Package ‘ider’
February 10, 2023

Type  Package
Title  Various Methods for Estimating Intrinsic Dimension
Version  0.1.1
Date  2023-02-10
Author  Hideitsu Hino
Maintainer  Hideitsu Hino <hideitsu.hino@gmail.com>
Depends  R (>= 4.2.0)
Description  An implementation of various methods for estimating intrinsic
dimension of vector-valued dataset or distance matrix. Most methods implemented
are based on different notion of fractal dimension such as the capacity
dimension, the box-counting dimension, and the information dimension.
License  GPL-2
Imports  FNN, stats, glm2
RoxygenNote  7.2.2
LazyData  true
NeedsCompilation  no
Repository  CRAN
Date/Publication  2023-02-10 10:20:02 UTC

R topics documented:

ider-package ............................................. 2
convU .................................................. 3
corint .................................................. 4
gendata ................................................ 5
handD ................................................... 7
lbmle .................................................... 8
mada ..................................................... 9
nni ....................................................... 10
pack ...................................................... 11
side ...................................................... 12

Index  14
**Description**

This package is used for estimating intrinsic dimension of a given dataset.

**Details**

In common data analysis situations, an observed datum is expressed by a p-dimensional vector. In general, the apparent data dimension \( p \) and its intrinsic dimension \( d \) are different. A basic assumption in many data analysis and machine learning methods is that the intrinsic dimension is low even when the apparent dimension is high and the data distribution is constrained onto a low dimensional manifold. Examples of such methods include manifold learning, subspace methods, and visualization and dimensionality reduction methods. The key to the success of dimensionality reduction, manifold learning and latent variable analysis lies in the accurate estimation of the intrinsic dimension of the dataset at hand. This package implements a number of intrinsic dimension estimation methods. Some functions are for estimating the global intrinsic dimension while others are capable of estimating both local and global intrinsic dimension.

The package has functions `corint`, `convU`, `lbmle`, `nni`, `pack` for estimating global intrinsic dimensions, and `mada`, `side` for estimating local intrinsic dimensions. A data generator `gendata` is included in the package.

**Author(s)**

Hideitsu Hino <hideitsu.hino@gmail.com>

**References**

convU

Intrinsic Dimension Estimation with Convergence Property of a U-statistics.

Description

convU estimates intrinsic dimension of given dataset based on the convergence property of U-statistics (smoothed correlation dimension) w.r.t. kernel bandwidth.

Usage

convU(x, maxDim = 5, DM = FALSE)
Arguments

\(x\)  
data matrix or distance matrix given by as.matrix(dist(x)).

\(\text{maxDim}\)  
maximum of the candidate dimension.

\(\text{DM}\)  
whether 'x' is distance matrix or not. logical.

Details

A variant of fractal dimension called the correlation dimension is considered. The correlation dimension is defined by the notion of the correlation integral, which is calculated by counting the number of pairs closer than certain threshold \(\epsilon\). The counting operation is replaced with the kernel smoothed version, and based on the convergence property of the resulting U-statistics, an intrinsic dimension estimator is derived.

Value

Estimated global intrinsic dimension.

Author(s)

Hideitsu Hino <hideitsu.hino@gmail.com>

References


Examples

```r
x <- gendata(DataName=’SwissRoll’,n=300)
estconvU <- convU(x=x)
print(estconvU)
```

---

**corint**

Intrinsic Dimension Estimation with Correlation Integral

Description

corint estimates intrinsic dimension of given dataset based on the correlation integral

Usage

corint(x, k1 = NULL, k2 = NULL, DM = FALSE, p = NULL)
Arguments

- **x**: data matrix or distance matrix given by `as.matrix(dist(x))`.
- **k1**: first k-NN parameter.
- **k2**: second k-NN parameter.
- **DM**: whether 'x' is distance matrix or not. logical.
- **p**: ambient dimension used for automatically define 'k1' and 'k2'.

Details

A variant of fractal dimension called the correlation dimension is considered. The correlation dimension is defined by the notion of the correlation integral, is calculated by using the power low for the definition of the correlation dimension.

Value

Estimated global intrinsic dimension.

Author(s)

Hideitsu Hino <hideitsu.hino@gmail.com>

References


Examples

```r
x <- gendata(DataName = 'SwissRoll', n = 300)
estcorint <- corint(x = x, k1 = 5, k2 = 10)
print(estcorint)
```

Description

gendata generates various artificial datasets for intrinsic dimension estimation experiments.

