# Package ‘RaProR’

August 6, 2019

**Version** 1.1-5

**Date** 2019-08-06

**Title** Calculate Sketches using Random Projections to Reduce Large Data Sets

**Imports** utils

**Description** Calculate sketches of a data set reducing the number of observations using random projections. These can be used for Bayesian or frequentist linear regression on large data sets as described in Geppert et. al (2017) <doi:10.1007/s11222-015-9608-z>.

**License** GPL (>= 3)

**NeedsCompilation** yes

**Author** Leo N. Geppert [aut, cre, cph],
Katja Ickstadt [aut],
Alexander Munteanu [aut],
Jens Quedenfeld [aut, cph],
Ludger Sandig [aut, cph],
Christian Sohler [aut]

**Maintainer** Leo N. Geppert <geppert@statistik.uni-dortmund.de>

**Repository** CRAN

**Date/Publication** 2019-08-06 08:50:02 UTC

## R topics documented:

## RaProR-package *Create a smaller substitute data set to perform linear regression*

### Description

This package can be used to calculate sketches of a data set that can be used to perform approximate classical or Bayesian linear regression. The sketch is a substitute data set of the same dimension but much smaller number of observations. The inference based on the sketch is much faster and is provably close to the exact inference. The calculation is done time- and space-efficiently in C. The two main functions are `sketch` for data sets that fit into the working memory and can be processed at once and `readinandsketch` for data sets that (potentially) do not fit into the working memory and will be read and sketched sequentially blockwise.

### Author(s)

LN Geppert, K. Ickstadt, A. Munteanu, J. Quedenfeld, L. Sandig, C. Sohler

### References

Geppert, L., Ickstadt, K., Munteanu, A., Quedenfeld, J., Sohler, C. (2017). Random projections for Bayesian regression. Statistics and Computing, 27(1), 79-101. doi:10.1007/s11222-015-9608-z

### Examples

```
# create a small simulated data set
# with 400 observations and
# 4 variables
set.seed(23)
x1 = rnorm(400, 10, 2)
x2 = rnorm(400, 5, 3)
x3 = rnorm(400, -2, 1)
x4 = rnorm(400, 0, 5)
y = 2.4 - 0.6 * x1 + 5.5 * x2 - 7.2 * x3 + 5.7 * x4 + rnorm(400)
# all in one data.frame
data = data.frame(x1, x2, x3, x4, y)

# linear model based on original data set
lm(y ~ ., data = data)

# Calculate an RAD/"R"-sketch with epsilon = 0.2
s1 = sketch(data, epsilon = 0.2, method = 'R', affine = TRUE)
dim(s1)
# very similar results, intercept should be omitted
lm(y ~ . - 1, data = s1)
```

## readinandsketch — *Create a sketch from a file containing a (very large) data set*

### Description

This function calculates a sketch of a file. The sketch can be used to perform approximate frequentist or Bayesian linear regression. The sketch is a substitute data set of the same dimension but much smaller number of observations. The analysis based on the sketch is much faster and its results are provably close to the results on the original data set. The file is read in sequentially, making it possible to sketch data sets that are too large to be loaded into R completely.

### Usage

```
readinandsketch(file, nrows = 50000, epsilon = NULL, obs_sketch = NULL,
                affine = TRUE, method= c("C", "S", "R"), header = FALSE,
                sep = "", col.names, skip = 0, warn = FALSE, ...)
```

### Arguments

`file` The name of a file that contains the (large) data set. The data set should consist of both the design matrix X and the vector Y, which contains the values of the dependent variable. The order is arbitrary.

`nrows` A positive integer, which controls, how many rows are read into the memory per iteration. Differs from use in `read.table` as the other rows will be read in subsequent iterations. For that reason, *nrows* has to be larger than 0.

`epsilon` Approximation error of the sketch (see Details). Only one of *epsilon* and *obs_sketch* can be used, if both are specified, currently *epsilon* is used and *obs_sketch* is ignored. Possible values for epsilon lie in the interval (0, 0.5].

`obs_sketch` Desired number of observations of the sketch (see Details). Only one of *epsilon* and *obs_sketch* can be used, if both are specified, currently *epsilon* is used and *obs_sketch* is ignored.

