Package 'slideimp'

May 9, 2026

Type Package

Title Numeric Matrices K-NN and PCA Imputation

Version 1.1.0

Description Fast k-nearest neighbors (K-NN) and principal component analysis (PCA) imputation algorithms for missing values in high-dimensional numeric matrices, i.e., epigenetic data. For extremely high-dimensional data with ordered features, a sliding window approach for K-NN or PCA imputation is provided. Additional features include group-wise imputation (e.g., by chromosome), hyperparameter tuning with repeated cross-validation, multi-core parallelization, and optional subset imputation. The K-NN algorithm is described in: Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. and Botstein, D. (1999) ``Imputing Missing Data for Gene Expression Arrays". The PCA imputation is an optimized version of the imputePCA() function from the 'missMDA' package described in: Josse, J. and Husson, F. (2016) <doi:10.18637/jss.v070.i01> ``missMDA: A Package for Handling Missing Values in Multivariate Data Analysis".

License GPL (>= 2)

URL https://github.com/hhp94/slideimp, https://hhp94.github.io/slideimp/

BugReports https://github.com/hhp94/slideimp/issues

Depends R (>= 4.3.0)

Imports bigmemory, carrier, checkmate, cli, collapse, mirai, Rcpp, stats

Suggests knitr, missMDA, RhpcBLASctl, rmarkdown, testthat (>= 3.0.0)

LinkingTo mlpack, Rcpp, RcppArmadillo, RcppEnsmallen, RcppThread

VignetteBuilder knitr

Config/Needs/website rmarkdown

Config/testthat/edition 3

Encoding UTF-8

Language en-US

RoxygenNote 7.3.3

NeedsCompilation yes

Author Hung Pham [aut, cre, cph] (ORCID:
<https://orcid.org/0000-0002-8271-9355>)

Maintainer Hung Pham <amser.hoanghung@gmail.com>

Repository CRAN

Date/Publication 2026-05-01 11:00:02 UTC

Contents

`col_vars` *Calculate Matrix Column Variances*

Description

Compute the sample variance for each column of a numeric matrix.

Usage

```
col_vars(obj, cores = 1)
```

Arguments

`obj`	A numeric matrix.
`cores`	Integer. Number of cores to use for parallel computation. Defaults to 1.

Details

Columns with fewer than two non-missing values are assigned NA.

Value

A numeric vector of column variances, named if obj has column names.

Examples

```
set.seed(123)
obj <- matrix(rnorm(5 * 10), ncol = 5)
obj[1, 1] <- NA
obj[1:8, 2] <- NA
obj[1:8, 3] <- NA
obj[9, 3] <- obj[10, 3]
obj[1:9, 4] <- NA
obj[, 5] <- NA
obj

col_vars(obj)
apply(obj, 2, var, na.rm = TRUE)
```

---

compute_metrics *Compute Prediction Accuracy Metrics*

Description

Compute prediction accuracy metrics for results from tune_imp().

Usage

```
compute_metrics(results, metrics = c("mae", "rmse"))

## S3 method for class 'data.frame'
compute_metrics(results, metrics = c("mae", "rmse"))

## S3 method for class 'slideimp_tune'
compute_metrics(results, metrics = c("mae", "rmse"))
```

Arguments

results	A slideimp_tune data frame from tune_imp(). Must contain a result list-column whose elements are data frames with truth and estimate columns.
metrics	Character vector of metric names to compute. Defaults to c("mae", "rmse"). Available metrics are "mae", "rmse", "mape", "bias", "rsq", and "rsq_trad".

Value

A data frame containing the original parameter columns along with unnested metric columns: `.metric`, `.estimator`, and `.estimate`.

Examples

```
set.seed(1234)
obj <- sim_mat(20, 30)$input

results <- tune_imp(
  obj = obj,
  parameters = data.frame(k = 5),
  .f = "knn_imp",
  n_reps = 1,
  num_na = 10,
  .progress = FALSE
)

compute_metrics(results)
```

---

group_imp *Grouped K-NN or PCA Imputation*

Description

Perform K-NN or PCA imputation independently within feature groups.

Usage

```
group_imp(
  obj,
  group,
  subset = NULL,
  allow_unmapped = FALSE,
  k = NULL,
  ncp = NULL,
  method = NULL,
  cores = 1,
  .progress = TRUE,
  min_group_size = NULL,
  colmax = NULL,
  post_imp = NULL,
  dist_pow = NULL,
  tree = NULL,
  scale = NULL,
  coeff.ridge = NULL,
  threshold = NULL,
```

```
  row.w = NULL,
  seed = NULL,
  nb.init = NULL,
  maxiter = NULL,
  miniter = NULL,
  solver = NULL,
  lobpcg_control = NULL,
  clamp = NULL,
  pin_blas = FALSE,
  na_check = TRUE,
  on_infeasible = c("error", "skip", "mean")
)
```

Arguments

obj	A numeric matrix with samples in rows and features in columns.
group	Specification of how features should be grouped for imputation. Accepted formats are: • A character scalar naming a supported Illumina platform; see Note. • A long-format data.frame with columns group and feature. • A list-column data.frame with a feature list-column. Optional list-columns are aux, for auxiliary feature names, and parameters, for group-specific parameter lists.
subset	Optional character vector of feature names to impute. If NULL, all grouped features are imputed. Features in a group but not in subset are demoted to auxiliary columns for that group. Groups left with zero features after demotion are dropped with a message.
allow_unmapped	Logical. If FALSE, every column in colnames(obj) must appear in group. If TRUE, columns with no group assignment are left untouched and are not used as auxiliary columns.
k	Integer or NULL. Number of nearest neighbors for K-NN imputation. If NULL, k may be supplied through group$parameters.
ncp	Integer or NULL. Number of components for PCA imputation. If NULL, ncp may be supplied through group$parameters.
method	Character or NULL. For K-NN imputation, one of "euclidean" or "manhattan". For PCA imputation, one of "regularized" or "EM". If NULL, the corresponding backend default is used unless overridden by group$parameters.
cores	Integer. Number of cores for K-NN imputation only. For PCA imputation, use mirai::daemons() to parallelize across groups.
.progress	Logical. If TRUE, show progress.
min_group_size	Integer or NULL. Minimum total number of columns per group, counting both features and auxiliary columns. Groups smaller than this are padded with randomly sampled columns from obj.
colmax	Numeric scalar between 0 and 1. Columns with a missing-data proportion greater than colmax are excluded from the main imputation method. Excluded

columns are left unchanged unless post_imp = TRUE, in which case remaining missing values are replaced by column means when possible.

