LMCLUS Documentation

Release 1.1.0

Art Wild

Aug 14, 2018

Contents

1	Linear Manifold Clustering	3
2	Parameters	5
3	Separations	9
4	Utilities	11

Many clustering algorithms are based on the concept that a cluster has a single center point. Clusters could be considered as groups of points compact around a linear manifold. A linear manifold of dimension 0 is a point. So clustering around a center point is a special case of linear manifold clustering.

Linear manifold clustering algorithm identifies subsets of the data which are embedded in arbitrary oriented lower dimensional linear manifolds, not nessesaraly zero dimensional. Minimal subsets of points are repeatedly sampled to construct trial a linear manifold and isolete points around it based of the proximity of points to the found manifold. Using top-down aproach, the linear manifold clustering alogorithm iterativly partitions dataset and discovers clusters embedded into low-dimensioanl linear subspaces¹.

LMCLUS.jl is a Julia package for linear manifold clustering.

Contents:

¹ Haralick, R. & Harpaz, R. "Linear manifold clustering in high dimensional spaces by stochastic search", Pattern recognition, Elsevier, 2007, 40, 2672-2684, DOI:10.1016/j.patcog.2007.01.020

Linear Manifold Clustering

Linear manifold clustering algorithm (LMCLUS) discovers clusters which are described by a following model:

$$x = \mu^{N \times 1} + B^{N \times K} \phi^{K \times 1} + \bar{B}^{N \times N - K} \epsilon^{N - K \times 1}$$

where N is a dimension of the dataset, K is dimension of the manifold, $\mu \in \mathbb{R}^N$ is a linear manifold translation vector, B is a matrix whose columns are orthonormal vectors that span \mathbb{R}^K , \overline{B} is a matrix whose columns span subspace orthogonal to spanned by columns of B, ϕ is a zero-mean random vector whose entries are i.i.d. from a support of linear manifold, ϵ is a zero-mean random vector with small variance independent of ϕ .

1.1 Clustering

This package implements the *LMCLUS* algorithm in the lmclus function:

lmclus(X, p)

Performs linear manifold clustering over the given dataset.

Parameters

- **X** The given sample matrix. Each column of X is a sample.
- **p** The clustering parameters as instance of *LMCLUSParameters*.

This function returns an LMCLUSResult instance.

1.2 Results

Let M be an instance of Manifold, n be the number of observations, and d be the dimension of the linear manifold cluster.

indim(M)

Returns a dimension of the observation space.

$\mathtt{outdim}(M)$

Returns a dimension of the linear manifold cluster which is the dimension of the subspace.

$\texttt{size}\left(M\right)$

Returns the number of points in the cluster which is the size of the cluster.

points(M)

Returns indexes of points assigned to the cluster.

mean(M)

Returns the translation vector μ which contains coordinates of the linear manifold origin.

projection(M)

Returns the basis matrix with columns corresponding to orthonormal vectors that span the linear manifold."

separation(M)

Returns the instance of Separation object.

1.3 Example

using LMCLUS

```
# Load test data, remove label column and flip
X = readdlm(Pkg.dir("LMCLUS", "test", "testData"), ',')[:,1:end-1]'
# Initialize clustering parameters with
# maximum dimensionality for clusters.
# I should be less then original space dimension.
params = LMCLUSParameters(5)
# perform clustering and returns a collection of clusters
clust = lmclus(X, params)
# pick the first cluster
M = manifold(clust, 1)
# obtain indexes of points assigned to the cluster
l = points(M)
# obtain the linear manifold cluster translation vector
mu = mean(M)
# get basis vectors that span manifold as columns of the returned matrix
B = projection(M)
# get separation properties
S = separation(M)
```