`affine` Boolean, choose TRUE if your model includes an intercept term and your data set does not contain a corresponding column. The corresponding column will be added as new left-most column of the sketch. If you do not want an added intercept term, choose FALSE.

`method` The sketching method to be used. Possible values are "R", "S", and "C". See Details.

`header` Boolean, if TRUE, the first line of the file is used as variable names, see `read.table`.

`sep` The field separator character, see `read.table`.

`col.names` An optional vector containing the variable names, see `read.table`.

`skip` integer: the number of lines of the data file to skip before beginning to read data, see `read.table`.

`warn` Boolean, if TRUE show a warning if the sketch will result in a matrix of larger dimension than the original matrix.

`...` Additional arguments that will be passed on to `read.table`.

## Details

This function reads a data set iteratively and calculates/updates a sketch of the read in data set. This sketch can then be used for frequentist or Bayesian linear regression, especially on large data sets. The functionality used here is the same as in `sketch`, but *readinandsketch* can also handle data sets that are too large to be loaded into the working memory.

In principle, *nrows* can be any positive integer value. If using the methods "R" or "C", small integer values will only lead to an increased running time. If using method "S", however, *nrows* has to be at least as large as the number of observations $k$ in the sketch, otherwise there will be an error.

If the number of observations in the data set is a multiple of *nrows*, there will be one additional empty run, where no data is read and a sketch of an empty data set is calculated. This does not influence the resulting sketch.

## Value

Returns a data frame, which contains both the sketched data frame SX and the sketched vector SY. The order of the columns is the same as in the original data set. If affine is TRUE, the corresponding intercept column is added as the new left-most column of the sketch. Please omit the standard intercept term from any models based on sketches in that case.

## Author(s)

LN Geppert, K. Ickstadt, A. Munteanu, J. Quedenfeld, L. Sandig, C. Sohler

## References

Geppert, L., Ickstadt, K., Munteanu, A., Quedenfeld, J., & Sohler, C. (2017). Random projections for Bayesian regression. Statistics and Computing, 27(1), 79-101. doi:10.1007/s11222-015-9608-z

## See Also

`sketch`

## Examples

```
## Not run:
  sketchC = readinandsketch(file.choose(), header = TRUE, sep = '\t',
  nrows = 10000, epsilon = 0.1, method = 'R')

## End(Not run)
```

## sketch *Create a sketch from a matrix, data.frame or file*

### Description

This function calculates a sketch of a data set, matrix or file. The sketch can be used to perform approximate frequentist or Bayesian linear regression. The sketch is a substitute data set of the same dimension but much smaller number of observations. The analysis based on the sketch is much faster and its results are provably close to the results on the original data set.

### Usage

```
sketch(data, file, epsilon = NULL, obs_sketch = NULL,
       warn = TRUE, affine = TRUE, method= c("R", "S", "C"), ...)
```

### Arguments

`data` Matrix or data.frame which contains both the design matrix X and the vector Y used for the linear regression, in any order. Either *data* or *file* has to be provided.

`file` The name of a file which contains the matrix or data.frame. Ignored, if *data* is specified. As in *data*, this file should contain both the design matrix X and the vector Y used in the linear regression, in any order. Either *file* or *data* has to be provided. If your data set is very large, consider using `readinandsketch` for more efficient sketching of the data.

`epsilon` Approximation error of the sketch (see Details). Only one of *epsilon* and *obs_sketch* can be used, if both are specified, currently *epsilon* is used and *obs_sketch* is ignored. Possible values for epsilon lie in the interval (0, 0.5].

`obs_sketch` Desired number of observations of the sketch (see Details). Only one of *epsilon* and *obs_sketch* can be used, if both are specified, currently *epsilon* is used and *obs_sketch* is ignored. If method "C" is chosen, the number is rounded up to the next power of two, see Details.

`warn` Boolean, if TRUE show a warning if the sketch will result in a matrix of larger dimension than the original matrix. Please note that a sketch with larger dimension than the original matrix will result in an error if method "S" is used.

`affine` Boolean, choose TRUE if your model includes an intercept term and your data set does not contain a corresponding column. The corresponding column will be added as new left-most column of the sketch. If you do not want an added intercept term, choose FALSE.

