Clustering Tendency

Assistant Professor :

Dr. Mohammad Javad Fadaeieslam

By :

Amir Shokri – Farshad Asgharzade – Sajad Dehghan – Amin Nazari

Introduction

- ➢Problems clustering algorithms :
  - ❑Most clustering algorithms **impose** a clustering structure on a data set X even though the vectors of X do not exhibit such a structure.
- ✓Solution :
  - ❑Before we apply any clustering algorithm on X , its must first be verified that X possesses a clustering structure.
- ❖**clustering tendency :** The problem of determining the presence or the absence of a clustering structure in X.
- Clustering tendency methods have been applied in various application areas, However, most of these methods are suitable only for $l = 2$. In the sequel, we discuss the problem in the general $l \geq 2$ case.
- ✓focus : methods that are suitable for detecting compact clusters (if any).

Clustering Tendency

- Clustering tendency is heavily based on hypothesis testing.
- Specifically is based on testing the **randomness (null) hypothesis (H0)** against the **clustering hypothesis (H2)** and the **regularity hypothesis (H1)**.

❖ **Randomness hypothesis :** the vectors of X are randomly distributed, according to the uniform distribution in the sampling window8 of X"(H0).

❖ **Clustering hypothesis:** the vectors of *X* are regularly spaced in the sampling window." This implies that, they are not too close to each other.

❖ **Regularity hypothesis :** the vectors of X form clusters.

- P(q|H0) , P(q|H01) , P(q|H2) are estimated via monte carlo simulations.
- If the **randomness** or the **regularity hypothesis** is accepted, methods alternative to clustering analysis should be used for the interpretation of the data set ***X***.

Clustering Tendency(cont.)

- There are **two** key points that have an important influence on the performance of many statistical tests used in clustering tendency:
  1) dimensionality of the data
  2) sampling window

➢ Problem sampling window : in practice, we do not know the sampling window

✓ ways to overcome this situation is :

1) use a periodic extension of the sampling window
2) sampling frame (extension of the sampling window)

❖ **sampling frame** : consider data in a smaller area inside the sampling window.

- With **sampling frame** , we overcome the boundary effects in the sampling frame by considering points outside it and inside the sampling frame, for the estimation of statistical properties.

Sampling Window

- A method for estimating the sampling window is to use the convex hull of the vectors in X.

➢ Problems : the distributions for the tests, derived using this sampling window :

1) depend on the specific data at hand.
2) high computational cost for computing the convex hull of **X**.

✓ **An alternative :** define the sampling window as the hypersphere centered at the mean point of **X** and including half of its vectors.

- **test statistics**, ***q***, suitable for the detection of clustering tendency :
    1) Generation of clustered data
    2) Generation of regularly spaced data

Generation of clustered data

- A well-known procedure for generating (compact) clustered data is the **Neyman–Scott** procedure :

  1) assumes that the sampling window is known

  2) The number of points in each cluster follows the Poisson distribution

➢ requires inputs :

1. total number of points ***N*** of the set
2. the intensity of the Poisson process
3. **spread parameter :** that controls the spread of each cluster around its center

Generation of clustered data(cont.)

- STEPS :

  I. randomly insert a point $\boldsymbol{y_i}$ in the sampling window, following the uniform distribution

  II. This point serves as the center of the ith cluster, and we determine its number of vectors, $\boldsymbol{n_i}$, using the ***Poisson distribution***.

  III. the $\boldsymbol{n_i}$ points around $\boldsymbol{y_i}$ are generated according to the normal distribution with mean y_i and covariance matrix $\boldsymbol{\delta^2 I}$.

- If a point turns out to be outside the sampling window, we ignore it and another one is generated.
- This procedure is repeated until **N** points have been inserted in the sampling window.

Generation of regularly spaced data

- Perhaps the simplest way to produce regularly spaced points is :
  - ❑define a lattice in the convex hull of X and to place the vectors at its vertices
  - ❑An alternative procedure, known as simple sequential inhibition (SSI)

I. The points $\boldsymbol{y_i}$ are inserted in the sampling window one at a time.

II. For each point we define a hypersphere of radius $\boldsymbol{r}$ centered at $\boldsymbol{y_i}$.

III. The next point can be placed anywhere in the sampling window in such a way that its hypersphere does not intersect with any of the hyperspheres defined by the previously inserted points.

➢The procedure stops :

- ❑a predetermined number of points have been inserted in the sampling window
- ❑no more points can be inserted in the sampling window, after say a few thousand trials

Generation of regularly spaced data(cont.)

