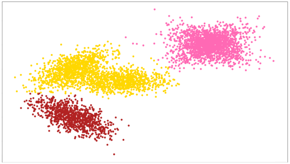


Merging clusters: a matter of scale

On the role of the scale parameter D



(A) Two clusterings (kmeans++, Tomato, etc)



(B) Meta-clusters as union of clusters

Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of D
  - Experiments
- Maximum Information Coefficient

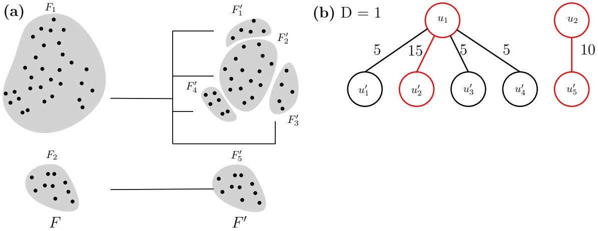
Comparing clusterings: previous work

REVISE REVISE REVISE REVISE

▷ 1-1 mapping of clusters: equivalent to the problem of computing a maximum weighted matching in weighted bipartite graph.

▷ Solution: solved in $O(n^2 \log n + nm)$

▷ Particular case of the D-family-matching problem for $D = 1$ – see later



Comparing clusterings: the Variation of Information

- A set Z of t items
- A clustering F of size r for Z: $F = \{F_1, \ldots, F_r\}$; $n_k = |F_k|$;

$p_k = n_k/t$.

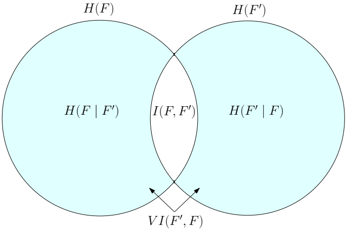
- A clustering F of size r' for Z: $F = \{F_1, \ldots, F_r\}$; $n'_k = |F'_{k'}|$;
- Overlap between two clusters: $p(k, k') = |F_k \cap F'_{k'}|/t$.
- Entropy of clustering: $H(F) = -\sum_{k=1,\ldots,r} p(k) \ln p(k)$

- Mutual information between F and F':

$$I(F, F') = \sum_{k} \sum_{k'} p(k, k') \ln \frac{p(k, k')}{p(k)p(k')}.$$

- Variation of information (VI):

$$VI(F, F') = H(F) + H(F') - 2I(F, F').$$



- Main properties:
  - VI is a metric
  - $VI(F, F') \leq \ln t$

▷`Ref: M. Meila, Journal of Multivariate Analysis, 2007`

Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
    - Problem
    - Jensen-Shannon divergence and discrepancy
    - Density based clustering
- Comparing clusterings
    - Motivation
    - Problem statement
    - Previous work
    - D-family matching: problem
    - Hardness
    - Algorithms
    - On the choice of D
    - Experiments
- Maximum Information Coefficient

Intersection graph

▷ Notations:

- Data: $Z = \{z_1, \dots, z_t\}$
- Clustering F of size r: $F = \{F_1, \dots, F_r\}$

$$F_i \subseteq Z, F_i \neq \emptyset \text{ and } F_i \cap F_j = \emptyset \text{ for every } i, j \in \{1, \dots, r\}, i \neq j.$$

- Clustering F' of size r': $F' = \{F'_1, \dots, F'_{r'}\}$

$$F'_i \subseteq Z, F'_i \neq \emptyset, \text{ and } F'_i \cap F'_j = \emptyset \text{ for every } i, j \in \{1, \dots, r'\}, i \neq j.$$

NB: a clustering may not contain all t items

Definition 1 (Intersection graph $G = (U, U', E, w)$ for F and F').

The set $U = \{u_1, \dots, u_r\}$: vertices of F
The set $U' = \{u'_1, \dots, u'_{r'}\}$: vertices of F'
Edges $E = \{\{u_i, u'_j\} \mid F_i \cap F'_j \neq \emptyset, 1 \leq i \leq r, 1 \leq j \leq r'\}$.
Edge weight of edge $e = \{u_i, u'_j\} \in E$ is $w_e = |F_i \cap F'_j|$.

D-family matching

▷ Let $D \in \mathbb{N}^+$: a constraint on the diameter of certain subgraph of the intersection graph

Definition 2. [D-family-matching for an intersection graph]
A family $\mathcal{S} = \{S_1, \ldots, S_k\}$, $k \geq 1$, such that

- for every $i, j \in \{1, \ldots, k\}$, if $i \neq j$, then: $S_i \subseteq V$, $S_i \neq \emptyset$, $S_i \cap S_j = \emptyset$,
- and the graph $G[S_i]$ induced by the set of nodes S_i has diameter at most D.

