Package 'ldr'

February 20, 2015

Type Package

Title Methods for likelihood-based dimension reduction in regression

Version 1.3.3

Depends R (>= 2.10), GrassmannOptim, Matrix

Date 2014-06-06

Author Kofi Placid Adragni, Andrew Raim

Maintainer Kofi Placid Adragni <kofi@umbc.edu>

Description Functions, methods, and data sets for fitting likelihood-based dimension reduction in regression, using principal fitted components (pfc), likelihood acquired directions (lad), covariance reducing models (core).

URL http://www.jstatsoft.org/v61/i03/

License GPL (>= 2)

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2014-10-29 16:36:14

R topics documented:

bf *Function to generate a basis function.*

Description

This function is to construct a data-matrix of basis function using the n response observations. The response can be continuous or categorical. The function returns a matrix of n rows and r columns. The number of columns r depends on the choice of basis function. Polynomial, piecewise polynomial continuous and discontinuous, and Fourier bases are implemented. For a polynomial basis, r is the degree of the polynomial.

Usage

```
bf(y, case = c("poly", "categ", "fourier", "pcont", "pdisc"),
degree = 1, nslices = 1, scale = FALSE)
```

Arguments

y	A response vector of n observations.
case	Take values "poly" for polynomial, "categ" for categorical, "fourier" for Fourier, "pcont" for piecewise continuous, and "pdisc" for piecewise discontinuous bases.
degree	For polynomial and piecewise polynomial bases, degree is the degree of the polynomial. With "pdisc", degree=0 corresponds to piecewise constant.
nslices	The number of slices for piecewise bases only. The range of the response is partitioned into nslices parts with roughly equal numbers of observations. See details on piecewise bases for more information.
scale	If TRUE, the columns of the basis function are scaled to have unit variance.

Details

The basis function f_y is a vector-valued function of the response $y \in R$. There is an infinite number of basis functions, including the polynomial, piecewise polynomial, and Fourier. We implemented the following:

1. Polynomial basis: $f_y = (y, y^2, ..., y^r)^T$. It corresponds to the "poly" argument of bf. The argument degree is r of the polynomial is provided by the user. The subsequent $n \times r$ data-matrix is column-wise centered.

2. Piecewise constant basis: It corresponds to pdisc with degree=0. It is obtained by first slicing the range of y into h slices $H_1, ..., H_k$. The k^{th} component of $f_y \in \mathrm{R}^{h-1}$ is $f_{y_k} = J(y \in H_k) - n_k/n, k = 1, ..., h-1$, where n_y is the number of observations in H_k, and J is the indicator function. We suggest using between two and fifteen slices without exceeding $n/5$.

3. Piecewise discontinuous linear basis: It corresponds to "pdisc" with degree=1. It is more elaborate than the piecewise constant basis. A linear function of y is fit within each slice. Let τ_i be the knots, or endpoints of the slices. The components of $f_y \in \mathrm{R}^{2h-1}$ are obtained with $f_{y_{(2i-1)}} = J(y \in H_i)$; $f_{y_{2i}} = J(y \in H_i)(y - \tau_{i-1})$ for $i = 1, 2, ..., h-1$ and $f_{y_{(2h-1)}} = J(y \in$

$H_h)(y - \tau_{h-1})$. The subsequent $n \times (2h-1)$ data-matrix is column-wise centered. We suggest using fewer than fifteen slices without exceeding $n/5$.

4. Piecewise continuous linear basis: The general form of the components f_{y_i} of $f_y \in \mathrm{R}^{h+1}$ is given by $f_{y_1} = J(y \in H_1)$ and $f_{y_{i+1}} = J(y \in H_i)(y - \tau_{i-1})$ for $i = 1, ..., h$.. The subsequent $n \times (h-1)$ data-matrix is column-wise centered. This case corresponds to `"pcont"` with degree=1. The number of slices to use may not exceed $n/5$.

5. Fourier bases: They consist of a series of pairs of sines and cosines of increasing frequency. A Fourier basis is given by $f_y = (\cos(2\pi y), \sin(2\pi y), ..., \cos(2\pi k y), \sin(2\pi k y))^T$. The subsequent $n \times 2k$ data-matrix is column-wise centered.

6. Categorical basis: It is obtained using `"categ"` option when y takes h distinct values $1, 2, ..., h$, corresponding to the number of sub-populations or sub-groups. The number of slices is naturally h. The expression for the basis is identical to piecewise constant basis.