post_imp — Logical. If TRUE, replace missing values remaining after the main imputation method with column means when possible.

dist_pow — Numeric. Power used to penalize more distant neighbors in the weighted average. dist_pow = 0 gives an unweighted average of the nearest neighbors.

tree — Logical. If FALSE, use brute-force K-NN. If TRUE, use ball-tree K-NN via mlpack.

scale — Logical. If TRUE, columns are scaled to unit variance.

coeff.ridge — Numeric. Ridge regularization coefficient. Only used when method = "regularized". Values less than 1 regularize less, moving closer to EM PCA. Values greater than 1 regularize more, moving closer to mean imputation.

threshold — Numeric. Convergence threshold.

row.w — Row weights, internally normalized to sum to 1. Can be:

- NULL: all rows are weighted equally.
- A numeric vector of positive weights with length nrow(obj).
- "n_miss": rows with more missing values receive lower weight.

seed — Integer, numeric, or NULL. Random seed for reproducibility.

nb.init — Integer. Number of random initializations. The first initialization is always mean imputation.

maxiter — Integer. Maximum number of iterations.

miniter — Integer. Minimum number of iterations.

solver — Character. Eigensolver selection. One of "auto", "exact", or "lobpcg". "exact" uses the exact solver. "lobpcg" uses the iterative LOBPCG solver with exact fallback. "auto" performs a short timed probe and chooses LOBPCG only if it is clearly faster than the exact solver. When nb.init > 1, the auto choice from the first PCA initialization is reused for subsequent PCA initializations.

lobpcg_control — A list of LOBPCG eigensolver control options, usually created by lobpcg_control(). A plain named list is also accepted. Ignored when solver = "exact".

clamp — Optional numeric vector of length 2 giving lower and upper bounds for PCA-imputed values. Use NULL for no clamping. Use c(0, 1) for DNA methylation beta values. Use c(lb, Inf) for only lower bound clamping, or c(-Inf, ub) for only upper bound clamping. Clamping is applied only to values imputed by the PCA step, not to observed values.

pin_blas — Logical. If TRUE, pin BLAS threads to 1 to reduce contention when using parallel PCA on systems linked with multithreaded BLAS.

na_check — Logical. If TRUE, check whether the returned matrix still contains missing values.

on_infeasible — Character. One of "error", "skip", or "mean". Controls behavior when a group is infeasible for imputation, for example when k or ncp exceeds the number of usable columns after applying colmax.

Details

group_imp() performs K-NN or PCA imputation on feature groups independently, which can substantially reduce runtime for large matrices.

Specify k and related arguments to use K-NN imputation, or ncp and related arguments to use PCA imputation. If both k and ncp are NULL, group$parameters must supply either k or ncp for every group.

Group-specific parameters in group$parameters take priority over global arguments. Global arguments fill in any missing values. All groups must use the same imputation method.

For method-specific arguments inherited from knn_imp() or pca_imp(), NULL means the backend default is used unless overridden by group$parameters.

Per-group k is capped at the number of usable columns in the group minus one. Per-group ncp is capped at the maximum feasible number of PCA components for that group's submatrix. A warning is issued when capping occurs.

Value

A numeric matrix of the same dimensions as obj, with missing values imputed. The returned object has class slideimp_results.

Parallelization

- K-NN: use the cores argument. If mirai daemons are active, cores is automatically set to 1 to avoid nested parallelism.
- PCA: use mirai::daemons() instead of cores.

When running PCA imputation in parallel with mirai, set pin_blas = TRUE in tune_imp() or group_imp() to prevent BLAS threads from oversubscribing CPU cores. This relies on RhpcBLASctl and works with OpenBLAS and MKL (typical on Linux, and on Windows after an OpenBLAS swap). pin_blas = TRUE may have no effect on macOS.

Performance tips

pca_imp() relies heavily on linear algebra. On Windows, the default BLAS shipped with R may be slow for large matrices. Advanced users can replace it with OpenBLAS.

PCA imputation speed depends on the eigensolver selected by solver and the convergence threshold threshold. The exact solver is selected with solver = "exact". The iterative LOBPCG solver is selected with solver = "lobpcg". The default, solver = "auto", performs a short timed probe and chooses LOBPCG only when it is clearly faster.

For large or approximately low-rank genomic matrices, it can be useful to benchmark solver = "exact" against solver = "lobpcg" on a representative subset, such as chromosome 22, before tuning accuracy-related parameters. For slide_imp(), this may include window_size and overlap_size.

The default threshold = 1e-6 is conservative. In many genomic datasets, threshold = 1e-5 can be faster while giving very similar imputed values. Check this on a representative subset before using the relaxed threshold in a full analysis.

See the pkgdown article Speeding up PCA imputation for a full workflow.

Note

A character scalar can be passed to `group` to name a supported Illumina platform, such as `"EPICv2"` or `"EPICv2_deduped"`. This requires the optional `slideimp.extra` package to be installed. Supported platforms are listed in the `slideimp_arrays` object in the `slideimp.extra` package.

