Package: dreval (via r-universe)

September 12, 2024

Type Package

Title Evaluate Reduced Dimension Representations

Version 0.1.5

Description Evaluate and compare multiple reduced dimension representations, based on how well they retain structure from the original data set.

License MIT + file LICENSE

Encoding UTF-8

Depends R (>= 3.6)

Imports SingleCellExperiment, stats, SummarizedExperiment, dplyr, wordspace, cluster, coRanking (>= 0.2.1), methods, ggplot2, grDevices, tidyr, magrittr

RoxygenNote 7.3.0

Suggests knitr, rmarkdown, testthat (>= 2.1.0), TENxPBMCData, cowplot, hexbin

VignetteBuilder knitr

URL https://github.com/csoneson/dreval

BugReports https://github.com/csoneson/dreval/issues

biocViews DimensionReduction, PrincipalComponent, Visualization

Repository https://csoneson.r-universe.dev

RemoteUrl https://github.com/csoneson/dreval

RemoteRef HEAD

RemoteSha 5c2cb03c24bf21919e7a5c7e2a07e769d377c3c5

Contents

`dreval` *Evaluate structure preservation in reduced dimension representations*

Description

Calculates a collection of metrics comparing one or more reduced dimension representations to a reference representation. The function takes a `SingleCellExperiment` object as input. The reference representation can be either one of the included assays or one of the reduced dimension representations. If an assay is used, reference distances can be calculated based on all or a subset of the features (rows). These distances are then compared to distances calculated from the specified reduced dimension representations, and several scores are returned. The execution time of the function depends strongly on both the number of retained variables (which affects the distance calculation in the reference space) and the number of samples that are randomly selected to use as the basis for the comparison. Since subsampling of the columns (via the `nSamples` argument) is random, setting the random seed is recommended to obtain reproducible results.

Usage

```
dreval(
  sce,
  dimReds = NULL,
  refType = "assay",
  refAssay = "logcounts",
  refDimRed = NULL,
  features = NULL,
  nSamples = NULL,
  distNorm = "none",
  refDistMethod = "euclidean",
  kTM = c(10, 100),
  labelColumn = NULL,
  verbose = FALSE
)
```

Arguments

`sce`	A `SingleCellExperiment` object.
`dimReds`	A character vector with the names of the reduced dimension representations from `sce` to include in the evaluation. If `NULL`, all reduced dimension representations are included.
`refType`	A character scalar, either "assay" or "dimred", specifying whether to use an assay or a reduced dimension representation of `sce` as the reference data source.
`refAssay`	A character scalar giving the name of the assay from `sce` to use as the basis for the distance calculations in the reference space, if `refType` if `"assay"`.
`refDimRed`	A character scalar specifying the reduced dimension representation to use as the reference data representation if `refType` is `"dimred"`.

features	A character vector giving the IDs of the features to use for distance calculations from the chosen assay. Will be matched to the row names of sce.
nSamples	A numeric scalar, giving the number of columns to subsample (randomly) from sce.
distNorm	A character scalar, indicating how the distance vectors in the reference and low-dimensional spaces should be normalized before they are compared. If set to "l2", the vectors are L2 normalized, if set to "median" they are divided by the median value times the square root of their length, and if set to any other value they are divided by the square root of their length, to avoid metrics scaling with the number of retained samples.
refDistMethod	A character scalar defining the distance measure to use in the reference space. Must be one of "euclidean", "manhattan", "maximum", "canberra" or "cosine". The distance in the low-dimensional representation will always be Euclidean.
kTM	An integer vector giving the number of neighbors to use for trustworthiness, continuity and Jaccard index calculations.
labelColumn	A character scalar defining a column of colData(sce) to use as the group assignments in the silhouette width calculations. If not provided, the silhouette widths are not calculated.
verbose	A logical scalar, indicating whether to print out progress messages.

Details

The following metrics are calculated:

- SpearmanCorrDist - The Spearman correlation between the reference distances and the Euclidean distances in the low-dimensional representation. Ranges from -1 to 1, higher values are better.
- PearsonCorrDist - The Pearson correlation between the reference distances and the Euclidean distances in the low-dimensional representation. Ranges from -1 to 1, higher values are better.
- KSstatDist - The Kolmogorov-Smirnov statistic comparing the distribution of distances in the reference space and in the low-dimensional representation. Ranges from 0 to 1, lower values are better.
- EuclDistBetweenDists - The Euclidean distance between the vector of distances in the reference space and those in the low-dimensional representation. Depending on the value of distNorm, distances are scaled before they are compared. Lower values are better.
- SammonStress - The Sammon stress (Sammon 1969). Depending on the value of distNorm, distances are scaled before they are compared. Lower values are better.
- Trustworthiness_kNN - The trustworthiness score (Venna & Kaski 2001), using NN nearest neighbors. The trustworthiness indicates to which degree we can trust that the points placed closest to a given sample in the low-dimensional representation are really close to the sample also in the reference space. Ranges from 0 to 1, higher values are better.
- Continuity_kNN - The continuity score (Venna & Kaski 2001), using NN nearest neighbors. The continuity indicates to which degree we can trust that the points closest to a given sample in the reference space are placed close to the sample also in the low-dimensional representation. Ranges from 0 to 1, higher values are better.

