Package 'distde'

July 31, 2014

Type Package

Title Dimension Estimators using Distances Between Points

Version 1.2

Date 2013-10-17

Author Kerstin Johnsson, Lund University

Maintainer Kerstin Johnsson <johnsson@maths.lth.se>

Depends manifgen (>= 1.0-1), yaImpute

Description Provides functions to do manifold dimension estimation with either translated Poisson distributions (a generalization of the Hill estimator of correlation dimension) or the Beardwood-Halton-Hammersley theorem.

License GPL (>=2)

URL http://www.maths.lu.se/staff/kerstin-johnsson/research/manifold-dimension-estimation/

LazyLoad yes

R topics documented:

kNN *Dimension Estimation with kNN Method*

Description

Estimates the intrinsic dimension of a data set using weighted average kNN distances.

Usage

```
kNN(data, k, ps, M, gamma = 2)
```

Arguments

`data`	data set with each row describing a data point.
`k`	number of distances to neighbors used at a time.
`ps`	vector with sample sizes; each sample size has to be larger than k and smaller than `nrow(data)`.
`M`	number of bootstrap samples for each sample size.
`gamma`	weighting constant.

Details

This is a somewhat simplified version of the kNN dimension estimation method described by Carter et al. (2010), the difference being that block bootstrapping is not used.

Value

A vector with two components:

`de`	the intrinsic dimension estimate (integer).
`residual`	the residual, see Carter et al. (2010).

Author(s)

Kerstin Johnsson, Lund University.

References

Carter, K.M., Raich, R. and Hero, A.O. (2010) On local intrinsic dimension estimation and its applications. *IEEE Trans. on Sig. Proc.*, **58**(2), 650-663.

Johnsson, K. (2011) Manifold dimension estimation for omics data analysis: Current methods and a novel approach. Master's Thesis, Lund University.

Examples

```
library(manifgen)

N <- 50
data <- hball(N, 5)

k <- 2
ps <- seq(max(k + 1, round(N/2)), N - 1, by = 3)
kNN(data, k, ps, M = 10, gamma = 2)
```

Noisefun *Transition Functions Describing Noise*

Description

Transition functions $f(s|r)$ describing the shift in lengths of vectors when Gaussian noise is added. Given a length r, $f(s|r)$ is the probability density for the length after noise is added to one endpoint.

Usage

```
dnoiseNcChi(r, s, sigma, k)
dnoiseGaussH(r, s, sigma, k)
dnoiseGaussB(r, s, sigma, k)
```

Arguments

r	length or vector of lengths of original vector.
s	length or vector of lengths of perturbed vector.
sigma	noise standard deviation.
k	dimension of noise.

Details

dnoiseNcChi is the true transition function density when the noise is Gaussian, the other transition functions are approximations of this. dnoiseGaussH is the Gaussian approximation used in Haro et al. and dnoiseGaussB is the best Gaussian approximation.

If Gaussian noise is added to both endpoints of the vector, sigma should be replaced by sqrt(2)*sigma.

Value

Vector of probability densities.

Note

Only r or s can be a vector.

Author(s)

Kerstin Johnsson, Lund University

References

Haro, G., Randall, G. and Sapiro, G. (2008) Translated Poisson Mixture Model for Stratification Learning. *Int. J. Comput. Vis.*, **80**, 358-374.

Examples

```
# High SNR, high-dim noise
sigma <- 0.05
x <- seq(0, 1.5, length.out = 200)
y <- dnoiseNcChi(x, s = .5, sigma, k = 20)
plot(x, y, type = 'l', main = 'Noise dim = 20')
y2 <- dnoiseGaussB(x, s = .5, sigma, k = 20)
lines(x, y2, col = 'red')
y3 <- dnoiseGaussH(x, s = .5, sigma, k = 20)
lines(x, y3, lty = 2)

# Low SNR
par(mfrow = c(2, 3))
sigma <- 0.2
x <- seq(0, 1.5, length.out = 200)
y <- dnoiseNcChi(x, s = .5, sigma, k = 4)
plot(x, y, type = 'l', main = 'Noise approximations')
y2 <- dnoiseGaussB(x, s = .5, sigma, k = 4)
lines(x, y2, col = 'red')
y3 <- dnoiseGaussH(x, s = .5, sigma, k)
lines(x, y3, lty = 2)

# High SNR, low-dim noise
sigma <- 0.05
x <- seq(0, 1.5, length.out = 200)
y <- dnoiseNcChi(x, s = .5, sigma, k = 4)
plot(x, y, type = 'l', main = 'Noise dim = 4')
y2 <- dnoiseGaussB(x, s = .5, sigma, k = 4)
lines(x, y2, col = 'red')
y3 <- dnoiseGaussH(x, s = .5, sigma, k)
lines(x, y3, lty = 2)
```

---

tp *Dimension Estimation via Translated Poisson Distributions*

Description

Estimates the intrinsic dimension of a data set using models of translated Poisson distributions. `tp` employs an approximation to avoid iteration (see Haro el al. (2008)), `tp_it` does five iterations to get the result.