Parameters

The clustering properties set in LMCLUSParameters instance, which is defined as follows:

```
type LMCLUSParameters
   min dim::Int
                                    # Minimum cluster dimension
   max_dim::Int
                                   # Maximum cluster dimension
                                   # Nominal number of resulting clusters
   number_of_clusters::Int
                                   # Fixed number of bins for the distance.
   hist_bin_size::Int
→histogram.
   min cluster size::Int
                                   # Minimum cluster size
   best_bound::Float64
                                   # Best bound
   error_bound::Float64
                                   # Error bound
   max_bin_portion::Float64
                                  # Maximum histogram bin size
                                   # Random seed
   random_seed::Int64
                                   # Sampling heuristic
   sampling_heuristic::Int
   sampling_factor::Float64
                                   # Sampling factor
   histogram_sampling::Bool
                                   # Sample points for distance histogram
                                   # Enable zero-dimensional manifold
   zero_d_search::Bool
⇔search
                                   # Manifold cluster basis alignment
   basis_alignment::Bool
                                    # Manifold dimensionality adjustment
   dim_adjustment::Bool
   dim_adjustment_ratio::Float64
                                   # Ratio of manifold principal subspace_
↔variance
                                    # Enable MDL heuristic
   mdl::Bool
   mdl_model_precision::Int
                                   # MDL model precision encoding constant
   mdl_data_precision::Int
                                   # MDL data precision encoding constant
   mdl_quant_error::Float64
                                   # Quantization error of a bin size
⇔calculation
   mdl_compres_ratio::Float64
                                    # Cluster compression ration
   log_level::Int
                                    # Log level (0-5)
end
```

Here is a description of algorithm parameters and their default values:

min_dim Low bound of a cluster manifold dimension. It cannot be larger then a dimension. It cannot be larger then a dimensionality of a dataset. 1 number_of_clusters Fxpected number of clusters. Re- 10 10 hist_bin_size Number of bins for a distance histogram heuristics. 0 hist_bin_size Number of bins in the distance histogram determined by parameter is set to zero. the number of bins in the distance histogram determined by parameter is set to zero. the number of bins in the distance histogram determined by parameter max_bin_portlan. 1.0 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used of a distance histogram. 1.0 error_bound Separation best bound value is used a minimal number of samples required to correctly identify a hicar manifold cluster. 1.0 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a hicar manifold cluster. 1.e-4 nandom_seed Ramdom number generator seed. If no a specified then RNG will be reinitial identify a discriminative discret a every ran. 0.1 sampling_heuristic The choice of furth sampling the a quantity exponential in max_clim and cluster_number of samples a parameters 2. Nill sample fixed number of points 3. the lesser of the previous two sampling factor used in the sam-pling heuristic (see above, options quantity exponentia	name	description	default
mension. mension. remain of a cluster manifold dimension. It cannot be larger then a dimensionality of a dataset. number, of clusters Expected number of clusters. Request for the sampling hearistics. 10 hist_bin_size Number of bins for a distance histogram of bins parameter is set to zero, the number of bins for a distance histogram. If this parameter is set to zero, the number of bins for a distance histogram determined by parameter inset of a collection of data points to be considered as a proper cluster. 0 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 1.0 max_bin_portion Separation best bound value is used for evaluating a goodness of segaration characterized by a discriminability and a depth between modes of a distance histogram. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear marifold cluster. Wate should be selected from a (0.11 range. 1e-4 random_seed Random number generator seed. If not specified then RNG will be rehiting a quantify exponential in max_dim and o cluster_number parameters 3 sampling_heuristic The choice of heuristic method: 3 sampling_heuristic Sampling factor used in the sampling and a cluster_number of samples as a procenter form a toni may. 3 sampling_heuristic (se above, options as a proper clustristic which will sample	min dim	Low bound of a cluster manifold di-	1
max_dim High bound of a cluster manifold dimension. It cannot a larger then a dimensionality of a cluster. Re- additionality of a distance his- clusters. 10 number_of_clusters Expected number of clusters. Re- larger dipensions. 10 hist_bin_size Number of bins for a distance histogram denome is set to tarace histogram denomical by parameter is set to tarace histogram denomical by parameter max_bin_portion. 10 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a gootness of separation characterized by a discriminiability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 10 random_seed Random number generator seed. If not specified then NKG will be erinitial identify a linear quantify exponential in max_dim and cluster_number of samples reparameters 2.2 will sample free quantify exponential in max_dim and cluster_number of samples as a percentage from a too.1 3 sampling_heuristic The choice of heuristic which will use a probabilistic heuristic which will sample for the sampling for a distance histogram construction, thus insproving as a percentage from a too samples as a percentage from		mension.	
number_of_clusters remension. It cannot be larger then a dimensionality of a diasase. number_of_clusters Expected number of clusters. Required for the sampling hearistics. 10 hist_bin_size Number of bins for a distance histogram determined by parameter is set to zero, the number of bins for a distance histogram determined by parameter max_bin_portion. 0 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of separation characterized by a distrininian lanuber of samples required to correctly identify a linear mamifold cluster. Walue should be selected from a (0,1) range. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear mamifold cluster. Walue should be selected from a (0,1) range. 0.1 random_seed Random number generator seed. If no sopecified then RNG will be reinitite it dilted at every run. 3 sampling_houristic The lesser of the previous two solution of points as a proper tails in max_d tim and cluster_number parameters 3. the lesser of the previous two solution in max_d tim and cluster_number of samples as a precenter runker runker sampling_factor Sampling fact used in the sampling for a distance histogram construction, thus improving a its performation and cluster runker runker runker 1.0 sampling factor Sampling factor us	max dim	High bound of a cluster manifold di-	
anumber_of_clusters Expected number of clusters. Re- qured for the sampling heuristics. 10 hist_bin_size Number of bins for a distance histogram. If this parameter is set to tarace histogram determined by pa- rameter max_bin_portlon. 0 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used or a distance histogram. 1.0 error_bound Separation best bound value is used a minimal number of samples re- quired to correctly identify a linear manifold cluster. 1.0 max_bin_portion Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. 1.1 max_bin_portion Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. 0.1 sampling_heuristic The choice of heuristic method: taltized at every ran. 3 sampling_heuristic The choice of heuristic method: not specified then RNG will be reini- taltized at every ran. 3 sampling_heuristic Sampling fortor used in the sam- parameters 0.01 sampling_heuristics which will sample fixed number of samples as a precentage from a to- ling theusistics (see above, options tal dataset size. 0.01 histogram_sampling Turus ong a sampling for a distance histogram_sampl	mux_um	mension It cannot be larger then a	
number_of_clusters Expected number of clusters. Re- qured for the sampling heuristics. 10 hist_bin_size Number of bins for a distance histogram determined by parameter is set to zero, the number of bins in the dis- tance histogram determined by parameter max_bin_port100. 0 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of sep- aration characterized by a discrim- inability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. 1e=-4 max_bin_portion Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. 0.1 random_seed Random number generator seed. If not specified then RNG will be reini- tialized at every run. 0 sampling_heuristic The choice of heuristic method: 1. algorithm will us a prob- abilistic heuristic which will sample fact number of points 3 sampling_factor Sampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to 1a dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sampl		dimensionality of a dataset	