▷ Comments:

- $D = 1$: matching
- $D = 2$: clusters as stars

▷ Notations:

- Set of all D-family matchings of a graph G: $\mathcal{S}_D(G)$

D-family matching problem

▷ Score $\Phi(\mathcal{S})$ of a D-family-matching $\mathcal{S}$:

$$\Phi(\mathcal{S}) = \sum_{i=1}^{k} \sum_{e \in E(G[S_i])} w_e. \tag{5}$$

▷ Remarks:

- The sum runs over **all** edges of a connected component. (Later: see algorithms based on spanning trees.)
- We wish to compute a D-family-matching which minimizes the inconsistencies.

Definition 3 (D-family-matching problem). Let $D \in \mathbb{N}^+$. Given an intersection graph G, the D-family-matching problem consists in computing

$$\text{(Opt score for a given } D\text{)} \quad \Phi_D(G) = \max_{\mathcal{S} \in \mathcal{S}_D(G)} \Phi(\mathcal{S}). \tag{6}$$

NB: Score with the diameter D stressed: $\Phi(\mathcal{S}^{D=d})$

D-family matching: role of the diameter, illustration

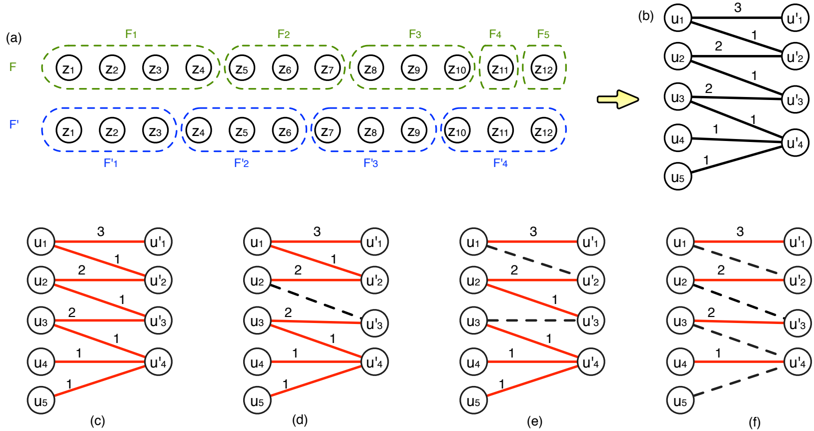


Figure: **Simple instance of the D-family-matching problem and solutions: panels (c,d,e,f) represent optimal solutions for different values of D. (a)** Simple instance of the D-family-matching problem with $t = 12$, $r = 5$, $r' = 4$, and so $n = 9$. The family F contains five sets and the family F' contains four sets. **(b)** Intersection graph G. **(c)** Optimal solution $\mathcal{S}$ for $D \geq 7$ with $\Phi(\mathcal{S}) = \Phi_D(G) = 12$. **(d)** Optimal solution $\mathcal{S}$ for $D = 3$ with $\Phi(\mathcal{S}) = \Phi_3(G) = 11$. **(e)** Optimal solution $\mathcal{S}$ for $D = 2$ with $\Phi(\mathcal{S}) = \Phi_2(G) = 9$. **(f)** Optimal solution $\mathcal{S}$ for $D = 1$ with $\Phi(\mathcal{S}) = \Phi_1(G) = 8$.

Notations, recap

Notation	Definition
$Z = \{z_1, \ldots, z_t\}$	Set of $t \geq 1$ elements
$F = \{F_1, \ldots, F_r\}$	Family of $r \geq 1$ disjoint subsets of Z
$F' = \{F'_1, \ldots, F'_{r'}\}$	Family of $r' \geq 1$ disjoint subsets of Z
$G = (V, E, w)$	Intersection graph of $n \geq 1$ nodes and $m \geq 1$ edges
$N_G(v) = \{v' \mid \{v, v'\} \in E\}$	Set of neighbors of node $v \in V$
$\Delta = \max_{v \in V} \lvert N_G(v) \rvert$	Maximum degree of G
$cc(G)$	Set of maximal connected components of G
$\mathcal{S} = \{S_1, \ldots, S_k\}$	D-family-matching
$\Phi(\mathcal{S}) = \sum_{i=1}^{k} \sum_{e \in E(G[S_i])} w_e$	Score of a D-family-matching $\mathcal{S}$
$\mathcal{S}_D(G)$	Set of all D-family-matching for G
$\Phi_D(G) = \max_{\mathcal{S} \in \mathcal{S}(G,D)} \Phi(\mathcal{S})$	Optimal score for the D-family-matching problem
$\mathcal{S}_D(G, T_r)$	Set of all D-family-matching constrained by T_r
$\Phi_D(G, T_r) = \max_{\mathcal{S} \in \mathcal{S}_D(G, T_r)} \Phi(\mathcal{S})$	Optimal score for the D-family-matching problem constrained by T_r

Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of D
  - Experiments
- Maximum Information Coefficient

Main result

Theorem 4. Let $D \geq 2$ be any integer. The decision version of the D-family-matching problem is NP-complete for :

- bipartite graphs of maximum degree 3;
- bipartite graphs of maximum degree 4 even if the maximum weight is constant.