See Also

`prep_groups()`, `knn_imp()`, `pca_imp()`

Examples

```
set.seed(1234)
to_test <- sim_mat(10, 20, perc_total_na = 0.05, perc_col_na = 1)
obj <- to_test$input
group <- to_test$col_group
head(group)

# Simple grouped K-NN imputation
results <- group_imp(obj, group = group, k = 2, .progress = FALSE)
results

# Impute only a subset of features
subset_features <- sample(to_test$col_group$feature, size = 10)
knn_subset <- group_imp(
  obj,
  group = group,
  subset = subset_features,
  k = 2,
  .progress = FALSE
)

# Use prep_groups() to inspect and edit per-group parameters
prepped <- prep_groups(colnames(obj), group)
prepped$parameters <- lapply(seq_len(nrow(prepped)), function(i) list(k = 2))
prepped$parameters[[2]]$k <- 4
knn_grouped <- group_imp(obj, group = prepped, .progress = FALSE)


# PCA imputation with mirai parallelism
mirai::daemons(2)
pca_grouped <- group_imp(obj, group = group, ncp = 2)
mirai::daemons(0)
pca_grouped
```

Description

Impute missing values in a numeric matrix using k-nearest neighbors (K-NN).

Usage

```
knn_imp(
  obj,
  k,
  colmax = 0.9,
  method = c("euclidean", "manhattan"),
  cores = 1,
  post_imp = TRUE,
  subset = NULL,
  dist_pow = 0,
  tree = FALSE,
  na_check = TRUE,
  .progress = FALSE
)
```

Arguments

obj	A numeric matrix with samples in rows and features in columns.
k	Integer. Number of nearest neighbors to use for K-NN imputation.
colmax	Numeric scalar between 0 and 1. Columns with a missing-data proportion greater than colmax are excluded from the main imputation method. Excluded columns are left unchanged unless post_imp = TRUE, in which case remaining missing values are replaced by column means when possible.
method	Character. K-NN imputation distance method: either "euclidean" or "manhattan".
cores	Integer. Number of cores to use for K-NN imputation. Defaults to 1.
post_imp	Logical. If TRUE, replace missing values remaining after the main imputation method with column means when possible.
subset	Optional character or integer vector specifying columns to target for imputation. If NULL, all eligible columns are targeted.
dist_pow	Numeric. Power used to penalize more distant neighbors in the weighted average. dist_pow = 0 gives an unweighted average of the nearest neighbors.
tree	Logical. If FALSE, use brute-force K-NN. If TRUE, use ball-tree K-NN via mlpack.
na_check	Logical. If TRUE, check whether the returned matrix still contains missing values.
.progress	Logical. If TRUE, show imputation progress.

Details

knn_imp() performs imputation column-wise, treating rows as observations and columns as features.

When `dist_pow > 0`, imputed values are computed as distance-weighted averages. Weights are inverse distances raised to the power of `dist_pow`.

If `tree = TRUE`, nearest neighbors are found with a ball tree via the `mlpack` package. This can be faster for some large, low-missingness data sets, but it requires initially filling missing values with column means, which can introduce bias when missingness is high.

Value

A numeric matrix of the same dimensions as `obj`, with missing values imputed. The returned object has class `slideimp_results`.

K-NN performance optimization

- `tree = FALSE` uses brute-force K-NN. This avoids the initial mean-filling step and is often faster for small to moderate datasets or high-dimensional data.
- `tree = TRUE` uses ball-tree K-NN. Consider this only when run time is prohibitive and missingness is low, for example less than 5%.
- Use `subset` when only specific columns need imputation.

References

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525. doi:10.1093/bioinformatics/17.6.520

Examples

```
set.seed(123)
obj <- sim_mat(20, 20, perc_col_na = 1)$input
sum(is.na(obj))

result <- knn_imp(obj, k = 10, .progress = FALSE)
result
```

---

lobpcg_control *LOBPCG Eigensolver Control Options*

Description

Construct a validated list of control options for the LOBPCG eigensolver used by `pca_imp()`. Most users do not need to call this directly.

Usage

```
lobpcg_control(warmup_iters = 10L, tol = 1e-09, maxiter = 20)
```

Arguments

warmup_iters	Integer. Number of warm-up iterations before the main LOBPCG solve. Must be non-negative.
tol	Numeric. Convergence tolerance for the LOBPCG eigensolver. Must be non-negative and finite.
maxiter	Integer. Maximum number of LOBPCG iterations. Must be non-negative. In pca_imp(), maxiter must be positive when solver = "auto" or solver = "lobpcg". Use solver = "exact" to force the exact solver.

Value

A named list of class "slideimp_lobpcg_control" containing warmup_iters, tol, and maxiter.

Examples

```
set.seed(123)
obj <- sim_mat(10, 10)$input

# Use all defaults
lobpcg_control()

# Override a single option
lobpcg_control(maxiter = 50)

# Force the exact solver from pca_imp()
pca_imp(obj, ncp = 2, solver = "exact")

# Pass directly to pca_imp()
pca_imp(obj, ncp = 2, lobpcg_control = lobpcg_control(tol = 1e-9))

# Or use a named list
pca_imp(obj, ncp = 2, lobpcg_control = list(maxiter = 50))
```

---

mat_miss *Column or Row Missing Counts and Proportions*

Description

Calculate the number or proportion of missing values per column or per row of a numeric matrix without allocating a full logical mask matrix.

Usage

```
mat_miss(obj, col = TRUE, prop = FALSE)
```

Arguments

obj	A numeric matrix.
col	Logical. If TRUE, compute column-wise. If FALSE, compute row-wise.
prop	Logical. If FALSE, return missing-value counts. If TRUE, return missing-value proportions.

Value

A numeric vector containing missing-value counts or proportions for columns or rows, named when the corresponding dimension names are present.

Examples

```
obj <- matrix(c(1, NA, 3, 4, NA, 6, NA, 8, 9), nrow = 3)
obj

# column missing counts
mat_miss(obj)

# row missing counts
mat_miss(obj, col = FALSE)

# column missing proportions
mat_miss(obj, prop = TRUE)
```

---

mean_imp_col *Column Mean Imputation*

Description

Impute missing values in a matrix by replacing them with the mean of their respective columns.

