Package ‘aroma.light’

January 18, 2020

Version 3.16.0

Depends R (>= 2.15.2)

Imports stats, R.methodsS3 (>= 1.7.1), R.oo (>= 1.22.0), R.utils (>= 2.9.0), matrixStats (>= 0.54.0)

Suggests princurve (>= 2.1.4)

Date 2019-08-28

Title Light-Weight Methods for Normalization and Visualization of Microarray Data using Only Basic R Data Types

Description Methods for microarray analysis that take basic data types such as matrices and lists of vectors. These methods can be used standalone, be utilized in other packages, or be wrapped up in higher-level classes.

License GPL (>= 2)

biocViews Infrastructure, Microarray, OneChannel, TwoChannel, MultiChannel, Visualization, Preprocessing


BugReports https://github.com/HenrikBengtsson/aroma.light/issues

LazyLoad TRUE

Encoding latin1

git_url https://git.bioconductor.org/packages/aroma.light

git_branch RELEASE_3_10

git_last_commit fc16179

git_last_commit_date 2019-10-29

Date/Publication 2020-01-17

Author Henrik Bengtsson [aut, cre, cph], Pierre Neuvial [ctb], Aaron Lun [ctb]

Maintainer Henrik Bengtsson <henrikb@braju.com>


R topics documented:

aroma.light-package .............................................. 2
1. Calibration and Normalization ................................. 4
averageQuantile .................................................... 7
backtransformAffine ............................................... 8
backtransformPrincipalCurve .................................... 9
calibrateMultiscan ............................................... 12
callNaiveGenotypes ........................................... 15
distanceBetweenLines ......................................... 16
fitIWPCA ............................................................... 19
fitNaiveGenotypes ............................................... 20
fitPrincipalCurve ............................................... 21
fitXYCurve ......................................................... 23
iwPCA ................................................................. 25
medianPolish .......................................................... 27
normalizeAffine .................................................... 29
normalizeAverage .................................................. 33
normalizeCurveFit ............................................... 34
normalizeDifferencesToAverage ............................... 37
normalizeFragmentLength .................................... 39
normalizeQuantileRank .......................................... 45
normalizeQuantileRank.matrix ................................. 46
normalizeQuantileSpline ....................................... 48
normalizeTumorBoost ........................................... 49
pairedAlleleSpecificCopyNumbers ............................ 51
plotDensity .......................................................... 52
plotMvsA .............................................................. 53
plotMvsAPairs ..................................................... 54
plotMvsMPairs .................................................... 55
plotXYCurve ......................................................... 55
robustSmoothSpline ............................................. 56
sampleCorrelations .............................................. 58
sampleTuples ....................................................... 59
wPCA ................................................................. 60

Index 64

aroma.light-package Package aroma.light

Description

Methods for microarray analysis that take basic data types such as matrices and lists of vectors. These methods can be used standalone, be utilized in other packages, or be wrapped up in higher-level classes.

Installation

To install this package, see http://bioconductor.org/packages/release/bioc/html/aroma.light.html.
To get started

For scanner calibration:

1. see \texttt{calibrateMultiscan} - scan the same array two or more times to calibrate for scanner effects and extended dynamical range.

To normalize multiple single-channel arrays all with the same number of probes/spots:

1. \texttt{normalizeAffine} - normalizes, on the intensity scale, for differences in offset and scale between channels.
2. \texttt{normalizeQuantileRank}, \texttt{normalizeQuantileSpline} - normalizes, on the intensity scale, for differences in empirical distribution between channels.

To normalize multiple single-channel arrays with varying number probes/spots:

1. \texttt{normalizeQuantileRank}, \texttt{normalizeQuantileSpline} - normalizes, on the intensity scale, for differences in empirical distribution between channels.

To normalize two-channel arrays:

1. \texttt{normalizeAffine} - normalizes, on the intensity scale, for differences in offset and scale between channels. This will also correct for intensity-dependent affects on the log scale.
2. \texttt{normalizeCurveFit} - Classical intensity-dependent normalization, on the log scale, e.g. loess normalization.

To normalize three or more channels:

1. \texttt{normalizeAffine} - normalizes, on the intensity scale, for differences in offset and scale between channels. This will minimize the curvature on the log scale between any two channels.

Further readings

Several of the normalization methods proposed in [1]-[7] are available in this package.

How to cite this package

Whenever using this package, please cite one or more of [1]-[7].

Wishlist

Here is a list of features that would be useful, but which I have too little time to add myself. Contributions are appreciated.

• At the moment, nothing.

If you consider to contribute, make sure it is not already implemented by downloading the latest "devel" version!

License

The releases of this package is licensed under GPL version 2 or newer.

NB: Except for the \texttt{robustSmoothSpline} method, it is alright to distribute the rest of the package under LGPL version 2.1 or newer.

The development code of the packages is under a private licence (where applicable) and patches sent to the author fall under the latter license, but will be, if incorporated, released under the "release" license above.
1. Calibration and Normalization

Description

In this section we give our recommendation on how spotted two-color (or multi-color) microarray data is best calibrated and normalized.

Author(s)
Henrik Bengtsson, Pierre Neuvial, Aaron Lun

References

Some of the reference below can be found at http://www.aroma-project.org/publications/.


1. Calibration and Normalization

Classical background subtraction

We do not recommend background subtraction in classical means where background is estimated by various image analysis methods. This means that we will only consider foreground signals in the analysis.

We estimate "background" by other means. In what is explain below, only a global background, that is, a global bias, is estimated and removed.

Multiscan calibration

In Bengtsson et al (2004) we give evidence that microarray scanners can introduce a significant bias in data. This bias, which is about 15-25 out of 65535, will introduce intensity dependency in the log-ratios, as explained in Bengtsson & Hössjer (2006).

In Bengtsson et al (2004) we find that this bias is stable across arrays (and a couple of months), but further research is needed in order to tell if this is true over a longer time period.

To calibrate signals for scanner biases, scan the same array at multiple PMT-settings at three or more (K >= 3) different PMT settings (preferably in decreasing order). While doing this, do not adjust the laser power settings. Also, do the multiscan without washing, cleaning or by other means changing the array between subsequent scans. Although not necessary, it is preferred that the array remains in the scanner between subsequent scans. This will simplify the image analysis since spot identification can be made once if images aligns perfectly.

After image analysis, read all K scans for the same array into the two matrices, one for the red and one for the green channel, where the K columns corresponds to scans and the N rows to the spots. It is enough to use foreground signals.

In order to multiscan calibrate the data, for each channel separately call \( X_c \leftarrow \text{calibrateMultiscan}(X) \) where \( X \) is the NxK matrix of signals for one channel across all scans. The calibrated signals are returned in the Nx1 matrix \( X_c \).

Multiscan calibration may sometimes be skipped, especially if affine normalization is applied immediately after, but we do recommend that every lab check at least once if their scanner introduce bias. If the offsets in a scanner is already estimated from earlier multiscan analyses, or known by other means, they can readily be subtracted from the signals of each channel. If arrays are still multiscanned, it is possible to force the calibration method to fit the model with zero intercept (assuming the scanner offsets have been subtracted) by adding argument center=FALSE.

Affine normalization

In Bengtsson & Hössjer (2006), we carry out a detailed study on how biases in each channel introduce so called intensity-dependent log-ratios among other systematic artifacts. Data with (additive) bias in each channel is said to be affinely transformed. Data without such bias, is said to be linearly (proportionally) transform. Ideally, observed signals (data) is a linear (proportional) function of true gene expression levels.

We do not assume proportional observations. The scanner bias is real evidence that assuming linearity is not correct. Affine normalization corrects for affine transformation in data. Without control spots it is not possible to estimate the bias in each of the channels but only the relative bias such that after normalization the effective bias are the same in all channels. This is why we call it normalization and not calibration.

In its simplest form, affine normalization is done by \( X_n \leftarrow \text{normalizeAffine}(X) \) where \( X \) is a Nx2 matrix with the first column holds the foreground signals from the red channel and the second holds the signals from the green channel. If three- or four-channel data is used these are added the same way. The normalized data is returned as a Nx2 matrix \( X_n \).
To normalize all arrays and all channels at once, one may put all data into one big NxK matrix where the K columns hold the all channels from the first array, then all channels from the second array and so on. Then $X_n \leftarrow \text{normalizeAffine}(X)$ will return the across-array and across-channel normalized data in the NxK matrix $X_n$ where the columns are stored in the same order as in matrix $X$.

Equal effective bias in all channels is much better. First of all, any intensity-dependent bias in the log-ratios is removed for all non-differentially expressed genes. There is still an intensity-dependent bias in the log-ratios for differentially expressed genes, but this is now symmetric around log-ratio zero.

Affine normalization will (by default and recommended) normalize all arrays together and at once. This will guarantee that all arrays are “on the same scale”. Thus, it not recommended to apply a classical between-array scale normalization afterward. Moreover, the average log-ratio will be zero after an affine normalization.

Note that an affine normalization will only remove curvature in the log-ratios at lower intensities. If a strong intensity-dependent bias at high intensities remains, this is most likely due to saturation effects, such as too high PMT settings or quenching.

Note that for a perfect affine normalization you should expect much higher noise levels in the log-ratios at lower intensities than at higher. It should also be approximately symmetric around zero log-ratio. In other words, a strong fanning effect is a good sign.

Due to different noise levels in red and green channels, different PMT settings in different channels, plus the fact that the minimum signal is zero, “odd shapes” may be seen in the log-ratio vs log-intensity graphs at lower intensities. Typically, these show themselves as non-symmetric in positive and negative log-ratios. Note that you should not see this at higher intensities.

If there is a strong intensity-dependent effect left after the affine normalization, we recommend, for now, that a subsequent curve-fit or quantile normalization is done. Which one, we do not know.

Why negative signals? By default, 5% of the normalized signals will have a non-positive signal in one or both channels. This is on purpose, although the exact number 5% is chosen by experience. The reason for introducing negative signals is that they are indeed expected. For instance, when measure a zero gene expression level, there is a chance that the observed value is (should be) negative due to measurement noise. (For this reason it is possible that the scanner manufacturers have introduced scanner bias on purpose to avoid negative signals, which then all would be truncated to zero.) To adjust the ratio (or number) of negative signals allowed, use for example normalizeAffine($X$, constraint=0.01) for 1% negative signals. If set to zero (or "max") only as much bias is removed such that no negative signals exist afterward. Note that this is also true if there were negative signals on beforehand.

Why not lowess normalization? Curve-fit normalization methods such as lowess normalization are basically designed based on linearity assumptions and will for this reason not correct for channel biases. Curve-fit normalization methods can by definition only be applied to one pair of channels at the time and do therefore require a subsequent between-array scale normalization, which is by the way very ad hoc.

Why not quantile normalization? Affine normalization can be though of a special case of quantile normalization that is more robust than the latter. See Bengtsson & Hössjer (2006) for details. Quantile normalization is probably better to apply than curve-fit normalization methods, but less robust than affine normalization, especially at extreme (low and high) intensities. For this reason, we do recommend to use affine normalization first, and if this is not satisfactory, quantile normalization may be applied.
Linear (proportional) normalization

If the channel offsets are zero, already corrected for, or estimated by other means, it is possible to normalize the data robustly by fitting the above affine model without intercept, that is, fitting a truly linear model. This is done done adding argument center=FALSE when calling normalizeAffine().

Author(s)

Henrik Bengtsson

averageQuantile

Gets the average empirical distribution

Description

Gets the average empirical distribution for a set of samples.

Usage

```r
## S3 method for class 'list'
averageQuantile(X, ...)
## S3 method for class 'matrix'
averageQuantile(X, ...)
```

Arguments

- `X`: A `list` with `K` numeric vectors, or a numeric `NxK` matrix. If a `list`, the vectors may be of different lengths.
- `...`: Not used.

Value

Returns a numeric vector of length equal to the longest vector in argument `X`.

Missing values

Missing values are excluded.

Author(s)

Parts adopted from Gordon Smyth (http://www.statsci.org/) in 2002 \& 2006. Original code by Ben Bolstad at Statistics Department, University of California.

See Also

- `normalizeQuantileRank`
- `normalizeQuantileSpline`
- `quantile`
backtransformAffine

Reverse affine transformation

Description
Reverse affine transformation.

Usage
## S3 method for class 'matrix'
backtransformAffine(X, a=NULL, b=NULL, project=FALSE, ...)

Arguments
X
An NxK matrix containing data to be backtransformed.

a
A scalar, vector, a matrix, or a list. First, if a list, it is assumed to contain the elements a and b, which are the used as if they were passed as separate arguments. If a vector, a matrix of size NxK is created which is then filled row by row with the values in the vector. Commonly, the vector is of length K, which means that the matrix will consist of copies of this vector stacked on top of each other. If a matrix, a matrix of size NxK is created which is then filled column by column with the values in the matrix (collected column by column). Commonly, the matrix is of size NxK, or NxL with L < K and then the resulting matrix consists of copies sitting next to each other. The resulting NxK matrix is subtracted from the NxK matrix X.

b
A scalar, vector, a matrix. A NxK matrix is created from this argument. For details see argument a. The NxK matrix X-a is divided by the resulting NxK matrix.

project
returned (K values per data point are returned). If TRUE, the backtransformed values "(X-a)/b" are projected onto the line L(a,b) so that all columns will be identical.

...
Not used.

Value
The ")(X-a)/b" backtransformed NxK matrix is returned. If project is TRUE, an N1 matrix is returned, because all columns are identical anyway.

Missing values
Missing values remain missing values. If projected, data points that contain missing values are projected without these.

Examples
X <- matrix(1:8, nrow=4, ncol=2)
X[2,2] <- NA

print(X)

# Returns a 4x2 matrix
backtransformPrincipalCurve

Reverse transformation of principal-curve fit

Description
Reverse transformation of principal-curve fit.