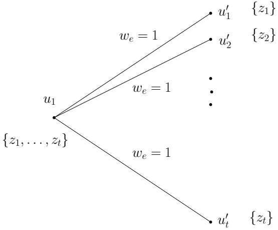
Moreover, the 2-family-matching problem is *APX*-hard for bipartite graphs of maximum degree 3 with unary weights.

▷ Open pb.: Is the D-family-matching problem in APX or not (constant factor approximation)?

- Nb: P $\neq$ NP: APX-hard pb. not in PTAS, i.e. no $(1+\varepsilon)$ approx

Greedy strategy on the diameter is not an option

Lemma 5. For any integer $n \geq 1$, then there exists an intersection graph $G = (V, E, w)$ composed of n nodes such that $\Phi_2(G)/\Phi_1(G) \geq n - 1$.



▷ One has:

- $\Phi(\mathcal{S}^{D=1}) = 1$ (one edge)
- $\Phi(\mathcal{S}^{D=2}) = t = n - 1$ (all edges)

Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of *D*
  - Experiments
- Maximum Information Coefficient

Trees: theorems

Theorem 6 (Computation of $\Phi_D(G)$ for trees). Let $D \in \mathbb{N}^+$. Consider any intersection tree $T = (V, E, w)$ of maximum degree $\Delta \geq 0$. Then, there exists an $O(D^2\Delta^2 n)$-time complexity algorithm for the D-family-matching problem for T.

Proof.
See black board. □

Theorem 7. For any $D \in \mathbb{N}^+$, the D-family-matching problem can be solved:

- in $O(Dn)$ time if G is a path;
- in $O(D^2n)$ time if G is a cycle(s) or a graph of maximum degree 2.

Proof.
See paper. □

Generic approach on spanning trees

Definition 8 (D-family-matching constrained by a tree). Let $G = (V, E, w)$ be an intersection graph and T be a spanning tree of G.
A D-family-matching for G *constrained by* T is a D-family-matching $\mathcal{S}$ for G such that all $S_i \in \mathcal{S}$ induces a connected subtree in T.
The set of all D-family-matching constrained by T is denoted $\mathcal{S}_D(G, T)$.

With this Def., we obtain the following sub-problem of D-family-matching:

Definition 9 (D-family-matching problem constrained by a tree). The D-family-matching problem consists in computing

$$\Phi_D(G, T) = \max_{\mathcal{S} \in \mathcal{S}_D(G,T)} \Phi(\mathcal{S}) \tag{7}$$

Generic algorithm for the D-family-matching problem

- Three ingredients:
  - **A property** $\Pi(\mathcal{M})$, depending on the set $\mathcal{M}$ of already computed D-family-matchings, represents the halting condition of the algorithm.
  - **A spanning tree generator** $\mathcal{R}(G, \lambda)$ computes the rooted spanning tree T^λ of G that is used at step $\lambda \geq 1$ by Algorithm $\mathcal{A}$.
  - **An algorithm** $\mathcal{A}(G, T^\lambda, D)$ computes a D-family-matching $\mathcal{S}^\lambda$ constrained by T^λ.

- Generic algorithm for the D-family-matching problem:

```
Require: An intersection graph $G = (V, E, w)$, an integer $D \geq 1$, a property
    $\Pi$, a spanning tree generator $\mathcal{R}$, and an algorithm $\mathcal{A}$.
$\mathcal{M} := \emptyset$, $\lambda := 0$
while $\neg\ \Pi(\mathcal{M})$ do
    $\lambda := \lambda + 1$; Compute the spanning tree $T^\lambda := \mathcal{R}(G, \lambda)$
    Compute $\mathcal{S}^\lambda$ by using Algorithm $\mathcal{A}(G, T^\lambda, D)$; $\mathcal{M} := \mathcal{M} \cup \mathcal{S}^\lambda$
return $\mathcal{S} \in \mathcal{M}$ of maximum score
```

Results on spanning trees

Lemma 10. Let $D \in \mathbb{N}^+$. Let G be any intersection graph. Then, there exists a rooted spanning tree T of G such that $\Phi_D(G) = \Phi_D(G, T)$.

Proof.
See black board. □

Lemma 11 (Computation of $\Phi_D(G, T)$). Let $D \in \mathbb{N}^+$. Let $G = (V, E, w)$ be any intersection graph and T be any spanning tree of G. Then, there exists a $O(2^{D\Delta \log_2(\Delta)} n)$-time algorithm for the D-family-matching problem for G constrained by T.