Usage

```
mean_imp_col(obj, subset = NULL, cores = 1)
```

Arguments

obj	A numeric matrix.
subset	Optional character or integer vector specifying columns to impute. If NULL, all columns are imputed.
cores	Integer. Number of cores to use for parallel computation. Defaults to 1.

Details

Columns with no observed values cannot be imputed by their column mean and are left unchanged.

Value

A numeric matrix of the same dimensions as obj, with missing values in the selected columns replaced by column means.

Examples

```
obj <- matrix(c(1, 2, NA, 4, NA, 6, NA, 8, 9, NA, NA, NA), nrow = 3)
colnames(obj) <- c("A", "B", "C", "D")
obj

# impute missing values with column means
mean_imp_col(obj)

# impute only specific columns by name
mean_imp_col(obj, subset = c("A", "C"))

# impute only specific columns by index
mean_imp_col(obj, subset = c(1, 3))
```

---

pca_imp *PCA Imputation for Numeric Matrices*

Description

Impute missing values in a numeric matrix using regularized or expectation-maximization PCA imputation.

Usage

```
pca_imp(
  obj,
  ncp = 2,
  scale = TRUE,
  method = c("regularized", "EM"),
  coeff.ridge = 1,
  row.w = NULL,
  threshold = 1e-06,
  seed = NULL,
  nb.init = 1,
  maxiter = 1000,
  miniter = 5,
  solver = c("auto", "exact", "lobpcg"),
  lobpcg_control = NULL,
  colmax = 0.9,
  post_imp = TRUE,
  na_check = TRUE,
  clamp = NULL
)
```

Arguments

obj	A numeric matrix with samples in rows and features in columns.
ncp	Integer. Number of principal components used to predict missing entries.
scale	Logical. If TRUE, columns are scaled to unit variance.
method	Character. PCA imputation method: either "regularized" or "EM".
coeff.ridge	Numeric. Ridge regularization coefficient. Only used when method = "regularized". Values less than 1 regularize less, moving closer to EM PCA. Values greater than 1 regularize more, moving closer to mean imputation.
row.w	Row weights, internally normalized to sum to 1. Can be: • NULL: all rows are weighted equally. • A numeric vector of positive weights with length nrow(obj). • "n_miss": rows with more missing values receive lower weight.
threshold	Numeric. Convergence threshold.
seed	Integer, numeric, or NULL. Random seed for reproducibility.
nb.init	Integer. Number of random initializations. The first initialization is always mean imputation.
maxiter	Integer. Maximum number of iterations.
miniter	Integer. Minimum number of iterations.
solver	Character. Eigensolver selection. One of "auto", "exact", or "lobpcg". "exact" uses the exact solver. "lobpcg" uses the iterative LOBPCG solver with exact fallback. "auto" performs a short timed probe and chooses LOBPCG only if it is clearly faster than the exact solver. When nb.init > 1, the auto choice from the first PCA initialization is reused for subsequent PCA initializations.
lobpcg_control	A list of LOBPCG eigensolver control options, usually created by lobpcg_control(). A plain named list is also accepted. Ignored when solver = "exact".
colmax	Numeric scalar between 0 and 1. Columns with a missing-data proportion greater than colmax are excluded from the main imputation method. Excluded columns are left unchanged unless post_imp = TRUE, in which case remaining missing values are replaced by column means when possible.
post_imp	Logical. If TRUE, replace missing values remaining after the main imputation method with column means when possible.
na_check	Logical. If TRUE, check whether the returned matrix still contains missing values.
clamp	Optional numeric vector of length 2 giving lower and upper bounds for PCA-imputed values. Use NULL for no clamping. Use c(0, 1) for DNA methylation beta values. Use c(lb, Inf) for only lower bound clamping, or c(-Inf, ub) for only upper bound clamping. Clamping is applied only to values imputed by the PCA step, not to observed values.

Details

This algorithm is based on missMDA::imputePCA() and is optimized for tall or wide numeric matrices.

Value

A numeric matrix of the same dimensions as `obj`, with missing values imputed. The returned object has class `slideimp_results`.

Performance tips

`pca_imp()` relies heavily on linear algebra. On Windows, the default BLAS shipped with R may be slow for large matrices. Advanced users can replace it with OpenBLAS.

PCA imputation speed depends on the eigensolver selected by `solver` and the convergence threshold `threshold`. The exact solver is selected with `solver = "exact"`. The iterative LOBPCG solver is selected with `solver = "lobpcg"`. The default, `solver = "auto"`, performs a short timed probe and chooses LOBPCG only when it is clearly faster.

For large or approximately low-rank genomic matrices, it can be useful to benchmark `solver = "exact"` against `solver = "lobpcg"` on a representative subset, such as chromosome 22, before tuning accuracy-related parameters. For `slide_imp()`, this may include `window_size` and `overlap_size`.

The default `threshold = 1e-6` is conservative. In many genomic datasets, `threshold = 1e-5` can be faster while giving very similar imputed values. Check this on a representative subset before using the relaxed threshold in a full analysis.

See the pkgdown article Speeding up PCA imputation for a full workflow.

References

Josse J, Husson F (2013). Handling missing values in exploratory multivariate data analysis methods. *Journal de la SFdS*, 153(2), 79-99.

Josse J, Husson F (2016). missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software*, 70(1), 1-31. doi:10.18637/jss.v070.i01

The PCA imputation algorithm is based on the original `missMDA` implementation by Francois Husson and Julie Josse.

Examples

```
set.seed(123)
obj <- sim_mat(10, 10)$input
sum(is.na(obj))
obj[1:4, 1:4]

# Randomly initialize missing values 5 times. The first initialization is
# mean imputation.
pca_imp(obj, ncp = 2, nb.init = 5, seed = 123)
```

prep_groups *Prepare Groups for Imputation*

Description

Normalize and validate a grouping specification for use with `group_imp()`.

