Package: datanugget (via r-universe)

October 20, 2026

Type Package

Title Create, Optimize, and Refine Data Nuggets

Version 1.5.0

Date 2026-08-12

Author Rituparna Dey [aut, cre], Yajie Duan [aut], Traymon Beavers [aut], Javier Cabrera [aut], Ge Cheng [aut], Kunting Qi [aut], Mariusz Lubomirski [aut]

Maintainer Rituparna Dey <rituparnadey525@gmail.com>

Description Creating, optimizing and refining data nuggets. Data nuggets reduce a large dataset into a small collection of nuggets of data, each containing a center (location), weight (importance), and scale (variability) parameter. Data nugget centers are selected based on a space-filling maximum-entropy scheme. Data nugget weights are created by counting the number observations closest to a given data nugget center. We then say the data nugget 'contains' these observations and the data nugget center is recalculated as the mean of these observations. Data nugget scales are created by calculating the trace of the covariance matrix of the observations contained within a data nugget divided by the dimension of the dataset. The optimal number of data nuggets is determined data-driven based on the relative second-order differences of propensity score indices. Data nuggets are refined by 'splitting' data nuggets which have high scales or elongated shapes (defined as the ratio of the two largest eigenvalues of the covariance matrix of the observations contained within the data nugget).

Depends R (>= 4.0), doSNOW (>= 1.0.16), doParallel(>= 1.0.17), foreach (>= 1.5.1), parallel (>= 4.0.5), Rfast(>= 2.0.7), mgcv(>= 1.9-4), ggplot2(>= 4.0.2)

License GPL-2

Encoding UTF-8

NeedsCompilation no

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Config/roxygen2/version 8.1.0

Config/pak/sysreqs cmake

Repository https://ritde-art.r-universe.dev

Date/Publication 2026-08-21 06:50:08 UTC

RemoteUrl https://github.com/cran/datanugget

RemoteRef HEAD

RemoteSha 53acfe021e932c2e8371bc3e3f038bc2090952f1

Contents

datanugget-package 2
AC 3
create.DN 4
create.DNcenters 7
getDN.means 8
optimal.DN 9
refine.DN 12

Index **16**

`datanugget-package` *Data Nuggets*

Description

This package contains functions to create, optimize and refine data nuggets which serve as representative samples of large datasets. The functions which perform these processes are create.DN, optimal.DN and refine.DN respectively.

Author(s)

Rituparna Dey, Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

AC *Calculate Arithmetic Complexicity of the Algorithm That Creates Data Nuggets*

Description

This function creates the centers of data nuggets from a random sample.

Usage

```
AC(x,
   R,
   delete.percent,
   DN.num1,
   DN.num2)
```

Arguments

x	A data matrix (of class matrix, data.frame, or data.table) containing only entries of class numeric.
R	The number of observations to sample from the data matrix when creating the initial data nugget centers. Must be of class numeric within [100,10000].
delete.percent	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1).
DN.num1	The number of initial data nugget centers to create. Must be of class numeric.
DN.num2	The number of data nuggets to create. Must be of class numeric.

Details

This function is used for calculating the arithmetic complexicity of the algorithm behind the create.DN function for the given parameter choices.

Value

my.AC	The arithmetic complexicity of the algorithm behind the create.DN function for the given parameter choices on a log10 scale.

Author(s)

Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
X = cbind.data.frame(rnorm(10^6),
                     rnorm(10^6),
                     rnorm(10^6),
                     rnorm(10^6),
                     rnorm(10^6))

my.AC = AC(x = X,
           R = 5000,
           delete.percent = .1,
           DN.num1 = 10^4,
           DN.num2 = 2000)
```

`create.DN` *Create Data Nuggets*

Description

This function draws a random sample of observations from a large dataset and creates data nuggets, a type of representative sample of the dataset, using a specified distance metric.

Usage