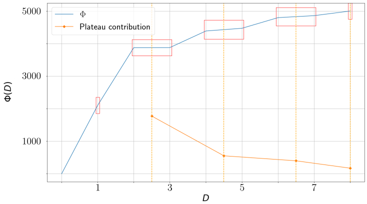
Proof.
See paper. □

Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
    - Problem
    - Jensen-Shannon divergence and discrepancy
    - Density based clustering
- Comparing clusterings
    - Motivation
    - Problem statement
    - Previous work
    - D-family matching: problem
    - Hardness
    - Algorithms
    - On the choice of D
    - Experiments
- Maximum Information Coefficient

Various strategies

- Three strategies:
  - (Stable plateaus) compute a set of non-overlapping plateaus optimizing a functional favoring *long* and *thin* plateaus.
  - (Prescribed num. plateaus) specify the number of plateaus to be obtained.
  - (Hierarchical plateaus) perform a hierarchical decomposition into plateaus, which is of interest if there are several *vertical scales*.



- Focus on:
  - local maxima of Φ
  - plateaus

Stable plateaus: long plateaus of small height

▷ Quality measures for a plateau:

- $\tau_w(y)$: positive increasing function for the plateau width
- $\tau_h(y)$: positive increasing function for the plateau height

▷ Given:

- Let D_G be the diameter of the intersection graph; we assume $\{(D, \Phi_D)\}_{D=1,\ldots,D_G}$
- Consider the set $\{\Phi_1(G), \ldots, \Phi_{D_G}(G)\}$
- Let $|I_x|$ is the size of plateau I_x

Definition 12. Determine $\mu \in \{1, \ldots, D_G\}$ plateaus (intervals) $I_1, \ldots, I_\mu$ of $[1, D_G]$ with

- $I_1 \cup \ldots \cup I_\mu = \{1, \ldots, D_G\}$, $I_x \cap I_{x'} = \emptyset$ for every $1 \leq x < x' \leq \mu$,
- such that the following function is minimum:

$$-\sum_{x=1}^{\mu} \frac{\tau_w(|I_x|)}{\tau_h(\max_{D,D' \in I_x \cap \mathbb{N}} \Phi_{D'}(G) - \Phi_D(G))}$$

Stable plateaus: construction

Theorem 13. There is an $O(D_G^2)$-time complexity algorithm that computes an optimal solution for the Tradeoff-plateau problem.

▷ Algorithm: blackboard

Hierarchical plateaus

▷ Dendogram of plateaus:

- For two consecutive plateaus, each consisting of a set of values $\{(D, \Phi_D)\}$: *coherence measure* for the union of these two plateaus: the maximum difference between any two values $\Phi_\cdot$ on these plateaus.
- Merge two plateaus realizing the minimum value then yields a dendogram.

▷ Formally: build a rooted tree $T = (V, E)$ representing the hierarchical plateaus

- One leaf per possible value of D; $D_G - 1$ internal nodes (including the root). That is, let $(I_1, I_2, \ldots, I_d)$ be the $d = D_G$ initial plateaus each composed of 1 point.
- Perform the aforementioned binary merge.

Statistical analysis: complementary topics

- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of *D*
  - Experiments
- Maximum Information Coefficient

Generic code and instantiation for experiments

▷ Implementation in the SBL:
`http://sbl.inria.fr/doc/D_family_matching-user-manual.html`.

▷ Implementation $STS(G, D)$ has the following ingredients:

- **(i)** the spanning tree generator $\mathcal{R}$ returns a *maximum spaning tree*, or a *random spanning tree*;
- **(ii)** the property $\Pi(\mathcal{M})$ returns true once we have computed a solution on the maximum spanning tree, as well as a solution on $n_i = (10,000)$ distinct random spanning trees (for a given n_i);
- **(iii)** $\mathcal{A}$: algorithm as in Theorem 6 with an additional step: edges for which both extremities belong to the same meta-cluster are added to the said meta-cluster. (In general, the intersection graph is indeed not a tree, so that such edges were unaccounted for.)
- The solution returned for a given graph G and a diameter D is the best yielded by the aforementioned $1 + n_i$ spanning trees.