Usage

```
prep_groups(
  obj_cn,
  group,
  subset = NULL,
  min_group_size = 0,
  allow_unmapped = FALSE,
  seed = NULL
)
```

Arguments

`obj_cn`	Character vector of column names from the data matrix, usually `colnames(obj)`.
`group`	Specification of how features should be grouped for imputation. Accepted formats are: • A character scalar naming a supported Illumina platform; see Note. • A long-format `data.frame` with columns `group` and `feature`. • A list-column `data.frame` with a `feature` list-column. Optional list-columns are `aux`, for auxiliary feature names, and `parameters`, for group-specific parameter lists.
`subset`	Optional character vector of feature names to impute. If `NULL`, all grouped features are imputed. Features in a group but not in `subset` are demoted to auxiliary columns for that group. Groups left with zero features after demotion are dropped with a message.
`min_group_size`	Integer or `NULL`. Minimum total number of columns per group, counting both features and auxiliary columns. Groups smaller than this are padded with randomly sampled columns from `obj_cn`. If `NULL` or `0`, no padding is performed.
`allow_unmapped`	Logical. If `FALSE`, every column in `colnames(obj)` must appear in `group`. If `TRUE`, columns with no group assignment are left untouched and are not used as auxiliary columns.
`seed`	Integer, numeric, or `NULL`. Random seed used when padding small groups.

Details

`prep_groups()` converts long-format or list-column group specifications into a validated `slideimp_tbl`, enforces feature and auxiliary-column set relationships, prunes dropped columns, and optionally pads small groups.

Let A be `obj_cn` and B be the union of all feature and auxiliary names in `group`. When `allow_unmapped = FALSE`, the function enforces $A \subseteq B$: every matrix column must appear somewhere in the grouping specification.

Elements in B but not in A, such as previously dropped probes, are pruned from each group. Groups left with zero features after pruning are removed with a diagnostic message.

Value

A data frame of class `slideimp_tbl` containing:

- `group`: original group labels, if provided, or sequential group labels.
- `feature`: a list-column of character vectors containing feature names.
- `aux`: a list-column of character vectors containing auxiliary names.
- `parameters`: a list-column of per-group configuration lists.

See Also

`group_imp()`

Examples

```
sim <- sim_mat(n = 10, p = 20)
prepped <- prep_groups(colnames(sim$input), sim$col_group)
prepped
```

---

print.slideimp_results

Print a `slideimp_results` *Object*

Description

Print the output of `knn_imp()`, `pca_imp()`, `group_imp()`, or `slide_imp()`.

Usage

```
## S3 method for class 'slideimp_results'
print(x, n = 6L, p = 6L, ...)
```

Arguments

`x`	A `slideimp_results` object.
`n`	Number of rows to print.
`p`	Number of columns to print.
`...`	Not used.

Value

x, invisibly.

Examples

```
set.seed(1234)
mat <- sim_mat(n = 10, p = 10)
result <- knn_imp(mat$input, k = 5, .progress = FALSE)
class(result)
print(result, n = 6, p = 6)
```

---

print.slideimp_sim *Print a* slideimp_sim *Object*

Description

Print the output of sim_mat().

Usage

```
## S3 method for class 'slideimp_sim'
print(x, n = 6L, p = 6L, ...)
```

Arguments

x	A slideimp_sim object.
n	Number of rows of each component to show.
p	Number of columns of input to show.
...	Not used.

Value

x, invisibly.

Examples

```
set.seed(123)
sim_data <- sim_mat(n = 50, p = 10, rho = 0.5)
class(sim_data)
print(sim_data)
```

print.slideimp_tbl *Print a* slideimp_tbl *Object*

Description

Print slideimp_tbl objects, which inherit from data.frame, with compact display of list-columns.

Usage

```
## S3 method for class 'slideimp_tbl'
print(x, n = NULL, ...)
```

Arguments

x	A slideimp_tbl object.
n	Number of rows to show. If NULL, a default is used.
...	Not used.

Value

x, invisibly.

Examples

```
sim <- sim_mat(n = 10, p = 20)
tbl <- prep_groups(colnames(sim$input), sim$col_group)
class(tbl)
print(tbl)
```

sample_na_loc *Sample Missing-Value Locations with Constraints*

Description

Sample matrix indices for NA injection while respecting row and column missingness limits and avoiding zero-variance columns.

Usage

```
sample_na_loc(
  obj,
  n_cols = NULL,
  n_rows = 2L,
  num_na = NULL,
  n_reps = 1L,
  rowmax = 0.9,
  colmax = 0.9,
  na_col_subset = NULL,
  max_attempts = 100
)
```

Arguments

obj	A numeric matrix.
n_cols	Integer or NULL. Number of columns to receive injected missing values. Must be supplied when num_na = NULL.
n_rows	Integer. Target number of missing values to inject per selected column.
num_na	Integer or NULL. Total number of missing values to inject per repetition. If supplied, n_cols is derived from num_na and n_rows, and missing values are distributed as evenly as possible across columns.
n_reps	Integer. Number of independent repetitions.
rowmax	Numeric scalar between 0 and 1. Maximum allowed missing-data proportion per row after injection.
colmax	Numeric scalar between 0 and 1. Maximum allowed missing-data proportion per column after injection.
na_col_subset	Optional integer or character vector restricting which columns are eligible for missing-value injection.
max_attempts	Integer. Maximum number of resampling attempts per repetition before giving up.

Details

The function uses a greedy stochastic search for valid NA locations. It ensures that:

- Total missingness per row and column does not exceed rowmax and colmax.
- At least two distinct observed values are preserved in every affected column.

Value

A list of length n_reps. Each element is a two-column integer matrix with row and column indices for sampled NA locations.

