

Package ‘LPCM’

July 2, 2014

Type Package

Title Local principal curve methods

Version 0.44-8

Date 2013-11-16

Author Jochen Einbeck and Ludger Evers

Maintainer Jochen Einbeck <jochen.einbeck@durham.ac.uk>

Depends R (>= 2.10)

Suggests scatterplot3d, lattice, dr

Description Fitting multivariate data patterns with local principal curves; including simple tools for data compression (projection), bandwidth selection, and measuring goodness-of-fit.

License GPL (>= 2)

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2013-11-16 17:28:18

R topics documented:

LPCM-package	2
calspeedflow	3
coverage	4
followx	7
gaia	8
gvessel	10
kernels.and.distances	11
lpc	12
lpc.control	14

<code>lpc.project</code>	16
<code>lpc.spline</code>	17
<code>lpc.spline.auxiliary.functions</code>	19
<code>ms</code>	20
<code>plot.lpc</code>	22
<code>print.lpc</code>	24
<code>Rc</code>	25
<code>unscale</code>	27

Index	28
--------------	-----------

LPCM-package	<i>Local principal curve methods</i>
--------------	--------------------------------------

Description

Fitting multivariate data patterns with local principal curves; including simple tools for data compression (projection), bandwidth selection, and measuring goodness-of-fit.

This package implements the techniques introduced in Einbeck, Tutz & Evers (2005), and successive related papers.

The main functions to be called by the user are

- `lpc`, for the estimation of the local centers of mass which make up the principal curve;
- `lpc.spline`, which is a smooth and fully parametrized cubic spline representation of the latter;
- `lpc.project`, which enables to compress data by projecting them orthogonally onto the curve;
- `lpc.coverage` and `Rc` for assessing goodness-of-fit;
- `lpc.self.coverage` for bandwidth selection;
- the generic `plot` and `print` methods for objects of class `lpc` and `lpc.spline`.

This package also contains some code for density mode detection ('local principal points') and mean shift clustering (as well as bandwidth selection in this context), which implements the methods presented in Einbeck (2011). See the help file for `ms`.

A second R package which will implement the extension of local principal curves to local principal surfaces and manifolds, as proposed in Einbeck, Evers & Powell (2010), is in preparation.

Details

Package:	LPCM
Type:	Package
License:	GPL (>=2)
LazyLoad:	yes

Acknowledgements

Contributions (in form of pieces of code, or useful suggestions for improvements) by Jo Dwyer, Mohammad Zayed, and Ben Oakley are gratefully acknowledged.

Author(s)

Jochen Einbeck and Ludger Evers

Maintainer: Jochen Einbeck <jochen.einbeck@durham.ac.uk>

References

Einbeck, J., Tutz, G., & Evers, L. (2005): Local principal curves, *Statistics and Computing* 15, 301-313.

Einbeck, J., Evers, L., & Powell, B. (2010): Data compression and regression through local principal curves and surfaces, *International Journal of Neural Systems*, 20, 177-192.

Einbeck, J. (2011): Bandwidth selection for nonparametric unsupervised learning techniques – a unified approach via self-coverage. *Journal of Pattern Recognition Research* 6, 175-192.

See Also

pcurve, princurve

calspeedflow

Speed-flow data from California.

Description

A ‘fundamental diagram’ with observations of speed and flow recorded from 9th of July 2007, 9am, to 10th of July 2007, 10pm, on Line 5 of the Californian Freeway SR57-N, VDS number 1202263. The data were originally measured in intervals of thirty seconds, and then aggregated over intervals of 5 minutes length.

Usage

```
data(calspeedflow)
```

Format

A data frame with 444 observations on the following 4 variables.

Date a factor with levels 07/09/2007... 07/10/2007.

Timestamp a factor with a timestamps in intervals of five minutes.

Lane5Flow a numeric vector of vehicle flow in vehicles per 5 minutes.

Lane5Speed a numeric vector of vehicle speed in miles per hour.

Source

Retrieved from PeMS.

References

Einbeck, J., and Dwyer, J. (2011). Using principal curves to analyze traffic patterns on freeways. *Transportmetrica* 7, 229-246.

Examples

```
data(calspeedflow)
plot(calspeedflow[,3:4])
```

coverage

Coverage and self-coverage plots.

Description

These functions compute coverages (for any principal object), and self-coverages (only for local principal curves, these may be used for bandwidth selection).

Usage

```
coverage.raw(X, vec, tau, weights=1, plot.type="p", print=FALSE,
             label=NULL,...)
```

```
coverage(X, vec, taumin=0.02, taumax, gridsize=25, weights=1,
         plot.type="o", print=FALSE,...)
```

```
lpc.coverage(object, taumin=0.02, taumax, gridsize=25, quick=TRUE,
             plot.type="o", print=FALSE, ...)
```

```
lpc.self.coverage(X, taumin=0.02, taumax=0.5, gridsize=25, x0=1,
                 way = "two", scaled=TRUE, weights=1, pen=2, depth=1,
                 control=lpc.control(boundary=0, cross=FALSE), quick=TRUE,
                 plot.type="o", print=FALSE, ... )
```

```
select.self.coverage(self, smin, plot.type="o", plot.segments=NULL)
```

Arguments

<code>X</code>	a $N \times d$ data matrix.
<code>object</code>	An object of type <code>lpc</code> or <code>lpc.spline</code> .
<code>vec</code>	A matrix with d columns. The rows contain the points which make up the fitted object.
<code>tau</code>	tube size.