Randomly edited clusterings: setup

- Initial random clusterings:
  - $(t = 1\ 000, r = 20)$ and $(t = 3\ 000, r = 50)$.
  - Generated with the Boltzmann sampler from Flajolet - Duchon et al
  - Due to the randomness, the process is repeated $N_r = 10$ times for each pair (t, r).
- Edited clusterings: a copy F' of a clustering F is edited in two steps
  - Union operations: e unions reduce the number of clusters to $r - e$
  - Jittering: for each cluster, a fraction τ of its items are distributed amongst the remaining $k - 1$ clusters uniformly at random.
- Values: 9 scenarios for edits and jitters
  - $e \in \{0, \lfloor r/4 \rfloor, \lfloor r/2 \rfloor\}$ and $\tau \in \{0.05, 0.1, 0.2\}$. (NB: for $e = 0$, F' is a jittered version of F (i.e. the numbers of clusters are identical.)
  - yields $N_r \times \#(t, r) \times \#e \times \#\tau = 180$ comparisons, which are ascribed to 9 scenarii (3 values for $e \times 3$ values for τ) denoted $EeJy$, where $y = 100\tau$.
- Comparson against VI: comparison of normalized scores $\in [0, 1]$:

$$s_\Phi = 1 - \Phi_D(\cdot)/t \text{ versus } s_{VI} = \mathrm{VI}/\log t.$$

Randomly edited clusterings: results for $(t = 1000, r = 20)$

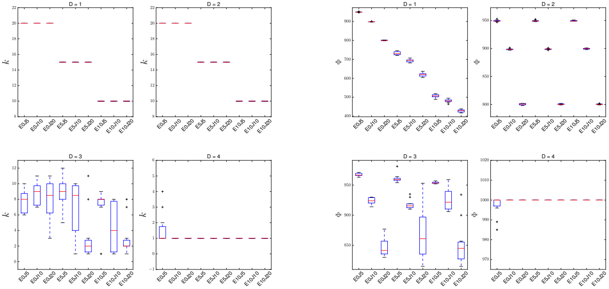


Figure: Algorithm $STS(G, D)$ for clusterings with $(t = 1\,000, r = 20)$. **(Left)** Best value for k as a function of the 9 scenarii. **(Right)** Scores s_Φ as a function of the 9 scenarii.

- $D \leq 2$: algo. finds the right number of clusters $\forall e$ (resp: 20, 15, 10)
- For $D = 2$: score $\Phi_D(\cdot)$ is almost perfect (≥ 800, wrt $t = 1000$)
- Across scenarii: scores hardly depend on the jitter level
- For $D = 3$: scores $\Phi_D(\cdot)$ varies significantly–but medians ok
- For $D = 4$: the algorithms output a full graph

Comparison with the Variation of Information: results

- **Method:** scatter plot of $s_{\Phi} = 1 - \Phi_D(\cdot)/t$ versus $s_{VI} = \mathrm{VI}/\log t$

NB: 1 symbol per scenario: copy number of a symbol: number of repeats.

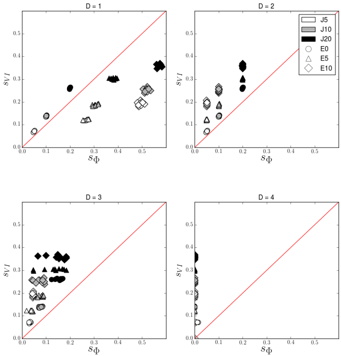
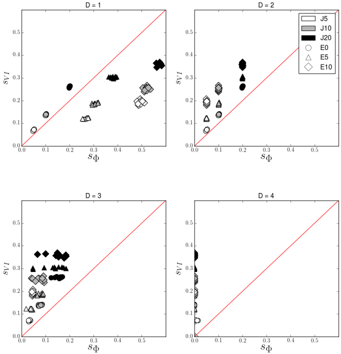


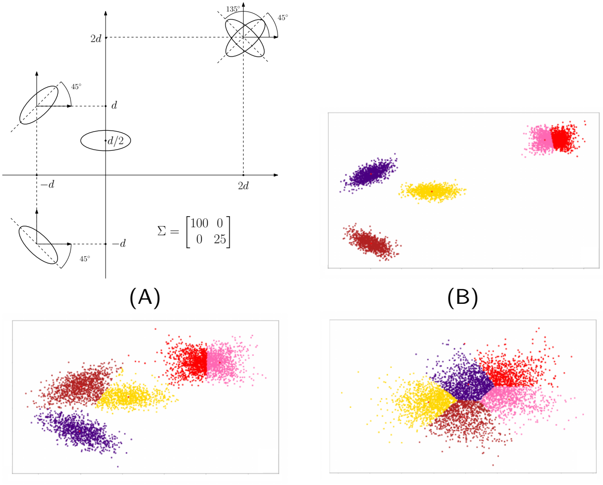
Figure: **Normalized score s_{VI} versus normalized score s_{Φ} of algorithm $STS(G, D)$.** Each marker is a different union scenario and each color represents a different jitter scenario following the legend on the upper right. We plot the $y = x$ function for reference.

Comparison with the Variation of Information: results



- $D = 2$: s_Φ corresponds to a matching.
- $D = 2$, two key differences with VI: $s_\Phi \leq s_{VI}$; s_Φ constant against union operations. Both s_{VI} and s_Φ are affected by jittering.
- $D = 3$: higher variability in s_Φ; dependence on jittering and # union operations.
- For $D = 4$: $s_\Phi = 0$ ie the full intersection graph reported.