Examples

```
set.seed(123)
mat <- matrix(runif(100), nrow = 10)

# Sample 5 missing values across 5 columns
locs <- sample_na_loc(mat, n_cols = 5, n_rows = 1)
locs

# Inject the missing values from the first repetition
mat[locs[[1]]] <- NA
mat
```

`sim_mat` *Simulate a Matrix with Metadata*

Description

`sim_mat()` generates random normal data with optional compound-symmetric column correlation. Values can optionally be scaled to the interval `[0, 1]` column-wise. The function also creates feature metadata for columns and sample metadata for rows, and can inject `NA` values into a specified proportion of matrix cells across a specified proportion of columns.

Usage

```
sim_mat(
  n = 100,
  p = 100,
  rho = 0.5,
  n_col_groups = 2,
  n_row_groups = 1,
  perc_total_na = 0.1,
  perc_col_na = 0.5,
  beta = TRUE
)
```

Arguments

`n`	Integer. Number of rows, interpreted as samples. Defaults to `100`.
`p`	Integer. Number of columns, interpreted as features. Defaults to `100`.
`rho`	Numeric. Compound-symmetric column correlation before optional scaling. Defaults to `0.5`.
`n_col_groups`	Integer. Number of groups to assign to features. Defaults to `2`.
`n_row_groups`	Integer. Number of groups to assign to samples. Defaults to `1`.
`perc_total_na`	Numeric scalar between `0` and `1`. Proportion of all matrix cells to set to `NA`. Defaults to `0.1`.

perc_col_na	Numeric scalar between 0 and 1. Proportion of columns across which injected NA values are spread. Defaults to 0.5.
beta	Logical. If TRUE, scale values to the interval [0, 1] column-wise.

Details

Generate a numeric matrix with optional row and column metadata and added missing values.

Value

An object of class slideimp_sim. This is a list containing:

- input: a numeric matrix of dimension $n \times p$ containing the simulated values and injected missing values.
- col_group: a data frame with p rows mapping each feature to a group.
- row_group: a data frame with n rows mapping each sample to a group.

Examples

```
set.seed(123)
sim_data <- sim_mat(n = 50, p = 10, rho = 0.5)
sim_data
```

---

slideimp_resolve_group

Resolve a Group Specification to a Data Frame

Description

Convert a group specification to the canonical data-frame form expected by prep_groups(). This S3 generic is exported so that extension packages, such as slideimp.extra, can register additional methods.

Usage

```
slideimp_resolve_group(x)

## S3 method for class 'data.frame'
slideimp_resolve_group(x)

## Default S3 method:
slideimp_resolve_group(x)
```

Arguments

x	A group specification. slideimp provides a method for data.frame objects. The optional slideimp.extra package provides a method for character platform names.

Note

This is primarily an extension hook for `slideimp.extra`.

Examples

```
df <- data.frame(feature = c("cg1", "cg2"), group = c(1, 1))
slideimp_resolve_group(df)
```

`slide_imp` *Sliding-Window K-NN or PCA Imputation*

Description

Perform sliding-window K-NN or PCA imputation on a numeric matrix whose columns are meaningfully ordered.

Usage

```
slide_imp(
  obj,
  location,
  window_size,
  overlap_size = 0,
  flank = FALSE,
  min_window_n,
  subset = NULL,
  dry_run = FALSE,
  k = NULL,
  cores = 1,
  dist_pow = 0,
  ncp = NULL,
  scale = TRUE,
  coeff.ridge = 1,
  threshold = 1e-06,
  seed = NULL,
  row.w = NULL,
  nb.init = 1,
  maxiter = 1000,
  miniter = 5,
  solver = c("auto", "exact", "lobpcg"),
  lobpcg_control = NULL,
  clamp = NULL,
  method = NULL,
  .progress = TRUE,
  colmax = 0.9,
  post_imp = TRUE,
```

```
  na_check = TRUE,
  on_infeasible = c("skip", "error", "mean")
)
```

Arguments

obj	A numeric matrix with samples in rows and features in columns.
location	A sorted numeric vector of length ncol(obj) giving the position of each column, such as genomic coordinates.
window_size	Numeric. Window width in the same units as location.
overlap_size	Numeric. Overlap between consecutive windows in the same units as location. Must be less than window_size. Ignored when flank = TRUE.
flank	Logical. If FALSE, imputation uses sliding windows across the full matrix. If TRUE, one window of width window_size is created for each feature listed in subset; overlap_size is ignored. subset must be supplied when flank = TRUE.
min_window_n	Integer. Minimum number of columns a window must contain to be considered for imputation. For non-dry runs, the selected k or ncp value must be feasible for the usable columns in each retained window after applying colmax.
subset	Optional character or integer vector specifying columns to impute. If NULL, all eligible columns are imputed. Required when flank = TRUE.
dry_run	Logical. If TRUE, skip imputation and return a slideimp_tbl describing the windows that would be used after all filtering rules are applied. In this mode, k and ncp are not required.
k	Integer. Number of nearest neighbors to use for K-NN imputation.
cores	Integer. Number of cores to use for K-NN imputation.
dist_pow	Numeric. Power used to penalize more distant neighbors in the weighted average. dist_pow = 0 gives an unweighted average of the nearest neighbors.
ncp	Integer. Number of principal components used to predict missing entries.
scale	Logical. If TRUE, columns are scaled to unit variance.
coeff.ridge	Numeric. Ridge regularization coefficient. Only used when method = "regularized". Values less than 1 regularize less, moving closer to EM PCA. Values greater than 1 regularize more, moving closer to mean imputation.
threshold	Numeric. Convergence threshold.
seed	Integer, numeric, or NULL. Random seed for reproducibility.
row.w	Row weights, internally normalized to sum to 1. Can be: • NULL: all rows are weighted equally. • A numeric vector of positive weights with length nrow(obj). • "n_miss": rows with more missing values receive lower weight.
nb.init	Integer. Number of random initializations. The first initialization is always mean imputation.
maxiter	Integer. Maximum number of iterations.

