

# Package ‘bigdist’

October 12, 2022

**Type** Package

**Title** Store Distance Matrices on Disk

**Version** 0.1.4

**Description** Provides utilities to compute, store and access distance matrices on disk as file-backed matrices provided by the 'bigstatsr' package. File-backed distance matrices are stored as a symmetric matrix to facilitate out-of-memory operations on file-backed matrix while the in-memory 'dist' object stores only the lower diagonal elements. 'disto' provides a unified interface to work with in-memory and disk-based distance matrices.

**URL** <https://github.com/talegari/bigdist>

**BugReports** <https://github.com/talegari/bigdist/issues>

**Imports** assertthat (>= 0.2.0), bigstatsr (>= 0.9.1), proxy (>= 0.4.21), furr (>= 0.1.0), utils,

**Depends** R (>= 3.4.0)

**Suggests** stats, spelling (>= 2.0), testthat (>= 2.0.1),

**License** GPL-3

**Encoding** UTF-8

**RoxygenNote** 6.1.0

**Language** en-US

**NeedsCompilation** no

**Author** Komala Sheshachala Srikanth [aut, cre],  
Florian Privé [ctb]

**Maintainer** Komala Sheshachala Srikanth <sri.teach@gmail.com>

**Repository** CRAN

**Date/Publication** 2019-03-16 14:13:30 UTC

## R topics documented:

as_bigdist . . . . .	2
bigdist . . . . .	3

bigdist_extract . . . . .	4
bigdist_replace . . . . .	5
bigdist_size . . . . .	6
bigdist_subset . . . . .	7
colStartIndex . . . . .	8
distIndex . . . . .	8
dist_ij_k . . . . .	9
dist_ij_k_ . . . . .	9
dist_k_ij . . . . .	10
dist_k_ij_ . . . . .	10
package_bigdist . . . . .	11

<b>Index</b>	<b>12</b>
--------------	-----------

---

as_bigdist	<i>Convert to bigdist</i>
------------	---------------------------

---

## Description

Generic to convert an object of class 'bigdist'

## Usage

```
as_bigdist(x, file, ...)
```

## Arguments

x	Object coercible to bigdist
file	File to write the 'bigdist' matrix
...	additional arguments

## Details

Writing distances to FBM can be parallelized by setting up a future backend

## Value

An object of class 'bigdist'

## Examples

```
temp3 <- as_bigdist(dist(mtcars), file = file.path(tempdir(), "temp_ex4"))
temp3
```

---

bigdist	<i>Read or Create a distance matrix on disk</i>
---------	---

---

### Description

Computes distances via `dist` and saves then as file-backed matrix(FBM) using **bigstatsr** package or connects existing FBM backup file on disk.

### Usage

```
bigdist(mat, file, method = "euclidean", type = "float")
```

### Arguments

mat	Numeric matrix. When missing, attempts to connect to existing backup file. See 'file' argument.
file	(string) Name of the backing file to be created or an existing backup file. Do not include trailing ".bk". See details for the backup file format.
method	(string or function) See method argument of <code>dist</code> . This ignored when mat is missing.
type	(string, default: 'float') Storage type of FBM. See <code>FBM</code> . This ignored when mat is missing.

### Details

`bigdist` class is a list where the key 'fbm' holds the FBM connection. The filename format is of the form `<someName>_<size>_<type>.bk` where size is the number of observations and type is the data type like 'double', 'float'.

**bigstatsr** package stores matrices on disk and allows efficient computation on them. The **disto** provides a unified frontend to read parts of distance matrices and apply functions over rows/columns. For efficient operations, write C++ functions to talk to **bigstatsr**'s `FBM`.

The distance computation and writing to FBM may be parallelized by setting a future backend

### Value

An object of class 'bigdist'.