Initial Constraints Description for a distance histogram durameter is set to serve, the number of bins in the distance histogram durameter is set to serve, the number of bins in the distance histogram durameter is set to rameter max_bin_portion. 0 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of separation hearts bound value is used a minimal number of samples required to correctly identify a linear manifold cluster. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1.e=4 max_bin_portion Sampling eneror bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0.1 random_seed Random number generator seed. If not specified then RNG will be reinitialized at every run. 0 sampling_heuristic The choice of heuristic method: 3 sampling factor 3. will sample fixed number of samples reparators seed. If not specified then RNG will be reinitialized at every run. 3 sampling_heuristic The choice of heuristic method: 3 sampling factor 3. will sample fixed number of samples a quantity exponential in max_chin and cluster_number of samples as a precentage from a to run of samples as a precentage from a to run tatataset ixe. 0.01 <t< td=""><td>number of clusters</td><td>Expected number of clusters <i>Re</i>-</td><td>10</td></t<>	number of clusters	Expected number of clusters <i>Re</i> -	10
hist_bin_size Number of bins for a distance hists 0 hist_bin_size Number of bins for a distance hists 0 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of separation characterized by a discriminability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1e=4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0.1 random_seed Random number generator seed. If not specified then RNG will be reinitial idized at every run. 0 sampling_heuristic The choice of heuristic method: 3 will sample a quantity exponential in max_bin distrace histogram. 3 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2. & 3) to determine a number of samples reparameters 0.01 sampling_factor Sampling factor used in the sampling nor an asampling for a distance histogram. 3 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2. & 3) to determine a number of samples a garretent genost as a percentage from a to raid data	humber_or_erusters	aured for the sampling heuristics	10
ind_dim_and Funder of his in the disparameter is set to serve, the number of bins in the disparameter is set to same thistogram durameter is same to same the same transmitter is set to same thistogram construction. 0 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of a samples a quantity exponential in max_dim and cluster_number of parameters is sampling for a distance histogram from a which a sampling for a distance histogram from a which a sampling for a distance histogram from a which a sampling for a distance histogram from a which a sampling for the isogram construction for deterom from durameter depends on a cluster_number and from a sampli is perform and isogram construction. 0.01 sampling for the isogram construction fore di	hist hin size	Number of bins for a distance bis-	0
Section 1 in the partners of bits in the distance histogram determined by parameter max_bin_portion.min_cluster_sizeMinimum size of a collection of data points to be considered as a proper cluster.best_boundSeparation best bound value is used for evaluating a goodness of separation characterized by a discriminability and a depth between modes of a distance histogram.error_boundSampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster.max_bin_portionSampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. Value should be selected from a (0.1) range.random_seedRandom number generator seed. If not specified then RNG will be reinititiatized at every run.sampling_houristicThe choice of heuristic method:sampling_houristicSampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples reparametersind cluster_runder3sampling_factorSampling factor used in the sampling heuristic (see above, options 2 & 3) to determine a quarity exponential in max_dim and cluster_runder of samples as a precentage from a toria dataset size.histogram_samplingTurns on a sampling for a distance histogram from a whole dataset, the algorithm draws as mala sample for the histogram construction, turk improving a its perform ance. This parameter evalue.istogram_samplingTurns on farce and of computing the distance histogram from a whole dataset, the algorithm draws as mala sample for the histogram construction, turk improving a its perform ance. This parameter depends on a cluster_runder value.ico_d_scarchTurn on/off zero dimensional mani-fol teres to th	liist_olii_size	togram If this parameter is set to	
itance histogram determined by parameter max_bin_portion. 20 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of separation characterized by a discriminability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1e=-4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0.1 max_bin_portion Sampling error bound determines a straininal number of samples required to correctly identify a linear manifold cluster. 0.1 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0 sampling_heuristic The choice of heuristic method: 1 3 sampling_heuristic I. algorithm will use a probabilistic heuristic which will sample a quantity exponential in max_dim and eluster_number of points 3 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3 to determine a number of samples as a percentage from a total dataset size. 6 histogram_sampling Turns on a sampling for a distance histogram from a whole dataset, the algor		zero the number of hins in the dis	
nameter max_bin_portion. 20 min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of separation characterized by a discriminability and adepth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1e=4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. Value should be selected from a (0.1) range. 0.1 random_seed Random number generator seed. If not specified then RNG will be reinitialized at every nm. 0 sampling_heuristic The choice of heuristic method: 3 .add addity and be a guantity exponential in max_dim and cluster_number parameters 3 .will sample fixed number of points 3 sampling_factor Sampling factor used in the sampling the distance histogram form a whole dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram form a whole dataset, the algorithm draws a small sample for the histogram forma whole dataset, the algorithm draws a small sample for the histogram forma whole dataset, the algorithm draws a small sample for the histogram forma whole dataset, the algorithm draws a small sample for the histogram forma whole dataset, the algor		tance histogram determined by na	