<code>taumin</code>	Minimal tube size.
<code>taumax</code>	Maximal tube size.
<code>weights</code>	An optional vector of weights. If weights are specified, then the coverage is the weighted mean of the indicator functions for falling within the tube. The function <code>lpc.coverage</code> does not have a <code>weights</code> argument, as it extracts the weights from the <code>\$weights</code> component of the fitted object.
<code>label</code>	Experimental option; don't use.
<code>gridsize</code>	The number of different tube sizes to consider.
<code>quick</code>	If TRUE, an approximate coverage curve is provided by computing distances between data points and the curve through the closest local centers or mass; whereas with FALSE we use the distances of the points when projected orthogonally onto the spline representation of the local principal curve. The latter takes considerably more computing time. The resulting coverage curves are generally very similar, but the quick version may deliver little spurious peaks occasionally.
<code>self</code>	An object of class <code>self</code> , or a matrix with two columns providing a self-coverage curve.
<code>smin</code>	Minimum coverage for bandwidth selection. Default: 1/3 for clustering, 2/3 for principal curves.
<code>plot.type</code>	If set to 0, no plotted output is given. Otherwise, an appropriate plot is provided, using the plotting type as specified.
<code>plot.segments</code>	A list with default <code>list(lty=c(1, 2, 3), lwd=c(2, 1, 1), lcol=c(3, 3, 3))</code> which specifies how (and how many) bandwidth candidates, in order of decreasing negative second derivative of self-coverage, are to be highlighted.
<code>print</code>	If TRUE, coverage values are printed on the screen as soon as computed. This is quite helpful especially if <code>gridsize</code> is large.
<code>x0, way, scaled, pen, depth, control</code>	LPC parameters as outlined in lpc and lpc.control .
<code>...</code>	Optional graphical parameters passed to the corresponding plotting functions.

Details

The function `coverage.raw` computes the coverage, i.e. the proportion of data points lying inside a circle or band with radius τ , for a fixed value `tau`. The whole coverage curve $C(\tau)$ is constructed through function `coverage`.

Functions `coverage.raw` and `coverage` can be used for any object fitted by an unsupervised learning technique (for instance, HS principal curves, or even clustering algorithms), while the functions prefixing with `lpc.` can only be used for local principal curves. The function `lpc.coverage` is a wrapper around `coverage` which takes directly a fitted `lpc` object, rather than a data matrix.

Function `select.self.coverage` extracts suitable bandwidths from the self-coverage curve, and produces a plot. The function is called from within `lpc.self.coverage`, but can also be called directly by the user (for instance, if the graphical output is to be reproduced, or if the minimum coverage `smin` is to be modified). The component `$select` contains the selected candidate bandwidths, in the order of strength of evidence provided by the self-coverage criterion (the best bandwidth comes first, etc.). A plot is produced as a by-product, which symbolizes the best bandwidth by a thick solid

line, the second-best by a dashed line, and the third-best by a dotted line. It is recommended to run the self-coverage functions with fixed starting points, as in the examples below.

See Einbeck (2011) for details. Note that the original publication by Einbeck, Tutz, and Evers (2005) uses 'quick' coverage curves.

Value

A list of items, and a plot (unless `plot.type=0`).

The function `lpc.self.coverage` produces an object of class `self`. The component `$select` recommends suitable bandwidths for the use in `lpc`, in the order of strength of evidence. These correspond to points of strong negative curvature (implemented via second differences) of the self-coverage curve.

Author(s)

J. Einbeck

References

Einbeck, J., Tutz, G., & Evers, L. (2005). Local principal curves. *Statistics and Computing* 15, 301-313.

Einbeck, J. (2011). Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *Journal of Pattern Recognition Research* 6, 175-192.

See Also

[lpc](#)

Examples

```
data(gvessel)
gvessel.self <- lpc.self.coverage(gvessel[,c(2,4,5)], x0=c(35, 1870,
6.3), print=FALSE, plot.type=0)
h <- select.self.coverage(gvessel.self)$select
gvessel.lpc <- lpc(gvessel[,c(2,4,5)], h=h[1], x0=c(35, 1870, 6.3))
lpc.coverage(gvessel.lpc, gridsize=10, print=FALSE)
```

```
data(calspeedflow)
fitms <- ms(calspeedflow[,3:4])
coverage(fitms$data, fitms$cluster.center)
```

`followx`*Fit an individual branch of a local principal curve.*

Description

Internal function of package **LPCM** called by `lpc`. Do not use!

Usage

```
followx(Xi, x0, h, t0, iter, way, weights, pen = 2, phi = 1,  
        lasteigenvector = 0, rho0 = 0.4, boundary=0.005,  
        convergence.at= 0.000001, cross=FALSE)
```

Arguments

`Xi`
`x0`
`h`
`t0`
`iter`
`way`
`weights`
`pen`
`phi`
`lasteigenvector`

`rho0`
`boundary`
`convergence.at`
`cross`

Author(s)

JE

See Also

[lpc](#)

gaia

Gaia data

Description

(Simulated) spectral decomposition of stellar objects, generated in the framework of the Gaia project.

Usage

```
data(gaia)
```

Format

A data frame with 8286 observations on the following 22 variables.

ID ID of the object

metallicity metallicity (abundance); that is proportion of matter other than hydrogen and helium relative to that of the sun.

gravity the surface gravity; that is acceleration due to gravity at the surface of the star.

temperature the 'effective' temperature (K); that is the temperature of the observable part of the stellar atmosphere.

band1 photon counts in band 1

band2 photon counts in band 2

band3 photon counts in band 3

band4 photon counts in band 4

band5 photon counts in band 5

band6 photon counts in band 6

band7 photon counts in band 7

band8 photon counts in band 8

band9 photon counts in band 9

band10 photon counts in band 10

band11 photon counts in band 11

band12 photon counts in band 12

band13 photon counts in band 13

band14 photon counts in band 14

band15 photon counts in band 15

band16 photon counts in band 16

Details

Gaia is an astrophysics mission of the European Space Agency (ESA) which will undertake a detailed survey of over 10^9 stars in our Galaxy and extragalactic objects. An important part of the scientific analysis of these data is the classification of all the objects as well as the estimation of stellar astrophysical parameters (effective stellar temperature, surface gravity, metallicity). This will be done on the basis of high-dimensional spectroscopic and astrometric data such as those ones given here.