```
create.DN(x,
          center.method = "mean",
          R = 5000,
          delete.percent = .1,
          DN.num1 = 10^4,
          DN.num2 = 2000,
          dist.metric = "euclidean",
          seed = 291102,
          no.cores = (parallel::detectCores() - 1),
          make.pbs = FALSE)
```

Arguments

x	A data matrix (of class matrix, data.frame, or data.table) containing only entries of class numeric.
center.method	The method used for creating data nugget centers. Must be 'mean' or 'random' or 'original'. 'mean' chooses the data nugget center to be the mean of all observations within that data nugget, 'random' chooses the data nugget center to be some random observation within that data nugget, and 'original' chooses the original data nugget centers generated by the final run of datanugget creation using create.DNcenters function. Default is 'mean'.
R	The number of observations to sample from the data matrix when creating the initial data nugget centers. Must be of class numeric within [100,10000]. Default is 5000.
delete.percent	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1). Default is 0.1.
DN.num1	The number of initial data nugget centers to create. Must be of class numeric. Default is 10^4.
DN.num2	The number of final data nuggets to create. Must be of class numeric. Default is 2000.
dist.metric	The distance metric used to create the initial centers of data nuggets. Must be 'euclidean' or 'manhattan'. Default is 'euclidean'.
seed	Random seed for replication. Must be of class numeric. Default is 291102.
no.cores	Number of cores used for parallel processing. If '0' then parallel processing is not used. Must be of class numeric.
make.pbs	Logical; whether to show a progress bar while the function runs. Default is FALSE.

Details

Data nuggets are a representative sample meant to summarize Big Data by reducing a large dataset to a much smaller dataset by eliminating redundant points while also preserving the peripheries of the dataset. Each data nugget is defined by a center (location), weight (importance), and scale (internal variability). This function creates data nuggets using Algorithm 1 provided in the reference.

Value

An object of class datanugget:

Data Nuggets	DN.num2 by (ncol(x)+3) data frame containing the information for the data nuggets created (index, center, weight and scale).
Data Nugget Assignments	Vector of length nrow(x) containing the data nugget assignment of each observation in x.

Author(s)

Rituparna Dey, Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
  ## small example
  X = cbind.data.frame(rnorm(10^3),
                       rnorm(10^3),
                       rnorm(10^3))

  suppressMessages({

    my.DN = create.DN(x = X,
                      R = 500,
                      delete.percent = .1,
                      DN.num1 = 500,
                      DN.num2 = 250,
                      no.cores = 0,
                      make.pbs = FALSE)

  })

  my.DN$`Data Nuggets`
  my.DN$`Data Nugget Assignments`

## Not run:

  ## large example
  X = cbind.data.frame(rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4))
  t1 <- Sys.time()
  my.DN = create.DN(x = X,
                    R = 5000,
                    delete.percent = .1,
                    DN.num1 = 10^4,
                    DN.num2 = 2000,
                    no.cores = 2)
  t2 <- Sys.time()
  my.DN$`Data Nuggets`
  my.DN$`Data Nugget Assignments`


## End(Not run)
```

`create.DNcenters` *Create Data Nugget Centers*

Description

This function creates the centers of data nuggets from a random sample.

Usage

```
create.DNcenters(RS,
                 delete.percent = 0.1,
                 DN.num,
                 dist.metric = "euclidean",
                 make.pbs = FALSE)
```

Arguments

`RS`	A data matrix (data frame, data table, matrix, etc) containing only entries of class numeric.
`delete.percent`	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1). Default value is 0.1.
`DN.num`	The number of data nuggets to create. Must be of class numeric.
`dist.metric`	The distance metric used to create the initial centers of data nuggets. Must be 'euclidean' or 'manhattan'. Default is "euclidean".
`make.pbs`	Logical; whether to show a progress bar while the function runs. Default is FALSE.

Details

This function is used for reducing a random sample to data nugget centers in the create.DN function. NOTE THAT THIS FUNCTION IS NOT DESIGNED FOR USE OUTSIDE OF THE create.DN FUNCTION.

Value

A data frame containing the selected data nugget centers, with DN.num rows and the same number of columns as `RS`. The returned data frame has an attribute named "kept" containing the row indices of the observations in RS that were retained as data nugget centers.

Author(s)

Rituparna Dey, Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

`getDN.means` *Calculates the means of a given Data Nugget*

Description

This function calculates the datanugget means of a given datanugget object.

Usage

```
getDN.means(x,
            DN)
```

Arguments

`x`	A data matrix (of class matrix, data.frame, or data.table) containing only entries of class numeric.
`DN`	A datanugget object created using the create.DN or refined using the refine.DN function.

Details

create.DN function generates Data nuggets with centers based on different centering methods viz., mean, random and original. For any datanugget object created by create.DN or refined by refine.DN, this function computes the means of the datanuggets.

Value

`out`	nrow(DN$'Data Nuggets') by ncol(x) data frame containing the means of the data nuggets.