min_cluster_size Minimum size of a collection of data points to be considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of separation characterized by a discriminability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1e=4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0.1 random_seed Random number generator seed. If not specified then RNG will be reinitable reinitability and cluster. 0 random_seed The choice of heuristic method: 1 sampling_heuristic The choice of heuristic which will use a probabilistic heuristic which will sample a quantity exponential in max_dim and cluster_number of points 3 sampling_factor Sampling factor used of computing heuristics (see above, options 2 & 3) to determine a number of samples a percentage from a total dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram from a whole dataset, the algorithm draws a small sample for the histogram from a whole dataset, the algorithm draws a small sample for the histogram from a whole dataset, the algorithm draws a small sample for the histogram from a whole dataset. false histogram_sampling Turns on a sampling run oruber value. <td></td> <td>ramatar may bin portion</td> <td></td>		ramatar may bin portion	
Imm_rinker_size Imministize to a considered as a proper cluster. 20 best_bound Separation best bound value is used for evaluating a goodness of separation characterized by a discriminability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0.1 random_seed Random number generator seed. If not specified then RNG will be reinitial dized at every ran. 0 sampling_heuristic The choice of heuristic method: 3 . algorithm will use a probabilistic heuristic which will sample factor used in max_dim and cluster_number parameters 3 . will sample factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a total dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram form a hotole dataset, the algorithm draws a small sample for the histogram form a hotole dataset, the algorithm draws a small sample for the histogram form a whole dataset, the algorithm draws a small sample for the histogram form a whole dataset, the algorithm draws a small sample for the histogram form a whole dataset, the algorithm draws a small sample for the histogram form a whole dataset size. false histogram_sampli ford diamber valu	min alustar siza	Minimum size of a collection of	3.0
and points to be considered as a proper cluster. best_bound Separation best bound value is used for evaluating a goodness of separation characterized by a discriminability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0.1 random_seed Random number of samples required to correctly identify a linear manifold cluster. 0 random_seed Random number generator seed. If 0 random_seed The choice of heuristic method: 3 1 1 1 3 sampling_heuristic The choice of heuristic method: 3 1 1 a quarith will use a probabilistic heuristic which will sample fixed number of points 3 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a totat dataset size. 6 histogram_sampling Turns on a sampling for a distance histogram mon a vhole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its performance. This parameter depends on a clusterer_number value. false	IIIII_clustel_size	data pointa to ha considered as a	20
best_bound Separation best bound value is used for evaluating a goodness of sep- aration characterized by a discrim- inability and a depth between modes of a distance histogram. 1.0 error_bound Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from a (0.1) range. 0.1 random_seed Random number generator seed. If not specified then RNG will be reini- tialized at every run. 0 sampling_heuristic The choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 3 sampling_factor Sampling factor used in the sam- pling heuristics (see above, options 3. the lesser of the previous two 0.01 sampling factor Sampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of parameters 0.01 histogram_sampling Turns on a sampling for a distance histogram_sampling histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value. Chapter 2. Parameters false		data points to be considered as a	
best_bound Separation best bound value is used for evaluating a goodness of sep- aration characterized by a discrim- inability and a depth between modes of a distance histogram. 1e-4 error_bound Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from a (0.1) range. 0.1 random_seed Random number generator seed. If not spectified than RNG will be reini- tialized at every run. 0 sampling_heuristic The choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dlim and cluster_number parameters 3 sampling_factor Sampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram_sampling are provential size. false histogram_sampling Turns on a sampling for a distance histogram_instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value. false zero_d_search Turn on/off zero dimensional mani- false false	heat have d	proper cluster.	1.0
Ior evaluating a goodness of sep- aration characterized by a discrim- inability and a depth between modes of a distance histogram.1e-4error_boundSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster.1e-4max_bin_portionSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from a (0,1) range.0.1random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and oluster_number parameters3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 3. the lesser of the previous two0.01sampling_factorSampling for a distance histogram_sampling0.01falsefalsefalsefor care	best_bound	Separation best bound value is used	1.0
aration characterized by a discriminality and a depth between modes of a distance histogram.1e-4error_boundSampling error bound determines quired to correctly identify a linear manifold cluster.1e-4max_bin_portionSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster.0.1max_bin_portionSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster.0.1random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_beuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters32. will sample fixed number of points 3. the lesser of the previous two0.01sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram_sampling the distance histogram from a whole dataset, the algorithm draws a small sample or samples a is perfor- mance. This parameter depends on a cluster_number value.false6Turn on/off zero dimensional mani- fold cazerbfalse		for evaluating a goodness of sep-	
error_bound Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. 1e-4 max_bin_portion Sampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from 40.1) range. 0.1 random_seed Random number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from 40.01 range. 0 sampling_heuristic The choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic the which will sample a quantity exponential in max_dlin and cluster_number parameters 3 sampling_factor Sampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size. 0.01 histogram_sampling Turns on a sampling for a distance tion, thus improving a its perfor a cluster_number value. false 6 mance. This parameter depends on a cluster_number value. false		aration characterized by a discrim-	