miniter	Integer. Minimum number of iterations.
solver	Character. Eigensolver selection. One of "auto", "exact", or "lobpcg". "exact" uses the exact solver. "lobpcg" uses the iterative LOBPCG solver with exact fallback. "auto" performs a short timed probe and chooses LOBPCG only if it is clearly faster than the exact solver. When nb.init > 1, the auto choice from the first PCA initialization is reused for subsequent PCA initializations.
lobpcg_control	A list of LOBPCG eigensolver control options, usually created by lobpcg_control(). A plain named list is also accepted. Ignored when solver = "exact".
clamp	Optional numeric vector of length 2 giving lower and upper bounds for PCA-imputed values. Use NULL for no clamping. Use c(0, 1) for DNA methylation beta values. Use c(lb, Inf) for only lower bound clamping, or c(-Inf, ub) for only upper bound clamping. Clamping is applied only to values imputed by the PCA step, not to observed values.
method	Character or NULL. For K-NN imputation, one of "euclidean" or "manhattan". For PCA imputation, one of "regularized" or "EM". If NULL, the corresponding backend default is used.
.progress	Logical. If TRUE, show imputation progress.
colmax	Numeric scalar between 0 and 1. Columns with a missing-data proportion greater than colmax are excluded from the main imputation method. Excluded columns are left unchanged unless post_imp = TRUE, in which case remaining missing values are replaced by column means when possible.
post_imp	Logical. If TRUE, replace missing values remaining after the main imputation method with column means when possible.
na_check	Logical. If TRUE, check whether the returned matrix still contains missing values.
on_infeasible	Character. One of "skip", "error", or "mean". Controls behavior when a window is infeasible for imputation, for example when k or ncp exceeds the number of usable columns after applying colmax.

Details

The sliding-window approach divides the input matrix into smaller segments based on location values and applies imputation to each window independently. Values in overlapping regions are averaged across windows to produce the final imputed result.

Two windowing modes are supported:

- flank = FALSE: greedily partition location into windows of width window_size with the requested overlap_size between consecutive windows.
- flank = TRUE: create one window per feature in subset, centered on that feature using the supplied window_size.

Specify k and related arguments to use knn_imp(), or ncp and related arguments to use pca_imp().

Value

If dry_run = FALSE, a numeric matrix of the same dimensions as obj, with missing values imputed. The returned object has class slideimp_results.

If dry_run = TRUE, a data frame of class slideimp_tbl with columns start, end, and window_n, plus subset_local and, when flank = TRUE, target.

Performance tips

pca_imp() relies heavily on linear algebra. On Windows, the default BLAS shipped with R may be slow for large matrices. Advanced users can replace it with OpenBLAS.

PCA imputation speed depends on the eigensolver selected by solver and the convergence threshold threshold. The exact solver is selected with solver = "exact". The iterative LOBPCG solver is selected with solver = "lobpcg". The default, solver = "auto", performs a short timed probe and chooses LOBPCG only when it is clearly faster.

For large or approximately low-rank genomic matrices, it can be useful to benchmark solver = "exact" against solver = "lobpcg" on a representative subset, such as chromosome 22, before tuning accuracy-related parameters. For slide_imp(), this may include window_size and overlap_size.

The default threshold = 1e-6 is conservative. In many genomic datasets, threshold = 1e-5 can be faster while giving very similar imputed values. Check this on a representative subset before using the relaxed threshold in a full analysis.

See the pkgdown article Speeding up PCA imputation for a full workflow.

Examples

```
set.seed(1234)

# Example data with 20 samples and 100 ordered columns
beta_matrix <- sim_mat(20, 100)$input
location <- 1:100

# First perform a dry run to inspect the calculated windows
window_statistics <- slide_imp(
  beta_matrix,
  location = location,
  window_size = 50,
  overlap_size = 10,
  min_window_n = 10,
  dry_run = TRUE,
  .progress = FALSE
)
window_statistics

# Sliding-window K-NN imputation
imputed_knn <- slide_imp(
  beta_matrix,
  location = location,
  k = 5,
  window_size = 50,
```

```
  overlap_size = 10,
  min_window_n = 10,
  .progress = FALSE
)
imputed_knn

# Sliding-window PCA imputation
imputed_pca <- slide_imp(
  beta_matrix,
  location = location,
  ncp = 2,
  window_size = 50,
  overlap_size = 10,
  min_window_n = 10,
  .progress = FALSE
)
imputed_pca

# K-NN imputation with flanking windows
imputed_flank <- slide_imp(
  beta_matrix,
  location = location,
  k = 2,
  window_size = 30,
  flank = TRUE,
  subset = c(10, 30, 70),
  min_window_n = 5,
  .progress = FALSE
)
imputed_flank
```

---

tune_imp *Tune Imputation Method Parameters*

Description

Tune imputation-method parameters by repeatedly masking observed values, imputing them, and comparing the imputed values with the original values.