Usage

## S3 method for class 'matrix'
backtransformPrincipalCurve(X, fit, dimensions=NULL, targetDimension=NULL, ...)

## S3 method for class 'numeric'
backtransformPrincipalCurve(X, ...)

print(backtransformAffine(X, a=c(1,5)))
# Returns a 4x2 matrix
print(backtransformAffine(X, b=c(1,1/2)))
# Returns a 4x2 matrix
print(backtransformAffine(X, a=matrix(1:4,ncol=1)))
# Returns a 4x2 matrix
print(backtransformAffine(X, a=matrix(1:3,ncol=1)))
# Returns a 4x2 matrix
print(backtransformAffine(X, a=matrix(1:2,ncol=1), b=c(1,2)))
# Returns a 4x1 matrix
print(backtransformAffine(X, b=c(1,1/2), project=TRUE))
# If the columns of X are identical, and a identity
# backtransformation is applied and projected, the
# same matrix is returned.
X <- matrix(1:4, nrow=4, ncol=3)
Y <- backtransformAffine(X, b=c(1,1,1), project=TRUE)
print(X)
print(Y)
stopifnot(sum(X[,1]-Y) <= .Machine$double.eps)

# If the columns of X are identical, and a identity
# backtransformation is applied and projected, the
# same matrix is returned.
X <- matrix(1:4, nrow=4, ncol=3)
X[,2] <- X[,2]*2; X[,3] <- X[,3]*3
print(X)
Y <- backtransformAffine(X, b=c(1,2,3))
print(Y)
Y <- backtransformAffine(X, b=c(1,2,3), project=TRUE)
print(Y)
stopifnot(sum(X[,1]-Y) <= .Machine$double.eps)
Arguments

X  An NxK **matrix** containing data to be backtransformed.

fit  An MxL principal-curve fit object of class principal_curve as returned by fitPrincipalCurve(). Typically \( L = K \), but not always.

dimensions  An (optional) subset of D dimensions all in \([1,L]\) to be returned (and back-transform).

targetDimension  An (optional) index specifying the dimension in \([1,L]\) to be used as the target dimension of the fit. More details below.

...  Passed internally to **smooth.spline**.

Details

Each column in X ("dimension") is backtransformed independently of the others.

Value

The backtransformed NxK (or NxD) **matrix**.

**Target dimension**

By default, the backtransform is such that afterward the signals are approximately proportional to the (first) principal curve as fitted by fitPrincipalCurve(). This scale and origin of this principal curve is not uniquely defined. If targetDimension is specified, then the backtransformed signals are approximately proportional to the signals of the target dimension, and the signals in the target dimension are unchanged.

**Subsetting dimensions**

Argument dimensions can be used to backtransform a subset of dimensions (K) based on a subset of the fitted dimensions (L). If \( K = L \), then both X and fit is subsetted. If \( K \neq L \), then it is assumed that X is already subsetted/expanded and only fit is subsetted.

See Also

fitPrincipalCurve()

Examples

```r
# Consider the case where K=4 measurements have been done
# for the same underlying signals 'x'. The different measurements
# have different systematic variation
#
# y_k = f(x_k) + eps_k; k = 1, ..., K.
# In this example, we assume non-linear measurement functions
#
# f(x) = a + b*x + x^c + eps(b*x)
#
# where 'a' is an offset, 'b' a scale factor, and 'c' an exponential.
# We also assume heteroscedastic zero-mean noise with standard
deviation proportional to the rescaled underlying signal 'x'.
```
Furthermore, we assume that measurements \( k=2 \) and \( k=3 \) undergo the same transformation, which may illustrate that they come from the same batch. However, when fitting the model below we will assume they are independent.

Transforms
\[
\begin{align*}
a &\leftarrow c(2, 15, 15, 3) \\
b &\leftarrow c(2, 3, 3, 4) \\
c &\leftarrow c(1, 2, 2, 1/2) \\
K &\leftarrow \text{length}(a)
\end{align*}
\]

The true signal
\[
\begin{align*}
N &\leftarrow 1000 \\
x &\leftarrow \text{rexp}(N)
\end{align*}
\]

The noise
\[
\begin{align*}
bX &\leftarrow \text{outer}(b,x) \\
E &\leftarrow \text{apply}(bX, \text{MARGIN}=2, \text{FUN}=\text{function}(x) \text{rnorm}(K, \text{mean}=0, \text{sd}=0.1*x))
\end{align*}
\]

The transformed signals with noise
\[
\begin{align*}
Xc &\leftarrow t(\text{sapply}(c, \text{FUN}=\text{function}(c) x^c)) \\
Y &\leftarrow a + bX + Xc + E \\
Y &\leftarrow t(Y)
\end{align*}
\]

Fit principal curve
\[
\begin{align*}
\text{fit} &\leftarrow \text{fitPrincipalCurve}(Y) \\
\text{flip} &\leftarrow (\text{rho} < 0) \\
\text{if} (\text{flip}) \\
&\begin{align*}
\text{fit$lambda} &\leftarrow \max(\text{fit$lambda}, \text{na.rm=TRUE})-\text{fit$lambda}
\end{align*}
\end{align*}
\]

Backtransform data according to model fit
\[
\begin{align*}
\text{YN1} &\leftarrow \text{backtransform Principal Curve}(Y, \text{fit=fit}) \\
\text{stopifnot}(\text{ncol}(\text{YN1}) == K)
\end{align*}
\]

Backtransform toward the first dimension
\[
\begin{align*}
\text{YN2} &\leftarrow \text{backtransform Principal Curve}(Y, \text{fit=fit}, \text{targetDimension=1}) \\
\text{stopifnot}(\text{ncol}(\text{YN2}) == K)
\end{align*}
\]

Backtransform toward the last (fitted) dimension
\[
\begin{align*}
\text{YN3} &\leftarrow \text{backtransform Principal Curve}(Y, \text{fit=fit}, \text{targetDimension=L})
\end{align*}
\]
calibrateMultiscan

Weighted affine calibration of a multiple re-scanned channel

description

Weighted affine calibration of a multiple re-scanned channel.

Usage

# S3 method for class 'matrix'
calibrateMultiscan(X, weights=NULL, typeOfWeights=c("datapoint"), method="L1", constraint="diagonal", satSignal=2^16 - 1, ..., average=median, deviance=NULL, project=FALSE, .fitOnly=FALSE)
Arguments

- **X**: An NxK matrix (K>=2) where the columns represent the multiple scans of one channel (a two-color array contains two channels) to be calibrated.
- **weights**: If NULL, non-weighted normalization is done. If data-point weights are used, this should be a vector of length N of data point weights used when estimating the normalization function.
- **typeOfWeights**: A character string specifying the type of weights given in argument weights.
- **method**: A character string specifying how the estimates are robustified. See iwpca() for all accepted values.
- **constraint**: Constraint making the bias parameters identifiable. See fitIWPCA() for more details.
- **satSignal**: Signals equal to or above this threshold is considered saturated signals.
- **average**: A function to calculate the average signals between calibrated scans.
- **deviance**: A function to calculate the deviance of the signals between calibrated scans.
- **project**: If TRUE, the calibrated data points projected onto the diagonal line, otherwise not. Moreover, if TRUE, argument average is ignored.
- **.fitOnly**: If TRUE, the data will not be back-transform.

Details

Fitting is done by iterated re-weighted principal component analysis (IWPCA).

Value

If average is specified or project is TRUE, an N x 1 matrix is returned, otherwise an N x K matrix is returned. If deviance is specified, a deviance N x 1 matrix is returned as attribute deviance. In addition, the fitted model is returned as attribute modelFit.

Negative, non-positive, and saturated values

Affine multiscan calibration applies also to negative values, which are therefore also calibrated, if they exist.

Saturated signals in any scan are set to NA. Thus, they will not be used to estimate the calibration function, nor will they affect an optional projection.

Missing values

Only observations (rows) in X that contain all finite values are used in the estimation of the calibration functions. Thus, observations can be excluded by setting them to NA.

Weighted normalization

Each data point/observation, that is, each row in X, which is a vector of length K, can be assigned a weight in [0,1] specifying how much it should affect the fitting of the calibration function. Weights are given by argument weights, which should be a numeric vector of length N. Regardless of weights, all data points are calibrated based on the fitted calibration function.
Robustness

By default, the model fit of multiscan calibration is done in $L_1$ (method="L1"). This way, outliers affect the parameter estimates less than ordinary least-square methods.

When calculating the average calibrated signal from multiple scans, by default the median is used, which further robustify against outliers.

For further robustness, downweight outliers such as saturated signals, if possible.

Tukey’s biweight function is supported, but not used by default because then a "bandwidth" parameter has to be selected. This can indeed be done automatically by estimating the standard deviation, for instance using MAD. However, since scanner signals have heteroscedastic noise (standard deviation is approximately proportional to the non-logged signal), Tukey’s bandwidth parameter has to be a function of the signal too, cf. loess. We have experimented with this too, but found that it does not significantly improve the robustness compared to $L_1$. Moreover, using Tukey’s biweight as is, that is, assuming homoscedastic noise, seems to introduce a (scale dependent) bias in the estimates of the offset terms.

Using a known/Previously estimated offset

If the scanner offsets can be assumed to be known, for instance, from prior multiscan analyses on the scanner, then it is possible to fit the scanner model with no (zero) offset by specifying argument center=FALSE. Note that you cannot specify the offset. Instead, subtract it from all signals before calibrating, e.g. Xc <- calibrateMultiscan(X-e,center=FALSE) where e is the scanner offset (a scalar). You can assert that the model is fitted without offset by stopifnot(all(attr(Xc,"modelFit")$adiag == 0)).

Author(s)

Henrik Bengtsson

References


See Also

1. Calibration and Normalization. normalizeAffine().

Examples

## Not run: # For an example, see help(normalizeAffine).
callNaiveGenotypes

Calls genotypes in a normal sample

Description

Calls genotypes in a normal sample.

Usage

```r
## S3 method for class 'numeric'
callNaiveGenotypes(y, cn=rep(2L, times = length(y)), ..., modelFit=NULL, verbose=FALSE)
```

Arguments

- `y`: A numeric vector of length J containing allele B fractions for a normal sample.
- `cn`: An optional numeric vector of length J specifying the true total copy number in \{0, 1, 2, \text{NA}\} at each locus. This can be used to specify which loci are diploid and which are not, e.g. autosomal and sex chromosome copy numbers.
- `...`: Additional arguments passed to `fitNaiveGenotypes()`.
- `modelFit`: A optional model fit as returned by `fitNaiveGenotypes()`.
- `verbose`: A logical or a Verbose object.

Value

Returns a numeric vector of length J containing the genotype calls in allele B fraction space, that is, in [0,1] where 1/2 corresponds to a heterozygous call, and 0 and 1 corresponds to homozygous A and B, respectively. Non called genotypes have value NA.

Missing and non-finite values

A missing value always gives a missing (NA) genotype call. Negative infinity (-Inf) always gives genotype call 0. Positive infinity (+Inf) always gives genotype call 1.

Author(s)

Henrik Bengtsson

See Also

Internally `fitNaiveGenotypes()` is used to identify the thresholds.

Examples

```r
layout(matrix(1:3, ncol=1))
par(mar=c(2,4,4,1)+0.1)

# A bimodal distribution
xAA <- rnorm(n=10000, mean=0, sd=0.1)
xBB <- rnorm(n=10000, mean=1, sd=0.1)
```
distanceBetweenLines

Finds the shortest distance between two lines.

distanceBetweenLines

Finds the shortest distance between two lines.
Consider the two lines
\[ x(s) = a_x + b_x s \quad \text{and} \quad y(t) = a_y + b_y t \]
in an K-space where the offset and direction vectors are \( a_x \) and \( b_x \) (in \( \mathbb{R}^K \)) that define the line \( x(s) \) (\( s \) is a scalar). Similar for the line \( y(t) \). This function finds the point \((s, t)\) for which \(|x(s) - x(t)|\) is minimal.
Usage

### Default S3 method:
distanceBetweenLines(ax, bx, ay, by, ...)

Arguments

- **ax, bx**: Offset and direction vector of length K for line \( z_x \).
- **ay, by**: Offset and direction vector of length K for line \( z_y \).
- **...**: Not used.

Value

Returns the a list containing

- **ax, bx**: The given line \( x(s) \).
- **ay, by**: The given line \( y(t) \).
- **s, t**: The values of \( s \) and \( t \) such that \( |x(s) - y(t)| \) is minimal.
- **xs, yt**: The values of \( x(s) \) and \( y(t) \) at the optimal point \( (s, t) \).
- **distance**: The distance between the lines, i.e. \( |x(s) - y(t)| \) at the optimal point \( (s, t) \).

Author(s)

Henrik Bengtsson

References


Examples

```r
for (zzz in 0) {
  # This example requires plot3d() in R.basic [http://www.braju.com/R/]
  if (!require(pkgName <- "R.basic", character.only=TRUE)) break

  layout(matrix(1:4, nrow=2, ncol=2, byrow=TRUE))

  # Lines in two-dimensions
  x <- list(a=c(1,0), b=c(1,2))
  y <- list(a=c(0,2), b=c(1,1))
  fit <- distanceBetweenLines(ax=x$a, bx=x$b, ay=y$a, by=y$b)

  xlim <- ylim <- c(-1,8)
  plot(NA, xlab="", ylab="", xlim=ylim, ylim=ylim)

  # Highlight the offset coordinates for both lines
  points(t(x$a), pch="+", col="red")
```
text(t(x$a), label=expression(a(x)), adj=c(-1,0.5))
points(t(y$a), pch="+", col="blue")

text(t(y$a), label=expression(a(y)), adj=c(-1,0.5))

v <- c(-1,1)*10;
xv <- list(x=x$a[1]+x$b[1]*v, y=x$a[2]+x$b[2]*v)
yv <- list(x=y$a[1]+y$b[1]*v, y=y$a[2]+y$b[2]*v)

lines(xv, col="red")
lines(yv, col="blue")

points(t(fit$xs), cex=2.0, col="red")
text(t(fit$xs), label=expression(x(s)), adj=c(+2,0.5))
points(t(fit$yt), cex=1.5, col="blue")
text(t(fit$yt), label=expression(y(t)), adj=c(-1,0.5))
print(fit)

############################################################
# Lines in three-dimensions
############################################################
x <- list(a=c(0,0,0), b=c(1,1,1)) # The 'diagonal'
y <- list(a=c(2,1,2), b=c(2,1,3)) # A 'fitted' line
fit <- distanceBetweenLines(ax=x$a, bx=x$b, ay=y$a, by=y$b)

xlim <- ylim <- zlim <- c(-1,3)
dummy <- t(c(1,1,1))*100;
# Coordinates for the lines in 3d
v <- seq(-10,10, by=1);
xv <- list(x=x$a[1]+x$b[1]*v, y=x$a[2]+x$b[2]*v, z=x$a[3]+x$b[3]*v)

for (theta in seq(30,140,length.out=3)) {
  plot3d(dummy, theta=theta, phi=30, xlab="", ylab="", zlab="",
         xlim=ylim, ylim=ylim, zlim=zlim)
  # Highlight the offset coordinates for both lines
  points3d(t(x$a), pch="+", col="red")
text3d(t(x$a), label=expression(a(x)), adj=c(-1,0.5))
  points3d(t(y$a), pch="+", col="blue")
text3d(t(y$a), label=expression(a(y)), adj=c(-1,0.5))

  # Draw the lines
  lines3d(xv, col="red")
  lines3d(yv, col="blue")

  # Draw the two points that are closest to each other
  points3d(t(fit$xs), cex=2.0, col="red")
text3d(t(fit$xs), label=expression(x(s)), adj=c(+2,0.5))
  points3d(t(fit$yt), cex=1.5, col="blue")
text3d(t(fit$yt), label=expression(y(t)), adj=c(-1,0.5))

  # Draw the distance between the two points
  lines3d(rbind(fit$xs,fit$yt), col="purple", lwd=2)
}

############################################################
# Lines in three-dimensions
############################################################
Robust fit of linear subspace through multidimensional data

Description

Robust fit of linear subspace through multidimensional data.