Author(s)

Rituparna Dey, Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
set.seed(55)
X = cbind.data.frame(rnorm(10^3),
                     rnorm(10^3),
                     rnorm(10^3))


my.DN = create.DN(x = X,
                  center.method = "original",
                  R = 500,
                  delete.percent = .1,
                  DN.num1 = 500,
                  DN.num2 = 250,
                  no.cores = 0,
                  make.pbs = FALSE)



my.DN.means = getDN.means(x = X,
                          DN = my.DN)
my.DN.means
```

`optimal.DN` *Optimal number of Data Nuggets*

Description

This function finds the optimal number data nuggets, and creates the nuggets using the `create.DN` function.

Usage

```
optimal.DN(x,
           center.method = "mean",
           dn.nos,
           R = 5000,
           delete.percent = .1,
           DN.num1 = 10^4,
           eps = 5e-3,
```

```
            dist.metric = "euclidean",
            seed = 291102,
            no.cores = (parallel::detectCores() - 1),
            make.pbs = FALSE)
```

Arguments

x	A data matrix (of class matrix, data.frame, or data.table) containing only entries of class numeric.
center.method	The method used for creating data nugget centers. Must be 'mean' or 'random' or 'original'. 'mean' chooses the data nugget center to be the mean of all observations within that data nugget, 'random' chooses the data nugget center to be some random observation within that data nugget, and 'original' chooses the original data nugget centers generated by the final run of datanugget creation using create.DNcenters function. Default is 'mean'.
dn.nos	The vector of candidate datanugget numbers. Must be a vector of length >= 3 and have numeric or integer entries.
R	The number of observations to sample from the data matrix when creating the initial data nugget centers. Must be of class numeric within [100,10000]. Default is 5000.
delete.percent	The proportion of observations to remove from the data matrix at each iteration when finding data nugget centers. Must be of class numeric and within (0,1). Default is 0.1.
DN.num1	The number of initial data nugget centers to create. Must be of class numeric. Default is 10^4.
eps	Stoppage tolerance on hitting the elbow. Default is 5e-3.
dist.metric	The distance metric used to create the initial centers of data nuggets. Must be 'euclidean' or 'manhattan'. Default is 'euclidean'.
seed	Random seed for replication. Must be of class numeric. Default is 291102.
no.cores	Number of cores used for parallel processing. If '0' then parallel processing is not used. Must be of class numeric.
make.pbs	Logical; whether to show a progress bar while the function runs. Default is FALSE.

Details

The optimal number data nuggets is data-driven based on the relative second-order differences of propensity score indices. That optimal number of data nuggets are created using the `create.DN` function.

Value

A list of 4 items:

opt.dn.no	Optimal data nugget number.
opt.dn	Final datanugget object based on optimal data nugget number.

`elbow.plot`	Elbow plot of Propensity Score Index vs Data Nugget number.
`diff.plot`	Relative second order differences vs Data Nugget number plot.

Author(s)

Rituparna Dey, Javier Cabrera

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
## small example
X = cbind.data.frame(rnorm(10^3),
                     rnorm(10^3),
                     rnorm(10^3))

suppressMessages({

  my.DN = optimal.DN(x = X,
                     dn.nos = seq(50, 500, by = 25),
                     R = 500,
                     delete.percent = .1,
                     DN.num1 = 500,
                     eps = 5e-5,
                     no.cores = 0,
                     make.pbs = FALSE)

})

my.DN$opt.dn.no
my.DN$opt.dn
my.DN$elbow.plot
my.DN$diff.plot

## Not run:

  ## large example
  X = cbind.data.frame(rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4))
  t1 <- Sys.time()
  my.DN = optimal.DN(x = X,
```

```
                        dn.nos = seq(1000, 10000, by = 1000),
                        R = 5000,
                        delete.percent = .1,
                        DN.num1 = 10^4,
                        eps = 5e-5,
                        no.cores = 2)
      t2 <- Sys.time()

      my.DN$opt.dn.no
      my.DN$opt.dn
      my.DN$elbow.plot
      my.DN$diff.plot


## End(Not run)
```

`refine.DN` *Refine Data Nuggets*

Description

This function refines the data nuggets found in an object of class datanugget created using the create.DN function or optimal.DN function.

Usage

```
refine.DN(x,
          DN,
          EV.tol = 0.9,
          max.splits = 5,
          min.nugget.size = 2,
          shape.split = FALSE,
          min.shape.size = 10,
          delta = 2,
          nstart = 25,
          seed = 291102,
          no.cores = (parallel::detectCores() - 1),
          make.pbs = FALSE)
```

Arguments

`x`	A data matrix (data frame, data table, matrix, etc.) containing only entries of class numeric.
`DN`	An object of class data nugget created using the create.DN or optimal.DN function.