More precisely, the spectral data come in form of photon counts ("fluxes") observed in (originally) 96 wavelength intervals ("bands"), see Bailer-Jones (2010) for more details. The data given here are a 16-dimensional subset created by binning/selecting from the 96 bands. The counts given here are standardized, i.e. they are divided by the total number of incoming photons over all filters (in other words, they add up to 1). Note that these data are simulated using computer models. The satellite which will collect the actual data will be launched in 2012.

The 16-d spectral data have been used in Einbeck, Evers and Bailer-Jones (2008) as well as Einbeck, Evers and Powell (2010) in order to predict the stellar temperature.

Source

Coryn Bailer-Jones (MPIA Heidelberg).

References

Bailer-Jones, C.A.L. (2010). The ILIUM forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from Gaia spectrophotometry, *Monthly Notices of the Royal Astronomical Society*, vol. 403, pp. 96-116.

Einbeck, J., Evers, L., and Bailer-Jones, C.A.L. (2008). Representing complex data using localized principal components with application to astronomical data. In: Gorban, A, Kegl, B, Wunsch, D, & Zinovyev, A: *Principal Manifolds for Data Visualization and Dimension Reduction; Lecture Notes in Computational Science and Engineering* 58, 180-204, ISSN/ISBN: 978-3-540-73749-0.

Einbeck, J., Evers, L., and Powell, B. (2010): Data compression and regression through local principal curves and surfaces, *International Journal of Neural Systems*, 20, 177-192.

Examples

```
data(gaia)
s <- sample(nrow(gaia),200)
library(lattice)
splom(gaia[s,5:20], cex=0.3,pscales=0)

gaia.pc <- princomp(gaia[s,5:20])
temp <- gaia$temperature
tempcol <- (temp[s]- min(temp[s]))/max(temp[s]- min(temp[s]))
library(scatterplot3d)
scatterplot3d(gaia.pc$scores[,c(2,1,3)], pch="+",
              color=rgb(sqrt(tempcol),0,1-sqrt(tempcol)))
# This is a 3D scatterplot of the first three principal component scores;
# with higher stellar temperatures shaded in red colour.
```

`gvessel`*North Atlantic Water Temperature Data.*

Description

These are observations taken over nine days in May 2000 by the German vessel Gauss in the North Atlantic.

Usage

```
data(gvessel)
```

Format

A data frame with 643 observations on the following 7 variables.

`day2g` an integer for the day at which the measurement was taken.

`salg` a numeric vector with measurements of salinity according to the PSS (Practical Salinity Scale).

`tempg` a numeric vector with measurements of water temperature in degrees Celsius.

`depthg` a numeric vector with the water depths (in meters) at which the measurements were taken.

`oxyg` a numeric vector with measurements of oxygen content (mm per litre of water)

`longg` longitude

`latg` latitude

Source

Retrieved by B. Powell from the World Ocean Database, http://www.nodc.noaa.gov/OC5/WOD/pr_wod.html.

References

Einbeck, J., Evers, L., and Powell, B. (2010): Data compression and regression through local principal curves and surfaces, *International Journal of Neural Systems*, 20, 177-192.

Examples

```
data(gvessel)
pairs(gvessel[,c(3,2,4,5)])
tcol <- (gvessel$tempg- min(gvessel$tempg))/(max(gvessel$tempg)- min(gvessel$tempg))
require(scatterplot3d)
scatterplot3d(gvessel[,2],gvessel[,4],gvessel[,5], color=rgb(tcol,0,1-tcol))
```

kernels.and.distances *Auxiliary kernel and distance functions.*

Description

Internal **LPCM** functions which are normally not to be called by the user.

Usage

```
kern(y, x = 0, h = 1)
kernd(X, x, h)
kdex(X, x, h)
distancevector(X, y, d = "euclid", na.rm = TRUE)
vecdist(X,Y)
mindist(X,y)
enorm(x)
```

Arguments

x	a number or vector.
y	a vector.
h	a bandwidth.
X	a matrix.
Y	a matrix.
d	type of distance measure (only 'euclid').
na.rm	...

Details

kern specifies the base kernel (by default Gaussian) used in lpc ; kernd is the corresponding multivariate product kernel. kdex is a pointwise multivariate kernel density estimator.

distancevector makes use of function vdisseuclid from R package **hopach** (but that package does not need to be loaded).

Author(s)

JE

References

Pollard, van der Laan, and Wall (2010). Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH). R package **hopach** version 2.9.1.

lpc

*Local principal curves***Description**

This is the main function which computes the actual local principal curve, i.e. a sequence of local centers of mass.