Usage

```r
## S3 method for class 'matrix'
fitIWPCA(X, constraint=c("diagonal", "baseline", "max"), baselineChannel=NULL, ..., 
aShift=rep(0, times = ncol(X)), Xmin=NULL)
```

Arguments

- `X`  
  N x K matrix where N is the number of observations and K is the number of dimensions (channels).
- `constraint`  
  A character string or a numeric value. If character it specifies which additional constraint to be used to specify the offset parameters along the fitted line;
  - If "diagonal", the offset vector will be a point on the line that is closest to the diagonal line (1,...,1). With this constraint, all bias parameters are identifiable.
  - If "baseline" (requires argument baselineChannel), the estimates are such that the bias and scale parameters of the baseline channel is 0 and 1, respectively. With this constraint, all bias parameters are identifiable.
  - If "max", the offset vector will be the point on the line that is as "great" as possible, but still such that each of its components is less than the corresponding minimal signal. This will guarantee that no negative signals are created in the backward transformation. If numeric value, the offset vector will be the point on the line such that after applying the backward transformation there are constraint*N. Note that constraint=0 corresponds approximately to constraint="max". With the latter two constraints, the bias parameters are only identifiable modulo the fitted line.
- `baselineChannel`  
  Index of channel toward which all other channels are conform. This argument is required if constraint="baseline". This argument is optional if constraint="diagonal" and then the scale factor of the baseline channel will be one. The estimate of the bias parameters is not affected in this case. Defaults to one, if missing.
- `...`  
  Additional arguments accepted by iwpca(). For instance, a N vector of weights for each observation may be given, otherwise they get the same weight.
- `aShift, Xmin`  
  For internal use only.
Details

This method uses re-weighted principal component analysis (IWPCA) to fit a the model $y_n = a + bx_n + eps_n$ where $y_n$, $a$, $b$, and $eps_n$ are vector of the $K$ and $x_n$ is a scalar.

The algorithm is: For iteration $i$: 1) Fit a line $L$ through the data close using weighted PCA with weights $\{w_n\}$. Let $r_n = \{r_{n,1}, ..., r_{n,K}\}$ be the $K$ principal components. 2) Update the weights as $w_n < -1/\sum_{k=2}^{K} (r_{n,k} + \epsilon_k)$ where we have used the residuals of all but the first principal component. 3) Find the point $a$ on $L$ that is closest to the line $D = (1, 1, ..., 1)$. Similarly, denote the point on $D$ that is closest to $L$ by $t = a * (1, 1, ..., 1)$.

Value

Returns a list that contains estimated parameters and algorithm details;

- **a**: A double vector $(a[1], ..., a[K])$ with offset parameter estimates. It is made identifiable according to argument constraint.
- **b**: A double vector $(b[1], ..., b[K])$ with scale parameter estimates. It is made identifiable by constraining $b[baselineChannel] == 1$. These estimates are independent of argument constraint.
- **adiag**: If identifiability constraint "diagonal", a double vector $(adiag[1], ..., adiag[K])$, where $adiag[1] = adiag[2] = ...adiag[K]$, specifying the point on the diagonal line that is closest to the fitted line, otherwise the zero vector.
- **eigen**: A KxK matrix with columns of eigenvectors.
- **converged**: TRUE if the algorithm converged, otherwise FALSE.
- **nbrOfIterations**: The number of iterations for the algorithm to converge, or zero if it did not converge.
- **t0**: Internal parameter estimates, which contains no more information than the above listed elements.
- **t**: Always NULL.

Author(s)

Henrik Bengtsson

See Also

This is an internal method used by the calibrateMultiscan() and normalizeAffine() methods. Internally the function iwpca() is used to fit a line through the data cloud and the function distanceBetweenLines() to find the closest point to the diagonal $(1,1, ..., 1)$.

---

**fitNaiveGenotypes**

*Fit naive genotype model from a normal sample*

Description

Fit naive genotype model from a normal sample.
Usage

```r
## S3 method for class 'numeric'
fitNaiveGenotypes(y, cn=rep(2L, times = length(y)), subsetToFit=NULL,
  flavor=c("density", "fixed"), adjust=1.5, ..., censorAt=c(-0.1, 1.1), verbose=FALSE)
```

Arguments

- **y**: A numeric vector of length J containing allele B fractions for a normal sample.
- **cn**: An optional numeric vector of length J specifying the true total copy number in \{0, 1, 2, N.A\} at each locus. This can be used to specify which loci are diploid and which are not, e.g. autosomal and sex chromosome copy numbers.
- **subsetToFit**: An optional integer or logical vector specifying which loci should be used for estimating the model. If NULL, all loci are used.
- **flavor**: A character string specifying the type of algorithm used.
- **adjust**: A positive double specifying the amount smoothing for the empirical density estimator.
- **...**: Additional arguments passed to `findPeaksAndValleys()`.
- **censorAt**: A double vector of length two specifying the range for which values are considered finite. Values below (above) this range are treated as -Inf (+Inf).
- **verbose**: A logical or a Verbose object.

Value

Returns a list of lists.

Author(s)

Henrik Bengtsson

See Also

To call genotypes see `callNaiveGenotypes()`. Internally `findPeaksAndValleys()` is used to identify the thresholds.

---

**Description**

Fit a principal curve in K dimensions.

**Usage**

```r
## S3 method for class 'matrix'
fitPrincipalCurve(X, ..., verbose=FALSE)
```
Arguments

X       An N\times K matrix (K\geq 2) where the columns represent the dimension.
...
verbose A logical or a Verbose object.

Value

Returns a principal_curve object (which is a list). See principal_curve for more details.

Missing values

The estimation of the normalization function will only be made based on complete observations, i.e. observations that contains no NA values in any of the channels.

Author(s)

Henrik Bengtsson

References


See Also

backtransformPrincipalCurve(). principal_curve.

Examples

# Simulate data from the model y <- a + bx + x^c + eps(bx)
J <- 1000
x <- rexp(J)
a <- c(2,15,3)
b <- c(2,3,4)
c <- c(1,2,1/2)
bx <- outer(b,x)
xc <- t(sapply(c, FUN=function(c) x^c))
eps <- apply(bx, MARGIN=2, FUN=function(x) rnorm(length(b), mean=0, sd=0.1*x))
y <- a + bx + xc + eps
y <- t(y)

# Fit principal curve through (y_1, y_2, y_3)
fit <- fitPrincipalCurve(y, verbose=TRUE)

# Flip direction of 'lambda'?
 rho <- cor(fit$lambda, y[,1], use="complete.obs")
flip <- (rho < 0)
if (flip) {
  fit$lambda <- max(fit$lambda, na.rm=TRUE) - fit$lambda
}
# Backtransform \((y_1, y_2, y_3)\) to be proportional to each other
\[ yN <- \text{backtransformPrincipalCurve}(y, \text{fit}=\text{fit}) \]

# Same backtransformation dimension by dimension
\[ yN2 <- y \]
for \((cc \ in 1:ncol(y))\) {
  \[ yN2[,cc] <- \text{backtransformPrincipalCurve}(y, \text{fit}=\text{fit}, \text{dimensions}=cc) \]
}  
\text{stopifnot(}\text{identical}(yN2, yN))

\[ \text{xlim} \leftarrow c(0, 1.04*\text{max}(x)) \]
\[ \text{ylim} \leftarrow \text{range}(c(y,N), \text{na.rm}=\text{TRUE}) \]

# Pairwise signals vs x before and after transform
\text{layout(matrix}(1:4, \text{nrow}=2, \text{byrow}=\text{TRUE}))
\text{par(mar=c(4,4,3,2)+0.1)}
for \((cc \ in 1:3)\) {
  \[ \text{ylab} \leftarrow \text{substitute}(y[c], \text{env=list}(c=cc)) \]
  \text{plot(NA, xlim=\text{xlim}, ylim=\text{ylim}, xlab="x", ylab=\text{ylab})}
  \text{abline(h=a[cc], \text{ly}=3)}
  \text{mtext(side=4, at=a[cc], \text{printf}("a=%g", a[cc]),}
  \text{ce=0.8, las=2, \text{line}=0, \text{adj=1.1, padj=0.2}}
  \text{points(x, y[,cc])}
  \text{points(x, yN[,cc], \text{col="tomato"}})
  \text{legend("topleft", \text{col=c("black", "tomato")}, \text{pch=19,}
  \text{c("original", "transformed"), bty="n")}
}  
\text{title(main="Pairwise signals vs x before and after transform", \text{outer}=\text{TRUE}, \text{line}=-2)}

# Pairwise signals before and after transform
\text{layout(matrix}(1:4, \text{nrow}=2, \text{byrow}=\text{TRUE}))
\text{par(mar=c(4,4,3,2)+0.1)}
for \((rr \ in 3:2)\) {
  \[ \text{ylab} \leftarrow \text{substitute}(y[c], \text{env=list}(c=rr)) \]
  for \((cc \ in 1:2)\) {
    \text{if (cc == rr) \{ }
      \text{plot.new()}
      \text{next}
    \}  
    \[ \text{xlab} \leftarrow \text{substitute}(y[c], \text{env=list}(c=cc)) \]
    \text{plot(NA, xlim=\text{ylim}, ylim=\text{ylim}, xlab=\text{xlab}, ylab=\text{ylab})}
    \text{abline(a=0, b=1, \text{ly}=2)}
    \text{points(y[,c(cc,rr)])}
    \text{points(yN[,c(cc,rr)], \text{col="tomato")}
    \text{legend("topleft", \text{col=c("black", "tomato")}, \text{pch=19,}
    \text{c("original", "transformed"), bty="n")}
  }  
  \text{title(main="Pairwise signals before and after transform", \text{outer}=\text{TRUE}, \text{line}=-2)}

---

\textit{fitXYCurve}  \hspace{1cm}  Fitting a smooth curve through paired \((x,y)\) data
Description

Fitting a smooth curve through paired (x,y) data.

Usage

```r
## S3 method for class 'matrix'
fitXYCurve(X, weights=NULL, typeOfWeights=c("datapoint"), method=c("loess", "lowess", "spline", "robustSpline"), bandwidth=NULL, satSignal=2^16 - 1, ...)
```

Arguments

- `X`: An Nx2 `matrix` where the columns represent the two channels to be normalized.
- `weights`: If `NULL`, non-weighted normalization is done. If data-point weights are used, this should be a `vector` of length N of data point weights used when estimating the normalization function.
- `typeOfWeights`: A `character` string specifying the type of weights given in argument `weights`.
- `method`: `character` string specifying which method to use when fitting the intensity-dependent function. Supported methods: "loess" (better than lowess), "lowess" (classic; supports only zero-one weights), "spline" (more robust than lowess at lower and upper intensities; supports only zero-one weights), "robustSpline" (better than spline).
- `bandwidth`: A `double` value specifying the bandwidth of the estimator used.
- `satSignal`: Signals equal to or above this threshold will not be used in the fitting.
- `...`: Not used.

Value

A named `list` structure of class `XYCurve`.

Missing values

The estimation of the function will only be made based on complete non-saturated observations, i.e. observations that contains no `NA` values nor saturated values as defined by `satSignal`.

Weighted normalization

Each data point, that is, each row in X, which is a vector of length 2, can be assigned a weight in [0,1] specifying how much it should affect the fitting of the normalization function. Weights are given by argument `weights`, which should be a `numeric vector` of length N.

Note that the lowess and the spline method only support zero-one {0,1} weights. For such methods, all weights that are less than a half are set to zero.

Details on loess

For `loess`, the arguments `family="symmetric",degree=1,span=3/4,control=loess.control(trace.hat="approx",iterations=5,surface="direct")` are used.

Author(s)

Henrik Bengtsson
Examples

```r
# Simulate data from the model y <- a + bx + x^c + eps(bx)
x <- rexp(1000)
a <- c(2,15)
b <- c(2,1)
c <- c(1.2)
bx <- outer(b,x)
xc <- t(sapply(c, FUN=function(c) x^c))
eps <- apply(bx, MARGIN=2, FUN=function(x) rnorm(length(x), mean=0, sd=0.1*x))
Y <- a + bx + xc + eps
Y <- t(Y)

lim <- c(0,70)
plot(Y, xlim=lim, ylim=lim)

# Fit principal curve through a subset of (y_1, y_2)
subset <- sample(nrow(Y), size=0.3*nrow(Y))
fit <- fitXYCurve(Y[subset,], bandwidth=0.2)
lines(fit, col="red", lwd=2)

# Backtransform (y_1, y_2) keeping y_1 unchanged
YN <- backtransformXYCurve(Y, fit=fit)
points(YN, col="blue")
abline(a=0, b=1, col="red", lwd=2)
```

---

description

Fits an R-dimensional hyperplane using iterative re-weighted PCA.