In all cases, the basis must be constructed such that $F^T F$ is invertible, where F is the $n \times r$ data-matrix with its ith row being f_y.

Value

`fy`	A matrix with `n` rows and `r` columns.
`scale`	Boolean. If TRUE, the columns of the output are standardized to have unit variance.

Author(s)

Kofi Placid Adragni <kofi@umbc.edu>

References

Adragni, KP (2009) PhD Dissertation, University of Minnesota.

Adragni, KP and Cook, RD (2009): Sufficient dimension reduction and prediction in regression. Phil. Trans. R. Soc. A 367, 4385-4405.

Cook, RD (2007): Fisher Lecture - Dimension Reduction in Regression (with discussion). Statistical Science, Vol. 22, 1–26.

Examples

```
data(bigmac)

# Piecewise constant basis with 5 slices
fy=bf(y=bigmac[,1], case="pdisc", degree=0, nslices=5)
fit1 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=fy, numdir=3, structure="aniso")
summary(fit1)

# Cubic polynomial basis
fy=bf(y=bigmac[,1], case="poly", degree=3)
fit2 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=fy, numdir=3, structure="aniso")
summary(fit2)

# Piecewise linear continuous with 3 slices
```

```
fy=bf(y=bigmac[,1], case="pcont", degree=1, nslices=3)
fit3 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=fy, numdir=3, structure="unstr")
summary(fit3)
```

---

bigmac *bigmac data*

Description

The data give average values in 1991 on several economic indicators for 45 world cities. All prices are in US dollars, using currency conversion at the time of publication.

Usage

```
data(bigmac)
```

Format

A data frame with 45 observations on the following 10 variables.

`BigMac` Minimum labor to buy a BigMac and fries

`Bread` Minimum labor to buy 1 kg bread

`BusFare` Lowest cost of 10k public transit

`EngSal` Electrical engineer annual salary, 1000s

`EngTax` Tax rate paid by engineer

`Service` Annual cost of 19 services

`TeachSal` Primary teacher salary, 1000s

`TeachTax` Tax rate paid by primary teacher

`VacDays` Average days vacation per year

`WorkHrs` Average hours worked per year

Source

Rudolf Enz, "Prices and Earnings Around the Globe", 1991 edition, Published by the Union Bank of Switzerland.

References

Cook, RD and Weisberg, S (2004). Applied Regression Including Computing and Graphics, New York: Wiley, http://www.stat.umn.edu/arc.

Examples

```
data(bigmac)
pairs(bigmac)
```

core *Covariance Reduction*

Description

Method to reduce sample covariance matrices to an informational core that is sufficient to characterize the variance heterogeneity among different populations.

Usage

```
core(X, y, Sigmas = NULL, ns = NULL, numdir = 2,
        numdir.test = FALSE, ...)
```

Arguments

X	Data matrix with n rows of observations and p columns of predictors. The predictors are assumed to have a continuous distribution.
y	Vector of group labels. Observations with the same label are considered to be in the same group.
Sigmas	A list object of sample covariance matrices corresponding to the different populations.
ns	A vector of number of observations of the samples corresponding to the different populations.
numdir	Integer between 1 and p. It is the number of directions to estimate for the reduction.
numdir.test	Boolean. If FALSE, core computes the reduction for the specific number of directions numdir. If TRUE, it does the computation of the reduction for the numdir directions, from 0 to numdir. Likelihood ratio test and information criteria are used to estimate the true dimension of the sufficient reduction.
...	Other arguments to pass to GrassmannOptim.

Details

Consider the problem of characterizing the covariance matrices $\Sigma_y, y = 1, ..., h$, of a random vector X observed in each of h normal populations. Let $S_y = (n_y - 1)\tilde{\Sigma}_y$ where $\tilde{\Sigma}_y$ is the sample covariance matrix corresponding to Σ_y, and n_y is the number of observations corresponding to y. The goal is to find a semi-orthogonal matrix $\Gamma \in R^{p\times d}, d < p$, with the property that for any two populations j and k

$$S_j|(\Gamma' S_j \Gamma = B, n_j = m) \sim S_k|(\Gamma' S_k \Gamma = B, n_k = m).$$