Usage

```
lpc(X, h, t0 = mean(h), x0, way = "two", scaled = TRUE,
    weights=1, pen = 2, depth = 1, control=lpc.control())
```

Arguments

X	data matrix with N rows (observations) and d columns (variables).
h	bandwidth. May be either specified as a single number, then the same bandwidth is used in all dimensions, or as a d -dimensional bandwidth vector. The default setting is 10 percent of the range in each direction. If <code>scaled = TRUE</code> then the bandwidth has to be specified in fractions of the data range, e.g. <code>h= c(0.2,0.1)</code> , rather than absolute values.
t0	scalar step length. Default setting is <code>t0=h</code> , if <code>h</code> is a scalar, and <code>t0=mean(h)</code> , if <code>h</code> is a vector.
x0	specifies the choice of starting points. The default choice <code>x0=1</code> will select one suitable starting point automatically (in form of a local density mode). The second built-in option <code>x0=0</code> will use all local density modes as starting points, hence produce as many branches as modes. Optionally, one can also set one or more starting points manually here. This can be done in form of a matrix, where each row corresponds to a starting point, or in form of a vector, where starting points are read in consecutive order from the entries of the vector. The starting point has always to be specified on the original data scale, even if <code>scaled=TRUE</code> . A fixed number of starting points can be enforced through option <code>mult</code> in <code>lpc.control</code> .
way	"one": go only in direction of the first local eigenvector, "back": go only in opposite direction, "two": go from starting point in both directions.
scaled	if TRUE, scales each variable by dividing through its range (see also the Notes section below).
weights	a vector of observation weights (can also be used to exclude individual observations from the computation by setting their weight to zero.)
pen	power used for angle penalization (see [1]). If set to 0, the angle penalization is switched off.
depth	maximum depth of branches (ϕ_{max} in [2]), restricted to the values 1,2 or 3 (The original LPC branch has depth 1. If, along this curve, a point features a high second local PC, this launches a new starting point, and the resulting branch has depth 2. If, along this branch, a point features a high second local PC, this launches a new starting point, and the resulting branch has depth 3.)

control Additional parameters steering particularly the starting-, boundary-, and convergence behavior of the fitted curve. See [lpc.control](#).

Value

A list of items:

LPC The coordinates of the local centers of mass of the fitted principal curve.

Parametrization

Curve parameters and branch labels for each local center of mass.

h The bandwidth used for the curve estimation.

t0 The constant t_0 used for the curve estimation.

starting.points

The coordinates of the starting point(s) used.

data The data frame used for curve estimation.

scaled Logical.

weights The vector of weights used for curve estimation.

control The settings used in `lpc.control()`

Misc Miscellanea.

Note

All values provided in the output refer to the scaled data, if `scaled=TRUE`. Use [unscale](#) to convert the results back to the original data scale.

The option `scaled=TRUE` scales the data by dividing each variable through their range. This differs from the usual scaling through the standard deviation as common for PCA, but we found the algorithm and the default bandwidth selection to work more reliably this way. If you wish to scale by the standard deviation, please do that by feeding the scaled data directly into the `lpc` function, i.e.

```
lpc(sweep(data, 2,sd(data), "/"),h, t0, ..., scaled=FALSE, ...).
```

Author(s)

J. Einbeck and L. Evers. See [LPCM-package](#) for further acknowledgements.

References

[1] Einbeck, J., Tutz, G., & Evers, L. (2005). Local principal curves. *Statistics and Computing* 15, 301-313.

[2] Einbeck, J., Tutz, G., & Evers, L. (2005): Exploring Multivariate Data Structures with Local Principal Curves. In: Weihs, C. and Gaul, W. (Eds.): *Classification - The Ubiquitous Challenge*. Springer, Heidelberg, pages 256-263.

Examples

```

data(calspeedflow)
lpc1 <- lpc(calspeedflow[,3:4])
plot(lpc1)

data(mussels, package="dr")
lpc2 <- lpc(mussels[,-3], x0=as.numeric(mussels[49,-3]), scaled=FALSE)
plot(lpc2, curvecol=2)

data(gaia)
s <- sample(nrow(gaia),200)
gaia.pc <- princomp(gaia[s,5:20])
lpc3 <- lpc(gaia.pc$scores[,c(2,1,3)], scaled=FALSE)
plot(lpc3, curvecol=2, type=c("curve", "mass"))

# Simulated letter 'E' with branched LPC
ex<- c(rep(0,40), seq(0,1,length=20), seq(0,1,length=20), seq(0,1,length=20))
ey<- c(seq(0,2,length=40), rep(0,20), rep(1,20), rep(2,20))
sex<-rnorm(100,0,0.01); sey<-rnorm(100,0,0.01)
eex<-rnorm(100,0,0.1); eey<-rnorm(100,0,0.1)
ex1<-ex+sex; ey1<-ey+sey
ex2<-ex+eex; ey2<-ey+eey
e1<-cbind(ex1,ey1); e2<-cbind(ex2,ey2)
lpc.e1 <- lpc(e1, h= c(0.1,0.1), depth=2, scaled=FALSE)
plot(lpc.e1, type=c("curve", "mass", "start"))

```

lpc.control

Auxiliary parameters for controlling local principal curves.

Description

This function bundles parameters controlling mainly the starting-, convergence-, boundary-, and stopping-behaviour of the local principal curve. It will be used only inside the `lpc()` function argument.

