

Package: MIML (via r-universe)

October 1, 2026

Type Package

Title Machine Learning Imputation, Clustering and Survival Analysis
for Longitudinal Proteomic Data

Version 0.1.0

Description Imputes missing biomarker measurements in a wide longitudinal serum panel with gradient-boosted decision trees, groups the completed panel by Bayesian consensus clustering, and compares the resulting patient subgroups by Kaplan-Meier, log-rank and Cox analysis. The imputation learner is described in Ke et al. (2017) <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree> and the clustering method in Lock and Dunson (2013) [doi:10.1093/bioinformatics/btt425](https://doi.org/10.1093/bioinformatics/btt425). Imputed values are conditional-mean predictions, so the procedure is a machine-learning single imputation; the completions carry no between-imputation variance and must not be pooled by Rubin's rules. Two panels from Gene Expression Omnibus accession 'GSE65622' are included, one for each survival endpoint.

License GPL-3

Encoding UTF-8

Depends R (>= 4.1.0)

Imports lightgbm (>= 3.3.0), data.table, survival, stats, utils

Suggests BCClong, testthat (>= 3.0.0)

Config/testthat/edition 3

LazyData true

LazyDataCompression xz

Config/roxygen2/version 8.1.0

NeedsCompilation no

Author Neelesh Kumar [aut, cre], Atanu Bhattacharjee [aut], Gajendra K. Vishwakarma [aut], Tanmoy Majumdar [aut]

Maintainer Neelesh Kumar <neeleesh2302@gmail.com>

Repository <https://neeleesh2302-creator.r-universe.dev>
Date/Publication 2026-09-30 13:54:55 UTC
RemoteUrl <https://github.com/cran/MIML>
RemoteRef HEAD
RemoteSha 9eb513f496b9081139542bf69d98ffdacb4b5c8f

Contents

MFS_data	2
miml_mfs	3
miml_rfs	4
RFS_data	6

Index	8
--------------	----------

MFS_data	<i>Metastasis-free survival panel (GSE65622)</i>
----------	--

Description

The same patients, visits and 507 biomarkers as [RFS_data](#), paired instead with metastasis-free survival. The default input to [miml_mfs\(\)](#).

Usage

MFS_data

Format

A data frame of 288 rows and 516 columns, identical in structure to [RFS_data](#) except that the survival column is MFS and event is the metastasis indicator. Twenty-six of the 80 patients record a metastasis.

Source

Gene Expression Omnibus accession [GSE65622](#), released April 2016.

See Also

[RFS_data](#), [miml_mfs\(\)](#)

Examples

```
data(MFS_data)
dim(MFS_data)
MFS_data[1:5, 1:6]
```

miml_mfs

*Metastasis-free survival analysis of the GSE65622 panel***Description**

Identical to `miml_rfs()` but run on `MFS_data` against metastasis-free survival. Twenty-six of the eighty patients record a metastasis, against five recurrences for the other endpoint, so this is the better powered of the two comparisons.

Usage

```
miml_mfs(
  data = NULL,
  m = 5L,
  maxit = 10L,
  nrounds = 200L,
  n_clusters = 2L,
  n_bcc_markers = 3L,
  predictors = c("manuscript", "all"),
  seed = 123L,
  verbose = TRUE
)
```

Arguments

<code>data</code>	Panel to analyse. Defaults to <code>MFS_data</code> .
<code>m</code>	Number of completed datasets. The published analysis used 5.
<code>maxit</code>	Refit cycles per completion. The published analysis used 10.
<code>nrounds</code>	Maximum boosting rounds. The published analysis used 200.
<code>n_clusters</code>	Number of consensus clusters. The published analysis used 2.
<code>n_bcc_markers</code>	Biomarkers carried into the clustering model, chosen by descending variance. The published analysis used 3.
<code>predictors</code>	Predictor pool for the imputation models. "manuscript" uses the patient identifier, the survival endpoint, the event indicator and the six phenotype columns, which is what the published analysis did; "all" lets biomarkers predict each other and drops the outcome from the pool. See the leakage note below.
<code>seed</code>	Random seed. The caller's RNG state is restored on exit.
<code>verbose</code>	Print progress.

Value

An object of class `miml_analysis`, as for `miml_rfs()`.

Predictor pool and outcome leakage

Under `predictors = "manuscript"` the survival time and the event indicator are inside the predictor pool, and no biomarker predicts another. Any association later estimated between the imputed values and survival is therefore inflated by construction. The default reproduces the published numbers; use `predictors = "all"` for new work.

Run time

At the defaults the imputation fits 507 biomarkers over 10 cycles for each of 5 completions and takes roughly an hour.