That is, given $\Gamma' S_g \Gamma$ and n_g, the conditional distribution of S_g must must depend on g. Thus $\Gamma' S_g \Gamma$ is sufficient to account for the heterogeneity among the population covariance matrices. The central subspace $\mathcal{S}$, spanned by the columns of Γ is obtained by optimizing the following log-likelihood function

$$L(\mathcal{S}) = c - \frac{n}{2}\log|\tilde{\Sigma}| + \frac{n}{2}\log|P_{\mathcal{S}}\tilde{\Sigma}P_{\mathcal{S}}| - \sum_{y=1}^{h}\frac{n_y}{2}\log|P_{\mathcal{S}}\tilde{\Sigma}_y P_{\mathcal{S}}|,$$

where c is a constant depending only on p and n_y, $\tilde{\Sigma}_y, y = 1, ..., h$, denotes the sample covariance matrix from population y computed with divisor n_y, and $\tilde{\Sigma} = \sum_{y=1}^{h}(n_y/n)\tilde{\Sigma}$. The optimization is carried over $\mathcal{G}_{(d,p)}$, the set of all d-dimensional subspaces in R^p, called Grassmann manifold of dimension $d(p-d)$.

The dimension d is to be estimated. A sequential likelihood ratio test and information criteria (AIC, BIC) are implemented, following Cook and Forzani (2008).

Value

This command returns a list object of class `ldr`. The output depends on the argument `numdir.test`. If `numdir.test=TRUE`, a list of matrices is provided corresponding to the `numdir` values (1 through `numdir`) for each of the parameters Γ, Σ, and Σ_g. Otherwise, a single list of matrices for a single value of `numdir`. A likelihood ratio test and information criteria are provided to estimate the dimension of the sufficient reduction when `numdir.test=TRUE`. The output of `loglik`, `aic`, `bic`, `numpar` are vectors with `numdir` elements if `numdir.test=TRUE`, and scalars otherwise. Following are the components returned:

`Gammahat`	Estimate of Γ.
`Sigmahat`	Estimate of overall Σ.
`Sigmashat`	Estimate of group-specific Σ_g's.
`loglik`	Maximized value of the CORE log-likelihood.
`aic`	Akaike information criterion value.
`bic`	Bayesian information criterion value.
`numpar`	Number of parameters in the model.

Note

Currently `loglik`, `AIC`, and `BIC` are computed up to a constant. Therefore, these can be compared relatively (e.g. two `loglik`'s can be subtracted to compute a likelihood ratio test), but they should not be treated as absolute quantities.

Author(s)

Andrew Raim and Kofi P Adragni, University of Maryland, Baltimore County

References

Cook RD and Forzani L (2008). Covariance reducing models: An alternative to spectral modelling of covariance matrices. Biometrika, Vol. 95, No. 4, 799–812.

See Also

`lad`, `pfc`

Examples

```
data(flea)
fit1 <- core(X=flea[,-1], y=flea[,1], numdir.test=TRUE)
summary(fit1)

## Not run:
data(snakes)
fit2 <- ldr(Sigmas=snakes[-3], ns=snakes[[3]], numdir = 4,
model = "core", numdir.test = TRUE, verbose=TRUE,
sim_anneal = TRUE, max_iter = 200, max_iter_sa=200)
summary(fit2)

## End(Not run)
```

`flea` *Flea-beetles data*

Description

Six measurements on each of three species of flea-beetles: concinna, heptapotamica, and heikertingeri.

Usage

```
data(flea)
```

Format

A data frame with 74 observations on the following 7 variables.

`species` a factor with levels `Concinna`, `Heikert.`, and `Heptapot.`

`tars1` width of the first joint of the first tarsus in microns (the sum of measurements for both tarsi).

`tars2` the same for the second joint.

`head` the maximal width of the head between the external edges of the eyes in 0.01 mm.

`aede1` the maximal width of the aedeagus in the fore-part in microns.

`aede2` the front angle of the aedeagus (1 unit = 7.5 degrees).

`aede3` the aedeagus width from the side in microns.

Source

Lubischew, AA "On the Use of Discriminant Functions in Taxonomy", Biometrics, Dec. 1962, pp. 455-477.

References

Dianne Cook and Deborah F. Swayne, Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi. URL: http://www.ggobi.org/book/data/flea.xml

Examples

```
data(flea)
```

`lad` *Likelihood Acquired Directions*

Description

Method to estimate the central subspace, using inverse conditional mean and conditional variance functions.