Usage

```r
## S3 method for class 'Var'
matrix
iwpca(X, w=NULL, R=1, method=c("symmetric", "bisquare", "tricube", "L1"), maxIter=30,
acc=1e-04, reps=0.02, fit0=NULL, ...)
```

Arguments

- **X**
  - N-times-K `matrix` where N is the number of observations and K is the number of dimensions.
- **w**
  - An N `vector` of weights for each row (observation) in the data matrix. If `NULL`, all observations get the same weight.
- **R**
  - Number of principal components to fit. By default a line is fitted.
- **method**
  - If "symmetric" (or "bisquare"), Tukey's biweight is used. If "tricube", the tricube weight is used. If "L1", the model is fitted in L1. If a `function`, it is used to calculate weights for next iteration based on the current iteration's residuals.
- **maxIter**
  - Maximum number of iterations.
acc  The (Euclidean) distance between two subsequent parameters fit for which the algorithm is considered to have converged.

reps  Small value to be added to the residuals before the the weights are calculated based on their inverse. This is to avoid infinite weights.

fit0  A list containing elements vt and pc specifying an initial fit. If NULL, the initial guess will be equal to the (weighted) PCA fit.

...  Additional arguments accepted by wpca().

Details

This method uses weighted principal component analysis (WPCA) to fit a R-dimensional hyper-plane through the data with initial internal weights all equal. At each iteration the internal weights are recalculated based on the "residuals". If method="L1", the internal weights are 1 / sum(abs(r) + reps). This is the same as method=function(r) 1/sum(abs(r)+reps). The "residuals" are orthogonal Euclidean distance of the principal components R,R+1,...,K. In each iteration before doing WPCA, the internal weights are multiplied by the weights given by argument w, if specified.

Value

Returns the fit (a list) from the last call to wpca() with the additional elements nbr0fIterations and converged.

Author(s)

Henrik Bengtsson

See Also

Internally wpca() is used for calculating the weighted PCA.

Examples

for (zzz in 0) {

# This example requires plot3d() in R.basic [http://www.braju.com/R/]
if (!require(pkgName <- "R.basic", character.only=TRUE)) break

# Simulate data from the model y <- a + bx + eps(bx)
  x <- rexp(1000)
  a <- c(2,15,3)
  b <- c(2,3,4)
  bx <- outer(b,x)
  eps <- apply(bx, MARGIN=2, FUN=function(x) rnorm(length(x), mean=0, sd=0.1*x))
  y <- a + bx + eps
  y <- t(y)

# Add some outliers by permuting the dimensions for 1/10 of the observations
  idx <- sample(1:nrow(y), size=1/10*nrow(y))
  y[idx,] <- y[idx,c(2,3,1)]

# Plot the data with fitted lines at four different view points
  opar <- par(mar=c(1,1,1,1)+0.1)
  N <- 4
  layout(matrix(1:N, nrow=2, byrow=TRUE))
  theta <- seq(0,270,length.out=N)

Median polish

\begin{verbatim}
phi <- rep(20, length.out=N)
xlim <- ylim <- zlim <- c(0,45);
persp <- list();
for (kk in seq_along(theta)) {
  # Plot the data
  persp[[kk]] <- plot3d(y, theta=theta[kk], phi=phi[kk], xlim=xlim, ylim=ylim, zlim=zlim)
}

# Weights on the observations
# Example a: Equal weights
w <- NULL
# Example b: More weight on the outliers (uncomment to test)
w <- rep(1, length(x)); w[idx] <- 0.8

# ...and show all iterations too with different colors.
maxIter <- c(seq(1,20,length.out=10),Inf)
col <- topo.colors(length(maxIter))
# Show the fitted value for every iteration
for (ii in seq_along(maxIter)) {
  # Fit a line using IWPCA through data
  fit <- iwpca(y, w=w, maxIter=maxIter[ii], swapDirections=TRUE)
  ymid <- fit$xMean
d0 <- apply(y, MARGIN=2, FUN=min) - ymid
d1 <- apply(y, MARGIN=2, FUN=max) - ymid
b <- fit$vt[1,]
y0 <- -b * max(abs(d0))
y1 <- b * max(abs(d1))
yline <- matrix(c(y0,y1), nrow=length(b), ncol=2)
yline <- yline + ymid

  for (kk in seq_along(theta)) {
    # Set pane to draw in
    par(mfg=c((kk-1) %% 2, (kk-1) %/% 2) + 1);
    # Set the viewpoint of the pane
    options(persp.matrix=persp[[kk]]);

    # Get the first principal component
    points3d(t(ymid), col=col[ii])
    lines3d(t(yline), col=col[ii])

    # Highlight the last one
    if (ii == length(maxIter))
      lines3d(t(yline), col="red", lwd=3)
  }
}
par(opar)
}
\end{verbatim}
Description
Median polish.

Usage
```r
## S3 method for class 'matrix'
medianPolish(X, tol=0.01, maxIter=10L, na.rm=NA, ..., .addExtra=TRUE)
```

Arguments
- `X` N-times-K matrix
- `tol` A numeric value greater than zero used as a threshold to identify when the algorithm has converged.
- `maxIter` Maximum number of iterations.
- `na.rm` If TRUE (FALSE), NAs are exclude (not exclude). If NA, it is assumed that X contains no NA values.
- `.addExtra` If TRUE, the name of argument X is returned and the returned structure is assigned a class. This will make the result compatible what medpolish returns.
- `...` Not used.

Details
The implementation of this method give identical estimates as medpolish, but is about 3-5 times more efficient when there are no NA values.

Value
Returns a named list structure with elements:
- `overall` The fitted constant term.
- `row` The fitted row effect.
- `col` The fitted column effect.
- `residuals` The residuals.
- `converged` If TRUE, the algorithm converged, otherwise not.

Author(s)
Henrik Bengtsson

See Also
medpolish.

Examples
```r
# Deaths from sport parachuting; from ABC of EDA, p.224:
deaths <- matrix(c(14,15,14, 7,4,7, 8,2,10, 15,9,10, 0,2,0), ncol=3, byrow=TRUE)
rownames(deaths) <- c("1-24", "25-74", "75-199", "200++", "NA")
colnames(deaths) <- 1973:1975
print(deaths)
```
### Description

Weighted affine normalization between channels and arrays.

This method will remove curvature in the M vs A plots that are due to an affine transformation of the data. In other words, if there are (small or large) biases in the different (red or green) channels, biases that can be equal too, you will get curvature in the M vs A plots and this type of curvature will be removed by this normalization method.

Moreover, if you normalize all slides at once, this method will also bring the signals on the same scale such that the log-ratios for different slides are comparable. Thus, do not normalize the scale of the log-ratios between slides afterward.

It is recommended to normalize as many slides as possible in one run. The result is that if creating log-ratios between any two channels and any slides, they will contain as little curvature as possible.

Furthermore, since the relative scale between any two channels on any two slides will be one if one normalizes all slides (and channels) at once it is possible to add or multiply with the same constant to all channels/arrays without introducing curvature. Thus, it is easy to rescale the data afterwards as demonstrated in the example.

### Usage

```r
## S3 method for class 'matrix'
normalizeAffine(X, weights=NULL, typeOfWeights=c("datapoint"), method="L1", constraint=0.05, satSignal=2^16 - 1, ..., .fitOnly=FALSE)
```

### Arguments

- **X**: An NxK matrix (K>=2) where the columns represent the channels, to be normalized.
- **weights**: If `NULL`, non-weighted normalization is done. If data-point weights are used, this should be a vector of length N of data point weights used when estimating the normalization function.
- **typeOfWeights**: A character string specifying the type of weights given in argument `weights`.
- **method**: A character string specifying how the estimates are robustified. See `iwpca()` for all accepted values.
- **constraint**: Constraint making the bias parameters identifiable. See `fitIWPCA()` for more details.

**normalizeAffine**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>satSignal</td>
<td>Signals equal to or above this threshold will not be used in the fitting.</td>
</tr>
<tr>
<td>...</td>
<td>Other arguments passed to <code>fitIWPCA()</code> and in turn <code>iwpca()</code>. For example, the weight argument of <code>iwpca()</code>. See also below.</td>
</tr>
<tr>
<td>.fitOnly</td>
<td>If <strong>TRUE</strong>, the data will not be back-transform.</td>
</tr>
</tbody>
</table>

**Details**

A line is fitted robustly through the \((y_R, y_G)\) observations using an iterated re-weighted principal component analysis (IWPCA), which minimized the residuals that are orthogonal to the fitted line. Each observation is down-weighted by the inverse of the absolute residuals, i.e. the fit is done in \(L_1\).

**Value**

A NxK matrix of the normalized channels. The fitted model is returned as attribute `modelFit`.

**Negative, non-positive, and saturated values**

Affine normalization applies equally well to negative values. Thus, contrary to normalization methods applied to log-ratios, such as curve-fit normalization methods, affine normalization, will not set these to NA.

Data points that are saturated in one or more channels are not used to estimate the normalization function, but they are normalized.

**Missing values**

The estimation of the affine normalization function will only be made based on complete non-saturated observations, i.e. observations that contains no `NA` values nor saturated values as defined by `satSignal`.

**Weighted normalization**

Each data point/observation, that is, each row in \(X\), which is a vector of length \(K\), can be assigned a weight in \([0,1]\) specifying how much it should affect the fitting of the affine normalization function. Weights are given by argument `weights`, which should be a numeric vector of length \(N\). Regardless of weights, all data points are normalized based on the fitted normalization function.

**Robustness**

By default, the model fit of affine normalization is done in \(L_1\) (method="L1"). This way, outliers affect the parameter estimates less than ordinary least-square methods.

For further robustness, downweight outliers such as saturated signals, if possible.

We do not use Tukey’s biweight function for reasons similar to those outlined in `calibrateMultiscan()`.

**Using known/previously estimated channel offsets**

If the channel offsets can be assumed to be known, then it is possible to fit the affine model with no (zero) offset, which formally is a linear (proportional) model, by specifying argument `center=FALSE`. In order to do this, the channel offsets have to be subtracted from the signals manually before normalizing, e.g. \(X_a < t(t(X)-a)\) where \(e\) is vector of length \(ncol(X)\). Then normalize by \(X_n < normalizeAffine(X_a,center=FALSE)\). You can assert that the model is fitted without offset by `stopifnot(all(attr(Xn,"modelFit")$adiag == 0))`. 

Author(s)
Henrik Bengtsson

References

See Also
calibrateMultiscan().

Examples

```
pathname <- system.file("data-ex", "PMT-RGData.dat", package="aroma.light")
rg <- read.table(pathname, header=TRUE, sep="\t")
nbr0fScans <- max(rg$slide)

gr <- as.list(rg)
for (field in c("R", "G"))
gr[[field]] <- matrix(as.double(gr[[field]]), ncol=nbr0fScans)
gr$slide <- gr$spot <- NULL
rg <- as.matrix(as.data.frame(rg))
colnames(rg) <- rep(c("R", "G"), each=nbr0fScans)

layout(matrix(c(1,2,0,3,4,0,5,6,7), ncol=3, byrow=TRUE))

rgC <- rg
for (channel in c("R", "G")) {
sidx <- which(colnames(rg) == channel)
channelColor <- switch(channel, R="red", G="green")

# The raw data
plotMvsAPairs(rg[,sidx])
title(main=paste("Observed", channel))
box(col=channelColor)

# The calibrated data
rgC[,sidx] <- calibrateMultiscan(rg[,sidx], average=NULL)
plotMvsAPairs(rgC[,sidx])
title(main=paste("Calibrated", channel))
box(col=channelColor)
}
```

# The average calibrated data

#
# Note how the red signals are weaker than the green. The reason
# for this can be that the scale factor in the green channel is
# greater than in the red channel, but it can also be that there
# is a remaining relative difference in bias between the green
# and the red channel, a bias that precedes the scanning.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
rgCA <- rg
for (channel in c("R", "G")) {
  sidx <- which(colnames(rg) == channel)
  rgCA[,sidx] <- calibrateMultiscan(rg[,sidx])
}
rgCAavg <- matrix(NA_real_, nrow=nrow(rgCA), ncol=2)
colnames(rgCAavg) <- c("R", "G")
for (channel in c("R", "G")) {
  sidx <- which(colnames(rg) == channel)
  rgCAavg[,channel] <- apply(rgCA[,sidx], MARGIN=1, FUN=median, na.rm=TRUE)
}
# Add some "fake" outliers
outliers <- 1:600
rgCAavg[outliers,"G"] <- 50000
plotMvsA(rgCAavg)
title(main="Average calibrated (AC)")

# Normalize data
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
weights <- rep(1, nrow(rgCAavg))
weights[outliers] <- 0.001
rgCANa <- normalizeAffine(rgCAavg, weights=weights)
# It is always ok to rescale the affine normalized data if its
done on (R,G); not on (A,M)! However, this is only needed for
# aesthetic purposes.
rgCANa <- rgCANa * 2^1.4
plotMvsA(rgCANa)
title(main="Normalized AC")

# Curve-fit (lowess) normalization
rgCANlw <- normalizeLowess(rgCAavg, weights=weights)
plotMvsA(rgCANlw, col="orange", add=TRUE)

# Curve-fit (loess) normalization
rgCANl <- normalizeLoess(rgCAavg, weights=weights)
plotMvsA(rgCANl, col="red", add=TRUE)

# Curve-fit (robust spline) normalization
rgCANrs <- normalizeRobustSpline(rgCAavg, weights=weights)
plotMvsA(rgCANrs, col="blue", add=TRUE)

legend(x=0.8, y=16, legend=c("affine", "lowess", "loess", "r. spline"), pch=19,
col=c("black", "orange", "red", "blue"), ncol=2, x.intersp=0.3, bty="n")
normalizeAverage

Rescales channel vectors to get the same average

Description

Rescales channel vectors to get the same average.

Usage

```r
## S3 method for class 'matrix'
normalizeAverage(x, baseline=1, avg=stats::median, targetAvg=2200, ...)
## S3 method for class 'list'
normalizeAverage(x, baseline=1, avg=stats::median, targetAvg=2200, ...)
```

Arguments

- `x`: A numeric NxK matrix (or list of length K).
- `baseline`: An integer in [1,K] specifying which channel should be the baseline.
- `avg`: A function for calculating the average of one channel.
- `targetAvg`: The average that each channel should have afterwards. If NULL, the baseline column sets the target average.
- `...`: Additional arguments passed to the avg function.

Value

Returns a normalized numeric NxK matrix (or list of length K).