Usage

```r
gendata(
    DataName = "SwissRoll",
    n = 300,
    p = NULL,
    noise = NULL,
    ol = NULL,
)```
curv = 1,
seed = 123,
sorted = FALSE
)

Arguments

DataName  Name of dataset, one of the following:
- SwissRoll: SwissRoll data, 2D manifold in 3D space.
- NDSwissRoll: Non-deformable SwissRoll data, 2D manifold in 3D space.
- Moebius: Moebius strip, 2D manifold in 3D space.
- SphericalShell: Spherical Shell, (p-1)-dimensional manifold in p-dimensional space.
- Sinusoidal: Sinusoidal data, 1D manifold in 3D space.
- Spiral: Spiral-shaped 1D manifold in 2D space.
- Cylinder: Cylinder-shaped 2D manifold in 3D space.
- SShape: S-shaped 2D manifold in 3D space.
- ldbl: LDB(line - disc - filled ball - line), embedded in 3D space (original dataset).

n  number of data points to be generated.

p  ambient dimension of the dataset.

noise  parameter to control noise level in the dataset. In many cases, it is used for sd of rnorm used inside the function.

ol  percentage of outliers, i.e., n * ol outliers are added to the generated dataset.

curv  a parameter to control the complexity of the embedded manifold.

seed  random number seed.

sorted  logical. If TRUE, the index of the generated dataset is sorted with respect to x-axis for the ease of visualization.

Details

This function generates various artificial datasets often used in manifold learning and dimension estimation researches. For some datasets, complexity of the shape is controlled by the parameter curv. The parameters noise and outlier are used for adding noise and/or outliers for the dataset.

Value

Data matrix. For ldbl dataset, it outputs a list composed of x: data matrix and tDim: true intrinsic dimension for each point.

Author(s)

Hideitsu Hino <hideitsu.hino@gmail.com>
Examples

```r
## global intrinsic dimension estimate
x <- gendata(DataName='SwissRoll')
estmle <- lbmle(x=x,k1=3,k2=5)
print(estmle)

## local intrinsic dimension estimate
tmp <- gendata(DataName='ldbl',n=1000)
x <- tmp$x
estmada <- mada(x=x,local=TRUE)
head(estmada)  ## estimated local intrinsic dimensions
head(tmp$tDim)  ## true local intrinsic dimensions
```

handD

### Hand ratation data

Description

Data from a QTL experiment on gravitropism in Arabidopsis, with data on 162 recombinant inbred lines (Ler x Cvi). The outcome is the root tip angle (in degrees) at two-minute increments over eight hours.

Usage

```r
data(handD)
```

Format

An object of class 'dist'.

References


H. Hino, J. Fujiki, S. Akaho, and N. Murata. 'Local Intrinsic Dimension Estimation by Generalized Linear Modeling', Neural Computation, 2017

Examples

```r
data(handD)
estmle <- lbmle(x=handD,DM=TRUE,k1=5,k2=10)
print(estmle)
```
lbmle

Maximum Likelihood Estimation of Intrinsic Dimension.

Description

lbmle estimate the intrinsic dimension of a given dataset.

Usage

lbmle(x = NULL, k1 = NULL, k2 = NULL, BC = TRUE, DM = FALSE, p = NULL)

Arguments

x data matrix or distance matrix given by as.matrix(dist(x)).
k1 first k-NN parameter.
k2 second k-NN parameter.
BC whether bias is corrected or not. logical.
DM whether 'x' is distance matrix or not. logical.
p ambient dimension used for automatically define 'k1' and 'k2'.

Details

The likelihood of the rate parameter of the Poisson process, which characterize the behaviour of
the distance from a point to another point in the given dataset, is considered, and the maximum
likelihood estimator (MLE) for the intrinsic dimension is derived. The original method proposed
by Levina and Bickel contains a known bias, and it is corrected by Mackay and Ghahramani. This
function implements both, with the default the bias corrected estimate.

Value

Estimated global intrinsic dimension.

Author(s)

Hideitsu Hino <hideitsu.hino@gmail.com>

References

E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. Advances in

Examples

x <- gendata(DataName='SwissRoll',n=300)
estmle <- lbmle(x=x,k1=3,k2=5)
print(estmle)
mada

**Manifold-Adaptive Local Dimension Estimation.**

**Description**

mada estimates local information dimension of given dataset based on the first order expansion of probability mass function.

**Usage**

mada(x, k = NULL, comb = "average", DM = FALSE, local = FALSE, maxDim = 5)

**Arguments**

- **x**: data matrix or distance matrix given by as.matrix(dist(x)).
- **k**: k-NN parameter.
- **comb**: 'average', 'median' or 'vote' for combining local estimates when global estimate is required.
- **DM**: whether 'x' is distance matrix or not. logical.
- **local**: logical. If TRUE, a vector of local dimensions at each sample point is returned.
- **maxDim**: maximum of the candidate dimensions.

**Details**

A variant of fractal dimension called the local information dimension is considered. The local information dimension is estimated by using the probability mass function. The function mada considers first order expansion of the probability mass around the inspection point, and it estimates the local information dimension by using two different radii from the inspection point.

**Value**

Estimated local or global intrinsic dimension.

**Author(s)**

Hideitsu Hino <hideitsu.hino@gmail.com>

**References**

Examples

```r
## local intrinsic dimension estimate
tmp <- gendata(DataName='ldbl',n=300)
x <- tmp$x
estmada <- mada(x=x,local=TRUE)
head(estmada) ## estimated local intrinsic dimensions by mada
head(tmp$tDim) ## true local intrinsic dimensions
```

### nni

**Intrinsic Dimensionality Estimation from Near-Neighbor Information.**

**Description**

nni estimates intrinsic dimension of given dataset based on the nearest-neighbor information.

**Usage**