Usage

```
tune_imp(
  obj,
  parameters = NULL,
  .f,
  na_loc = NULL,
  num_na = NULL,
  n_reps = 1,
```

```
  n_cols = NULL,
  n_rows = 2,
  rowmax = 0.9,
  colmax = 0.9,
  na_col_subset = NULL,
  max_attempts = 100,
  .progress = TRUE,
  cores = 1,
  location = NULL,
  pin_blas = FALSE
)
```

Arguments

obj	A numeric matrix.
parameters	A data.frame specifying parameter combinations to tune. Each column should be a parameter accepted by .f, excluding obj. List-columns are supported for complex parameters. Duplicate rows are removed. NULL is treated as a single parameter set with no additional arguments, which is useful for functions whose required arguments all have defaults.
.f	One of "knn_imp", "pca_imp", or "slide_imp", or a custom imputation function.
na_loc	Optional predefined missing-value locations. Accepted formats are a two-column integer matrix of row and column indices, a numeric vector of linear positions, or a list whose elements are either of those formats.
num_na	Integer or NULL. Total number of missing values to inject per repetition. If supplied, n_cols is derived from num_na and n_rows, and missing values are distributed as evenly as possible across columns. Ignored when na_loc is supplied.
n_reps	Integer. Number of independent repetitions.
n_cols	Integer or NULL. Number of columns to receive injected missing values per repetition. Must be supplied when both num_na and na_loc are NULL, unless the automatic default applies. Ignored when num_na or na_loc is supplied.
n_rows	Integer. Target number of missing values to inject per selected column. Ignored when na_loc is supplied.
rowmax	Numeric scalar between 0 and 1. Maximum allowed missing-data proportion per row after injection.
colmax	Numeric scalar between 0 and 1. Maximum allowed missing-data proportion per column after injection.
na_col_subset	Optional integer or character vector restricting which columns are eligible for random missing-value injection. Ignored when na_loc is supplied.
max_attempts	Integer. Maximum number of resampling attempts per repetition before giving up.
.progress	Logical. If TRUE, show progress during tuning.
cores	Integer. Number of cores to use for K-NN and sliding-window K-NN imputation. For other methods, use mirai::daemons().

location	Numeric vector of column locations. Required when .f = "slide_imp".
pin_blas	Logical. If TRUE, pin BLAS threads to 1 during parallel tuning to reduce thread contention.

Details

Built-in methods can be selected by passing .f = "knn_imp", .f = "pca_imp", or .f = "slide_imp". A custom function can also be supplied. Custom functions must accept obj as their first argument and return a numeric matrix with the same dimensions as obj.

When .f is a character string, columns in parameters are validated against the selected method:

- "knn_imp" requires k.
- "pca_imp" requires ncp.
- "slide_imp" requires window_size, overlap_size, and min_window_n, plus exactly one of k or ncp.

To tune parameters for grouped imputation, tune knn_imp() or pca_imp() on representative groups, then pass the selected parameters to group_imp().

The top-level rowmax and colmax arguments control random missing-value injection performed by sample_na_loc(). To tune or pass an imputation method's own colmax argument, include a colmax column in parameters.

Tuning results can be summarized with compute_metrics() or evaluated with external packages such as yardstick.

Value

A data frame of class slideimp_tune containing:

- columns originally provided in parameters;
- param_set, an integer ID for each unique parameter combination;
- rep_id, an integer repetition index;
- result, a list-column where each element is a data frame containing truth and estimate columns;
- error, a character column containing the error message if the iteration failed, otherwise NA.

Parallelization

- K-NN: use the cores argument. If mirai daemons are active, cores is automatically set to 1 to avoid nested parallelism.
- PCA: use mirai::daemons() instead of cores.

When running PCA imputation in parallel with mirai, set pin_blas = TRUE in tune_imp() or group_imp() to prevent BLAS threads from oversubscribing CPU cores. This relies on RhpcBLASctl and works with OpenBLAS and MKL (typical on Linux, and on Windows after an OpenBLAS swap). pin_blas = TRUE may have no effect on macOS.

Performance tips

`pca_imp()` relies heavily on linear algebra. On Windows, the default BLAS shipped with R may be slow for large matrices. Advanced users can replace it with OpenBLAS.

PCA imputation speed depends on the eigensolver selected by `solver` and the convergence threshold `threshold`. The exact solver is selected with `solver = "exact"`. The iterative LOBPCG solver is selected with `solver = "lobpcg"`. The default, `solver = "auto"`, performs a short timed probe and chooses LOBPCG only when it is clearly faster.

For large or approximately low-rank genomic matrices, it can be useful to benchmark `solver = "exact"` against `solver = "lobpcg"` on a representative subset, such as chromosome 22, before tuning accuracy-related parameters. For `slide_imp()`, this may include `window_size` and `overlap_size`.

The default `threshold = 1e-6` is conservative. In many genomic datasets, `threshold = 1e-5` can be faster while giving very similar imputed values. Check this on a representative subset before using the relaxed threshold in a full analysis.

See the pkgdown article Speeding up PCA imputation for a full workflow.

Examples

```
set.seed(123)

# Simulate some data
obj <- sim_mat(10, 50)$input

# Tune K-NN imputation with random missing-value injection.
# Use larger `num_na` and `n_reps` values for real analyses.
params_knn <- data.frame(k = c(2, 4))
results <- tune_imp(
  obj,
  params_knn,
  .f = "knn_imp",
  n_reps = 1,
  num_na = 10,
  .progress = FALSE
)
compute_metrics(results)

# Tune with fixed missing-value positions
na_positions <- list(
  matrix(c(1, 2, 3, 1, 1, 1), ncol = 2),
  matrix(c(2, 3, 4, 2, 2, 2), ncol = 2)
)

results_fixed <- tune_imp(
  obj,
  data.frame(k = 2),
  .f = "knn_imp",
  na_loc = na_positions,
  .progress = FALSE
)
```

```
# Custom imputation function
custom_fill <- function(obj, val = 0) {
  obj[is.na(obj)] <- val
  obj
}

tune_imp(
  obj,
  data.frame(val = c(0, 1)),
  .f = custom_fill,
  num_na = 10,
  .progress = FALSE
)


# Parallel tuning with mirai
mirai::daemons(2)

parameters_custom <- data.frame(mean = c(0, 1), sd = c(1, 1))

custom_imp <- function(obj, mean, sd) {
  na_pos <- is.na(obj)
  obj[na_pos] <- stats::rnorm(sum(na_pos), mean = mean, sd = sd)
  obj
}

results_p <- tune_imp(
  obj,
  parameters_custom,
  .f = custom_imp,
  n_reps = 1,
  num_na = 10,
  .progress = FALSE
)

mirai::daemons(0)
```

Index