Package ‘harmony’
November 14, 2022

Title  Fast, Sensitive, and Accurate Integration of Single Cell Data

Version  0.1.1

Description  Implementation of the Harmony algorithm for single cell integration, described in Kor-
sunsky et al <doi:10.1038/s41592-019-0619-0>. Package includes a standalone Harmony func-
tion and interfaces to external frameworks.

URL  software.broadinstitute.org/harmony

License  GPL-3

Encoding  UTF-8

RoxygenNote  7.2.1

Depends  R(>= 3.4.0), Rcpp

LazyData  true

LinkingTo  Rcpp, RcppArmadillo, RcppProgress

Imports  dplyr, cowplot, tidyr, ggplot2, irlba, Matrix, methods,
tibble, rlang

Suggests  SingleCellExperiment, Seurat (>= 4.1.1), testthat, knitr,
markdown

VignetteBuilder  knitr

NeedsCompilation  yes

Author  Ilya Korsunsky [cre, aut] (<https://orcid.org/0000-0003-4848-3948>),
Nghia Millard [aut] (<https://orcid.org/0000-0002-0518-7674>),
Jean Fan [aut, ctb] (<https://orcid.org/0000-0002-0212-5451>),
Kamil Slowikowski [aut, ctb] (<https://orcid.org/0000-0002-2843-6370>),
Miles Smith [ctb],
Soumya Raychaudhuri [aut] (<https://orcid.org/0000-0002-1901-8265>)

Maintainer  Ilya Korsunsky <ilya.korsunsky@gmail.com>

Repository  CRAN

Date/Publication  2022-11-14 09:20:08 UTC


### R topics documented:

- `cell_lines` .................................................. 2
- `cell_lines_small` ........................................... 2
- `harmony` ....................................................... 3
- `HarmonyMatrix` ............................................. 3
- `moe_ridge_get_betas` ...................................... 5
- `RunHarmony` ................................................. 6

#### Index

<table>
<thead>
<tr>
<th>cell_lines</th>
<th>List of metadata table and scaled PCs matrix</th>
</tr>
</thead>
</table>

**Description**

List of metadata table and scaled PCs matrix

**Usage**

`cell_lines`

**Format**

- meta_data: data.table of 9478 rows with defining dataset and cell_type
- scaled_pcs: data.table of 9478 rows (cells) and 20 columns (PCs)

**Source**

[https://www.10xgenomics.com](https://www.10xgenomics.com)

---

<table>
<thead>
<tr>
<th>cell_lines_small</th>
<th>Same as <code>cell_lines</code> but smaller (300 cells).</th>
</tr>
</thead>
</table>

**Description**

Same as `cell_lines` but smaller (300 cells).

**Usage**

`cell_lines_small`

**Format**

An object of class list of length 2.

**Source**

[https://www.10xgenomics.com](https://www.10xgenomics.com)
Harmony: fast, accurate, and robust single cell integration.

**Description**

Algorithm for single cell integration.

**Usage**

1. `?HarmonyMatrix` to run Harmony on gene expression or PCA embeddings matrix.
2. `?RunHarmony` to run Harmony on Seurat or SingleCellExperiment objects.

**Useful links**

2. Read the manuscript online.

---

**HarmonyMatrix**

*Main Harmony interface*

**Description**

Use this to run the Harmony algorithm on gene expression or PCA matrix.