```r
nni(x, k1 = 2, k2 = 30, DM = FALSE, eps = 0.01, p = NULL)
```

**Arguments**

- `x`: data matrix or distance matrix given by as.matrix(dist(x)).
- `k1`: first k-NN parameter.
- `k2`: second k-NN parameter.
- `DM`: whether 'x' is distance matrix or not. logical.
- `eps`: accuracy parameter.
- `p`: ambient dimension used for automatically define 'k1' and 'k2'.

**Details**

First order expansion of the probability mass function is considered, then the distribution of the nearest-neighbor points from the inspection point is modeled by the Poisson distribution. The average of the nearest-distance is expressed by intrinsic dimension to be estimated.

**Value**

Estimated global intrinsic dimension.

**Author(s)**

Hideitsu Hino <hideitsu.hino@gmail.com>

**References**


Examples

```r
x <- gendata(DataName='SwissRoll', n=300)
estnni <- nni(x=x)
print(estnni)
```

pack

Intrinsic Dimension Estimation Using Packing Numbers.

Description

pack estimates intrinsic dimension of given dataset based on the packing number.

Usage

```r
pack(x, k1 = NULL, k2 = NULL, greedy = TRUE, eps = 0.01, DM = FALSE)
```

Arguments

- `x`: data matrix or distance matrix given by `as.matrix(dist(x))`.
- `k1`: first radius parameter. If one of `k1` or `k2` is `NULL`, then both are automatically determined from the input data.
- `k2`: second radius parameter.
- `greedy`: logical. If `TRUE`, then a greedy algorithm is used for estimating the packing number. If `FALSE`, then a hierarchical clustering algorithm is used instead.
- `eps`: accuracy parameter.
- `DM`: whether `'x'` is distance matrix or not. logical.

Details

A variant of fractal dimension called the capacity dimension is considered. The capacity dimension is defined by using the notion of covering number, which is hard to calculate in general. In this function, the packing number of the data space is used as the surrogate of the covering number. The packing number is estimated by greedy manner or by hierarchical clustering.

Value

Estimated global intrinsic dimension.

Author(s)

Hideitsu Hino <hideitsu.hino@gmail.com>

References


Examples

```r
x <- gendata(DataName='SwissRoll',n=300)
estpackG <- pack(x=x,greedy=TRUE)  ## estimate the packing number by greedy method
print(estpackG)
estpackC <- pack(x=x,greedy=FALSE) ## estimate the packing number by cluttering
print(estpackC)
```

**Description**

`side` is a Higher-order Information Dimension Estimator, which estimates local information dimension of given dataset based on the polynomial regression with Poisson error structure.

**Usage**

```r
side(
x,
maxDim = 5,
DM = FALSE,
local = FALSE,
method = "disc",
comb = "average"
)
```

**Arguments**

- `x`  
  data matrix or distance matrix given by as.matrix(dist(x)).
- `maxDim`  
  maximum of the candidate dimensions.
- `DM`  
  whether ‘x’ is distance matrix or not. logical.
- `local`  
  logical. If TRUE, a vector of local dimensions at each sample point is returned.
- `method`  
  algorithm to estimate intrinsic dimension. ‘disc’ for discrete dimension estimation. ‘cont’ for continuous dimension estimation with MLE by Newton-method.
- `comb`  
  ‘average’, ‘median’ or ‘vote’ for combining local estimates when global estimate is required.

**Details**

A variant of fractal dimension called the local information dimension is considered. The local information dimension is estimated by using the probability mass function. The function `side` considers higher-order expansion of the probability mass around the inspection point, and it estimates the local information dimension by fitting a generalized linear model with Poisson error structure and an identity link function. There are two methods for dimension estimation: the first method tries different dimensions and adopt the one with maximum likelihood, while the second method directly maximises the likelihood with respect to the intrinsic dimension. The former returns an
integer-valued dimension estimate, and the latter returns a real-valued estimate. The result of the former method is used as an initial value for the latter method in numerical optimization. Slow but more accurate than mada in some cases.

Value

Estimated local or global intrinsic dimension.

Author(s)

Hideitsu Hino <hideitsu.hino@gmail.com>

References

H. Hino, J. Fujiki, S. Akaho, and N. Murata. 'Local Intrinsic Dimension Estimation by Generalized Linear Modeling', Neural Computation, 2017

Examples

```r
## local intrinsic dimension estimate
tmp <- gendata(DataName='ldbl', n=300)
x <- tmp$x
set.seed(999)
idx <- c(sample(which(tmp$tDim==1)[1:10],3), sample(which(tmp$tDim==2)[1:30],3))
estmada <- mada(x=x[1:100,], local=TRUE)
estmada[idx] ## estimated local intrinsic dimensions by mada
tmp$tDim[idx] ## true local intrinsic dimensions
estside <- side(x=x[1:100,], local=TRUE)
estside[idx] ## estimated local intrinsic dimensions by side
```
Index

* datasets
  handD, 7

convU, 3

corint, 4

gendata, 5

handD, 7

ider (ider-package), 2
ider-package, 2

lbmle, 8

mada, 9

nni, 10

pack, 11

side, 12