`method` The sketching method to be used. Possible values are "R", "S", and "C", where "R" is the default. See Details.

`...` Additional arguments passed on to `read.table` if *file* is specified.

## Details

Let X be a $(n \times d)$-matrix and Y a $(n \times 1)$-vector. This function produces an implicit matrix S and efficiently performs the multiplication, which embeds X and Y into a lower dimension $k$, with $k \ll n$. The value of k depends on the method used. For "R" and "S", the formula is $k = \left\lceil \frac{d \cdot \ln(d)}{\varepsilon^2} \right\rceil$, for "C", this changes to $k = 2^{\lceil \log_2 s \rceil}$, where $s = \left\lceil \frac{d^2}{20 \cdot \varepsilon^2} \right\rceil$, that is the smallest power of two larger than s.

The function outputs the sketched data frame and vector SX and SY. These can be used to conduct frequentist or Bayesian linear regression on a smaller data set, saving running time and memory. The results are guaranteed to be close to the results that would have been obtained on the original data set in the sense that the original likelihood is closely approximated by the likelihood on the sketched data set in the case of classical linear regression, or, in the Bayesian case the original posterior is closely approximated by the posterior on the sketched data set.

When using methods "R" and "C", it is possible for the sketch to be of a larger dimension than the original matrix. When using method "S", this will result in an error. Such cases occur when the number of variables is relatively large compared to the number of observations.

For more details, please refer to Geppert et al. (2017) and the references mentioned therein.

## Value

Returns a data frame, which contains both the sketched data frame SX and the sketched vector SY. The order of the columns is the same as in the original data set. If affine is TRUE, the corresponding intercept column is added as the new left-most column of the sketch. Please omit the standard intercept term from any models based on sketches in that case.

## Author(s)

LN Geppert, K. Ickstadt, A. Munteanu, J. Quedenfeld, L. Sandig, C. Sohler

## References

Geppert, L., Ickstadt, K., Munteanu, A., Quedenfeld, J., & Sohler, C. (2017). Random projections for Bayesian regression. Statistics and Computing, 27(1), 79-101. doi:10.1007/s11222-015-9608-z

## See Also

readinandsketch

## Examples

```
# create a small simulated data set
# with 400 observations and
# 4 variables
set.seed(23)
x1 = rnorm(400, 10, 2)
x2 = rnorm(400, 5, 3)
x3 = rnorm(400, -2, 1)
x4 = rnorm(400, 0, 5)
y = 2.4 - 0.6 * x1 + 5.5 * x2 - 7.2 * x3 + 5.7 * x4 + rnorm(400)
```

```
# all in one data.frame
data = data.frame(x1, x2, x3, x4, y)

# linear model based on original data set
lm(y ~ ., data = data)

# Calculate a RAD/"R"-sketch with epsilon = 0.2
s1 = sketch(data, epsilon = 0.2, method = 'R', affine = TRUE)
dim(s1)
# very similar results, intercept should be omitted
lm(y ~ . - 1, data = s1)

# use option "obs_sketch" to fix the new number of observations
s2 = sketch(data, obs_sketch = 200, method = 'R', affine = TRUE)
dim(s2)
# some more differences as sketch is smaller
lm(y ~ . - 1, data = s2)

# calculate SRHT/"S"-sketch
s3 = sketch(data, epsilon = 0.2, method = 'S', affine = TRUE)
dim(s3)
lm(y ~ . - 1, data = s3)

# calculate CW/"C"-sketch
s4 = sketch(data, epsilon = 0.2, method = 'C', affine = TRUE)
dim(s4)
# sketch is smaller, because the number of variables is very small
# CW-sketches require a lot more observations compared to RAD/SRHT
# when number of variables increases
lm(y ~ . - 1, data = s4)

# same simulated data set, but with intercept added to data.frame
data2 = data.frame(x0 = 1, x1, x2, x3, x4, y)
lm(y ~ . - 1, data = data2)

# Same as s1, but now option affine = FALSE is adequate
s5 = sketch(data2, epsilon = 0.2, method = 'R', affine = FALSE)
dim(s5)
lm(y ~ . - 1, data = s5)
```

# Index