See Also

[miml_rfs\(\)](#), [MFS_data](#)

Examples

```
data(MFS_data)

small <- MFS_data[1:60, c("Patient_ID", "MFS", "event",
                        names(MFS_data)[4:6])]
miml_mfs(small, m = 1, maxit = 1, nrounds = 3,
         n_clusters = 0, verbose = FALSE)

sub <- MFS_data[, c("Patient_ID", "MFS", "event", names(MFS_data)[4:23])]
res <- miml_mfs(sub, m = 1, maxit = 2, nrounds = 20)
res

## the published analysis is miml_mfs() at its defaults.
```

miml_rfs

Recurrence-free survival analysis of the GSE65622 panel

Description

Runs the published MIML workflow on [RFS_data](#): LightGBM imputation of every incomplete biomarker, Bayesian consensus clustering of the completed panel, and a Kaplan-Meier, log-rank and Cox comparison of the resulting clusters against recurrence-free survival.

Usage

```
miml_rfs(
  data = NULL,
  m = 5L,
  maxit = 10L,
```

```

nrounds = 200L,
n_clusters = 2L,
n_bcc_markers = 3L,
predictors = c("manuscript", "all"),
seed = 123L,
verbose = TRUE
)

```

Arguments

<code>data</code>	Panel to analyse. Defaults to RFS_data .
<code>m</code>	Number of completed datasets. The published analysis used 5.
<code>maxit</code>	Refit cycles per completion. The published analysis used 10.
<code>nrounds</code>	Maximum boosting rounds. The published analysis used 200.
<code>n_clusters</code>	Number of consensus clusters. The published analysis used 2.
<code>n_bcc_markers</code>	Biomarkers carried into the clustering model, chosen by descending variance. The published analysis used 3.
<code>predictors</code>	Predictor pool for the imputation models. "manuscript" uses the patient identifier, the survival endpoint, the event indicator and the six phenotype columns, which is what the published analysis did; "all" lets biomarkers predict each other and drops the outcome from the pool. See the leakage note below.
<code>seed</code>	Random seed. The caller's RNG state is restored on exit.
<code>verbose</code>	Print progress.

Details

Imputed values are the conditional-mean predictions of the fitted LightGBM models. No residual is drawn and no donor is sampled, so the completions carry no between-imputation variance and **must not be combined by Rubin's rules**. This is machine-learning single imputation.

Value

An object of class `miml_analysis`: a list with `endpoint`, `missingness`, `completed`, `clusters`, `survival` and `settings`.

Predictor pool and outcome leakage

Under `predictors = "manuscript"` the survival time and the event indicator are inside the predictor pool, and no biomarker predicts another. Any association later estimated between the imputed values and survival is therefore inflated by construction. The default reproduces the published numbers; use `predictors = "all"` for new work.

Run time

At the defaults the imputation fits 507 biomarkers over 10 cycles for each of 5 completions and takes roughly an hour.

See Also

[miml_mfs\(\)](#), [RFS_data](#)

Examples

```
data(RFS_data)

## small and fast: three biomarkers, one cycle, no clustering
small <- RFS_data[1:60, c("Patient_ID", "RFS", "event",
                        names(RFS_data)[4:6])]
miml_rfs(small, m = 1, maxit = 1, nrounds = 3,
         n_clusters = 0, verbose = FALSE)

## twenty biomarkers, two cycles, clustered and compared by survival
sub <- RFS_data[, c("Patient_ID", "RFS", "event", names(RFS_data)[4:23])]
res <- miml_rfs(sub, m = 1, maxit = 2, nrounds = 20)
res

## the published analysis is miml_rfs() at its defaults: 507 biomarkers,
## 10 cycles, 5 completions, roughly an hour.
```

RFS_data

Recurrence-free survival panel (GSE65622)

Description

Longitudinal serum proteomic measurements from patients with locally advanced rectal cancer undergoing multimodal treatment, paired with recurrence-free survival. The default input to [miml_rfs\(\)](#).

Usage

```
RFS_data
```

Format

A data frame of 288 rows and 516 columns: 80 patients observed at three or four sampling points each. Patient_ID is an integer identifier repeated across a patient's visits; RFS is recurrence-free survival in days measured *from that visit*, so it decreases across a patient's rows; event is the recurrence indicator, constant within a patient. Then 507 protein biomarkers named as in the original submission, then six phenotype columns: diarrhea_grade, sampling_point, slide_print_batch, trg_score, ypn_stage, ypt_stage.

Biomarker names are not syntactic R names, so quote them or use `[[ ]]`. About 10.3 percent of the biomarker measurements are missing. Five of the 80 patients record a recurrence.

Source

Gene Expression Omnibus accession [GSE65622](#), released April 2016.

See Also

[MFS_data](#), [miml_rfs\(\)](#)

Examples

```
data(RFS_data)
dim(RFS_data)
RFS_data[1:5, 1:6]
```

Index

* **datasets**

MFS_data, [2](#)

RFS_data, [6](#)

MFS_data, [2](#), [3](#), [4](#), [7](#)

miml_mfs, [3](#)

miml_mfs(), [2](#), [6](#)

miml_rfs, [4](#)

miml_rfs(), [3](#), [4](#), [6](#), [7](#)

RFS_data, [2](#), [4](#), [5](#), [6](#), [6](#)