- MeanJaccard_kNN - The mean Jaccard index (over all samples), comparing the set of NN nearest neighbors in the reference space and those in the low-dimensional representation. Ranges from 0 to 1, higher values are better.
- MeanSilhouette_X - If a `labelColumn` X is supplied, the mean silhouette score (Rousseeuw 1987) across all samples, with the grouping given by this column and the distances obtained from the low-dimensional representation. Ranges from -1 to 1, higher values are better.
- coRankingQlocal - Q_local, defined as the average LCMC over the values to the left of the maximum, following the dimRed/coRanking package implementations (Kraemer et al 2018, Lee and Verleysen 2009, Chen and Buja 2009). Measures the preservation of local distances, higher values are better.
- coRankingQglobal - Q_global, defined as the average LCMC over the values to the right of the maximum, following the dimRed/coRanking package implementations (Kraemer et al 2018, Lee and Verleysen 2009, Chen and Buja 2009). Measures the preservation of global distances, higher values are better.

Value

A list with two elements:

- scores - A `data.frame` with values of all evaluation metrics, across the dimension reduction methods. In addition to the metrics, it contains the dimensionality of the respective reduced dimension representations, and the value of K giving the highest value of LCMC (used for the calculations of Qlocal and Qglobal, see Kraemer et al 2018, Lee and Verleysen 2009, Chen and Buja 2009).
- plots - A list of ggplot objects, representing diagnostic plots.

Author(s)

Charlotte Soneson

References

Venna J., Kaski S. (2001). Neighborhood preservation in nonlinear projection methods: An experimental study. In Dorffner G., Bischof H., Hornik K., editors, Proceedings of ICANN 2001, pp 485–491. Springer, Berlin.

Lee J.A., Verleysen M. (2009). Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 72 (7-9):1431-1443.

Chen L., Buja A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. Journal of the American Statistical Association 104:209-219.

Kraemer G., Reichstein M., Mahecha M.D. (2018). dimRed and coRanking - Unifying dimensionality reduction in R. The R Journal 10 (1):342-358.

Sammon J.W. Jr (1969). A nonlinear mapping for data structure analysis. IEEE Transactions on Computers C18(5):401-409.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53-65.

Examples

```
data(pbmc3ksub)
dre <- dreval(sce = pbmc3ksub, nSamples = 150)
```

`dreval-pkg` *dreval*

Description

dreval evaluates and compares multiple reduced dimension representations, based on how well they retain structure from the original data set.

`pbmc3ksub` *Expression profiles for 2,700 PBMCs*

Description

This data set contains expression profiles for 2,700 PBMCs. The original data set was obtained from the TENxPBMCData package (pbmc3k data set). This data set was subset to around 1,800 highly expressed genes, normalized using the scran package, and several dimensionality reduction methods were applied.

Author(s)

Charlotte Soneson

`plotRankSummary` *Plot summary of the dimension reduction method ranking across metrics*

Description

For each metric, rank the evaluated reduced dimension representations by performance, and plot a summary of the overall ranking. Metrics evaluating local and global structure preservations are colored in red and blue, respectively.

Usage

```
plotRankSummary(
  dreSummary,
  metrics = NULL,
  sortBars = "decreasing",
  scoreType = "rank",
  tiesMethod = "average"
)
```

Arguments

dreSummary	A data.frame with the values of the evaluation metrics, typically the "scores" element of the output of dreval().
metrics	A character vector with the metrics to include in the summary. Must be a subset of the column names of dreSummary. If NULL, all metrics will be used. It can also be "global" or "local", in which case all the global or local metrics, respectively, will be used.
sortBars	A character scalar indicating whether/how to sort the bars in the output. Either "decreasing", "increasing" or "none" (in which case the input order will be used).
scoreType	A character scalar indicating what type of values to show in the plot. Either "rank" or "rescale". If set to "rank", the representations will be ranked for each metric (with the best one assigned the highest rank). If set to "rescale", the scores for each metric will first, if necessary, be inverted so that a high (positive) value corresponds to better performance, and then be linearly rescaled, mapping the lowest score to 1 and the highest to P, where P is the number of evaluated representations. If the original scores are approximately equally spaced between the highest and lowest observed values, this gives similar results as setting scoreType to "rank". However, if some of the scores are very similar to each other, the "rescale" approach allows them to get a similar rank score rather than forcing a uniform difference between successive scores.
tiesMethod	A character scalar indicating how ties are handled if scoreType is "rank". Should be one of the values accepted by base::rank ("average", "first", "last", "random", "max", "min").

Value

Nothing is returned, but a plot is generated.

Author(s)

Charlotte Soneson

Examples

```
data(pbmc3ksub)
dre <- dreval(sce = pbmc3ksub, nSamples = 150)
plotRankSummary(dre$scores)
```

Index