On the separability of clusters and D: setup



(A)

(B)

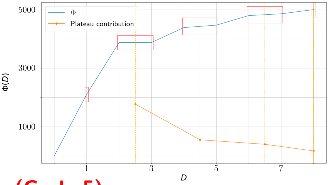
(C)

(D)

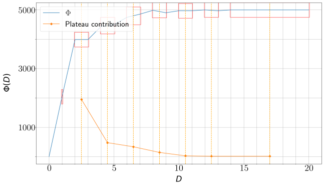
Figure: **Parameterized dataset: mixture of 5 Gaussian blobs**. **(A)** Relative position of the five Gaussian blobs: function of d **(B, C, D)** $t = 5,000$, $d = 50, 20, 5$. Samples clustered with `k-means++` ($k = 5$).

On the separability of clusters and D: plateaus

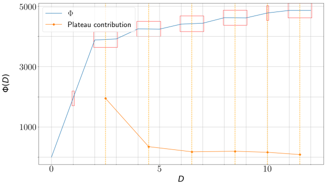
▷ **(A, d=50)**



▷ **(B, d=20)**

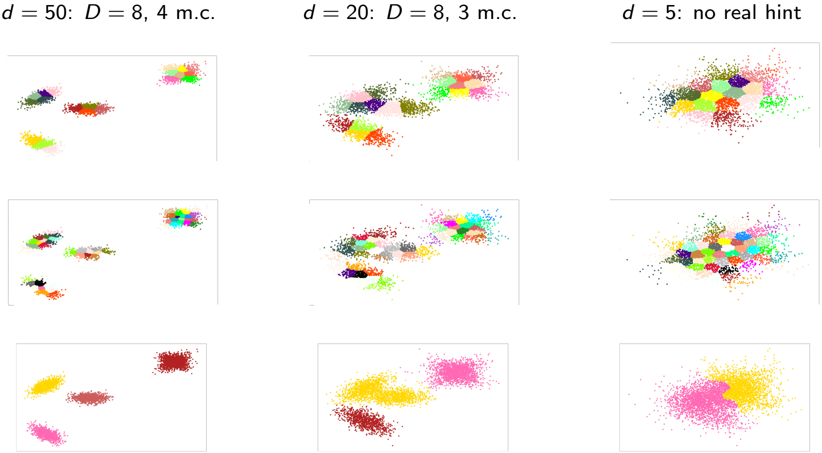


▷ **(C, d=5)**



- **(A)** $d = 50$, $k = 4$ meta-clusters suggested for $D = 8$.
- **(B)** $d = 20$, $k = 3$ meta-clusters suggested for $D = 8$.
- **(C)** $d = 5$ No obvious choice for the number of meta-clusters.

On the separability of clusters and D: illustrations



$d = 50$: $D = 8$, 4 m.c.

$d = 20$: $D = 8$, 3 m.c.

$d = 5$: no real hint

Final words on the choice of D

Fom Strehl et al (JMLR 2002): "*In fact, the* `right` *number of clusters in a dataset often depends on the* `scale` *at which the dataset is inspected*".

- Parameter D acts as a scale parameter providing information of the structure of the intersection graph.
- When this graph is dense or has a specific topology (star-shaped), trivial values of Φ are obtained for small values of D, and a unit change of D may trigger an abrupt change of Φ. However, in more complex situations, large values of D may be required.
- As a general strategy to choose D, we suggest identifying drops in Φ when decreasing D. Indeed, for any range of D corresponding to a *plateau* for Φ, the most significant value for D is the smallest one.

Outlook

- Interesting complexity issues: open
- Useful tool, available from `https://sbl.inria.fr/doc/D_family_matching-user-manual.html`
- Interesting connexions with model clustering in deep learning – amongst others

▷Ref: Cazals et al, ACM J. of Experimental Algorithms, 2019
▷Ref: Interactive Naming for Explaining Deep Neural Networks: A Formative Study M Hamidi-Haines, Z Qi, A Fern, F Li arXiv preprint arXiv, 2018.