Usage

```
lad(X, y, numdir = NULL, nslices = NULL, numdir.test = FALSE, ...)
```

Arguments

`X`	Data matrix with `n` rows of observations and `p` columns of predictors. The predictors are assumed to have a continuous distribution.
`y`	Response vector of `n` observations, possibly categorical or continuous. It is assumed categorical if `nslices=NULL`.
`numdir`	Integer between 1 and p. It is the number of directions of the reduction to estimate. If not provided then it will equal the number of distinct values of the categorical response.
`nslices`	Integer number of slices. It must be provided if `y` is continuous, and must be less than n. It is used to discretize the continuous response.
`numdir.test`	Boolean. If `FALSE`, `core` computes the reduction for the specific number of directions `numdir`. If `TRUE`, it does the computation of the reduction for the `numdir` directions, from 0 to `numdir`.
`...`	Other arguments to pass to `GrassmannOptim`.

Details

Consider a regression in which the response Y is discrete with support $S_Y = \{1, 2, ..., h\}$. Following standard practice, continuous response can be sliced into finite categories to meet this condition. Let $X_y \in R^p$ denote a random vector of predictors distributed as $X|(Y = y)$ and assume that $X_y \sim N(\mu_y, \Delta_y)$, $y \in S_Y$. Let $\mu = E(X)$ and $\Sigma = \mathrm{Var}(X)$ denote the marginal mean and variance of X and let $\Delta = E(\Delta_Y)$ denote the average covariance matrix. Given n_y independent observations of $X_y, y \in S_Y$, the goal is to obtain the maximum likelihood estimate of the d-dimensional central subspace $\mathcal{S}_{Y|X}$, which is defined informally as the smallest subspace such that Y is independent of X given its projection $P_{\mathcal{S}_{Y|X}} X$ onto $\mathcal{S}_{Y|X}$.

Let $\tilde{\Sigma}$ denote the sample covariance matrix of X, let $\tilde{\Delta}_y$ denote the sample covariance matrix for the data with $Y = y$, and let $\tilde{\Delta} = \sum_{y=1}^{h} m_y \tilde{\Delta}_y$ where m_y is the fraction of cases observed with $Y = y$. The maximum likelihood estimator of $\mathcal{S}_{Y|X}$ maximizes over $\mathcal{S} \in \mathcal{G}_{(d,p)}$ the log-likelihood function

$$L(\mathcal{S}) = \frac{n}{2}\log|P_{\mathcal{S}}\tilde{\Sigma}P_{\mathcal{S}}|_0 - \frac{n}{2}\log|\tilde{\Sigma}| - \frac{1}{2}\sum_{y=1}^{h} n_y \log|P_{\mathcal{S}}\tilde{\Delta}_y P_{\mathcal{S}}|_0,$$

where $|A|_0$ indicates the product of the non-zero eigenvalues of a positive semi-definite symmetric matrix A, $P_{\mathcal{S}}$ indicates the projection onto the subspace $\mathcal{S}$ in the usual inner product, and $\mathcal{G}_{(d,p)}$ is the set of all d-dimensional subspaces in R^p, called Grassmann manifold. The desired reduction is then $\hat{\Gamma}^T X$. Once the dimension of the reduction subspace is estimated, the columns of $\hat{\Gamma}$ are a basis for the maximum likelihood estimate of $\mathcal{S}_{Y|X}$.

The dimension d of the sufficient reduction is to be estimated. A sequential likelihood ratio test, and information criteria (AIC, BIC) are implemented, following Cook and Forzani (2009).

Value

This command returns a list object of class `ldr`. The output depends on the argument `numdir.test`. If `numdir.test=TRUE`, a list of matrices is provided corresponding to the `numdir` values (1 through `numdir`) for each of the parameters Γ, Δ, and Δ_y; otherwise, a single list of matrices for a single value of `numdir`. The output of `loglik`, `aic`, `bic`, `numpar` are vectors of `numdir` elements if `numdir.test=TRUE`, and scalars otherwise. Following are the components returned:

`R`	The reduction data-matrix of X obtained using the centered data-matrix X. The centering of the data-matrix of X is such that each column vector is centered around its sample mean.
`Gammahat`	Estimate of Γ
`Deltahat`	Estimate of Δ
`Deltahat_y`	Estimate of Δ_y
`loglik`	Maximized value of the LAD log-likelihood.
`aic`	Akaike information criterion value.
`bic`	Bayesian information criterion value.
`numpar`	Number of parameters in the model.

Author(s)

Kofi Placid Adragni <kofi@umbc.edu>

References

Cook RD, Forzani L (2009). Likelihood-based Sufficient Dimension Reduction, J. of the American Statistical Association, Vol. 104, No. 485, 197–208.