Author(s)

Henrik Bengtsson
normalizeCurveFit  
Weighted curve-fit normalization between a pair of channels

Description

Weighted curve-fit normalization between a pair of channels.

This method will estimate a smooth function of the dependency between the log-ratios and the log-intensity of the two channels and then correct the log-ratios (only) in order to remove the dependency. This method is also known as intensity-dependent or lowess normalization.

The curve-fit methods are by nature limited to paired-channel data. There exist at least one method trying to overcome this limitation, namely the cyclic-lowess [1], which applies the paired curve-fit method iteratively over all pairs of channels/arrays. Cyclic-lowess is not implemented here.

We recommend that affine normalization [2] is used instead of curve-fit normalization.

Usage

```r
## S3 method for class 'matrix'
normalizeCurveFit(X, weights=NULL, typeOfWeights=c("datapoint"),
    method=c("loess", "lowess", "spline", "robustSpline"), bandwidth=NULL,
    satSignal=2^16 - 1, ...)
## S3 method for class 'matrix'
normalizeLoess(X, ...)
## S3 method for class 'matrix'
normalizeLowess(X, ...)
## S3 method for class 'matrix'
normalizeSpline(X, ...)
## S3 method for class 'matrix'
normalizeRobustSpline(X, ...)
```

Arguments

- **X**: An Nx2 matrix where the columns represent the two channels to be normalized.
- **weights**: If NULL, non-weighted normalization is done. If data-point weights are used, this should be a vector of length N of data point weights used when estimating the normalization function.
- **typeOfWeights**: A character string specifying the type of weights given in argument weights.
- **method**: character string specifying which method to use when fitting the intensity-dependent function. Supported methods: "loess" (better than lowess), "lowess" (classic; supports only zero-one weights), "spline" (more robust than lowess at lower and upper intensities; supports only zero-one weights), "robustSpline" (better than spline).
- **bandwidth**: A double value specifying the bandwidth of the estimator used.
- **satSignal**: Signals equal to or above this threshold will not be used in the fitting.
- **...**: Not used.
Details

A smooth function $c(A)$ is fitted through data in $(A, M)$, where $M = \log_2(y_2/y_1)$ and $A = 1/2 \times \log_2(y_2 \times y_1)$. Data is normalized by $M < -M - c(A)$.

Loess is by far the slowest method of the four, then lowess, and then robust spline, which iteratively calls the spline method.

Value

A Nx2 matrix of the normalized two channels. The fitted model is returned as attribute `modelFit`.

Negative, non-positive, and saturated values

Non-positive values are set to not-a-number (NaN). Data points that are saturated in one or more channels are not used to estimate the normalization function, but they are normalized.

Missing values

The estimation of the normalization function will only be made based on complete non-saturated observations, i.e. observations that contains no NA values nor saturated values as defined by `satSignal`.

Weighted normalization

Each data point, that is, each row in $X$, which is a vector of length 2, can be assigned a weight in $[0,1]$ specifying how much it should affect the fitting of the normalization function. Weights are given by argument `weights`, which should be a numeric vector of length N. Regardless of weights, all data points are normalized based on the fitted normalization function.

Note that the lowess and the spline method only support zero-one \{0,1\} weights. For such methods, all weights that are less than a half are set to zero.

Details on loess

For `loess`, the arguments family="symmetric", degree=1, span=3/4, control=loess.control(trace.hat="approximate", iterations=5, surface="direct") are used.

Author(s)

Henrik Bengtsson

References


See Also

normalizeAffine().
Examples

```r
pathname <- system.file("data-ex", "PMT-RGData.dat", package="aroma.light")
rg <- read.table(pathname, header=TRUE, sep="\t")
nbrOfScans <- max(rg$slide)

erg <- as.list(rg)
for (field in c("R", "G"))
  rg[[field]] <- matrix(as.double(rg[[field]]), ncol=nbrOfScans)
rg$slide <- rg$spot <- NULL
rg <- as.matrix(as.data.frame(rg))

colnames(rg) <- rep(c("R", "G"), each=nbrOfScans)

layout(matrix(c(1,2,0,3,4,0,5,6,7), ncol=3, byrow=TRUE))

rgC <- rg
for (channel in c("R", "G")) {
  sidx <- which(colnames(rg) == channel)
  channelColor <- switch(channel, R="red", G="green")

  # The raw data
  # - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  plotMvsAPairs(rg[,sidx])
  title(main=paste("Observed", channel))
  box(col=channelColor)

  # The calibrated data
  # - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  rgC[,sidx] <- calibrateMultiscan(rg[,sidx], average=NULL)
  plotMvsAPairs(rgC[,sidx])
  title(main=paste("Calibrated", channel))
  box(col=channelColor)
} # for (channel ...)

# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# The average calibrated data
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
rgCA <- rg
for (channel in c("R", "G")) {
  sidx <- which(colnames(rg) == channel)
  rgCA[,sidx] <- calibrateMultiscan(rg[,sidx])
}

rgCAavg <- matrix(NA_real_, nrow=nrow(rgCA), ncol=2)
colnames(rgCAavg) <- c("R", "G")
for (channel in c("R", "G")) {
  sidx <- which(colnames(rg) == channel)
  rgCAavg[,sidx] <- calibrateMultiscan(rg[,sidx])
}
```

normalizeDifferencesToAverage

Rescales channel vectors to get the same average

rgCAavg[,channel] <- apply(rgCA[,sidx], MARGIN=1, FUN=median, na.rm=TRUE)

# Add some "fake" outliers
outliers <- 1:600
rgCAavg[outliers,"G"] <- 50000

plotMvsA(rgCAavg)
title(main="Average calibrated (AC)")

# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# Normalize data
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# Weight-down outliers when normalizing
weights <- rep(1, nrow(rgCAavg))
weights[outliers] <- 0.001

# Affine normalization of channels
rgCANa <- normalizeAffine(rgCAavg, weights=weights)
# It is always ok to rescale the affine normalized data if its
# done on (R,G); not on (A,M)! However, this is only needed for
# esthetic purposes.
rgCANa <- rgCANa *2^1.4
plotMvsA(rgCANa)
title(main="Normalized AC")

# Curve-fit (lowess) normalization
rgCANlw <- normalizeLowess(rgCAavg, weights=weights)
plotMvsA(rgCANlw, col="orange", add=TRUE)

# Curve-fit (loess) normalization
rgCANl <- normalizeLoess(rgCAavg, weights=weights)
plotMvsA(rgCANl, col="red", add=TRUE)

# Curve-fit (robust spline) normalization
rgCANrs <- normalizeRobustSpline(rgCAavg, weights=weights)
plotMvsA(rgCANrs, col="blue", add=TRUE)

legend(x=0,y=16, legend=c("affine", "lowess", "loess", "r. spline"), pch=19,
       col=c("black", "orange", "red", "blue"), ncol=2, x.intersp=0.3, bty="n")

plotMvsMPairs(cbind(rgCANa, rgCANlw), col="orange", xlab=expression(M[affine]))
title(main="Normalized AC")
plotMvsMPairs(cbind(rgCANa, rgCANl), col="red", add=TRUE)
plotMvsMPairs(cbind(rgCANa, rgCANrs), col="blue", add=TRUE)
abline(a=0, b=1, lty=2)
legend(x=-6,y=6, legend=c("lowess", "loess", "r. spline"), pch=19,
       col=c("orange", "red", "blue"), ncol=2, x.intersp=0.3, bty="n")
normalizeDifferencesToAverage

Description

Rescales channel vectors to get the same average.

Usage

## S3 method for class 'list'
normalizeDifferencesToAverage(x, baseline=1, FUN=median, ...)

Arguments

x 
A numeric list of length K.

baseline 
An integer in [1,K] specifying which channel should be the baseline. The baseline channel will be almost unchanged. If NULL, the channels will be shifted towards median of them all.

FUN 
A function for calculating the average of one channel.

... 
Additional arguments passed to the avg function.

Value

Returns a normalized list of length K.

Author(s)

Henrik Bengtsson

Examples

# Simulate three shifted tracks of different lengths with same profiles
ns <- c(A=2, B=1, C=0.25)*1000
xx <- lapply(ns, FUN=function(n) { seq(from=1, to=max(ns), length.out=n) })
zz <- mapply(seq_along(ns), ns, FUN=function(z,n) rep(z,n))

yy <- list(
  A = rnorm(ns["A"], mean=0, sd=0.5),
  B = rnorm(ns["B"], mean=5, sd=0.4),
  C = rnorm(ns["C"], mean=-5, sd=1.1)
)

yy <- lapply(yy, FUN=function(y) {
  n <- length(y)
  y[1:(n/2)] <- y[1:(n/2)] + 2
  y[1:(n/4)] <- y[1:(n/4)] - 4
  y
})

# Shift all tracks toward the first track
yyN <- normalizeDifferencesToAverage(yy, baseline=1)

# The baseline channel is not changed
stopifnot(identical(yy[[1]], yyN[[1]]))

# Get the estimated parameters
fit <- attr(yyN, "fit")

# Plot the tracks
normalizeFragmentLength

Normalizes signals for PCR fragment-length effects

Description

Normalizes signals for PCR fragment-length effects. Some or all signals are used to estimated the normalization function. All signals are normalized.

Usage

```r
## Default S3 method:
normalizeFragmentLength(y, fragmentLengths, targetFcns=NULL, subsetToFit=NULL,
onMissing=c("ignore", "median"), .isLogged=TRUE, ..., .returnFit=FALSE)
```

Arguments

- `y` A numeric vector of length `K` of signals to be normalized across `E` enzymes.
- `fragmentLengths` An integer `KxE` matrix of fragment lengths.
- `targetFcns` An optional list of `E` functions; one per enzyme. If `NULL`, the data is normalized to have constant fragment-length effects (all equal to zero on the log-scale).
- `subsetToFit` The subset of data points used to fit the normalization function. If `NULL`, all data points are considered.
- `onMissing` Specifies how data points for which there is no fragment length is normalized. If "ignore", the values are not modified. If "median", the values are updated to have the same robust average as the other data points.
- `.isLogged` A logical.
- `...` Additional arguments passed to `lowess`.
- `returnFit` A logical.

Value

Returns a numeric vector of the normalized signals.

Multi-enzyme normalization

It is assumed that the fragment-length effects from multiple enzymes added (with equal weights) on the intensity scale. The fragment-length effects are fitted for each enzyme separately based on units that are exclusively for that enzyme. If there are no or very such units for an enzyme, the assumptions of the model are not met and the fit will fail with an error. Then, from the above single-enzyme fits the average effect across enzymes is the calculated for each unit that is on multiple enzymes.
Target functions

It is possible to specify custom target function effects for each enzyme via argument `targetFcns`. This argument has to be a list containing one function per enzyme and ordered in the same order as the enzyme are in the columns of argument `fragmentLengths`. For instance, if one wish to normalize the signals such that their mean signal as a function of fragment length effect is constantly equal to 2200 (or the intensity scale), the use `targetFcns=function(fl,...) log2(2200)` which completely ignores fragment-length argument `fl` and always returns a constant. If two enzymes are used, then use `targetFcns=rep(list(function(fl,...) log2(2200)),2).` Note, if `targetFcns` is `NULL`, this corresponds to `targetFcns=rep(list(function(fl,...) 0),ncol(fragmentLengths)).` Alternatively, if one wants to only apply minimal corrections to the signals, then one can normalize toward target functions that correspond to the fragment-length effect of the average array.

Author(s)

Henrik Bengtsson

References


Examples

```r
# Example 1: Single-enzyme fragment-length normalization of 6 arrays
# Number samples
I <- 9
# Number of loci
J <- 1000
# Fragment lengths
fl <- seq(from=100, to=1000, length.out=J)
# Simulate data points with unknown fragment lengths
hasUnknownFL <- seq(from=1, to=J, by=50)
fl[hasUnknownFL] <- NA
# Simulate data
y <- matrix(0, nrow=J, ncol=I)
maxY <- 12
for (kk in 1:I) {
  k <- runif(n=1, min=3, max=5)
  mu <- function(fl) {
    mu <- rep(maxY, length(fl))
    ok <- !is.na(fl)
    mu[ok] <- mu[ok] - fl[ok]^(1/k)
    mu
  }
  eps <- rnorm(J, mean=0, sd=1)
  y[,kk] <- mu(fl) + eps
}
```
# Normalize data (to a zero baseline)
yN <- apply(y, MARGIN=2, FUN=function(y) {
    normalizeFragmentLength(y, fragmentLengths=fl, onMissing="median")
})

# The correction factors
rho <- y-yN
print(summary(rho))
# The correction for units with unknown fragment lengths
# equals the median correction factor of all other units
print(summary(rho[hasUnknownFL,]))

# Plot raw data
layout(matrix(1:9, ncol=3))
xlim <- c(0,max(fl, na.rm=TRUE))
ylim <- c(0,max(y, na.rm=TRUE))
xlab <- "Fragment length"
ylab <- expression(log2(theta))
for (kk in 1:I) {
    plot(fl, y[,kk], xlim=xlim, ylim=ylim, xlab=xlab, ylab=ylab)
    ok <- (is.finite(fl) & is.finite(y[,kk]))
    lines(lowess(fl[ok], y[ok,kk]), col="red", lwd=2)
}

# Plot normalized data
layout(matrix(1:9, ncol=3))
ylim <- c(-1,1)*max(y, na.rm=TRUE)/2
for (kk in 1:I) {
    plot(fl, yN[,kk], xlim=xlim, ylim=ylim, xlab=xlab, ylab=ylab)
    ok <- (is.finite(fl) & is.finite(y[,kk]))
    lines(lowess(fl[ok], yN[ok,kk]), col="blue", lwd=2)
}

# Example 2: Two-enzyme fragment-length normalization of 6 arrays
set.seed(0xbeef)
# Number samples
I <- 5

# Number of loci
J <- 3000

# Fragment lengths (two enzymes)
fl <- matrix(0, nrow=J, ncol=2)
fl[,1] <- seq(from=100, to=1000, length.out=J)
fl[,2] <- seq(from=1000, to=100, length.out=J)