Usage

```

lpc.control(iter =100, cross=TRUE,
            boundary = 0.005, convergence.at = 0.00001,
            mult=NULL, ms.h=NULL, ms.sub=30,
            pruning.thresh=0.0, rho0=0.4)

```

Arguments

<code>iter</code>	Maximum number of iterations on either side of the starting point within each branch.
<code>cross</code>	Logical parameter. If TRUE, curves are stopped when they come too close to an existing branch. Used in the self-coverage function.

boundary	This boundary correction [2] reduces the bandwidth adaptively once the relative difference of parameter values between two centers of mass falls below the given threshold. This measure delays convergence and enables the curve to proceed further into the end points. If set to 0, this boundary correction is switched off.
convergence.at	This forces the curve to stop if the relative difference of parameter values between two centers of mass falls below the given threshold. If set to 0, then the curve will always stop after exactly <code>iter</code> iterations.
mult	numerical value which enforces a fixed number of starting points. If the number given here is larger than the number of starting points provided at <code>x0</code> , then the missing points will be set at random (For example, if $d = 2$, <code>mult=3</code> , and <code>x0=c(58.5, 17.8, 80,20)</code> , then one gets the starting points (58.5, 17.8), (80,20), and a randomly chosen third one. Another example for such a situation is <code>x0=NULL</code> with <code>mult=1</code> , in which one random starting point is chosen). If the number given here is smaller the number of starting points provided at <code>x0</code> , then only the first <code>mult</code> starting points will be used.
ms.h	sets the bandwidth (vector) for the initial mean shift procedure which finds the local density modes, and, hence, the starting points for the LPC. If unspecified, the bandwidth <code>h</code> used in function <code>lpc</code> is used here too.
ms.sub	proportion of data points (default=30) which are used to initialize mean shift trajectories for the mode finding. In fact, we use $\min(\max(\text{ms.sub}, \text{floor}(\text{ms.sub} \cdot N / 100)), 10 \cdot \text{ms.sub})$ trajectories.
pruning.thresh	Prunes branches corresponding to higher-depth starting points if their density estimate falls below this threshold. Typically, a value between 0.0 and 1.0. The setting 0.0 means no pruning.
rho0	A numerical value which steers the birth process of higher-depth starting points. Usually, between 0.3 and 0.4 (see reference [1]).

Value

A list of the nine specified input parameters, which can be read by the `control` argument of the `lpc` function.

Author(s)

JE

References

- [1] Einbeck, J., Tutz, G. & Evers, L. (2005): Exploring Multivariate Data Structures with Local Principal Curves. In: Weihs, C. and Gaul, W. (Eds.): Classification - The Ubiquitous Challenge. Springer, Heidelberg, pages 256-263.
- [2] Einbeck, J. and Zayed, M. (2011). Some asymptotics for localized principal components and curves. Working paper, Durham University. Unpublished.

Examples

```
data(calspeedflow)
fit1 <- lpc(calspeedflow[,c(3,4)], x0=c(50,60), scaled=TRUE,
  control=lpc.control(iter=20, boundary=0))
plot(fit1, type=c("curve", "start", "mass"))
```

lpc.project

*Projection onto LPC***Description**

Projects a new observation onto the spline representation of the local principal curve.

Usage

```
lpc.project(object, newdata, ...)
```

Arguments

object	Object of class lpc or lpc.spline.
newdata	A data frame containing the new data to be projected.
...	Additional arguments to be passed to lpc.project.spline.

Value

closest.pi	Projection index of projected point(s) (in cubic spline parametrization).
closest.or.pi	Projection index of projected point(s) (in terms of the original LPC parametrization).
closest.coords	Coordinates of projected data point(s)
closest.dist	Euclidean distance between data point(s) and their projected counterpart(s).
closest.branch	ID of branch onto which the data point was projected (the IDs get allocated in the output component \$Parametrization of function lpc).

Note

The parametrization of the cubic spline function is not exactly the same as that of the original LPC. The reason is that the latter uses Euclidean distances between centers of masses, while the former uses the arc length along the cubic spline. The differences are normally quite small, though.

Author(s)

J. Einbeck and L. Evers

References

Einbeck, J., Evers, L. & Hinchliff, K. (2010): Data compression and regression based on local principal curves. In A. Fink, B. Lausen, W. Seidel, and A. Ultsch (Eds), *Advances in Data Analysis, Data Handling, and Business Intelligence*, Heidelberg, pp. 701–712, Springer.

See Also

[lpc](#), [lpc.spline](#)

Examples

```
data(gvessel)
gvessel.lpc <- lpc(gvessel[,c(2,4,5)], scaled=TRUE, h=0.11, x0=c(35, 1870, 6.3))
lpc.project(gvessel.lpc, newdata=data.frame(salg=35,dephtg= 2000,oxyg=6))
```

lpc.spline

Representing local principal curves through a cubic spline.

Description

Fis a natural cubic spline component-wise through the series of local centers of mass. This provides a continuous parametrization in terms of arc length distance, which can be used to compute a projection index for the original or new data points.

Usage

```
lpc.spline(lpcobject, optimize = TRUE, compute.Rc=FALSE,
           project=FALSE, ...)
```

Arguments

lpcobject	Object of class lpc.
optimize	Boolean. If TRUE, optimize is used to find the point on the curve with minimum distance. Otherwise, data points are only projected onto the closest knot.
compute.Rc	Boolean. If TRUE, the goodness-of-fit measure R_C suggested in [1] is computed and returned (using the scaled data, if scaled=TRUE in lpcobject).
project	Boolean. If TRUE, projections onto curve are computed.
...	Additional arguments to be passed to lpc.project.spline

Details

See reference [2].

Value

knots.pi	LPC parameters (in cubic spline parametrization) at position of the knots of the spline function (these are not identical to the LPC mass points!)
knots.coords	Coordinates of the spline knots.
closest.pi	Parameter of the projected data points.
closest.coords	Coordinates of projected data points.
closest.dist	Euclidean distance between original and projected data point.
closest.branch	ID Number of the branch on which the data point was projected (the IDs are given in the output of function lpc).
Rc	Value of R_C .
project	repeats the input value of project.
lpcobject	returns the provided object lpcobject.
splinefun	returns the cubic spline function (generated by lpc.splinefun).

Warning

Careful with options project and compute.Rc - they can take rather long if the data set is large!

Note

The parametrization of the cubic spline function is not exactly the same as that of the original LPC. The reason is that the latter uses Euclidean distances between centers of masses, while the former uses the arc length along the cubic spline. However, the differences are normally quite small.