See Also

`core`, `pfc`

Examples

```
data(flea)
fit <- lad(X=flea[,-1], y=flea[,1], numdir=2, numdir.test=TRUE)
summary(fit)
plot(fit)
```

---

ldr *Likelihood-based Dimension Reduction*

Description

Main function of the package. It creates objects of one of classes `core`, `lad`, or `pfc` to estimate a sufficient dimension reduction subspace using covariance reducing models (CORE), likelihood acquired directions (LAD), or principal fitted components (PFC).

Usage

```
ldr(X, y = NULL, fy = NULL, Sigmas = NULL, ns = NULL,
    numdir = NULL, nslices = NULL, model = c("core", "lad", "pfc"),
    numdir.test = FALSE, ...)
```

Arguments

`X`	Design matrix with `n` rows of observations and `p` columns of predictors. The predictors are assumed to have a continuous distribution.
`y`	The response vector of length `n`. It can be continuous or categorical.
`fy`	Basis function to be obtained using `bf` or defined by the user. It is a function of `y` alone and has independent column vectors. It is used exclusively with `pfc`. See `bf` for detail.
`Sigmas`	A list object of sample covariance matrices corresponding to the different populations. It is used exclusively with `core`.
`ns`	A vector of number of observations of the samples corresponding to the different populations.
`numdir`	The number of directions to be used in estimating the reduction subspace. When calling `pfc`, the dimension `numdir` must be less than or equal to the minimum of `p` and `r`, where `r` is the number of columns of `fy`. When calling `lad` and `y` is continuous, `numdir` is the number of slices to use.
`nslices`	Number of slices for a continuous response. It is used exclusively with `lad`.
`model`	One of the following: `"pfc"`, `"lad"`, `"core"`.
`numdir.test`	Boolean. If `FALSE`, the chosen model fits with the provided `numdir`. If `TRUE`, the model is fit for all dimensions less or equal to `numdir`.
`...`	Additional arguments for specific models and/or Grassmannoptim.

Details

Likelihood-based methods to sufficient dimension reduction are model-based inverse regression approaches using the conditional distribution of the p-vector of predictors X given the response $Y = y$. Three methods are implemented in this package: covariance reduction (CORE), principal fitted components (PFC), and likelihood acquired directions (LAD). All three assume that $X|(Y = y) \sim N(\mu_y, \Delta_y)$.

For CORE, given a set of h covariance matrices, the goal is to find a sufficient reduction that accounts for the heterogeneity among the population covariance matrices. See the documentation of `"core"` for details.

For PFC, $\mu_y = \mu + \Gamma\beta f_y$, with various structures of Δ. The simplest is the isotropic (`"iso"`) with $\Delta = \delta^2 I_p$. The anisotropic (`"aniso"`) PFC model assumes that $\Delta = \mathrm{diag}(\delta_1^2, ..., \delta_p^2)$, where the conditional predictors are independent and on different measurement scales. The unstructured (`"unstr"`) PFC model allows a general structure for Δ. Extended structures are considered. See the help file of `pfc` for more detail.

LAD assumes that the response Y is discrete. A continuous response is sliced into finite categories to meet this condition. It estimates the central subspace $\mathcal{S}_{Y|X}$ by modeling both μ_y and Δ_y. See `lad` for more detail.

Value

An object of one of the classes `core`, `lad`, or `pfc` . The output depends on the model used. See `pfc`, `lad`, and `core` for further detail.

Author(s)

Kofi Placid Adragni <kofi@umbc.edu>

References

Adragni, KP and Cook, RD (2009): Sufficient dimension reduction and prediction in regression. Phil. Trans. R. Soc. A 367, 4385-4405.

Cook, RD (2007): Fisher Lecture - Dimension Reduction in Regression (with discussion). Statistical Science, 22, 1–26.

Cook, R. D. and Forzani, L. (2008a). Covariance reducing models: An alternative to spectral modelling of covariance matrices. Biometrika 95, 799-812.

Cook, R. D. and Forzani, L. (2008b). Principal fitted components for dimension reduction in regression. Statistical Science 23, 485–501.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. Journal of the American Statistical Association, Vol. 104, 485, pp 197–208.