error_boundSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster.1e-4max_bin_portionSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from a (0.1) range.0.1random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of sampling bears to construct to a dataset size.0.01histogram_samplingTurns on a sampling for a distance tataset, the algorithm draws a small sample for the histogram. Instead of computing the distance histogram form a whole dataset, the algorithm draws a small sample for the histogram construc-false6mance. This parameter value.falsezero_d_searchTurn on/off zero dimensional mani- fold eazerbfalse		inability and a depth between modes	
error_bound Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. 0.1 max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. Value should be selected from a (0.1) range. 0.1 random_seed Random number generator seed. If not specified then RNG will be reinitid identify a linear manifold cluster. Value should be selected from a (0.1) range. 0 sampling_heuristic The choice of heuristic method: 1. algorithm will use a probabilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 3 2. will sample fixed number of points 3. the lesser of the previous two 0.01 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a total dataset, the algorithm draws a small sample for the histogram construction, thus improving a its perform a total dataset. He algorithm draws a whole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its perform a cluster_number value. false 6 mance. This parameter value. false		of a distance histogram.	
a minimal number of samples required to correctly identify a linear manifold cluster.0.1max_bin_portionSampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. Value should be selected from a (0,1) range.0.1random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 3. the lesser of the previous two0.01sampling_factorSampling factor used in the sam- pling heuristics (see above, options a 3. to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter value.falsegero_d_searchTurn on/off zero dimensional mani- fold earset.false	error_bound	Sampling error bound determines	le-4
quired to correctly identify a linear manifold cluster.max_bin_portionSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from a (0,1) range.0.1random_seedRandom number generator seed. If not specified then RNG will be reini- titalized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic withich will sample a quantity exponential in max_dim and cluster_number parameters 2. will sample fixed number of points 3. the lesser of the previous two3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram_number of asample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.false6Turn on/off zero dimensional mani- fold searchfalse		a minimal number of samples re-	
max_bin_portion Sampling error bound determines a minimal number of samples required to correctly identify a linear manifold cluster. Value should be selected from a (0,1) range. 0.1 random_seed Random number generator seed. If not specified then RNG will be reinitialized at every run. 0 sampling_heuristic The choice of heuristic method: 3 sampling_heuristic The choice of heuristic which will use a probabilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 3 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples a percentage from a total dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram. Instead of computing the distance histogram construction, thus improving a its performance. This parameter depends on a cluster_number value. false zero_d_search Turn on/off zero dimensional manified cluster depends on a cluster_number value. false		quired to correctly identify a linear	
max_bin_portionSampling error bound determines a minimal number of samples re- quired to correctly identify a linear manifold cluster. Value should be selected from a (0,1) range.0.1random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 2. will sample fixed number of points 3. the lesser of the previous two3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram_n. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tin, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- fold searchfalse		manifold cluster.	
a minimal number of samples required to correctly identify a linear manifold cluster. Value should be selected from a (0,1) range.random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 2. will sample fixed number of points 3. the lesser of the previous two3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram_sampling for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.6Chapter 2. Parameters falsezero_d_searchTurn on/off zero dimensional mani- fold searchTurn on/off zero dimensional mani- fold searchfalse	max_bin_portion	Sampling error bound determines	0.1
quired to correctly identify a linear manifold cluster. Value should be selected from a (0,1) range.random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters32. will sample fixed number of points 3. the lesser of the previous two0.01sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram_samplingfalse6mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- fold searchfalse		a minimal number of samples re-	
manifold cluster. Value should be selected from a (0,1) range.random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram_load cluster_number range and bistogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- fold searchfalse		quired to correctly identify a linear	
selected from a (0,1) range. random_seed Random number generator seed. If not specified then RNG will be reini- italized at every run. 0 sampling_heuristic The choice of heuristic method: 3 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 3 2. will sample fixed number of points 0.01 sampling_factor Sampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram_sampling false 6 macc. This parameter depends on a cluster_number value. false zero_d_search Turn on/off zero dimensional mani- fold search false		manifold cluster. Value should be	
random_seedRandom number generator seed. If not specified then RNG will be reini- tialized at every run.0sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 2. will sample fixed number of points 3. the lesser of the previous two3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- fold searchfalse		selected from $a(0,1)$ range.	
not specified then RNG will be reinitialized at every run.sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 2. will sample fixed number of points 3. the lesser of the previous two3sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram_for the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- fold searchfalse	random_seed	Random number generator seed. If	0
tialized at every run.sampling_heuristicThe choice of heuristic method: 1. algorithm will use a prob- abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters3algorithm will use a prob- abilistic heuristic which will sample fixed number of points3algorithm will sample fixed number of points0.01algorithm euristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.6zero_d_searchTurn on/off zero dimensional mani- fold caracefalse		not specified then RNG will be reini-	