Usage

```
tp(data, indexes = 1:dim(data)[1], k, dnoise, sigma, n,
   loc = FALSE, verbose = FALSE)
tp2(data, indexes = 1:dim(data)[1], k, dnoise, sigma, n,
    loc = FALSE, verbose = FALSE)
```

```
tp_it(data, indexes = 1:dim(data)[1], k, dnoise, sigma, n,
      init = 5, loc = FALSE, verbose = FALSE)
```

Arguments

data	data set with each row describing a data point.
indexes	the indexes of the data points for which local dimension estimation should be made.
k	the number of neighbors that should be used for dimension estimation for each point.
dnoise	a function or a name of a function giving the translation density.
sigma	(estimated) standard deviation of the (isotropic) noise.
n	dimension of the noise.
init	initial value for iterative procedure.
loc	should local estimates be returned?
verbose	should intermediate estimates be printed out?

Details

tp uses the same approximation as Haro et al. (2008) of a certain integral, tp2 multiplies the integrand with r, which guarantees that the integral converges. However, the approximation in tp2 does not result in error terms that are symmetric around r as in tp.

Value

If loc = FALSE a vector with two components:

de	the overall dimension estimate
likelihood	this is NA in the current implementation.

If loc = FALSE a matrix with two columns corresponding to the two components above. Row i corresponds to local dimension estimate at point data[indexes[i],].

Author(s)

Kerstin Johnsson, Lund University.

References

Haro, G., Randall, G. and Sapiro, G. (2008) Translated Poisson Mixture Model for Stratification Learning. *Int. J. Comput. Vis.*, **80**, 358-374.

See Also

tp_glob, tp_loc

Examples

```
require(manifgen)
data <- hball(1000, d = 7, n = 13, sd = 0.01)
tp(data, 1:100, 10, dnoiseGaussH, 0.01, 13)
tp(data, 1:100, 10, dnoiseNcChi, 0.01, 13)
tp2(data, 1:100, 10, dnoiseGaussH, 0.01, 13)
tp2(data, 1:100, 10, dnoiseNcChi, 0.01, 13)
tp_it(data, 1:100, 10, dnoiseGaussH, 0.01, 13)
tp_it(data, 1:100, 10, dnoiseNcChi, 0.01, 13)
```

---

tp_glob *Dimension Estimation via Translated Poisson Distributions*

Description

Estimates the intrinsic dimension of a data set using models of translated Poisson distributions. `tp` and `tp2` employs an approximation to avoid iteration (see Haro el al. (2008)), `tp_it` does five iterations to get the result. However, opposed to Haro et al. (2008) these functions uses distances between data points in the whole data set instead of local dimension estimation.

Usage

```
tp_glob(data, k, dnoise, sigma, n)
tp2_glob(data, k, dnoise, sigma, n)
tp_itglob(data, k, dnoise, sigma, n, init = 5)
```

Arguments

`data`	data set with each row describing a data point.
`k`	the number of distances between data points that should be used for dimension estimation.
`dnoise`	a function or a name of a function giving the transition density.
`sigma`	(estimated) standard deviation of the (isotropic) noise.
`n`	dimension of the noise.
`init`	initial value for iterative procedure.

Details

`tp_glob` uses the same approximation as Haro et al. (2008) of a certain integral, `tp2_glob` multiplies the integrand with `r`, which guarantees that the integral converges. However, the approximation in `tp2_glob` does not result in error terms that are symmetric around `r` as in `tp_glob`.

Value

A vector with two components:

`de`	the dimension estimate.
`likelihood`	this is `NA` in the current implementation.

Author(s)

Kerstin Johnsson, Lund University

References

Haro, G., Randall, G. and Sapiro, G. (2008) Translated Poisson Mixture Model for Stratification Learning. *Int. J. Comput. Vis.*, **80**, 358-374.

See Also

tp, tp_loc

Examples

```
require(manifgen)
data <- hball(1000, d = 7, n = 13, sd = 0.01)
tp_glob(data, 500, dnoiseGaussH, 0.01, 13)
tp_glob(data, 500, dnoiseNcChi, 0.01, 13)
tp2_glob(data, 500, dnoiseGaussH, 0.01, 13)
tp2_glob(data, 500, dnoiseNcChi, 0.01, 13)
tp_itglob(data, 500, dnoiseGaussH, 0.01, 13)
tp_itglob(data, 500, dnoiseNcChi, 0.01, 13)
```

---

tp_loc *Local Dimension Estimation via Translated Poisson Distributions*

Description

Performs local dimension estimation of a neighborhood centered at the origin, using models of translated Poisson distributions.

Usage

```
tp_loc(data, dnoise, sigma, n)
tp2_loc(data, dnoise, sigma, n)
```

Arguments

data	data set with each row describing a data point.
dnoise	a function or a name of a function giving the translation density.
sigma	(estimated) standard deviation of the (isotropic) noise.
n	dimension of the noise.

Details

tp_loc uses the same approximation as Haro et al. (2008) of a certain integral, tp2_loc multiplies the integrand with r, which guarantees that the integral converges. However, the approximation in tp2_loc does not result in error terms that are symmetric around r as in tp_loc.

Value

A vector with two components:

de	the dimension estimate.
likelihood	this is NA in the current implementation.

Author(s)

Kerstin Johnsson, Lund University

References

Haro, G., Randall, G. and Sapiro, G. (2008) Translated Poisson Mixture Model for Stratification Learning. *Int. J. Comput. Vis.*, **80**, 358-374.

See Also

tp, tp_glob

Examples

```
require(manifgen)
data <- cuthplane(50, d = 7, n = 13, sd = 0.01)
tp_loc(data, dnoiseNcChi, 0.1, 3)
tp_loc(data, dnoiseGaussH, 0.1, 3)
tp2_loc(data, dnoiseNcChi, 0.1, 3)
tp2_loc(data, dnoiseGaussH, 0.1, 3)
```

Index