See Also

`pfc`, `lad`, `core`

Examples

```
data(bigmac)
fit1 <- ldr(X=bigmac[,-1], y=bigmac[,1], fy=bf(y=bigmac[,1], case="pdisc",
        degree=0, nslices=5), numdir=3, structure="unstr", model="pfc")
summary(fit1)
plot(fit1)

fit2 <- ldr(X=bigmac[,-1], y=bigmac[,1], fy=bf(y=bigmac[,1], case="poly",
        degree=2), numdir=2, structure="aniso", model="pfc")
summary(fit2)
plot(fit2)

fit3 <- ldr(X=as.matrix(bigmac[,-1]), y=bigmac[,1], model="lad", nslices=5)
summary(fit3)
plot(fit3)
```

---

ldr.slices *Function to slice continuous response.*

Description

Divides a vector of length n into slices of approximately equal size. It is used to construct the piecewise bases, and internally used in lad functions.

Usage

```
ldr.slices(y, nslices = 3)
```

Arguments

y	a vector of length n.
nslices	the number of slices, no larger than n.

Details

The number of observations per slice m is computed as the largest integer less or equal to n/nslices. The n observations of y are ordered in the increasing order. The first set of first m observations is allocated to the first slice, the second set is allocated into the second slice, and so on.

Value

Returns a named list with four elements as follows:

bins	Slices with their observations
nslices	The actual number of slices produced.
slice.size	The number of observations in each slice.
slice.indicator	Vector of length n indicating the slice number of each observed response value.

Author(s)

Kofi Placid Adragni <kofi@umbc.edu>

References

Cook, RD and Weisberg, S (1999), Applied Regression Including Computing and Graphics, New York: Wiley.

OH *OH dataset*

Description

The hydroxyl OH group activity of compounds from molecular descriptors.

Usage

```
data(OH)
```

Format

A data frame with 719 observations on 294 descriptors/predictors. The response is act.

Source

The dataset was provided by Tomas Oberg.

Examples

```
data(OH)
```

---

pfc *Principal fitted components*

Description

Principal fitted components model for sufficient dimension reduction. This function estimates all parameters in the model.

Usage

```
pfc(X, y, fy = NULL, numdir = NULL, structure = c("iso", "aniso",
    "unstr", "unstr2"), eps_aniso = 1e-3, numdir.test = FALSE, ...)
```

Arguments

X	Design matrix with n rows of observations and p columns of predictors. The predictors are assumed to have a continuous distribution.
y	The response vector of n observations, continuous or categorical.
fy	Basis function to be obtained using bf or defined by the user. It is a function of y alone and has r independent column vectors. See bf, for detail.
numdir	The number of directions to be used in estimating the reduction subspace. The dimension must be less than or equal to the minimum of r and p. By default numdir=$\min\{r, p\}$.
structure	Structure of var(X|Y). The following options are available: "iso" for isotropic (predictors, conditionally on the response, are independent and on the same measurement scale); "aniso" for anisotropic (predictors, conditionally on the response, are independent and on different measurement scales); "unstr" for unstructured variance. The fourth structure "unstr2" refers to an extended PFC model with an heterogenous error structure.
eps_aniso	Precision term used in estimating var(X|Y) for the anisotropic structure.
numdir.test	Boolean. If FALSE, pfc fits with the numdir provided only. If TRUE, PFC models are fit for all dimensions less than or equal to numdir.
...	Additional arguments to Grassmannoptim.

Details

Let X be a column vector of p predictors, and Y be a univariate response variable. Principal fitted components model is an inverse regression model for sufficient dimension reduction. It is an inverse regression model given by $X|(Y = y) \sim N(\mu + \Gamma\beta f_y, \Delta)$. The term Δ is assumed independent of y. Its simplest structure is the isotropic (iso) with $\Delta = \delta^2 I_p$, where, conditionally on the response, the predictors are independent and are on the same measurement scale. The sufficient reduction is $\Gamma^T X$. The anisotropic (aniso) PFC model assumes that Δ =diag$(\delta_1^2, ..., \delta_p^2)$, where the conditional predictors are independent and on different measurement scales. The unstructured (unstr) PFC model allows a general structure for Δ. With the anisotropic and unstructured Δ, the sufficient reduction is $\Gamma^T\Delta^{-1}X$. it should be noted that $X \in R^p$ while the data-matrix to use is in $R^{n\times p}$.