# Let 1/2 of the units be on both enzymes
fl[seq(from=1, to=J, by=4),1] <- NA
fl[seq(from=2, to=J, by=4),2] <- NA

# Let some have unknown fragment lengths
hasUnknownFL <- seq(from=1, to=J, by=15)
normalizeFragmentLength

fl[hasUnknownFL,] <- NA

# Sty/Nsp mixing proportions:
rho <- rep(1, I)
rho[1] <- 1/3; # Less Sty in 1st sample
rho[3] <- 3/2; # More Sty in 3rd sample

# Simulate data
z <- array(0, dim=c(J,2,I))
maxLog2Theta <- 12
for (ii in 1:I) {
  # Common effect for both enzymes
  mu <- function(fl) {
    k <- runif(n=1, min=3, max=5)
    mu <- rep(maxLog2Theta, length(fl))
    ok <- is.finite(fl)
    mu[ok] <- mu[ok] - fl[ok]^((1/k)
    mu
  }
  # Calculate the effect for each data point
  for (ee in 1:2) {
    z[,ee,ii] <- mu(fl[,ee])
  }
  # Update the Sty/Nsp mixing proportions
  ee <- 2
  z[ee,ii] <- rho[ee]*z[,ee,ii]
  # Add random errors
  for (ee in 1:2) {
    eps <- rnorm(J, mean=0, sd=1/sqrt(2))
    z[,ee,ii] <- z[,ee,ii] + eps
  }
}

hasFl <- is.finite(fl)

unitSets <- list(
  nsp = which( hasFl[,1] & !hasFl[,2]),
  sty = which(!hasFl[,1] & hasFl[,2]),
  both = which( hasFl[,1] & hasFl[,2]),
  none = which(!hasFl[,1] & !hasFl[,2])
)

# The observed data is a mix of two enzymes
theta <- matrix(NA_real_, nrow=J, ncol=I)

# Single-enzyme units
for (ee in 1:2) {
  uu <- unitSets[[ee]]
  theta[uu,] <- 2^z[uu,ee,]
 }

# Both-enzyme units (sum on intensity scale)
uu <- unitSets$both
theta[uu,] <- (2^z[uu,1,]+2^z[uu,2,])/2

# Missing units (sample from the others)
uu <- unitSets$none
theta[uu,] <- apply(theta, MARGIN=2, sample, size=length(uu))

# Calculate target array
thetaT <- rowMeans(theta, na.rm=TRUE)
targetFcns <- list()
for (ee in 1:2) {
  uu <- unitSets[[ee]]
  fit <- lowess(fl[uu,ee], log2(thetaT[uu]))
  class(fit) <- "lowess"
  targetFcns[[ee]] <- function(fl, ...) {
    predict(fit, newdata=fl)
  }
}

# Fit model only to a subset of the data
subsetToFit <- setdiff(1:J, seq(from=1, to=J, by=10))

# Normalize data (to a target baseline)
thetaN <- matrix(NA_real_, nrow=J, ncol=I)
fits <- vector("list", I)
for (ii in 1:I) {
  lthetaNi <- normalizeFragmentLength(log2(theta[,ii]), targetFcns=targetFcns,
    fragmentLengths=fl, onMissing="median", subsetToFit=subsetToFit, .returnFit=TRUE)
  fits[[ii]] <- attr(lthetaNi, "modelFit")
  thetaN[,ii] <- 2^lthetaNi
}

# Plot raw data
xlim <- c(0, max(fl, na.rm=TRUE))
ylim <- c(0, max(log2(theta), na.rm=TRUE))
Mlim <- c(-1,1)*4
xlab <- "Fragment length"
ylab <- expression(log2(theta))
Mlab <- expression(M==log[2](theta/theta[R]))
layout(matrix(1:(3*I), ncol=I, byrow=TRUE))
for (ii in 1:I) {
  plot(NA, xlim=xlim, ylim=ylim, xlab=xlab, ylab=ylab, main="raw")
  # Single-enzyme units
  for (ee in 1:2) {
    # The raw data
    uu <- unitSets[[ee]]
    points(fl[uu,ee], log2(theta[uu,ii]), col=ee+1)
  }
  # Both-enzyme units (use fragment-length for enzyme #1)
  uu <- unitSets$both
  points(fl[uu,1], log2(theta[uu,ii]), col=3+1)
for (ee in 1:2) {
  # The true effects
  uu <- unitSets[[ee]]
  lines(lowess(fl[uu,ee], log2(theta[uu,ii])), col="black", lwd=4, lty=3)

  # The estimated effects
  fit <- fits[[ii]][[ee]]$fit
  lines(fit, col="orange", lwd=3)

  muT <- targetFcns[[ee]](fl[uu,ee])
  lines(fl[uu,ee], muT, col="cyan", lwd=1)
}
}

# Calculate log-ratios
thetaR <- rowMeans(thetaN, na.rm=TRUE)
M <- log2(thetaN/thetaR)

# Plot normalized data
for (ii in 1:I) {
  plot(NA, xlim=xlim, ylim=Mlim, xlab=xlab, ylab=Mlab, main="normalized")
  # Single-enzyme units
  for (ee in 1:2) {
    # The normalized data
    uu <- unitSets[[ee]]
    points(fl[uu,ee], M[uu,ii], col=ee+1)
  }
  # Both-enzyme units (use fragment-length for enzyme #1)
  uu <- unitSets$both
  points(fl[uu,1], M[uu,ii], col=3+1)
}

ylim <- c(0,1.5)
for (ii in 1:I) {
  data <- list()
  for (ee in 1:2) {
    # The normalized data
    uu <- unitSets[[ee]]
    data[[ee]] <- M[uu,ii]
  }
  uu <- unitSets$both
  if (length(uu) > 0) data[[3]] <- M[uu,ii]
  uu <- unitSets$none
  if (length(uu) > 0) data[[4]] <- M[uu,ii]

  cols <- seq_along(data)+1
  plotDensity(data, col=cols, xlim=Mlim, xlab=Mlab, main="normalized")
  abline(v=0, lty=2)
}
normalizeQuantileRank

Normalizes the empirical distribution of one or more samples to a target distribution

Description

Normalizes the empirical distribution of one of more samples to a target distribution.

The average sample distribution is calculated either robustly or not by utilizing either \texttt{weightedMedian()} or \texttt{weighted.mean()}. A weighted method is used if any of the weights are different from one.

Usage

\begin{verbatim}
## S3 method for class 'numeric'
normalizeQuantileRank(x, xTarget, ties=FALSE, ...)
## S3 method for class 'list'
normalizeQuantileRank(X, xTarget=NULL, ...)
## Default S3 method:
normalizeQuantile(x, ...)
\end{verbatim}

Arguments

\begin{itemize}
\item \texttt{x, X}  a numeric vector of length \textit{N} or a list of length \textit{N} with numeric vectors. If a list, then the vectors may be of different lengths.
\item \texttt{xTarget}  The target empirical distribution as a sorted numeric vector of length \textit{M}. If \texttt{NULL} and \texttt{X} is a list, then the target distribution is calculated as the average empirical distribution of the samples.
\item \texttt{ties}  Should ties in \texttt{x} be treated with care or not? For more details, see "limma:normalizeQuantiles".
\item \texttt{...}  Not used.
\end{itemize}

Value

Returns an object of the same shape as the input argument.

Missing values

Missing values are excluded when estimating the "common" (the baseline). Values that are \texttt{NA} remain \texttt{NA} after normalization. No new \texttt{NAs} are introduced.

Weights

Currently only channel weights are support due to the way quantile normalization is done. If signal weights are given, channel weights are calculated from these by taking the mean of the signal weights in each channel.

Author(s)

See Also

To calculate a target distribution from a set of samples, see `averageQuantile()`. For an alternative empirical density normalization methods, see `normalizeQuantileSpline()`.

Examples

```r
# Simulate ten samples of different lengths
N <- 10000
X <- list()
for (kk in 1:8) {
  rfcn <- list(rnorm, rgamma)[[sample(2, size=1)]]
  size <- runif(1, min=0.3, max=1)
  a <- rgamma(1, shape=20, rate=10)
  b <- rgamma(1, shape=10, rate=10)
  values <- rfcn(size*N, a, b)

  # "Censor" values
  values[values < 0 | values > 8] <- NA

  X[[kk]] <- values
}

# Add 20% missing values
X <- lapply(X, FUN=function(x) {
  x[sample(length(x), size=0.20*length(x))] <- NA
  x
})

# Normalize quantiles
Xn <- normalizeQuantile(X)

# Plot the data
layout(matrix(1:2, ncol=1))
xlim <- range(X, na.rm=TRUE)
plotDensity(X, lwd=2, xlim=xlim, main="The original distributions")
plotDensity(Xn, lwd=2, xlim=xlim, main="The normalized distributions")
```

normalizeQuantileRank.matrix

Normalizes the empirical distribution of a set of samples to a common target distribution.

Description

Normalizes the empirical distribution of a set of samples to a common target distribution. The average sample distribution is calculated either robustly or not by utilizing either `weightedMedian()` or `weighted.mean()`. A weighted method is used if any of the weights are different from one.

Usage

```r
## S3 method for class 'matrix'
normalizeQuantileRank(X, ties=FALSE, robust=FALSE, weights=NULL,
                      typeOfWeights=c("channel", "signal"), ...)
```
Arguments

- **X**: a numerical NxK matrix with the K columns representing the channels and the N rows representing the data points.
- **robust**: If TRUE, the (weighted) median function is used for calculating the average sample distribution, otherwise the (weighted) mean function is used.
- **ties**: Should ties in x be treated with care or not? For more details, see "limma:normalizeQuantiles".
- **weights**: If NULL, non-weighted normalization is done. If channel weights, this should be a vector of length K specifying the weights for each channel. If signal weights, it should be an NxK matrix specifying the weights for each signal.
- **typeOfWeights**: A character string specifying the type of weights given in argument weights.
- **...**: Not used.

Value

Returns an object of the same shape as the input argument.

Missing values

Missing values are excluded when estimating the "common" (the baseline). Values that are NA remain NA after normalization. No new NAs are introduced.

Weights

Currently only channel weights are support due to the way quantile normalization is done. If signal weights are given, channel weights are calculated from these by taking the mean of the signal weights in each channel.

Author(s)

Adopted from Gordon Smyth (http://www.statsci.org/) in 2002 \\& 2006. Original code by Ben Bolstad at Statistics Department, University of California. Support for calculating the average sample distribution using (weighted) mean or median was added by Henrik Bengtsson.

See Also

median, weightedMedian, mean() and weighted.mean. normalizeQuantileSpline().

Examples

```r
# Simulate three samples with on average 20% missing values
N <- 10000
X <- cbind(rnorm(N, mean=3, sd=1),
           rnorm(N, mean=4, sd=2),
           rgamma(N, shape=2, rate=1))
X[sample(3*N, size=0.20*3*N)] <- NA

# Normalize quantiles
Xn <- normalizeQuantile(X)

# Plot the data
layout(matrix(1:2, ncol=1))
xlim <- range(X, Xn, na.rm=TRUE)
plotDensity(X, lwd=2, xlim=xlim, main="The three original distributions")
plotDensity(Xn, lwd=2, xlim=xlim, main="The three normalized distributions")
```
normalizeQuantileSpline

Normalizes the empirical distribution of one or more samples to a target distribution

Description

Normalizes the empirical distribution of one or more samples to a target distribution. After normalization, all samples have the same average empirical density distribution.

Usage

## S3 method for class 'numeric'
normalizeQuantileSpline(x, w=NULL, xTarget, sortTarget=TRUE, robust=TRUE, ...)
## S3 method for class 'matrix'
normalizeQuantileSpline(X, w=NULL, xTarget=NULL, sortTarget=TRUE, robust=TRUE, ...)
## S3 method for class 'list'
normalizeQuantileSpline(X, w=NULL, xTarget=NULL, sortTarget=TRUE, robust=TRUE, ...)

Arguments

x, X  A single ($K = 1$) numeric vector of length $N$, a numeric $N \times K$ matrix, or a list of length $K$ with numeric vectors, where $K$ represents the number of samples and $N$ the number of data points.
w  An optional numeric vector of length $N$ of weights specific to each data point.
xTarget  The target empirical distribution as a sorted numeric vector of length $M$. If NULL and $X$ is a list, then the target distribution is calculated as the average empirical distribution of the samples.
sortTarget  If TRUE, argument xTarget will be sorted, otherwise it is assumed to be already sorted.
robust  If TRUE, the normalization function is estimated robustly.
...  Arguments passed to (smooth.spline or robustSmoothSpline).

Value

Returns an object of the same type and dimensions as the input.

Missing values

Both argument $X$ and xTarget may contain non-finite values. These values do not affect the estimation of the normalization function. Missing values and other non-finite values in $X$, remain in the output as is. No new missing values are introduced.

Author(s)

Henrik Bengtsson
**normalizeTumorBoost**

Normalizes allele B fractions for a tumor given a match normal

**Description**

TumorBoost [1] is a normalization method that normalizes the allele B fractions of a tumor sample given the allele B fractions and genotypes of a matched normal. The method is a single-sample (single-pair) method. It does not require total copy-number estimates. The normalization is done such that the total copy number is unchanged afterwards.

**Usage**

```r
## S3 method for class 'numeric'
normalizeTumorBoost(betaT, betaN, muN=callNaiveGenotypes(betaN), preserveScale=FALSE, flavor=c("v4", "v3", "v2", "v1"), ...)
```

**References**


**See Also**

The target distribution can be calculated as the average using `averageQuantile()`.

Internally either `robustSmoothSpline(robust=TRUE)` or `smooth.spline(robust=FALSE)` is used.

An alternative normalization method that is also normalizing the empirical densities of samples is `normalizeQuantileRank()`. Contrary to this method, that method requires that all samples are based on the exact same set of data points and it is also more likely to over-correct in the tails of the distributions.