❖packing density : A measure of the degree of fulfillment of the sampling window

- which is defined as : $\rho = \frac{L}{V} V_r$

❖ $\frac{L}{V}$ is the average number of points per unit volume

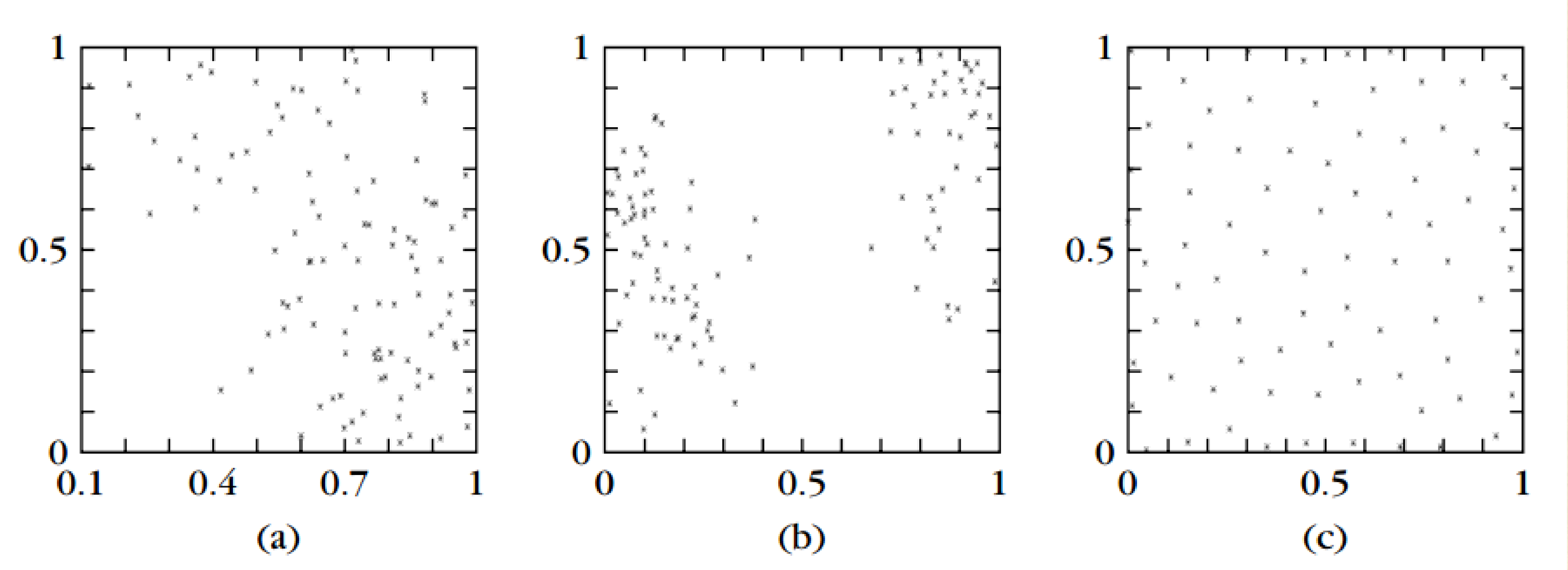
❖ V_r is the volume of a hypersphere of radius r

❖ V_r can be written as : $V_r = Ar^l$

- where A is the volume of the l-dimensional hypersphere with unit radius, which is given by :

$$A = \frac{\pi^{\frac{l}{2}}}{\Gamma(\frac{l}{2} + 1)}$$

Example



- **(a) and (b) :** Clustered data sets produced by the Neyman–Scott process

- **(c) :** Regularly spaced data produced by the SSI model

Tests for Spatial Randomness

- Several tests for spatial randomness have been proposed in the literature. All of them assume knowledge of the sampling window :
  - The scan test
  - the quadrat analysis
  - the second moment structure
  - the interpoint distances
- provide us with tests for clustering tendency that have been extensively used when $l = 2$.
- **three** methods for determining **clustering tendency** that are well suited for the general $l \geq 2$ case. All these methods require knowledge of the sampling window :

1) **Tests Based on Structural Graphs**
2) **Tests Based on Nearest Neighbor Distances**
3) **A Sparse Decomposition Technique**

1) ***Tests Based on Structural Graphs***

- based on the idea of the ***minimum spanning tree (MST)***

➢ Steps :

I. determine the convex region where the vectors of X lie.

II. generate M vectors that are uniformly distributed over a region that approximates the convex region found before (usually M = N). These vectors constitute the set X

III. find the MST of X ∪ X and we determine the number of edges, q, that connect vectors of X with vectors of X.

✓ If X contains clusters, then we expect q to be small.