The error structure of the extended structure has the following form

$$\Delta = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T,$$

where Γ_0 is the orthogonal completion of Γ such that (Γ, Γ_0) is a $p \times p$ orthogonal matrix. The matrices $\Omega \in R^{d\times d}$ and $\Omega_0 \in R^{(p-d)\times(p-d)}$ are assumed to be symmetric and full-rank. The sufficient reduction is $\Gamma^T X$. Let $\mathcal{S}_\Gamma$ be the subspace spanned by the columns of Γ. The parameter space of $\mathcal{S}_\Gamma$ is the set of all d dimensional subspaces in R^p, called Grassmann manifold and denoted by $\mathcal{G}_{(d,p)}$. Let $\hat{\Sigma}$, $\hat{\Sigma}_{\text{fit}}$ be the sample variance of X and the fitted covariance matrix, and let $\hat{\Sigma}_{\text{res}} = \hat{\Sigma} - \hat{\Sigma}_{\text{fit}}$. The MLE of $\mathcal{S}_\Gamma$ under unstr2 setup is obtained by maximizing the log-likelihood

$$L(\mathcal{S}_U) = -\log|U^T\hat{\Sigma}_{\text{res}}U| - \log|V^T\hat{\Sigma}V|$$

over $\mathcal{G}_{(d,p)}$, where V is an orthogonal completion of U.

The dimension d of the sufficient reduction must be estimated. A sequential likelihood ratio test is implemented as well as Akaike and Bayesian information criterion following Cook and Forzani (2008)

Value

This command returns a list object of class `ldr`. The output depends on the argument `numdir.test`. If `numdir.test=TRUE`, a list of matrices is provided corresponding to the `numdir` values (1 through `numdir`) for each of the parameters μ, β, Γ, Γ_0, Ω, and Ω_0. Otherwise, a single list of matrices for a single value of `numdir`. The outputs of `loglik`, `aic`, `bic`, `numpar` are vectors of `numdir` elements if `numdir.test=TRUE`, and scalars otherwise. Following are the components returned:

`R`	The reduction data-matrix of X obtained using the centered data-matrix X. The centering of the data-matrix of X is such that each column vector is centered around its sample mean.
`Muhat`	Estimate of μ.
`Betahat`	Estimate of β.
`Deltahat`	The estimate of the covariance Δ.
`Gammahat`	An estimated orthogonal basis representative of $\hat{\mathcal{S}}_\Gamma$, the subspace spanned by Γ.
`Gammahat0`	An estimated orthogonal basis representative of $\hat{\mathcal{S}}_{\Gamma_0}$, the subspace spanned by Γ_0.
`Omegahat`	The estimate of the covariance Ω if an extended model is used.
`Omegahat0`	The estimate of the covariance Ω_0 if an extended model is used.
`loglik`	The value of the log-likelihood for the model.
`aic`	Akaike information criterion value.
`bic`	Bayesian information criterion value.
`numdir`	The number of directions to estimate.
`numpar`	The number of parameters in the model.
`evalues`	The first `numdir` largest eigenvalues of $\hat{\Sigma}_{\text{fit}}$.

Author(s)

Kofi Placid Adragni <kofi@umbc.edu>

References

Adragni, KP and Cook, RD (2009): Sufficient dimension reduction and prediction in regression. Phil. Trans. R. Soc. A 367, 4385-4405.

Cook, RD (2007): Fisher Lecture - Dimension Reduction in Regression (with discussion). Statistical Science, 22, 1–26.

Cook, RD and Forzani, L (2008): Principal fitted components for dimension reduction in regression. Statistical Science 23, 485–501.

See Also

`core`, `lad`

Examples

```
data(bigmac)
fit1 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=bf(y=bigmac[,1], case="poly",
        degree=3),numdir=3, structure="aniso")
summary(fit1)
plot(fit1)

fit2 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=bf(y=bigmac[,1], case="poly",
        degree=3), numdir=3, structure="aniso", numdir.test=TRUE)
summary(fit2)
```

---

screen.pfc *Adaptive Screening of Predictors*

Description

Given a set of p predictors and a response, this function selects all predictors that are statistically related to the response at a specified significance level, using a flexible basis function.

Usage

```
screen.pfc(X, fy, cutoff=0.1)
```

Arguments

X	Matrix or data frame with n rows of observations and p columns of predictors of continuous type.
fy	Function of y. Basis function to be used to capture the dependency between individual predictors and the response. See bf for detail.
cutoff	The level of significance to be used for the cutoff, by default 0.1.