**Examples**

```r
# Simulate three samples with on average 20% missing values
N <- 10000
X <- cbind(rnorm(N, mean=3, sd=1),
          rnorm(N, mean=4, sd=2),
          rgamma(N, shape=2, rate=1))
X[sample(3*N, size=0.20*3*N)] <- NA

# Plot the data
layout(matrix(c(1,0,2:5), ncol=2, byrow=TRUE))
xlim <- range(X, na.rm=TRUE)
plotDensity(X, lwd=2, xlim=xlim, main="The three original distributions")

Xn <- normalizeQuantile(X)
plotDensity(Xn, lwd=2, xlim=xlim, main="The three normalized distributions")
plotXYCurve(X, Xn, xlim=xlim, main="The three normalized distributions")

Xn2 <- normalizeQuantileSpline(X, xTarget=Xn[,1], spar=0.99)
plotDensity(Xn2, lwd=2, xlim=xlim, main="The three normalized distributions")
plotXYCurve(X, Xn2, xlim=xlim, main="The three normalized distributions")
```
**normalizeTumorBoost**

**Arguments**

- `betaT`, `betaN` Two numeric vectors each of length J with tumor and normal allele B fractions, respectively.
- `muN` An optional vector of length J containing normal genotypes calls in `(0, 1/2, 1, NA)` for `(AA, AB, BB)`.
- `preserveScale` If TRUE, SNPs that are heterozygous in the matched normal are corrected for signal compression using an estimate of signal compression based on the amount of correction performed by TumorBoost on SNPs that are homozygous in the matched normal.
- `flavor` A character string specifying the type of correction applied. Not used.

**Details**

Allele B fractions are defined as the ratio between the allele B signal and the sum of both (all) allele signals at the same locus. Allele B fractions are typically within `[0, 1]`, but may have a slightly wider support due to for instance negative noise. This is typically also the case for the returned normalized allele B fractions.

**Value**

Returns a numeric vector of length J containing the normalized allele B fractions for the tumor. Attribute `modelFit` is a list containing model fit parameters.

**Flavors**

This method provides a few different "flavors" for normalizing the data. The following values of argument `flavor` are accepted:

- v4: (default) The TumorBoost method, i.e. Eqns. (8)-(9) in [1].
- v3: Eqn (9) in [1] is applied to both heterozygous and homozygous SNPs, which effectively is v4 where the normalized allele B fractions for homozygous SNPs becomes 0 and 1.
- v2: ...
- v1: TumorBoost where correction factor is forced to one, i.e. \( \eta_j = 1 \). As explained in [1], this is a suboptimal normalization method. See also the discussion in the paragraph following Eqn (12) in [1].

**Preserving scale**

As of **aroma.light v1.33.3** (March 30, 2014), argument `preserveScale` no longer has a default value and has to be specified explicitly. This is done in order to change the default to **FALSE** in a future version, while minimizing the risk for surprises.

Allele B fractions are more or less compressed toward a half, e.g. the signals for homozygous SNPs are slightly away from zero and one. The TumorBoost method decreases the correlation in allele B fractions between the tumor and the normal **conditioned on the genotype**. What it does not control for is the mean level of the allele B fraction **conditioned on the genotype**.

By design, most flavors of the method will correct the homozygous SNPs such that their mean levels get close to the expected zero and one levels. However, the heterozygous SNPs will typically keep the same mean levels as before. One possibility is to adjust the signals such as the mean levels of the heterozygous SNPs relative to that of the homozygous SNPs is the same after as before the normalization.
pairedAlleleSpecificCopyNumbers

If argument `preserveScale=TRUE`, then SNPs that are heterozygous (in the matched normal) are corrected for signal compression using an estimate of signal compression based on the amount of correction performed by TumorBoost on SNPs that are homozygous (in the matched normal).

The option of preserving the scale is not discussed in the TumorBoost paper [1], which presents the `preserveScale=FALSE` version.

**Author(s)**

Henrik Bengtsson, Pierre Neuvial

**References**


**Examples**

```r
library(R.utils)

# Load data
pathname <- system.file("data-ex/TumorBoost,fracB,exampleData.Rbin", package="aroma.light")
data <- loadObject(pathname)
attachLocally(data)
pos <- position/1e6
muN <- genotypeN

layout(matrix(1:4, ncol=1))
par(mar=c(2.5,4,0.5,1)+0.1)
ylim <- c(-0.05, 1.05)
col <- rep("#999999", length(muN))
col[muN == 1/2] <- "#000000"

# Allele B fractions for the normal sample
plot(pos, betaN, col=col, ylim=ylim)

# Allele B fractions for the tumor sample
plot(pos, betaT, col=col, ylim=ylim)

# TumorBoost w/ naive genotype calls
betaTN <- normalizeTumorBoost(betaT=betaT, betaN=betaN, preserveScale=FALSE)
plot(pos, betaTN, col=col, ylim=ylim)

# TumorBoost w/ external multi-sample genotype calls
betaTNx <- normalizeTumorBoost(betaT=betaT, betaN=betaN, muN=muN, preserveScale=FALSE)
plot(pos, betaTNx, col=col, ylim=ylim)
```

**pair**

Calculating tumor-normal paired allele-specific copy number stratified on genotypes
Description
Calculating tumor-normal paired allele-specific copy number stratified on genotypes. The method is a single-sample (single-pair) method. It requires paired tumor-normal parent-specific copy number signals.

Usage
```r
## S3 method for class 'single.Var'
pairedAlleleSpecificCopyNumbers(thetaT, betaT, thetaN, betaN,
  muN=callNaiveGenotypes(betaN), ...)
```

Arguments
- `thetaT, betaT`: Theta and allele-B fraction signals for the tumor.
- `thetaN, betaN`: Total and allele-B fraction signals for the matched normal.
- `muN`: An optional vector of length J containing normal genotypes calls in (0,1/2,1,NA) for (AA,AB,BB).
- `...`: Not used.

Value
Returns a `data.frame` with elements CT, betaT and muN.

Author(s)
Pierre Neuvial, Henrik Bengtsson

See Also
This definition of calculating tumor-normal paired ASCN is related to how the `normalizeTumorBoost()` method calculates normalized tumor BAFs.

---

plotDensity

Plots density distributions for a set of vectors

Description
Plots density distributions for a set of vectors.

Usage
```r
## S3 method for class 'data.frame'
plotDensity(X, ..., xlab=NULL)
## S3 method for class 'matrix'
plotDensity(X, ..., xlab=NULL)
## S3 method for class 'numeric'
plotDensity(X, ..., xlab=NULL)
## S3 method for class 'list'
plotDensity(X, W=NULL, xlim=NULL, ylim=NULL, xlab=NULL,
  ylab="density (integrates to one)", col=1:length(X), lty=NULL, lwd=NULL, ...,
  add=FALSE)
```
**plotMvsA**

Arguments

- **X**: A single of `list` of numeric vectors or `density` objects, a numeric matrix, or a numeric data.frame.
- **W**: (optional) weights of similar data types and dimensions as X.
- **xlim,ylim**: character vector of length 2. The x and y limits.
- **xlab,ylab**: character string for labels on x and y axis.
- **col**: The color(s) of the curves.
- **lty**: The types of curves.
- **lwd**: The width of curves.
- **...**: Additional arguments passed to `density, plot, and lines`.
- **add**: If `TRUE`, the curves are plotted in the current plot, otherwise a new is created.

Author(s)

Henrik Bengtsson

See Also

Internally, `density` is used to estimate the empirical density.

---

**plotMvsA**

*Plot log-ratios vs log-intensities*

Description

Plot log-ratios vs log-intensities.

Usage

```r
## S3 method for class 'matrix'
plotMvsA(X, Alab="A", Mlab="M", Alim=c(0, 16), Mlim=c(-1, 1) * diff(Alim) * aspectRatio, 
aspectRatio=1, pch=".", ..., add=FALSE)
```

Arguments

- **X**: Nx2 matrix with two channels and N observations.
- **Alab, Mlab**: Labels on the x and y axes.
- **Alim, Mlim**: Plot range on the A and M axes.
- **aspectRatio**: Aspect ratio between Mlim and Alim.
- **pch**: Plot symbol used.
- **...**: Additional arguments accepted by `points`.
- **add**: If `TRUE`, data points are plotted in the current plot, otherwise a new plot is created.

Details

Red channel is assumed to be in column one and green in column two. Log-ratio are calculated taking channel one over channel two.
Value

Returns nothing.

Author(s)

Henrik Bengtsson

---

plotMvsAPairs

Plot log-ratios/log-intensities for all unique pairs of data vectors

Usage

```r
## S3 method for class 'matrix'
plotMvsAPairs(X, Alab="A", Mlab="M", Alim=c(0, 16), Mlim=c(-1, 1) * diff(Alim), pch=".", ..., add=FALSE)
```

Arguments

- `X`: NxK `matrix` where N is the number of observations and K is the number of channels.
- `Alab,Mlab`: Labels on the x and y axes.
- `Alim,Mlim`: Plot range on the A and M axes.
- `pch`: Plot symbol used.
- `...`: Additional arguments accepted by `points`.
- `add`: If `TRUE`, data points are plotted in the current plot, otherwise a new plot is created.

Details

Log-ratios and log-intensities are calculated for each neighboring pair of channels (columns) and plotted. Thus, in total there will be K-1 data set plotted.

The colors used for the plotted pairs are 1, 2, and so on. To change the colors, use a different color palette.

Value

Returns nothing.

Author(s)

Henrik Bengtsson
plotMvsMPairs

Plot log-ratios vs log-ratios for all pairs of columns.

Usage

```r
## S3 method for class 'matrix'
plotMvsMPairs(X, xlab="M", ylab="M", xlim=c(-1, 1) * 6, ylim=xlim, pch=".", ...
```

Arguments

- `X`: Nx2K matrix where N is the number of observations and 2K is an even number of channels.
- `xlab, ylab`: Labels on the x and y axes.
- `xlim, ylim`: Plot range on the x and y axes.
- `pch`: Plot symbol used.
- `...`: Additional arguments accepted by `points`.
- `add`: If TRUE, data points are plotted in the current plot, otherwise a new plot is created.

Details

Log-ratio are calculated by over paired columns, e.g. column 1 and 2, column 3 and 4, and so on.

Value

Returns nothing.

Author(s)

Henrik Bengtsson

plotXYCurve

Plot the relationship between two variables as a smooth curve.

Usage

```r
## S3 method for class 'numeric'
plotXYCurve(x, y, col=1L, lwd=2, dlwd=1, dcol=NA, xlim=NULL, ylim=xlim, xlab=NULL, ylab=NULL, curveFit=smooth.spline, ..., add=FALSE)
## S3 method for class 'matrix'
plotXYCurve(X, Y, col=seq_len(nrow(X)), lwd=2, dlwd=1, dcol=NA, xlim=NULL, ylim=xlim, xlab=NULL, ylab=NULL, curveFit=smooth.spline, ..., add=FALSE)
```
Arguments

\(x, y, X, Y\) Two \textit{numeric vectors} of length \(N\) for one curve (\(K=1\)), or two \textit{numeric} \(NxK\) matrices for \(K\) curves.

col The color of each curve. Either a scalar specifying the same value of all curves, or a \textit{vector} of \(K\) curve-specific values.

dlwd The line width of each curve. Either a scalar specifying the same value of all curves, or a \textit{vector} of \(K\) curve-specific values.

dcol The width of each density curve.

xlim, ylim The \(x\) and \(y\) plotting limits.

\textit{curveFit} The \textit{function} used to fit each curve. The two first arguments of the function must take \(x\) and \(y\), and the function must return a \textit{list} with fitted elements \(x\) and \(y\).

... Additional arguments passed to \textit{lines} used to draw each curve.

add If \texttt{TRUE}, the graph is added to the current plot, otherwise a new plot is created.

Value

Returns nothing.

Missing values

Data points \((x,y)\) with non-finite values are excluded.

Author(s)

Henrik Bengtsson

---

\texttt{robustSmoothSpline} \hspace{1cm} \textit{Robust fit of a Smoothing Spline}

Description

Fits a smoothing spline robustly using the \(L_1\) norm. Currently, the algorithm is an \textit{iterative reweighted smooth spline} algorithm which calls \texttt{smooth.spline}(\(x, y, w, \ldots\)) at each iteration with the weights \(w\) equal to the inverse of the absolute value of the residuals for the last iteration step.

Usage

\begin{verbatim}
## Default S3 method:
robustSmoothSpline(x, y=NULL, w=NULL, ..., minIter=3, maxIter=max(minIter, 50),
method=c("L1", "symmetric"), sdCriteria=2e-04, reps=1e-15, tol=1e-06 * IQR(x),
plotCurves=FALSE)
\end{verbatim}
robustSmoothSpline

Arguments

- **x**: a vector giving the values of the predictor variable, or a list or a two-column matrix specifying x and y. If x is of class smooth.spline then x$x is used as the x values and x$yin are used as the y values.
- **y**: responses. If y is missing, the responses are assumed to be specified by x.
- **w**: a vector of weights the same length as x giving the weights to use for each element of x. Default value is equal weight to all values.
- **...**: Other arguments passed to smooth.spline.
- **minIter**: the minimum number of iterations used to fit the smoothing spline robustly. Default value is 3.
- **maxIter**: the maximum number of iterations used to fit the smoothing spline robustly. Default value is 25.
- **method**: the method used to compute robustness weights at each iteration. Default value is "L1", which uses the inverse of the absolute value of the residuals. Using "symmetric" will use Tukey’s biweight with cut-off equal to six times the MAD of the residuals, equivalent to lowess.
- **sdCriteria**: Convergence criteria, which the difference between the standard deviation of the residuals between two consecutive iteration steps. Default value is 2e-4.
- **reps**: Small positive number added to residuals to avoid division by zero when calculating new weights for next iteration.
- **tol**: Passed to smooth.spline (R >= 2.14.0).
- **plotCurves**: If TRUE, the fitted splines are added to the current plot, otherwise not.

Value

Returns an object of class smooth.spline.

Author(s)

Henrik Bengtsson

See Also

This implementation of this function was adopted from smooth.spline of the stats package. Because of this, this function is also licensed under GPL v2.

Examples

data(cars)
attach(cars)
plot(speed, dist, main = "data(cars) & robust smoothing splines")

# Fit a smoothing spline using L_2 norm
cars.spl <- smooth.spline(speed, dist)
lines(cars.spl, col = "blue")

# Fit a smoothing spline using L_1 norm
cars.rspl <- robustSmoothSpline(speed, dist)
lines(cars.rspl, col = "red")

# Fit a smoothing spline using L_2 norm with 10 degrees of freedom
lines(smooth.spline(speed, dist, df=10), lty=2, col = "blue")

# Fit a smoothing spline using L_1 norm with 10 degrees of freedom
lines(robustSmoothSpline(speed, dist, df=10), lty=2, col = "red")

legend(5,120, c(
paste("smooth.spline [C.V.] => df =",round(cars.spl$df,1)),
paste("robustSmoothSpline [C.V.] => df =",round(cars.rspl$df,1)),
"standard with s( * , df = 10)","robust with s( * , df = 10)"
), col = c("blue","red","blue","red"), lty = c(1,1,2,2), bg='bisque')

---

sampleCorrelations  
Calculates the correlation for random pairs of observations

Description

Calculates the correlation for random pairs of observations.