EV.tol	A value designating the percentile for finding the corresponding quantile that will designate how large the largest eigenvalue of the covariance matrix of a data nugget can be before it must be split. Must be of class numeric and within (0,1). Default is 0.9.
max.splits	A value designating the maximum amount of attempts that will be made to split data nuggets according to their largest eigenvalue before the algorithm breaks. Must be of class numeric and non-negative. Default is 5.
min.nugget.size	A value designating the minimum amount of observations a data nugget created from a split must contain. Must be of class numeric and with be greater than 1. Default is 2.
shape.split	Logical; whether to refine the nuggets of elongated shape i.e., perform shape splitting. Default is FALSE.
min.shape.size	A value designating the minimum weight for a data nugget to be shape split. Must be at least 10. Default is 10.
delta	A value defining a data nugget to be elongated when its first two largest eigenvalues, EV1 and EV2 satisfy EV1/EV2 > delta. Must be gtreater than 1. Default is 2.
nstart	The number of random starts used by the kmeans algorithm when splitting. Default is 25.
seed	Random seed for replication. Must be of class numeric. Default is 291102.
no.cores	Number of cores used for parallel processing. If '0' then parallel processing is not used. Must be of class numeric.
make.pbs	Print progress bars? Must be TRUE or FALSE.

Details

Data nuggets can be refined by splitting the nuggets with unusually large within-data-nugget variability. For each nugget, the largest eigenvalue of its covariance matrix, EV.1 serves as the variability measure. EV.tol defines the percentile threshold and if a nugget's EV.1 exceeds that threshold, it is split into two by K-Means algorithm. However, if either of the two data nuggets created by this split have less than the designated minimum data nugget size (min.nugget.size), then the split is cancelled and the data nugget remains as is. This function refines data nuggets using Algorithm 2 provided in the reference.

Another version of refinement is shape based splitting. A data nugget is considered to be very elongated in shape, if the ratio of the first to the second largest eigenvalues of its covariance matrix, EV.1 / EV.2 is greater than a specified threshold, delta (default value is 2). Those data nuggets are split using the same K-Means clustering.

Value

An object of class datanugget:

Data Nuggets	DN.num by (ncol(x)+3) data frame containing the information for the data nuggets created (index, center, weight, scale).
Data Nugget Assignments	Vector of length nrow(x) containing the data nugget assignment of each observation in x.

Author(s)

Rituparna Dey, Traymon Beavers, Javier Cabrera, Mariusz Lubomirski

References

Beavers, T. E., Cheng, G., Duan, Y., Cabrera, J., Lubomirski, M., Amaratunga, D., & Teigler, J. E. (2024). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Journal of Computational and Graphical Statistics, 1-21.

Cherasia, K. E., Cabrera, J., Fernholz, L. T., & Fernholz, R. (2022). Data Nuggets in Supervised Learning. *In Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler* (pp. 429-449). Cham: Springer International Publishing.

Examples

```
## small example
X = cbind.data.frame(rnorm(10^3),
                     rnorm(10^3),
                     rnorm(10^3))

suppressMessages({

  my.DN = create.DN(x = X,
                    R = 500,
                    delete.percent = .1,
                    DN.num1 = 500,
                    DN.num2 = 250,
                    no.cores = 0,
                    make.pbs = FALSE)

  my.DN2 = refine.DN(x = X,
                     DN = my.DN,
                     no.cores = 0,
                     make.pbs = FALSE)

})

my.DN2$`Data Nuggets`
my.DN2$`Data Nugget Assignments`

## Not run:

  ## large example
  X = cbind.data.frame(rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4),
                       rnorm(5*10^4))

  my.DN = create.DN(x = X,
                    R = 5000,
                    delete.percent = .1,
```

```
                        DN.num1 = 10^4,
                        DN.num2 = 2000,
                        no.cores = 2)

      my.DN2 = refine.DN(x = X,
                         DN = my.DN,
                         no.cores = 2)

      my.DN2$`Data Nuggets`
      my.DN2$`Data Nugget Assignments`


## End(Not run)
```

Index

AC, 3

create.DN, 4
create.DNcenters, 7

datanugget-package, 2

getDN.means, 8

optimal.DN, 9

refine.DN, 12