### Examples

```
# basics of 'bigdist'
# create a random matrix
set.seed(1)
amat <- matrix(rnorm(1e3), ncol = 10)
td <- tempdir()

# create a bigdist object with FBM (file-backed matrix) on disk
temp <- bigdist(mat = amat, file = file.path(td, "temp_ex1"))
```

```

temp
temp$fbm$backingfile
temp$fbm[1, 2]

# connect to FBM on disk as a bigdist object
temp2 <- bigdist(file = file.path(td, "temp_ex1_100_float"))
temp2
temp2$fbm[1,2]

# check the size of bigdist object
bigdist_size(temp)

# bigdist accessors

# ij
bigdist_extract(temp, 1, 2)
bigdist_extract(temp, 1:2, 3:4)
bigdist_extract(temp, 1:2, 3:4, product = "inner")
dim(bigdist_extract(temp, 1:2,))
dim(bigdist_extract(temp, , 3:4))

# k (lower trianle indexing)
bigdist_extract(temp, k = 3:7)

# bigdist replacers

# ij
bigdist_replace(temp, 1, 2, 10)
bigdist_extract(temp, 1, 2)
bigdist_replace(temp, 1:2, 3:4, 11:12)
bigdist_extract(temp, 1:2, 3:4, product = "inner")

# k (lower trianle indexing)
bigdist_replace(temp, k = 3:7, value = 51:55)
bigdist_extract(temp, k = 3:7)

# subset a bigdist object
temp_subset <- bigdist_subset(temp, index = 21:30, file = file.path(td, "temp_ex2"))
temp_subset
temp_subset$fbm$backingfile

# convert a dist object(in memory) to a bigdist object
temp3 <- as_bigdist(dist(mtcars), file = file.path(td, "temp_ex3"))
temp3

```

---

bigdist\_extract

*Extract parts of bigdist*


---

## Description

Extract parts of bigdist

**Usage**

```
bigdist_extract(x, i, j, k, product = "outer")
```

**Arguments**

x	Object of class 'bigdist'
i	(integer vector) row positions
j	(integer vector) column positions
k	(integer vector) positions
product	(string) One among: 'inner', 'outer'(default)

**Details**

In k-mode, both i and j should be missing and k should not be missing. In ij-mode, k should be missing and both i and j are optional. If i or j are missing, they are interpreted as all values of i or j (similar to matrix or dataframe subsetting).

**Value**

A matrix or vector of distances when product is 'outer' and 'inner' respectively

**Examples**

```
set.seed(1)
amat <- matrix(rnorm(1e3), ncol = 10)
td <- tempdir()
temp <- bigdist(mat = amat, file = file.path(td, "temp_ex6"))
bigdist_extract(temp, 1, 2)
bigdist_extract(temp, 1:2, 3:4)
bigdist_extract(temp, 1:2, 3:4, product = "inner")
dim(bigdist_extract(temp, 1:2,))
dim(bigdist_extract(temp, , 3:4))
```

---

bigdist_replace	<i>Replace parts of bigdist</i>
-----------------	---------------------------------

---

**Description**

Replace parts of bigdist

**Usage**

```
bigdist_replace(x, i, j, value, k)
```

**Arguments**

x	Object of class 'bigdist'
i	(integer vector) row positions
j	(integer vector) column positions
value	(integer/numeric vector) Values to replace
k	(integer vector) positions

**Details**

There are two modes to specify the positions:

- ij-mode where i and j are specified and k is missing. If i or j are missing, they are interpreted as all values of i or j (similar to matrix or dataframe subsetting). Lengths of i, j are required to be same. If 'value' is singleton, then it is extended to the length of i or j. Else, 'value' should have same length as i or j.
- k-mode where k is present and both i and k are missing. k is the positions in the dist object. If 'value' is singleton, then it is extended to the length of k. Else, 'value' should have same length as k.

**Value**

bigdist object

**Examples**

```
set.seed(1)
amat <- matrix(rnorm(1e3), ncol = 10)
td <- tempdir()
temp <- bigdist(mat = amat, file = file.path(td, "temp_ex7"))
bigdist_replace(temp, 1, 2, 10)
bigdist_extract(temp, 1, 2)
bigdist_replace(temp, 1:2, 3:4, 11:12)
bigdist_extract(temp, 1:2, 3:4, product = "inner")
```

---

bigdist\_size

*Size of bigdist*

---

**Description**

Size of bigdist

**Usage**

```
bigdist_size(x)
```

**Arguments**

x                    Object of class 'bigdist'

**Examples**

```
set.seed(1)
amat <- matrix(rnorm(1e3), ncol = 10)
td <- tempdir()
temp <- bigdist(mat = amat, file = file.path(td, "temp_ex5"))
bigdist_size(temp)
```

---

bigdist_subset	<i>Subset parts of bigdist</i>
----------------	--------------------------------

---

**Description**

Subset parts of bigdist

**Usage**

```
bigdist_subset(x, index, file)
```

**Arguments**

x                    Object of class 'bigdist'

index                Indexes of the bigdist to be subset as a bigdist

file                 (string) Name of the backing file to be created. Do not include trailing ".bk". See details for the backup file format.

**Details**

The filename format is of the form <some name>\_<size>\_<type>.bk where size is the number of observations and type is the data type like 'double', 'float'.