Usage

```r
## S3 method for class 'matrix'
sampleCorrelations(X, MARGIN=1, pairs=NULL, npairs=max(5000, nrow(X)), ...)
```

Arguments

- **X**: An NxK matrix where N >= 2 and K >= 2.
- **MARGIN**: The dimension (1 or 2) in which the observations are. If MARGIN==1 (==2), each row (column) is an observation.
- **pairs**: If a Lx2 matrix, the L index pairs for which the correlations are calculated. If NULL, pairs of observations are sampled.
- **npairs**: The number of correlations to calculate.
- **...**: Not used.

Value

Returns a double vector of length npairs.

Author(s)

Henrik Bengtsson

References


See Also

`sample()`
Examples

# Simulate 20000 genes with 10 observations each
X <- matrix(rnorm(n=20000), ncol=10)

# Calculate the correlation for 5000 random gene pairs
cor <- sampleCorrelations(X, npairs=5000)
print(summary(cor))

---

sampleTuples

Sample tuples of elements from a set

Description

Sample tuples of elements from a set. The elements within a sampled tuple are unique, i.e. no two elements are the same.

Usage

```r
## Default S3 method:
sampleTuples(x, size, length, ...)
```

Arguments

- `x`: A set of elements to sample from.
- `size`: The number of tuples to sample.
- `length`: The length of each tuple.
- `...`: Additional arguments passed to `sample()`.

Value

Returns a NxK matrix where N = size and K = length.

Author(s)

Henrik Bengtsson

See Also

`sample()`.

Examples

```r
pairs <- sampleTuples(1:10, size=5, length=2)
print(pairs)

triples <- sampleTuples(1:10, size=5, length=3)
print(triples)

# Allow tuples with repeated elements
quadruples <- sampleTuples(1:3, size=5, length=4, replace=TRUE)
print(quadruples)
```
Light-weight Weighted Principal Component Analysis

Description

Calculates the (weighted) principal components of a matrix, that is, finds a new coordinate system (not unique) for representing the given multivariate data such that i) all dimensions are orthogonal to each other, and ii) all dimensions have maximal variances.

Usage

```r
## S3 method for class 'matrix'
w pca(x, w=NULL, center=TRUE, scale=FALSE, method=c("dgesdd", "dgesvd"),
    swapDirections=FALSE, ...)
```

Arguments

- `x`: An N x K matrix.
- `w`: An N vector of weights for each row (observation) in the data matrix. If NULL, all observations get the same weight, that is, standard PCA is used.
- `center`: If TRUE, the (weighted) sample mean column vector is subtracted from each column in mat, first. If data is not centered, the effect will be that a linear subspace that goes through the origin is fitted.
- `scale`: If TRUE, each column in mat is divided by its (weighted) root-mean-square of the centered column, first.
- `method`: If "dgesdd" LAPACK's divide-and-conquer based SVD routine is used (faster [1]). If "dgesvd", LAPACK's QR-decomposition-based routine is used.
- `swapDirections`: If TRUE, the signs of eigenvectors that have more negative than positive components are inverted. The signs of corresponding principal components are also inverted. This is only of interest when for instance visualizing or comparing with other PCA estimates from other methods, because the PCA (SVD) decomposition of a matrix is not unique.
- `...`: Not used.

Value

Returns a list with elements:

- `pc`: An N x K matrix where the column vectors are the principal components (a.k.a. loading vectors, spectral loadings or factors etc).
- `d`: An K vector containing the eigenvalues of the principal components.
- `vt`: An K x K matrix containing the eigenvector of the principal components.
- `xMean`: The center coordinate.

It holds that x == t(t(fit$pc %*% fit$vt) + fit$xMean).
Method

A singular value decomposition (SVD) is carried out. Let $X = \text{mat}$, then the SVD of the matrix is $X = UDV'$, where $U$ and $V$ are orthogonal, and $D$ is a diagonal matrix with singular values. The principal returned by this method are $UD$.

Internally `La.svd()` (or `svd()`) of the `base` package is used. For a popular and well written introduction to SVD see for instance [2].

Author(s)

Henrik Bengtsson

References


See Also

For a iterative re-weighted PCA method, see `iwpca()`. For Singular Value Decomposition, see `svd()`. For other implementations of Principal Component Analysis functions see (if they are installed): `prcomp` in package `stats` and `pca()` in package `pcurve`.

Examples

```r
for (zzz in 0) {
    # This example requires plot3d() in R.basic [http://www.braju.com/R/]
    if (!require(pkgName <- "R.basic", character.only=TRUE)) break

    # A first example
    # Simulate data from the model y <- a + bx + eps(bx)
    x <- rexp(1000)
    a <- c(2,15,3)
    b <- c(2,3,15)
    bx <- outer(b,x)
    eps <- apply(bx, MARGIN=2, FUN=function(x) rnorm(length(x), mean=0, sd=0.1*x))
    y <- a + bx + eps
    y <- t(y)

    # Add some outliers by permuting the dimensions for 1/3 of the observations
    idx <- sample(1:nrow(y), size=1/3*nrow(y))
    y[idx,] <- y[idx,c(2,3,1)]

    # Down-weight the outliers W times to demonstrate how weights are used
    W <- 10

    # Plot the data with fitted lines at four different view points
    N <- 4
    theta <- seq(0,180,length.out=N)
    phi <- rep(30, length.out=N)
```

```
# Use a different color for each set of weights
col <- topo.colors(W)

opar <- par(mar=c(1,1,1,1)+0.1)
layout(matrix(1:N, nrow=2, byrow=TRUE))
for (kk in seq_along(theta)) {
  # Plot the data
  plot3d(y, theta=theta[kk], phi=phi[kk])

  # First, same weights for all observations
  w <- rep(1, length=nrow(y))

  for (ww in 1:W) {
    # Fit a line using IWPCA through data
    fit <- wpca(y, w=w, swapDirections=TRUE)

    # Get the first principal component
    ymid <- fit$xMean
    d0 <- apply(y, MARGIN=2, FUN=min) - ymid
    d1 <- apply(y, MARGIN=2, FUN=max) - ymid
    b <- fit$vt[,1]
    y0 <- -b * max(abs(d0))
    y1 <- b * max(abs(d1))
    yline <- matrix(c(y0, y1), nrow=length(b), ncol=2)
    yline <- yline + ymid
    points3d(t(ymid), col=col)
    lines3d(t(yline), col=col)

    # Down-weight outliers only, because here we know which they are.
    w[idx] <- w[idx]/2
  }

  # Highlight the last one
  lines3d(t(yline), col="red", lwd=3)
}

par(opar)
)
# for (zzz in 0)
rm(zzz)

if (dev.cur() > 1) dev.off()

# -----------------------------------------------
# A second example
# -----------------------------------------------
# Data
x <- c(1,2,3,4,5)
y <- c(2,4,3,3,6)

opar <- par(bty="L")
opalette <- palette(c("blue", "red", "black"))
xlim <- ylim <- c(0,6)
# Plot the data and the center mass
plot(x,y, pch=16, cex=1.5, xlim=xlim, ylim=ylim)
points(mean(x), mean(y), cex=2, lwd=2, col="blue")

# Linear regression y ~ x
fit <- lm(y ~ x)
abline(fit, lty=1, col=1)

# Linear regression y ~ x through without intercept
fit <- lm(y ~ x - 1)
abline(fit, lty=2, col=1)

# Linear regression x ~ y
fit <- lm(x ~ y)
c <- coefficients(fit)
b <- 1/c[2]
a <- -b*c[1]
abline(a=a, b=b, lty=1, col=2)

# Linear regression x ~ y through without intercept
fit <- lm(x ~ y - 1)
b <- 1/coefficients(fit)
abline(a=0, b=b, lty=2, col=2)

# Orthogonal linear "regression"
fit <- wpca(cbind(x,y))
b <- fit$vt[1,2]/fit$vt[1,1]
a <- fit$xMean[2]-b*fit$xMean[1]
abline(a=a, b=b, lwd=2, col=3)

# Orthogonal linear "regression" without intercept
fit <- wpca(cbind(x,y), center=FALSE)
b <- fit$vt[1,2]/fit$vt[1,1]
a <- fit$xMean[2]-b*fit$xMean[1]
abline(a=a, b=b, lty=2, lwd=2, col=3)

legend(xlim[1],ylim[2], legend=c("lm(y~x)", "lm(y~x-1)", "lm(x~y)", "lm(x~y-1)", "pca", "pca w/o intercept"), lty=rep(1:2,3),
       lwd=rep(c(1,1,2),each=2), col=rep(1:3,each=2))

palette(opalette)
par(opar)
Index

*Topic algebra
  - distanceBetweenLines, 16
  - fitIWPCA, 19
  - iwpca, 25
  - medianPolish, 27
  - wpca, 60

*Topic documentation
  1. Calibration and Normalization, 4

*Topic methods
  - averageQuantile, 7
  - backtransformAffine, 8
  - backtransformPrincipalCurve, 9
  - calibrateMultiscan, 12
  - callNaiveGenotypes, 15
  - fitIWPCA, 19
  - fitNaiveGenotypes, 20
  - fitPrincipalCurve, 21
  - fitXYCurve, 24
  - iwpca, 25
  - medianPolish, 27
  - normalizeAffine, 29
  - normalizeAverage, 33
  - normalizeCurveFit, 34
  - normalizeDifferencesToAverage, 37
  - normalizeQuantileRank, 45
  - normalizeQuantileRank.matrix, 46
  - normalizeQuantileSpline, 48
  - normalizeTumorBoost, 49
  - pairedAlleleSpecificCopyNumbers, 51
  - plotDensity, 52
  - plotMvsA, 53
  - plotMvsAPairs, 54
  - plotMvsMPairs, 55
  - plotXYCurve, 55
  - sampleCorrelations, 58
  - wpca, 60

*Topic multivariate
  - averageQuantile, 7
  - normalizeQuantileRank, 45
  - normalizeQuantileRank.matrix, 46
  - normalizeQuantileSpline, 48
  - plotXYCurve, 55

*Topic nonparametric
  - averageQuantile, 7
  - normalizeFragmentLength, 39
  - normalizeQuantileRank, 45
  - normalizeQuantileRank.matrix, 46
  - normalizeQuantileSpline, 48
  - plotXYCurve, 55

*Topic package
  - aroma.light-package, 2

*Topic robust
  - averageQuantile, 7
  - normalizeFragmentLength, 39
  - normalizeQuantileRank, 45
  - normalizeQuantileRank.matrix, 46
  - normalizeQuantileSpline, 48
  - plotXYCurve, 55
  - robustSmoothSpline, 56

*Topic smooth
  - robustSmoothSpline, 56

*Topic utilities
  - sampleCorrelations, 58
  - sampleTuples, 59

  1. Calibration and Normalization, 4, 14

  - callNaiveGenotypes, 15
  - calibrateMultiscan, 3, 12, 20, 30, 31
  - character, 13, 19, 21, 24, 29, 34, 47, 50, 53
  - data.frame, 52, 53
  - density, 53
  - distanceBetweenLines, 16, 20
  - double, 20, 21, 24, 34, 58
  - FALSE, 20, 28, 50
  - findPeaksAndValleys, 21
  - fitIWPCA, 13, 19, 29, 30

64
INDEX

fitNaiveGenotypes, 15, 20
fitPrincipalCurve, 10, 21
fitXYCurve, 23
function, 13, 25, 33, 38–40, 56

Inf, 15, 21
integer, 21, 33, 38, 39
iwpca, 13, 19, 20, 29, 30, 61

lines, 53, 56
list, 7, 8, 17, 20–22, 24, 26, 28, 33, 38–40, 45, 48, 50, 53, 56, 57, 60
loess, 14, 24, 35
logical, 15, 21, 22, 39
lowess, 39, 57

matrix, 7, 8, 10, 13, 19, 20, 22, 24, 25, 28–30, 33–35, 39, 47, 48, 53–60
mean, 47
median, 47
medianPolish, 27
medpolish, 28

NA, 13, 15, 22, 24, 28, 30, 35, 45, 47, 50, 52
NaN, 35
normalizeAffine, 3, 14, 20, 29, 35
normalizeAverage, 33
normalizeCurveFit, 3, 34
normalizeDifferencesToAverage, 37
normalizeFragmentLength, 39
normalizeLoess (normalizeCurveFit), 34
normalizeLowess (normalizeCurveFit), 34
normalizeQuantile
(normalizeQuantileRank), 45
normalizeQuantileRank, 3, 7, 45, 49
normalizeQuantileRank.matrix, 46
normalizeQuantileSpline, 3, 7, 46, 47, 48
normalizeRobustSpline
(normalizeCurveFit), 34
normalizeSpline (normalizeCurveFit), 34
normalizeTumorBoost, 49, 52
NULL, 13, 20, 21, 24–26, 29, 33, 34, 38–40, 45, 47, 48, 58, 60
numeric, 7, 13, 15, 19, 21, 24, 28, 30, 33, 35, 38, 39, 45, 48, 50, 53, 56

pairedAlleleSpecificCopyNumbers, 51
plot, 53
plotDensity, 52
plotMvsA, 53
plotMvsAPairs, 54
plotMvsMPairs, 55
plotXYCurve, 55

points, 53–55
prcomp, 61
principal_curve, 22
quantile, 7
robustSmoothSpline, 48, 49, 56
sample, 58, 59
sampleCorrelations, 58
sampleTuples, 59
smooth.spline, 10, 48, 49, 57
svd, 61
TRUE, 8, 13, 20, 28, 30, 47, 48, 50, 53–57, 60
vector, 7, 8, 13, 15–17, 19–21, 24, 25, 29, 30, 34, 35, 39, 45, 47, 48, 50, 52, 53, 56–58, 60
Verbose, 15, 21, 22
weighted.mean, 47
weightedMedian, 47
w pca, 26, 60