sampling_heuristic The choice of heuristic method: 3 1. algorithm will use a probabilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters 3 2. will sample fixed number of points 3. the lesser of the previous two 0.01 sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a total dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its performance. This parameter depends on a cluster_number value. false zero_d_search Turn on/off zero dimensional manifold eagree false		tialized at every run.	
1. algorithm will use a probabilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parameters2. will sample fixed number of points3. the lesser of the previous twosampling_factorSampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.histogram_samplingTurns on a sampling for a distance histogram_samplingTurns on a sampling for a distance histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.zero_d_searchTurn on/off zero dimensional mani- fol search	sampling_heuristic	The choice of heuristic method:	3
abilistic heuristic which will sample a quantity exponential in max_dim and cluster_number parametersquantity exponential in max_dim and cluster_number parameters2. will sample fixed number of points2. will sample fixed number of points0.01sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- falsefalse		1. algorithm will use a prob-	
will sample a quantity exponential in max_dim and cluster_number parameters2. will sample fixed number of points3. the lesser of the previous twosampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset sizehistogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- falsefalse		abilistic heuristic which	
exponential in max_dim and cluster_number parameters2. will sample fixed number of points3. the lesser of the previous twosampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.zero_d_searchTurn on/off zero dimensional mani- fold search		will sample a quantity	
andcluster_number parametersandcluster_number parameters2. will sample fixed number of points3. the lesser of the previous twosampling_factorSampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- fold casarchfalse		exponential in max dim	
parameters 2. will sample fixed number of points 3. the lesser of the previous two0.01sampling_factorSampling factor used in the sam- pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.0.01histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.falsezero_d_searchTurn on/off zero dimensional mani- fold searchfalse		and cluster number	
2. will sample fixed number of points3. the lesser of the previous twosampling_factorSampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.zero_d_searchTurn on/off zero dimensional mani- fold search		parameters	
points3. the lesser of the previous twosampling_factorSampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size.histogram_samplingTurns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value.zero_d_searchTurn on/off zero dimensional mani- fold search		2. will sample fixed number of	
3. the lesser of the previous two sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a total dataset size. histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its performance. This parameter depends on a cluster_number value. zero_d_search		points	
sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a total dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its performance. This parameter depends on a cluster_number value. false cluster_number value. Chapter 2. Parameters fold carers false false		3. the lesser of the previous two	
sampling_factor Sampling factor used in the sampling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a total dataset size. 0.01 histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its performance. This parameter depends on a cluster_number value. false zero_d_search Turn on/off zero dimensional manifield search false		1	
pling heuristics (see above, options 2 & 3) to determine a number of samples as a percentage from a to- tal dataset size. histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its performance. This parameter depends on a cluster_number value. zero_d_search	sampling_factor	Sampling factor used in the sam-	0.01
1 & 3) to determine a number of samples as a percentage from a total dataset size. histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its performance. This parameter depends on a cluster_number value. zero_d_search Turn on/off zero dimensional manifold search	1 C-	pling heuristics (see above, options	
samples as a percentage from a to- tal dataset size. histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value. false Zero_d_search Turn on/off zero dimensional mani- fold search false		2 & 3 to determine a number of	
histogram_sampling Turns on a sampling for a distance histogram. Instead of computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value. false 6 Turn on/off zero dimensional mani- fold search false		samples as a percentage from a to-	
histogram_sampling Turns on a sampling for a distance false histogram. Instead of computing the distance histogram from a whole false dataset, the algorithm draws a small sample for the histogram construc- for the histogram construc- 6 mance. This parameter depends on a cluster_number value. Chapter 2. Parameters zero_d_search Turn on/off zero dimensional mani- false		tal dataset size.	
and game in the stand of a computing for a construction is to a computing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construction, thus improving a its performance. This parameter depends on a cluster_number value. Chapter 2. Parameters a cluster_number value. Turn on/off zero dimensional manifield search false	histogram sampling	Turns on a sampling for a distance	false
analogium Instant of comparing the distance histogram from a whole dataset, the algorithm draws a small sample for the histogram construc- tion, thus improving a its perfor- mance. This parameter depends on a cluster_number value. 6 Chapter 2. Parameters a cluster_number value. zero_d_search Turn on/off zero dimensional mani- fold search		histogram. Instead of computing	
attack		the distance histogram from a whole	
6 a cluster_number value. zero_d_search Turn on/off zero dimensional mani- fold search		dataset the algorithm draws a small	
6 Chapter 2. Parameters zero_d_search Turn on/off zero dimensional mani- fold search		sample for the histogram construc-	
6 Instrume a rus perior mance. This parameter depends on a cluster_number value. Chapter 2. Parameters zero_d_search Turn on/off zero dimensional mani- fold search false		tion thus improving a its perfor-	
a cluster_number value. zero_d_search Turn on/off zero dimensional mani- fold search	6	mance. This parameter depends on	Chapter 2. Parameters
zero_d_search Turn on/off zero dimensional mani- fold search false		a cluster number value	•
fold search	zero d search	Turn on/off zero dimensional mani	false
	Zero_u_searen	fold search	T0796