Statistical analysis: complementary topics

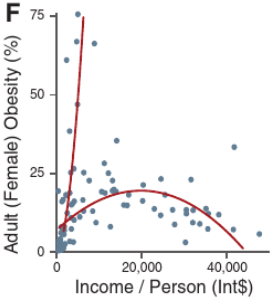
- Beyond Two-Sample-Tests
  - Problem
  - Jensen-Shannon divergence and discrepancy
  - Density based clustering
- Comparing clusterings
  - Motivation
  - Problem statement
  - Previous work
  - D-family matching: problem
  - Hardness
  - Algorithms
  - On the choice of D
  - Experiments
- Maximum Information Coefficient

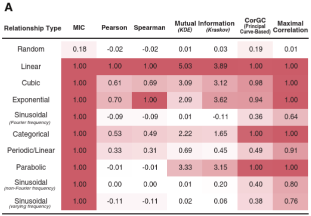
Detecting Novel Associations in Large Data Sets

The Maximal Information Coefficient (MIC)

Science 334, December 2011

9 authors, led by Michael Mitzenmacher and Pardis Sabeti, Harvard





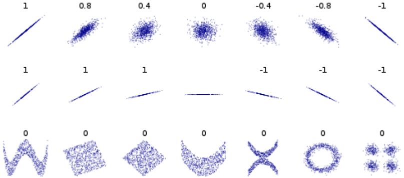
A

Relationship Type	MIC	Pearson	Spearman	Mutual Information		CorGC (Principal Curve-Based)	Maximal Correlation
				(KDE)	*(Kraskov)*		
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal *(Fourier frequency)*	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal *(non-Fourier frequency)*	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal *(varying frequency)*	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

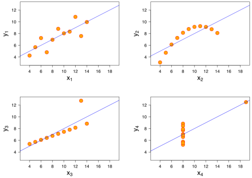
Correlations in 2D: the Pearson correlation coefficient

Does the knowledge of X provide information on Y?

- ▷ The Pearson coeff.: $\rho = cov(X, Y)/(\sigma_X \sigma_Y)$



- ▷ Anscombe's quartet: $\rho = 0.816$

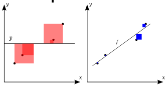


- ▷ Properties of ρ:
  - Coupled to linear regression $f_i = \alpha x_i + \beta$
    - $\alpha = \rho \sigma_Y / \sigma_X$
    - coeff. of determination: $R^2 = \rho^2$
  - Not invariant to rotations
  - Spearman's coeff: Pearson on ranks: for monotonic correlations

- ▷ Coeff of determination

  R^2: variance explained by the model

  $1 - R^2$: unexplained var. / noise level



$R^2 = 1 - \frac{AreaResiduals}{AreaVariance}$

http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
http://en.wikipedia.org/wiki/Correlation_and_dependence

http://en.wikipedia.org/wiki/Coefficient_of_determination

Information Theory: Key Quantities

▷ Entropy of the r.v. X

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

$$H(X) \leq \log \mid \mathcal{X} \mid$$

▷ Joint and Conditional entropies

$$H(X, Y) = -\sum p(x, y) \log p(x, y)$$

$$H(Y \mid X) = \sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x)$$

▷ Relative entropy: Kullback-Leibler divergence of two distributions on the same proba. space:

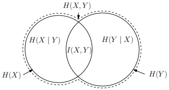
Def:
$D(P, Q) = \sum_{x \in \mathcal{X}} p(x) \log(p(x)/q(x))$

Prop.: $D(P, Q) \geq 0$

▷ Mutual information

$$I(X, Y) = \sum p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

$$\begin{cases} I(X, Y) & = D(p(x, y), p(x)p(y)) \\ & = H(X) - H(X \mid Y) \\ & = H(X) + H(Y) - H(X, Y) \end{cases}$$



Notes: (i)Poincare formula for I
(ii) I as correlation: common entropy

http://en.wikipedia.org/wiki/Mutual_information
http://en.wikipedia.org/wiki/Kullback-Leibler_divergence

Maximal Information Coefficient (MIC): Definition

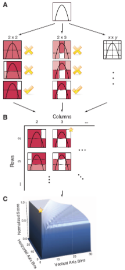
- ▷ Input: a 2D point cloud $D = \{(x_i, y_i)\}_{i=1,\dots,n}$ and its bounding box
- ▷ Grids: $G_{x,y}$: grids of size $x \times y$ not necessarily regular
- ▷ Joint proba/marginal of $D_{|G}$: fraction of samples, out of n, in a cell/row/column

▷ Def of MIC:

$$I^*(D,x,y) = \max_{G \in G_{x,y}} I(D_{|G}) \qquad (8)$$

$$M_{xy} = \frac{I^*(D,x,y)}{\log \min(x,y)} \qquad (9)$$

$$MIC = \max_{xy < B(n) = n^{1-\epsilon}} M_{xy} \qquad (10)$$



Note: For the normalization of Eq. (9): $\log \min(x, y)$ rather than n: # cells sub-linear, see Eq. (10).