❖ small values of q indicate the presence of clusters.

❖ large values of q indicate a regular arrangement of the vectors of X.

1) Tests Based on Structural Graphs(cont.)

- ***mean value*** of ***q*** and the ***variance*** of ***q*** under the null (randomness) hypothesis, conditioned on e, are derived:

  - $E(q|H_0) = \frac{2MN}{M+N}$

  - $var(q|e, H_0) = \frac{2MN}{L(L-1)}\left[\frac{2MN-L}{L}\right] + \frac{e-L+2}{(L\_2)(L-3)}[L(L-1) - 4MN + 2]$

- where ***L*** **=** ***M*** **+** ***N*** and **e** = the number of pairs of the MST edges that share a node.
- if ***M, N*** →∞ and ***M/N*** is away from **0** and ∞ , the pdf of the statistic is approximately given by the standard normal distribution.

Tests Based on Structural Graphs(cont.)

❖ Formula :

$$q` = \frac{q - E(q|H_0)}{\sqrt{var(q|e, H_0)}}$$

if **q`** is less than the $\boldsymbol{\rho}$ -percentile of the standard normal distribution:

- ➢ reject H_0 at significance level $\boldsymbol{\rho}$

✓ This test exhibits high power against **clustering tendency** and little power against **regularity**.

2) ***Tests Based on Nearest Neighbor Distances***

- The tests rely on the distances between the vectors of $\boldsymbol{X}$ and a number of vectors which are randomly placed in the **sampling window**.

- **Two** tests of this kind are :
  1) **The Hopkins test**
     - ➢This statistic compares the nearest neighbor distribution of the points in $\boldsymbol{X_1}$ with that from the points in $\boldsymbol{X}$.
  2) **The Cox–Lewis test**
     - ➢It follows the setup of the previous test with the exception that $\boldsymbol{X_1}$ need not be defined.

2_1)The Hopkins Test

➢ Definitions:

❖ $X` \{yi, i\ 1, \ldots, M\}, M << N$: a set of vectors that are randomly distributed in the sampling window, following the uniform distribution.

❖ $X_1 \subset X$: a set of $\boldsymbol{M}$ randomly chosen vectors of $\boldsymbol{X}$.

❖ d_j : the distance from $\boldsymbol{y_j} \in X`$ to its closest vector in $\boldsymbol{X_1}$, denoted by $\boldsymbol{x_j}$,

❖ j : the distance from $\boldsymbol{X_j}$ to its closest vector in $\boldsymbol{X_1} - \{\boldsymbol{X_j}\}$.

- The Hopkins statistic involves the **lth** powers of $\boldsymbol{d_j}$ and $\boldsymbol{\delta_j}$ and it is defined as:

$$\boldsymbol{h} = \frac{\sum_{j=1}^{M} d_j^l}{\sum_{j=1}^{M} d_j^l + \sum_{j=1}^{M} \delta_j^l}$$

2_1)The Hopkins Test (cont.)

✓ ***Values of h :***

❖**Large values :** large values of h indicate the presence of a clustering structure in X.

❖**Small values :** small values of h indicate the presence of regularly spaced points.

❖**h = ½ :** a value around 1/2 is an indication that the vectors of X are randomly distributed over the sampling window.

- if the generated vectors are distributed according to a **Poisson random process** and all nearest neighbor distances are **statistically independent**:

✓h (under H_0) follows a beta distribution, with (M, M) parameters

2_2)The Cox–Lewis test

- Definitions:

- x_j :For each $y_j \in X$` we determine its closest vector in X
- x_i : the vector closest to X_j in X-$\{x_j\}$
- d_i : be the distance between y_j and x_j
- δ_i : distance between x_j and x_i
- M : be the number of such y_j's

2-2) The Cox–Lewis test(cont.)

- We consider all y_j's for which $2d_j/\delta_i$ is greater than or equal to **one**.
- Finally, we define the statistic :

$$R = \frac{1}{M}\sum_{j=1}^{M`} R_j$$

✓**Values of R :**

❖ ***Small values :*** indicate the presence of a clustering structure in ***X***

❖ ***large values :*** indicate a regular structure in ***X***.

❖ ***R values around the mean :*** indicate that the vectors of ***X*** are randomly arranged in the sampling window.