**Usage**

```r
HarmonyMatrix(
  data_mat,  
  meta_data,  
  vars_use,  
  do_pca = TRUE,  
 npcs = 20,
  theta = NULL,
  lambda = NULL,
  sigma = 0.1,
  nclust = NULL,
  tau = 0,
  block.size = 0.05,
  max.iter.harmony = 10,
  max.iter.cluster = 200,
  epsilon.cluster = 1e-05,
  epsilon.harmony = 1e-04,
  plot_convergence = FALSE,
  return_object = FALSE,
  verbose = TRUE,
)```

reference_values = NULL,
cluster_prior = NULL
)

Arguments

data_mat  Matrix of normalized gene expression (default) or PCA embeddings (see do_pca). Cells can be rows or columns.

meta_data  Either (1) Dataframe with variables to integrate or (2) vector with labels.

vars_use  If meta_data is dataframe, this defined which variable(s) to remove (character vector).

do_pca  Whether to perform PCA on input matrix.

npcs  If doing PCA on input matrix, number of PCs to compute.

theta  Diversity clustering penalty parameter. Specify for each variable in vars_use. Default theta=2. theta=0 does not encourage any diversity. Larger values of theta result in more diverse clusters.

lambda  Ridge regression penalty parameter. Specify for each variable in vars_use. Default lambda=1. Lambda must be strictly positive. Smaller values result in more aggressive correction.

sigma  Width of soft kmeans clusters. Default sigma=0.1. Sigma scales the distance from a cell to cluster centroids. Larger values of sigma result in cells assigned to more clusters. Smaller values of sigma make soft kmeans cluster approach hard clustering.

nclust  Number of clusters in model. nclust=1 equivalent to simple linear regression.

tau  Protection against overclustering small datasets with large ones. tau is the expected number of cells per cluster.

block.size  What proportion of cells to update during clustering. Between 0 to 1, default 0.05. Larger values may be faster but less accurate.

max.iter.harmony  Maximum number of rounds to run Harmony. One round of Harmony involves one clustering and one correction step.

max.iter.cluster  Maximum number of rounds to run clustering at each round of Harmony.

epsilon.cluster  Convergence tolerance for clustering round of Harmony. Set to -Inf to never stop early.

epsilon.harmony  Convergence tolerance for Harmony. Set to -Inf to never stop early.

plot_convergence  Whether to print the convergence plot of the clustering objective function. TRUE to plot, FALSE to suppress. This can be useful for debugging.

return_object  (Advanced Usage) Whether to return the Harmony object or only the corrected PCA embeddings.

verbose  Whether to print progress messages. TRUE to print, FALSE to suppress.
moe_ridge_get_betas

-reference_values
   (Advanced Usage) Defines reference dataset(s). Cells that have batch variables values matching reference_values will not be moved.

-cluster_prior
   (Advanced Usage) Provides user defined clusters for cluster initialization. If the number of provided clusters C is less than K, Harmony will initialize K-C clusters with kmeans. C cannot exceed K.

Value

By default, matrix with corrected PCA embeddings. If return_object is TRUE, returns the full Harmony object (R6 reference class type).

Examples

```r
## By default, Harmony inputs a normalized gene expression matrix
## Not run:
harmony_embeddings <- HarmonyMatrix(exprs_matrix, meta_data, 'dataset')
## End(Not run)

## Harmony can also take a PCA embeddings matrix
data(cell_lines_small)
pca_matrix <- cell_lines_small$scaled_pcs
meta_data <- cell_lines_small$meta_data
harmony_embeddings <- HarmonyMatrix(pca_matrix, meta_data, 'dataset',
   do_pca=FALSE)