Author(s)

J. Einbeck and L. Evers

References

- [1] Einbeck, J., Tutz, G., and Evers, L. (2005). Local principal curves. *Statistics and Computing* 15, 301-313.
- [2] Einbeck, J., Evers, L. & Hinchliff, K. (2010): Data compression and regression based on local principal curves. In A. Fink, B. Lausen, W. Seidel, and A. Ultsch (Eds), *Advances in Data Analysis, Data Handling, and Business Intelligence*, Heidelberg, pp. 701–712, Springer.

See Also

[lpc](#)

Examples

```
data(gvessel)
gvessel.lpc <- lpc(gvessel[,c(2,4,5)], h=0.11, x0=c(35, 1870, 6.3))
gvessel.spline <- lpc.spline(gvessel.lpc)
plot(gvessel.spline, lwd=2)
```

`lpc.spline.auxiliary.functions`*Auxiliary functions for spline fitting and projection.*

Description

Internal functions of package **LPCM** called by `lpc.spline` and others. These will rarely be called directly by the user.

Usage`lpc.splinefun(lpcobject)``lpc.fit.spline(lpcsl, num.knots = 100)``lpc.spline.eval(lpcsl, or.pi, branch = 0)``lpc.project.spline(lpcsl, newdata, num.knots = 100, optimize = TRUE)``lpc.curve.length(lpcsl, or.pi, branch = 0, total.subdivisions = 10000,
min.subdivisions = 100)`**Arguments**

<code>lpcobject</code>	Object of type <code>lpc</code> .
<code>lpcsl</code>	Object generated by <code>lpc.splinefun</code> .
<code>num.knots</code>	number of spline knots
<code>or.pi</code>	original projection index
<code>branch</code>	branch ID
<code>newdata</code>	new data frame
<code>optimize</code>	Boolean.
<code>total.subdivisions</code>	total number of subdivisions for arc length computation.
<code>min.subdivisions</code>	minimum number of subdivisions for arc length computation.

Author(s)

L. Evers and J. Einbeck

See Also

[lpc.spline](#)

ms *Mean shift clustering.*

Description

Functions for mean shift, iterative mean shift, mean shift clustering, and bandwidth selection for mean shift clustering (based on self-coverage). The main function is `ms` which, for a given bandwidth, detects the local modes ('local principal points') and performs the clustering.

These functions implement the techniques presented in Einbeck (2011).

Usage

```
meanshift(X, x, h)
ms.rep(X, x, h, plotms=1, thresh= 0.00000001, iter=200)
ms(X, h, subset, thr=0.0001, scaled= TRUE, iter=200, plotms=2,
  or.labels=NULL, ...)
ms.self.coverage(X, taumin=0.02, taumax=0.5, gridsize=25,
  thr=0.0001, scaled=TRUE, cluster=FALSE, plot.type="o",
  or.labels=NULL, print=FALSE, ...)
```

Arguments

<code>X</code>	data matrix.
<code>h</code>	bandwidth (by default, 10 percent of the data range).
<code>x</code>	point from which we wish to shift to the local mean.
<code>subset</code>	vector specifying a subset of 1:n, where n is the sample size. This allows to run the iterative mean shift procedure only from a subset of points (if unspecified, 1:n is used here, i.e. each data point serves as a starting point).
<code>scaled</code>	logical (if TRUE, each variable is divided by its range).
<code>taumin, taumax, gridsize</code>	determine the grid of bandwidths to investigate.
<code>thresh, iter</code>	mean shift iterations are stopped when the mean shift length (relative to the length of <code>x</code>) falls below <code>thresh</code> , or after <code>iter</code> iterations (whatever event happens first).
<code>thr</code>	adjacent mean shift clusters are merged if their relative distance falls below this threshold.
<code>cluster</code>	if TRUE, distances are always measured to the cluster to which an observation is assigned, rather than to the nearest cluster.
<code>plotms, plot.type, or.labels, ...</code>	graphical parameters.
<code>print</code>	if TRUE, coverage values are printed on the screen as soon as computed. This is quite helpful especially if <code>gridsize</code> is large.

Details

The methods implemented here can be used for density mode estimation, clustering, and the selection of starting points for the LPC algorithm.

Chen (1995) showed that, if the mean shift is computed iteratively, the resulting sequence of local means converges to a mode of the estimated density function. By assigning each data point to the mode to which it has converged, this turns into a clustering technique.

The concepts of coverage and self-coverage, which were originally introduced in the principal curve context, adapt straightforwardly to this setting.

The goodness-of-fit measure R_C can also be applied in this context. For instance, a value of $R_C = 0.8$ means that, after the clustering, the mean absolute residual length has been reduced by 80% (compared to the distances to the overall mean).

Value

The main function `ms` produces an object of class `ms`, with components:

<code>cluster.center</code>	a matrix which gives the coordinates of the estimated density modes (i.e., of the mean-shift based cluster centers).
<code>cluster.label</code>	assigns each data point to the cluster center to which its mean shift trajectory has converged.
<code>closest.label</code>	assigns each data point to the closest cluster center in terms of Euclidean distance.
<code>data</code>	the data frame (scaled if <code>scaled=TRUE</code>).
<code>scaled</code>	boolean.
<code>scaled.by</code>	the data were scaled by dividing each variable through the values provided in this vector.

For all other functions, use `names()`.

Author(s)

J. Einbeck. See [LPCM-package](#) for further acknowledgements.

References

Chen, Y. (1995). Mean Shift, Mode Seeking, and Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17, 790-799.

Einbeck, J. (2011). Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *Journal of Pattern Recognition Research* 6, 175-192.