**Examples**

```
set.seed(1)
amat <- matrix(rnorm(1e3), ncol = 10)
td <- tempdir()
temp <- bigdist(mat = amat, file = file.path(td, "temp_ex8"))
temp_subset <- bigdist_subset(temp, index = 21:30, file = file.path(td, "temp_ex9"))
temp_subset
temp_subset$fbm$backingfile
```

---

colStartIndex	<i>Get the column start index</i>
---------------	-----------------------------------

---

**Description**

Get the index of the dist object corresponding to some column start in the symmetric form of the distance matrix

**Usage**

```
colStartIndex(i, size)
```

**Arguments**

i	Column number
size	Size of the dist object

**Value**

An index of dist object

**Examples**

```
colStartIndex(2, 10)
```

---

distIndex	<i>Compute distance between a row and its subsequent rows</i>
-----------	---

---

**Description**

Computes distance between row  $i$  and rows  $i + 1, i + 2, \dots, n$  where  $n$  is the total number of rows

**Usage**

```
distIndex(i, mat, method, size)
```

**Arguments**

i	row index
mat	numeric matrix
method	method is passed to <a href="#">dist</a>
size	Number of rows of mat

**Value**

(numeric vector) Vector of  $n - i$  distances where  $n$  is the total number of rows



---

dist_ij_k	<i>Vectorized version of dist_ij_k_</i>
-----------	---

---

**Description**

Convert ij indexes to k indexes for a dist object

**Usage**

```
dist_ij_k(i, j, size)
```

**Arguments**

i	row indexes
j	column indexes
size	value of size attribute of the dist object

**Value**

k indexes

**Examples**

```
dist_ij_k(1:3, 4:6, 10)
```

---

dist_ij_k_	<i>Convert ij index to k index</i>
------------	------------------------------------

---

**Description**

Convert ij index to k index for a dist object

**Usage**

```
dist_ij_k_(i, j, size)
```

**Arguments**

i	row index
j	column index
size	value of size attribute of the dist object

**Value**

k index

**Examples**

```
dist_ij_k_(1, 3, 10)
```

---

dist_k_ij	<i>Vectorized version of dist_k_ij_</i>
-----------	---

---

**Description**

Convert kth indexes to ij indexes of a dist object

**Usage**

```
dist_k_ij(k, size)
```

**Arguments**

k	kth indexes
size	value of size attribute of the dist object

**Value**

ij indexes as 2\*n matrix where n is length of k vector

**Examples**

```
dist_k_ij(4:6, 10:12)
```

---

dist_k_ij_	<i>Convert kth index to ij index</i>
------------	--------------------------------------

---

**Description**

Convert kth index to ij index of a dist object

**Usage**

```
dist_k_ij_(k, size)
```

**Arguments**

k	kth index
size	value of size attribute of the dist object

**Value**

ij index as a length two integer vector

**Examples**

```
dist_k_ij_(4, 10)
```

---

package_bigdist	<i>Package: bigdist</i>
-----------------	-------------------------

---

**Description**

bigdist package facilitates storing distance matrices on disk as 'file-backed matrix' (FBM) using **bigstatsr** package. The FBM stores a symmetric matrix. Each distance is stored as a 'float/double' approximately requiring 4 or 8 bytes. The resulting file size will approximately be  $8 * \text{size}^2$  where size is nrow/ncol of the data matrix. Working with bigdist package requires working knowledge of **bigstatsr** package.

- bigdist: (function) Create or connect to distance matrix on disk.
- as\_bigdist: (method) Convert to 'bigdist' class.

**disto** package provides a unified interface to distance matrices in-memory (class 'dist') or on disk (class 'bigdist').

**Author(s)**

**Maintainer:** Komala Sheshachala Srikanth <sri.teach@gmail.com>

Other contributors:

- Florian Privé <florian.prive.21@gmail.com> [contributor]

**See Also**

Useful links:

- <https://github.com/talegari/bigdist>
- Report bugs at <https://github.com/talegari/bigdist/issues>

# Index

`_PACKAGE (package_bigdist)`, 11

`as_bigdist`, 2

`bigdist`, 3

`bigdist_extract`, 4

`bigdist_replace`, 5

`bigdist_size`, 6

`bigdist_subset`, 7

`colStartIndex`, 8

`dist`, 3, 8

`dist_ij_k`, 9

`dist_ij_k_`, 9

`dist_k_ij`, 10

`dist_k_ij_`, 10

`distIndex`, 8

`FBM`, 3

`package_bigdist`, 11

`package_bigdist-package`  
`(package_bigdist)`, 11