2.1 Suggestions

Particular settings could impact performance of the algorithm:

- If you want a persistent clustering results fix a random_seed parameter. By default, RNG is reinitialized every time when algorithm runs.
- If a dimensionality of the data is low, a histogram sampling could speeding up calculations.
- Value 1 of sampling_heuristic parameter should not be used if parameter max_dim is large, as it will generate a very large number of samples.
- Increasing value of max_bin_portion parameter could improve an efficiency of the clustering partitioning, but as well could degrade overall performance of the algorithm.

2.2 Parallelization

This implementation of LMCLUS algorithm uses parallel computations during a manifold sampling stage. You need add additional workers before executing the algorithm.

Separations

When linear manifold is formed, a distance from every point of dataset to the manifold is calculated, and a histograms of point distances to each trial manifold are computed. If the resulting histogram contains multiple modes then the mode near zero is isolated in histogram¹. The isolated part of histogram is used to determine a separation criteria, and the data points are partitioned from the rest of the dataset on the basis of such separation.

The separation properties defined in Separation instance, which is defined as follows:

```
type Separation
   depth::Float64
                                # Separation depth (depth between separated_
→histogram modes)
   discriminability::Float64
                                # Separation discriminability (width between
→ separated histogram modes)
   threshold::Float64
                                # Distance threshold value
   globalmin::Int
                                # Global minimum as histogram bin index
   hist_range::Vector{Float64} # Histogram ranges
   hist_count::Vector{UInt32} # Histogram counts
   bin_index::Vector{UInt32}
                                # Point to bin assignments
end
```

Separation criteria and distance threshold value can be accessed through following functions:

criteria(S)

Returns separation criteria value which is product of depth and discriminability.

threshold(S)

Returns distance threshold value for separation calculated on histogram of distances. It is used to determine which points belong to formed cluster.

References

1

^{10.} Kittler & J. Illingworth: "Minimum Error Thresholding", Pattern Recognition, Vol 19, nr 1. 1986, pp. 41-47, DOI:10.1016/0031-3203(86)90030-0

Utilities

Linear Manifold Clustering Algorithm relies on multiple search and optimization methods:

kittler (*X*, *bins*, *tol*)

A minimum error thresholding method for multimodal histograms¹.

otsu(X, bins)

A gray-level thresholding method for multimodal histograms².

mdl(M, X; Pm = 32, Pd = 16, T = :Empirical, = 1e-4)

Performs calculation of the minimum description length for the linear manifold cluster.

Parameters

- M Linear manifold cluster description as Manifold type instance.
- X Linear manifold cluster data as Matrix with points as its columns.
- **Pm** Precision encoding constant for the model, i.e. number of bits required for encoding on element of the model description. Default value is 32 which corresponds to Float32.
- Pd Precision encoding constant for the data.
- **T** Type of a dataset encoding model as symbol: :Gausian, :Uniform, :Empirical.
- - Error tolerance for bin quantization used in an empirical model encoding

Returns number of bits required to encode linear manifold cluster with the MDL schema.

¹

^{10.} Kittler & J. Illingworth: "Minimum Error Thresholding", Pattern Recognition, Vol 19, nr 1. 1986, pp. 41-47, DOI:10.1016/0031-3203(86)90030-0

²

^{14.} Otsu: "A threshold selection method from gray-level histograms", Automatica, 1975, 11, 285-296, DOI:10.1109/TSMC.1979.4310076

References

Notes:

All methods implemented in this package adopt the column-major convention: in a data matrix, each column corresponds to a sample/observation, while each row corresponds to a feature (variable or attribute).

References

Index

С

criteria() (built-in function), 9

I

indim() (built-in function), 3

Κ

kittler() (built-in function), 11

L

lmclus() (built-in function), 3

Μ

mdl() (built-in function), 11 mean() (built-in function), 4

0

otsu() (built-in function), 11 outdim() (built-in function), 3

Ρ

points() (built-in function), 4
projection() (built-in function), 4

S

separation() (built-in function), 4
size() (built-in function), 4

Т

threshold() (built-in function), 9