Hopkins vs Cox–Lewis

1) **The Hopkins test**

- This test exhibits high power against regularity for a hypercubic sampling window and periodic boundaries, for $l = 2, \ldots, 5$.
- However, its power is limited against ***clustering tendency***.

2) **The Cox–Lewis test**

- This test is less intuitive than the previous one. It was first proposed for the two-dimensional case and it has been extended to the general $l \geq 2$ dimensional case.
- This test exhibits inferior performance compared with the Hopkins test against the clustering alternative.
- However, this is not the case against the ***regularity hypothesis.***

3) ***A Sparse Decomposition Technique***

- This technique begins with the data set ***X*** and sequentially removes vectors from it until no vectors are left.
- A sequential decomposition D of X is a partition of X into ***L****1*, . . . **,** ***Lk*** sets, such that the order of their formation matters. Li's are also called decomposition layers.

I. *We denote by* ***MST(X)****.* ***S(X)*** be the set derived from ***X*** according to the following procedure. Initially, ***S(X) =Ø***.

II. move an end point ***x*** of the longest edge,***e***,of the ***MST (X)*** to ***S(X).***

III. mark this point and all points that lie at a distance less than or equal to ***b*** from ***x***, where ***b*** is the length of ***e***.

IV. determine the unmarked point , ***y ∈ X*** , that lies closer to ***S(X)*** and we move it to ***S(X)***.

V. mark all the unmarked vectors that lie at a distance no greater than ***b*** from ***y***.

VI. apply the same procedure for all the unmarked vectors of ***X***.

VII. The procedure terminates : when all vectors are marked.

3) A Sparse Decomposition Technique(cont.)

- Let us define $\boldsymbol{R}(\boldsymbol{X}) \equiv \boldsymbol{X} \ \ \boldsymbol{S}(\boldsymbol{X})$**.** Setting $\mathbf{X} = \mathbf{R}^{\mathbf{0}}(\mathbf{X})$, we define :

$$\mathbf{L_i} = \mathbf{S}(\mathbf{R}^{\mathbf{i-1}}(\mathbf{X}))\ ,\ \ \mathbf{i=1,2,\ldots,k}$$

- where ***k*** is the smallest integer such that $\boldsymbol{R}^{\mathbf{0}}(\boldsymbol{X}) = \boldsymbol{\emptyset}$.
- The index ***i*** denotes the so-called *decomposition layer*. Intuitively speaking, the procedure sequentially "peels" *X* until all of its vectors have been removed.
- The information that becomes available to us after the application of the decomposition procedure is :
  - ❑(a) the number of decomposition layers k
  - ❑(b) the decomposition layers Li
  - ❑(c) the cardinality, li, of the Li decomposition layer, i =1, . . . , k
  - ❑(d) the sequence of the longest MST edges used in deriving the decomposition layers

3) A Sparse Decomposition Technique(cont.)

- The **decomposition procedure** gives different results, when :
  - the vectors of ***X*** are clustered
  - the vectors of ***X*** regularly spaced or randomly distributed in the sampling window
- Based on this observation we may define statistical indices utilizing the information associated with this decomposition procedure.
- For example, it is expected that the number of decomposition layers, k, is smaller for random data than it is for clustered data. Also, it is smaller for regularly spaced data than for random data .

Another tests

- Exist Several tests that rely on the preceding information ,One such statistic that exhibits good performance is the so-called P statistic, which is defined as follows:

$$\mathbf{P} = \prod_{\mathbf{i}=1}^{\mathbf{K}} \frac{\mathbf{l_i}}{\mathbf{n_i} - \mathbf{l_i}}$$

- where $\boldsymbol{n_i}$ is the number of points in $\mathbf{R^{i-1}(X)}$. In words,each factor of **P** is the ratio of the removed to the remaining points at each decomposition stage.
- Finally, tests for clustering tendency for the cases in which ordinal proximity matrices are in use have also been proposed.
  - ❑Most of them are based on graph theory concepts.

Conclusion

➢ The Basic steps of the clustering tendency philosophy are :

I. Definition of a test statistics q suitable for the detection of clustering tendency.

II. Estimation of the pdf of q under the null (H0) hypothesis, p(q|H0)

III. Estimation of p(q|H1) and p(q|H2) (they are necessary for measuring the power of q (the probability of making a correct decision when H0 is rejected)against the regularity and the clustering tendency hypotheses).

IV. Evaluation of q for the data set at hand ,X, and examination whether it lies in the critical interval of p(q|H0), which corresponds to a predetermined significance level *p*