Details

For each predictor X_j, write the equation

$$X_j = \mu + \phi f_y + \epsilon$$

where f_y is a flexible basis function provided by the user. The basis function is constructed using the function bf. The screening procedure uses a test statistic on the null hypothesis $\phi = 0$ against the alternative $\phi \neq 0$. Given the r components of the basis function f_y, the above model is a linear model where X_j is the response and f_y constitutes the predictors. The hypothesis test on ϕ is essentially an F-test. Specifically, given the data, let $\hat{\phi}$ be the ordinary least squares estimator of ϕ. We consider the usual test statistic

$$F_j = \frac{n-r-1}{r} \cdot \frac{\sum_{i=1}^{n}[(X_{ji} - \bar{X}_{j\cdot})^2 - (X_{ji} - \bar{X}_{j\cdot} - \hat{\phi}_j \mathbf{f}_{y_i})^2]}{\sum_{i=1}^{n}(X_{ji} - \bar{X}_{j\cdot} - \hat{\phi}_j \mathbf{f}_{y_i})^2}$$

where $\bar{X}_{j\cdot} = \sum_{i=1}^{n} X_{ji}/n$. The statistic F_j follows an F distribution with $(r, n-r-1)$ degrees of freedom. The sample size n is expected to be larger than r.

Value

Return a data frame object with p rows corresponding to the variables with the following columns

F	F statistic for testing the above hypotheses.
P-value	The p-value of the test statistic. The F test has 1 and n-2 degrees of freedom
Index	Index of the variable, as its position j.

Author(s)

Kofi Placid Adragni <kofi@umbc.edu>

References

Adragni, KP and Cook, RD (2008) Discussion on the Sure Independence Screening for Ultrahigh Dimensional Feature Space of Jianqing Fan and Jinchi Lv (2007) Journal of the Royal Statistical Society Series B, 70, Part5, pp1:35

Examples

```
data(OH)
X <- OH[, -c(1,295)]; y=OH[,295]

# Correlation screening
out <- screen.pfc(X, fy=bf(y, case="poly", degree=1))
head(out)

# Special basis function
out1 <- screen.pfc(X, fy=scale(cbind(y, sqrt(y)), center=TRUE, scale=FALSE))
head(out1)

# Piecewise constant basis with 10 slices
out2 <- screen.pfc(X, fy=bf(y, case="pdisc", degree=0, nslices=10))
head(out2)
```

---

snakes *Snakes data*

Description

Genetic covariance matrices for six genetic traits of two female garter snake populations, one from a coastal and the other from inland site in northern California. The data set was initially studied by Phillips and Arnold (1999).

Usage

```
data(snakes)
```

Format

List format of 3 components.

`snakes[[1]]` sample genetic covariance matrix for the inland population, obtained.

`snakes[[2]]` sample genetic covariance matrix for the coastal population.

`snakes[[3]]` vector of sample sizes, respectively for inland and coastal samples.

Details

Both genetic variance-covariances are obtained on six traits of the snakes.

References

Phillips P, Arnold S (1999). Hierarchical Comparison of Genetic variance-Covariance matrix using the Flury Hierarchy." Evolution, 53, 1506–1515.

Examples

```
data(snakes)
```

`structure.test` *Test of covariance structure for PFC models*

Description

Information criterion and likelihood ratio test for the structure of the covariance matrix of PFC models.

Usage

```
structure.test(object1, object2)
```

Arguments

`object1`	An object of class `pfc`
`object2`	A second object of class `pfc`, fitted exactly as for `object1` except for the covariance structure Δ.

Details

Consider two PFC models M_1 and M_2, with the same parameters, except for the conditional covariance that is Δ_1 for M_1 and Δ_2 for M_2 such that model M_1 is nested in model M_2. We implemented the likelihood ratio test for the hypotheses: $H_0 : \Delta = \Delta_1$ versus $H_a : \Delta = \Delta_2$. The test is implemented for the isotropic, anisotropic, and the unstructured PFC models. One may test isotropic against either anisotropic or unstructured, or test anisotropic against unstructured. The degrees of freedom are given by the difference in the number of parameters in the covariances. Information criterion AIC and BIC are also provided.

Author(s)

Kofi Placid Adragni <kofi@umbc.edu>

Examples

```
data(bigmac)
fit1 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=bf(y=bigmac[,1], case="poly",
        degree=3), numdir=3, structure="iso")
fit2 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=bf(y=bigmac[,1], case="poly",
        degree=3), numdir=3, structure="aniso")
fit3 <- pfc(X=bigmac[,-1], y=bigmac[,1], fy=bf(y=bigmac[,1], case="poly",
        degree=3), numdir=3, structure="unstr")
structure.test(fit1, fit3)
structure.test(fit2, fit3)
```

Index