See Also

[Rc](#), [lpc.self.coverage](#)

Examples

```

data(faithful)
# Mean shift clustering with user-defined bandwidth (5 percent of data range)
fit <- ms(faithful, h=0.05)

# Goodness-of-fit
coverage(fit$data, fit$cluster.center)
Rc(fit)

# Bandwidth selection via self-coverage
## Not run: foo <- ms.self.coverage(faithful,gridsize= 50, taumin=0.1,
taumax=0.5, plot.type="o")
h <- select.self.coverage(foo)$select
fit <- ms(faithful,h=h[1])
## End(Not run)

```

plot.lpc

Plotting local principal curves

Description

Takes an object of class `lpc` or `lpc.spline` and plots any subset of the following components of the local principal curve: Centers of mass; the curve connecting the local centers of mass; the cubic spline representation of the curve; the projections onto the curve; the starting points.

Usage

```

## S3 method for class 'lpc'
plot(x, type, unscale = TRUE, lwd = 1, datcol = "grey60",
      datpch = 21, masscol = NULL, masspch = 15, curvecol = 1, splinecol = 3,
      projectcol = 4, startcol = NULL, startpch=NULL,...)
## S3 method for class 'lpc.spline'
plot(x, type, unscale = TRUE, lwd = 1, datcol = "grey60",
      datpch = 21, masscol = NULL, masspch = 15, curvecol = 1, splinecol = 3,
      projectcol = 4, startcol = NULL, startpch=NULL,...)

```

Arguments

<code>x</code>	an object of class <code>lpc</code> or <code>lpc.spline</code> .
<code>type</code>	a vector of type <code>c("mass", "spline", ...)</code> with possible entries <code>mass</code> , <code>curve</code> , <code>spline</code> , <code>project</code> , <code>start</code> .
<code>unscale</code>	if <code>TRUE</code> , then data (and all fitted componens) are scaled back to their original scale; otherwise the scaled data are plotted (only relevant if <code>scaled=TRUE</code> in the fitted object).
<code>lwd</code>	width of curves.
<code>datcol</code>	color of data points.

datpch	plotting symbol for data points.
masscol	color of centers of mass (see below).
masspch	plotting symbol for centers of mass.
curvecol	color of the curve interpolating the local centers of mass (this is the "local principal curve").
splinecol	color of the spline representation of the local principal curve.
projectcol	color of projections onto the spline representation of the local principal curve.
startcol	color of the plotted starting points.
startpch	plotting symbol for starting points; needs to be either a single symbol, or a vector of symbols of the same length as the number of starting points.
...	further arguments passed to plot or scatterplot3d.

Value

A 2D plot, 3D plot, or a pairs plot (depending on the data dimension d).

The most flexible plotting option is `masscol`. Depending on the length of the specified vector, this will be interpreted differently. If a scalar is provided, the corresponding color will be given to all centers of mass. If the length of the vector is larger than 1, then this option will assign different colours to different depths, or different branch numbers, or to individual data points, depending on the length. The default setting is assigning colours according to depth, in the order red, blue, black.

With increasing dimension d , less plotting options tend to be supported. The nicest plots are obtained for $d = 2$ and $d = 3$.

Warning

This function computes all missing information (if possible), so computation will take the longer the less informative the given object is, and the more advanced aspects are asked to plot!

Author(s)

JE

References

Einbeck, J., Tutz, G., and Evers, L. (2005). Local principal curves. *Statistics and Computing* 15, 301-313.

Einbeck, J., Evers, L. & Hinchliff, K. (2010): Data compression and regression based on local principal curves. In A. Fink, B. Lausen, W. Seidel, and A. Ultsch (Eds), *Advances in Data Analysis, Data Handling, and Business Intelligence*, Heidelberg, pp. 701–712, Springer.

See Also

[lpc](#), [lpc.spline](#)

Examples

```
data(calspeedflow)
lpc1 <- lpc(calspeedflow[,3:4])
plot(lpc1, type=c("spline","project"), lwd=2)
```

print.lpc

Printing output for lpc and lpc.spline objects

Description

Takes an object of class `lpc` or `lpc.spline` and displays some standard output.

Usage

```
## S3 method for class 'lpc'
print( x, digits = max(3, getOption("digits") - 3), ...)
## S3 method for class 'lpc.spline'
print( x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

<code>x</code>	an object of class <code>lpc</code> or <code>lpc.spline</code> .
<code>digits</code>	not yet in use.
<code>...</code>	further arguments.

Value

Some short text.

Author(s)

JE

See Also

[lpc](#)

Examples

```
data(calspeedflow)
lpc1 <- lpc(calspeedflow[,3:4])
print(lpc1)
lpc2 <- lpc.spline(lpc1)
print(lpc2)
```

Rc *Measuring goodness-of-fit for principal objects.*

Description

These functions compute the ‘coverage coefficient’ R_C for local principal curves, local principal points (i.e., kernel density estimates obtained through iterated mean shift), and other principal objects.

Usage