## Output is a matrix of corrected PC embeddings
dim(harmony_embeddings)
harmony_embeddings[seq_len(5), seq_len(5)]

## Finally, we can return an object with all the underlying data structures
harmony_object <- HarmonyMatrix(pca_matrix, meta_data, 'dataset',
   do_pca=FALSE, return_object=TRUE)
dim(harmony_object$Y) ## cluster centroids
dim(harmony_object$R) ## soft cluster assignment
dim(harmony_object$Z_corr) ## corrected PCA embeddings
head(harmony_object$O) ## batch by cluster co-occurence matrix
```

moe_ridge_get_betas Get beta Utility

Description

Utility function to get ridge regression coefficients from trained Harmony object
RunHarmony

Usage

moe_ridge_get_betas(harmonyObj)

Arguments

harmonyObj  Trained harmony object. Get this by running HarmonyMatrix function with return_object=TRUE.

Value

Returns nothing, modifies object in place.

RunHarmony

Harmony single cell integration

Description

Run Harmony algorithm with Seurat and SingleCellAnalysis pipelines.

Usage

RunHarmony(object, group.by.vars, ...)

## S3 method for class 'Seurat'
RunHarmony(
  object,
  group.by.vars,
  reduction = "pca",
  dims.use = NULL,
  theta = NULL,
  lambda = NULL,
  sigma = 0.1,
  nclust = NULL,
  tau = 0,
  block.size = 0.05,
  max.iter.harmony = 10,
  max.iter.cluster = 20,
  epsilon.cluster = 1e-05,
  epsilon.harmony = 1e-04,
  plot_convergence = FALSE,
  verbose = TRUE,
  reference_values = NULL,
  reduction.save = "harmony",
  assay.use = NULL,
  project.dim = TRUE,
  ...
)
## S3 method for class 'SingleCellExperiment'
RunHarmony(
  object,
  group.by.vars,
  dims.use = NULL,
  theta = NULL,
  lambda = NULL,
  sigma = 0.1,
  nclust = NULL,
  tau = 0,
  block.size = 0.05,
  max.iter.harmony = 10,
  max.iter.cluster = 20,
  epsilon.cluster = 1e-05,
  epsilon.harmony = 1e-04,
  plot_convergence = FALSE,
  verbose = TRUE,
  reference_values = NULL,
  reduction.save = "HARMONY",
  ...
)

### Arguments

- **object**: Pipeline object. Must have PCA computed.
- **group.by.vars**: Which variable(s) to remove (character vector).
- **dims.use**: Which PCA dimensions to use for Harmony. By default, use all.
- **theta**: Diversity clustering penalty parameter. Specify for each variable in group.by.vars. Default theta=2. theta=0 does not encourage any diversity. Larger values of theta result in more diverse clusters.
- **lambda**: Ridge regression penalty parameter. Specify for each variable in group.by.vars. Default lambda=1. Lambda must be strictly positive. Smaller values result in more aggressive correction.
- **sigma**: Width of soft kmeans clusters. Default sigma=0.1. Sigma scales the distance from a cell to cluster centroids. Larger values of sigma result in cells assigned to more clusters. Smaller values of sigma make soft kmeans cluster approach hard clustering.
- **nclust**: Number of clusters in model. nclust=1 equivalent to simple linear regression.
- **tau**: Protection against overclustering small datasets with large ones. tau is the expected number of cells per cluster.
- **block.size**: What proportion of cells to update during clustering. Between 0 to 1, default 0.05. Larger values may be faster but less accurate.
max.iter.harmony
Maximum number of rounds to run Harmony. One round of Harmony involves one clustering and one correction step.

max.iter.cluster
Maximum number of rounds to run clustering at each round of Harmony.

epsilon.cluster
Convergence tolerance for clustering round of Harmony. Set to -Inf to never stop early.

epsilon.harmony
Convergence tolerance for Harmony. Set to -Inf to never stop early.

plot_convergence
Whether to print the convergence plot of the clustering objective function. TRUE to plot, FALSE to suppress. This can be useful for debugging.

verbose
Whether to print progress messages. TRUE to print, FALSE to suppress.

reference_values
(Advanced Usage) Defines reference dataset(s). Cells that have batch variables values matching reference_values will not be moved.

reduction.save
Keyword to save Harmony reduction. Useful if you want to try Harmony with multiple parameters and save them as e.g. 'harmony_theta0', 'harmony_theta1', 'harmony_theta2'

assay.use
(Seurat V3 only) Which assay to run PCA on if no PCA present?

project.dim
Project dimension reduction loadings. Default TRUE.

Value
Seurat (version 3) object. Harmony dimensions placed into dimensional reduction object harmony. For downstream Seurat analyses, use reduction='harmony'.
SingleCellExperiment object. After running RunHarmony, the corrected cell embeddings can be accessed with reducedDim(object, "Harmony").
Index

* datasets
  - cell_lines, 2
  - cell_lines_small, 2

  cell_lines, 2
  cell_lines_small, 2

  harmony, 3
  HarmonyMatrix, 3

  moe_ridge_get_betas, 5

  RunHarmony, 6