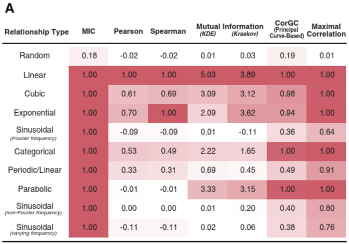
▷ Elementary properties:

- $M_{xy} \in [0,1]$
- $MIC(X,Y) = MIC(Y,X)$
- MIC invariant to order preserv. transf.
  grids determined by abscissa / ordinates
- MIC not invariant to rotations
  cf $y = x$ vs $y = c$

▷ Note: exploring all grids ∼ enclosing the data in a *tube*

MIC, illustrations (I): the functional noiseless case

- ▷ Ideal scores: (almost) one

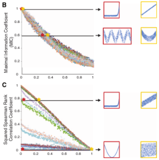


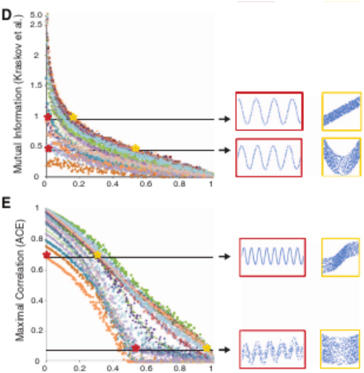
A

Relationship Type	MIC	Pearson	Spearman	Mutual Information		CorGC	Maximal Correlation
				(KDE)	*(Kraskov)*	(Principal Curve-Based)	
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal *(Fourier frequency)*	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal *(non-Fourier frequency)*	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal *(varying frequency)*	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

MIC, illustrations (II): the functional noisy case

▷ Testing 27 functions with uniform vertical noise: $MIC = \textit{function of}\,(1 - R^2)$
with R^2 the determination coeff of the data relative to the noiseless function

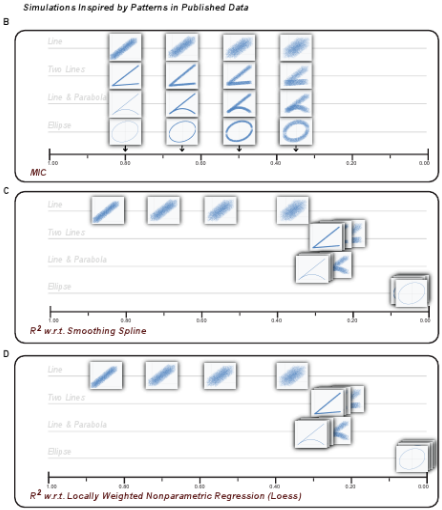




Bottomline is → $MIC \sim R^2$: easy comparison of ≠ functions

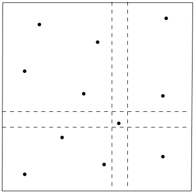
MIC, illustrations (III): the non functional noisy case

- MIC also degrades *smoothly* as a function of the noise level



Simulations Inspired by Patterns in Published Data

Details of the Definition



- ▷ Grid resolution B(n)
  - – too low: searching for simple patterns
  - – too high: high scores even for random data
    samples isolated in cell/column
- ▷ Normalization
  - – grids with $\neq$ dimensions have $\neq$ mutual information
  - – normalizing by $\log \min(x, y)$:
    Comparing grids of $\neq$ dimensions
    Ensures that almost all
    noiseless functions get MIC of one
    (finite union of differentiable curves)

MIC: Theorems

(Ten pages of proofs in the Supplemental)

- ▷ Thm 1. If X and Y independent R.V.: ApproxMIC converges to 0 in probability when $n \to \infty$
If X and Y are not independent R.V.: MIC bounded away from 0 almost surely.

- ▷ Thm 2. For any joint distribution (X, Y), MIC computed with a number of cells $B(n) = n^{1+\varepsilon}$ would yield $MIC \to 1$ almost surely.

- ▷ Thm 3. Let D consist of n samples drawn according to a distribution $(X, f(X))$, with f nowhere constant on $[0, 1]$. Then $MIC \to 1$ almost surely.

- ▷ Thm 4. If the support of $(x(t), y(t))$ is a finite union of smooth curves, nowhere flat (critical points of measure 0), then $MIC > 1 - \varepsilon$ for large n.

- ▷ Thm 5. MIC of a noisy functional $(X, f(X) + E_h)$, with E_h uniform noise in $[-h, h]$, is lower bounded by a (complex) functional of the R^2 between $f(X)$ and $f(X) + E_h$.

More Ingredients

– Computing MIC: algorithm ApproxMIC uses 2D dynamic programming

– p-value calculation for H0: *X and Y are statistically independent*
  Create surrogate datasets created with random permutations
    (eg of X for Y fixed)

– $MIC - \rho^2$ as a natural measure of linear dependence:
  Since MIC behaves as R^2 for functional relationships

– Symmetry of the matrix M_{xy}: hints at monotony
  Maximum Asymmetry Score $| m_{xy} - m_{yx} |$
    Hints at periodic relationships with non constant period

– Software MINE: `http://www.exploredata.net/`