```
Rc(x, ...)
```

```
## S3 method for class 'lpc'
```

```
Rc(x, ...)
```

```
## S3 method for class 'lpc.spline'
```

```
Rc(x, ...)
```

```
## S3 method for class 'ms'
```

```
Rc(x, ...)
```

```
base.Rc(data, closest.coords, type="curve")
```

Arguments

x	an object used to select a method.
...	Further arguments passed to or from other methods (not needed yet).
data	A data matrix.
closest.coords	A matrix of coordinates of the projected data.
type	For principal curves, don’t modify. For principal points, set "points".

Details

Rc computes the coverage coefficient R_C , a quantity which estimates the goodness-of-fit of a fitted principal object. This quantity can be interpreted similar to the coefficient of determination in regression analysis: Values close to 1 indicate a good fit, while values close to 0 indicate a ‘bad’ fit (corresponding to linear PCA).

For objects of type `lpc`, `lpc.spline`, and `ms`, S3 methods are available which use the generic function `Rc`. This, in turn, calls the base function `base.Rc`, which can also be used manually if the fitted object is of another class. In principle, function `base.Rc` can be used for assessing goodness-of-fit of any principal object provided that the coordinates (`closest.coords`) of the projected data are available. For instance, for HS principal curves fitted via `princurve`, this information is contained in component `$s`, and for a k-means object, say `fitk`, this information can be obtained via `fitk$centers[fitk$cluster,]`. Set `type="points"` in the latter case.

The function `Rc` attempts to compute all missing information, so computation will take the longer the less informative the given object `x` is. Note also, `Rc` looks up the option `scaled` in the fitted

object, and accounts for the scaling automatically. Important: If the data were scaled, then do NOT unscale the results by hand in order to feed the unscaled version into `base.Rc`, this will give a wrong result.

In terms of methodology, these functions compute R_C directly through the mean reduction of absolute residual length, rather than through the area above the coverage curve.

These functions do currently not account for observation weights, i.e. R_C is computed through the unweighted mean reduction in absolute residual length (even if weights have been used for the curve fitting).

Acknowledgements

Contributions (in form of pieces of code, or useful suggestions for improvements) by Jo Dwyer, Mohammad Zayed, and Ben Oakley are gratefully acknowledged.

Author(s)

J. Einbeck and L. Evers.

References

Einbeck, Tutz, and Evers (2005). Local principal curves. *Statistics and Computing* 15, 301-313.

Einbeck (2011). Bandwidth selection for nonparametric unsupervised learning techniques – a unified approach via self-coverage. *Journal of Pattern Recognition Research* 6, 175-192.

See Also

[lpc.spline](#), [codems](#), [coverage](#).

Examples

```
data(calspeedflow)
lpc1 <- lpc.spline(lpc(calspeedflow[,3:4]), project=TRUE)
Rc(lpc1)
# is the same as:
base.Rc(lpc1$lpcobject$data, lpc1$closest.coords)

ms1 <- ms(calspeedflow[,3:4],plotms=0)
Rc(ms1)
# is the same as:
base.Rc(ms1$data, ms1$cluster.center[ms1$closest.label,], type="points")
```

unscale	<i>Unscaling local principal objects.</i>
---------	---

Description

unscale takes an object of type `lpc`, `lpc.spline`, or `ms`, which had been fitted using option `scaled=TRUE`, and transforms the scaled components back to the original data scale.

Usage

```
unscale(x, ...)  
  
## S3 method for class 'lpc'  
unscale(x,...)  
## S3 method for class 'lpc.spline'  
unscale(x,...)  
## S3 method for class 'ms'  
unscale(x,...)
```

Arguments

<code>x</code>	an object used to select a method.
<code>...</code>	Further arguments passed to or from other methods (not needed yet).

Value

A list of relevant items, such as `LPC`, `start`, `cluster.centers`, etc., which gives the unscaled versions of these quantities (some of them may carry the value `NULL`, if the corresponding information was not available from `x`).

Author(s)

JE

See Also

[lpc](#), [lpc.spline](#), [ms](#)

Examples

```
data(gvessel)  
unscale(lpc(gvessel[,c(2,4,5)], h=0.11, x0=c(35, 1870, 6.3)) )
```

Index

- *Topic **datasets**
 - calspeedflow, 3
 - gaia, 8
 - gvessel, 10
- *Topic **multivariate**
 - coverage, 4
 - lpc, 12
 - lpc.project, 16
 - lpc.spline, 17
 - LPCM-package, 2
 - ms, 20
 - plot.lpc, 22
 - print.lpc, 24
 - Rc, 25
 - unscale, 27
- *Topic **smooth**
 - lpc, 12
 - lpc.spline, 17
 - LPCM-package, 2
- base.Rc (Rc), 25
- calspeedflow, 3
- coverage, 4, 26
- distancevector (kernels.and.distances), 11
- enorm (kernels.and.distances), 11
- followx, 7
- gaia, 8
- gvessel, 10
- kdex (kernels.and.distances), 11
- kern (kernels.and.distances), 11
- kernd (kernels.and.distances), 11
- kernels.and.distances, 11
- lpc, 2, 5–7, 12, 17, 18, 23, 24, 27
- lpc.control, 5, 13, 14
- lpc.coverage, 2
- lpc.coverage (coverage), 4
- lpc.curve.length
 - (lpc.spline.auxiliary.functions), 19
- lpc.fit.spline
 - (lpc.spline.auxiliary.functions), 19
- lpc.project, 2, 16
- lpc.project.spline
 - (lpc.spline.auxiliary.functions), 19
- lpc.self.coverage, 2, 21
- lpc.self.coverage (coverage), 4
- lpc.spline, 2, 17, 17, 19, 23, 26, 27
- lpc.spline.auxiliary.functions, 19
- lpc.spline.eval
 - (lpc.spline.auxiliary.functions), 19
- lpc.splinefun
 - (lpc.spline.auxiliary.functions), 19
- LPCM (LPCM-package), 2
- LPCM-package, 2
- meanshift (ms), 20
- mindist (kernels.and.distances), 11
- ms, 2, 20, 26, 27
- plot.lpc, 22
- print.lpc, 24
- Rc, 2, 21, 25
- select.self.coverage (coverage), 4
- unscale, 13, 27
- vecdist (kernels.and.distances), 11